Equipping Computational Pathology Systems with Artifact Processing Pipelines: A Showcase for Computation and Performance Trade-offs

Neel Kanwal*[1†], Farbod Khoraminia[2†], Umay Kiraz[3,4†], Andrés Mosquera-Zamudio[5†], Carlos Monteagudo[5], Emiel A.M. Janssen[3,4], Tahlita C.M. Zuiverloon[2], Chunmig Rong[1], Kjersti Engan[1]

[1]Department of Electrical Engineering and Computer Science, University of Stavanger, 4021 Stavanger, Norway.
[2]Department of Urology, University Medical Center Rotterdam, Erasmus MC Cancer Institute, 1035 GD Rotterdam, The Netherlands.
[3]Department of Pathology, Stavanger University Hospital, 4011 Stavanger, Norway.
[4]Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, 4021 Stavanger, Norway.
[5]Department of Pathology, INCLIVA Biomedical Research Institute, and University of Valencia, 46010 Spain
* Corresponding author(s): {neel.kanwal, kjersti.engan}@uis.no.

[†]These authors contributed equally to this work.

Abstract

Background: Histopathology is a gold standard for cancer diagnosis. It involves extracting tissue specimens from suspicious areas to prepare a glass slide for a microscopic examination. However, histological tissue processing procedures result in the introduction of artifacts, which are ultimately transferred to the digitized version of glass slides, known as whole slide images (WSIs). Artifacts are diagnostically irrelevant areas and may result in wrong predictions from deep learning (DL) algorithms. Therefore, detecting and excluding artifacts in the computational pathology (CPATH) system is essential for reliable automated diagnosis.

Methods: In this paper, we propose a mixture of experts (MoE) scheme for detecting five notable artifacts, including damaged tissue, blur, folded tissue, air bubbles, and histologically irrelevant blood from WSIs. First, we train independent binary DL models as experts to capture particular artifact morphology. Then, we ensemble their predictions using a fusion mechanism. We apply probabilistic thresholding over the final probability distribution to improve the sensitivity of the MoE. We developed four DL pipelines to evaluate computational and performance trade-offs. These include two MoEs and two multiclass models of state-of-the-art deep convolutional neural networks (DCNNs) and vision transformers (ViTs). These DL pipelines are quantitatively and qualitatively evaluated on external and out-of-distribution (OoD) data to assess generalizability and robustness for artifact detection application.
Results: We extensively evaluated the proposed MoE and multiclass models. DCNNs-based MoE and ViTs-based MoE schemes outperformed simpler multiclass models and were tested on datasets from different hospitals and cancer types, where MoE using (MobiletNet) DCNNs yielded the best results. The proposed MoE yields 86.15 % F1 and 97.93% sensitivity scores on unseen data, retaining less computational cost for inference than MoE using ViTs. This best performance of MoEs comes with relatively higher computational trade-offs than multiclass models. Furthermore, we apply post-processing to create an artifact segmentation mask, a potential artifact-free RoI map, a quality report, and an artifact-refined WSI for further computational analysis. During the qualitative evaluation, pathologists assessed the predictive performance of MoEs over OoD WSIs. They rated artifact detection and artifact-free area preservation, where the highest agreement translated to the Cohen kappa of 0.82, indicating substantial agreement for the overall diagnostic usability of the DCNN-based MoE scheme.
Conclusions: The proposed artifact detection pipeline will not only ensure reliable CPATH predictions but may also provide quality control. In this work, the best-performing pipeline for artifact detection is MoE with DCNNs. Our detailed experiments show that there is always a trade-off between performance and computational complexity, and no straightforward DL solution equally suits all types of data and applications. The code and dataset for training and development can be found online at Github and Zenodo, respectively.

Keywords: Computational Pathology, Deep Learning, Histological Artifacts, Mixture of Experts, Vision Transformer, Whole Slide Images

1 Introduction

Cancer develops in organs when genetic mutations in normal cells trigger their transformation into tumor cells. This transformation may be triggered by frequent exposure to carcinogens, a class of substances (chemical, biological, or physical), or several other factors that have the potential to cause cancer [1]. Diagnosing cancer accurately and efficiently is critical for medical treatment and a reduced mortality rate, given its status as one of the deadliest diseases worldwide, with a projected estimate of 29 million deaths by 2040 [2, 3]. Histopathology is considered a gold standard for identifying the presence of cancerous cells, which involves examining tissue samples under

a microscope using a histological glass slide [4]. However, this manual inspection and laboratory procedure is not without its pitfalls, as it is labor-intensive, subjective, and can be affected by inter- and intra-observer variability [5, 6]. Furthermore, the projected rise in cancer cases and the shortage of pathologists are significant issues that may lead to delayed diagnosis and treatment, resulting in a severe impact on clinical decision-making [7]. Therefore, streamlining the traditional diagnostic process through digitization and automation can provide timely diagnosis, improved treatment decisions, and efficacy [3]. Digital pathology (DP) has the potential to overcome these challenges by providing rapid diagnosis and smooth sharing of secondary opinions [8]. In fact, in the last decade, there has been a five-fold growth in DP research and development [9, 10]. This increase in the adoption of DP in clinical practice enables computation over the digitized version of histological slides, commonly called whole slide images (WSIs).

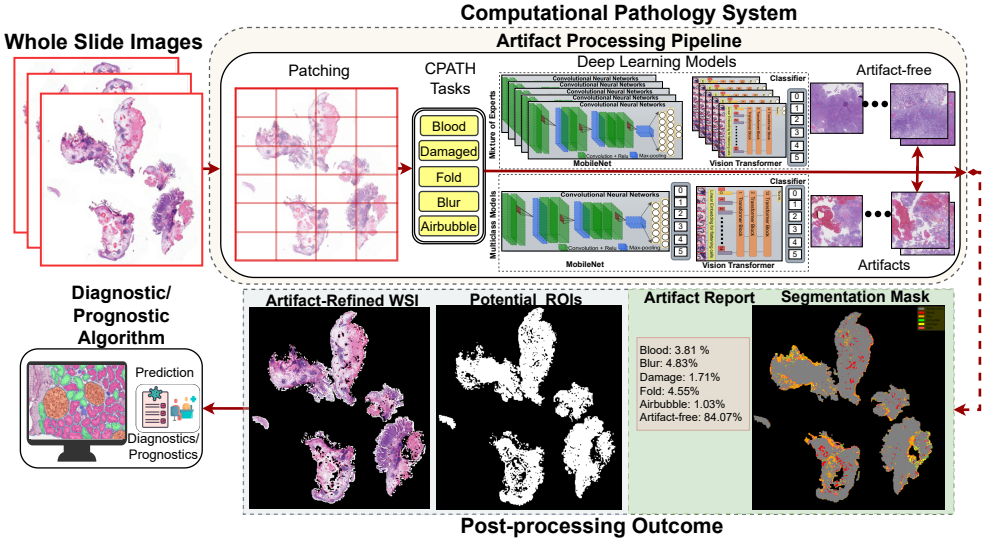

Fig. 1: **An overview of computational pathology (CPATH) system equipped with artifact processing pipeline.** Whole slide images (WSIs) are split into small sub-images (patches) to make them computationally tractable for deep learning (DL) models. These patches are fed to a mixture of experts (MoE) or multiclass models composed of state-of-the-art DL architectures to perform different CPATH classification tasks. Only patches with histological relevance can flow further for the downstream tasks. Finally, predictions are post-processed to produce different outcomes, such as a segmentation map, artifact report for quality control, region of interest mask, and artifact-free WSI for the diagnostic or prognostic algorithm to make a final clinical prediction.

Computational pathology (CPATH) systems have the potential to unfold information embedded in WSIs by automated systems based on AI and image processing

[10–12]. The seamless integration of CPATH with DP can enhance diagnostic or prognostic methodologies and save pathologists‘ time [6, 13]. However, artifacts that appear during the histological slide preparation are ultimately transferred to the WSIs [14–16]. Artifacts are diagnostically irrelevant areas, and pathologists usually ignore these areas during manual inspection, but unfortunately, the presence of histological artifacts can hamper the performance of CPATH systems during automated diagnosis [10, 17]. Therefore, it is essential to equip the CPATH system with an *artifact detection pipeline* to exclude artifacts and ensure the flow of histologically relevant tissue for diagnostic or prognostic algorithms, as illustrated in Figure 1. Thus, a CPATH system with artifact processing capacity will not only increase the likelihood of reliable and accurate predictions but also provide quality control (QC) for laboratory procedures, identifying weaknesses during the histotechnical stages (see review [10]) in acquiring WSIs.

In recent years, deep learning (DL) approaches have garnered more attention from the CPATH community due to their ability to extract hidden patterns in histological data [18–21]. Popular DL architectures such as deep convolution neural networks (DCNNs) and vision transformers (ViTs) have widely been used as state-of-the-art (SOTA) to distinguish tissue patterns for different cancer types and perform image classification and segmentation tasks [16, 19, 22]. Some researches [23, 24] demonstrate that DCNNs perform better on small datasets, thanks to the inductive bias, which helps them to learn spatial relevance effectively. While other works [25–27] argue in favor of ViTs, showing that they are highly robust, attend to overall structural information, and are less biased towards textures. Nevertheless, both DL architectures may suffer from overfitting, poor generalization, and reproducibility issues, leading to overconfident predictions on new (external) data. To address these problems, ensembles of DL models (a.k.a. deep ensembles) have been used to overcome the weakness of an individual model [28–30]. Ensemble methods combine the prediction of independent models using averaging or majority voting. A mixture of experts (MoE) is an extended method that trains DL for a sub-task and then combines the predictions dynamically to obtain a nuanced prediction. In short, the MoE approach consists of multiple DCNNs or ViTs, experts on each subclass, to achieve improved results. MoEs benefit in terms of reproducibility by reducing the variance of predictions but augmenting computational expense [31]. In contrast, the multiclass approach can be computationally efficient but does not involve the strength of multiple models, which are adaptive for looking into different aspects of data. Based on these arguments, the choice between DL approaches depends on application requirements. This raises a fundamental question: *how to build an effective artifact detection DL approach for CPATH systems with suitable trade-offs between computational complexity and performance?*

An effective DL approach for artifact detection applications (our case) might be created using MoEs, one DL model for each artifact class, or multiclass models with multiple output classes. In this paper, we propose the MoE-based DL approach, which uses a fusion mechanism to integrate predictions from experts and apply probabilistic thresholding to improve the sensitivity. We establish several DL pipelines using the MoE and multiclass models for detecting notable artifacts (i.e., damaged tissue, blur, folded tissue, air bubbles, and diagnostically irrelevant blood) from histological WSIs (see Figure 1). Our DL pipelines produce four outcomes for the input WSI: i) an

Artifact segmentation map; ii) an artifact report for QC using six classes (five artifacts and artifact-free area); iii) An artifact-free mask with potential regions of interest (ROIs) with diagnostic relevance; and iv) an artifact-refined WSI for the diagnostic algorithm.

Our contributions to this work are summarized below:

- We develop four DL models (referred to as DL pipeline throughout the paper), with SOTA DCNNs (MobileNet [32]) and ViTs (ViT-tiny [33]), using MoE and a multiclass approach.
- We evaluate the computational complexity of the pipelines and systematically choose a learned probability threshold for maximizing the sensitivity of DL models in external validation.
- We conduct a qualitative and quantitative evaluation of out-of-distribution data (from different cancer types) and assess the efficiency of the proposed MoE scheme for detecting artifacts and QC.

The paper is structured as follows: Section 2 presents recent studies involving DL approaches for computational pathology, along with related work for detecting artifacts. Section 3 provides data material descriptions. Section 4 explains pre-processing for creating datasets, the proposed method, post-processing, evaluation metrics, and implementation details. Section 5 discusses results for performance and computational complexity. Finally, section 6 concludes this work and discusses future directions for smooth integration of artifact processing pipelines in CPATH systems.

2 Related Work

2.1 Deep Learning for Computational Pathology

Deep learning (DL) approaches have gained popularity in the CPATH community [21, 34–36]. In recent years, several works [20, 37–40] have used popular DL architectures for diagnosis and prognostic algorithms. FDA-approved PAIGE [41] is an example of such a DL-based algorithm for prostate cancer. These works can be roughly divided into two branches such as DCNN-based (MobileNet [32], DenseNet [42], ResNet [43], or GoogleNet [44], etc.) or ViT-based (ViT-Tiny([33], DINO [45], or SwinTransformer [46] etc.) approaches.

In the first branch, Srinidhi *et al.* [47] comprehensively reviewed different DL approaches for developing disease-specific classification algorithms using histological images. Riasatian *et al.* [48] applied transfer learning over DCNNs to classify various tumor types and accomplished remarkable results using three public histopathology datasets. Talo [49] demonstrated that pre-trained ResNet [43] and DenseNet [42] to achieve better accuracy than traditional methods in the literature for classifying grayscale and color histopathological images. Similarly, Wang *et al.* [50] proposed a DCNN-based method based on GoogleNet [44] to locate tumors in breast and colon images using complex example-guided training for WSI analysis. Among other DCNN works, Meng *et al.* [29] compared several architectures for classification and segmentation problems on a cervical histopathology dataset. Their approach found the best results for precancerous lesions using ResNet-101 [43]. For the same task,

MobileNet [32] was the fastest. Wang *et al.* [51] performed multi-class breast cancer classification in their two-stage dependency-based framework. A MobileNet [32] was used as a backbone to extract the features in the first stage. Then, the MobileNet [32] backbone was modified to perform sub-type classification for benign and malignant categories. Gandomkar *et al.* [38] deployed ResNet [43] for classifying breast histology images into benign or malign and then identified them among several sub-types using a meta-classifier based on a decision tree.

Works in the second branch used ViTs, which have emerged as new SOTA, leveraging attention mechanisms to improve shape understanding and generalizablity [26] [27]. Stegmüller *et al.* [40] developed ViT-based ScoreNet for breast cancer classification. Their approach attended to some regions in the WSI for faster processing based on image semantics. Wessel*et al.* [39] used DINO [45] for predicting overall and disease-specific survival in renal cell carcinoma. Zidan *et al.* [46] introduced a ViT-based cascaded architecture for segmenting glands, nuclei, and stroma in colorectal cancer. Gao *et al.* [52] proposed instance-based ViT to capture global and local features for subtyping papillary renal carcinoma, achieving better performance over selected ROIs.

Unsurprisingly, in both branches, most of these DL algorithms were trained and tested on manually annotated clean data (with diagnostic relevance) and overlooked the impact of potential noise (histological artifacts) during the inference stage or unseen scenarios.

Schomig *et al.* [53], in their stress-testing study, showed that the accuracy of the prostate cancer DL algorithm was negatively affected by the presence of artifacts and resulted in more false positives. Even the presence of artifacts in the training data may result in poor learning by DL models, as they add irrelevant features to the data [10, 54]. Wright*et al.* [17] demonstrated that removing images with artifacts improved the accuracy of DL models by a significant margin. Laleh *et al.* [55] emphasized the need for robustness of DL-based CPATH systems against artifacts for their widespread clinical adaptability. Artifact processing pipeline that can detect, extract, and eliminate non-relevant patches in WSIs before running a diagnostic algorithm, avoiding any detrimental effect on downstream image analysis [11, 17, 56]. Therefore, it is essential to equip CPATH systems with artifact detection ability, which is also the focus of this work, to obtain reliable predictions [17, 57, 58].

2.2 Detection of Histological Artifacts

Most researches focus on reducing color variations and image augmentations during the pre-processing phase in CPATH literature [59, 60]. Detection of artifacts is often an underrepresented aspect of WSI pre-processing [10]. Compared to color normalization techniques, there remains a scarcity of research detecting notable artifacts before feeding histologically relevant data to the diagnostic algorithms. While some works [17, 61–63] have relied on quickly identifying faulty WSIs by doing QC at low magnification. Avanki *et al.* [64] proposed a quality estimation method by combining blurriness, contrast, brightness, etc., to accept or discard WSI based on a reference. HistoQC [65] provides content-based evaluation for finding outliers in a cohort of WSIS. Bahlmann *et al.* [63] exploited texture features and stain absorption to separate

diagnostically relevant and irrelevant regions. However, artifacts appearing in diagnostically relevant areas are likely to be missed. Apart from their limitations with lower magnification, they were validated on specific staining and tissue types. Therefore, artifact detection methods need to be extended to higher magnification. Moreover, artifact detection methods that can identify specific artifacts are desirable for QC, as some artifacts, like a blur, can be avoided by re-scanning glass slides or de-blurring techniques.

Earlier works for artifact detection relied on traditional image processing and color-space transformation approaches. Gao *et al.* [66] detected blurry areas using 44 handcrafted (local statistics, brightness, etc) features. Hashimoto *et al.* [67] combined image sharpness and noise information to create a regression model for out-of-focus detection. For folded tissue detection, Palokangas*et al.* [68] transformed red, green, and blue (RGB) images to hue, saturation, and intensity (HSI) to apply k-means clustering over the different saturation and intensity values. Bautista and Yagi [69] detected folds using RGB shift with fixed thresholding on luminance and saturation values to enhance color structure in thick (folded) areas. Kothari *et al.* [57] introduced a rank-sum approach that used connectivity descriptors and image features to discard folded tissues. Their approach used two adaptive thresholds on saturation and intensity ranges. Chadaj *et al.* [70] separated uninformative blood (hemorrhage) from blood vessels using cyan, magenta, yellow, and black (CYMK) color space and morphology. Mercan *et al.* [71] proposed a k-means method to classify blood patches using local binary patterns extracted from stains and L*a*b histograms. A detailed review of other artifact detection works can be found in Kanwal *et al.* [10]. Since color-based approaches can heavily underperform when exposed to data from different cohorts with stain variation, data-driven DL approaches are needed to resolve the challenges.

Among recent works using DL-based approaches, Albuquerque *et al.* [72] compared several DCNNs for detecting out-of-focus areas in their ordinal classification problem. Kohlberger *et al.* [73] proposed ConvFocus to quantify and localize blurry areas in WSI. Wetteland *et al.* [74, 75] proposed a segmentation model to find blood and damaged tissue in bladder cancer WSIs. Clymer*et al.* [76] developed a two-stage method to detect blood at low resolution using RetinaNet and later Xception CNN for subsequent classification. Babie *et al.* [77] used SOTA DCNNs with SVM, KNN, and decision tree classifiers to separate folded tissues from normal tissue in a binary fashion. Kanwal *et al.* [78] used several DCNNs to assess the impact of color normalization over blood and damaged tissue detection. In another work [16], they trained ViT-Tiny [33] for air bubble detection using knowledge distillation. All these works relied on training a single network to classify one or two artifacts against an artifact-free class. It is a well-known problem that DL models suffer from poor generalization, robustness, and overconfident predictions over out-of-distribution (OoD) data [79–81]. Thus, the high variance in the prediction of DL models needs to be addressed, especially when deployed in a critical application. A prominent DL technique, "deep ensembles," resolves these problems by training several baseline DL architectures and combining the resultant predictions to increase accuracy and OoD performance [5]. However, the success of the ensemble method relies on several factors, such as how baseline models are trained and integrated. The most widely used ensemble techniques

include averaging and majority voting [31]. It is worth noting that a simple aggregation using averaging methods or majority voting is not a smart choice and is very sensitive to biased baseline models [31].

A mixture of experts (MoE) may address this shortcoming by combining base learners, which are experts on detecting particular artifact morphology. Unlike deep ensemble, where all models are trained on the same data, in MoE, each DL model is trained for a sub-task to master specific aspects of data, resulting in improved robustness. To the best of our knowledge, this is the first work to provide a comprehensive DL-based artifact processing pipeline that takes the entire WSI, preprocess, inference, and post-process and excels in both artifact detection and QC applications.

3 Data Materials

This section details the histological data used for training and validating DL models. The following in-house (private) datasets are used for the experiments.

3.1 Training and Development Data

We obtained 55 glass slides of bladder cancer resection biopsies from the Erasmus Medical Centre (EMC) in Rotterdam, The Netherlands. These glass slides were formalin-fixed and stained with Hematoxylin (purple) and Eosin (pink) (H&E) dyes. The slides were scanned with a Hamamatsu Nanozoomer 2.0HT at 40× and saved in *ndpi* format with a pixel size of 0.227 μm × 0.227 μm. These WSIs were properly anonymized to preserve patient privacy, and all ethical requirements were followed before the dataset was created. A non-pathologist who had received training for this task manually annotated five artifacts (blurry areas, folded tissues, blood hemorrhage, air bubbles, and damaged tissue). The rest of the tissue was marked as an artifact-free region. Not all WSIs were extensively labeled as distinct tissue types since this histological data is not used for any task other than artifact detection. However, each WSI had at least one annotation for RoI or the artifact region (our case). In the later sections, we refer to this cohort as EMC_{dev}. A detailed description of the prepared dataset and its availability is mentioned in Section 4.1.

3.2 External Validation Data

We have used the following datasets for inference only to assess the generalizability and robustness of artifact processing pipelines.

3.2.1 EMC Cohort:

This dataset is a collection of high-risk non-muscle invasive bladder cancer WSIs from a multi-center cohort provided by Erasmus MC, Rotterdam, The Netherlands. These WSIs with *MRXS* format were prepared with H&E staining and scanned with a 3DHistech P100 scanner at 80x magnification. A few WSIs were selected and annotated (by FK) based on the presence of artifacts and annotated them to test their generalization ability. We have used a 40x magnification level for inference as the models are trained

at a similar level. We will refer to this dataset as EMC_{inf}, and it is a different cohort than the above-mentioned EMC_{dev}.

3.2.2 SUH Cohort:

This dataset is a private triple-negative breast cancer cohort of 258 surgical specimens. This dataset contains H&E WSIs prepared from surgical specimens and collected by the Stavanger University Hospital (SUH) in Norway between 1978 and 2004. The WSIs are in *ndpi* format and scanned using the Hamamatsu NanoZoomer S60 at 40× magnification. An expert breast pathologist (UK) selected and annotated a few WSIs based on the severity of the presence of these artifacts. We have used these WSIs to test DL pipelines over cancer types different from the ones they are trained on. We will refer to this dataset as SUH_{inf}.

3.2.3 INCLIVA Cohort:

This dataset is prepared by the department of anatomical pathology of the Hospital Clínico Universitario de Valencia, Spain, and is a collection between 1988 and 2020. The prepared WSIs are from skin cancer biopsies with Spitz tumors and were scanned with Roche's Ventana iScan HT at 40x magnification. WSIs were saved in *tiff* format. An expert dermatopathologist (AM) selected and annotated a few WSIs with artifacts from this cohort to validate the proposed pipeline over the external cohort. We will refer to this dataset as $INCLIVA_{inf}$

Table 1: Breakdown of the number of patches obtained in each class of the dataset $\mathcal{D}$, obtained form EMC_{dev} after preprocessing.

(label) Class	(35 WSIs) Training	(10 WSIs) Validation	(10 WSIs) Test	Total
(0) Artifact free	5,249	1,441	965	7,655
(1) Blood	16,743	4,186	1,409	22,338
(2) Blur	5,661	754	1,137	7,552
(3) Air bubbles	2,499	1,175	846	4520
(4) Damaged Tissue	2,577	332	1,023	3,932
(5) Folded Tissue	998	114	131	1,243

4 Proposed Method

This section describes the data pre-processing, the proposed method for MoE, post-processing, evaluation metrics, and implementation details for the DL pipelines.

Figure 1 gives a graphical overview of the proposed DL method for detecting histological artifacts in WSIs. We proceed with the artifact detection task in two steps. First, we train binary and multiclass models for patch-wise classification. The binary models are trained to detect one particular artifact, i.e., blur against artifact-free. The multiclass models provide output with six classes (five artifacts and artifact-free). In the second step, we used these trained binary models to create a sort of MoE for

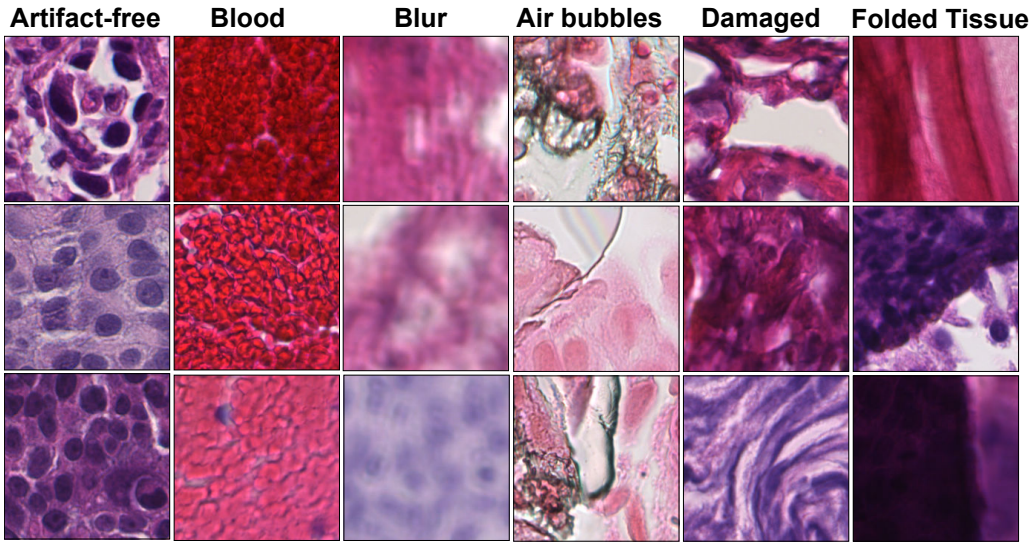

Fig. 2: Examples of artifact-free and artifact-classes patches in our prepared dataset $\mathcal{D}$ from EMC_{dev}, and extracted at 40x magnification.

inference and post-processing the predictions. We deploy multiclass models directly. We combine predictions from each expert in MoE by fusing their outputs. We apply a probability threshold for maximizing sensitivity for detecting notable artifacts and providing artifact-free WSI with diagnostic potential. A detailed description of the proposed method is given below.

4.1 Pre-processing

We have used the EMC_{dev} cohort to prepare the dataset. This included WSIs from this cohort, which were divided into 35/10/10 training, validation, and test WSIs to prevent data leakage.

Let a WSI at magnification level 40x (sometimes known as 400x) be denoted by $I^{40x}_{\text{WSI}(i)}$. Since I^{40x}_{WSI} are huge gigapixel images, it is not possible to process the entire WSI in compute memory at once. To make computation feasible, most CPATH systems first tile or patch the WSI, or the RoI, before processing it further. The initial step in the patching procedure was to separate the foreground tissue from the background (white) areas irrelevant for image analysis. Foreground/background separation is usually done with a low-resolution version of the image, which can later be interpolated to be used with the full-resolution image. We obtained tissue foreground by transforming the RGB (red, green, and blue) color space to HSV (hue, saturation, and value). Later, Otsu thresholding was performed on the value channel to separate the foreground-containing tissue from the background. We set a uniform patch-coordinate sampling grid over the extracted foreground. Patches having at least 70% overlap with the annotation area (R) were retrieved after the extracted foreground was tiled across the grid with a non-overlapping stride, as depicted in Figure 1.

Assuming $\mathcal{T} : I^{40x}_{\text{WSI(i)}\in R} \rightarrow \{\boldsymbol{x}^i_j; j = 1 \cdots J\}$ denotes the tiling process, which gives a set of J patches over R. Here, $\mathbf{x}^i_j \in \mathbb{R}^{W\times H\times C}$ corresponds to patch j with coordinates (x_{i0}, y_{i0}) from WSI_i and H, W, and C represent the width, height, and channels of the patch, respectively. We refer to this prepared dataset from EMC_{dev} as $\mathcal{D} = (\mathbf{X}, \mathbf{y}) = \{(\mathbf{x}_n, \mathbf{y}_n)\}^N_{n=1}$ containing N patches. Here $\mathbf{x}_n$ is n-th instance with $224\times224\times3$ pixels and $y_n \in \{0, ..., k\}$, where $k = 1$ for binary and $k = 5$ for multiclass dataset formulation. For instance, in a multiclass dataset, '0' represents artifact-free class, and {1,2,3,4,5} corresponds to blood, blur, air bubbles, damaged tissue, and folded tissue classes, respectively. Table 1 shows the breakdown of patches in each subset of the dataset $\mathcal{D}$ and Figure 2 shows example instances for all classes obtained from I^{40x}_{WSI}. This training and development dataset is made publicly available and can be downloaded from Zenodo.

4.2 Feature Extractors and Classifiers

The feature extractor and classifier are two significant components of any DL model. Feature extractors are crucial in DL algorithms as they help identify critical features in the data. In short, it reduces the dimensionality of the image and facilitates classification from a vector. Based on artifact detection works in the literature [16, 78], we have selected two popular DL architectures as feature extractors due to their smaller parametric size and faster inference: i) DCNN-based MobileNetv3 [32] architecture, and ii) Vision transformer-based ViT-Tiny [33] architecture.

MobileNetv3: MobileNetV3 is a SOTA DCNN architecture proposed by Howard*et al.* [32] and is part of the family of computationally efficient models for small devices by Google. The basic building blocks of MobileNetV3 include depth-wise separable convolutions and inverted residual blocks designed to reduce computational complexity and improve accuracy. MobileNetV3 is optimized through a combination of hardware-aware network architecture search and novel architecture advances, including the use of hard-swish activation and squeeze-and-excitation modules [32]. This architecture is released in different variants. The large architecture variant (used in this work) has a 5.4M parameter and is lightweight and efficient, making it suitable for computationally efficient image classification pipelines.

Vision Transformer: Vision Transformers (ViTs) have gained attention as a new SOTA for image recognition tasks [26, 27]. ViT architecture breaks down an input image into a series of smaller patches, linearly embeds each patch, adds position embeddings, and then feeds the resulting sequence of vectors to a standard Transformer encoder [82]. This Transformer encoder consists of a stack of identical layers. It uses a self-attention mechanism that allows it to focus on different parts of the input by computing a weighted sum of the input features based on their similarity. We use a lightweight and efficient variant of the ViT architecture, ViT-Tiny [33], with 6M parameters for faster inference.

We apply transfer learning to train DL models and update model parameters at each epoch. Assume ϕ represents our feature extractor with θ_f parameters. Then, for the input patch $(\mathbf{x}_n)$ with ground truth (y_n), we get a flattened feature embedding (a_n) using;

$$\phi_{\theta_f}(\mathbf{x}_n) = a_n \quad where \quad a_n = \{a_1, a_1, ..., a_z\} \tag{1}$$

For patch-wise classification, we train classifiers in a binary and multiclass fashion. We appended a three-layer fully connected (FC) classifier (C_{θ_c}) at the end of the feature extractor. Let us denote our DL models with notation ψ_θ, where $\theta = \theta_f \cup \theta_c$, denotes the parameter set of both the feature extractor and the classifier. To obtain the output probability vector ($\boldsymbol{P}_{\boldsymbol{y_n}}$) for the input patch, we apply softmax (σ) to the output logits of the classifier as shown in Eq. (2). For instance, binary models predict (artifact vs. artifact-free), and multiclass models predict (5 artifact classes vs. artifact-free), as shown in Eq. (3).

$$\boldsymbol{P}_{\boldsymbol{y_n}}(\mathbf{x}_n) = \psi_\theta(\mathbf{x}_n) = \sigma(C_{\theta_c}(\phi_{\theta_f}(\mathbf{x}_n))) = \sigma(C_{\theta_c}(a_n)) \tag{2}$$

$$\boldsymbol{P}_{\boldsymbol{y_n}} = \begin{cases} [p_{y_0}, p_{y_1}]^T & \text{if} \quad \text{binary} \\ [p_{y_0}, p_{y_1}, p_{y_2}, p_{y_3}, p_{y_4}, p_{y_5}]^T & \text{if} \quad \text{multiclass} \end{cases} \tag{3}$$

Here, y_{p_0} is the probability of being an artifact-free class. In the binary model, y_{p_1} corresponds to artifact class and in the multiclass model $[y_{p_1}, y_{p_2}, y_{p_3}, y_{p_4}, y_{p_5}]$ are predicted probabilities for blood, blur, air bubbles, damaged tissue, and folded tissue classes respectively. Finally, We calculate cross-entropy loss between the ground truth and the prediction, back-propagate this loss, and update model parameters, θ, at each epoch based on the experimental setup explained in Sec 4.7.

$$L_{CE}(y_n, P_{y_n}) = \begin{cases} -y_n \cdot log(p_{y_0}) + (1 - y_n) \cdot log(1 - p_{y_0}) & \text{for binary} \\ -\sum_{i=0}^{k} y_n \cdot log(p_{y_i}) & \text{for multiclass} \end{cases} \tag{4}$$

To obtain final predictions ($\hat{P}_{y_n}$) for classes, we apply argmax to $\boldsymbol{P}_{\boldsymbol{y_n}}$.

$$\hat{\boldsymbol{P}}_{\boldsymbol{y_n}} = argmax(\boldsymbol{P}_{\boldsymbol{y_n}}) \tag{5}$$

At the inference stage, we establish four DL pipelines using combinations of trained models, i.e., multiclass models (with MobileNetv3 and ViT-Tiny) and MoEs (combining binary MobileNetv3 and ViT-Tiny), as explained further in the following sections.

4.3 Mixture of Experts

The "mixture of experts (MoE)" DL approach is often confused with deep ensembles. A deep ensemble combines DL models trained on the same data using different seed initializations or hyperparameters to learn different aspects of the data [81]. Unlike deep ensemble, in MoE, each DL model is trained for a specific task (blur, fold, blood, folded tissue, and damaged tissue detection) to become a specialist in particular

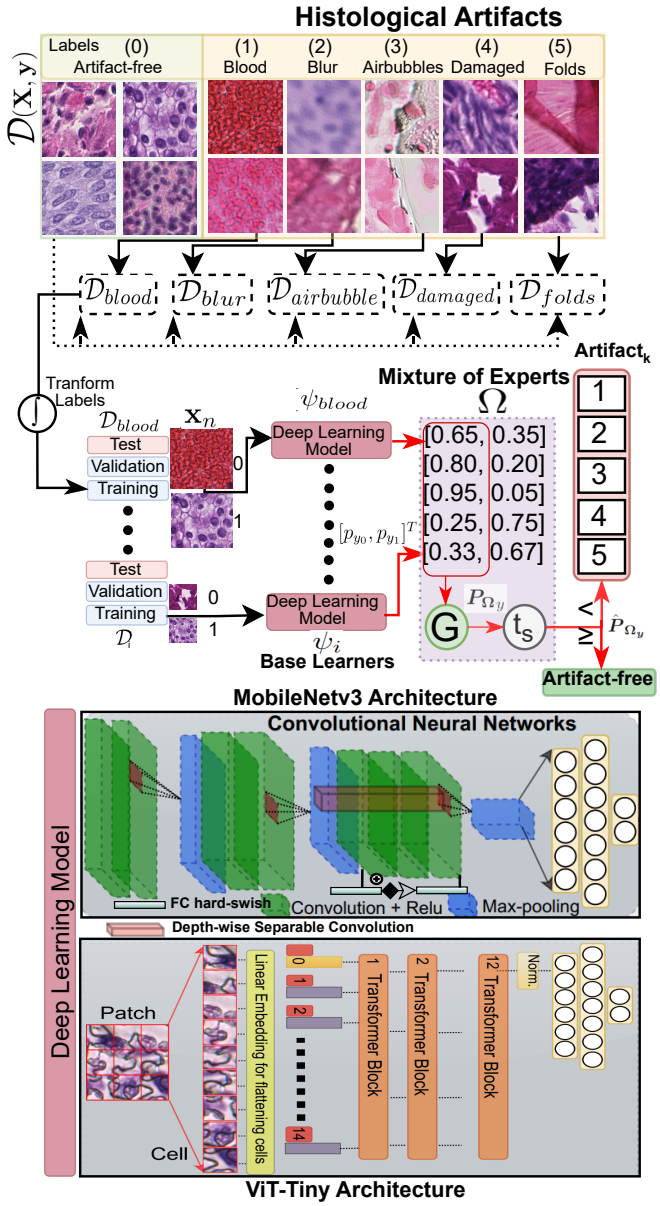

Fig. 3: **An overview of the mixture of experts (MoE) formation for artifact detection**. Five base learners (either MobileNetv3 or ViT-Tiny deep learning architectures) are trained on overlapping sub-datasets to learn the distinct morphology of each artifact. Labels are transformed to take the artifact class as a negative class. A fusion function integrates output from all experts to form a predictive probability distribution for the final prediction. A meta-learned probability threshold is applied to maximize the sensitivity of the MoE.

aspects of data. Instead of applying simple majority voting like deep ensembles, a gating mechanism forms the final prediction, incorporating output from diverse experts and resulting in improved robustness.

Our proposed DL scheme is a kind of MoE where we integrate five identical DL architectures (also called base learners or experts) after training on the parts of the data (similar to bagging). Bagging offers the advantage of reducing variance, thus eliminating overfitting by training models on subsets of data. This parallel and data-independent training strategy avoids affecting the results of other experts. We form two MoE-based DL pipelines, namely ViTs-based MoE and DCNNs-based MoE, by choosing five base learners (either DCNN or ViT architectures as explained in Sec 4.2). All these experts are trained on five overlapping subsets, $\{\mathcal{D}_{blood}, \mathcal{D}_{blur}, \mathcal{D}_{airbubble}, \mathcal{D}_{damaged}, \mathcal{D}_{folds}\} \in \mathcal{D}$. Here, each sub-dataset contains a distinct artifact class and the same artifact-free class as shown in Figure 3. For simplicity, we transform ground truth labels as a positive class with the label '1' for artifact-free and a negative class with the label '0' for the artifact class.

The contingent MoE model, Ω, forms a single prediction using the aggregation function (G). G is similar to gating, which combines the output probabilities of the experts using a fusion approach. In short, the proposed approach formulates MoE trained on individual artifact morphology detection tasks. For artifact models $\psi_i \in \{\psi_{blood}, \psi_{blur}, \psi_{airbubble}, \psi_{damage}, \psi_{fold}\}$, we only utilize the prediction for negative class (P_{ψ_0}) (a.k.a probability of being an artifact), and fuse binary outputs for Ω as shown in Eq. (7).

$$\boldsymbol{P}_{\boldsymbol{\Omega}_y} = G(\psi_{blood}, \psi_{blur}, \psi_{airbubble}, \psi_{damage}, \psi_{fold}) \tag{6}$$

$$\boldsymbol{P}_{\boldsymbol{\Omega}_y} = \begin{cases} 1 - max(P_{\psi_{i_0}}) & \text{for artifact-free (positive) class} \\ max(P_{\psi_{i_0}}) & \text{for artifact (negative) class} \end{cases} \tag{7}$$

To evaluate the final prediction ($\hat{\boldsymbol{P}}_{\boldsymbol{\Omega}_y}$), we adopt a form of meta-learning by placing a constraint on maximizing the sensitivity of the model for the positive (artifact-free) class. Therefore, we introduce a probability threshold, t_s, to handle previously unseen tissue morphology and avoid misclassifying artifact-free patches with potential diagnostic relevance. In other words, if the probability of being a positive class in $\boldsymbol{P}_{\boldsymbol{\Omega}_y}$ is higher than t_s, then we assign *artifact-free* label to the patch as shown in Eq. (8). Here, t_s would help to efficiently minimize false negatives without re-training models with a new cohort of WSIs with different tissue types or staining. We determine the best value of t_s by maximizing the true positive rate (sensitivity) in the receiver operating characteristic (ROC) curve over the validation data.

$$\hat{\boldsymbol{P}}_{\boldsymbol{\Omega}_y} = \begin{cases} Artifact - free & \text{if} \quad P_{\Omega_{y_0}} \geq t_s \\ Artifact_k & \text{Otherwise} \quad k \in \{1, 2, 3, 4, 5\} \end{cases} \tag{8}$$

4.4 Multiclass Models

In case of multiclass models (ψ_{multi}) with predicted probability distribution $\boldsymbol{P}_{\boldsymbol{\psi}_{\boldsymbol{y_i}}} \forall \quad i \in \{0, 1, 2, 3, 4, 5\}$. We find the probability threshold (t_s) by maximizing sensitivity similar to MoE (see Sec. 4.3). In other words, if the predicted probability for the artifact-free class is higher than t_s, then the patch is assigned *artifact-free* label. Otherwise, the artifact label with the highest probability value is assigned (see Eq. (9)).

$$\hat{\boldsymbol{P}}_{\boldsymbol{\psi}_{\boldsymbol{multiy}}} = \begin{cases} Artifact - free \quad \text{with} \quad p_{\psi_{y_0}} \quad \text{if} \quad p_{\psi_{y_0}} \geq t_s \\ Artifact_k \quad \text{with} \quad p_{\psi_{y_k}} \quad max(p_{\psi_{y_1}}, p_{\psi_{y_2}}, ..., p_{\psi_{y_k}}) \quad \text{Otherwise} \end{cases} \tag{9}$$

4.5 Post-processing

At the inference stage, we utilize predictions for both artifact detection and QC applications, as illustrated in the post-processing part of Figure 1. Since the predictions of DL models are patch-based, we need to stitch patches back to see the overall view of the tissue in the WSI structure. However, stitching smaller patches introduces boundary artifacts (blockish appearance) [4]. To avoid this problem, we turn to the matrix-filling approach.

For patch x_i with coordinates (x_0, y_0), the next consecutive patch $x_{(i+1)}$ holds the difference of sampling stride (s) with coordinates $(x_1, y_1) = (x_{0+s}, y_{0+s})$. Here, s equals the patch size owing to a uniform, non-overlapping grid. For the segmentation map, we use a matrix (M), a downscale version of the original resolution, to assign predicted class k.

$$\begin{aligned} M[x_0 : x_0 + s, y_0 : y_0 + s] = k \quad \text{where} \quad s = 224 \quad \text{(patch-size)} \\ M[x_1 : x_1 + s, y1 : y_1 + s] = k \quad \text{where} \quad k = \{0, 1, ..5\} \end{aligned} \tag{10}$$

Since M is down-scaled to sampling stride size, every filled box can be seen as a pixel in the final segmentation map (see 1 in Figure 4). We use filled M for the artifact report to calculate the percentage of predicted patches with artifact class k over the total patches N_{tot} in the foreground. See 2 in Figure 4 for an example artifact report for QC.

$$Per_k = \frac{N_k}{N_{tot}} * 100\% \quad \text{where} \quad N_k = \text{predicted with class} \quad k \tag{11}$$

We denote the artifact-free post-processed region as ρ. It measures the usefulness of the WSI and can be compared against a predefined threshold τ for assessing its suitability (accepting or discarding) for developing DL algorithms.

$$\rho = \frac{\text{Number of artifact-free pixels} \quad (N_{k_0})}{\text{Total number of pixels in the foreground} \quad (N_{tot})} \tag{12}$$

To highlight the histologically relevant region, we binarize M to M_ρ and treat all artifact classes as a single class as shown in Eq. (13). The binary mask (M_ρ) indicates the potentially histologically relevant RoI (see 3 in Figure 4). Later, we apply a morphological closing operation to remove small holes in the final mask.

$$M_{\rho_{(i,j)}} = \begin{cases} 1, & \text{if} \quad M_{(i,j)} = k_0 \quad \text{(artifact-free)} \\ 0, & \text{Otherwise} \end{cases} \tag{13}$$

Finally, obtain artifact-free WSI by performing the Hadamard product between M_ρ and the original WSI ($I \in \mathbb{R}^{m \times n}$) with the dimensions of $m \times n$ (see Eq (14)). Using the nearest interpolation, we resize the M_ρ mask to $m \times n$. Let's denote the element at the i-th row and j-th column of M_ρ as $M_\rho(i, j)$, and the corresponding element in I as $I(i, j)$. This element-wise operation between M_ρ and I removes any regions or areas with the presence of artifacts (4 in Figure 4) and $I_{artifact-free}$ can be written as:

$$(I \odot M_\rho)_{ij} = \begin{bmatrix} M_{\rho_{(1,1)}} \cdot I_{(1,1)} & M_{\rho_{(1,2)}} \cdot I_{(1,2)} & \dots & M_{\rho_{(1,n)}} \cdot I_{(1,n)} \\ M_{\rho_{(2,1)}} \cdot I_{(2,1)} & M_{\rho_{(2,2)}} \cdot I_{(2,2)} & \dots & M_{\rho_{(2,n)}} \cdot I_{(2,n)} \\ \vdots & \vdots & \ddots & \vdots \\ M_{\rho_{(m,1)}} \cdot I_{(m,1)} & M_{\rho_{(m,2)}} \cdot I_{(m,2)} & \dots & M_{\rho_{(m,n)}} \cdot I_{(m,n)} \end{bmatrix} \tag{14}$$

4.6 Evaluation Metrics

For performance comparison, we report accuracy, sensitivity, and the F1-score. Let TP, FN, FP, and TN denote true positive and false negative, false positive, and true negative predictions, respectively. Here, a positive class refers to a patch without artifacts (artifact-free patch). Then, the confusion matrix (CM) is a tabular representation of the model's predictions using TP, FN, FP, and TN. Accuracy is the proportion of correct predictions to the total number of predictions and is defined as $Acc. = (TP + TN)/(TP + FN + FP + TN)$. Sensitivity, also known as recall, measures the proportion of actual positives correctly identified by the model and is termed $Sens. = TP/(TP + FN)$. High sensitivity is essential to retaining potentially relevant (artifact-free) RoIs for the diagnostic algorithm. On the other hand, specificity $Specs. = TP/(TP + FP)$, quantifies the performance of a model in distinguishing negative instances from those falsely labeled as positive. In our application, high specificity filters out irrelevant information (artifacts) appearing in relevant RoIs. The F1 score is the harmonic mean of precision and recall and is calculated as $F1 = 2 \cdot (\text{precision} \cdot \text{recall})(\text{precision} + \text{recall})$, where precision $= TP/(TP + FP)$. For overall segmentation, dice co-efficient is reported. Dice scores the overlap between the predicted segmentation and the ground truth and ranges from 0 to 1, where 1 indicates perfect overlap between the predicted and ground truth segmentation. We use model weights with the lowest validation loss during the training to report these evaluation metrics.

For computational complexity evaluation, we have considered FLOPS, parameters, and inference time. FLOPS measures the number of floating-point operations required by a specific algorithm. The number of parameters refers to the learnable parameters in the model that are used to perform operations, where high parameters result in more FLOPS. Finally, inference time is the time the DL model consumes to make predictions over a patch. These metrics, combined, provide a comprehensive understanding of the

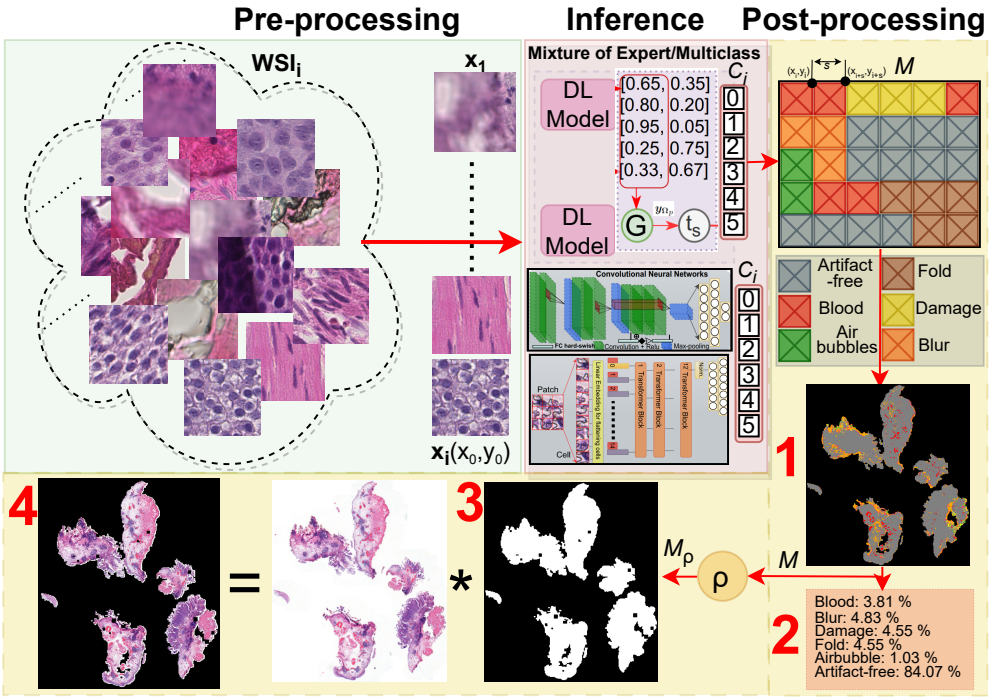

Fig. 4: Overview of deep learning (DL) pipeline emphasizing the post-processing stage during the inference. *Pre-processing:* The whole slide image (WSI) is split, and every patch is stored with its corresponding coordinate. *Inference:* Every patch is assigned a label using a mixture of experts or multiclass DL models. *Post-processing:* The matrix-based filling method assigns a color to every pixel (in the downscaled version of WSI) at the corresponding coordinate location. Post-processing provides: 1). Segmentation map; 2). Artifact report for quality control; 3). Artifact-free region of interest map, and 4). Artifact-refined WSI for computational analysis.

DL model's performance and computational efficiency, which are crucial for assessing the practical applicability in real-world scenarios.

4.7 Implementation Details

The code was implemented using Python. The patch extraction was accomplished using the Pyvips [1] library. During the patching, We used torch multiprocessing [2] to carry out process pooling for faster pre-processing. The extracted patches were standardized to the mean and standard deviation of ImageNet [83] due to transfer learning over ImageNet weights. To compensate for the scarcity of labeled data, augmentation is applied at each epoch during the training [34, 84]. We used random geometric

[1] https://libvips.github.io/pyvips/
[2] https://pytorch.org/docs/stable/multiprocessing.html

transformations, including rotations and flips both horizontally and vertically. Our DL models consist of a feature extractor and a classifier with three fully connected (FC) layers. We used state-of-the-art architectures MobileNetv3 [32] and ViT-Tiny [33] as backbones for feature extractors. MobileNetv3 was borrowed from the Pytorch [3] DL framework, and ViT-Tiny was taken from the Timm [4] library. Both of these backbones were initialized with ImageNet weights. We referred to best hyperparameter settings from works [16, 58, 78], and fixed final parameters to cross-entropy loss, SGD optimizer, ReduceLRonPlateau scheduler initialized with 0.01, batch size of 128, and early stopping of 20 epoch over the validation loss to avoid overfitting, drop out of 0.2 and fixed random seed for reproducibility. All training and inference experiments were done on Nvidia A100 40GB GPU. The source code is available at Github.

5 Experimental Results and Discussion

This section presents experimental results for training and validating DL pipelines of the EMC cohort and discusses their performance on validation, testing, and external data.

Table 2: **Performance of artifact processing pipelines on the validation set of EMC_{dev} cohort 3.1.** Various DL pipelines, including the mixture of experts (MoE) and multiclass models using SOTA DCNN and ViT architectures, are deployed. A simple binary formulation is used for a fair comparison, and accuracy for the artifact-free class is reported. The best results are marked in bold, and the second-best results are underlined in each column.

DL architecture		$Acc.(\%)$	$F1$	$Acc._{afree}$	$F1_{afree}$	$Sens._{afree}$
DCNNs	MoE	92.08	91.87	<u>97.82</u>	<u>88.66</u>	90.12
	Multiclass	93.48	93.43	94.96	78.64	**96.89**
	Binary	<u>95.92</u>	<u>95.26</u>	-	-	<u>94.68</u>
ViTs	MoE	94.81	94.53	**97.84**	**89.06**	91.92
	Multiclass	94.29	94.48	96.79	83.80	86.84
	Binary	**97.45**	**97.46**	-	-	87.25

5.1 Validation on the EMC_{dev} Cohort

This experiment aims to evaluate the performance of the proposed MoE and multiclass models for artifact detection. These pipelines consist of four DL approaches using MoE and multiclass models based on DCNNs (MobileNetv3 [32]) and ViTs (ViT-Tiny [33]). For simplicity, we will refer to DCNNs or ViTs in the discussion. For a baseline comparison, we also trained binary classification models (DCNN and ViT) using the entire EMC_{dev} dataset in a binary fashion. In other words, we wanted to

[3] https://github.com/pytorch/pytorch
[4] https://timm.fast.ai/

compare the benefits and drawbacks of the simpler classification model against a MoE and their computational and performance trade-offs for efficient DL pipelines.

We will first focus on discussing the performance aspect. Table 2 presents classification results over the EMC_{dev} validation subset. We have reported metrics for artifact-free classes to compare them fairly against baseline (binary) models. For better classification performance, we desire high sensitivity to avoid misclassifying artifact-free patches as artifacts and retain potential histologically relevant tissue for automated diagnostics. This is because the artifact detection application is not affected by one artifact class being classified as another. In the end, patches with the presence of any artifacts will be excluded from downstream (diagnostic) applications. Though the baseline models yield the best overall accuracy, they relatively underperform and exhibit lower sensitivity in classifying the artifact-free class. The MoEs outperform multiclass models and baseline models in detecting artifact-free class. Overall, both MoE pipelines give superior results for the positive class and avoid false negatives. However, the DCNN-based multiclass model gives the best sensitivity score. To present an unbiased view, we test MoEs and multiclass models on unseen data from the same EMC_{dev} cohort.

Table 3: **Generalization results on the test set of EMC_{dev} cohort 3.1.** The table presents results over unseen data, with and without probabilistic thresholding. All metrics are calculated for the classification performance over artifact-free class. The best results in each column are marked in bold, and the second-best results are underlined.

DL architecture		Without probabilistic threshold				With probabilistic threshold	
		Acc. (%)	F1	Sens.	t_s	F1	Sens.
DCNNs	MoE	**97.82**	<u>88.66</u>	89.12	0.326	**86.15**	**97.93**
	Multiclass	93.58	85.21	**94.72**	0.341	83.53	95.47
ViTs	MoE	<u>95.61</u>	**88.91**	<u>90.45</u>	0.052	<u>84.90</u>	<u>97.83</u>
	Multiclass	92.55	82.51	89.94	0.015	70.15	96.54

We present generalization results in Table 3. The table reports mixed results when probabilistic thresholding is not applied. To improve the sensitivity over new data, we learn a probability threshold (t_s) using ROC curves of the validation set (see Section 4.3), as displayed in Figure 5. We target a 98% sensitivity and obtain different t_s values for each DL pipeline, as reported in Table 3. Interestingly, the DCNN-based pipelines assign higher probability scores to the artifact-free class, indicating better confidence and stronger learning of histologically relevant morphology than the ViT-based models. Figure 6 reflects similar insight that ViT-based pipelines carry weak differentiation between artifacts and artifact-free patches (see black dotted line). It is fascinating to see that probabilistic thresholding significantly improves the ability to detect artifact-free class, hinting that the proposed MoEs would be the best choice with the least false negatives.

To evaluate the computational aspect, Table 4 indicates the computational complexity of all four DL pipelines. Undoubtedly, MoEs have nearly five times more parameters and lower throughput than multiclass models. This is because each MoE

combines five binary experts. Comparatively, DCNN-based pipelines can be efficient at the inference stage due to very little patch processing time per second. We have to make a trade-off in selection, either choosing multiclass DCNN with better computational efficiency but relatively lower performance or based on the best performance. We prioritize classification performance and opt for the two best-performing DL pipelines from Table 3; therefore, we will use MoEs for the following experiments.

Table 4: **A comparative analysis of computational complexity.** Lower values of parameters and flops indicate computationally efficient models, and higher throughput is desired for faster inference.

DL Pipelines	Parameters (M)⇓	Flops (B)⇓	Throughput (p/sec.)⇑
MoE (DCNNs)	17.65	1.13	178
MoE (ViTs)	27.62	5.38	128
Multiclass (DCNN)	3.53	0.22	832
Multiclass (ViT)	5.53	1.08	419

Table 5: **Results for quantitative evaluation for assessing the robustness of the proposed mixture of experts (MoE) approach.** Qualitative evaluation is performed on external (out-of-distribution) data. The table reports classification performance corresponding to patch-wise classification and dice scores for overall segmentation maps obtained through artifact processing pipelines.

DL Pipeline	Cohort	WSIs	$F1_{afree}$	$Sens._{afree}$	$Spec._{afree}$	Dice
MoE of DCNNs	EMC_{inf}	s1	92.86	93.48	53.76	0.909
		s2	89.11	89.61	52.71	0.784
	SUH_{inf}	s3	70.91	55.07	99.09	0.487
		s4	85.51	79.78	44.57	0.572
	$INCLIVA_{inf}$	s5	60.05	43.99	80.53	0.532
		s6	37.39	23.55	98.97	0.506
MoE of ViTs	EMC_{inf}	s1	93.17	93.01	60.92	0.939
		s2	89.34	87.97	63.18	0.795
	SUH_{inf}	s3	68.79	54.51	79.56	0.367
		s4	87.97	85.63	26.38	0.482
	$INCLIVA_{inf}$	s5	78.92	66.02	79.71	0.559
		s6	45.49	42.49	42.91	0.412

5.2 Quantitative Evaluation

We perform this experiment to assess the robustness of DL pipelines over external (OoD) data. For this purpose, we chose six WSIs (s1-6) from external validation data (see Section 3.2). Note that all these WSIs were prepared and scanned by different

laboratories and scanning hardware. Thus, they exhibit vast differences in staining, tissue types, and image acquisition protocols, as displayed in Figure 7. We did not incorporate color normalization in the artifact processing pipeline due to their additive computational cost and latency [78].

Quantitative assessment is crucial to objectively evaluate the numerical performance, enabling us to compare both proposed MoEs of DCNNs and ViTs. We require histological correctness that only an expert can provide in the form of ground truths. Therefore, all WSIs were roughly annotated by FK, UK, and AM for different artifacts. Table 5 presents the results for classification and segmentation performance. Since certain artifacts, such as folded tissue, have blurry areas surrounded [10]; one artifact class is likely to be predicted as another. Thus, for simplicity purposes, we report metrics for artifact-free (positive) classes only.

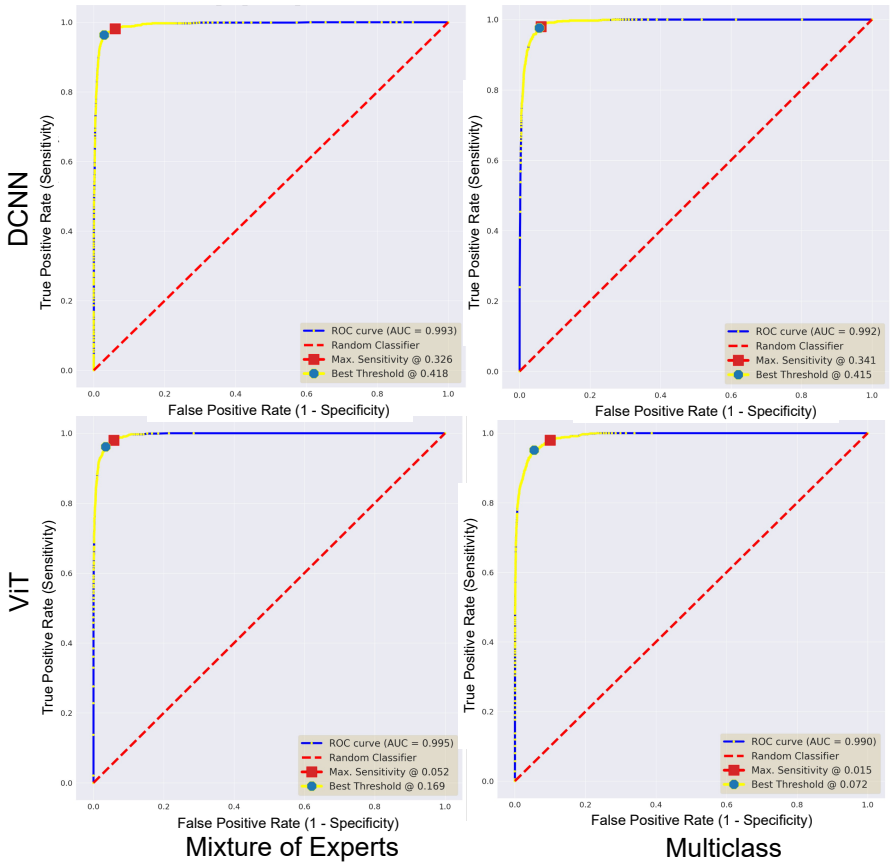

Fig. 5: **ROC curves for deep learning pipelines over the validation subset.** All plots highlight the area under the curves (AUC) score and best probability thresholds for maximizing F1 and sensitivity metrics.

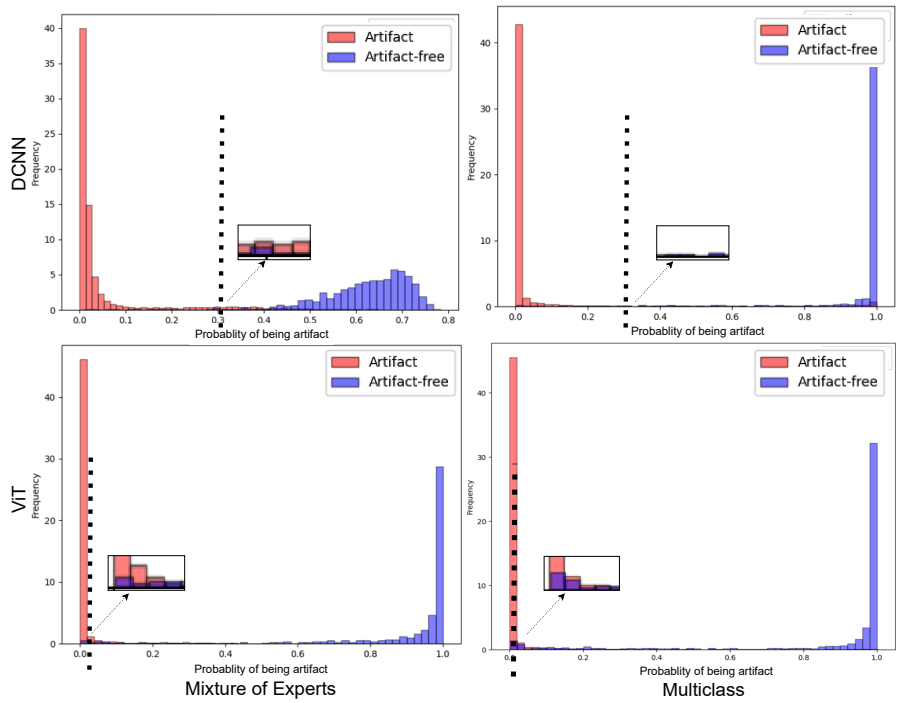

Fig. 6: **Classification plots for deep learning pipelines over the validation subset.** All subplots highlight the delineation (black dotted line) with the estimated value of t_s for probabilistic thresholding.

Both MoE pipelines experience a drop in sensitivity over breast cancer (SUH_{inf}) and skin cancer ($INCLIVA_{inf}$) WSIs. This behavior could be due to misclassifying ambiguous regions or susceptibility to specific tissue types. Since SUH_{inf} and $INCLIVA_{inf}$ WSIs are OoD data for our DL pipelines, it is interesting to see that we get high specificity scores. In short, both pipelines ensure that most of the actual artifacts present in the data are accurately flagged. Dice score in Table 5 shows good segmentation results on the EMC_{inf} cohort. Nevertheless, EMC_{inf} is bladder cancer tissue and may carry more similarity in structural appearance.

Quantitative metrics can miss subtle nuances masked by overall performance scores. Therefore, we observe false predictions of both DCNNs-based MoE and ViT-based MoE over better performance (s_1, s_4, and s_5) and the worst (s_2, s_3, and s_6) performance in OoD data. Figures 8 and 9 show ground truths and predictions masks for the better results in each cohort, and Figures 10 and 11) shows the same for the worst results in each cohort. Both MoEs densely predict artifacts in all three examples. Here, false negative instances pertain to regions identified as artifacts but were labeled artifact-free. Conversely, false positives are cases classified as artifact-free but were labeled as any artifact class. Figure 10 highlights that DCNN-based MoE might be overdoing their job predicting certain artifacts like air bubbles. For instance, in s_6,

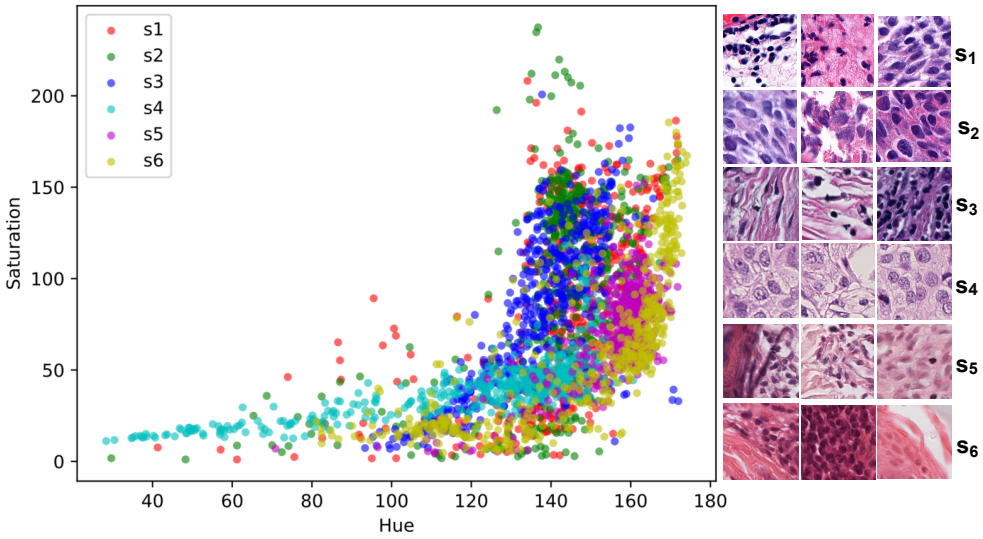

Fig. 7: **Hue-Saturation plot shows massive variation in the external (out-of-distribution) data.** Random patches from all six WSIs (s1-6) are used to calculate hue and saturation values to observe the depth of H&E staining. WSI acquisition procedures from different laboratories and scanning hardware affect the final appearance of histological images (as shown on the right).

the entire WSI has a hazy appearance, with air trapped under most of the tissue. The false predictions for s_6 show that those examples lack cellular features. Likewise, for false positives in the case of s_2, those specific examples were the boundary of another artifact region and contained some presence of blood. In case s_2 and s_3, annotations had some noise, and with the chosen mask overlap, the obtained ground truth was not accurate enough. On the other hand, the ViT-based MoE (in Figure 11) appears to be slightly overdoing damage detection. In most false predictions here, we might be dealing with potentially noisy and imprecise ground truth annotations. Therefore, relying on only quantitative analysis is not concrete and conclusive. We require a thorough qualitative analysis by field experts to scrutinize further the strengths and weaknesses of both MoEs in detecting artifacts.

5.3 Qualitative Evaluation

In this experiment, we perform qualitative evaluations by three field experts to delve deeper into the DL pipelines' behavior and see the holistic view after artifact refinement. While quantitative metrics provide valuable numerical insights into a model's performance, they often fall short of capturing the intricacies of segmentation results. Therefore, assessing whether the model was misclassified due to genuine limitations or imperfections in the ground truth is vital.

Three field experts (P1, P2, and P3) assessed segmentation maps for six WSIs (s1-s6) from three cohorts used in the above experiment. They scored them based on visual interpretation, including how well artifacts were detected, how artifact-free regions were preserved, and the overall diagnostic usability of WSIs after the artifact

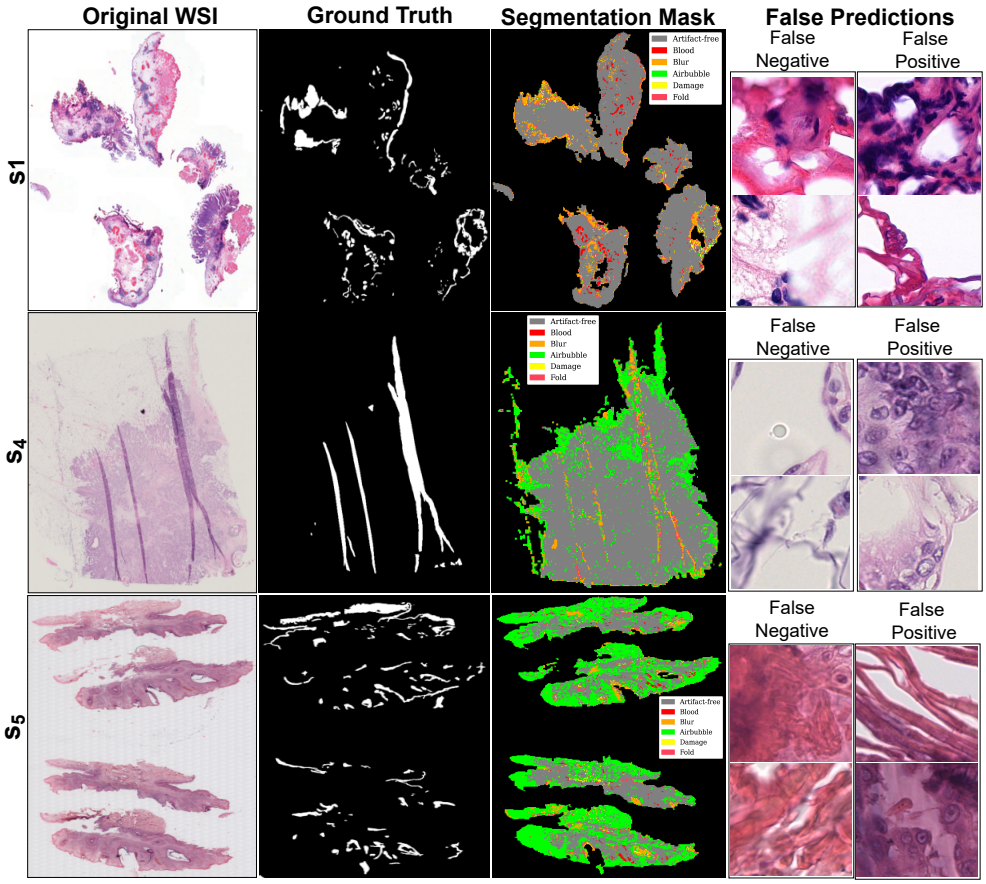

Fig. 8: **Visualization of DCNNs-based mixture of experts' predictions with better performance over out-of-distribution data.** Image shows original WSIs (s1,s4, and s5) along with ground truth for artifacts (combined), artifact segmentation map, and a few examples of false predictions. False negative refers to patches detected as artifacts but were artifact-free, and false positive refers to patches detected as artifact-free but belonged to any artifact class.

processing, where field experts scored them from 1 (worst) to 10 (best). Each expert who rated these WSIs was a domain specialist on a specific cancer type (See box plot in Figure 12). Figure 12 represents the score variability for each task across the six WSIs. The central line in each box represents the median, while the box's upper and lower edges correspond to the interquartile range.

Cohen's Kappa coefficient measures the agreement between experts, where '1' indicates perfect agreement between experts and '0' indicates agreement no better than chance. Figure 13 reveals levels of agreement for each assessment category among the different pairs of experts for DCNNs-based MoE and ViT-based MoE. Vertical dotted lines present the average consensus across three assessment categories for each pair (in

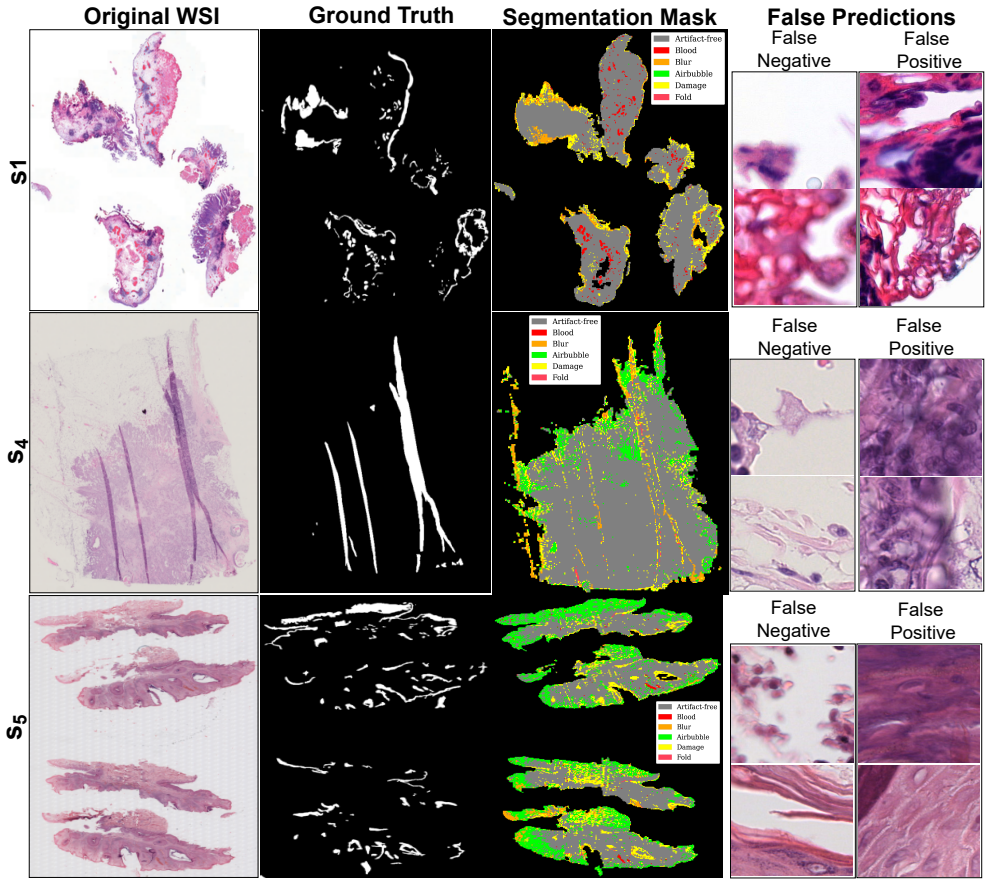

Fig. 9: **Visualization of ViTs-based mixture of experts' predictions with better performance over out-of-distribution data.** Image shows original WSIs (s2,s3, and s6) along with ground truth for artifacts (combined), artifact segmentation map, and a few examples of false predictions. False negative refers to patches detected as artifacts but were artifact-free, and false positive refers to patches detected as artifact-free but belonged to any artifact class.

corresponding color). Both subplots highlight substantial agreement for overall usability and high average agreement between P1 and P2 (red dashed line) in Figure 13. In contrast, artifact-free preservation has relatively lower agreement, echoing similar findings across all pairs. Based on the remarks obtained from field experts (see Figure 12), generally, better results were obtained for bladder cancer WSIs (s_1,s_2). Although MoEs were too sensitive in detecting blurry areas, their folded and damaged regions were well segmented. In breast cancer WSIs (s_3,s_4), adipose tissue was predicted as air bubbles (with DCNNs-based MoE) or damaged (with ViTs-based MoE). Note that the training data did not include adipose tissue, primarily fat cells. This situation can be more evident in breast samples because there is more adipose tissue in

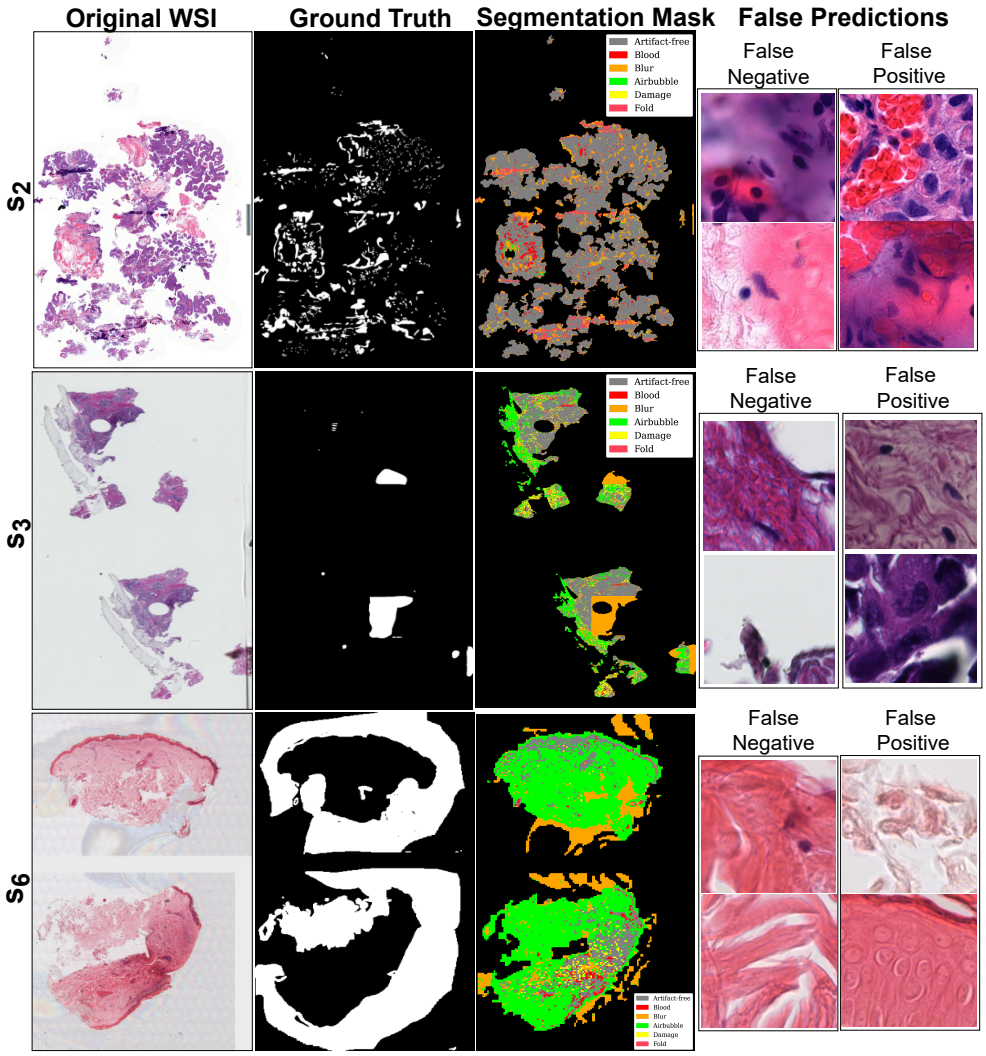

Fig. 10: **Visualization of DCNNs-based mixture of experts' predictions with worst performance over out-of-distribution data.** Image shows original WSIs (s2,s3, and s6) along with ground truth for artifacts (combined), artifact segmentation map, and a few examples of false predictions. False negative refers to patches detected as artifacts but were artifact-free, and false positive refers to patches detected as artifact-free but belonged to any artifact class.

them than in other cancer types. While adipose tissue can provide valuable contextual information and aid in certain aspects of diagnosis, its absence does not necessarily preclude accurate assessment of breast cancer.

The particular examples of skin cancer WSIS (s_5,s_6) had significant air bubbles, leaving a hazy and unclear appearance over the foreground tissue. At the same time,

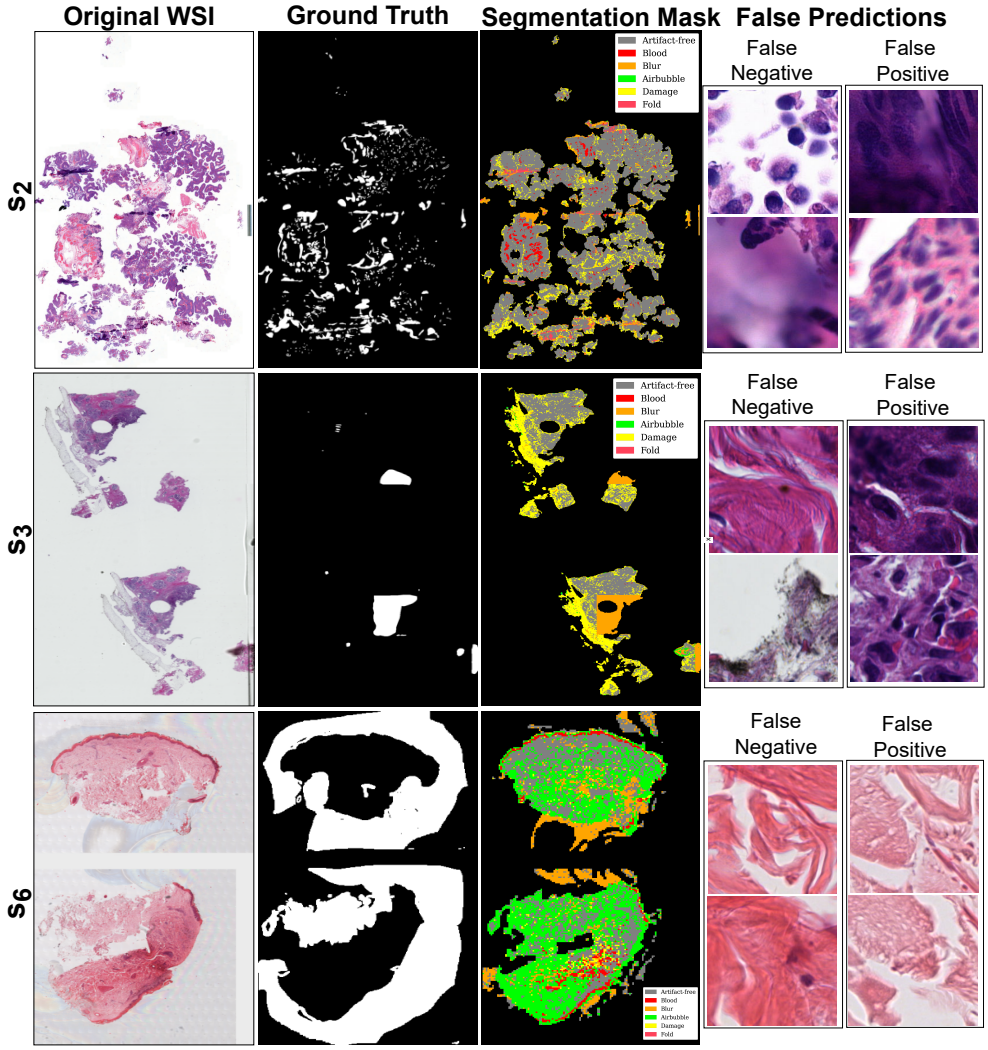

Fig. 11: **Visualization of ViTs-based mixture of experts' predictions with worst performance over out-of-distribution data.** Image shows original WSIs (s2, s3, and s6) along with ground truth for artifacts (combined), artifact segmentation map, and a few examples of false predictions. False negative refers to patches detected as artifacts but were artifact-free, and false positive refers to patches detected as artifact-free but belonged to any artifact class.

both artifact processing pipelines were overdoing air bubble prediction, and the epidermis was predicted as blood. The performance of both MoEs is worst in these cases; one of the reasons could be the severity of artifacts and significant variation in staining in the WSI. While there is generally substantial agreement among field experts for overall diagnostic usability, there are areas, such as artifact-free preservation, where

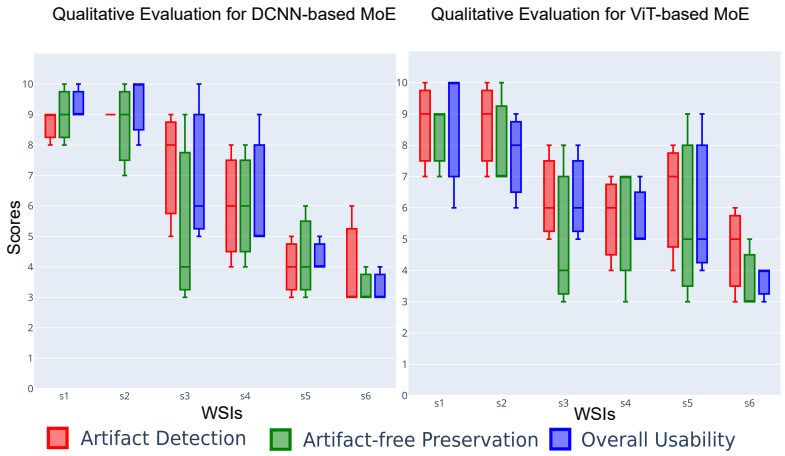

Fig. 12: **Scores for qualitative evaluation by field experts (P1, P2 and P3) for different Tasks.** The boxplot provides a visual representation of the experts' assessments for predictions of OoD WSIs. The scores were provided on a scale of 1 to 10, with higher scores indicating better performance.

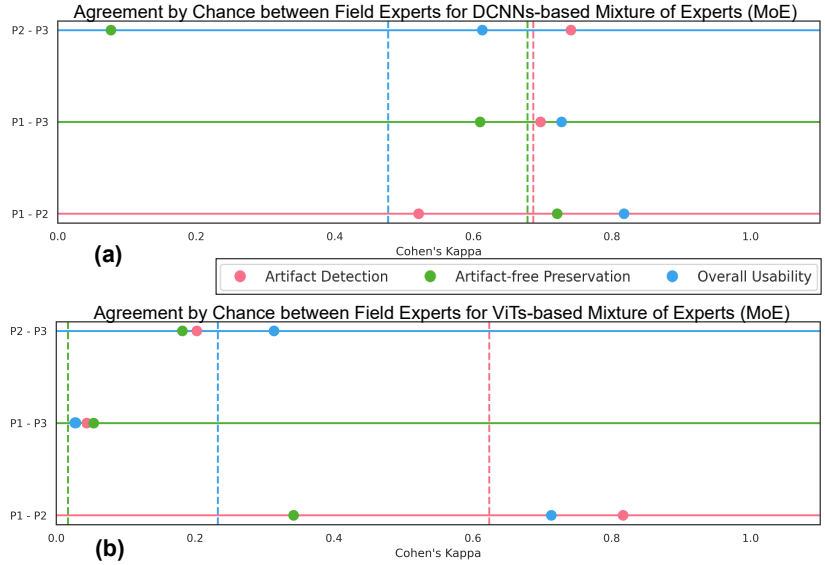

Fig. 13: **Qualitative evaluation of artifact detection by the mixture of experts (MoE) models over OoD data.** Plots (a) represent Cohen's kappa score (on the x-axis) for DCNNs-based MoE and a pair of field experts on the y-axis, and (b) show scores for ViTs-based MoE. Both subplots show agreement by chance for each task. Each pair's average agreement of all three tasks is plotted as a vertical dashed line.

discrepancies emerge and may be more challenging to achieve. Moreover, considering inter-rater variability, DCNNs-based MoE indicates potential effectiveness for artifact detection and overall diagnostic usability.

By triangulating quantitative and qualitative analysis findings, we conclude that DCNNs-based MoE provides better generalizability and robustness with the trade-off of higher computational cost.

6 Conclusion and Future Directions

In this work, we established end-to-end deep learning (DL) pipelines, taking whole slide images (WSIs) as input and providing artifact-refined WSIs to enable computational pathology (CPATH) systems to make reliable predictions. For the development of DL pipelines, we propose a mixture of experts (MoE) scheme and multiclass models. The MoE scheme uses five base learners (experts) with underlying state-of-the-art DL architectures (MobileNetv3 or ViT-Tiny). The MoE captures the intricacies of different artifact morphologies and dynamically combines predictions using a fusion mechanism to generate predictive probability distribution. Later, a meta-learned probabilistic threshold is applied to improve sensitivity for histologically relevant regions. In rigorous experiments, we performed generalizability and robustness tests over DL pipelines by testing on external cohorts of different tissue types. During the investigation, we found that the MoE scheme with underlying DCNNs attains the best classification and segmentation performance with some computational trade-offs compared to multiclass models. However, if high inference speed is the desired requirement, then multiclass models are a better choice with some degree of performance trade-off. Furthermore, during the qualitative evaluation, field experts rated the outcomes and achieved a substantial agreement for the overall usability of DCNNs-based MoE.

Our artifact processing DL pipelines can provide various outcomes, such as a segmentation map, artifact report, artifact-free mask with potential region of interest with the histologically relevance, and artifact-refined WSI for further computational analysis. Overall, the proposed DL solution is efficient and has a great advantage in equipping the CPATH system with the necessary tools to isolate anomalies (or noise) from affecting automated clinical applications.

The proposed work has a limitation in that the DL models were trained on a dataset prepared from a single cohort of data. In future work, we will overcome these limitations by pooling datasets from different cohorts in training and adopting an active learning strategy to adapt meta-learned thresholding parameters for improved sensitivity. Also, by adopting tailored fusion mechanisms for different cancer types. Moreover, artifact-refined WSIs can be tested with the corresponding diagnostic or prognostic algorithms to assess the usefulness of artifact processing pipelines for clinical practice.

Data and Code Availability

The code is available at Github. The training and development dataset can be downloaded from Zenodo.

Abbreviations

WSI	Whole slide image
DL	Deep learning
CPATH	Computational pathology
MoE	Mixture of experts
SOTA	state-of-the-art
DCNN	Deep convolutional neural networks
ViT	Vision transformer
OoD	Out-of-distribution
DP	Digital pathology
QC	Quality control
RGB	Red, Green, Blue
HSI	Hue, Saturation, Intensity
SVM	Support vector machine
H&E	Hematoxylin and Eosin
EMC	Erasmus medical centre
SUH	Stavanger University Hospital

Compliance with Ethical Standards

This study was performed in line with the principles of the Declaration of Helsinki. The patients/participants provided written informed consent to use their data for secondary purposes. The Erasmus MC Medical Research Committee granted approval from the Institutional Review Board under the reference MEC-2018-1097. The Stavanger University Hospital's data is approved by the Norwegian Regional Committee for Medical and Health Research Ethics under REC, 2010/1241. INCLIVA Biomedical Research Institute granted approval from the Research Ethics Committee (CEIm) of the Hospital Clínico Universitario of Valencia, Spain, under the reference CEIm-2020/114.

Declaration of competing interest

This research work is objective and unbiased. The authors have no relevant financial or non-financial interests to disclose.

CRediT authorship contribution statement

N.K: Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing - Original Draft, Writing - Review & Editing, Visualization; **F.K:** Formal analysis; Investigation; Data Resources; Writing - Review & Editing; **U.K:** Formal analysis; Investigation; Data Resources; Writing-Review & Editing; **A.M:** Formal analysis; Investigation; Data Resources; Writing - Review & Editing; **C.M:** Formal analysis; Data Resources; Writing - Review & Editing; Funding acquisition; **E.J:** Formal analysis; Data Resources; Writing - Review & Editing; Funding acquisition; **T.Z:** Formal analysis; Data Resources; Writing - Review & Editing; Funding acquisition;

C.R: Writing - Review & Editing; **K.E:** Conceptualization; Methodology; Investigation; Supervision; Writing - Review & Editing; Project administration; Funding acquisition.

Acknowledgments

The European Union's Horizon 2020 research and innovation program (CLARIFY) financially supports this research work under the Marie Skłodowska-Curie grant agreement 860627.

References

[1] National Cancer Institute: Environmental Carcinogens and Cancer Risk. https://www.cancer.gov/about-cancer/causes-prevention/risk/substances/carcinogens. Accessed on August 31, 2023 (2015)

[2] World Cancer Research Fund International: Differences in cancer incidence and mortality across the globe. https://www.wcrf.org/differences-in-cancer-incidence-and-mortality-across-the-globe/. Accessed on August 31, 2023 (2023)

[3] Pulumati, A., Pulumati, A., Dwarakanath, B.S., Verma, A., Papineni, R.V.: Technological advancements in cancer diagnostics: Improvements and limitations. Cancer Reports **6**(2), 1764 (2023)

[4] Khened, M., Kori, A., Rajkumar, H., Krishnamurthi, G., Srinivasan, B.: A generalized deep learning framework for whole-slide image segmentation and analysis. Scientific reports **11**(1), 11579 (2021)

[5] Zhu, C., Song, F., Wang, Y., Dong, H., Guo, Y., Liu, J.: Breast cancer histopathology image classification through assembling multiple compact cnns. BMC medical informatics and decision making **19**(1), 1–17 (2019)

[6] Kanwal, N., Amundsen, R., Hardardottir, H., Janssen, E.A., Engan, K.: Detection and localization of melanoma skin cancer in histopathological whole slide images. In: 2023 31st European Signal Processing Conference (EUSIPCO), pp. 1128–1135 (2023). IEEE

[7] Car, L.T., Papachristou, N., Bull, A., Majeed, A., Gallagher, J., El-Khatib, M., Aylin, P., Rudan, I., Atun, R., Car, J., *et al.*: Clinician-identified problems and solutions for delayed diagnosis in primary care: a prioritize study. BMC family practice **17**, 1–9 (2016)

[8] Pallua, J., Brunner, A., Zelger, B., Schirmer, M., Haybaeck, J.: The future of pathology is digital. Pathology-Research and Practice **216**(9), 153040 (2020)

[9] Inc, D.S..R.S.: Digital Science and Research Solutions Inc. https://app.dimensions.ai/analytics/publication/overview/timeline?search_mode=content&or_f

acet_year=2018&or_facet_year=2019&or_facet_year=2020&or_facet_year=2021&or_facet_year=2022&or_facet_year=2023&search_text=Digital%20Pathology&search_type=kws&search_field=full_search. Query: "CPATH" OR "Computational Pathology" OR "Digital Pathology" (accessed: August 2023)

[10] Kanwal, N., Pérez-Bueno, F., Schmidt, A., Molina, R., Engan, K.: The devil is in the details: Whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation. a review. IEEE Access (2022)

[11] Campanella, G., Rajanna, A.R., Corsale, L., Schüffler, P.J., Yagi, Y., Fuchs, T.J.: Towards machine learned quality control: A benchmark for sharpness quantification in digital pathology. Computerized Medical Imaging and Graphics **65**, 142–151 (2018)

[12] Hosseini, M.S., Bejnordi, B.E., Trinh, V.Q.-H., Chan, L., Hasan, D., Li, X., Yang, S., Kim, T., Zhang, H., Wu, T., et al.: Computational pathology: a survey review and the way forward. Journal of Pathology Informatics, 100357 (2024)

[13] Louis, D.N., Gerber, G.K., Baron, J.M., Bry, L., Dighe, A.S., Getz, G., Higgins, J.M., Kuo, F.C., Lane, W.J., Michaelson, J.S., *et al.*: Computational pathology: an emerging definition. Archives of pathology & laboratory medicine **138**(9), 1133–1138 (2014)

[14] Taqi, S.A., Sami, S.A., Sami, L.B., Zaki, S.A.: A review of artifacts in histopathology. Journal of oral and maxillofacial pathology: JOMFP **22**(2), 279 (2018)

[15] Bindhu, P., Krishnapillai, R., Thomas, P., Jayanthi, P.: Facts in artifacts. Journal of oral and maxillofacial pathology: JOMFP **17**(3), 397 (2013)

[16] Kanwal, N., Eftestøl, T., Khoraminia, F., Zuiverloon, T.C., Engan, K.: Vision transformers for small histological datasets learned through knowledge distillation. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 167–179 (2023). Springer

[17] Wright, A.I., Dunn, C.M., Hale, M., Hutchins, G.G., Treanor, D.E.: The effect of quality control on accuracy of digital pathology image analysis. IEEE Journal of Biomedical and Health Informatics **25**(2), 307–314 (2020)

[18] Tabatabaei, Z., Colomer, A., Engan, K., Oliver, J., Naranjo, V.: Residual block convolutional auto encoder in content-based medical image retrieval. In: 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), pp. 1–5 (2022). IEEE

[19] Chen, C., Chen, C., Ma, M., Ma, X., Lv, X., Dong, X., Yan, Z., Zhu, M., Chen, J.: Classification of multi-differentiated liver cancer pathological images based on deep learning attention mechanism. BMC Medical Informatics and Decision Making **22**(1), 1–13 (2022)

[20] Fuster, S., Khoraminia, F., Kiraz, U., Kanwal, N., Kvikstad, V., Eftestøl, T., Zuiverloon, T.C.M., Janssen, E.A.M., Engan, K.: Invasive cancerous area detection in non-muscle invasive bladder cancer whole slide images. In: 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), pp. 1–5 (2022)

[21] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)

[22] Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., Miao, Y.: Review of image classification algorithms based on convolutional neural networks. Remote Sensing **13**(22), 4712 (2021)

[23] Lu, Z., Xie, H., Liu, C., Zhang, Y.: Bridging the gap between vision transformers and convolutional neural networks on small datasets. Advances in Neural Information Processing Systems **35**, 14663–14677 (2022)

[24] Zhu, H., Chen, B., Yang, C.: Understanding why vit trains badly on small datasets: An intuitive perspective. arXiv preprint arXiv:2302.03751 (2023)

[25] Atabansi, C.C., Nie, J., Liu, H., Song, Q., Yan, L., Zhou, X.: A survey of transformer applications for histopathological image analysis: New developments and future directions. BioMedical Engineering OnLine **22**(1), 96 (2023)

[26] Naseer, M.M., Ranasinghe, K., Khan, S.H., Hayat, M., Shahbaz Khan, F., Yang, M.-H.: Intriguing properties of vision transformers. Advances in Neural Information Processing Systems **34**, 23296–23308 (2021)

[27] Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., Veit, A.: Understanding robustness of transformers for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10231–10241 (2021)

[28] Hsu, S.-T., Su, Y.-J., Hung, C.-H., Chen, M.-J., Lu, C.-H., Kuo, C.-E.: Automatic ovarian tumors recognition system based on ensemble convolutional neural network with ultrasound imaging. BMC Medical Informatics and Decision Making **22**(1), 298 (2022)

[29] Meng, Z., Zhao, Z., Li, B., Su, F., Guo, L.: A cervical histopathology dataset for computer aided diagnosis of precancerous lesions. IEEE Transactions on Medical Imaging **40**(6), 1531–1541 (2021)

[30] Abe, T., Buchanan, E.K., Pleiss, G., Zemel, R., Cunningham, J.P.: Deep ensembles work, but are they necessary? Advances in Neural Information Processing Systems **35**, 33646–33660 (2022)

[31] Mohammed, A., Kora, R.: A comprehensive review on ensemble deep learning: Opportunities and challenges. Journal of King Saud University-Computer and Information Sciences (2023)

[32] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., *et al.*: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314–1324 (2019)

[33] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357 (2021). PMLR

[34] Morales, S., Engan, K., Naranjo, V.: Artificial intelligence in computational pathology–challenges and future directions. Digital Signal Processing **119**, 103196 (2021)

[35] Bulten, W., Kartasalo, K., Chen, P.-H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., Boven, H., Vink, R., *et al.*: Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. Nature medicine **28**(1), 154–163 (2022)

[36] Khoraminia, F., Fuster, S., Kanwal, N., Olislagers, M., Engan, K., Leenders, G.J., Stubbs, A.P., Akram, F., Zuiverloon, T.C.: Artificial intelligence in digital pathology for bladder cancer: Hype or hope? a systematic review. Cancers **15**(18), 4518 (2023)

[37] Gay, J., Harlin, H., Wetzer, E., Lindblad, J., Sladoje, N.: Texture-based oral cancer detection: A performance analysis of deep learning approaches. In: 3rd NEUBIAS Conference (2019)

[38] Gandomkar, Z., Brennan, P.C., Mello-Thoms, C.: Mudern: Multi-category classification of breast histopathological image using deep residual networks. Artificial Intelligence in Medicine **88**, 14–24 (2018) https://doi.org/10.1016/j.artmed.2018.04.005

[39] Wessels, F., Schmitt, M., Krieghoff-Henning, E., Nientiedt, M., Waldbillig, F., Neuberger, M., Kriegmair, M.C., Kowalewski, K.-F., Worst, T.S., Steeg, M., *et al.*: A self-supervised vision transformer to predict survival from histopathology in renal cell carcinoma. World Journal of Urology **41**(8), 2233–2241 (2023)

[40] Stegmüller, T., Bozorgtabar, B., Spahr, A., Thiran, J.-P.: Scorenet: Learning non-uniform attention and augmentation for transformer-based histopathological image classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 6170–6179 (2023)

[41] Perincheri, S., Levi, A.W., Celli, R., Gershkovich, P., Rimm, D., Morrow, J.S.,

Rothrock, B., Raciti, P., Klimstra, D., Sinard, J.: An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy. Modern Pathology **34**(8), 1588–1595 (2021)

[42] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)

[43] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[44] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9 (2015)

[45] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9650–9660 (2021)

[46] Zidan, U., Gaber, M.M., Abdelsamea, M.M.: Swincup: Cascaded swin transformer for histopathological structures segmentation in colorectal cancer. Expert Systems with Applications **216**, 119452 (2023)

[47] Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: A survey. Medical Image Analysis, 101813 (2020)

[48] Riasatian, A., Babaie, M., Maleki, D., Kalra, S., Valipour, M., Hemati, S., Zaveri, M., Safarpoor, A., Shafiei, S., Afshari, M., Rasoolijaberi, M., Sikaroudi, M., Adnan, M., Shah, S., Choi, C., Damaskinos, S., Campbell, C.J., Diamandis, P., Pantanowitz, L., Kashani, H., Ghodsi, A., Tizhoosh, H.R.: Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. Medical Image Analysis **70**, 102032 (2021)

[49] Talo, M.: Automated classification of histopathology images using transfer learning. Artificial Intelligence in Medicine **101**, 101743 (2019)

[50] Wang, Y., Peng, T., Duan, J., Zhu, C., Liu, J., Ye, J., Jin, M.: Pathological image classification based on hard example guided cnn. IEEE Access **8**, 114249–114258 (2020)

[51] Wang, C., Gong, W., Cheng, J., Qian, Y.: Dblcnn: Dependency-based lightweight convolutional neural network for multi-classification of breast histopathology images. Biomedical Signal Processing and Control **73**, 103451 (2022)

[52] Gao, Z., Hong, B., Zhang, X., Li, Y., Jia, C., Wu, J., Wang, C., Meng, D., Li, C.: Instance-based vision transformer for subtyping of papillary renal cell carcinoma in histopathological image. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24, pp. 299–308 (2021). Springer

[53] Schömig-Markiefka, B., Pryalukhin, A., Hulla, W., Bychkov, A., Fukuoka, J., Madabhushi, A., Achter, V., Nieroda, L., Büttner, R., Quaas, A., *et al.*: Quality control stress test for deep learning-based diagnostic model in digital pathology. Modern Pathology **34**(12), 2098–2108 (2021)

[54] Linmans, J., Raya, G., Laak, J., Litjens, G.: Diffusion models for out-of-distribution detection in digital pathology. Medical Image Analysis **93**, 103088 (2024)

[55] Ghaffari Laleh, N., Truhn, D., Veldhuizen, G.P., Han, T., Treeck, M., Buelow, R.D., Langer, R., Dislich, B., Boor, P., Schulz, V., *et al.*: Adversarial attacks and adversarial robustness in computational pathology. Nature communications **13**(1), 5711 (2022)

[56] Kanwal, N., Engan, K.: Extract, detect, eliminate: Enhancing reliability and performance of computational pathology through artifact processing pipelines. Science Talks (2024)

[57] Kothari, S., Phan, J.H., Wang, M.D.: Eliminating tissue-fold artifacts in histopathological whole-slide images for improved image-based prediction of cancer grade. Journal of pathology informatics **4**(1), 22 (2013)

[58] Kanwal, N., López-Pérez, M., Kiraz, U., Zuiverloon, T.C., Molina, R., Engan, K.: Are you sure it's an artifact? artifact detection and uncertainty quantification in histological images. Computerized Medical Imaging and Graphics **112**, 102321 (2024)

[59] Salvi, M., Acharya, U.R., Molinari, F., Meiburger, K.M.: The impact of pre-and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis. Computers in Biology and Medicine **128**, 104129 (2021)

[60] Pérez-Bueno, F., Vega, M., Naranjo, V., Molina, R., Katsaggelos, A.K.: Super gaussian priors for blind color deconvolution of histological images. In: 2020 IEEE International Conference on Image Processing (ICIP), pp. 3010–3014 (2020). IEEE

[61] Ameisen, D., Deroulers, C., Perrier, V., Bouhidel, F., Battistella, M., Legrès, L., Janin, A., Bertheau, P., Yunès, J.B.: Towards better digital pathology workflows: Programming libraries for high-speed sharpness assessment of Whole Slide

Images. Diagnostic Pathology **9**(1), 1–7 (2014)

[62] Shrestha, P., Kneepkens, R., Vrijnsen, J., Vossen, D., Abels, E., Hulsken, B.: A quantitative approach to evaluate image quality of whole slide imaging scanners. Journal of pathology informatics **7** (2016)

[63] Bahlmann, C., Patel, A., Johnson, J., Ni, J., Chekkoury, A., Khurd, P., Kamen, A., Grady, L., Krupinski, E., Graham, A., *et al.*: Automated detection of diagnostically relevant regions in h&e stained digital pathology slides. In: Medical Imaging 2012: Computer-Aided Diagnosis, vol. 8315, p. 831504 (2012). International Society for Optics and Photonics

[64] Avanaki, A.R.N., Espig, K.S., Xthona, A., Lanciault, C., Kimpe, T.R.L.: Automatic image quality assessment for digital pathology. In: Tingberg, A., Lång, K., Timberg, P. (eds.) Breast Imaging, pp. 431–438. Springer, Cham (2016)

[65] Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M., Madabhushi, A.: Histoqc: an open-source quality control tool for digital pathology slides. JCO clinical cancer informatics **3**, 1–7 (2019)

[66] Gao, D., Padfield, D., Rittscher, J., McKay, R.: Automated training data generation for microscopy focus classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 446–453 (2010). Springer

[67] Hashimoto, N., Bautista, P.A., Yamaguchi, M., Ohyama, N., Yagi, Y.: Referenceless image quality evaluation for whole slide imaging. Journal of pathology informatics **3** (2012)

[68] Palokangas, S., Selinummi, J., Yli-Harja, O.: Segmentation of folds in tissue section images. In: 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 5641–5644 (2007). IEEE

[69] Bautista, P.A., Yagi, Y.: Detection of tissue folds in whole slide images. Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009, 3669–3672 (2009)

[70] Swiderska-Chadaj, Z., Markiewicz, T., Cierniak, S., Koktysz, R.: Automatic quantification of vessels in hemorrhoids whole slide images. In: 2016 17th International Conference Computational Problems of Electrical Engineering (CPEE), pp. 1–4 (2016). IEEE

[71] Mercan, E., Aksoy, S., Shapiro, L.G., Weaver, D.L., Brunye, T., Elmore, J.G.: Localization of diagnostically relevant regions of interest in whole slide images. In: 2014 22nd International Conference on Pattern Recognition, pp. 1179–1184 (2014). IEEE

[72] Albuquerque, T., Moreira, A., Cardoso, J.S.: Deep ordinal focus assessment for whole slide images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 657–663 (2021)

[73] Kohlberger, T., Liu, Y., Moran, M., Chen, P.-H.C., Brown, T., Hipp, J.D., Mermel, C.H., Stumpe, M.C.: Whole-slide image focus quality: Automatic assessment and impact on ai cancer detection. Journal of pathology informatics **10** (2019)

[74] Wetteland, R., Engan, K., Eftestøl, T., Kvikstad, V., Janssen, E.A.M.: Multiclass tissue classification of whole-slide histological images using convolutional neural networks. ICPRAM 2019 - Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods, 320–327 (2019)

[75] Wetteland, R., Engan, K., Eftestøl, T., Kvikstad, V., Janssen, E.A.: A multiscale approach for whole-slide image segmentation of five tissue classes in urothelial carcinoma slides. Technology in Cancer Research & Treatment **19**, 1533033820946787 (2020)

[76] Clymer, D., Kostadinov, S., Catov, J., Skvarca, L., Pantanowitz, L., Cagan, J., LeDuc, P.: Decidual vasculopathy identification in whole slide images using multiresolution hierarchical convolutional neural networks. The American Journal of Pathology **190**(10), 2111–2122 (2020)

[77] Babaie, M., Tizhoosh, H.R.: Deep features for tissue-fold detection in histopathology images. In: European Congress on Digital Pathology, pp. 125–132 (2019). Springer

[78] Kanwal, N., Fuster, S., Khoraminia, F., Zuiverloon, T.C., Rong, C., Engan, K.: Quantifying the effect of color processing on blood and damaged tissue detection in whole slide images. In: 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), pp. 1–5 (2022). IEEE

[79] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330 (2017). PMLR

[80] Linmans, J., Elfwing, S., Laak, J., Litjens, G.: Predictive uncertainty estimation for out-of-distribution detection in digital pathology. Medical Image Analysis **83**, 102655 (2023)

[81] Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. Communications of the ACM **64**(3), 107–115 (2021)

[82] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint

arXiv:2010.11929 (2020)

[83] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE ICCV, pp. 248–255 (2009). Ieee

[84] Wetzer, E.: Representation learning and information fusion: Applications in biomedical image processing. PhD thesis, Acta Universitatis Upsaliensis (2023)