I. J. Engineering and Manufacturing, **2026, 5, 158-178**
Published Online on October 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijem.2026.05.09

Attention-Guided Deep Learning Framework for Ovarian Cancer Subtype Classification

Vijay H. Kalmani*
Department of Computer Science and Engineering, Kasegaon Education Society's, Rajarambapu Institute of Technology, Shivaji University, Sakharale, Maharashtra – 415414, India
E-mail: vijaykalmani@gmail.com
ORCID iD: https://orcid.org/0000-0003-0738-3211
*Corresponding author

Nagaraj V. Dharwadkar*
Department of Computer Science, Central University of Karnataka, Kadaganchi, Aland Road, Kalaburagi Dist. - 585 367, Karnataka, India
E-mail: dharwadkarn@cuk.ac.in
ORCID iD: https://orcid.org/0000-0003-3017-0011
*Corresponding author

Amol C. Adamuthe
Department of Information Technology, Kasegaon Education Society's, Rajarambapu Institute of Technology, Shivaji University, Sakharale, Maharashtra – 415414, India
E-mail: amol.admuthe@gmail.com
ORCID iD: https://orcid.org/0000-0003-2483-7051

Altaf Husain
RUCKUS Networks, Sunnyvale, CA 94089, USA
E-mail: althaf.atf@gmail.com
ORCID iD: https://orcid.org/0009-0009-6615-8736

Received: July 17, 2026; Revised: August 17, 2026; Accepted: August 27, 2026; Published: October 8, 2026

Abstract: Ovarian cancer histotype classification is challenging because of substantial morphological heterogeneity and subtle subtype-specific features. This work presents a controlled evaluation of high-resolution pathology representations and slide-level aggregation methods for automated classification of five ovarian cancer subtypes from whole-slide histopathology images. Precomputed CONCH patch embeddings were aggregated using mean pooling, max pooling, gated attention-based multiple-instance learning, and a max-pooling cascade. The models were evaluated on 513 non-TMA whole-slide images from UBC-OCEAN using leakage-controlled five-fold cross-validation, with each slide receiving exactly one out-of-fold prediction. Gated ABMIL achieved a balanced accuracy of 81.39% and a macro-F1 score of 81.96%. Max pooling produced the highest numerical performance, with a balanced accuracy of 81.44%, macro-F1 of 82.59% (95% CI: 78.84-86.26%), macro-AUROC of 96.99%, and macro-AUPRC of 90.79%. However, paired slide-level bootstrap comparisons found no statistically significant differences among the CONCH aggregation strategies after Holm correction. Compared with the EfficientNet-B0 thumbnail baseline, CONCH max pooling improved macro-F1 by 30.30 percentage points and balanced accuracy by 25.70 percentage points, with both paired bootstrap confidence intervals excluding zero. Attention weights enabled visualization of influential patches, although these regions were not independently validated by pathologists. The findings show that high-resolution pathology foundation-model representations support ovarian cancer subtyping, while greater aggregation complexity does not necessarily improve performance. External multi-institutional validation is required before clinical generalizability can be established.

Index Terms: Ovarian cancer, Computational pathology, Whole-slide images, Pathology foundation model, Multiple-instance learning, Gated attention, Histopathology classification

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

1. Introduction

1.1. Clinical and Computational Importance of Ovarian Cancer Subtyping

Ovarian cancer is associated with high mortality, primarily because of its clinically silent progression and substantial histological and molecular heterogeneity. Accurate subtyping is important for treatment selection and prognostic assessment because the major subtypes differ in their clinical behaviour and response to therapy. Both high-grade serous and endometrioid cancers are sensitive to platinum-based therapies, whereas clear cell and mucinous subtypes are resistant to platinum-based therapies [1,2].

With recent advances in computing methods and the integration of multi-omics data, such as gene expression, microRNA profiling, and methylation patterns [3,4], such classification will become more accurate in subtyping and better able to identify actionable targets. Such models will allow targeted therapies, such as the use of poly (ADP-ribose) polymerase (PARP) inhibitors, in cancers with homologous recombination deficiencies [5]. Therefore, accurate classification is required to assess prognosis, treatment options, and diagnosis.

Notwithstanding recent advancements in deep learning, subtype classification in ovarian cancer is still plagued by imbalanced classes, stain variability, and poor interpretability of predictions. To address these challenges, we investigate an attention-guided framework that uses high-resolution pathology features and slide-level feature aggregation for ovarian cancer subtype classification. Our methodological approach aims to improve the classification accuracy, interpretability, and generalizability of histopathological image analysis.

1.2. Shortcomings of Conventional Histopathology

Historically, the reference standard of histopathological diagnosis has been affected by several limitations, particularly inter-observer variability and subjective interpretation in borderline or equivocal cases [6,7]. These limitations may delay diagnosis or lead to inconsistent decisions that directly influence treatment.

Recent advancements include machine learning-assisted second-harmonic generation imaging coupled with DESI-MSI, which has indicated an objective and reproducible inspection of tissue morphology and metabolism [6,8]. However, these techniques are not yet ready for routine clinical practice; therefore, greater attention should be given to developing scalable and automated approaches.

1.3. Deep Learning in Histopathological Image Analysis

Deep learning has enabled automated feature extraction from histopathological images and has the potential to support reproducible cancer classification. Convolutional neural networks (CNNs) can learn complex morphological patterns directly from tissue images, thereby reducing dependence on manually designed features [9-12]. However, in ovarian cancer, the reliability of deep-learning-based subtyping is affected by limited sample sizes, imbalanced subtype distributions, stain variability, and insufficient model interpretability. These limitations motivate the development of subtype-specific approaches that combine robust data preparation and spatially focused feature learning.

1.4. Challenges in Deep Learning-Based Ovarian Subtyping

However, the application of deep learning to ovarian cancer subtype classification remains challenging for several reasons. One of the greatest challenges is the class imbalance problem, in which extremely rare subtypes are underrepresented in the dataset, resulting in a biased model [13,14]. Moreover, the significant morphological variability that exists both among and within subtypes restricts generalization [15].

If we consider the integration of high-dimensional genomic data, the overfitting risk also increases, and we require techniques such as ensemble methods and dimensionality reduction [16,17]. These are some common concerns that point towards the requirement of properly structured preprocessing pipelines and explainable models.

1.5. Requirement for Advanced Augmentation and Interpretability

Another significant bottleneck in computer pathology is stain variability across laboratories. This creates domain shifts that compromise the robustness of the model. Methods such as RandStainNA, which fuses stain augmentation and normalization, enable the design of stain-invariant models and promote cross-domain generalizability [18].

In addition, attention mechanisms, such as the Convolutional Block Attention Module and Self-Attentive Adversarial Stain Normalization, assist deep networks in concentrating on areas that are important in terms of diagnostics and enhancing interpretability and instilling confidence [19,20].

1.6. Objective of this Study

The objective of this work was to evaluate high-resolution pathology foundation-model representations for five-subtype ovarian carcinoma classification from whole-slide histopathology images. Frozen CONCH embeddings were used to represent tissue patches, and gated attention-based multiple-instance learning was applied to aggregate the variable number of patch features available for each slide. The gated attention model was compared under identical data partitions with mean pooling, max pooling, and a two-stage max-pooling cascade. Additional comparisons were

conducted against ImageNet-pretrained ResNet18, ResNet50, ResNeXt50-32x4d, EfficientNet-B0, and ConvNeXt-Tiny models trained on downsampled WSI thumbnails. Performance was assessed using leakage-controlled five-fold out-of-fold predictions at the slide level, with balanced accuracy and macro-F1 as the principal metrics, supported by macro-AUROC, macro-AUPRC, classwise measures, bootstrap confidence intervals, and paired statistical comparisons.

1.7. Summary of the Article

The remainder of this paper is organized as follows. Section 2 reviews related work and positions the present research. Section 3 describes the dataset, high-resolution feature representation, aggregation models, baseline classifiers, and evaluation protocol. Section 4 presents the experimental, classwise, comparative, and statistical results. Section 5 discusses the principal findings and limitations, and Section 6 concludes the article and outlines directions for future work.

2. Related Work

2.1. Deep Learning for Ovarian Cancer Histopathology subtyping

Recent studies have investigated several deep-learning strategies for ovarian cancer histopathology subtyping. An EfficientNet-B0 feature extractor combined with a fine-tuned k-nearest-neighbour classifier was evaluated on 725 images representing clear-cell carcinoma, endometrioid carcinoma, high-grade serous carcinoma, low-grade serous carcinoma, and mucinous carcinoma [21]. Although the study reported 100% validation and test accuracy, the use of a relatively small 80:20 split, absence of external validation, and limited information about patient-level separation make the generalizability of this result uncertain. OVANet combined modified VGG19 and InceptionV3 backbones with squeeze-and-excitation and spatial-attention mechanisms [22]. The original collection of 508 images was expanded to approximately 13,024 images through augmentation, and the model reportedly achieved 99.01% accuracy. However, the unclear validation protocol and substantial dependence on augmented images make it difficult to determine whether this performance would transfer to independent cases. A rehearsal-based continual-learning approach subsequently applied a Vision Transformer with task-specific classification heads and memory buffers [23]. It reported validation and test accuracies of 96.2% and 95.4%, respectively, with subtype-level F1 scores above 0.92. Nevertheless, incomplete reporting of the dataset, imaging unit, and external evaluation limits direct comparison with slide-level histopathology studies. Overall, these findings demonstrate the feasibility of automated ovarian cancer subtyping, but their consistently high reported performance must be interpreted in relation to dataset size, augmentation, prediction unit, and validation design.

2.2. High-Resolution and Foundation-Model Representations in Computational Pathology

The gigapixel scale of whole-slide images makes direct processing computationally difficult and has encouraged the development of patch-based computational pathology methods. In these approaches, tissue regions are divided into smaller patches, encoded individually, and subsequently combined to form a slide-level representation. Transformer-based and hierarchical architectures extend conventional patch processing by modelling broader contextual relationships among tissue regions at one or more spatial scales [24,25].

Pathology foundation models use large-scale pretraining to produce transferable representations of histopathology patches. Models such as CONCH and UNI can provide general-purpose features for downstream diagnostic, prognostic, and biomarker-prediction tasks, reducing dependence on task-specific feature learning from limited annotated datasets [24]. CONCH additionally employs contrastive vision-language pretraining to align histopathology images with descriptive pathology text, whereas UNI is based on large-scale self-supervised visual pretraining. These approaches illustrate the transition from features pretrained on natural images to representations learned directly from diverse histopathology data [25].

At the whole-slide level, Prov-GigaPath was pretrained using approximately 1.3 billion image tiles obtained from 171,189 whole-slide images. Its LongNet-based slide encoder was designed to process extensive patch sequences and capture broader tissue context [26]. Other approaches, such as SeqShort, reduce long patch sequences before transformer-based classification, thereby providing a more computationally manageable route to whole-slide modelling [27].

Compared with classifiers operating on low-resolution slide thumbnails, high-resolution patch representations preserve more detailed cellular and tissue morphology. Large-scale pathology pretraining may also facilitate adaptation to downstream tasks with limited labelled data. Nevertheless, performance can remain sensitive to domain shifts, staining differences, scanner variation, computational requirements, and under-representation of rare tumors [24]. These limitations are particularly relevant to ovarian histopathology, where uncommon subtypes and substantial within- and between-class morphological heterogeneity complicate model development and validation [28,29]. Accordingly, the present work uses pretrained CONCH patch embeddings to retain high-resolution tissue information and evaluates slide-level aggregation separately, rather than assuming that a more complex aggregation method will necessarily improve classification.

2.3. Attention-Based Multiple-Instance Learning

Multiple-instance learning is widely used for weakly supervised analysis of whole-slide images because diagnostic labels are generally available for complete slides rather than individual tissue patches. Each slide is represented as a variable-length bag of patch embeddings, which are aggregated to obtain a slide-level prediction. In gated attention-based multiple-instance learning, parallel nonlinear transformations generate a trainable attention score for every patch, and the normalized scores are used to construct a weighted slide representation. Recent ovarian histopathology studies have combined this approach with pathology foundation-model features for five-subtype whole-slide classification [29,30].

Gated ABMIL differs from convolutional attention modules such as CBAM and SE-Net. CBAM applies channel and spatial attention within a convolutional feature map, whereas SE-Net recalibrates the relative importance of feature channels. These mechanisms refine representations within an image or tissue patch but do not independently aggregate a variable number of patches into a whole-slide prediction. Gated ABMIL instead operates across patch embeddings at the bag level and estimates their relative contributions to the slide representation [24,25].

Transformer-based MIL provides another slide-level modelling strategy by allowing patch embeddings to interact through self-attention. This can capture relationships among tissue regions that are not explicitly represented by the weighted-sum formulation of gated ABMIL. However, the computational cost of self-attention increases substantially when a slide contains thousands of patches, and its multi-head attention patterns can be difficult to interpret. Recent ovarian WSI studies have therefore evaluated transformer-based aggregation alongside mean pooling, max pooling, gated ABMIL, CLAM, and dual-stream MIL methods [31,32].

Gated ABMIL produces patch-level weights that can be projected onto the corresponding slide coordinates for visual examination. Nevertheless, high attention does not establish that a patch is biologically or diagnostically meaningful, because attention weights may also respond to staining variation, tissue-preparation artifacts, or other spurious features. Attention maps should therefore be treated as visualizations of model behavior rather than as clinically validated explanations unless they are independently assessed by pathologists [29,30].

Slide-level aggregation can also be performed without trainable attention. Mean pooling treats all patch representations equally and is computationally efficient, although diagnostically informative regions may be diluted by less relevant tissue. Max pooling retains the strongest response within each feature dimension but may be sensitive to isolated artifacts. Hierarchical classification divides the multiclass problem into successive decision stages, potentially improving discrimination among related subtypes at the cost of additional complexity. Recent evidence does not establish that one aggregation strategy is consistently superior across ovarian datasets and feature extractors [31]. Consequently, the present work compares mean pooling, max pooling, gated ABMIL, and a two-stage max-pooling cascade using identical CONCH embeddings, fold assignments, optimization settings, and slide-level OOF evaluation.

2.4. Research Gap and Positioning of the Present Work

Recent ovarian histopathology research has demonstrated that high-resolution patch representations obtained from pathology foundation models can support five-subtype whole-slide classification [29]. Encoders such as UNI and H-optimus-0 have achieved strong internal performance, while attention-based multiple-instance learning has enabled weakly supervised classification from whole-slide images [29]. However, reported performance depends substantially on the development cohort, feature extractor, prediction unit, and validation design. Performance can decline when models are evaluated on external data [29,30], indicating that internal cross-validation alone is insufficient to establish cross-institutional generalizability.

Recent evidence also indicates that the choice of feature extractor may influence ovarian subtype-classification performance more strongly than the aggregation method [31]. Mean pooling, max pooling, gated ABMIL, clustering-constrained attention, dual-stream MIL, transformer-based aggregation, and hierarchical classification have been investigated. Nevertheless, controlled comparisons using identical feature representations, data partitions, optimization procedures, and statistical analyses remain limited [31]. Class imbalance is consistently observed in five-subtype ovarian histopathology datasets, making balanced accuracy and subtype-level recall important complements to ordinary accuracy [29,31]. However, precision, recall, and F1 scores for uncommon subtypes are not reported consistently across studies.

Attention heatmaps are frequently used to visualize tissue regions that influence slide-level predictions [29,30]. However, such maps are seldom validated independently by pathologists, and formal evaluation of probability calibration and predictive uncertainty remains uncommon in ovarian WSI classification [24,29]. The established evidence and the aspects that remain insufficiently examined are summarized in Table 1.

As shown in Table 1, the principal limitation is not the absence of increasingly complex architectures, but the limited controlled evidence concerning how representation and aggregation choices affect slide-level ovarian subtype classification. The present work addresses part of this gap by comparing CONCH patch embeddings aggregated through mean pooling, max pooling, gated ABMIL, and a two-stage max-pooling cascade under the same leakage-controlled five-fold OOF protocol. Each of the 513 eligible whole-slide images contributes exactly one OOF prediction, and performance is assessed using balanced accuracy, macro-F1, macro-AUROC, macro-AUPRC, classwise measures, bootstrap confidence intervals, and paired statistical comparisons.

Table 1. Summary of existing approaches and research gaps in five-subtype ovarian cancer classification from whole-slide images.

Aspect	Established evidence	Gap or insufficiently examined aspect
High-resolution representation	Pathology foundation-model embeddings preserve detailed tissue morphology and support ovarian cancer subtype classification [29].	The effects of patch magnification, multiscale representation, and slide-wide spatial context have not been evaluated consistently across ovarian histopathology datasets [29, 31].
Leakage-controlled evaluation	Stratified cross-validation, independent holdout testing, and external evaluation have been used in comparatively rigorous ovarian WSI studies [29, 30].	Patient-level separation, correlated slides, batch effects, and other potential sources of leakage are not documented uniformly across studies [29, 31].
Feature aggregation	Mean pooling, max pooling, gated ABMIL, CLAM, DSMIL, transformer-based aggregation, and hierarchical classification have been investigated [31].	Direct comparisons using identical embeddings, partitions, optimization procedures, and statistical analyses remain limited for five-subtype ovarian classification [31]
Class imbalance	Balanced accuracy and subtype-level recall are increasingly used to assess performance on imbalanced ovarian WSI datasets [29, 31].	Rare-subtype precision, recall, and F1 scores are not reported consistently, and dedicated imbalance-handling methods require more systematic evaluation [29, 31].
Interpretability and uncertainty	Attention weights and heatmaps are used to visualize tissue regions influencing slide-level predictions [29, 30].	Attention maps are rarely validated independently by pathologists, while probability calibration and formal uncertainty quantification remain underexamined [24,29].
External validation	Some models have been evaluated on independent institutional cohorts or publicly available ovarian cancer WSI datasets [29, 30].	Large, geographically diverse cohorts covering different scanners, staining protocols, and patient populations remain limited [24, 29].

The work is positioned as a controlled empirical evaluation rather than the introduction of a new attention mechanism. It does not resolve the need for geographically diverse external validation [29,30], calibrated uncertainty estimation [24], or independent pathological verification of attention-derived visualizations [29,30]. These aspects remain limitations of the present analysis and priorities for future investigation.

3. Materials and Methodology

3.1. Dataset

This study used the labelled histopathological image dataset from the Ovarian Cancer Subtype Classification and Outlier Detection (UBC-OCEAN) competition on Kaggle, organized by the Artificial Intelligence in Medicine Lab at the University of British Columbia [33]. The present analysis includes 513 non-TMA whole-slide images representing five histological subtypes of ovarian cancer: clear cell carcinoma (CC), endometrioid carcinoma (EC), high-grade serous carcinoma (HGSC), low-grade serous carcinoma (LGSC), and mucinous carcinoma (MC).

The class distribution is imbalanced, with HGSC constituting the largest class and LGSC and MC the smallest classes. All 513 eligible whole-slide images were included in leakage-controlled five-fold cross-validation. In each fold, the aggregation classifier was fitted using only the training slides and then applied to the corresponding validation slides. Predictions from the five non-overlapping validation folds were pooled so that each slide contributed exactly one out-of-fold prediction. No separate 25-image holdout result is reported in the revised analysis.

The high-resolution feature source contained 1,307,237 tissue-patch embeddings across the 513 whole-slide images, with a median of 2,274 patches per slide. Table 2 summarizes the dataset and its role in the revised evaluation.

Table 2. Dataset composition and evaluation coverage.

Dataset component	Number	Role
Eligible non-TMA whole-slide images	513	Five-fold model development and OOF evaluation
High-resolution tissue patches	1,307,237	Source of pretrained CONCH embeddings
Median patches per slide	2,274	Variable-length slide representation
Pooled OOF predictions	513 per completed model	Final slide-level evaluation

Tissue microarray evaluation was not included because compatible CONCH features were unavailable. This subset is therefore treated as pending external evaluation rather than as a completed holdout test.

3.2. Dataset Partitioning and Evaluation Protocol

The 513 eligible non-TMA whole-slide images were partitioned into five stratified folds while preserving the distribution of the five ovarian cancer subtypes as closely as possible. In each cross-validation iteration, four folds were used to fit the slide-level aggregation classifier, and the remaining fold was used exclusively for validation. Each slide appeared in a validation fold exactly once and was not used to train the model that generated its prediction.

Precomputed CONCH feature bags were used as fixed, label-independent representations of the tissue patches. Model fitting, parameter estimation, and slide-level classification were performed independently within each fold. Checks of slide identifiers and fold assignments confirmed that no slide was shared between the training and validation subsets of the same fold.

The validation predictions from all five folds were concatenated to produce exactly 513 out-of-fold predictions for

every completed model. Principal performance metrics were calculated from these pooled slide-level predictions rather than from augmented patches or the average of patch-level results. Fold-level results were retained to describe variation across the five partitions, while 95% confidence intervals were estimated by resampling at the slide level. No separate 25-image holdout result was used in the revised evaluation.

3.3. High-Resolution Feature Extraction and Slide Representation

The revised analysis used precomputed patch embeddings from the official SMMILe OCEAN CONCH archive. The source extraction code identifies the encoder as CONCH ViT-B/16, although the exact checkpoint revision used to create the downloaded archive could not be independently verified. Each patch was represented by a 512-dimensional float32 feature vector. The CONCH encoder remained frozen, and only the slide-level aggregation models were trained. No stain normalization, stain augmentation, or geometric augmentation was applied by the local MIL pipeline.

The source archive reports 512×512-pixel patches extracted at an apparent magnification of 20×. The corresponding extraction specification uses a nominal stride of 512 pixels, indicating non-overlapping patches; however, the exact archive-generation command and slide metadata were unavailable for independent confirmation. Tissue regions were identified using a binary whole-slide mask in the official extraction pipeline, but its archive-specific threshold could not be authenticated. These source-reported settings are therefore distinguished from the parameters verified directly from the downloaded feature bags.

For slide *i*, the resulting feature bag was represented as

$$\mathbf{H}_i = \left[\mathbf{h}_{i1}, \mathbf{h}_{i2}, \ldots, \mathbf{h}_{im_i}\right]^{\mathsf{T}} \in \mathbb{R}^{m_i \times 512}, \tag{1}$$

where m is the number of patches available for that slide. The 513 bags contained 1,307,237 patches in total, with a median of 2,274 patches per slide and a range of 112-8,380. No L2 normalization or additional feature standardization was applied locally. During model fitting and internal validation, bags containing more than 512 patches were reduced by deterministic sampling without replacement. Complete bags were used to generate the final outer-fold out-of-fold predictions. The feature-bag construction process is illustrated in Fig. 1.

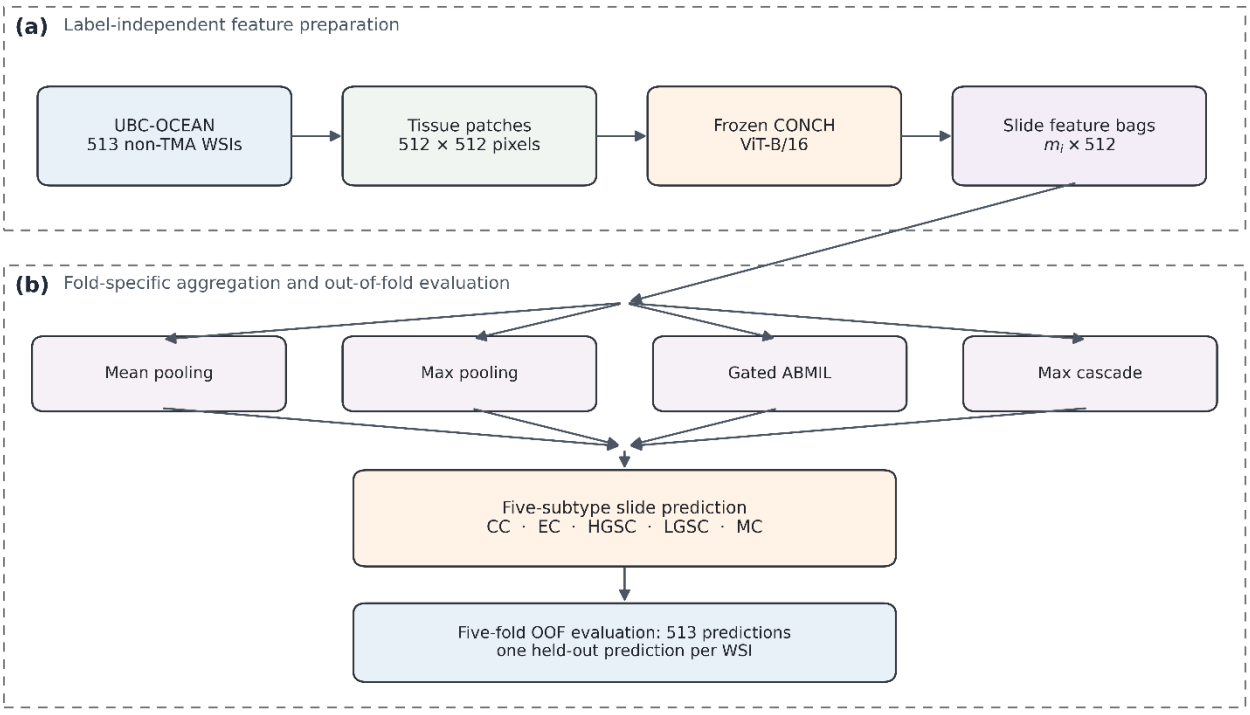


Fig. 1. CONCH-MIL ovarian subtype classification pipeline. (a) Label-independent construction of variable-length slide feature bags using the frozen CONCH encoder. (b) Fold-specific comparison of mean pooling, max pooling, gated ABMIL, and max-cascade aggregation for the classification of CC, EC, HGSC, LGSC, and MC. The five validation partitions produced exactly one out-of-fold prediction for each of the 513 eligible whole-slide images.

The verified feature-bag characteristics are summarized in Table 3.

Table 3. Verified characteristics of the high-resolution CONCH feature bags.

Parameter	Verified value
Feature source	Official SMMILe OCEAN CONCH archive
Encoder architecture	CONCH ViT-B/16, as reported by the source code
Embedding dimension and type	512, float32
Patch size	512×512 pixels
Apparent magnification	20×, source-reported
Nominal stride	512 pixels, source-reported
Number of WSIs	513
Total number of patches	1,307,237
Patches per WSI	Median 2,274; range 112–8,380
Feature-bag shape	$\mathbf{H}_i \in \mathbb{R}^{m_i \times 512}$
Coordinate-array shape	$\mathbb{Z}^{m_i \times 2}$
Storage format	One NPZ file per WSI
Local feature normalization	None
Training and internal-validation cap	Maximum 512 patches per WSI
Outer-fold OOF inference	Complete feature bag
CONCH fine-tuning	None; precomputed embeddings were frozen
Local image augmentation	None

The verified procedure used to construct and validate each slide-level feature bag is summarized in Algorithm 1.

Algorithm 1 Construction of a Slide-Level CONCH Feature Bag

Require: Locked WSI manifest, verified CONCH archive, and slide identifier i

Ensure: Validated slide-level feature bag $\mathbf{H}_i$

Verify the manifest and archive checksums
Match slide i to its corresponding archive member
Read the patch-feature matrix and convert it to float32
Read the patch indices and parse the integer spatial coordinates
Verify that each patch contains a 512-dimensional feature vector
Confirm agreement between the numbers of features and coordinates
Attach the slide identifier, subtype label, fold, encoder name, patch size, source information, and checksums
Save the feature matrix and coordinates as one NPZ file
Check for empty, non-finite, all-zero, or coordinate-duplicated bags
if the bag is used for model fitting or internal validation **then**
Deterministically sample at most 512 patches without replacement
else
Retain the complete bag for outer-fold OOF inference
end if
return $\mathbf{H}_i$

3.4. Data Leakage Prevention

Data leakage was controlled using a locked five-fold manifest covering all 513 eligible non-TMA whole-slide images. In each iteration, four folds were used for model fitting and the remaining fold was used exclusively for out-of-fold prediction. Slide identifiers in the training and validation partitions were checked for overlap, and each slide appeared in a validation fold exactly once. The five held-out partitions were subsequently concatenated to obtain exactly 513 slide-level out-of-fold predictions for every completed model.

CONCH feature extraction was performed independently of the subtype labels, and the downloaded embeddings remained fixed during model training. No data-dependent normalization was fitted across the complete dataset, and no stain or geometric augmentation was applied by the revised pipeline. During model fitting and internal validation, deterministic slide-specific sampling was used when a bag contained more than 512 patches. For final outer-fold prediction, the trained model was applied to the complete feature bag of each held-out slide.

Additional checks confirmed that the prediction files contained the same 513 unique slide identifiers, consistent subtype labels, finite probability values, and exactly one prediction per slide. Principal metrics were calculated from the concatenated out-of-fold predictions, while confidence intervals were estimated by resampling at the slide level. Patient-level independence could not be independently verified because authenticated patient identifiers were not available. Therefore, the leakage assessment is restricted to the slide-level evidence recorded in the project.

The leakage-controlled generation of slide-level out-of-fold predictions is summarized in Algorithm 2.

Algorithm 2 Leakage-Controlled Out-of-Fold Evaluation

Require: Locked slide manifest D, number of folds $K = 5$, and aggregation model M
Ensure: One out-of-fold prediction for each of the 513 slides

Verify the manifest checksum and unique slide identifiers
Partition D according to the locked stratified folds
Initialize an empty out-of-fold prediction set P
for $k \leftarrow 1$ to K **do**
Assign fold k to $\mathrm{D}_{\mathrm{val}}^{(k)}$
Assign the remaining $K-1$ folds to $\mathrm{D}_{\mathrm{train}}^{(k)}$
Verify that $\mathrm{D}_{\mathrm{train}}^{(k)} \cap \mathrm{D}_{\mathrm{val}}^{(k)} = \varnothing$
Load the fixed CONCH feature bags for the training slides
Deterministically cap training and internal-validation bags at 512 patches when necessary
Fit aggregation model $M^{(k)}$ without using the held-out fold
Load the complete feature bag for every slide in $\mathrm{D}_{\mathrm{val}}^{(k)}$
Generate slide-level class probabilities using $M^{(k)}$
Add the held-out predictions to P
end for
Verify that P contains 513 unique slide identifiers
Verify that every slide occurs exactly once in P
Verify label consistency and finite class probabilities
Compute pooled OOF metrics from P
Estimate confidence intervals by slide-level bootstrap resampling
return P and the corresponding evaluation metrics

3.5. Training Configuration and Checkpoint Selection

The executed training configuration was verified from the model code and saved checkpoints. No grid-search, random-search, or multi-seed experiment records were found; therefore, the revised analysis does not claim systematic hyperparameter optimization. A single predefined configuration was applied consistently across the five outer folds. Within each outer-training partition, 15% of the slides were assigned to a stratified internal-validation subset using random seed 42. This internal subset was used only for learning-rate scheduling, early stopping, and checkpoint selection.

Table 4. Verified training configuration for the CONCH-MIL models.

Parameter	Verified value
Loss	Class-weighted cross-entropy
Class weighting	Balanced weights from internal-training labels
Optimizer	AdamW
Learning rate	3×10^{-4}
Weight decay	1×10^{-4}
Maximum epochs	100
Internal-validation fraction	15% of outer-training slides
Checkpoint-selection metric	Internal-validation macro-F1
Early-stopping patience	12 epochs
Flat-model scheduler	ReduceLROnPlateau
Scheduler factor and patience	0.5 and 4 epochs
Cascade scheduler	None
Gradient clipping	Global norm of 1.0
Bag handling	One WSI per optimizer step
Training patch cap	512 patches per WSI
Outer-fold inference	Complete feature bag
Random seed	42
Independent repetitions	One seed per model and fold

The flat mean-pooling, max-pooling, and gated-ABMIL models were optimized using AdamW with a learning rate of 3×10^{-4} and weight decay of 1×10^{-4}. Class-weighted cross-entropy was used, with balanced weights calculated only from the internal-training labels. Training was limited to 100 epochs and stopped when internal-validation macro-F1 did

not improve for 12 epochs. A ReduceLROnPlateau scheduler reduced the learning rate by a factor of 0.5 after four epochs without improvement. Gradient norms were clipped at 1.0, and one slide-level feature bag was processed per optimizer step. The max-cascade stages used the same optimizer, loss, epoch limit, and early-stopping rule but did not use a learning-rate scheduler.

Each model was fitted once for every outer fold using seed 42. Thus, mean pooling, max pooling, and gated ABMIL each involved five fitted models, while the two-stage cascade involved ten separately fitted heads. Independent repeated-seed experiments were not performed. The verified training configuration is summarized in Table 4.

Checkpoint selection was nested within each outer fold because the internal-validation subset contained only outer-training slides. In contrast, the final comparison of mean pooling, max pooling, gated ABMIL, and the max cascade was based on their pooled outer-fold predictions and was not a nested model-family selection procedure. This distinction is retained when interpreting the comparative results.

3.6. Slide-Level Aggregation Models

All flat aggregation models received a variable-length slide feature bag $\mathbf{H}_i \in \mathbb{R}^{m_i \times 512}$. Each patch embedding was first projected from 512 to 256 dimensions using a linear layer, rectified linear unit activation, and dropout with a rate of 0.50:

$$\mathbf{h}_{ij} = \mathrm{Dropout}_{0.5}\left[\mathrm{ReLU}\left(\mathbf{W}_p \mathbf{x}_{ij} + \mathbf{b}_p\right)\right], \qquad \mathbf{h}_{ij} \in \mathbb{R}^{256}. \tag{2}$$

Mean pooling calculated the arithmetic mean of the projected patch features, whereas max pooling retained the elementwise maximum response:

$$\mathbf{z}_i^{\mathrm{mean}} = \frac{1}{m_i}\sum_{j=1}^{m_i}\mathbf{h}_{ij}, \qquad \mathbf{z}_i^{\max} = \max_{1\le j\le m_i}\mathbf{h}_{ij}. \tag{3}$$

Both operations produced a 256-dimensional slide representation. A second dropout layer with a rate of 0.50 was applied before the final linear classifier, which generated five logits ordered as CC, EC, HGSC, LGSC, and MC.

Gated Attention-Based Multiple-Instance Learning

The gated-ABMIL model used two parallel attention branches to assign a normalized weight to every projected patch [27]. For patch *j* from slide *i*, the two 128-dimensional gated representations were calculated as

$$\mathbf{v}_{ij} = \tanh\left(\mathbf{V}\mathbf{h}_{ij} + \mathbf{b}_V\right), \qquad \mathbf{u}_{ij} = \sigma\left(\mathbf{U}\mathbf{h}_{ij} + \mathbf{b}_U\right). \tag{4}$$

The branch outputs were multiplied elementwise, and a bias-free linear layer produced one attention logit per patch:

$$s_{ij} = \mathbf{w}^{\mathsf{T}}\left(\mathbf{v}_{ij} \odot \mathbf{u}_{ij}\right). \tag{5}$$

Attention was normalized across the patch dimension:

$$a_{ij} = \frac{\exp(s_{ij})}{\sum_{r=1}^{m_i}\exp(s_{ir})}, \qquad \sum_{j=1}^{m_i} a_{ij} = 1. \tag{6}$$

For padded inputs, invalid patch logits were assigned the minimum representable value before softmax normalization. The attention-weighted slide representation was then obtained as

$$\mathbf{z}_i^{\max} = \sum_{j=1}^{n} a_{ij}\,\mathbf{h}_{ij}, \mathbf{z}_i^{\max} \in \mathbb{R}^{256}. \tag{7}$$

The five-class logits and predicted probabilities were calculated as

$$\ell_l = W_c Dropout_{0.5}(\mathbf{z}_l^{\mathrm{stt}}) + \mathbf{b}_c, \tilde{\mathbf{p}}_l = softmax(\ell_l). \tag{8}$$

The executed attention-hidden dimension was 128. This value was verified from the source code and checkpoint tensors and replaces the unused value of 256 recorded in the generic configuration file. The gated-ABMIL architecture and the change in feature dimensions are illustrated in Fig. 2.

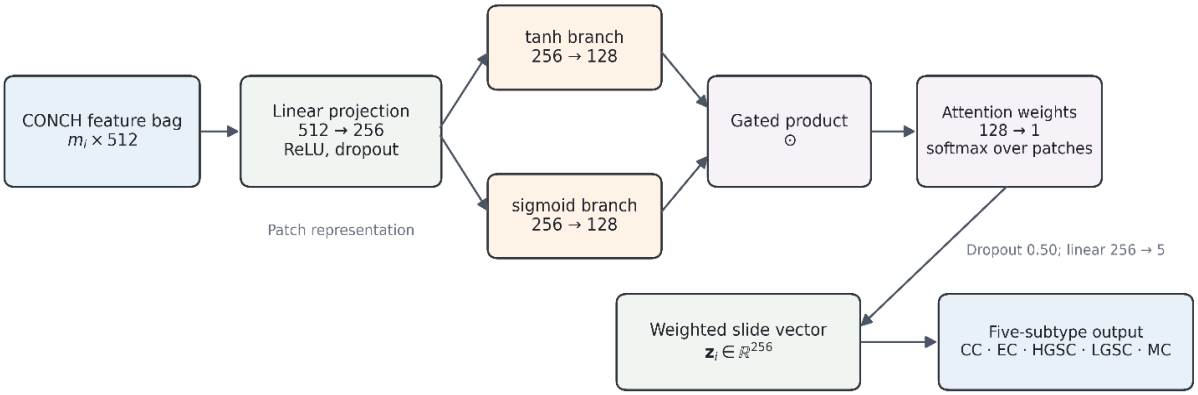


Fig. 2. Gated-ABMIL architecture used for slide-level ovarian cancer subtype classification. Each 512-dimensional CONCH patch embedding was projected to 256 dimensions. Parallel tanh and sigmoid branches produced 128-dimensional gated features, from which normalized patch-attention weights were obtained. The weighted slide representation was classified into CC, EC, HGSC, LGSC, and MC.

Probabilistic Max Cascade

The max cascade comprised two independently trained max-pooling heads. Stage 1 separated HGSC from all non-HGSC subtypes, while Stage 2 classified the non-HGSC slides as CC, EC, LGSC, or MC. Both heads were evaluated for every held-out slide, and their probabilities were combined without hard routing:

$$P(\text{HGSC}) = P_1(\text{HGSC}), \qquad P(c) = P_1(\text{nonHGSC})P_2(c|\ \text{nonHGSC}), \tag{9}$$

where $c \in \{\text{CC}, \text{EC}, \text{LGSC}, \text{MC}\}$. The composed probabilities were ordered as CC, EC, HGSC, LGSC, and MC and were verified to sum to one within a numerical tolerance of 10^{-6}.

The dimensions and trainable parameter counts of the four aggregation approaches are summarized in Table 5.

Table 5. Feature dimensions and trainable parameters of the slide-level aggregation models.

Model	Feature dimensions	Parameters
Mean pooling	$m_i \times 512 \rightarrow m_i \times 256 \rightarrow 256 \rightarrow 5$	132,613
Max pooling	$m_i \times 512 \rightarrow m_i \times 256 \rightarrow 256 \rightarrow 5$	132,613
Gated ABMIL	$m_i \times 512 \rightarrow m_i \times 256 \rightarrow m_i \times 128 \rightarrow 256 \rightarrow 5$	198,533
Max cascade, Stage 1	$m_i \times 512 \rightarrow m_i \times 256 \rightarrow 256 \rightarrow 2$	131,842
Max cascade, Stage 2	$m_i \times 512 \rightarrow m_i \times 256 \rightarrow 256 \rightarrow 4$	132,356
Max cascade, combined	Two independently trained max-pooling heads with a composed five-class output	264,198

3.7. Implementation Details

The revised experiments were implemented in PyTorch. The preserved thumbnail-baseline environment contained PyTorch 2.6.0 with CUDA 12.4 and TorchVision 0.21.0. The CONCH-MIL pipeline loaded one float32 NPZ feature bag per whole-slide image, while the thumbnail baselines used RGB images processed through TorchVision-compatible transformations. Model checkpoints, fold-level predictions, training histories, and pooled out-of-fold predictions were stored separately for every model and fold.

All experiments used the same locked five-fold manifest and random seed 42. Python, NumPy, PyTorch, and CUDA random-number generators were initialized using this seed. The thumbnail CNNs used mixed-precision CUDA training, whereas the MIL models processed one slide feature bag per optimizer step. Gradient norms were clipped at 1.0 in both pipelines. Complete slide-level feature bags were used for outer-fold MIL prediction, while validation thumbnails were evaluated using deterministic centre crops.

The executed settings recorded in the source code and checkpoints were treated as authoritative whenever they differed from generic configuration files. In particular, the thumbnail CNN batch size was 16 rather than 32, the gated-attention hidden dimension was 128 rather than 256, and the optimizer was AdamW rather than Adam. No executed grid search or independent repeated-seed experiment was identified; consequently, neither is claimed in the revised manuscript. The model-specific configurations are reported in Tables 4, 5, and 6.

3.8. Thumbnail-Based CNN Baselines

Five pretrained CNN architectures were evaluated as image-level baselines: ResNet18, ResNet50, ResNeXt50-32x4d, EfficientNet-B0, and ConvNeXt-Tiny. These models operated on downsampled whole-slide thumbnails and were included to provide a conventional image-classification comparison with the high-resolution CONCH-MIL

models. They were not trained on the extracted high-resolution tissue patches.

ResNet18 and ResNet50 use residual connections to support optimization at different network depths [34]. ResNeXt50-32x4d extends the residual design using grouped convolutions and increased cardinality [35]. EfficientNet-B0 provides a comparatively compact architecture based on compound scaling, whereas ConvNeXt-Tiny represents a more recent convolutional design incorporating architectural principles associated with modern vision models. All five implementations were obtained from TorchVision and initialized using ImageNet-1K pretrained weights.

Each official UBC-OCEAN thumbnail was converted to RGB and resized with its aspect ratio preserved so that the shorter side contained 256 pixels. During training, a random 224 × 224 crop was generated using a scale range of 0.8-1.0. Random horizontal and vertical flips, rotations within ±15°, translations of up to 3%, and scaling between 0.95 and 1.05 were applied only to the training images. Validation thumbnails were resized to a shorter side of 256 pixels and centre-cropped to 224 × 224. Both training and validation inputs were normalized using the ImageNet channel means (0.485, 0.456, 0.406) and standard deviations (0.229, 0.224, 0.225).

The original classifier of each network was replaced with a five-output layer ordered as CC, EC, HGSC, LGSC, and MC. All backbone and classifier parameters were fine-tuned. ResNet18, ResNet50, ResNeXt50-32x4d, and ConvNeXt-Tiny used no additional classifier dropout. EfficientNet-B0 retained its original classifier dropout rate of 0.20. The verified classifier dimensions and parameter counts are summarized in Table 6.

Table 6. Architecture of the thumbnail-based CNN baselines.

Model	Pretraining	Five-class head	Trainable parameters
ResNet18	ImageNet-1K V1	Linear 512 → 5	11,179,077
ResNet50	ImageNet-1K V1	Linear 2048 → 5	23,518,277
ResNeXt50-32x4d	ImageNet-1K V1	Linear 2048 → 5	22,990,149
EfficientNet-B0	ImageNet-1K V1	Dropout 0.20; Linear 1280 → 5	4,013,953
ConvNeXt-Tiny	ImageNet-1K V1	LayerNorm; Linear 768 → 5	27,823,973

The five CNN baselines used the same locked stratified folds as the CONCH-MIL models. Training used class-weighted cross-entropy, with weights calculated separately from the training portion of each fold according to

$$w_c = \frac{N_{\text{train}}}{5n_c}, \tag{10}$$

where N_{train} is the number of training slides and n_c is the number belonging to class c. The models were optimized using AdamW with a learning rate of 3×10^{-4} and weight decay of 1×10^{-4}. The batch size was 16, and training was limited to 50 epochs. A ReduceLROnPlateau scheduler monitored validation macro-F1 with a patience of two epochs, while early stopping used the same metric with a patience of seven epochs. Gradient norms were clipped at 1.0, mixed-precision training was enabled on CUDA, and all experiments used random seed 42. Each architecture was fitted once within each of the five folds; independent repeated-seed training was not performed.

4. Results

4.1. Pooled Out-of-Fold Performance of Thumbnail CNN Baselines

The five thumbnail-based CNN baselines were evaluated using the same locked five-fold protocol. Predictions from the five held-out folds were combined so that each model produced exactly 513 unique slide-level out-of-fold predictions. Performance was calculated from these pooled predictions rather than from training accuracy or the mean of patch-level results.

Table 7 presents the pooled results. Among the thumbnail models, EfficientNet-B0 achieved the highest accuracy, balanced accuracy, and macro-F1. ConvNeXt-Tiny produced the highest macro-AUPRC, while EfficientNet-B0 achieved the highest macro-AUROC. The difference between ordinary accuracy and balanced accuracy across the models also illustrates the importance of reporting class-balanced measures for the imbalanced five-subtype dataset.

Table 7. Pooled out-of-fold performance of the thumbnail-based CNN baselines.

Model	Accuracy	Balanced accuracy	Macro-F1	Macro-AUROC	Macro-AUPRC
ResNet18	0.4269	0.4761	0.4322	0.7559	0.4845
ResNet50	0.4561	0.4888	0.4526	0.7481	0.4576
ResNeXt50-32x4d	0.4620	0.5163	0.4721	0.7757	0.5239
EfficientNet-B0	**0.5127**	**0.5573**	**0.5229**	**0.7901**	0.5208
ConvNeXt-Tiny	0.4951	0.5158	0.5039	0.7824	**0.5273**

Although EfficientNet-B0 was the strongest thumbnail baseline according to the principal macro-F1 measure, its performance remained constrained by the low-resolution input representation. These models therefore provide an image-level reference for evaluating whether high-resolution CONCH feature bags preserve information that is lost when a complete whole-slide image is reduced to a single thumbnail.

4.2. Performance of CONCH Aggregation Models

The high-resolution CONCH feature bags were evaluated using mean pooling, max pooling, gated ABMIL, and a probabilistic two-stage max-pooling cascade. Each aggregation model produced exactly 513 unique slide-level OOF predictions. Balanced accuracy and macro-F1 were treated as the principal measures because the five ovarian cancer subtypes were not equally represented. The pooled results are presented in Table 8.

Table 8. Pooled out-of-fold performance of the CONCH aggregation models.

Aggregation model	Balanced accuracy	Macro-F1
Mean pooling	0.7930	0.8020
Max pooling	**0.8144**	**0.8259**
Gated ABMIL	0.8139	0.8196
Max-pooling cascade	0.8101	0.8146

Max pooling achieved the highest pooled balanced accuracy and macro-F1, although gated ABMIL produced numerically similar performance. Max pooling achieved a macro-F1 of 0.8259 with a 95% confidence interval of 0.7884-0.8626 and a balanced accuracy of 0.8144 with a 95% confidence interval of 0.7736-0.8552. Its macro-AUROC and macro-AUPRC were 0.9699 and 0.9079, respectively. The complete confidence intervals for max pooling and gated ABMIL are reported in Table 9.

Gated ABMIL improved macro-F1 by 0.0177 relative to mean pooling, but the 95% confidence interval for this difference extended from –0.0189 to 0.0533 and included zero. The observed advantage of gated attention over mean pooling was therefore uncertain. Similarly, the max-pooling cascade did not improve upon flat max pooling. Max pooling was consequently treated as the numerically strongest parameter-efficient aggregation strategy, while gated ABMIL was retained as the principal attention-guided model.

Overall, the relatively small differences among the four CONCH aggregation strategies indicate that increasing aggregation complexity did not produce a clear predictive advantage on this cohort. These findings motivate the paired statistical comparisons reported in Section 4.8.

4.3. Comparative Out-of-Fold Performance

Fig. 3 compares the principal slide-level metrics of the thumbnail CNN baselines and the high-resolution CONCH aggregation models. The plotted estimates were calculated from the pooled out-of-fold predictions, with each of the 513 slides contributing exactly one held-out prediction per completed model. Fold-level observations and slide-level bootstrap confidence intervals are shown where available.

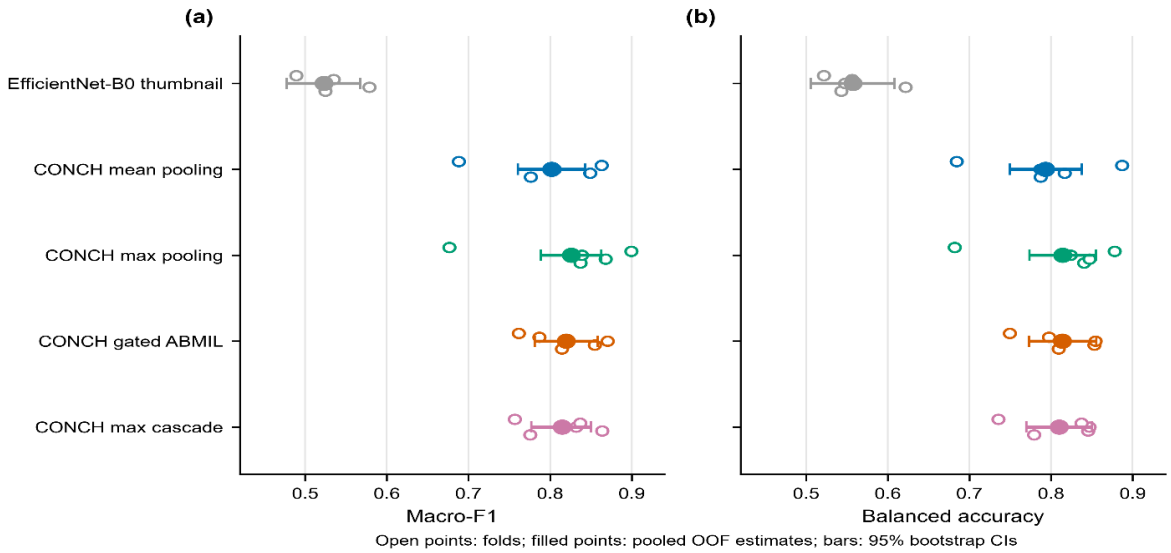


Fig. 3. Pooled out-of-fold performance of the strongest thumbnail baseline and four CONCH aggregation models. Circles denote macro-F1, squares denote balanced accuracy, and horizontal bars show 95% slide-level bootstrap confidence intervals.

The CONCH-based models consistently produced higher macro-F1 and balanced accuracy than the thumbnail CNN baselines. Relative to EfficientNet-B0, the strongest thumbnail baseline, max pooling increased macro-F1 from 0.5229 to 0.8259, an absolute difference of 0.3030. Balanced accuracy increased from 0.5573 to 0.8144, an absolute difference of 0.2571. These differences indicate that retaining high-resolution patch representations was more consequential than changing among the evaluated thumbnail CNN backbones.

Within the CONCH models, max pooling produced the highest numerical macro-F1, followed by gated ABMIL, the max cascade, and mean pooling. The relatively small differences among these aggregation methods contrast with the larger separation between the thumbnail and high-resolution representations. In particular, gated attention did not demonstrate a clear predictive advantage over the simpler pooling strategies.

4.4. Overall and Classwise Out-of-Fold Performance

Overall and classwise performance was calculated exclusively from pooled slide-level out-of-fold predictions. Both max pooling and gated ABMIL produced 513 predictions for 513 unique slides, with the class order fixed as CC, EC, HGSC, LGSC, and MC. Confidence intervals were estimated from 2,000 true-class-stratified slide-level bootstrap resamples using seed 20260819 and the 2.5th and 97.5th percentiles. The same bootstrap samples were used for both models.

Table 9 reports the overall results. Max pooling achieved a macro-F1 of 0.8259 and balanced accuracy of 0.8144, while gated ABMIL achieved corresponding values of 0.8196 and 0.8139. The confidence intervals overlapped substantially across all reported measures, and the two models therefore showed similar overall performance. Max pooling remained numerically stronger for accuracy, macro-precision, macro-F1, macro-AUROC, and macro-AUPRC.

Table 9. Pooled slide-level OOF performance with 95% bootstrap confidence intervals.

Metric	CONCH max pooling	CONCH gated ABMIL
Accuracy	0.8460 [0.8168, 0.8752]	0.8402 [0.8090, 0.8713]
Balanced accuracy	0.8144 [0.7736, 0.8552]	0.8139 [0.7727, 0.8556]
Macro-precision	0.8442 [0.8085, 0.8822]	0.8265 [0.7892, 0.8681]
Macro-recall	0.8144 [0.7736, 0.8552]	0.8139 [0.7727, 0.8556]
Macro-F1	0.8259 [0.7884, 0.8626]	0.8196 [0.7813, 0.8582]
Macro-specificity	0.9551 [0.9466, 0.9638]	0.9544 [0.9453, 0.9633]
Macro-AUROC	0.9699 [0.9608, 0.9790]	0.9673 [0.9566, 0.9775]
Macro-AUPRC	0.9079 [0.8812, 0.9352]	0.8975 [0.8662, 0.9294]

The classwise results are summarized in Table 10. Both models classified CC and HGSC comparatively well. EC was the most challenging subtype, with a recall of 0.6891 for both models. Max pooling provided higher HGSC and MC recall, whereas gated ABMIL provided higher CC and LGSC recall. These differences indicate that the similar macro-level results were accompanied by different subtype-specific error patterns.

Table 10. Classwise performance of the two principal CONCH models.

Model	Class	Support	Precision	Recall	F1	AUROC	AUPRC
Max pooling	CC	94	0.9341	0.9043	0.9189	0.9939	0.9818
Max pooling	EC	119	0.7961	0.6891	0.7387	0.9297	0.8281
Max pooling	HGSC	217	0.8458	0.9355	0.8884	0.9658	0.9570
Max pooling	LGSC	42	0.9118	0.7381	0.8158	0.9831	0.8895
Max pooling	MC	41	0.7333	0.8049	0.7674	0.9773	0.8833
Gated ABMIL	CC	94	0.9355	0.9255	0.9305	0.9939	0.9773
Gated ABMIL	EC	119	0.7593	0.6891	0.7225	0.9294	0.8030
Gated ABMIL	HGSC	217	0.8571	0.9124	0.8839	0.9660	0.9583
Gated ABMIL	LGSC	42	0.7805	0.7619	0.7711	0.9680	0.8644
Gated ABMIL	MC	41	0.8000	0.7805	0.7901	0.9792	0.8848

The high macro-specificity values should be interpreted together with recall and precision because specificity was calculated in a one-vs-rest setting, where the negative class was considerably larger for each subtype. Similarly, the high AUROC values do not remove the need to examine precision, recall, F1, and AUPRC for the less frequent subtypes. No external or TMA validation is claimed from these results.

4.5. Receiver Operating Characteristic and Precision-Recall Analysis

One-vs-rest receiver operating characteristic (ROC) and precision-recall (PR) curves were generated from the 513 pooled out-of-fold probability predictions of the gated-ABMIL model. ROC analysis describes the relationship between sensitivity and false-positive rate across decision thresholds, whereas PR analysis is particularly informative for the less

frequent ovarian cancer subtypes because it relates precision to recall.

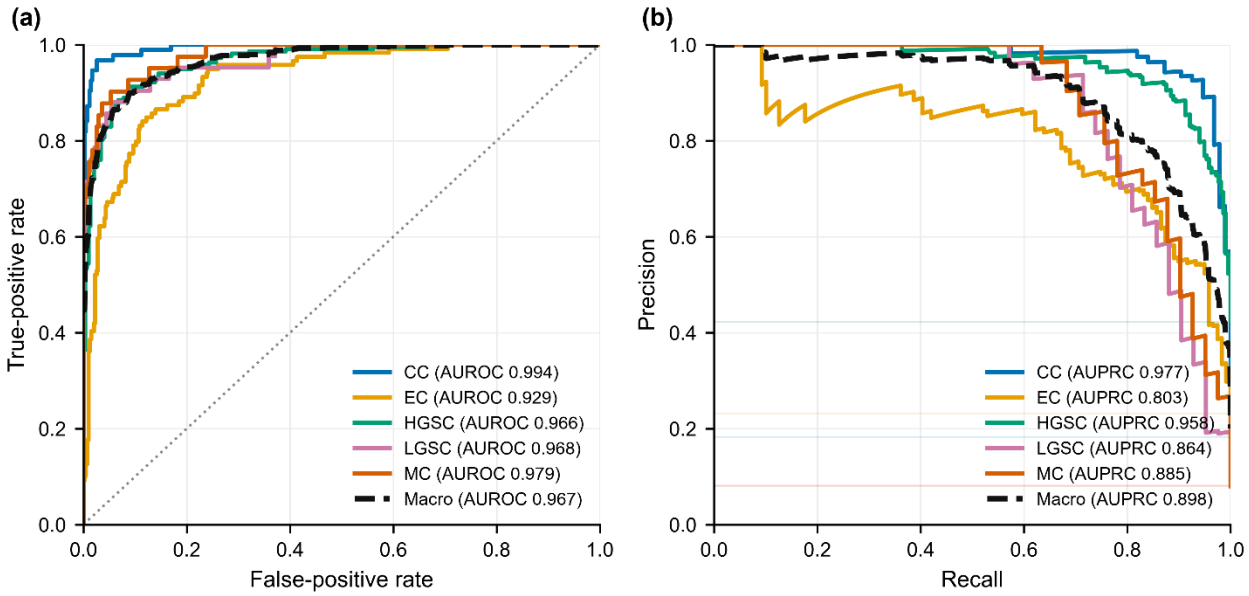


Fig. 4. One-vs-rest discrimination of gated ABMIL using pooled out-of-fold predictions. (a) ROC curves. (b) Precision-recall curves for CC, EC, HGSC, LGSC, and MC.

As shown in Fig. 4., the classwise AUROC values were 0.9939 for CC, 0.9294 for EC, 0.9660 for HGSC, 0.9680 for LGSC, and 0.9792 for MC. The corresponding AUPRC values were 0.9773, 0.8030, 0.9583, 0.8644, and 0.8848, respectively. The gated model achieved a macro-AUROC of 0.9673 and a macro-AUPRC of 0.8975.

CC showed the strongest threshold-independent discrimination, whereas EC had the lowest AUROC and AUPRC. The reduction was more apparent in the PR analysis, which is consistent with the lower precision and recall observed for EC. Although the overall discrimination was strong, these subtype-level differences show that high macro-AUROC alone does not imply uniform classification performance across all five subtypes.

4.6. Out-of-Fold Confusion Matrix and Classwise Recall

The pooled confusion matrix and classwise recall of gated ABMIL were calculated from 513 unique out-of-fold predictions. Each row represents the true subtype, while each column represents the predicted subtype. Unlike the former 25-image holdout matrix, this analysis includes every eligible non-TMA whole-slide image exactly once.

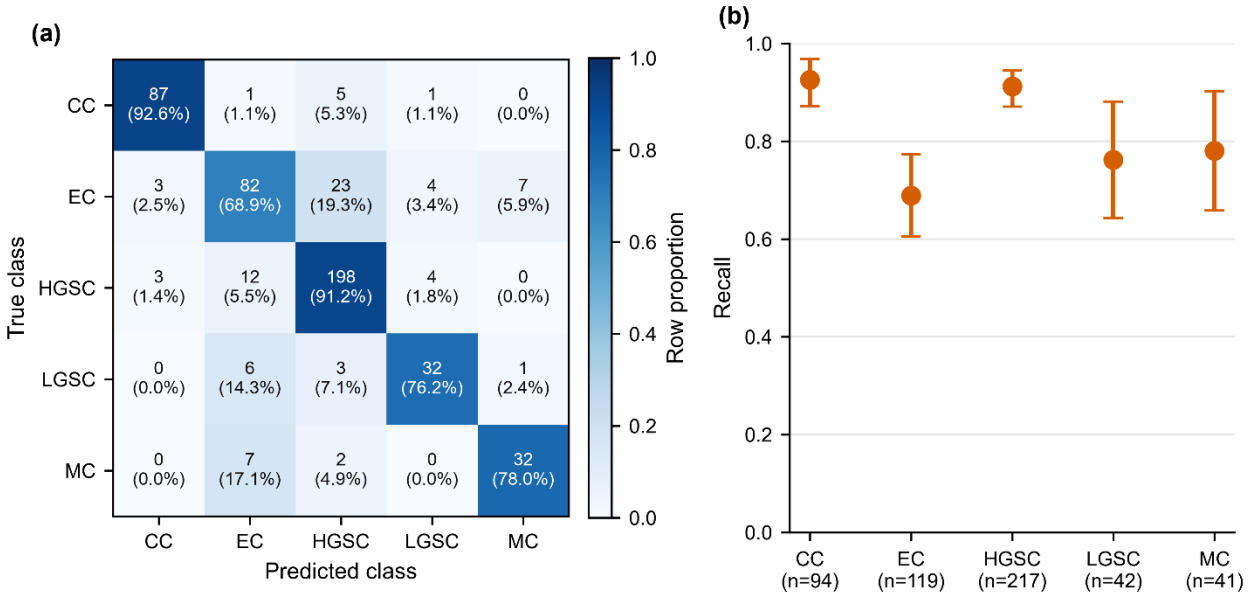


Fig. 5. Classwise performance of gated ABMIL using 513 pooled out-of-fold predictions. (a) Confusion matrix showing counts and row-normalized percentages. (b) Classwise recall with 95% slide-level bootstrap confidence intervals; class supports are shown below the subtype labels.

The classwise confusion pattern is shown in Fig. 5. The model correctly classified 87 of 94 CC slides, 82 of 119 EC slides, 198 of 217 HGSC slides, 32 of 42 LGSC slides, and 32 of 41 MC slides. Corresponding recalls were 0.9255 for CC, 0.6891 for EC, 0.9124 for HGSC, 0.7619 for LGSC, and 0.7805 for MC.

EC was the most difficult subtype to identify, with 23 EC slides classified as HGSC and seven classified as MC. HGSC also accounted for the largest source of error among misclassified CC slides. In contrast, CC and HGSC achieved recalls above 0.90. These findings demonstrate that the overall accuracy of 0.8402 was accompanied by meaningful subtype-specific variation, which would have been obscured by reporting accuracy alone.

No performance claim is made for the former 25-image holdout set. The confusion matrix, accuracy, and classwise results in the revised manuscript are derived exclusively from leakage-controlled slide-level out-of-fold predictions.

4.7. Contextual Comparison with Previous Ovarian Histopathology Studies

Table 11 places the present results in the context of recent studies addressing five-subtype ovarian carcinoma classification. Direct numerical ranking is inappropriate because the studies differ in their datasets, prediction units, feature extractors, validation protocols, and reported metrics. In particular, results obtained from an image-level train-test split are not directly comparable with pooled slide-level out-of-fold estimates or evaluations on independent external cohorts.

Table 11. Contextual comparison with recent five-subtype ovarian histopathology classification studies.

Study	Dataset and prediction unit	Model	Validation and reported result	Comparability consideration
Behera et al. [31]	UBC-OCEAN-derived collection; 725 images; image-level classification	EfficientNet-B0 feature extraction with fine-KNN	80:20 split; 100% accuracy reported; classwise AUROCs of 0.94, 0.78, 0.69, 0.92, and 0.94	Single image-level split without reported external validation; not directly comparable with slide-level OOF evaluation
Sai and Punn [21]	UBC-OCEAN; WSI-level five-subtype classification	K-stage Transformer MIL	Reported balanced accuracy of 0.933 and AUROC of 0.990	Validation details and several macro-averaged metrics were not reported
Breen et al. [30]	1,864 development WSIs with internal and two external test cohorts	ABMIL using H-optimus-0 and other pathology foundation-model features	Reported balanced accuracies of 0.89 internally and 0.97 and 0.74 on the two external cohorts	Different development cohort and five-fold ensemble evaluation; provides stronger external-validation evidence
Song et al. [29]	UBC-OCEAN; 510 quality-controlled WSIs	Phikon-v2, CTransPath, and UNI features with six aggregation methods and hierarchical classification	Stratified five-fold cross-validation; reported balanced accuracy ranged from 0.538 to 0.925 across feature-extractor and classification settings	Most closely related comparison, but preprocessing, embeddings, retained slides, and aggregation implementations differ
Chun et al. [32]	Multicentre cohort; 319 slides from 152 patients; slide- and patient-level predictions	Attention-based MIL using UNI features	Five-fold cross-validation with external TCGA evaluation; reported accuracy of 0.918 at slide level and 0.900 at patient level	Different cohort composition and prediction units; the reported F1 ranges are classwise rather than macro-averaged results
Present work	UBC-OCEAN; 513 non-TMA WSIs; one pooled OOF prediction per slide	Frozen CONCH features with mean pooling, max pooling, gated ABMIL, and a max-pooling cascade	Leakage-controlled five-fold OOF evaluation; max pooling achieved balanced accuracy 0.8144 (95% CI: 0.7736-0.8552), macro-F1 0.8259 (0.7884-0.8626), and macro-AUROC 0.9699 (0.9608-0.9790)	Controlled comparison of aggregation methods using identical CONCH embeddings and folds; no independent external validation

The comparison demonstrates that high performance has been reported using both pathology foundation-model representations and transformer- or attention-based MIL approaches [29-32]. Nevertheless, the reported results also reflect substantial differences in cohort composition and evaluation design. Behera et al. used an image-level 80:20 split [21], whereas Song et al. evaluated multiple feature extractors and aggregation methods using stratified WSI-level cross-validation [31]. Breen et al. additionally employed independent external cohorts and observed dataset-dependent performance [29]. These protocol differences prevent a statistically defensible conclusion that one method universally outperforms the others.

The contribution of the present study is therefore the controlled evaluation of four slide-level aggregation strategies using the same frozen CONCH patch representations, locked folds, optimization procedure, and pooled OOF assessment. Under these conditions, max pooling produced the highest macro-F1, although its confidence intervals overlapped those of gated ABMIL and the other CONCH aggregation approaches. This finding suggests that the high-resolution feature representation contributed substantially to classification performance and that greater aggregation complexity did not provide a clear advantage on this cohort. External multi-institutional validation remains necessary before clinical generalizability can be established.

4.8. Paired Statistical Comparison of Model Performance

Differences between models were evaluated using their paired slide-level OOF predictions. The statistical unit was one WSI, and all comparisons used the same 513 slide identifiers, true labels, fold assignments, and class order. A total of 2,000 class-stratified paired bootstrap samples were generated using a fixed random seed of 20260819. The same resampled slide indices were applied to both models within each comparison, thereby preserving the paired evaluation design.

For a performance metric denoted by *M*, the difference between models *A* and B was calculated as

$$\Delta_M = M_A - M_B. \tag{11}$$

The confidence interval at the 95% level was obtained from the 2.5th and 97.5th percentiles of the bootstrap distribution. The two-sided bootstrap *p*-value was calculated as

$$p = 2\min\left\{\Pr\left(\Delta_M^* \le 0\right), \Pr\left(\Delta_M^* \ge 0\right)\right\}, \tag{12}$$

with values greater than 1 truncated to 1. Holm correction was applied across the prespecified family of 16 comparisons, comprising four model contrasts and four performance metrics. The macro-F1 and balanced-accuracy comparisons most relevant to model selection are reported in Table 12.

Table 12. Paired slide-level bootstrap comparisons of model performance.

Comparison	Metric	Difference	95% bootstrap CI	*p*-value	Holm-adjusted *p*
Max pooling minus gated ABMIL	Macro-F1	0.0062	[−0.0274, 0.0379]	0.734	1.000
Max pooling minus gated ABMIL	Balanced accuracy	0.0005	[−0.0342, 0.0327]	0.984	1.000
Gated ABMIL minus mean pooling	Macro-F1	0.0177	[−0.0189, 0.0533]	0.339	1.000
Gated ABMIL minus mean pooling	Balanced accuracy	0.0208	[−0.0177, 0.0563]	0.293	1.000
Max pooling minus mean pooling	Macro-F1	0.0239	[−0.0103, 0.0588]	0.155	1.000
Max pooling minus mean pooling	Balanced accuracy	0.0213	[−0.0147, 0.0575]	0.254	1.000
Max pooling minus max cascade	Macro-F1	0.0112	[−0.0206, 0.0438]	0.513	1.000
Max pooling minus max cascade	Balanced accuracy	0.0043	[−0.0284, 0.0388]	0.807	1.000

Max pooling produced the highest pooled macro-F1 among the four CONCH aggregation approaches. However, none of its paired differences from gated ABMIL, mean pooling, or the max cascade was statistically significant after Holm correction. All corresponding confidence intervals included zero. Accordingly, the observed ordering should be interpreted as a numerical ranking on the present cohort rather than evidence that max pooling is consistently superior.

Fold-level standardized effect sizes were available for macro-F1. For paired fold differences Δ_k, Cohen's d_z was calculated as

$$d_z = \frac{\bar{\Delta}}{s_\Delta}, \tag{13}$$

where $\bar{\Delta}$ and s_Δ denote the mean and sample standard deviation of the five paired outer-fold differences. The resulting values were 0.088 for max pooling versus gated ABMIL, 0.350 for gated ABMIL versus mean pooling, 1.038 for max pooling versus mean pooling, and 0.189 for max pooling versus the max cascade. These estimates should be interpreted cautiously because they are based on only five folds. In particular, the apparently large fold-standardized effect for max pooling versus mean pooling was not supported by the slide-level bootstrap analysis, whose confidence interval included zero.

An additional paired comparison was conducted between CONCH max pooling and the EfficientNet-B0 thumbnail classifier because both methods generated OOF predictions for the same 513 slides. Max pooling improved macro-F1 by 0.3030 (95% CI: 0.2532-0.3559; $p<0.001$) and balanced accuracy by 0.2570 (95% CI: 0.2007-0.3173; $p<0.001$). Because this comparison was calculated separately and was not included in the prespecified 16-comparison family, a Holm-adjusted *p*-value is not reported. These differences support the benefit of high-resolution CONCH patch representations over a single downsampled WSI thumbnail under the present evaluation protocol.

4.9. Comparison of Slide-Level Aggregation Strategies

The influence of slide-level feature aggregation was examined by comparing mean pooling, max pooling, gated ABMIL, and a two-stage max-pooling cascade. All four approaches used the same frozen 512-dimensional CONCH

patch embeddings, locked five-fold partitions, training patch cap, and leakage-controlled OOF evaluation protocol. Consequently, differences among these models primarily reflect the aggregation and classification strategies rather than changes in the underlying patch representation or data partitioning.

Mean pooling calculated the arithmetic mean of the projected patch features, whereas max pooling retained the largest response within each projected feature dimension. Gated ABMIL learned normalized patch-level attention weights before constructing the slide representation. The max-pooling cascade first separated HGSC from non-HGSC cases and subsequently classified non-HGSC slides as CC, EC, LGSC, or MC. The probabilities from the two cascade stages were combined to obtain a five-class prediction.

Table 13. Pooled out-of-fold performance of the CONCH aggregation strategies.

Aggregation strategy	**Accuracy**	Balanced accuracy	Macro-F1	Macro-AUROC	Macro-AUPRC
Mean pooling	–	0.7930	0.8020	–	–
Max pooling	0.8460	0.8144	0.8259	0.9699	0.9079
Gated ABMIL	0.8402	0.8139	0.8196	0.9673	0.8975
Max-pooling cascade	–	0.8101	0.8146	–	–

As shown in Table 13, max pooling obtained the highest pooled macro-F1 and balanced accuracy, followed closely by gated ABMIL and the max-pooling cascade. Mean pooling produced the lowest values among the four CONCH approaches. Nevertheless, the differences between max pooling and the other aggregation strategies were small, and their paired bootstrap confidence intervals included zero, as reported in Table 12. The results therefore do not establish a statistically significant advantage for any single CONCH aggregation strategy after correction for multiple comparisons.

The comparison indicates that a relatively simple max-pooling operator can remain competitive when applied to strong high-resolution pathology foundation-model features. Gated attention and hierarchical classification did not provide a clear aggregate advantage on this cohort, although they may offer different subtype-specific behaviour or interpretability properties.

5. Discussion and Limitations

The principal finding is that high-resolution CONCH patch representations supported substantially stronger five-subtype classification than ImageNet-pretrained CNNs operating on downsampled WSI thumbnails. CONCH max pooling improved macro-F1 and balanced accuracy over the strongest thumbnail baseline, EfficientNet-B0, and the corresponding paired bootstrap confidence intervals excluded zero. This result suggests that preserving local cellular and tissue morphology was more consequential on this cohort than changing among conventional thumbnail-level CNN backbones. However, this comparison reflects both the feature-pretraining source and the input representation: CONCH was pretrained on pathology data and operated on high-resolution tissue patches, whereas the CNN baselines were pretrained on natural images and received a single downsampled slide thumbnail. The observed difference should therefore not be attributed exclusively to either resolution or encoder pretraining.

Among the CONCH models, max pooling achieved the highest numerical macro-F1 and balanced accuracy, but its differences from mean pooling, gated ABMIL, and the max-pooling cascade were not statistically significant after Holm correction. The result does not establish that max pooling is universally superior; rather, it indicates that greater aggregation complexity did not provide a measurable overall advantage under the present experimental conditions. Strong pretrained patch embeddings may already encode much of the subtype-relevant morphology, allowing a simple elementwise maximum to retain discriminative signals. Conversely, gated attention may distribute weight across several informative regions and could remain useful for visual examination or subtype-specific behaviour even without improving the aggregate metrics.

The classwise analysis revealed that performance was not uniform across the five subtypes. CC and HGSC achieved comparatively high recall, whereas EC was the most difficult subtype for both principal CONCH models. A substantial proportion of EC errors involved prediction as HGSC, indicating residual morphological overlap or insufficient representation of discriminative EC patterns. The less frequent LGSC and MC classes also exhibited wider uncertainty and lower recall than the strongest classes. These findings demonstrate why ordinary accuracy and macro-AUROC should be interpreted together with balanced accuracy, macro-F1, AUPRC, and classwise measures in an imbalanced multiclass setting.

The present findings are broadly consistent with recent ovarian histopathology research showing that pathology foundation-model features can support strong WSI-level subtype classification [29]. Recent controlled evaluations have also indicated that feature-extractor choice may influence performance more strongly than the downstream aggregation strategy [31]. Nevertheless, direct numerical ranking against previous studies is not appropriate because published investigations differ in dataset composition, quality-control procedures, patch magnification, prediction unit, validation design, and external-testing availability. The contextual comparison in Table 11 should therefore be interpreted as

evidence of methodological positioning rather than as a performance leaderboard.

Several limitations should be considered. First, the evaluation used a single public dataset and did not include an independent external or cross-institutional cohort. Generalizability across laboratories, scanners, staining protocols, specimen types, and patient populations therefore remains unknown. Second, authenticated patient identifiers were unavailable, and patient-level independence could not be verified. The leakage audit establishes slide-level separation but cannot exclude the possibility of correlated slides from the same patient. Third, the analysis used precomputed CONCH embeddings, and the exact checkpoint revision and archive-generation configuration could not be independently authenticated. The CONCH encoder was frozen, so the effects of task-specific fine-tuning were not evaluated.

Fourth, bags were deterministically capped at 512 patches during model fitting and internal validation, although complete bags were used for outer-fold prediction. This approach reduced computational requirements but may have excluded informative or heterogeneous regions during training. Fifth, only one random seed was used for each model-fold combination, and independent repeated-seed experiments were not available. The resulting analysis therefore does not quantify variability due to random initialization or training order. Sixth, the final comparison among aggregation families was based on pooled outer-fold predictions rather than a fully nested model-selection procedure. Although every reported OOF prediction was generated by a model that did not train on that slide, the numerical selection of max pooling should be regarded as an empirical comparison within this cohort.

Finally, the attention weights generated by gated ABMIL were not independently reviewed against pathologist annotations. They indicate which patches influenced the model but should not be treated as biological explanations or clinically validated regions of interest. Formal calibration, uncertainty-based referral rules, and prospective workflow evaluation were also outside the scope of this work. Moreover, feature-extraction time, per-slide inference latency, memory consumption, computational cost, and processing throughput were not benchmarked systematically. These evaluations, together with external validation and pathological review, are required before the method can be considered for clinical deployment.

6. Conclusion and Future Work

This work presented a leakage-controlled comparison of mean pooling, max pooling, gated ABMIL, and a two-stage max-pooling cascade for five-subtype ovarian cancer classification using frozen high-resolution CONCH patch representations. Max pooling ranked highest numerically, although no CONCH aggregation strategy demonstrated a statistically significant advantage after correction for multiple comparisons. In contrast, the high-resolution CONCH-based approach outperformed classification based on downsampled whole-slide thumbnails, indicating that the quality of the patch representation may be more influential than increased aggregation complexity on this cohort. The analysis is limited by the absence of independent external validation, unavailable patient and demographic metadata, unverified patient-level independence, deterministic patch subsampling during training, and the use of a single frozen feature encoder. Future work should include geographically and institutionally diverse cohorts, evaluate scanner and stain domain shifts, compare additional pathology foundation models and multiscale representations, incorporate calibrated uncertainty estimation, and obtain independent pathological validation of attention-highlighted tissue regions. Clinical-readiness studies should also quantify feature-extraction time, per-slide inference latency, GPU memory requirements, computational cost, and processing throughput under representative digital-pathology workflows before clinical integration is considered.

All the Declarations and Statements

Author Contributions Statement

Vijay H. Kalmani contributed to conceptualization, methodology, formal analysis, validation, interpretation of the results, manuscript preparation, review, and editing.

Nagaraj V. Dharwadkar contributed to methodology, software development, data processing, experimentation, validation, and visualization.

Amol C. Adamuthe contributed to conceptualization, experimental design, technical supervision, result interpretation, manuscript review, and editing.

Altaf Husain contributed to data curation, investigation, validation, literature review, and manuscript preparation. All authors read and approved the final manuscript.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Declaration
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability Statement
The UBC Ovarian Cancer Subtype Classification and Outlier Detection (UBC-OCEAN) dataset used in this study is publicly available through the Kaggle competition repository [33] at https://www.kaggle.com/competitions/UBC-OCEAN. Access to and use of the dataset are subject to the terms specified by the data provider.

Ethical Declarations
This study was a non-invasive image collection of hand gestures with verbal consent of the participants. There was no involvement of human subjects or animals.

Acknowledgements
NA

Declaration of Generative AI in Scholarly Writing
Language polishing and grammar were done with the help of Paperpal AI. There was no use of a generative AI to create content, generate ideas, analyse data, generate figures, or write scientifically.

Abbreviations
This manuscript uses the following abbreviations:

AI – Artificial Intelligence
BdSL – Bangla Sign Language
CNN – Convolutional Neural Network DL – Deep Learning
YOLO – You Only Look Once mAP – mean Average Precision IoU – Intersection over Union TP – True Positive
FP – False Positive TN – True Negative FN – False Negative
F1-score – Harmonic mean of precision and recall SE – Squeeze and Excitation
SHAP – SHapley Additive exPlanations FPN – Feature Pyramid Network
CSP – Cross Stage Partial
C2f – Cross Stage Feature Fusion DFL – Distribution Focal Loss AMP – Automatic Mixed Precision GPU – Graphics Processing Unit CPU – Central Processing Unit
VRAM – Video Random Access Memory AABB – Axis-Aligned Bounding Box JPG – Joint Photographic Experts Group EXIF – Exchangeable Image File Format FPS – Frames Per Second

Appendix A/B/C…, with appendix title
None.

References

[1] I. Tsibulak and C. Fotopoulou, “Tumor biology and impact on timing of surgery in advanced epithelial ovarian cancer,” *International Journal of Gynecological Cancer*, vol. 33, no. 10, pp. 1627–1632, Aug. 2023.
[2] Z. Liu, C. Jing, and F. Kong, “From clinical management to personalized medicine: Novel therapeutic approaches for ovarian clear cell cancer,” *Journal of Ovarian Research*, vol. 17, no. 1, Feb. 2024.
[3] J. Xu et al., “A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data,” *BMC Bioinformatics*, vol. 20, no. 1, Oct. 2019.
[4] D. Zhang, P. Chen, C.-H. Zheng, and J. Xia, “Identification of ovarian cancer subtype-specific network modules and candidate drivers through an integrative genomics approach,” *Oncotarget*, vol. 7, no. 4, pp. 4298–4309, Dec. 2015.
[5] D. P. Cook and B. C. Vanderhyden, “Ovarian cancer and the evolution of subtype classifications using transcriptional profiling,” *Biology of Reproduction*, vol. 101, no. 3, pp. 645–658, Jun. 2019.
[6] G. Wang et al., “Machine learning-based rapid diagnosis of human borderline ovarian cancer on second-harmonic generation images,” *Biomedical Optics Express*, vol. 12, no. 9, p. 5658, Sep. 2021.
[7] B. Ziyambe et al., “A deep learning framework for the prediction and diagnosis of ovarian cancer in pre and post-menopausal women,” *Diagnostics*, vol. 13, no. 10, p. 1703, May 2023.
[8] M. L. D´oria et al., “Epithelial ovarian carcinoma diagnosis by desorption electrospray ionization mass spectrometry imaging,” *Scientific Reports*, vol. 6, no. 1, Dec. 2016.
[9] A. Titoriya and S. Sachdeva, “Breast cancer histopathology image classification using AlexNet,” in *Proceedings of the 4th International Conference on Information Systems and Computer Networks (ISCON)*, 2019, pp. 708–712.
[10] J. Sun and A. Binder, “Comparison of deep learning architectures for H&E histopathology images,” in *Proceedings of the International Conference on Big Data Analytics and Applications (ICBDAA)*, 2017, pp. 43–48.

[11] M. M. Srikantamurthy, V. P. S. Rallabandi, D. B. Dudekula, S. Natarajan, and J. Park, "Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid CNN-LSTM-based transfer learning," *BMC Medical Imaging*, vol. 23, no. 1, Jan. 2023.
[12] S. Tabibu, P. K. Vinod, and C. V. Jawahar, "Pan-renal cell carcinoma classification and survival prediction from histopathology images using deep learning," *Scientific Reports*, vol. 9, no. 1, Jul. 2019.
[13] Y. Guo, S. Liu, Z. Li, and X. Shang, "Towards the classification of cancer subtypes by using cascade deep forest model in gene expression data," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017, pp. 1664–1669.
[14] "BCDForest: A boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data," *BMC Bioinformatics*, vol. 19, no. S5, Apr. 2018.
[15] S. P. Blagden, "Harnessing pandemonium: The clinical implications of tumor heterogeneity in ovarian cancer," *Frontiers in Oncology*, vol. 5, Jun. 2015.
[16] R. Chen, L. Yang, S. Goodison, and Y. Sun, "Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data," *Bioinformatics*, vol. 36, no. 5, pp. 1476–1483, 2020.
[17] J. Wang, Z. Zhang, and Y. Wang, "Utilizing feature selection techniques for AI-driven tumor subtype classification: Enhancing precision in cancer diagnostics," *Biomolecules*, vol. 15, no. 1, p. 81, Jan. 2025.
[18] Y. Shen, Y. Luo, D. Shen, and J. Ke, "RandStainNA: Learning stain-agnostic features from histology slides by bridging stain augmentation and normalization," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, ser. Lecture Notes in Computer Science. Springer, 2022, pp. 212–221.
[19] L. Aldakhil, H. Alhasson, and S. Alharbi, "Attention-based deep learning approach for breast cancer histopathological image multi-classification," *Diagnostics*, vol. 14, no. 13, p. 1402, Jul. 2024.
[20] A. Shrivastava et al., "Self-attentive adversarial stain normalization," in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, ser. Lecture Notes in Computer Science. Springer, 2021, pp. 120–140.
[21] S. K. Behera, A. Das, and P. K. Sethy, "Deep fine-KNN classification of ovarian cancer subtypes using EfficientNet-B0 extracted features: A comprehensive analysis," *Journal of Cancer Research and Clinical Oncology*, vol. 150, no. 7, p. 361, Jul. 2024.
[22] T. A. Fahim, F. B. Alam, and K. T. Ahmmed, "OVANet: Dual attention mechanism based new deep learning framework for diagnosis and classification of ovarian cancer subtypes from histopathological images," *IEEE Access*, vol. 12, pp. 131 942–131 953, 2024.
[23] Z. Tarek and E. Hassan, "Rehearsal-based continual learning for robust ovarian cancer subtype classification under catastrophic forgetting," *Journal of Engineering and Applied Science*, vol. 73, no. 1, Apr. 2026.
[24] H. Al-Asi et al., "Pathology foundation models: Evolution, current landscape, challenges and opportunities from a technical and clinical perspective," *Bioengineering*, vol. 13, no. 5, p. 577, May 2026.
[25] B. Zhang et al., "Foundation models in cancer pathology: Techniques, applications, and future directions," *Research*, vol. 9, 2026.
[26] H. Xu et al., "A whole-slide foundation model for digital pathology from real-world data," *Nature*, vol. 630, no. 8015, pp. 181–188, May 2024.
[27] J. I. Pisula and K. Bozek, "Efficient WSI classification with sequence reduction and transformers pretrained on text," *Scientific Reports*, vol. 15, no. 1, p. 5612, Feb. 2025.
[28] S. Orsulic, J. John, A. E. Walts, and A. Gertych, "Computational pathology in ovarian cancer," *Frontiers in Oncology*, vol. 12, p. 924945, Jul. 2022.
[29] J. Breen, K. Allen, K. Zucker, L. Godson, N. M. Orsi, and N. Ravikumar, "A comprehensive evaluation of histopathology foundation models for ovarian cancer subtype classification," *npj Precision Oncology*, vol. 9, no. 1, p. 33, Jan. 2025.
[30] J. Chun et al., "Deep learning-based diagnosis of epithelial ovarian cancer from whole-slide histopathology images," *Diagnostics*, vol. 16, no. 10, p. 1470, May 2026.
[31] H. J. Song, Y. S. Cho, and Y. S. Kim, "Comparative evaluation of feature extractors, aggregation strategies, and classification hierarchies for ovarian cancer subtype classification in whole slide images," *Diagnostics*, vol. 16, no. 10, p. 1570, May 2026.
[32] M. J. Sai and N. S. Punn, "KS-TMIL: A K-stage transformer approach with multiple instance learning model for ovarian cancer subtype classification," *Computers in Biology and Medicine*, p. 110726, 2025.
[33] A. Bashashati and H. Farahani, "UBC ovarian cancer subtype classification and outlier detection (UBC-OCEAN)," Kaggle Competition, 2023, online; accessed: 20 August 2026. [Online]. Available: https://www.kaggle.com/competitions/UBC-OCEAN
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
[35] S. Xie, R. Girshick, P. Doll´ar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 5987–5995.*

Authors' Profiles



Vijay H. Kalmani, is a Professor in Department of Computer Science and Engineering at Rajarambapu Institute of Technology, Rajaramnagar, Sangli, India. He obtained his M.Tech in Computer Science and Engineering from Gogte Institute of Technology, Belagavi, affiliated with Visvesvaraya Technological University, Belagavi and earned his Ph.D. in Computer Science and Engineering from Suresh Gyan Vihar University, Jaipur, India, in 2016. His areas of expertise included Cloud Security, Artificial Intelligence, and Machine Learning. He successfully guided three research scholars towards the completion of their Ph.D.



Nagaraj V. Dharwadkar obtained his BE in Computer Science and Engineering in 2000 from the Karnataka University Dharwad, received his M.Tech in Computer Science and Engineering in 2006 from VTU, Belgaum and PhD in Computer Science and Engineering in 2014 from the National Institute of Technology, Warangal. He is Associate Professor in the department of Computer Science and at the Central University of Karnataka, Kadaganchi, India. He has an experience of 24 years in teaching at professional institutions across India and published 100 papers in Various international journals and conferences. His areas of research interest are Multimedia security, image processing, data mining, and machine learning.



Amol C. Adamuthe is an IEEE Senior Member. Presently, he is the professor and head of the Department of Information Technology at Rajarambapu Institute of Technology, Sangli, Maharashtra, India. He obtained his Ph.D. from Mumbai University and specializes in technology forecasting, optimization algorithms, soft computing, and cloud computing. He has written more than 45 papers in referred journals and conferences. He received the Young Engineers Award 2018–19 in recognition of his contributions to engineering from the Institution of Engineers (India). Further, he contributes to research and professional development as Associate Editor of the Journal of Electrical Engineering & Technology (JEET) and Convener of IEEE ICONAT.



Altaf Husain is an industry technology leader with over 25 years of experience in wireless networking, network security, and intelligent systems. His research interests include computer vision, medical imaging, and the application of deep learning to real-world engineering challenges. He is a member of the IEEE with publications at international venues spanning computer vision, signal processing, and network security.

How to cite this paper

Vijay H. Kalmani, Nagaraj V. Dharwadkar, Amol C. Adamuthe, Altaf Husain, "Attention-Guided Deep Learning Framework for Ovarian Cancer Subtype Classification", International Journal of Engineering and Manufacturing (IJEM), Vol.16, No.5, pp. 158-178, 2026. DOI:10.5815/ijem.2026.05.09