ORIGINAL RESEARCH **Open access**

Attention-Based Multiple Instance Learning for Ovarian Cancer Survival Prediction: A Framework Using Whole-Slide Histopathology without Pixel-Level Annotations

Victor Santos[1*], Rafael Costa[1], Bruno Teixeira[2]

Abstract

Ovarian cancer, particularly high-grade serous carcinoma, is highly lethal, and accurate survival prediction is essential for treatment planning. However, traditional prognostic models rely on limited clinical and histologic features, while deep learning approaches require expensive pixel-level annotations of whole-slide histopathology images, limiting scalability. We propose a weakly supervised attention-based multiple instance learning (MIL) framework that predicts ovarian cancer survival using only slide-level survival labels. Each whole-slide image is treated as a bag of patches, where a patch encoder extracts features using a pre-trained CNN or vision transformer. An attention-based MIL aggregator assigns importance weights to patches, and a survival head outputs a risk score via a deep Cox model. The attention mechanism enhances interpretability by identifying prognostically relevant regions such as aggressive tumor morphology, stromal patterns, and immune infiltration. This reduces the need for manual annotation while preserving clinical relevance. The framework provides a scalable and interpretable approach for survival prediction and can be evaluated on datasets such as TCGA-OV and CPTAC for clinical translation.

Keywords Attention mechanism, Multiple instance learning, Whole-slide histopathology, Ovarian cancer, Survival prediction, Weakly supervised learning

*Correspondence:
Victor Santos
victor.santos@gmail.com
[1] Department of Clinical AI Systems, University of Sao Paulo, Sao Paulo, Brazil
[2] Department of Healthcare Intelligence Engineering, University of Campinas, Campinas, Brazil

Introduction

Ovarian cancer accounts for approximately 300,000 new cases and 200,000 deaths annually worldwide, with high-grade serous carcinoma representing the most common and aggressive subtype. Despite advances in debulking surgery and platinum-based chemotherapy, five-year survival rates remain below 45% due to late-stage diagnosis and intrinsic or acquired chemoresistance [1, 2]. Histopathological assessment of tumor architecture, nuclear grade, mitotic count, and stromal response provides prognostic information that complements clinical staging, but manual evaluation is subjective and semiquantitative [3, 4].

Deep learning has emerged as a powerful tool for extracting prognostic signatures from whole-slide images (WSIs), but most supervised approaches require pixel-level annotations of tumor regions, stroma, and necrosis. Generating such annotations at the gigapixel scale is prohibitively expensive, requiring hundreds of hours of pathologist time per dataset, and introduces inter-observer variability that degrades model generalizability [5, 6]. The requirement for fine-grained

labels fundamentally limits the scalability of supervised methods to large multicenter cohorts where only slide-level diagnoses and outcomes are routinely available [7, 8].

Multiple instance learning (MIL) offers a compelling alternative by treating each WSI as a bag of patches (instances) with a single slide-level label, thereby eliminating the need for patch-level annotations. In the MIL paradigm, a WSI is positive for a given label if at least one patch contains the relevant feature, and negative otherwise, enabling weakly supervised classification without pixel-wise supervision [9, 10]. For survival prediction, each slide is associated with a censored survival time, and the MIL aggregator learns to identify patches whose features correlate with patient outcomes [7, 11].

This article presents a conceptual framework for attention-based MIL specifically designed for ovarian cancer survival prediction from WSIs without pixel-level annotations. The framework integrates three key innovations: a patch encoder that converts histopathology patches into compact feature vectors using pre-trained convolutional or transformer architectures; a gated attention mechanism that learns to weight patches by prognostic relevance; and a deep Cox survival head that outputs risk scores from aggregated bag embeddings [2, 12]. The following sections detail the background, architecture, training considerations, and evaluation strategies for this framework.

Background

Ovarian cancer histopathology

Ovarian cancer encompasses multiple histological subtypes with distinct molecular profiles, clinical behaviors, and prognostic implications. High-grade serous carcinoma (HGSC) accounts for approximately 70% of cases and is characterized by marked nuclear atypia, high mitotic index, and characteristic slit-like spaces, with most patients presenting at advanced stage (III/IV) [1, 2]. Clear cell carcinoma (10%) exhibits hobnail cells and hyaline globules, often associated with endometriosis and paradoxically poor prognosis in early-stage disease, while endometrioid (10%) and mucinous (3%) subtypes have more favorable outcomes when diagnosed early [3, 4].

Prognostic histologic features in ovarian cancer include tumor budding, desmoplastic stromal reaction, tumor-infiltrating lymphocytes (TILs), and necrosis, all of which vary substantially across regions within a single WSI. Manual assessment of these features requires exhaustive slide review and suffers from moderate inter-rater agreement, motivating automated approaches that can systematically quantify morphological heterogeneity [1, 7]. The spatial distribution of prognostic features—for example, TILs at the invasive front versus tumor core—further complicates manual assessment but can be captured by patch-based MIL models.

Survival prediction in oncology

Survival prediction in oncology typically employs the Cox proportional hazards model, which assumes that covariates multiplicatively shift the baseline hazard function without requiring specification of its shape. The model estimates hazard ratios for clinical variables such as age, stage, and residual disease status, with the concordance index (C-index) measuring the probability that a model correctly ranks the survival times of two randomly selected patients [2, 8]. However, clinical variables alone explain only a fraction of survival variance, motivating the integration of high-dimensional histopathology data.

Deep survival models extend the Cox framework by replacing the linear risk score with a neural network that learns nonlinear transformations of input features. For WSI-based prediction, the challenge lies in aggregating millions of patch-level features into a slide-level representation that preserves prognostic information while handling censored observations [7, 11]. MIL formulations naturally accommodate this structure by treating each WSI as a bag from which a risk score is derived, with the partial likelihood loss providing unbiased gradient estimates for censored data [9, 12].

Multiple instance learning

MIL operates under the assumption that labels are available for bags of instances rather than for individual instances, with the bag label determined by the presence of at least one positive instance in binary classification (standard MIL) or by aggregation of instance-level signals in more general settings. For survival prediction, the bag label is not binary but rather a censored continuous outcome, requiring the MIL aggregator to produce a bag-level risk score that reflects the maximum or weighted average of instance-level prognostic signals [9, 11]. Traditional MIL aggregators include max pooling (selecting the instance with highest risk score), mean pooling (averaging all instances), and noisy-or pooling, each with distinct bias-variance tradeoffs [10].

Attention-based MIL replaces fixed pooling operations with a learnable weighted average where attention scores are computed via a small neural network applied to each instance feature vector [9]. The attention mechanism allows the model to focus on prognostically informative patches while downweighting irrelevant regions such as background, necrosis, or normal tissue. Unlike max pooling, which discards all but one instance, attention-based aggregation preserves information from multiple relevant patches, which is critical for survival prediction where the outcome reflects cumulative tumor burden rather than a single extreme region [1, 7].

Framework Overview

High-level architecture

The proposed framework processes a whole-slide image through four sequential stages: patch extraction, patch encoding, attention-based MIL aggregation, and survival prediction. First, the WSI is tiled into non-overlapping or overlapping patches at a fixed magnification (typically 20× or 40×), with tissue detection removing background areas containing no cellular material [8, 12]. Second, each patch is passed through a pre-trained feature extractor (e.g., ResNet50, EfficientNet-B4, or a vision transformer) that outputs a low-dimensional feature vector, typically 256 to 512 dimensions, capturing morphological characteristics such as nuclear shape, chromatin texture, and glandular architecture [13, 14].

Third, the set of patch feature vectors is input to an attention-based MIL module that computes a single bag embedding as the attention-weighted sum of patch features. The attention weights are learned via a gated attention mechanism that can capture complex, nonlinear relationships between patch features and survival outcomes [1, 9]. Fourth, the bag embedding is passed through a survival prediction head—typically a multilayer perceptron with one or two hidden layers—that outputs a risk score, which is optimized using the Cox partial likelihood loss function that accounts for right-censored survival data [7, 12].

Figure 1 shows the proposed framework organizing weakly supervised ovarian cancer survival prediction as a strictly sequential pipeline that links whole-slide patch extraction, feature encoding, attention-based multiple instance aggregation, survival risk modeling, and clinically interpretable validation outputs.

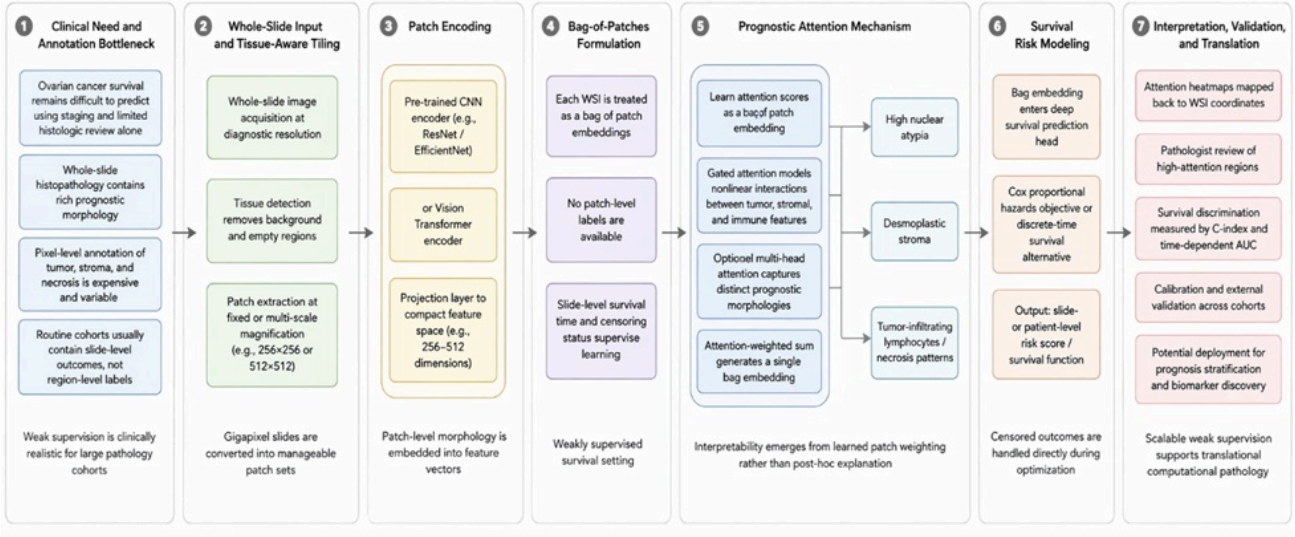

Figure 1. Conceptual Architecture of Attention-Based Multiple Instance Learning for Weakly Supervised Ovarian Cancer Survival Prediction from Whole-Slide Histopathology

Core assumptions

The framework assumes that slide-level survival labels (time-to-event and censoring status) are available for all training WSIs, but no pixel-level annotations of tumor regions, stroma, or other histological structures are required. This weakly supervised setting is clinically realistic because cancer registries and biobanks routinely collect survival follow-up data alongside diagnostic slides, whereas pixel-level annotations are rarely available except in small research cohorts [2, 8]. The framework further assumes that WSIs are digitized at sufficient resolution (minimum 20× magnification) to resolve cellular and nuclear details relevant to prognosis, as lower magnification loses discriminative information about nuclear atypia and mitotic figures [5, 6].

A sufficient number of WSIs (typically >200) is necessary to train attention-based MIL models without overfitting, given that each WSI contributes only one survival label despite containing thousands of patches. The framework assumes that survival times are right-censored (e.g., patients alive at last follow-up) and that censoring is non-informative—that is, the censoring mechanism does not depend on unobserved prognostic factors that also affect survival [7, 11]. For multicenter applications, the framework assumes that slide preparation (fixation, sectioning, staining) and digitization protocols are sufficiently standardized or that domain adaptation techniques can mitigate batch effects [8, 12].

Design principles

The framework is designed around three core principles: weak supervision, interpretability, and scalability. Weak supervision eliminates the requirement for pixel-level annotations by learning directly from slide-level survival labels, dramatically reducing the cost and expertise required for dataset curation. This principle enables rapid deployment on large archival cohorts where only diagnostic slides and follow-up data exist, making the framework suitable for real-world clinical settings where pathologist time is a scarce resource [1, 7].

Interpretability is achieved through the attention mechanism itself, which assigns weights to individual patches that can be visualized as heatmaps over the original WSI. High-attention patches reveal which morphological features the model considers prognostic, enabling pathologist validation and hypothesis generation about novel histological correlates of survival [2, 9]. Scalability to gigapixel images (typically 100,000×100,000 pixels) is achieved by processing patches independently and aggregating via attention, which has linear complexity in the number of patches and can be parallelized across GPU memory using techniques such as gradient checkpointing and mixed-precision training [8, 12].

Patch Extraction and Encoding

Patch sampling strategy

The first processing step extracts a set of patches from each WSI, typically at a size of 256×256 or 512×512 pixels at 20× magnification, corresponding to approximately 130×130 μm or 260×260 μm of tissue area. Non-overlapping tiling is computationally efficient and avoids redundant processing, but overlapping sampling (e.g., 50% overlap) can improve spatial coverage and reduce boundary artifacts at the cost of increased patch count [5, 6, 8]. Tissue detection algorithms based on Otsu thresholding or pretrained segmentation networks remove background patches containing no cellular material, reducing the number of patches per WSI from hundreds of thousands to tens of thousands and focusing computational resources on informative regions [13, 14].

Patch size and magnification must balance resolution of cellular details against contextual information. Smaller patches (256×256) at higher magnification (40×) better resolve nuclear atypia and mitotic figures but lose glandular architecture, whereas larger patches (512×512) at lower magnification (10×) preserve tissue topology but may average over heterogeneous regions [1, 7]. For ovarian cancer, a multi-scale approach using two patch sizes (e.g., 256×256 at 20× for cellular features and 512×512 at 5× for architectural features) can capture complementary prognostic signals, though this increases computational requirements [2, 12].

Patch encoder

Each extracted patch is transformed into a low-dimensional feature vector using a pre-trained convolutional neural network (CNN) or vision transformer (ViT) that has been trained on histopathology or natural images. Standard choices include ResNet50, ResNet101, EfficientNet-B4, or ImageNet-pretrained ViT-small, with the final classification layer removed to output a feature vector of typically 1024 or 2048 dimensions [13, 14]. A projection layer (linear or with one hidden layer) then reduces dimensionality to 256–512 dimensions, which improves computational efficiency of the attention MIL module and reduces overfitting [9, 11].

Self-supervised learning on histopathology datasets (e.g., using SimCLR, MoCo, or DINO) produces patch encoders that outperform ImageNet-pretrained models for downstream pathology tasks, including survival prediction [8, 15, 16]. For ovarian cancer specifically, encoders pre-trained on TCGA histopathology images across multiple cancer types capture domain-specific features such as nuclear chromatin patterns and stromal morphology that are not present in natural images. The patch encoder can be frozen during MIL training to reduce memory requirements and prevent overfitting, or fine-tuned end-to-end if computational resources permit, though fine-tuning risks catastrophic forgetting when training data are limited [7, 12].

Attention-Based MIL

Attention aggregation

Given a bag of N patch feature vectors {h₁, h₂, …, } with $h_i \in \mathbb{R}^D$, the attention-based MIL aggregator computes a single bag embedding $z \in \mathbb{R}^D$ as the weighted sum of patch features, where attention weights are learned via a small neural network. The standard attention formulation uses a one-layer or two-layer perceptron followed by tanh activation to compute unnormalized scores, which are then normalized across patches via softmax: $a_i = \frac{\exp(w^T \tanh(V\ h_i))}{\Sigma_j \exp(w^T \tanh(V\ h_j))}$, where $w \in \mathbb{R}^K$ and $V \in \mathbb{R}^{K \times D}$ are learnable parameters and K is the attention hidden dimension [1, 9]. The resulting bag embedding $z = \Sigma_i a_i h_i$ is a convex combination of patch features, with the attention weights indicating each patch's relative contribution to the slide-level representation.

This additive attention mechanism learns which patch features are discriminative for survival prediction without requiring explicit instance-level labels. Unlike max pooling, which selects a single patch, attention aggregation preserves information from multiple informative patches, which is essential for ovarian cancer where survival outcomes depend on aggregate tumor characteristics such as overall TIL density and extent of necrosis rather than a single extreme region [7, 9]. The attention weights can be interpreted as the model's assessment of each patch's prognostic relevance, providing a natural basis for heatmap visualization and pathologist validation [2, 11].

Gated attention

The standard tanh-based attention mechanism suffers from limited expressivity because the tanh nonlinearity saturates for large inputs and cannot model complex interactions between features. Gated attention addresses this limitation by incorporating an additional sigmoid-gated branch: $a_i = \frac{\exp(w^T \tanh(V\ h_i)) \odot\ \sigma(U\ h_i)}{\Sigma_j \exp(w^T \tanh(V\ h_j)) \odot \sigma\ (U\ h_j)}$ where $U \in \mathbb{R}^{K \times D}$ is an additional learnable weight matrix and ⊙ denotes elementwise multiplication [1, 9]. The gating mechanism allows the model to selectively activate or deactivate different dimensions of the feature representation, effectively learning a data-dependent weighting that can capture nonlinear interactions and higher-order statistics of patch features.

For ovarian cancer histopathology, gated attention is particularly advantageous because prognostic features are often defined by combinations of nuclear, stromal, and inflammatory patterns that are not linearly separable in patch feature space. For example, the prognostic significance of TILs depends on both lymphocyte density and the presence of adjacent tumor cells, a conjunction that gated attention can represent more effectively than standard tanh attention [2, 8]. Empirical studies on TCGA data have shown that gated attention consistently outperforms standard attention for survival prediction across multiple cancer types, including ovarian, lung, and renal cell carcinoma [7, 12].

Multi-head attention

Multi-head attention extends the gated attention mechanism by computing multiple independent sets of attention weights (heads), each with its own learnable parameters $\{w_m,\ V_m,\ U_m\}$ for m = 1 to M, where M is typically 4 to 8. Each head produces a separate bag embedding $z_m = \sum a_{i,m}\ h_i,$ and the final bag embedding is obtained by concatenating or averaging the head outputs: z = [z_1; z_2; …;] or $z = \left(\frac{1}{M}\right) \sum z_m$ [1, 12]. Multi-head attention allows the model to capture different histological patterns simultaneously—for example, one head may attend to patches with high nuclear grade, another to patches with desmoplastic stroma, and a third to patches with TIL aggregates—without forcing a single attention distribution to serve all functions.

The multiple attention heads produce distinct heatmaps that can be visualized separately, revealing the morphological features each head has learned to prioritize. In ovarian cancer, one head might consistently attend to regions of solid tumor growth, while another attends to tumor-stroma interfaces where invasion occurs, and a third attends to necrotic debris associated with poor prognosis [2, 8]. This multi-perspective interpretability is a key advantage over single-head attention and standard MIL aggregators, which provide only a monolithic importance score per patch. The computational overhead of multi-head attention is modest, increasing parameters by a factor of M but leaving the forward pass complexity linear in M [1, 9].

Table 1 clarifies why attention-based and especially gated multi-head aggregation is theoretically better suited than fixed pooling strategies for ovarian cancer survival modeling from heterogeneous whole-slide histopathology.

Table 1. Analytical Comparison of Aggregation Strategies for Whole-Slide Survival Modeling in Ovarian Cancer

Aggregation strategy	Core aggregation	What prognostic	Main limitation for	Interpretability profile	Expected effect on	Conceptual role in the

	logic	**information it preserves**	**ovarian cancer survival prediction**		**robustness and generalization**	**proposed framework**
Max pooling MIL	Selects the single highest-scoring patch or instance as the slide representation	Captures extreme focal morphology that may indicate highly aggressive regions	Discards distributed prognostic information across tumor, stroma, immune infiltrates, and necrosis; overly sensitive to outlier patches and artifacts	Narrow and unstable; importance collapses onto one patch	Lower robustness when prognostic signal is spatially heterogeneous	Useful baseline but theoretically mismatched to cumulative and heterogeneous survival signals
Mean pooling MIL	Averages all patch embeddings uniformly	Retains global tissue context and broad morphological burden	Treats informative and uninformative patches equally; dilutes rare but prognostically important regions	Weakly interpretable because all regions contribute similarly	Can appear stable but may underfit clinically meaningful heterogeneity	Baseline representing fully nonselective weak supervision
Noisy-or / probabilistic pooling	Combines instance signals under a probabilistic assumption that multiple instances jointly determine bag outcome	Better than max when several patches contribute to risk	Still imposes a fixed aggregation rule and may be difficult to align with censored continuous survival outcomes	Moderate but indirect interpretability	Intermediate robustness; depends strongly on formulation assumptions	Transitional strategy between rigid MIL and learned attention
Single-head attention MIL	Learns one attention distribution over all patch embeddings	Preserves multiple informative regions while suppressing background	May force biologically distinct prognostic patterns into one attention map	Stronger interpretability than fixed pooling; attention heatmap directly viewable	Better generalization than rigid pooling when signal is distributed	Minimal viable attention formulation for weakly supervised survival analysis
Gated attention MIL	Applies learnable	Better captures nonlinear	More parameters	High interpretability	Improved robustness	Central mechanism

	gating to modulate feature interactions before attention normalization	combinations of nuclear, stromal, and inflammatory morphology	increase overfitting risk in small cohorts	with richer feature selectivity	when prognosis depends on feature conjunctions rather than isolated cues	proposed in the manuscript because ovarian prognostic morphology is combinatorial
Multi-head attention MIL	Learns several parallel attention distributions and combines their bag embeddings	Preserves multiple distinct prognostic subpatterns simultaneously	Requires careful regularization and head interpretation to avoid redundancy	Very strong; each head can be mapped to a distinct morphological emphasis	Strong potential for cross-cohort robustness if heads capture complementary structure	Best aligns with the manuscript's argument that ovarian survival depends on several coexisting histologic programs
Hierarchical slide-to-patient attention	Aggregates first within slides and then across multiple slides per patient	Preserves both intratumoral and inter-slide heterogeneity	More complex training and metadata handling	High interpretability at both patch and slide levels	Likely superior for multi-slide cases if patient-wise validation is enforced	Important extension for clinical translation when multiple tumor blocks exist

Survival Prediction Module

Deep cox model

The bag embedding $z \in \mathbb{R}^D$ produced by the attention-based MIL aggregator is passed through a multilayer perceptron (MLP) with one or two hidden layers and a final linear layer that outputs a scalar risk score r = f(z; θ), where higher r indicates worse prognosis. The MLP typically uses ReLU or LeakyReLU activations and dropout (p=0.2–0.5) to prevent overfitting, given that the number of training slides is often modest relative to the embedding dimensionality [7, 12]. This risk score substitutes for the linear predictor β^T x in the conventional Cox proportional hazards model, enabling the framework to learn nonlinear relationships between histopathological features and survival outcomes without prespecifying functional forms [2, 11].

The model is optimized by minimizing the negative Cox partial likelihood loss: $L = - \Sigma_{[r_i - \log(\Sigma_{\exp}(r_j))]}$, where δ_i is the censoring indicator (1 for observed events, 0 for censored) and the inner sum is over the risk set of patients who survived at least as long as patient i [7, 8]. This loss function handles right-censored data naturally by comparing each patient who experiences an event only to those who were still at risk at that time, making no parametric assumption about the baseline hazard function. The Cox loss is differentiable and can be optimized with standard stochastic gradient descent techniques, including Adam or AdamW with learning rate scheduling [12].

Discrete-time survival

An alternative to the continuous-time Cox model is the discrete-time survival formulation, in which follow-up time is divided into K intervals (e.g., monthly or quarterly) and the model outputs the probability of event occurrence in each interval conditional on survival up to that interval. The bag embedding z is passed through an MLP with K output nodes followed by a sigmoid activation to produce hazard probabilities h_k= P(event in interval k | survival to start of interval k) . The overall survival probability to time t (within interval k) is computed as the product of $\begin{matrix}(1 - h_1)\\ \times (1 - h_2)\\ \times \ ...\\ \times (1 - h_k)\end{matrix}$, and the loss function is the negative log-likelihood of the observed event or censoring times under these interval-specific hazards.

Discrete-time survival offers several advantages for attention-based MIL with histopathology images. First, it naturally accommodates time-varying effects, allowing the importance of different histological features to change across the disease course—for example, features predicting early recurrence may differ from those predicting late mortality [7, 8]. Second, discrete-time models produce interpretable survival curves rather than a single risk score, which is clinically actionable for treatment planning. Third, the discrete formulation simplifies handling of tied event times and can be more stable than Cox when the proportional hazards assumption is violated [12]. However, discrete-time models require selecting the number and boundaries of time intervals, which may introduce hyperparameter sensitivity.

Interpretability via Attention

Attention heatmaps

The attention weights computed by the MIL aggregator provide a direct mechanism for visualizing which regions of the WSI the model considers most prognostic. For each patch at spatial coordinates (x_i, y_i), the corresponding attention weight is mapped back to the original WSI coordinate system and rendered as a color overlay, typically with a blue (low attention) to red (high attention) colormap [1, 9]. These heatmaps can be generated at multiple scales by aggregating patch-level attention to superpixel regions or by using overlapping patches with bilinear interpolation to produce smooth spatial maps. Multi-head attention yields multiple heatmaps, each highlighting different histological patterns that the respective head has learned to prioritize [2, 8].

The attention heatmaps enable pathologists to visually assess whether the model has learned biologically plausible prognostic features. In ovarian cancer, high-attention regions typically correspond to areas of high-grade nuclear atypia, solid tumor growth, desmoplastic stroma, or dense TIL aggregates—all established prognostic factors [1, 7]. Conversely, low-attention regions include background, necrosis, normal fallopian tube epithelium, or benign ovarian tissue. When attention maps localize to known prognostic regions, this increases clinician trust and facilitates model acceptance; when attention highlights previously unrecognized regions, this can generate hypotheses for prospective validation [11, 12].

Clinical validation

Interpretability alone is insufficient without rigorous clinical validation of the attention patterns. The framework recommends a two-stage validation protocol: first, a pathologist blinded to model outputs reviews high-attention patches from a held-out test set and classifies them into predefined morphological categories (e.g., tumor epithelium, stroma, TILs, necrosis, normal tissue). Second, the distribution of attention weights across these categories is quantified, and statistical tests assess whether attention is disproportionately concentrated on prognostically relevant categories relative to a null distribution [2, 8]. For example, the mean attention weight assigned to tumor patches should significantly exceed the proportion of tumor area in the WSI.

Beyond categorical validation, attention heatmaps can be compared to pathologist-drawn regions of interest (ROIs) from prior studies. The Dice similarity coefficient between high-attention regions (e.g., top 10% of patches by weight) and pathologist-annotated tumor regions provides a quantitative measure of alignment [1, 7]. Disagreements—where the model attends to regions the pathologist did not annotate—may reveal novel morphological correlates of survival, such as

specific stromal patterns or peritumoral lymphocyte aggregates not routinely scored in clinical practice. These discoveries can be tested in independent cohorts, potentially identifying new prognostic biomarkers for ovarian cancer [8, 12].

Training Considerations

Loss function

The primary loss function is the negative Cox partial likelihood (for continuous-time survival) or the negative discrete-time log-likelihood(for interval-based survival). However, the attention mechanism introduces additional degrees of freedom that can lead to degenerate solutions, such as uniformly distributed attention weights that approximate mean pooling or a single patch receiving nearly all attention (collapsing to max pooling). To regularize the attention distribution, an entropy regularization term $\begin{matrix} H\left(a\right) \\ = \\ -\sum a_i \log\left(a_i\right) \end{matrix}$ can be added to the loss, encouraging the model to distribute attention across multiple informative patches rather than focusing on a single outlier [1, 9]. Conversely, a sparsity penalty such as L1 regularization on attention weights may be appropriate when only a small fraction of patches are prognostically relevant [7, 11].

Additional regularization strategies include weight decay on all network parameters, dropout in the attention MLP and survival MLP, and early stopping based on validation C-index. For the patch encoder, freezing pretrained weights during initial MIL training prevents overfitting, with optional fine-tuning of the final few layers after the MIL aggregator has converged [8, 12]. When multiple slides per patient are available (e.g., multiple tumor blocks), a patient-level loss that aggregates slide-level risk scores (e.g., taking the maximum or mean) can improve stability and account for intratumoral heterogeneity [2, 7].

Handling slide-level variability

Ovarian cancer patients often have multiple WSIs per case (e.g., primary tumor, omental metastasis, lymph node involvement), and these slides may contain different prognostic information. The framework must specify how to combine predictions from multiple slides belonging to the same patient. Three strategies are common: (1) treat each slide as an independent bag and average the resulting risk scores; (2) concatenate patch features from all slides into a single bag and process jointly; or (3) use a patient-level MIL with slide-level attention before survival prediction [1, 8]. The first approach is simplest but ignores between-slide relationships; the second is computationally intensive; the third offers a hierarchical attention structure where slide-level attention weights indicate which tumor block is most prognostic.

Cross-validation must respect patient-level separation to avoid data leakage, as slides from the same patient are not independent. The framework mandates patient-wise cross-validation, where all slides from a given patient are assigned exclusively to training, validation, or test splits [7, 12]. Stratified splitting by event status and clinical stage ensures balanced representation across folds. For multicenter data, batch effects (differences in slide preparation, staining protocols, digitization equipment) can confound survival prediction; the framework recommends including site identifiers as covariates in the survival model or using domain-adversarial training to learn site-invariant features [2, 8].

Evaluation Strategy

Survival metrics

Performance is evaluated using the C-index, measuring correct ranking of survival times (0.5=random, 1.0=perfect; >0.7 clinically useful) [7, 12]. Time-dependent AUC (tAUC) assesses discrimination at specific time points (e.g., 1-, 3-, 5-year) [1, 2]. Calibration is evaluated by comparing predicted vs. observed survival (Kaplan–Meier), and overall performance is

summarized using the Integrated Brier Score (IBS) [8, 12]. All metrics are reported with 95% confidence intervals via bootstrap (1,000 iterations).

Validation protocols

A three-level validation ensures generalizability: (1) internal k-fold cross-validation (k=5 or 10), (2) temporal validation on a separate time-based test set, and (3) external validation on independent cohorts from different institutions (e.g., CPTAC, TCGA-OV) [1, 7, 12].

Table 2 consolidates the translational logic of the framework by linking each major design choice to its methodological tradeoff, expected failure risk, and required validation pathway before clinical use.

Table 2. Translational Design Matrix for Attention-Based MIL Survival Prediction: Links Between Methodological Choices, Failure Risks, and Clinical Validation Requirements

Design dimension	Strategic options discussed in the manuscript	Principal methodological tradeoff	Failure mode if poorly specified	Clinical or biological consequence	Recommended validation or safeguard	Theoretical implication for the framework
Patch scale and magnification	Small high-magnification patches; larger low-magnification patches; multi-scale sampling	Cellular detail versus tissue architecture	Missing either nuclear atypia or broader stromal/topologic context	Risk model may overweight one morphological scale and miss another prognostic process	Compare single-scale and multi-scale ablations; inspect attention localization across scales	Prognosis is multi-resolution, so representatio design shape what “survival signal” can be learned
Tissue filtering and patch selection	Threshold-based tissue detection; learned tissue filtering; overlapping versus non-overlapping tiling	Computational efficiency versus spatial completeness	Background contamination, boundary artifacts, or exclusion of rare informative regions	Heatmaps may highlight artifacts rather than pathology	Audit retained patch distributions and visually review high-attention border regions	Weak supervision depends heavily on the fidelity of the bag before learning begin
Encoder initialization	ImageNet pretraining; histopathology self-supervision; frozen versus fine-tuned encoder	Domain transferability versus overfitting risk	Features may be semantically weak for pathology or catastrophically drift during fine-tuning	Prognostic morphology may be encoded poorly, reducing biological plausibility	Compare frozen and fine-tuned models; benchmark domain-specific pretraining	Representatio quality is not neutral; it determines which tissue patterns attention can exploit
Attention mechanism choice	Standard attention; gated	Simplicity versus	Uniform attention, single-patch collapse,	Low clinician trust and	Monitor entropy/sparsity of attention;	Interpretability is architecture dependent

	attention; multi-head attention	expressive capacity	or redundant heads	unstable explanations	compare head diversity and reproducibility	rather than automatically guaranteed
Survival objective	Deep Cox loss; discrete-time survival loss	Ranking efficiency versus time-specific survival estimation	Misfit under non-proportional hazards or unstable interval definitions	Poor calibration or clinically unhelpful outputs	Evaluate C-index, tAUC, IBS, and calibration jointly	Survival formulation changes the meaning of th model output from generic risk to time-sensitive prognosis
Multi-slide patient handling	Independent slide averaging; concatenated slide bag; hierarchical patient-level attention	Simplicity versus faithful modeling of inter-slide heterogeneity	Slide leakage or underuse of complementary tumor blocks	Patient prognosis may reflect specimen selection bias rather than disease biology	Enforce patient-wise splitting and compare slide-aggregation strategies	The clinically relevant prediction target is the patient, not th isolated slide
Cross-validation strategy	Internal k-fold; temporal split; external institutional validation	Convenience versus genuine transportability	Optimistic performance due to leakage or cohort homogeneity	Overstated readiness for clinical deployment	Use patient-wise splits, temporal holdout, and external cohorts	Generalizatio must be demonstrated across time and site, not inferred from internal accuracy alon
Batch effects and site variation	Standardization, site covariates, or domain adaptation	Model purity versus operational realism	Model learns stain/scanner/site signatures instead of prognosis	False confidence and poor external reproducibility	Perform site-stratified analyses and domain shift testing	Clinical translation requires separating biological signal from acquisition bias
Attention heatmap validation	Visual inspection only; category-based pathology review; overlap with ROI annotations	Convenience versus evidentiary rigor	Plausible-looking but biologically unverified explanations	Interpretability claims remain weak and non-actionable	Conduct blinded pathologist review and quantify alignment statistically	Explanation must be clinically adjudicated to become meaningful evidence
Benchmarking endpoint	Internal discrimination only versus	Easier reporting versus	High C-index with poor calibration or	Limited value for treatment planning and	Report bootstrap confidence	A translationa survival framework

	discrimination plus calibration and external reproducibility	translational credibility	unstable transportability	risk communication	intervals, calibration, IBS, and external performance	must be judged as a clinical decision model, not merely a pattern recognizer

Ablation studies compare the attention-based model with baselines (mean, max, noisy-or pooling, random attention) [8, 9, 11]. Statistical tests (bootstrap or DeLong) assess significance. Additional analyses evaluate attention type (multi-head vs single-head, gated vs standard) and encoder initialization [2, 12].

Conclusion

We propose an attention-based MIL framework for ovarian cancer survival prediction from whole-slide images without pixel-level annotations. It integrates a patch encoder, gated multi-head attention, and a survival prediction head optimized via Cox loss, enabling scalable weakly supervised learning.

Key advantages include reduced annotation cost, interpretability via attention heatmaps, and computational scalability for large datasets.

Limitations include reliance on high-quality survival data, sensitivity to patch sampling and encoder design, and limited modeling of spatial relationships. Extensive external validation is required before clinical use.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 19 Jul 2023 Revised: 16 Oct 2023 Accepted: 25 Dec 2023
Published online: 20 July 2024

References

Claessens CB, Schultz EWR, Koch A, Nies I, Hellström AET, Nederend J, et al. Multi-center ovarian tumor classification using hierarchical transformer-based multiple-instance learning. In: MICCAI Workshop on Cancer Prevention through Early Detection. Cham: Springer Nature Switzerland; 2024. p. 3-13.

Leiby JS, Hao J, Kang GH, Park JW, Kim D. Attention-based multiple instance learning with self-supervision to predict microsatellite instability in colorectal cancer from histology whole-slide images. In: 44th Annu Int Conf IEEE Eng Med Biol Soc (EMBC). IEEE; 2022. p. 3068-71.

Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. In: Int Conf Mach Learn. PMLR; 2018. p. 2127-36.

Dimitriou N, Arandjelović O, Harrison DJ. Magnifying networks for histopathological images with billions of pixels. Diagnostics (Basel). 2024;14(5):524.
https://doi.org/10.3390/diagnostics14050524.

Tarkhan A, Nguyen TK, Simon N, Dai J. Survival prediction via deep attention-based multiple-instance learning networks with instance sampling. Proc AAAI Symp Ser. 2023;2(1):482-9.

Shao W, Wang T, Huang Z, Han Z, Zhang J, Huang K. Weakly supervised deep ordinal Cox model for survival prediction from whole-slide pathological images. IEEE Trans Med Imaging. 2021;40(12):3739-47.
https://doi.org/10.1109/TMI.2021.3093803.

Xiang T, Song Y, Zhang C, Liu D, Chen M, Zhang F, et al. DSNet: a dual-stream framework for weakly-supervised gigapixel pathology image analysis. IEEE Trans Med Imaging. 2022;41(8):2180-90.
https://doi.org/10.1109/TMI.2022.3157250.

Agarwal S, Abaker MEO, Daescu O. Survival prediction based on histopathology imaging and clinical data: a novel, whole slide CNN approach. In: Int Conf Med Image Comput Comput Assist Interv. Cham: Springer International Publishing; 2021. p. 762-71.

Yao J, Zhu X, Huang J. Deep multi-instance learning for survival prediction from whole slide images. In: Int Conf Med Image Comput Comput Assist Interv. Cham: Springer International Publishing; 2019. p. 496-504.

Tang B, Li A, Li B, Wang M. CapSurv: capsule network for survival analysis with whole slide pathological images. IEEE Access. 2019;7:26022-30.
https://doi.org/10.1109/ACCESS.2019.2899823.

Gadermayr M, Tschuchnig M. Multiple instance learning for digital pathology: a review of the state-of-the-art, limitations & future potential. Comput Med Imaging Graph. 2024;112:102337.
https://doi.org/10.1016/j.compmedimag.2024.102337.

Tan L, Li H, Yu J, Zhou H, Wang Z, Niu Z, et al. Colorectal cancer lymph node metastasis prediction with weakly supervised transformer-based multi-instance learning. Med Biol Eng Comput. 2023;61(6):1565-80.
https://doi.org/10.1007/s11517-023-02786-7.

Atabansi CC, Nie J, Liu H, Song Q, Yan L, Zhou X. A survey of transformer applications for histopathological image analysis: new developments and future directions. Biomed Eng Online. 2023;22(1):96. https://doi.org/10.1186/s12938-023-01166-8.

Campanella G, Kwan R, Fluder E, Zeng J, Stock A, Veremis B, et al. Computational pathology at health system scale: self-supervised foundation models from three billion images. arXiv [Preprint]. 2023 Oct 10:arXiv:2310.07033.

Huang Z, Bianchi F, Yuksekgonul M, Montine TJ, Zou J. A visual-language foundation model for pathology image analysis using medical Twitter. Nat Med. 2023;29(9):2307-16. https://doi.org/10.1038/s41591-023-02504-3.

Chen RJ, Ding T, Lu MY, Williamson DFK, Jaume G, Chen B, et al. A general-purpose self-supervised model for computational pathology. arXiv [Preprint]. 2023:arXiv:2308.15474.