HistoEffiCrossFormer: Cross-Attention CNN–Transformer Fusion with Multi-Scale Tokens for Ovarian Cancer Histopathology Classification

Roseline Oluwaseun Ogundokun ogundokunroseline1@gmail.com
Rotimi-Williams Bello sirbrw@yahoo.com
Pius Adewale Owolawi owolawpa@tut.ac.za
Chunling Tu duc@tut.ac.za
Sunday Agbolade sjagbolade@gmail.com
Tosho Abdulahi AbdulRahman abdulrahman.t@kwarastatepolytechnic.edu.ng
Department of Computer Systems Engineering
Tshwane University of Technology (TUT)
Pretoria, South Africa
Department of Computer Science
Redeemer University, Ede
Osun State, Nigeria
Department of Multimedia Engineering
Kaunas University of Technology
Kaunas, Lithuania

Editors: Sakinat Folorunso, Roseline Ogundokun, and Francisca Oladipo

Abstract

Ovarian cancer remains a leading cause of gynecologic cancer mortality, and histopathology is central to definitive diagnosis. We propose HistoEffiCrossFormer, a lightweight CNN–Transformer hybrid for five-class classification of ovarian cancer histopathology images. The pipeline uses Macenko-inspired stain normalisation, EfficientNet-B0 feature extraction, squeeze-and-excitation channel attention, multi-scale tokenisation (1×1 and 3×3 convolutions), a 2-layer transformer encoder, and cross-attention fusion between a learnable [CLS] query and contextualised tokens. We benchmark against SqueezeNet, ShuffleNetV2, AlexNet, and Xception under identical preprocessing and AdamW training. HistoEffiCrossFormer reaches 0.8267 test accuracy with AUC 0.974, outperforming the lightweight CNN baselines (0.7067–0.7867 accuracy) and closely matching Xception's AUC (0.975) while remaining competitive in accuracy (0.840). These findings indicate that cross-attention token fusion can enhance discrimination in the presence of stain variability and motivate further validation in external cohorts and whole-slide pipelines. We also provide an end-to-end experimental workflow that links preprocessing, training, and ROC-based evaluation, enabling reproducible, fair comparisons.

Keywords: computational pathology, ovarian cancer, histopathology, EfficientNet, transformer, cross-attention, ROC-AUC

1. Introduction

The deadliest of all gynaecological cancers is ovarian cancer, which has a significant contribution from the broad range of patients with an advanced disease at the time of diagnosis and from the molecular heterogeneity of the cancer cells (Bucur et al., 2024). Histopathology using H and E (hematoxylin and eosin)-stained slides remains the most definitive way to determine whether a lesion is malignant and to characterise its morphological features. As digital pathology continues to transform physical glass slides into WSIs, it will increasingly utilise computational decision support in place of manual microscopy processes (Ogundokun et al., 2025; Hoque et al., 2024).

Machine learning in histopathology has many challenges. The H and E-stained slides used for pathology are not only from different laboratories or scanned by different scanners, but they also have variability in the way they are prepared based on the staining technique; therefore, multiple sources of variability have been found to contribute to the domain shifting in the slides and to have an adverse effect on generalisation (Ogundokun et al., 2025; Folorunso et al., 2024). Additionally, the number of diagnostically significant parameters spans a continuum from the very small scales of individual cells (e.g., the nucleus, chromatin, and the matrix in which they reside) to the very large scale of the overall tissue (e.g., stroma or parenchyma) and back again. CNNs are good at detecting small morphological features but lack long-range dependencies, so they are not as well-suited to capturing contextual relationships between small and large features. Self-attention models, such as transformers, are effective at capturing global contextual relationships, but they are more difficult to implement in practice due to the physical characteristics of medical images (Tan and Le, 2019).

This paper presents HistoEffiCrossFormer, a lightweight hybrid architecture to classify 5 categories of ovarian cancer histopathology. HistoEffiCrossFormer fuses EfficientNet-B0 with squeeze and excitation (SE) channel attention, converts deep features into multi-scale tokens using parallel 1×1 and 3×3 convolutions, processes tokens through a compact 2-layer transformer encoder, then performs cross-attention fusion to allow a learnable CLS query to attend across contextualised tokens. A Macenko-inspired method is used to perform stain normalisation and standard augmentation to help ensure insensitivity to variation in colour and intensity.

The study examines 3 research questions: (RQ1) Does HistoEffiCrossFormer improve test accuracy vs compact CNN baselines and Xception? (RQ2) Does HistoEffiCrossFormer improve discrimination measured by ROC-AUC? (RQ3) What practical lessons can we learn to create efficient, stable computational pathology pipelines? These questions correspond to common translation issues noted in the literature on AI in pathology, particularly regarding how well the system can generalise and be evaluated rigorously.

2. Related Works

A literature review of Computational Pathology by (Hosseini et al., 2024) has shown that algorithmic performance in papers has increased rapidly, but reproducible and deployable clinical systems remain difficult to achieve. At present, there are regular umbrella surveys that cover the entire pipeline from slide digitisation to patching, through model development and assessment, and they consistently point out bottlenecks (such as data governance

and access, the total number of WSIs, the cost of labelling images and limited external validation). As a consequence of these broad reviews, there is a trade-off between the breadth of what they cover to include multiple different tasks and families of models, but there is frequently no resolution on which architectural choices will work best when compute, data, and annotation resources are limited, or an organ-specific target (like ovarian cancer histopathology) (Hosseini et al., 2024). Therefore, both colour variability and stain normalisation issues are considered first-priority problems in Digital Pathology. It has been extensively documented, through a comprehensive review of stain normalisation techniques (and experimental comparisons), that both stain and acquisition variability reduce not only the accuracy of computer-aided diagnosis but also that of human diagnosis. Thus, it has been proposed that normalisation of stains be adopted as a standardisation approach. Another more recent review that focuses specifically on deep learning stain normalisation has classified stain normalisation methods as either GAN-based or non-GAN-based. This classification is indicative of the rapid changes in this field. A major limitation of studies comparing different methods of stain normalisation is the inherent difficulty of generalisation, as study populations often comprise different datasets, reference targets, and evaluation strategies. Furthermore, methods that rely on selecting an appropriate reference for stain normalisation will produce artefacts if there is a large difference between the stain in question and its reference source.

The use of transformers and attention-based modelling is also a rapidly developing field. A well-publicised survey of transformer networks for medical imaging, containing over 125 publications, suggests that transformers offer an alternative to traditional convolutional neural networks for capturing global context (which CNNs can find difficult to model) due to their self-attention mechanisms. Despite their advantages, they also come with several practical limitations, such as typically requiring larger data sets for training and large images requiring more variable computation time when using naïve attention-based transformers.

Volumes of surveys specifically looking at transformers for histopathology suggest that both patch-based and WSI (whole slide image) methodologies are increasingly accepted practices in industry, but continue to predominantly focus on the utilisation of larger model architectures, complex training protocols, and infrastructure dependencies that may not necessarily apply to most clinical laboratories around the world. The limitation seen most often in the surveys above is that "context" in pathology often requires connecting patterns across many patches within a slide, rather than just within a single fixed-size crop. This pushes models to develop smarter aggregation and training methods (Shamshad et al., 2023).

Multiple instance learning (MIL) has become a standard method to accommodate the need to work with whole-slide packs and weak slide-level labels. A recent study on multiple-instance learning (MIL) summarises several popular deep MIL concepts and indicates that many challenges remain to be overcome. Some of the needed improvements include robustness to domain shift, the quality of explanations (i.e., which regions contributed most to the slide label), and an evaluation design suited to a clinical context rather than to the easier-to-use patch-level benchmarks. The primary drawback of using image-level experiments like the present one is that most MIL reviews only go into detail on slide-level learning. They serve to motivate future work and to highlight the importance of careful patch- or image-level modelling when using only tile-level labels (Gadermayr and Tschuchnig, 2024).

Reviews of ovarian cancer indicate a gap in the ability of treatment to be translated into practice. A systematic review of the use of artificial intelligence (AI) in ovarian cancer histopathology employed a specific methodology to assess the risk of bias in the literature. Specific issues, such as incomplete reporting and bias, made it difficult to determine whether strong metrics reported in studies would hold true in other settings. Additionally, reviews employing broader searches of AI-created literature on ovarian cancer, including those related to the laboratory, imaging, and treatment of patients with ovarian cancer, all noted that differences in the datasets used and the methods applied posed major hurdles to the effective implementation of the therapies. The systematic review of the literature on AI methods for identifying ovarian cancer (published between 2020 and the current date of 2025) illustrates similar methodological sophistication but also demonstrates significant heterogeneity and a need for improved translational capabilities.

The consistent findings of these reviews underscore a clear need: the ovarian cancer field requires reproducible, compact architectures that properly address staining variability, capture multi-scale morphology (and long-distance context), and have benchmarks for evaluation against both lower and higher standards. To address this need, we propose the HistoEffiCrossFormer, an architecture that provides stain normalisation, multi-scale tokenisation, and cross-attention fusion, and evaluate its effectiveness using matched training compared with four baseline methods.

3. Materials and Methods

The experiment used an Image dataset of Ovarian Carcinoma, where the images contain histological data on the structural and functional characteristics of cancerous ovarian tissue samples. The results of classifier generation have been recorded in a 5-class dataset, based on existing clinical literature on how these cancer types are classified according to Pathology Review Classification Systems. Once created, a training/validation/test split was used to train the classifier. Experimental processing from start to finish, from the dataset creation to the eventual ROC Compliant Comparison of the 5 Class Classifier, is depicted in Figure 1.

As part of the experience, all images were/are subjected to some degree of pre-processing, including Mean Stain Normalisation and other procedures that were derived from, or that are like, the Macenko Method. Each image has been normalised to have equal average intensity and standard deviation; all images were then scaled/mapped to some common maximum and minimum intensity; all images have also been resized to 384x384 pixels. During the training phase, random horizontal and vertical flipping is employed in conjunction with very minimal colour alteration; during the evaluation phase, all images will undergo deterministic resizing and no colour augmentation. Finally, as with other techniques described above, all images were normalised to the same mean and standard deviation based on the ImageNet dataset, resulting in all images having the same (and normal) mean and standard deviation. (Note: The original method, as described by Macenko, was based on the optical density and vector method of stain normals. However, because this coding method for using mean normalisation and then standardising each per image is more appropriately classified as being a "Macenko-inspired" version of the method rather than an exact replication).

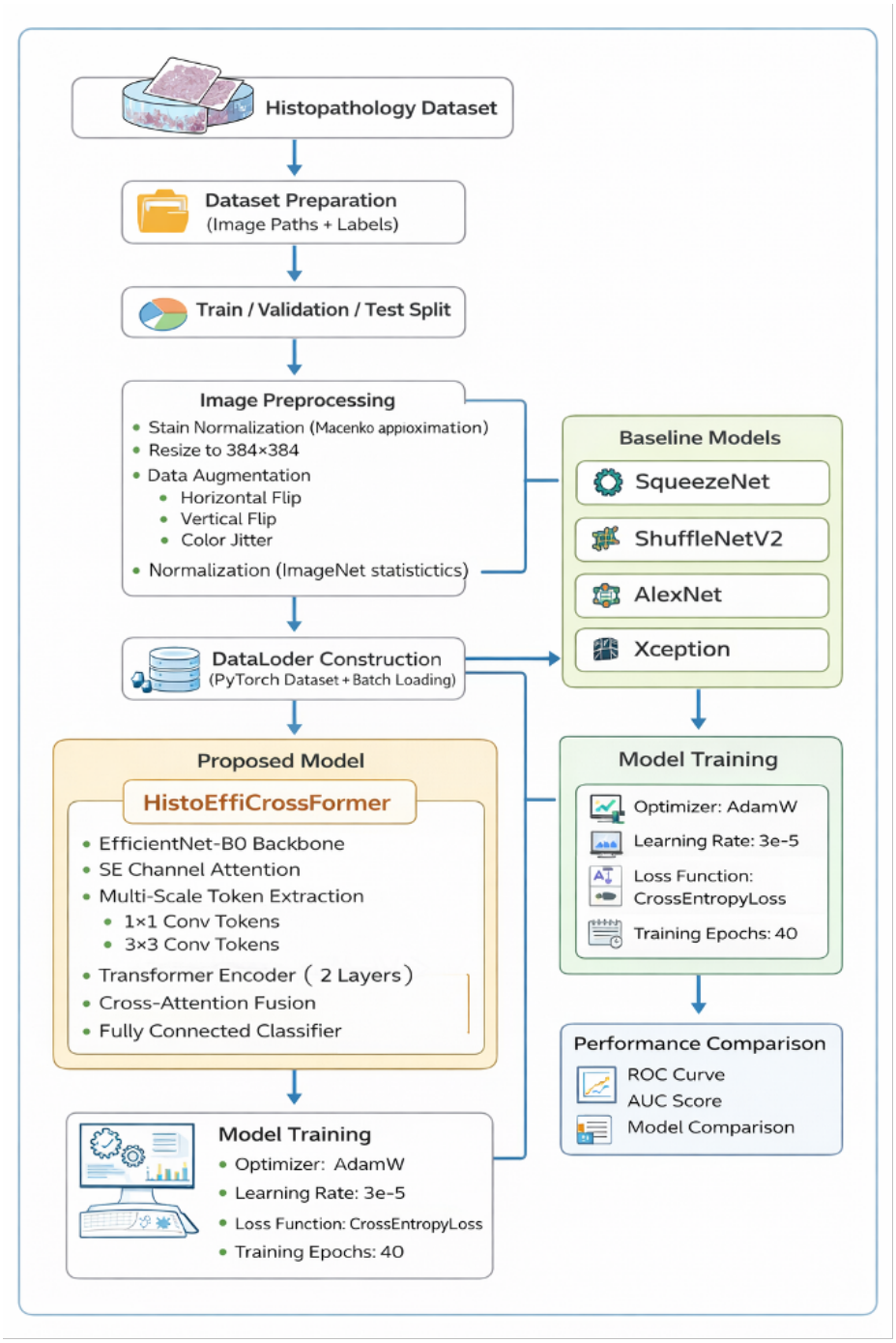

Figure 1: Proposed Study Experimental Flow Diagram

Model Architecture: Figure 2 presents the proposed HistoEffiCrossFormer architecture. An EfficientNet-B0 backbone produces a deep feature map with channel dimension C=1280. An SE block reweights channels to emphasise informative morphology. Multi-scale tokeni-

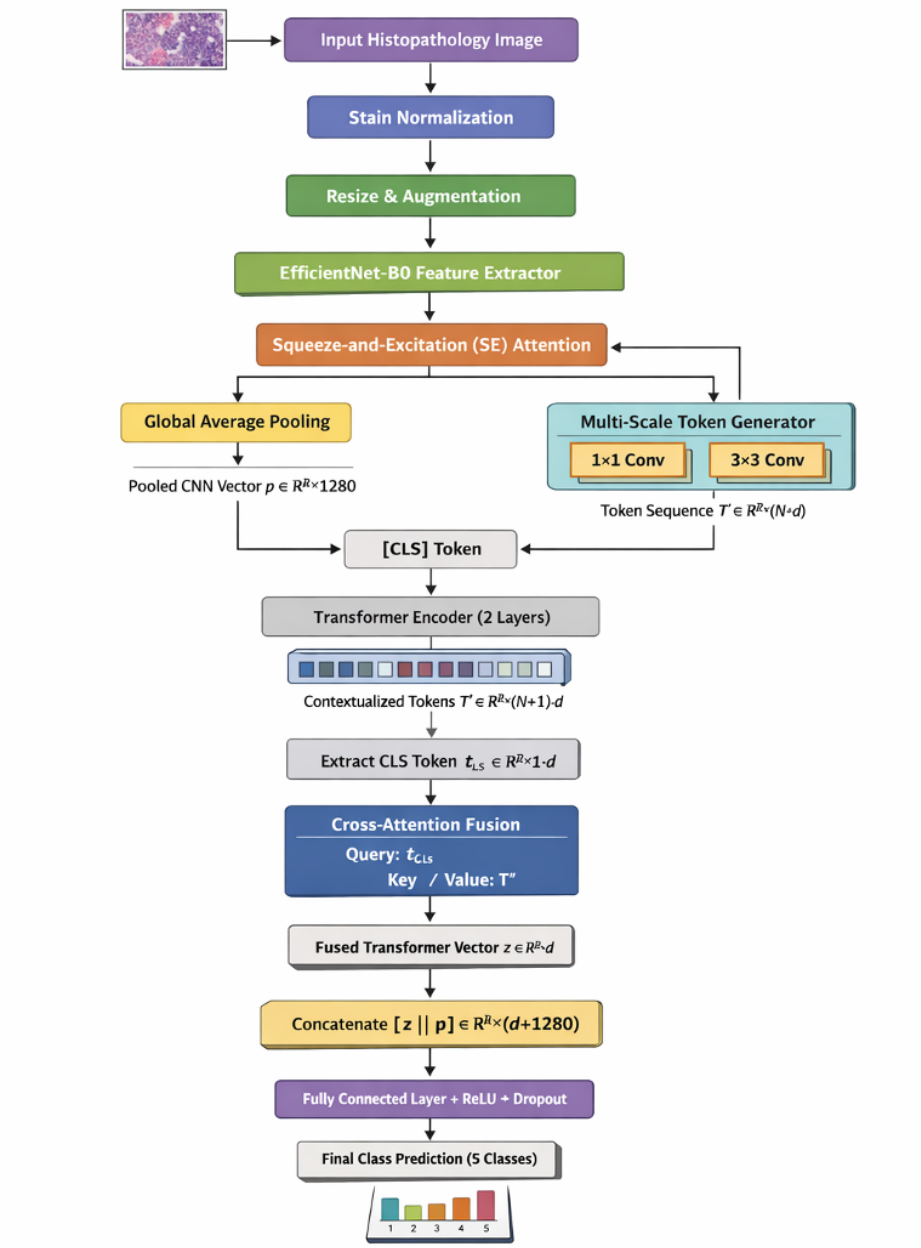

Figure 2: Proposed HistoEffiCrossFormer Architecture

sation projects features to the token dimension d=192 using parallel 1×1 and 3×3 convolutions; tokens are concatenated, a learnable [CLS] token is prepended, and a compact two-layer transformer encoder contextualises the sequence. Cross-attention fusion uses the [CLS] token as query over all contextualised tokens (keys/values) to produce a fused vector z. In parallel, global average pooling yields a pooled CNN vector p. The final classifier consumes the concatenation $[z \parallel p]$.

Baselines: Four baselines are used for comparison: SqueezeNet, ShuffleNetV2, AlexNet, and Xception, each adapted to a five-class output and initialised with ImageNet-pretrained weights (as indicated by the "DEFAULT" pretrained weights setting in the code). These baselines span a range from compact models (SqueezeNet, ShuffleNetV2) to historically influential CNNs (AlexNet) and a strong depthwise-separable architecture (Xception). Training hyperparameters: Table 1 summarises the training configuration explicit in the provided code (values not shown in the snippet, such as batch size, are not asserted here). AdamW is used because decoupled weight decay is widely discussed as improving generalisation behaviour for adaptive optimisers relative to naïve L2 regularisation in Adam-style updates.

Table 1: Model Training and Preprocessing Hyperparameters

Component	**Setting**
Dataset	Ovarian cancer histopathology images (5 classes)
Input size	384×384
Stain normalization	Macenko-inspired approximation (per-image mean/std normalisation + min–max rescaling)
Training augmentation	Random horizontal flip; random vertical flip; colour jitter $(0.3, 0.3, 0.3, 0.05)$
Evaluation augmentation	None (deterministic resize only)
Tensor normalization	ImageNet mean/std
Optimizer	AdamW
Learning rate	3×10^{-5}
Loss	CrossEntropyLoss
Epochs	40
Classification dropout	0.25
Token dimension (d)	192
Transformer encoder	2 layers, 4 heads
Cross-attention fusion	Multi-head attention, 4 heads

Evaluation metrics: Performance is reported as test accuracy and ROC-AUC. ROC analysis is a standard way to visualise threshold tradeoffs, and AUC provides a threshold-agnostic summary. For multi-class ROC, the pipeline uses one-vs-rest label binarisation and produces a micro-averaged ROC curve by flattening class-wise scores.

4. Results and Discussion

Figure 3 presents ROC curves for the proposed model and four baselines, while Table 2 reports test accuracy. Both are computed from softmax probabilities using one-vs-rest label binarisation and a micro-averaged ROC constructed by flattening class-wise scores.

Table 2: Test Accuracy Comparison on the Ovarian Cancer Histopathology Dataset

Model	Test Accuracy
HistoEffiCrossFormer	0.826667
SqueezeNet	0.760000
ShuffleNetV2	0.706667
AlexNet	0.786667
Xception	0.840000

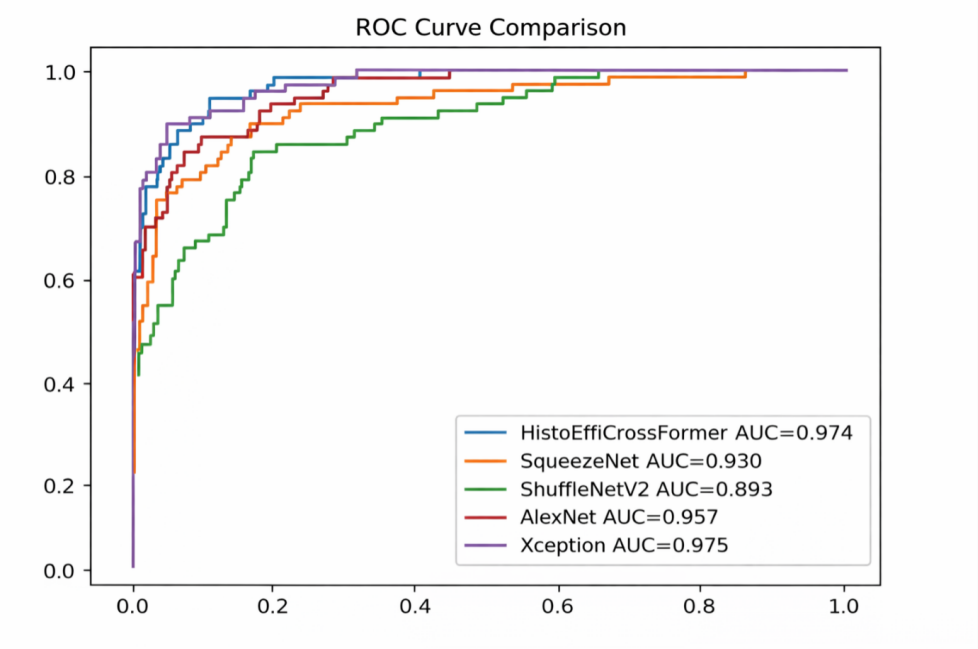

Figure 3: ROC Curve Comparison (Micro-Averaged) with AUC Values

RQ1: Accuracy comparison. Table 1 shows that HistoEffiCrossFormer achieves a test accuracy of 0.8267, outperforming SqueezeNet (0.7600), ShuffleNetV2 (0.7067), and AlexNet (0.7867). Xception is the top baseline in terms of accuracy, at 0.8400, only 0.0133 points above the proposed model. Since the results come from a single split and no uncertainty estimates are provided, the appropriate conclusion is that the proposed model is strongly competitive in terms of accuracy and clearly better than the lightweight CNN baselines in this experimental setting.

RQ2: Discrimination via ROC-AUC. In Figure 3, HistoEffiCrossFormer reaches AUC=0.974, exceeding SqueezeNet (0.930), ShuffleNetV2 (0.893), and AlexNet (0.957), and closely matching Xception (0.975). There is a near tie between HistoEffiCrossFormer and Xception, suggesting they have nearly identical ability to separate predicted scores globally across nearly all thresholds. This is important because the area under the receiver operating characteristic curve (AUC) summarises discriminative power at all possible operating points, while accuracy is only evaluated based on a single argmax decision rule (the single pre-

dicted classes). Additionally, small changes in calibration (or margin) for the top two-class predictions can alter accuracy without significantly affecting AUC.

RQ3: Interpretation and Limitations. The observed performance improvements compared to compact CNNs are in alignment with the design intent of utilising (1) local morphology encoders (based on EfficientNet), (2) explicit channel re-weighting (using SE), and (3) token-based context modelling (using both a transformer architecture and cross-attention fusion) as a hybrid approach to pathology imaging characterised by high variance. Hybrid approaches like this are also consistent with survey conclusions indicating that attention can augment convolutional locality by embedding higher-level information into the architecture while leveraging the CNN's inductive-bias properties for morphological and texture representation.

Several limitations of generalising these findings exist. Firstly, micro-averaged ROC may mask poor performance on certain classes, such as minority or visually subtle classes; therefore, to derive clinically meaningful conclusions from this analysis, a confusion matrix and per-class AUCs must be computed. Furthermore, stain handling remains a primary risk factor for external validity in digital pathology. The current processing pipeline employs a known, approximate stain standardisation (i.e., the stains used during the production of the analysed image), and the consistency of stain normalisation processes has been shown to potentially impact the outcome. Finally, the experiment was performed at the image level, whereas implementation in real clinical practice requires slide-level inferences, and there is very limited external validation across services, laboratories, and scanners, as detailed in several systematic reviews above.

5. Clinical and Practical Implications

From the perspective of clinical practice, we consider methodologies such as HistoEffiCross-Former best suited to assist in decision-making rather than to serve as a primary mode of patient diagnosis. Classifiers that have good ROC-AUC characteristics may help prioritise cases based on their likelihood of accurate diagnosis (for review only), assist with prioritisation of cases for triage purposes (where appropriate), and provide an alternative measure of second-read probability when used to interpret results in conjunction with an actual pathologist's evaluation and the clinical environment they are working in. The published literature regarding the application of AI tools to pathology emphasises that the usefulness and success of those tools depend as much upon how well they've been validated by reputable organisations, integrated into workflows, and how likely they are to fail as it does upon their measure of effectiveness (i.e., their ROC-AUC).

Digital pathology platforms for the supply and retrieval of digital pathology data are maturing to the point that validation studies have shown whole-slide images to be non-inferior to conventional light microscopy for making a primary diagnosis. Consequently, the question regarding the deployment of digital pathology will now have to shift from "will we be able to use digital slides?" to "can an artificial intelligence support tool demonstrate its ability to generalise, can it be audited for compliance, and can it ensure safety when utilised in routine clinical practice?"

The experimental flow in Figure 1 provides a framework for implementing validated protocols by making sure everything is the same between experiments (known as "pre-

analytics"), using a consistent way of training (whether via multiple splits of your training data or cross validation), testing against a like-data source (e.g., using data collected in another lab to test generalisability, using data collected over time; e.g., testing the system using results from previous trials), and reporting total accuracy/overall errors and individual class accuracy/overall errors (per class; rate of incorrect test). In coordination with these goals, approved guidance/standards should be utilised, such as TRIPOD/AI (for predictive) and CONSORT/AI/SPIRIT (for clinical trials).

6. Conclusion and Future Works

This study proposed HistoEffiCrossFormer, a compact CNN–Transformer hybrid for ovarian cancer histopathology classification that integrates SE attention, multi-scale tokenisation, and cross-attention fusion atop an EfficientNet-B0 backbone. In the provided benchmark, the model improved over lightweight CNN baselines in both accuracy and AUC and achieved micro-averaged AUC (0.974) essentially indistinguishable from Xception (0.975), while remaining close in accuracy. Future work should prioritize four directions that are emphasized repeatedly in the review literature: (i) per-class reporting to expose failure modes masked by micro-averages; (ii) robustness studies across laboratories, scanners, and staining protocols (including more faithful stain normalization and domain adaptation); (iii) whole-slide inference via patch aggregation or MIL aligned with real workflows; and (iv) interpretability and calibration so outputs support accountable clinical use. Given the rapid emergence of pathology foundation and multimodal models, a further practical step is to explore pathology-specific pretraining rather than relying only on ImageNet transfer learning.

References

[1] C. Bucur, I. Balescu, S. Petrea, B. Gaspar, L. Pop, V. Varlas, and N. Bacalbasa. Artificial intelligence in ovarian cancers—from diagnosis to treatment: A literature review. *Journal of Mind and Medical Sciences*, 11(2):36, 2024. doi: 10.22543/2392-7674.1514.

[2] S. O. Folorunso, J. B. Awotunde, Y. P. Rangaiah, and R. O. Ogundokun. Efficientnets transfer learning strategies for histopathological breast cancer image analysis. *International Journal of Modeling, Simulation, and Scientific Computing*, 15(02):2441009, 2024. doi: 10.1142/S1793962324410095.

[3] M. Gadermayr and M. Tschuchnig. Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential. *Computerized Medical Imaging and Graphics*, 112:102337, 2024. doi: 10.1016/j.compmedimag.2024.102337.

[4] M. Z. Hoque, A. Keskinarkaus, P. Nyberg, and T. Seppänen. Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison. *Information Fusion*, 102:101997, 2024. doi: 10.1016/j.inffus.2023.101997.

[5] M. S. Hosseini, B. E. Bejnordi, V. Q. H. Trinh, L. Chan, D. Hasan, X. Li, and K. N. Plataniotis. Computational pathology: A survey review and the way forward. *Journal of Pathology Informatics*, 15:100357, 2024. doi: 10.1016/j.jpi.2024.100357.

[6] R. Ogundokun, P. Owolawi, and C. Tu. Optimized deep feature learning with hybrid ensemble soft voting for early breast cancer histopathological image classification. *Computers, Materials & Continua*, 84(3):4869, 2025. doi: 10.32604/cmc.2025.066418.

[7] R. O. Ogundokun, P. A. Owolawi, M. O. Adebiyi, and O. Ishola. Deep feature extraction with convolutional autoencoder and ensemble learning for multiclass breast cancer histopathology image classification. *NIPES Journal of Science and Technology Research Special Issue*, 7(1):591–595, 2025.

[8] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu. Transformers in medical imaging: A survey. *Medical Image Analysis*, 88:102802, 2023. doi: 10.1016/j.media.2023.102802.

[9] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.