Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Are you sure it's an artifact? Artifact detection and uncertainty quantification in histological images

Neel Kanwal [a,*], Miguel López-Pérez [b], Umay Kiraz [d,e], Tahlita C.M. Zuiverloon [c], Rafael Molina [b], Kjersti Engan [a]

[a] Department of Electrical Engineering and Computer Science, University of Stavanger, 4021 Stavanger, Norway
[b] Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain
[c] Department of Urology, University Medical Center Rotterdam, Erasmus MC Cancer Institute, 1035 GD Rotterdam, The Netherlands
[d] Department of Pathology, Stavanger University Hospital, 4011 Stavanger, Norway
[e] Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, 4021 Stavanger, Norway

ARTICLE INFO

Keywords:
Artifact detection
Computational pathology
Deep learning
Gaussian processes
Whole Slide Images

ABSTRACT

Modern cancer diagnostics involves extracting tissue specimens from suspicious areas and conducting histotechnical procedures to prepare a digitized glass slide, called Whole Slide Image (WSI), for further examination. These procedures frequently introduce different types of artifacts in the obtained WSI, and histological artifacts might influence Computational Pathology (CPATH) systems further down to a diagnostic pipeline if not excluded or handled. Deep Convolutional Neural Networks (DCNNs) have achieved promising results for the detection of some WSI artifacts, however, they do not incorporate uncertainty in their predictions. This paper proposes an uncertainty-aware Deep Kernel Learning (DKL) model to detect blurry areas and folded tissues, two types of artifacts that can appear in WSIs. The proposed probabilistic model combines a CNN feature extractor and a sparse Gaussian Processes (GPs) classifier, which improves the performance of current state-of-the-art artifact detection DCNNs and provides uncertainty estimates. We achieved 0.996 and 0.938 F1 scores for blur and folded tissue detection on unseen data, respectively. In extensive experiments, we validated the DKL model on unseen data from external independent cohorts with different staining and tissue types, where it outperformed DCNNs. Interestingly, the DKL model is more confident in the correct predictions and less in the wrong ones. The proposed DKL model can be integrated into the preprocessing pipeline of CPATH systems to provide reliable predictions and possibly serve as a quality control tool.

1. Introduction

Cancer is one of the leading causes of death worldwide, with nearly 0.6 million estimated deaths and 1.9 million new cases diagnosed in 2022 in the United States alone (Siegel et al., 2022). Cancer develops when normal cells undergo genetic modifications that cause them to convert into tumor cells. This transition is frequently triggered by exposure to carcinogens, which are agents (chemical, biological, or physical) capable of causing cancer (National Cancer Institute, 2015). Histopathology is a gold standard in cancer diagnosis where pathologists conduct a microscopic examination of tissue samples mounted on a glass slide to identify malignancy by evaluating cellular characteristics and tissue morphology and possibly assigning a cancer grade or stage to the patient (Morales et al., 2021; Tabatabaei et al., 2022). Such glass slides are acquired through a series of histological laboratory procedures, including steps like dehydration, fixation, clearing, embedding, sectioning, mounting, and staining. The glass slide is scanned in modern digital pathology, forming a Whole Slide Image (WSI). This entire histotechnical procedure is manual to a large degree. During a single or combination of these steps, imperfect handling of the tissue specimen may result in the introduction of artifacts on the obtained WSI (Rastogi et al., 2013; Taqi et al., 2018; Rolls et al., 2008). Artifacts are areas with no diagnostic relevance because the tissue is altered or damaged in its appearance.

Some artifacts may appear on the tissue specimen due to complications arising in the biopsy procedure of specific organs (such as blood and damaged tissue in trans-urethral resection of the bladder tumor or burnt areas due to cauterization). Other artifacts appear during the preparation of the WSI. These artifacts include folded tissues due to imperfections in placing the section during the mounting process, air bubbles (air trapped under the coverslip) from mounting, or blur introduced in the scanning process. A more comprehensive description

[*] Corresponding author.
E-mail address: neel.kanwal@uis.no (N. Kanwal).

https://doi.org/10.1016/j.compmedimag.2023.102321
Received 10 July 2023; Received in revised form 8 November 2023; Accepted 12 December 2023
Available online 20 December 2023

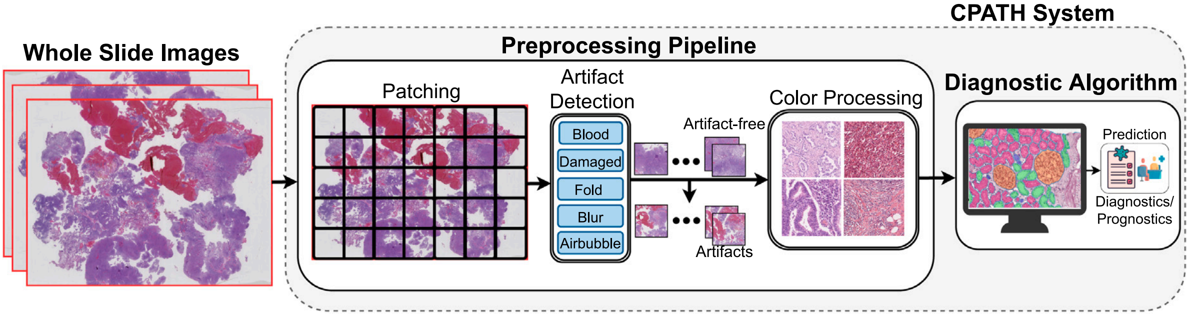


Fig. 1. A patch-based Computational Pathology (CPATH) system with a preprocessing pipeline to tackle artifacts in Whole Slide Images (WSIs). Artifact detection models ensure the flow of only histologically relevant patches to the diagnostic algorithm for a reliable prediction.

of these steps and the artifacts they can introduce is provided in Kanwal et al. (2022b). Pathologists are trained to focus on the diagnostically relevant parts of a glass slide or WSI and would usually ignore these areas during manual inspection (Kanwal et al., 2022a; Bindhu et al., 2013; Kanwal et al., 2023b). Blurry regions and folded tissues are the most common artifacts in WSIs, prepared from specimens of most cancer biopsies (Palokangas et al., 2007; Babaie and Tizhoosh, 2019; Salvi et al., 2021), and we focus only on these two artifacts in this paper. Blur results from unaligned focus (i.e., scanning profile) during the scanning of glass slides and may cause a complete loss of visual features in regions of the WSI (Kanwal et al., 2022b; Janowczyk et al., 2019). Similarly, folds (i.e., folded tissues) appear due to the placement of tissue over itself during the mounting stage, undermining cellular visibility with a thicker appearance, making it irrelevant during diagnostic inspection (Kanwal et al., 2022b; Bancroft and Gamble, 2008).

A Computational Pathology (CPATH) system is an automated system for diagnostics, prognostics, or segmentation and visualization of WSI, usually built on image processing and artificial intelligence. Most CPATH tasks lack large enough datasets for training and rely on manually annotated diagnostically relevant regions (Kanwal et al., 2023). The performance of CPATH systems deteriorate with the presence of histological artifacts because they represent noise and clinically irrelevant areas. Thus, removing irrelevant regions from the data (i.e., noise) can improve prediction quality (Kanwal et al., 2022b,a; Wright et al., 2020). A stress-testing study (Schömig-Markiefka et al., 2021) on prostate cancer shows how the presence of artifacts results in false positive predictions, leading to a substantial loss in diagnostic accuracy in Deep Learning (DL) models. Therefore, it is beneficial to automatically detect and exclude artifacts in a preprocessing pipeline before running a DL-based diagnostic algorithm, as illustrated in Fig. 1.

Deep Convolutional Neural Networks (DCNNs) are popular choices for biomedical image analysis in CPATH systems (Ho et al., 2021; Kanwal et al., 2023a; Tomasetti et al., 2020). They have demonstrated their success in numerous medical imaging challenges (Kanwal et al., 2022a; Bulten et al., 2022; Del Amor et al., 2023). Although DCNNs provide high accuracy, they often suffer from over-fitting and over-confident predictions due to their deterministic nature (Nguyen et al., 2015; William F et al., 2021). DCNNs provide a single-point estimate and do not model confidence in their predictions. However, confidence in predictions is advantageous in medical applications where decision-making involves human life (Abdar et al., 2021). In contrast to DCNNs, probabilistic classifiers provide a distribution, quantifying the certainty of the classifier on the predicted distribution. These probabilistic models also work well with reduced datasets (Wu et al., 2021b). One of the most popular probabilistic models is Gaussian Processes (GPs) (Williams and Rasmussen, 2006). GPs learn distributions over functions and provide a confidence measure of their prediction. This confidence is extracted from its predictive variance. A higher predictive variance implies a wider distribution and depicts weak confidence (and vice versa). This fact makes GPs trustworthy, and for this reason, they are getting increased attention for CPATH tasks (Kandemir, 2015; Haußmann et al., 2017; Esteban et al., 2019). However, GPs alone cannot handle images directly and rely on a prior feature extraction step (López-Pérez et al., 2022).

Deep probabilistic neural architectures combine the advantages of DCNNs and probabilistic models. The most extended approach is Monte Carlo (MC) dropout during the inference, which performs approximate Bayesian inference on DCNNs (Gal and Ghahramani, 2016). However, the MC dropout approach is not computationally efficient since it requires multiple runs. Recently, the combination of deep parametric and non-parametric approaches has attracted increasing attention. Deep Kernel Learning (DKL) consolidates the feature extraction power of CNNs with the modeling capacity of GPs to learn a distribution from the feature space (Wilson et al., 2016a). This combination gives the DKL model good classification performance and uncertainty modeling, making it interesting for the medical imaging community (Wu et al., 2021c).

The topic of artifact detection in WSI has not been given very much attention in the literature. One reason might be that no publicly available datasets exist for most histological artifacts. To the best of our knowledge, an uncertainty-aware artifact detection model has not been proposed yet in the CPATH literature. Depending on the nature of the CPATH task, uncertainty estimation gives the option of only removing what the model is confident is an artifact or only using regions where the model is confident that it is not an artifact.

From these observations, we propose an uncertainty-aware DKL model that combines a custom DenseNet (Huang et al., 2017) feature extractor with a sparse GP classifier to detect histological artifacts, blur and folds, and provide uncertainty estimates for the predictions. Fig. 2 illustrates the pipeline of the proposed DKL model, which integrates DenseNet CNN and GP classifier to exploit the full advantage by training them in an end-to-end fashion. Throughout the paper, we use a *CNN* term for feature extractors and *DCNN* as an umbrella term for state-of-the-art (SOTA) CNN-based architectures with a fully connected (FC) classifier at the end. The main contributions of this paper are summarized below.

1. We propose the DKL model for detecting folded tissue and blurry areas against artifact-free tissue in WSIs. The prime benefit of the DKL model is its ability to provide uncertainty estimates along with the prediction. In short, this combination formulates a reliable approach that tells how confident the model is in the detected artifacts.
2. We perform extensive experiments to evaluate the proposed model against SOTA DCNN and the baseline (a custom variant of DenseNet CNN with FC classifier) models in terms of performance, robustness, and computational complexity. The models are trained, validated, and tested on a cohort from one lab; In addition, it is tested on different cohorts from other labs.

This paper is organized as follows: Section 2 provides an overview of works using DCNN and GP approaches in medical images, along with

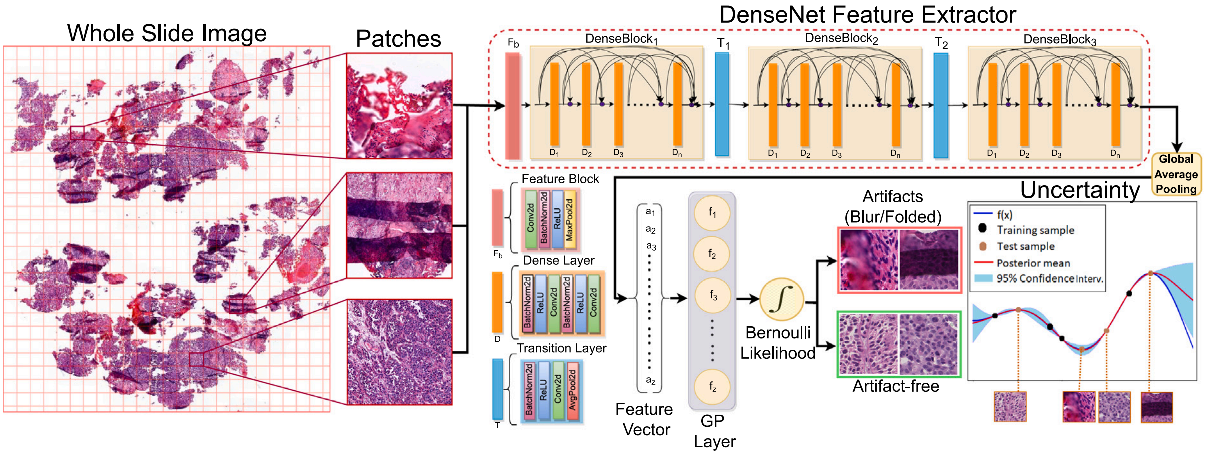


Fig. 2. An introductory overview of our proposed uncertainty-aware artifacts detection method (deep kernel learning model). The Whole Slide Image is split into patches of a predefined size. A DenseNet feature extractor with three blocks and varying dense layers was used to extract hidden representations. The global average pooling layer converts the obtained feature map to a vector. The Gaussian Process (GP) layer is applied to the feature vector to estimate a multivariate Gaussian distribution over the training data. Finally, the GP predicts the presence of artifacts in the patch using Bernoulli likelihood and calculates the predictive variance for the test samples. The sparse GP and the DenseNet CNN are trained jointly in an end-to-end fashion.

recent ones on artifact detection. Section 3 presents the method for training and inference of the proposed DKL model. Section 4 details the datasets, experimental setup, and evaluation metrics. Section 5 discusses the classification and uncertainty quantification results of the DKL and baseline models. Finally, Section 6 presents the conclusions of this paper and discusses future directions for preprocessing pipelines in artifact detection and quality control applications.

2. Related work

Deep learning (DL) approaches have demonstrated the most promising performance in medical imaging challenges (Bulten et al., 2022). Specifically, popular DL architectures, such as DenseNet (Huang et al., 2017), ResNet (He et al., 2016), GoogleNet (Szegedy et al., 2015) or MobileNet (Howard et al., 2017), DCNNs have been used in CPATH with SOTA performance (Kanwal et al., 2022a). Previously, several uncertainty quantification methods for DCNNs have been proposed (see reviews Abdar et al., 2021; Loftus et al., 2022). Unfortunately, uncertainty estimates obtained through these approximation methods may not be meaningful since DCNNs do not incorporate the flexibility of probabilistic modeling, such as GPs (Wu et al., 2021b).

2.1. Artifact detection

Recent works on WSIs using DL for diagnostic or prognostic prediction tasks have relied mainly on the manual selection of artifact-free Regions of Interest (ROIs) with diagnostic relevance (Priego-Torres et al., 2020; Urdal et al., 2017). However, manual selection is laborious and time-consuming. Since artifacts included during training or inference might deteriorate prediction results (Kanwal et al., 2023b); it is beneficial to equip CPATH systems with tools to deal with the presence of artifacts (Kanwal et al., 2023b, 2022b; Wetteland et al., 2021; Campanella et al., 2018).

Some publications exploited the stain absorption and texture features for finding diagnostically relevant regions (Mercan et al., 2014; Bahlmann et al., 2012). However, they did not take into account the artifacts appearing within diagnostically relevant regions. Artifact detection can be done before applying color normalization methods; thus, prior color processing may not be necessary (Kanwal et al., 2022a).

Blur is an artifact introduced in the scanning process and may affect downstream image features (Janowczyk et al., 2019; Wu et al., 2015). Gao et al. (2010) presented a method for detecting out-of-focus areas by handcrafting 44 extensive features (i.e., local statistics, wavelets, and contrast) to train an AdaBoost classifier. Hashimoto et al. (2012) utilized a linear combination of image noise and sharpness information to locate blur. Their non-reference method applied a regression model to evaluate the quality of mouse embryo WSIs. Wu et al. (2015) detected blurry patches in their workflow by training KNN, Naive Bayes, SVM, and random forest classifiers on the pixel-level distribution of local and global metrics. They found that selected local metrics outperformed global metrics for all four classifiers. Recent methods have relied on automatic feature extraction by CNNs. Campanella et al. (2018) in their framework evaluated the extent of blur using a random forest model on ResNet features. In a DCNN-based approach, Albuquerque et al. (2021) trained seven SOTA architectures on the FocusPath (Hosseini et al., 2019) dataset for the ordinal blur regression task and compared outcomes with knowledge-driven methods. DeepFocus (Senaras et al., 2018) used an analogous approach with CNN to analyze blurry regions in WSIs with different stains (i.e., CD10, CD21, Ki67, H&E). Other focus quality assessment frameworks like ConvFocus (Kohlberger et al., 2019) and FocusLiteNN (Wang et al., 2020) utilized CNN-based models to quantify and localize blurry areas in WSIs.

Folded tissue artifacts appear darker after staining due to the thickness of the additional tissue layer. The tissue layers on top of each other make it hard to assess the individual cells and features (Kanwal et al., 2022b). Previous works for tissue fold detection mostly relied on image enhancement and thresholding methodologies (Palokangas et al., 2007; Kothari et al., 2013; Bautista and Yagi, 2009). Palokangas et al. (2007) used color-space transformation to convert Red, Green, and Blue (RGB) images to Hue, Saturation, Intensity (HSI) color-space in their multistage method. First, they obtained the difference between saturation and intensity channels and then used the K-means clustering algorithm to separate folded tissue pixels from non-folded ones. In a similar approach, Bautista and Yagi (2009) used color information with a fixed threshold at the difference between saturation and luminance values of each pixel. Pixels with a difference higher than the threshold were used to apply adaptive RGB shifting to enhance the color structure in the folded areas. Unfortunately, using a fixed threshold may not be effective in histological images due to stain variations, especially between different cohorts (from different labs). To account for stain variations, Kothari et al. (2013) proposed a rank-sum method with neighborhood criterion to find image features and connectivity descriptors in low magnification images. Their method used two adaptive thresholds based on the differences in saturation and intensity ranges. Working with a similar observation that folded tissues contain high-saturation and low-intensity values, Shakhawat et al. (2020) explored Gray-level Co-occurrence Matrix (GLCM) based features to train a binary SVM classifier. Unfortunately, folded tissue detection methods based on colorimetric and texture features may not be reliable due to variations in staining and tissue type. To overcome this limitation, Babaie and Tizhoosh (2019) extracted deep features using well-known CNN architectures and trained SVM, KNN, and decision tree classifiers for folded tissue detection. Their method was developed and tested on a private dataset.

All the mentioned artifact detection works provide predictions as a single-point estimate. In high-risk scenarios, such as medical image analysis, reliable systems that can estimate the *uncertainty* in the predictions are highly desirable (Abdar et al., 2021; Olsson et al., 2022).

2.2. Gaussian processes in medical images

The capacity of GPs to handle and estimate uncertainty has led to their adoption in various biomedical image applications. Toledo-Cortés et al. (2020) proposed a hybrid GP with DL for diabetic retinopathy diagnosis. Their model combined the representational power of DL with the ability of GPs to learn from small datasets. They indicated that uncertainty quantification led to a more robust model, which improved the interpretability of the method. Wu et al. (2021a) combined an attention CNN with a GP for multiple instance learning in computerized tomography scans to detect intracranial hemorrhages. Their work demonstrated the improvement provided by adding GPs to the model. This approach was later extended to Deep Gaussian Processes (DGPs) in López-Pérez et al. (2022), confirming the promising performance of deep probabilistic models based on GPs.

GPs have been proposed for some CPATH tasks as well. Haußmann et al. (2017) proposed GPs for multiple instance learning to localize Barrett's cancer from tissue microarray histopathology images. Esteban et al. (2019) applied shallow and DGPs for prostate cancer detection in CPATH and compared them to VGG19 (Simonyan and Zisserman, 2015), Xception (Chollet, 2017) and Inception v3 (Szegedy et al., 2016). They found that incorporating morphological and texture features into GPs enabled them to achieve performance comparable to SOTA DCNNs. López-Pérez et al. (2021) proposed GPs to learn from crowds in histological breast cancer classification, using deep features from VGG16 (Simonyan and Zisserman, 2015) and labels provided by medical students. They showed that GPs with crowdsourced labels were competitive with DL approaches using expert labels. Later, this crowdsourcing model was extended to DGPs, highlighting the promising performance of deep probabilistic models (López-Pérez et al., 2023). Similarly, Kandemir (2015) proposed a model for asymmetric transfer learning based on DGPs.

The main drawback of these works is that they performed the classification in two separate steps: first, they trained the CNN for feature extraction, and second, they trained the GP with the obtained features. Hence, they did not take full advantage of both modules in an end-to-end way. Wu et al. (2021c) recently proposed a hybrid model using GP trained for a univariate regression task. Their method achieved better results than the baseline with a linear output layer, suggesting that end-to-end trained hybrid models can offer improved performance and uncertainty estimates compared to DCNNs using MC dropout (Gal and Ghahramani, 2016).

3. Method

This section provides an overview of the methodology. Section 3.1 reviews the DenseNet architecture used for feature extraction. Section 3.2 provides a brief introduction to GPs. Section 3.3 describes our proposed DKL model for detecting artifacts.

3.1. Feature extractor: DenseNet

The Densely Connected Convolutional Networks (DenseNet) is a family of architectures proposed as an extension of the ResNet (He et al., 2016) architecture for image classification (Huang et al., 2017). Its architecture contains dense blocks connected by transition layers, as illustrated in the feature extractor part of Fig. 2. Every dense block is composed of multiple layers that perform summation operations on the output of previous layers and forward it to the next layers in the same block. In other words, every layer receives all the feature maps from previous layers. Transition layers help to concatenate varying-size feature maps from every dense block. These operations preserve the context from earlier layers to improve generalization and overcome the problem of vanishing gradient. The final feature map obtained from the last dense block is transformed into a feature vector by applying global average pooling. This feature vector has low complexity and provides smoother decision boundaries.

3.2. Probabilistic classifier: Gaussian processes

GPs define a prior probability distribution over functions. This probability distribution, when combined with a set of observations, updates its prior knowledge and is capable of providing a posterior distribution for each (seen or unseen) sample. This posterior can be used for decision-making and uncertainty quantification in classification (our case) and regression problems. Regarding classification problems, GPs yield intractable inference to compute the posterior distribution (because the observation model and the Gaussian prior are not conjugated). To solve this problem, Stochastic Variational inference for Gaussian Processes (SVGPs) performs approximate inference in classification problems. See further details of SVGP in Hensman et al. (2015) and an intuitive review in Lopez-Perez et al. (2021). SVGP aims to approximate the posterior by solving an optimization problem. That is, maximizing the Evidence Lower BOund (ELBO). The inference is performed by Monte Carlo sampling and allows training in mini-batches, which scales to larger datasets. Finally, the predictions are computed with the learned approximated posterior distribution.

3.3. The proposed deep kernel learning model

GPs cannot directly handle high-dimensional data, for instance, images. The proposed approach combines a feature extraction step using CNNs with the GPs to leverage its advantages. In short, the DKL model has two consecutive parts: a three-block DenseNet CNN and a sparse GPs classifier. The modules are trained jointly end-to-end. In this work, we follow the inference procedure similar to the one proposed in Wilson et al. (2016b). DenseNet acts as a feature extractor (a.k.a. the backbone), computing relevant features for artifact detection and providing low dimensional data to sparse GPs for uncertainty-aware predictions, as demonstrated in Fig. 2.

Let $\mathcal{D} = (\mathbf{X}, \mathbf{y}) = \{(\mathbf{x}_n, y_n)\}_{n=1,\dots,N}$ be our dataset, where each $\mathbf{x}_n \in \mathbb{R}^{W\times H\times C}$ is a histological patch from a WSI and W, H, C represents height, width and channels, respectively. $y_n \in \{0, 1\}$ is 1 if $\mathbf{x}_n$ contains artifacts and 0 otherwise. The feature extractor model g consists of a DenseNet network with learnable parameters $\boldsymbol{\Phi}$ followed by global average pooling converting feature maps to vector embeddings. For a patch $\mathbf{x}_n$, we define the feature embedding $\mathbf{a}_n = \{a_1, a_2, \dots, a_z\}$ as:

$$g_{\boldsymbol{\Phi}}(\mathbf{x}_n) = \mathbf{a}_n. \tag{1}$$

Algorithm 1 : Training Deep Kernel Learning for Artifact Detection

```
Input: Patches $\mathbf{X}$, artifact labels $\mathbf{y}$, and number of epochs $E$.
Output: Variational $\{\mathbf{m}_j, \mathbf{S}_j\}_{j=1}^{z}$ and model $\{\boldsymbol{\Phi}, \boldsymbol{\Theta}, R\}$ parameters.
  for $j = 1$ to $z$ do
    $\mathbf{m}_j \leftarrow \mathbf{0}; \mathbf{S}_j \leftarrow \mathbf{I}_M$
  end for
  $\boldsymbol{\Theta}, \boldsymbol{\Phi} \leftarrow$ Random Initialization
  $R \leftarrow \mathbf{1}_z = (1, ..., 1) \in \mathbb{R}^z$
  for $e = 1$ to $E$ do
    1. $\mathbf{A} \leftarrow g_{\boldsymbol{\Phi}}(\mathbf{X})$.
    2. Draw $T$ samples from the posterior distribution with Eqs. (4) and (5) using $\boldsymbol{\Theta}, \mathbf{A}, \{\mathbf{m}_j, \mathbf{S}_j\}_{j=1}^{z}$.
    3. Approximate the likelihood with Eq. (3) using $R$ and the $T$ samples.
    4. Compute the prior term $\mathrm{KL}(\mathrm{q}(\mathbf{u})||\mathrm{p}(\mathbf{u}))$ using $\boldsymbol{\Theta}, \{\mathbf{m}_j, \mathbf{S}_j\}_{j=1}^{z}$.
    5. Calculate the ELBO in Eq. (6) adding the likelihood plus the prior term.
    6. Update $\boldsymbol{\Phi}, \boldsymbol{\Theta}, \{\mathbf{m}_j, \mathbf{S}_j\}_{j=1}^{z}, R$ using the Adam optimizer.
  end for
  return Optimal parameters: $\{\boldsymbol{\Phi}, \boldsymbol{\Theta}, \{\mathbf{m}_j, \mathbf{S}_j\}_{j=1}^{z}, R\}$.
```

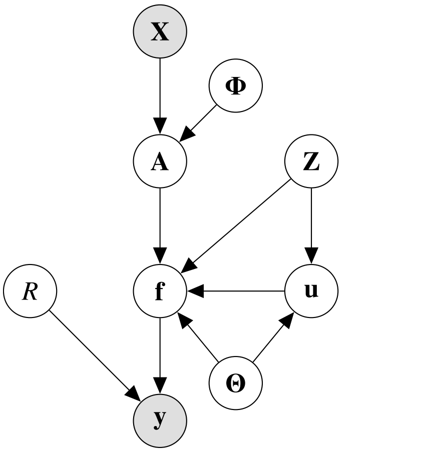


Fig. 3. Probabilistic graphical model of the proposed Deep Kernel Learning (DKL). Dark circles stand for observed variables, while light circles stand for latent variables. We obtain embeddings $\mathbf{A}$ from our histology patches $\mathbf{X}$ using the trainable parameters $\boldsymbol{\Phi}$ of our DenseNet CNN. Then, we consider a weighted sum of the sparse GPs $\mathbf{f}$ with the weights R to obtain the class label $\mathbf{y}$ (if there is an artifact in the patch or not). For scalability, we define M inducing points $\mathbf{u}$ over the inducing locations $\mathbf{Z}$. $\boldsymbol{\Theta}$ stands for the (trainable) kernel parameters of the GP.

The size of the feature embedding is z. In the experimental section, we study the influence of this size. We denote the set of all the embeddings as $\mathbf{A} \in \mathbb{R}^{N \times z}$.

The objective here is to learn z independent GPs: $\{\mathbf{f}_j\}_{j=1}^{z}$, one GP for each feature of $\mathbf{A}$. Each GP follows a prior multivariate Gaussian distribution,

$$\mathbf{f}_j \sim \mathcal{N}(\mathbf{0}, K_{\mathbf{AA}}), \qquad \forall j \in \{1, \ldots, z\}, \tag{2}$$

where the covariance matrix $K_{\mathbf{AA}}$ is defined by a Squared Exponential (SE) kernel function $k_{\boldsymbol{\Theta}}(\cdot, \cdot)$. We denote the learnable kernel parameters of the GPs as $\boldsymbol{\Theta}$. We omit the kernel parameters in the equations and their dependence on j for clarity.

Notice that each GP has n components (one per each observed sample), we denote by $\mathbf{f}_{:,n} \in \mathbb{R}^z$ the vector of independent GPs for the nth instance. The likelihood of our model is a Bernoulli distribution where the parameter is a weighted sum of the GPs using a vector of learnable weights $R \in \mathbb{R}^z$ to exploit cross-dimensional correlations,

$$\mathrm{p}(y_n | R, \mathbf{f}_{:,n}) = \frac{1}{1 + e^{-R^{\top} f_{:,n}}}. \tag{3}$$

DKL uses stochastic variational inference but it assumes some approximations to lighten the model and make sampling more efficient. First, a set of M inducing points $\mathbf{u}_j$ with inducing locations $\mathbf{Z}$ is introduced per each GP. They follow a prior distribution $\mathrm{q}(\mathbf{u}) = \prod_j \mathcal{N}(\mathbf{u}_j | \mathbf{0}, K_{\mathbf{ZZ}})$ and an approximated posterior distribution $\mathrm{q}(\mathbf{u}) = \prod_j \mathcal{N}(\mathbf{u}_j | \mathbf{m}_j, \mathbf{S}_j)$. Using local kernel interpolation, the latent function $\mathbf{f}$ is defined as a deterministic function of $\mathbf{u}$. The GP samples are computed directly by the computation,

$$\mathbf{f}_j = K_{\mathbf{AZ}} K_{\mathbf{ZZ}}^{-1} \mathbf{u}_j. \tag{4}$$

The inducing points $\mathbf{u}_j$ are also reparametrized using a Cholesky decomposition. Specifically, $\mathbf{S}_j = \mathbf{L}_j^{\top} \mathbf{L}_j$. The sampling procedure is given by,

$$\mathbf{u}_j^{(t)} = \mathbf{m}_j + \mathbf{L}_j \epsilon^{(t)}, \qquad \epsilon^{(t)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{5}$$

The inducing locations are placed on a grid taking advantage of the Toeplitz and circulant structures. For further details on the model and sampling procedures, see Wilson et al. (2016b). We approximate the ELBO by taking T samples of $\mathbf{f}$,

$$\mathcal{L}(q) \simeq \frac{N}{T \times B} \sum_{t=1}^{T} \sum_{n=1}^{B} \mathbb{E}_{\mathrm{p}(\mathbf{f}_{:,n}|\mathbf{u})\mathrm{q}(\mathbf{u})} \log \mathrm{p}(y_n | R, \mathbf{f}_{:,n}^{(t)}) - \mathrm{KL}(\mathrm{q}(\mathbf{u}) \parallel \mathrm{p}(\mathbf{u})). \tag{6}$$

This objective function enables training in mini-batches of size B, and it can be optimized with SGD or Adam. Regarding the first term, we draw T samples to estimate $\log \mathrm{p}(y_n | R, \mathbf{f}_{:,n})$. The second term can be computed in a closed form because it is the Kullback–Leibler (KL) divergence between two multivariate Gaussians. Fig. 3 presents the graphical model and Algorithm 1 summarizes the DKL model training process.

In the inference stage, for a new unseen image $\mathbf{x}_*$ we obtain the embedding vector from $g_{\boldsymbol{\Phi}}(\mathbf{x}_*) = \mathbf{a}_*$, and draw T samples $\{f_*^{(t)}\}_{t=1}^{T}$ using Eq. (5) for each GP. Then, we approximate the predictive mean $\hat{p}_*$ of y_* with

$$\hat{p}_* = \frac{1}{T} \sum_{t=1}^{T} \mathrm{p}(\mathbf{y}_* | R, \mathbf{f}_{:,*}^{(t)}). \tag{7}$$

Since the posterior of y_* follows a Bernoulli distribution, the predictive variance is given by

$$\hat{\sigma}_*^2 = \hat{p}_*(1 - \hat{p}_*). \tag{8}$$

This variance can be further split into two terms: the aleatoric uncertainty and the epistemic uncertainty (Kwon et al., 2020),

$$\hat{\sigma}_*^2 \approx \underbrace{\frac{1}{T} \sum_{t=1}^{T} \left(\mathrm{p}(\mathbf{y}_* | R, \mathbf{f}_{:,*}^{(t)}) - \mathrm{p}(\mathbf{y}_* | R, \mathbf{f}_{:,*}^{(t)})^2 \right)}_{\hat{\sigma}_{*_{al}}^2 =: \text{ aleatoric uncertainty}} + \underbrace{\frac{1}{L} \sum_{t=1}^{T} (\mathrm{p}(\mathbf{y}_* | R, \mathbf{f}_{:,*}^{(t)}) - \hat{p}_*)^2}_{\hat{\sigma}_{*_{ep}}^2 =: \text{ epistemic uncertainty}}. \tag{9}$$

```
Algorithm 2 : Inference with Deep Kernel Learning for Artifact Detection
Input: New patch $\mathbf{x}_*$, optimal variational $\{\mathbf{m}_j, \mathbf{S}_j\}_{j=1}^{z}$ and model parameters $\{\boldsymbol{\Phi}, \boldsymbol{\Theta}\}$.
Output: Predictive mean $\hat{p}_*$, variance $\hat{\sigma}_*^2$ and the epistemic uncertainty $\hat{\sigma}_{*_{ep}}^2$.
  1. $\mathbf{a}_* \leftarrow g_{\boldsymbol{\Phi}}(\mathbf{x}_*)$.
  2. Draw $T$ samples from the likelihood term using Eqs. (3), (4), and (5).
  3. Compute the predictive mean $\hat{p}_*$ with Eq. (7).
  4. Compute the predictive variance $\hat{\sigma}_*^2$ with Eq. (8).
  5. Compute the predictive epistemic uncertainty $\hat{\sigma}_{*_{ep}}^2$ with Eq. (9).
  return $\hat{p}_*$, $\hat{\sigma}_*^2$, $\hat{\sigma}_{*_{ep}}^2$.
```

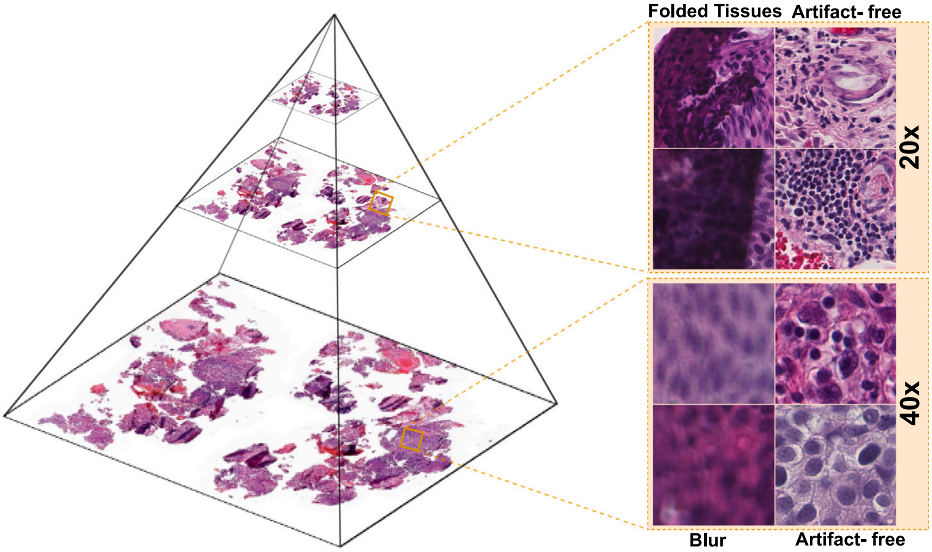


Fig. 4. A depiction of Whole Slide Image (WSI) pyramid with multiple magnification levels. Patches for the blur class are extracted at 40× magnification, and patches for folded tissues are extracted at 20× magnification for a broader field of view along with the corresponding artifact-free class.

Table 1
A breakdown of the number of WSIs and patches obtained from the EMC dataset for blur, folded tissue, and artifact-free classes in each subset.

	Training	Validation	Test	Total
No. of WSIs	35	10	10	55
Blur	5661	754	1137	7552
Artifact-free (40×)	5249	1441	965	7655
Folded tissue	478	130	138	746
Artifact-free (20×)	513	140	131	784

The aleatoric is irreducible and is inherent to randomness in the data and the epistemic uncertainty refers to the noise in the model's parameters because of the lack of knowledge (or data). In the experiments, we will use the predictive variance, which combines these two sources of uncertainty, and the epistemic uncertainty to further examine the uncertainty of our model on new samples. High values indicate a high uncertainty. We summarize the whole inference procedure with the DKL model in the Algorithm 2.

4. Data and experimental details

This section provides description of the histological data and the experiments carried out to validate our DKL model. Section 4.1 details the data and its preparation. Section 4.2 provides the implementation details, and finally, Section 4.3 explains the metrics used to evaluate the proposed model.

4.1. Data materials

4.1.1. EMC dataset

We have analyzed 55 glass slides of bladder tumor resections from the Erasmus Medical Center (EMC), Rotterdam, The Netherlands. The glass slides were fixed with formalin and stained with Hematoxylin and Eosin (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. WSIs were anonymized, and all ethical guidelines were followed before creating the dataset. The dataset was divided into 35/10/10 for training, validation, and test sets at the WSI level. A non-pathologist trained for the task manually annotated the WSIs for blurry areas, folded tissues, and artifact-free tissue areas. There were at least two areas annotated in each WSI for artifact and artifact-free regions, but none of the WSIs were densely annotated into different tissue types.

This in-house dataset was used for training and validating proposed DKL models. Since CNN cannot process the entire WSI at once, we split the WSIs further into sub-images (patches) (Kanwal et al., 2022b). In the first step, binary thresholding with the Otsu method was applied to the HSV-transformed image to perform the foreground-background segmentation of the WSIs. Later, the obtained foreground was used to extract patches of 224 × 224 pixels with at least 70% overlap with the annotation mask (blur, fold, artifact-free). Tissue folds were patched at 20× magnification as they need a broader field of view for better visibility, and blurry regions were patched at 40× magnification. The number of patches extracted for the artifact-free class was fixed in each WSI to avoid significant class imbalance. Fig. 4 shows some examples of extracted patches for both artifact classes with the artifact-free class at the corresponding magnification. A further breakdown of the number of patches in each class with both magnification levels is presented in Table 1.

Table 2
The configuration and hyper-parameters used for the training of the proposed DKL model (above) and baseline model (below).

Architecture	Parameter	Value
DKL model	Input patch size	224 × 224 × 3
	Patch magnifications in WSI	20 × (for folded tissue) & 40 × (for blurry areas)
	Number of denseblocks in CNN	3
	Structure of dense layers	[6, 6, 6]; [8, 8, 8]; [10, 10, 10]
	Denseblock configuration	growth_rate = 12, compression = 0.5, num_init_features = 24
	Inducing points for GP	{64, 128, 256, 384, 512}
	Learning rate	0.001 Initialized for ReduceLROnPlateau scheduler
	Batch size	32
	Kernel	Squared-Exponential
	Optimization algorithm	Adam with weight decay = 0.0001
	Activation function	ReLu
	Objective function	Variational ELBO with β = 0.5
Baseline model	Configuration of FC layers	[512, 128, 2]
	Activation function	ReLu
	Loss function	Focal Loss with γ = 2, α = 0.25
	Scheduler	ReduceLRonPlateau with patience = 5
	Dropout regularization	0.2
	Early-stopping regularization	10 epochs

4.1.2. TCGA focus dataset

The Cancer Genome Atlas (TCGA) Focus[1] is a publicly available dataset with patches processed from 1000 WSIs by the National Cancer Institute (NCI), the United States. The dataset contains a wide spectrum of stain and texture variation due to its preparation from 52 different organ types. TCGA Focus comprises 14,371 patches of size 1024 × 1024 pixels with 11,328 in-focus and 3043 out-of-focus labels (Wang et al., 2020). Due to the availability of binary focus labels for ROI, we have used this dataset as an external evaluation benchmark for DKL models on blur detection task.

4.1.3. FocusPath dataset

FocusPath[2] is another public dataset that contains 8640 patches of 1024 × 1024 pixels (Hosseini et al., 2019). These patches are extracted from diverse WSIs, stained with nine different chemical dyes, and scanned with a 40× magnification lens of a Huron Tissue Scope LE1.2. at 0.25 μm/pixel resolution. Every patch has a ground truth class for a focal level (i.e., corresponding to its absolute z-level score). Label 0 corresponds to the lowest extent of blurriness, whereas label 13 indicates the highest degree of blur. We utilize this external dataset to determine the generalization ability of DKL and baseline models to detect blur in unseen data.

4.1.4. SUH dataset

We have also used 4 Hematoxylin, Eosin, and Saffron (HES) stained bladder biopsy WSIs from the University Hospital of Stavanger (SUH) in Norway. These WSIs were scanned with a Leica SCN400 at 40× magnification and stored in *scn* format. A non-pathologist trained for the task annotated folded tissue regions. Later, patches of 224 × 224 pixels were extracted at 20× magnification in a similar fashion as described in 4.1.1. We have tested folded tissue detection models on this external dataset to assess their classification ability on WSIs with different staining.

4.2. Implementation details

Patch extraction was performed using Histolab (Colling et al., 2019) Python library. The obtained patches were normalized by re-scaling to ImageNet (Deng et al., 2009) mean and standard deviation. Augmentation was done at every training epoch by applying random geometric transformations, including rotation, horizontal and vertical flips with a probability of 0.5. For validation on external cohorts, patches were center cropped to 224 × 224 pixel size.

We trained distinct models for binary classification of fold and blur, allowing for comparison to existing works in the literature. Models using DCNN architectures were implemented using Pytorch (Paszke et al., 2019). All DCNN architectures were initialized with ImageNet (Deng et al., 2009) weights to benefit from transfer learning. The fully connected (FC) layers from these pretrained DCNNs were replaced with a custom three-layer FC classifier, and initialized with random weights. Hyper-parameters were explored through a grid search for improved validation metrics. The final chosen parameters were *Adam* optimizer (Kingma and Ba, 2014) with weight decay of 0.01, *ReduceLROnPlateau* scheduler with a learning rate of 0.01 and patience of 5, batch size of 32, dropout of 0.2 and Focal loss (Lin et al., 2017). We applied an early stopping of 10 epochs on validation loss to avoid overfitting. Lastly, we used the best model weights on the validation set to report evaluation metrics.

The proposed DKL method was implemented using GPytorch (Gardner et al., 2018) library. A three-block DenseNet CNN was used as the backbone for the DKL models. This customized CNN was initialized with random weights due to the unavailability of ImageNet weights for this structure of DenseNet. Later, global average pooling was applied to obtain a feature vector from the CNN feature map. We used inducing points in the range $[64 - 512]$ and an SE kernel (Rasmussen, 2003). For DKL training, the ELBO was optimized using Adam.

For a fair comparison, we also developed a baseline model using a customized three-block DenseNet CNN with the FC classifier. This baseline model is a smaller version of SOTA DenseNet DCNN to establish an equitable comparison against the counterpart DKL model (using similar CNN with GP classifier). All chosen hyper-parameters are summarized in Table 2. To further assess the stochastic dominance of our proposed model against the baseline model, we borrowed the Almost Stochastic Order (ASO) test (Del Barrio et al., 2018; Dror et al., 2019) from the deep-significance[3] Python library. We performed retraining of DKL and baseline models five times by fixing the hyperparameters and changing the seed in every round. All experiments were conducted on NVIDIA GeForce RTX 3090 with 24 GB.

[1] https://zenodo.org/record/3910757#.YtZ-lnZBwnA.

[2] https://zenodo.org/record/3926181#.YtaEAnZBwnA.

[3] https://deep-significance.readthedocs.io/en/latest/.

Table 3
DCNN models on the validation and test set of the EMC cohort 4.1.1. Four SOTA DCNN architectures are trained on the EMC cohort. All architecture backbones were initialized with ImageNet weights. The number of training parameters and the output feature size are reported to show the computational complexity of each model. For blur detection, we report results from the paper of (*) Albuquerque et al. (2021) and results from (†) Babaie and Tizhoosh (2019) are reported for tissue fold detection. Note that these models are trained on different data and under different experimental setup. We provide validation results from literature with identical blur and fold detection works for referential comparison only. The best results are highlighted in bold. Dash (–) indicates results not reported in the published works.

CPATH task	Architecture	Trainable parameters	Feature size (z)	Validation/Test results				Acc. (%) from previous works
				Accuracy (%)	F1	AUC-ROC	MCC	
Blur detection	ResNet	60.19 M	2048	97.6/99.1	0.964/0.992	0.969/0.991	0.947/0.983	88.7*
	DenseNet	28.68 M	2208	97.9/92.1	**0.980**/0.993	**0.981**/0.992	**0.962**/0.982	–
	GoogleNet	13 M	1024	97.2/99.3	0.958/0.993	0.967/0.992	0.937/0.985	92.4*
	MobileNet	0.54 M	960	**98.2/99.5**	0.974/**0.995**	0.980/**0.994**	0.960/**0.989**	94.4*
Fold detection	ResNet	60.19 M	2048	97.4/91.8	0.972/0.921	0.971/0.91	0.95/0.85	94.6†
	DenseNet	28.68 M	2208	96.3/92.2	0.962/0.929	0.961/0.919	0.923/0.853	96.7†
	GoogleNet	13 M	1024	**98.5/92.5**	**0.985**/0.932	**0.985**/0.923	**0.971/0.861**	93.7†
	MobileNet	0.54 M	960	97.8/92.9	0.977/**0.935**	0.978/0.927	0.955/0.864	-

4.3. Evaluation metrics

We evaluated the performance of each model using four different metrics; Accuracy, F1-score, Area Under the Receiver Operating Characteristic (AUC-ROC), and Mathew Correlation Coefficient (MCC). Let FP, FN, TP, and TN represent false positive, false negative, true positive, and true negative in the confusion matrix (CM), respectively. The ratio of correctly classified patches to the total patches can be defined as $Accuracy = (TP{+}TN)/(TP{+}TN{+}FP{+}FN)$. The F1 score can be more informative than accuracy when FN and FP are crucial. It is defined as: $F1 = 2 \cdot (\text{precision} \cdot \text{recall})(\text{precision} + \text{recall})$ where recall = sensitivity $= TP/(TP + FN)$ and precision $= TP/(TP + FP)$.

AUC-ROC measures the capability to distinguish between classes regardless of the decision threshold. A higher AUC score indicates a better classifier for the task. Finally, MCC is an informative measure in binary classification over an imbalanced dataset (Chicco and Jurman, 2020). MCC, as defined in Eq. (10), achieves a high score in $[-1, 1]$ when the model correctly classifies positive and negative instances from the dataset.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}. \tag{10}$$

To evaluate the statistical significance of our results, we utilized the ASO (Del Barrio et al., 2018; Dror et al., 2019) test in five random runs of both the baseline and the proposed DKL models. We used the test on the MCC metric with a significance level of $\alpha = 0.95$. ASO returns a value showing the amount of violation in stochastic order, where a value less than 0.5 shows the DKL model is statistically better than the baseline model. Finally, we calculate trainable parameters, feature size, and inference time per patch to highlight the computational complexity of the models.

5. Results and discussion

This section presents and discusses the experimental results on validation, test, and external datasets for two different binary classification problems; blur and fold detection. First, we study well-known SOTA DCNN architectures (Section 5.1). Then, we present the proposed DKL models, asses their generalization capability and compare them with counterpart baselines and literature works (mentioned in Section 5.2). Furthermore, we compare the performance of the best-performing DKL model and baseline on external datasets to evaluate their robustness (Section 5.3). Finally, we study their confidence and compare the uncertainty estimates of the DKL model and the baseline using MC dropout (Section 5.4).

5.1. State-of-the-art DCNNs on the EMC cohort

The purpose of this experiment is to evaluate the performance of four SOTA DCNN architectures, namely, ResNet152 (He et al., 2016), GoogleNet (Szegedy et al., 2015), MobileNetv3 (Howard et al., 2019), and DenseNet161 (Huang et al., 2017), on blur and fold detection problems. These popular architectures are used in the existing publications (Babaie and Tizhoosh, 2019; Albuquerque et al., 2021) with similar classification tasks. Table 3 displays validation and test results on the EMC dataset, allowing us to assess both the performance and generalization capability. The DCNN models tend to perform relatively well on the blur detection task compared to fold detection, probably because there is a greater availability of blur data. Our DCNN models, with carefully chosen hyper-parameters, outperform previously reported methods in the literature (Babaie and Tizhoosh, 2019; Albuquerque et al., 2021) in terms of reported accuracy for both detection tasks. Note, these results cannot be compared directly as the models were trained and tested on different data. However, they correspond to identical artifact detection tasks and are therefore reported for comparative reference.

Comparing the architectures, we can observe that MobileNet, despite having fewer parameters, produces acceptable classification results. Heavier architectures such as DenseNet and GoogleNet achieve the highest F1, AUC-ROC, and MCC scores for blur and fold detection on the validation set. Additionally, Table 3 suggests that DCNNs with a large number of parameters, like ResNet, require more data even with the knowledge from ImageNet initialization. The results also reveal that in the case of fold detection, every DCNN architecture generalizes poorly, with metrics significantly dropping from validation to test. This is due to the overfitting tendencies of DCNN models on relatively small datasets. Surprisingly, DenseNet exhibits the most pronounced overfitting compared to other architectures in both classification tasks.

In the following experiment, we will test if the DKL model has better generalizing capabilities compared to SOTA DCNNs. Therefore we want to proceed with DenseNet CNN as the backbone, as we observed that DenseNet has more improvement potential in the generalization.

5.2. Evaluation of DKL model on the EMC cohort

In this experiment, we present the results of the proposed DKL model, in terms of performance, generalizability, and computational complexity. Traditional DCNN architectures have a large number of trainable parameters and relatively large feature vectors/embedding that work as the input for a classifier of fully connected layers. For example, DenseNet161 (Huang et al., 2017) contains four dense blocks, and the resulting feature embedding has a large size (z = 2208) (see Table 3). Such a large feature vector size, z, is not convenient for the DKL model. Consequently, we have chosen to utilize a smaller, customized version of DenseNet with three dense blocks to explore

Table 4
DKL models on the validation and test set of the EMC Cohort 4.1.1. Our proposed DKL models (with GP classifiers) and baseline models (with FC classifiers) are tested with three variants of Densenet feature extractors (used as a backbone). The best results for each task are highlighted in bold.

CPATH task	Subset	Architectures	Accuracy (%)	F1	AUC-ROC	MCC
Blur detection	Validation	DenseNet$_{(6,6,6)}$ + FC	96.53	0.948	0.956	0.922
		DenseNet$_{(6,6,6)}$ + GP	97.79	0.968	0.975	0.951
		DenseNet$_{(8,8,8)}$ + FC	96.54	0.949	0.96	0.923
		DenseNet$_{(8,8,8)}$ + GP	97.70	0.968	0.974	0.951
		DenseNet$_{(10,10,10)}$ + FC	94.61	0.926	0.955	0.888
		DenseNet$_{(10,10,10)}$ + GP	**98.13**	**0.972**	**0.981**	**0.958**
	Test	DenseNet$_{(10,10,10)}$ + FC	97.33	0.975	0.971	0.947
		DenseNet$_{(10,10,10)}$ + GP	**99.52**	**0.996**	**0.995**	**0.990**
Fold detection	Validation	DenseNet$_{(6,6,6)}$ + FC	97.77	0.977	0.956	0.978
		DenseNet$_{(6,6,6)}$ + GP	99.25	**0.992**	**0.992**	**0.985**
		DenseNet$_{(8,8,8)}$ + FC	98.88	0.988	0.989	0.978
		DenseNet$_{(8,8,8)}$ + GP	**99.63**	0.988	0.988	0.977
		DenseNet$_{(10,10,10)}$ + FC	97.40	0.973	0.974	0.948
		DenseNet$_{(10,10,10)}$ + GP	98.88	0.984	0.989	0.970
	Test	DenseNet$_{(6,6,6)}$ + FC	91.41	0.923	0.911	0.839
		DenseNet$_{(6,6,6)}$ + GP	**93.28**	**0.938**	**0.930**	**0.873**

Table 5
Computational complexity of DKL and baseline models. For top-performing DKL and baseline models on blur and fold detection tasks, trainable parameters, output feature size, inference time per patch, and the outcomes of Almost Stochastic Order (ASO) test on both classification tasks are presented.

Architectures	Trainable parameters	Feature size (z)	Inference time (ms)[†]	ASO for Blur/Fold
DenseNet$_{(6,6,6)}$ + FC	0.308 M	132	136.6	0.453/0.224
DenseNet$_{(6,6,6)}$ + GP	2.354 M		3.648	
DenseNet$_{(10,10,10)}$ + FC	0.543 M	216	266.6	0.339/0.051
DenseNet$_{(10,10,10)}$ + GP	3.933 M		6.501	

† Time is estimated by a hundred runs of MC dropout for the (CNN+FC) baseline model and drawing a hundred samples for the DKL model.

Table 6
Testing robustness of DKL models on external datasets. This table presents the results of DKL and corresponding baseline models using the best-performing DenseNet backbones on blur and fold detection tasks. The evaluation is performed on the unseen data from external datasets (TCGA Focus 4.1.2 & SUH 4.1.4). The best results for every cohort are marked in bold.

Dataset (Artifact)	Architecture	Accuracy (%)	F1	MCC
TCGAFocus (Blur)	DenseNet$_{(10,10,10)}$ + FC	67.19	0.286	0.09
	DenseNet$_{(10,10,10)}$ + GP	**69.66**	**0.443**	**0.301**
SUH (Fold)	DenseNet$_{(6,6,6)}$ + FC	**89.9**	**0.904**	**0.814**
	DenseNet$_{(6,6,6)}$ + GP	87.9	0.887	0.780

the discrimination power of GPs with reasonable-sized feature vectors. Each block has x layers, where we let $x \in \{6, 8, 10\}$, and denote the model architecture as DenseNet$_{(x,x,x)}$. In other words, the DKL model comprised of DenseNet$_{(x,x,x)}$ with GP classifier, and the baseline model contains DenseNet$_{(x,x,x)}$ with FC classifier. The baseline model used in this experiment is a smaller variant of the SOTA DenseNet161 (Huang et al., 2017).

Table 4 shows the results of the DKL and baseline models on the validation set. DKL models demonstrate better performance for both blur and fold detection tasks across all DenseNet configurations and consistently outperform the baselines. The best-performing architecture for DKL models for blur and fold detection are DenseNet$_{(10,10,10)}$ and DenseNet$_{(6,6,6)}$, respectively. Compared to the SOTA results in Table 3, the proposed DKL model demonstrates competitive performance, even with the smaller DenseNet CNN.

To be able to assess the generalization capabilities, Table 4 also shows the results on the test set of the EMC cohort with the best-performing DenseNet backbones over the validation set. DKL shows very good generalization and significant improvements compared to deterministic SOTA DCNN and the baselines when the training data is limited. Moreover, the results of the ASO for both tasks (in Table 5) highlight the stochastic dominance of DKL models over baseline models. Especially on smaller datasets (like our fold dataset), the classification performance of the DKL model holds notable statistical superiority. Overall, the DKL models outperformed on unseen data from the same distribution (same cohort), providing a similar or better metric compared to SOTA DCNN architectures (recall Table 3) and baseline models for both artifact detection tasks.

5.2.1. Computational cost and scalability

Table 5 presents the computational complexity of the best-performing DKL models which underscores their advantage over SOTA DCNNs. The table highlights that more dense layers result in more trainable parameters and larger feature vectors (z). Top-performing DKL model for blur detection has nearly eight times fewer parameters compared to its counterpart SOTA DCNN in Table 3 and obtains the same AUC score (see Table 4). Similarly, the top-performing DKL model for fold detection, with roughly seventeen times smaller feature vector size (z) relative to SOTA DenseNet DCNN in Table 3, yields a higher MCC score than both the baseline (see Table 4) and DenseNet DCNN in Table 3. Moreover, Table 5 also exhibits that applying MC dropout to

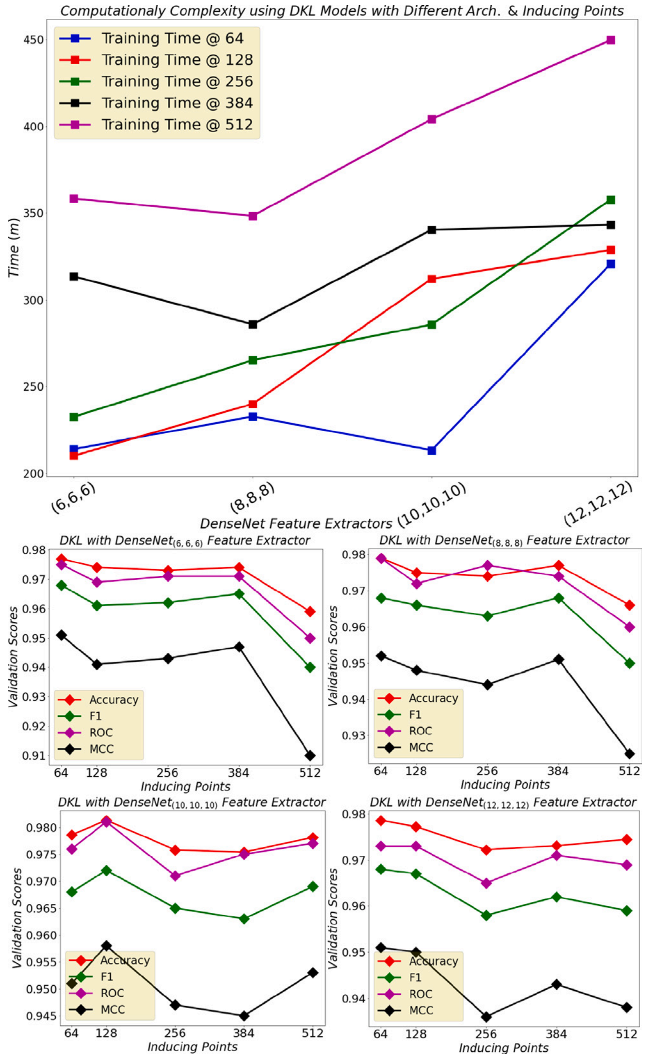


Fig. 5. DKL models trained on fold dataset using different configurations to assess the scalability. To the top, the running time for training is plotted against the increase in the size of the CNN network/ feature vector. Bottom: validation scores of different metrics are plotted as a function of the number of inducing points. The four subplots are for different depths of DenseNet CNN used with the GP classifier.

the baselines (or DCNNs) is computationally expensive and has a clear disadvantage with significantly higher inference time.

To further explore the scalability cost of the sparse GP classifier, we analyze DKL models using DenseNet$_{(x,x,x)}$ with $x \in \{6, 8, 10, 12\}$, resulting in feature vector sizes $z \in \{132, 174, 216, 258\}$ and different number of inducing points (defined in Section 3.3), $M \in \{64, 128, 256, 384, 512\}$. Fig. 5 illustrates that an increasing number of dense layers, i.e., size of feature vector/number of estimated GPs, or inducing points result in higher computational requirements for training the DKL model. However, this increase does not inflict a significant improvement in classification performance, which shows an exciting side of the probabilistic classifier to perform well with fewer features. In general, a high number of the inducing points, which is often considered necessary for a better approximation of the GP classifier, does not reflect better results in our small datasets. Overall, the end-to-end training of the DKL model is found to be more computationally expensive compared to training the baseline or DCNN.

In summary, introducing DKL, which can give us uncertainty measures in addition to predictions, does not cost us anything in terms of performance, and it generalizes equally well or better than SOTA DCNN and corresponding baseline models.

5.3. Validation on external data

Since DCNNs often underperform when they experience data from different sources, it would be interesting to evaluate our DKL and baseline models on external datasets. In this experiment, we choose the best DKL and baseline models from Table 4 to investigate their robustness on patches with different tissue types and staining compared to the training data, i.e., different cohorts from other labs and other diseases. Table 6 shows inference results for the TCGA Focus dataset (blur detection) and the SUH dataset (fold detection). In the case of the TCGA focus dataset, which carries a more vast texture deviation than the EMC cohort, both baseline and DKL models suffer from a decline in their classification ability. However, the DKL model performs better than the baseline. On the contrary, DKL models for the fold detection task on the SUH dataset lag behind the baseline model by a slight margin.

Fig. 6 displays t-SNE plots for features extracted using DenseNet feature extractors trained using EMC data and illustrates how feature extractors find relevant features on the external datasets. In the blur detection task, we notice that the feature extractor from the baseline model provides significantly overlapped features, which is also the case for the DKL feature extractor, but a bit less so. For fold detection, both feature extractors separate the classes well, but the baseline model seems slightly better, consistent with the results in Table 6.

The FocusPath dataset provides ordinal regression labels for different levels of blur. Therefore, we evaluate how many patches are classified as blurry for each ordinal blur label, as demonstrated in Fig. 7. The baseline model (with DenseNet$_{(10,10,10)}$) assigns more blur labels to patches with lower ordinal labels (less blurry ones). This fact may be due to the overconfident predictions of the baseline on external unseen data. Although there is a massive variation in the staining, the DKL model performs reasonably well and starts detecting blur in more than half of the samples after the sixth ordinal label. We also notice that patches with the greatest extent of a blur (in labels 12 & 13) are detected perfectly by both baseline and DKL models. This experiment asserts the higher robustness of the DKL model on new unseen data with a high degree of stain variations.

5.4. Uncertainty quantification for blur detection on the EMC cohort

In order to quantify and assess the uncertainty in the predictions over the EMC cohort, we selected our best blur detection DKL model (with DenseNet$_{(10,10,10)}$ from Table 4) for this section. For the DKL model, we take a hundred samples from the posterior distribution to estimate the predictive mean p_* (given by Eq. (7)) and the epistemic uncertainty $\hat{\sigma}^2_{*ep}$ (given by Eq. (9)). To compare with MC dropout method applied to DCNNs, we performed a hundred forward passes (inference) over the baseline model to approximate uncertainty estimates.

The first experiment compares the predictive mean and uncertainty estimates of the baseline and DKL models. Fig. 8 shows the predictive mean, and the shadow region represents the confidence interval $p_* \pm 2\hat{\sigma}_{*ep}$ plotted for fifteen random patches from the EMC test set. It can be observed that the DKL model produces tight confidence boundaries and high probabilities in the correctly classified patches. Also, in less confident samples, the interval is consistently wider. The baseline model has lower confidence for most of the samples in both classes, and its confidence estimates lack consistency compared to our DKL model. In the true negative subplot in Fig. 8 (bottom), we can see that the baseline misclassifies sample nr. 3, which has an artifact-free label. Furthermore, the baseline model is quite confident in this incorrect prediction. This experiment indicates that the uncertainty estimates of

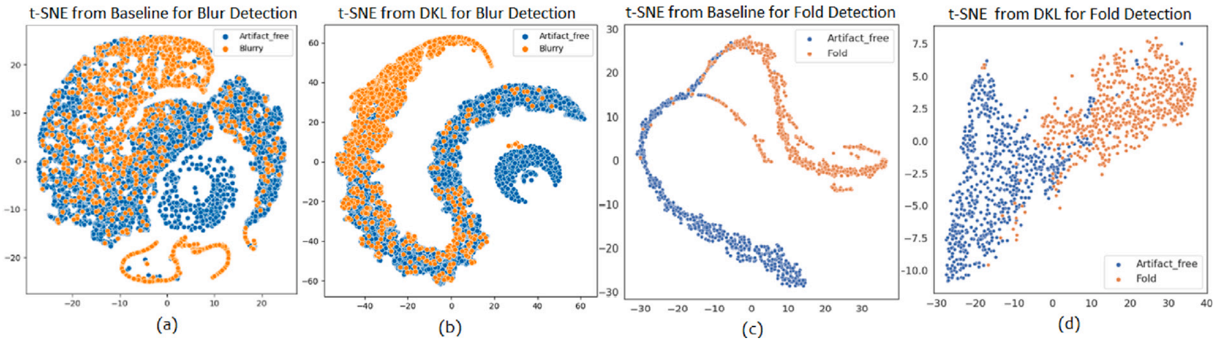
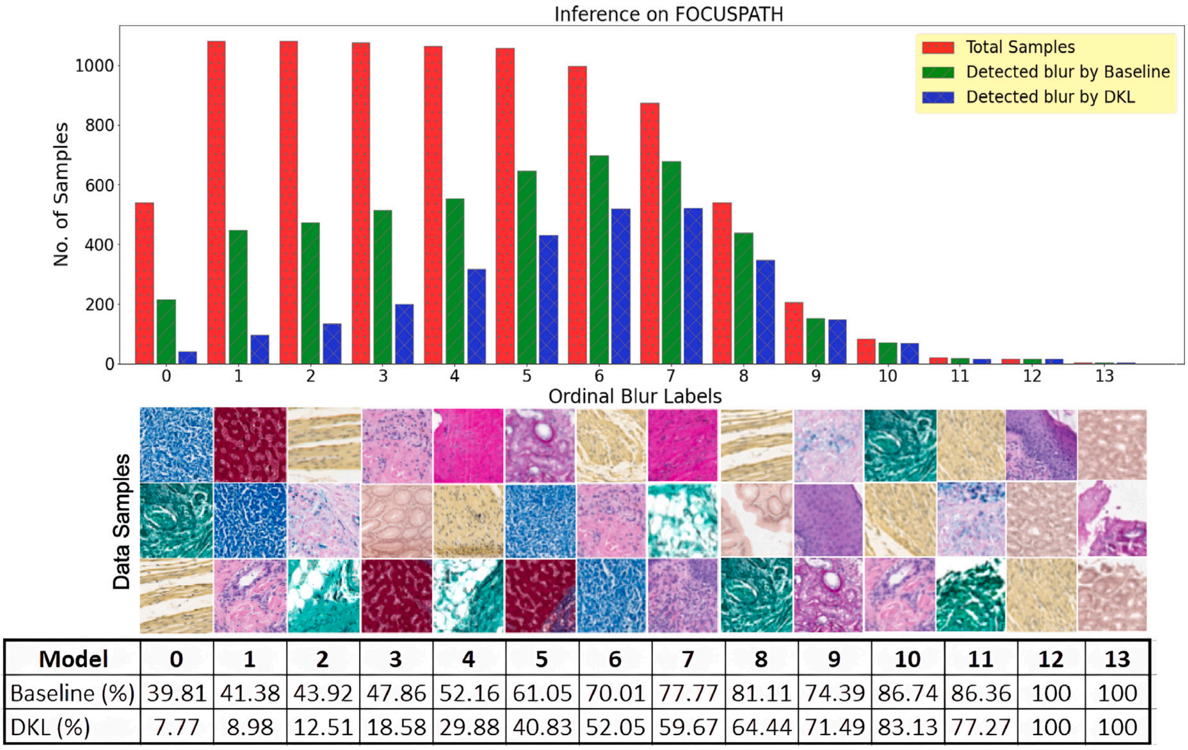


Fig. 6. A t-SNE demonstration of features from inference using baseline and DKL models. DenseNet CNN feature extractors from best-performing models in the previous experiment (DenseNet$_{(10,10,10)}$ for blur and DenseNet$_{(6,6,6)}$ for fold) are used to extract features from external datasets. Orange refers to the patches with artifacts, and blue for artifact-free patches. Plots (a) & (b) compare features for the blur detection task on the TCGA Focus dataset, and plots (c) & (d) compare features for the fold detection task on the SUH dataset.



Model	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Baseline (%)	39.81	41.38	43.92	47.86	52.16	61.05	70.01	77.77	81.11	74.39	86.74	86.36	100	100
DKL (%)	7.77	8.98	12.51	18.58	29.88	40.83	52.05	59.67	64.44	71.49	83.13	77.27	100	100

Fig. 7. DKL and baseline models tested on publicly available FocusPath dataset 4.1.3. The dataset is labeled in an ordinal fashion where label 0 corresponds to the lowest degree of blur and label 13 corresponds to the highest degree of blur. Red bars indicate the total number of samples for each label. The green and blue bars show samples predicted to be blurred by baseline and DKL models, respectively.

Table 7
Uncertainty estimate of best-performing blur DKL model on artifact class. Predictive epistemic uncertainty is calculated over only artifact class from the test set of the EMC cohort 4.1.1, and mean and the standard deviation is reported. Accuracy and F1-score are calculated over the entire test set.

Testset class	Accuracy (%)	F1	Uncertainty mean ($\times 10^{-6}$)	Uncertaintystd. ($\times 10^{-9}$)
Blur	99.52	0.995	7.96	2.26
Fold	98.63	0.944	60.41	77.43

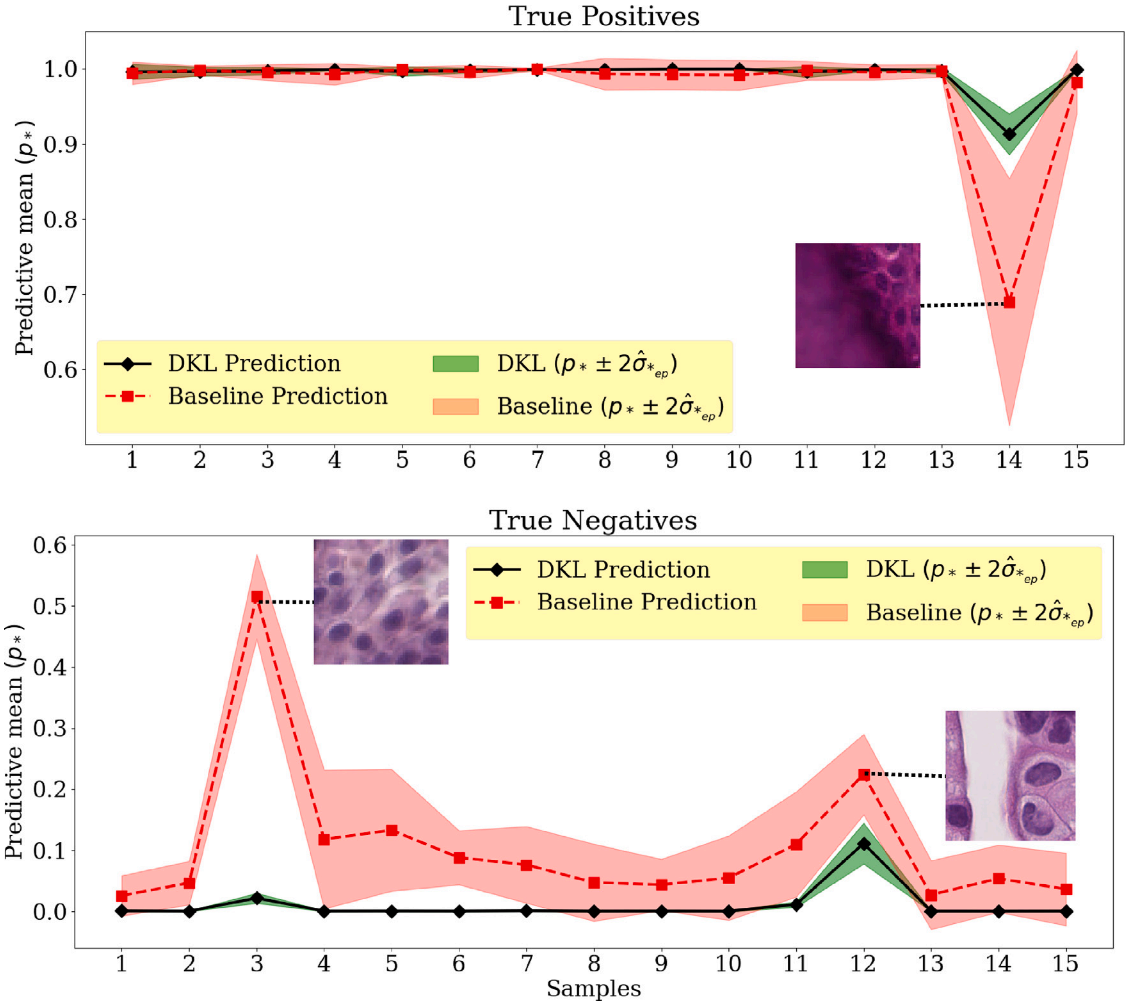


Fig. 8. A comparison of uncertainty estimates of fifteen random patches and their prediction by the DKL and baseline models. The figure was created using best-performing blur models on the test set from the EMC cohort 4.1.1. The red and black lines show the probability of blur class from the baseline and DKL models, respectively. The first row shows the predictions by the baseline and DKL models on the patches with blur labels. The second row shows predictions on patches with artifact-free labels. The predictive epistemic uncertainty of both models is shown across their predictive mean.

DKL models are more meaningful than the ones obtained by applying MC dropout to the baseline models or DCNNs.

To further visualize the uncertainty quantification, we plot the prediction of the DKL model over four random patches from each category of the Confusion Matrix (CM). Fig. 9 depicts the predictive mean over each example along with the predictive variance given by Eq. (8). Regarding the predictive mean and the variance, the DKL model is able to accurately classify True Positive (TP) and True Negative (TN) samples with significant confidence, as indicated by the predicted probabilities being close to 1 and 0, respectively, and lower variance for these predictions. Besides, a False Positive (FP) sample is predicted closer to the decision boundary, and the confidence interval around this prediction is wider, indicating lower confidence in false classifications. Interestingly, the number of False Negatives (FN) is zero, which shows that the DKL model did not miss any artifact patch. In conclusion, the DKL model is able to identify cases where it is certain of the correct classification and less certain when it makes mistakes.

The last experiment on uncertainty shows the behavior of our blur detector when it finds folded artifacts in practice. That is, the model trained on blur artifacts is used to classify patches with folded tissue artifacts on unseen data from the same cohort. We calculate the predictive epistemic uncertainty of each sample. Then, we report the mean and standard deviation of variance across all the samples in the artifact class as shown in Table 7. The accuracy and F1 scores were calculated over the entire EMC test set. We observe that our blur detector is able to identify the fold patches as artifacts. This is probably due to learning well the morphology of the artifact-free patches. Besides, many folded tissue patches contain blur due to unaligned lens focus. On average, the predictive epistemic uncertainty is nearly eight times higher on the fold test set with a bigger standard deviation across samples. The top part of Fig. 10 illustrates the t-SNE plots for the feature extractor of the blur DKL model used on the test set. We see that the discriminative capability is very good for both blur and fold. At the bottom of Fig. 10 the prediction for fifteen random samples of blur and fold are plotted. We see that the DKL model has higher uncertainty in predicting the folded tissue class as an artifact compared to the learned distribution, i.e. the blur class, as can be expected.

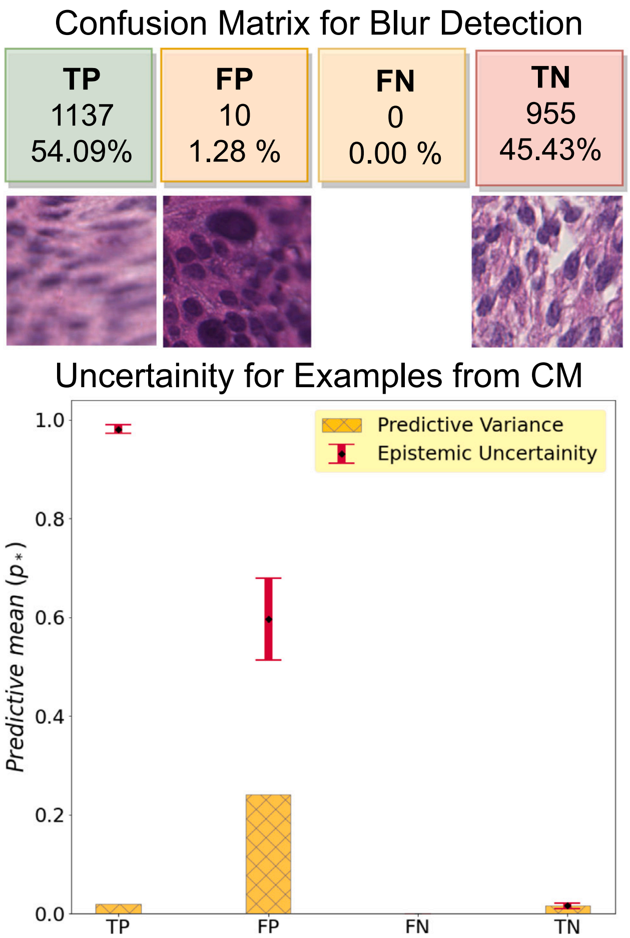


Fig. 9. Predictive mean and variance for samples from the EMC Cohort. (Top) shows DKL predictions of blurry patches on the test set of the EMC cohort. Confusion Matrix (CM) shows the total classified patches in each category along with an example patch. (Bottom) The predictive variance plot shows an example patch CM (shown on the top) with predictive variance.

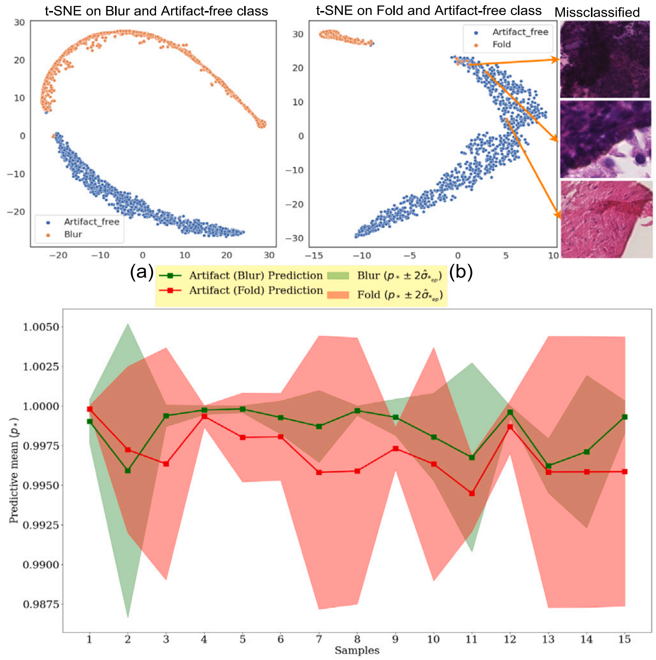


Fig. 10. Best-performing blur DKL model used for folded tissue (unseen artifacts) prediction on the EMC Cohort. (Top) t-SNE scatter plots showing how the DenseNet feature extractor from the DKL model is able to distinguish blur (subplot (a)) and folded tissue (subplot (b)) class from artifacts-free images. (Bottom) predictive mean and epistemic uncertainty for fifteen random patches from blur and folded tissue class of the EMC test set.

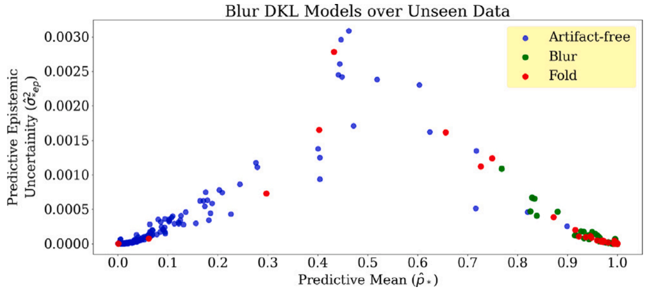


Fig. 11. Predictive mean and epistemic uncertainty over unseen (blur) and unseen (fold) data from the EMC test set. We see that more unseen fold patches are closer to the decision boundary and have higher uncertainty than blurred ones.

Fig. 11 shows the values of predictive mean and epistemic uncertainty for all the patches in the EMC test set, again using the DKL model trained for classifying blur vs. artifact-free. The figure shows that patches predicted with strong probability, i.e., close to 0 or 1 (far from the decision boundary), have lower variance (i.e., less predictive epistemic uncertainty). On the contrary, patches near the decision boundary ($0.4 < \hat{p}_* < 0.6$) have a higher predictive uncertainty, which agrees with the predictive variance decomposition given by Eq. (9). A larger proportion of fold patches are situated near the decision boundary compared to blurred ones, indicating that the model's confidence in detecting new artifacts is lower, as anticipated.

6. Conclusion & future work

This paper proposes an uncertainty-aware method for artifact detection in histopathological WSIs. The proposed DKL model combines the feature extraction power of DenseNet CNN and the probabilistic modeling of GPs. We trained several models in an end-to-end fashion with a varying depth of the DenseNet feature extractor for blur and folded tissue detection tasks.

We analyzed the DKL model against state-of-the-art DCNNs and baselines (DenseNet-based custom feature extractors with fully connected classifiers). We discovered that the DKL model outperforms DCNNs and baseline models, generalizes well to unseen data, and is robust to new data with different stains and tissue types. The proposed DKL model is computationally efficient in providing meaningful confidence in their predictions compared to applying the Monte Carlo dropout to DCNNs. Interestingly, the DKL model, trained on one artifact class, is able to correctly detect other artifacts, maintaining higher uncertainty in dubious cases or wrong predictions. This fact implies that the proposed DKL model is reliable and can measure how *sure* it is in its predictions.

In future work, we will combine these artifact detection models in the preprocessing pipeline of CPATH systems to run diagnostic algorithms reliably. Moreover, artifact detection methods may also benefit existing quality control approaches (Janowczyk et al., 2019; Shrestha et al., 2016) that overlook the presence of artifacts and only incorporate sharpness, contrast, noise, metadata, and color properties for evaluating the usability of WSIs for developing CPATH systems.

Data and code availability

The code and data is publicly available on Github.

CRediT authorship contribution statement

Neel Kanwal: Conceptualization, Software, Experiments, Writing – original draft, Writing – review & editing, Visualization. **Miguel López-Pérez:** Conceptualization, Methodology, Visualization, Investigation, Writing – original draft, Writing – review & editing. **Umay Kiraz:** Providing data, Writing – review. **Tahlita C.M. Zuiverloon:** Providing data, Writing – review. **Rafael Molina:** Conceptualization, Validation, Writing – review & editing, Supervision. **Kjersti Engan:** Investigation, Validation, Writing – review & editing, Supervision, Funding.

Declaration of competing interest

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

Acknowledgments

This research is financially supported by CLARIFY, the Fondo Europeo de Desarrollo Regional (FEDER)/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento *y* Universidades, under Grant B-TIC-324-UGR20 and Project PID2022-140189OB-C22 funded by MCIN/AEI /10.13039/501100011033 and FEDER. CLARIFY is the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska Curie grant agreement No 860627. Miguel López-Pérez was supported by the University of Granada postdoctoral program "Contrato Puente".

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V., Nahavandi, S., 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf. Fusion 76, 243–297.

Albuquerque, T., Moreira, A., Cardoso, J.S., 2021. Deep ordinal focus assessment for Whole Slide Images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 657–663.

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

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

Bancroft, J.D., Gamble, M., 2008. Theory and Practice of Histological Techniques. Elsevier health sciences.

Bautista, P.A., Yagi, Y., 2009. Detection of tissue folds in Whole Slide Images. In: Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009. IEEE, pp. 3669–3672.

Bindhu, P.R., Krishnapillai, R., Thomas, P., Jayanthi, P., 2013. Facts in artifacts. J. Oral Maxillofac. Pathol. 17 (3), 397–401.

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

Campanella, G., Rajanna, A.R., Corsale, L., Schüffler, P.J., Yagi, Y., Fuchs, T.J., 2018. Towards machine learned quality control: A benchmark for sharpness quantification in digital pathology. Comput. Med. Imaging Graph. 65, 142–151.

Chicco, D., Jurman, G., 2020. The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom. 21 (1), 1–13.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1800–1807.

Colling, R., Pitman, H., Oien, K., Rajpoot, N., Macklin, P., in Histopathology Working Group, C.-P.A., Bachtiar, V., Booth, R., Bryant, A., Bull, J., et al., 2019. Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice. J. Pathol. 249 (2), 143–150.

Del Amor, R., Silva-Rodríguez, J., Naranjo, V., 2023. Labeling confidence for uncertainty-aware histology image classification. Comput. Med. Imaging Graph. 107, 102231.

Del Barrio, E., Cuesta-Albertos, J.A., Matrán, C., 2018. An optimal transportation approach for assessing almost stochastic order. In: The Mathematics of the Uncertain. Springer, pp. 33–44.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255.

Dror, R., Shlomov, S., Reichart, R., 2019. Deep dominance - how to properly compare deep neural models. In: Korhonen, A., Traum, D.R., Màrquez, L. (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics, pp. 2773–2785.

Esteban, Á.E., López-Pérez, M., Colomer, A., Sales, M.A., Molina, R., Naranjo, V., 2019. A new optical density granulometry-based descriptor for the classification of prostate histological images using shallow and deep Gaussian processes. Comput. Methods Programs Biomed. 178, 303–317.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. PMLR, pp. 1050–1059.

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

Gardner, J.R., Pleiss, G., Bindel, D., Weinberger, K.Q., Wilson, A.G., 2018. GPyTorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In: Advances in Neural Information Processing Systems, Vol. 31.

Hashimoto, N., Bautista, P.A., Yamaguchi, M., Ohyama, N., Yagi, Y., 2012. Referenceless image quality evaluation for whole slide imaging. J. Pathol. Inform. 3.

Haußmann, M., Hamprecht, F.A., Kandemir, M., 2017. Variational bayesian multiple instance learning with gaussian processes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6570–6579.

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

Hensman, J., Matthews, A., Ghahramani, Z., 2015. Scalable variational Gaussian process classification. In: Lebanon, G., Vishwanathan, S.V.N. (Eds.), Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics. In: Proceedings of Machine Learning Research, vol. 38, PMLR, San Diego, California, USA, pp. 351–360.

Ho, D.J., Yarlagadda, D.V., D'Alfonso, T.M., Hanna, M.G., Grabenstetter, A., Ntiamoah, P., Brogi, E., Tan, L.K., Fuchs, T.J., 2021. Deep multi-magnification networks for multi-class breast cancer image segmentation. Comput. Med. Imaging Graph. 88, 101866.

Hosseini, M.S., Zhang, Y., Plataniotis, K.N., 2019. Encoding visual sensitivity by MaxPol convolution filters for image sharpness assessment. IEEE Trans. Image Process. 28 (9), 4510–4525.

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

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

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

Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M., Madabhushi, A., 2019. HistoQC: an open-source quality control tool for digital pathology slides. JCO Clin. Cancer Inform. 3, 1–7.

Kandemir, M., 2015. Asymmetric transfer learning with deep gaussian processes. In: International Conference on Machine Learning. PMLR, pp. 730–738.

Kanwal, N., Amundsen, R., Hardardottir, H., Janssen, E.A., Engan, K., 2023a. Detection and localization of melanoma skin cancer in histopathological Whole Slide Images. In: 2023 31st European Signal Processing Conference (EUSIPCO). IEEE, pp. 1128–1135.

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

Kanwal, N., Janssen, E.A., Engan, K., 2023. Balancing privacy and progress in artificial intelligence: anonymization in histopathology for biomedical research and education. In: 2023 1st International Conference on Frontiers of Artificial Intelligence, Ethics, and Multidisciplinary Applications (FAIEMA). Springer Singapore, arXiv preprint 2307.09426.

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

Kanwal, N., Trygve, E., Farbod, K., Tahlita CM, Z., Kjersti, E., 2023b. Vision transformers for small histological datasets learned through knowledge distillation. In: Advances in Knowledge Discovery and Data Mining: 27th Pacific-Asia Conference, PAKDD 2023, Osaka, Japan, May 25–28, 2023, Proceedings. Springer, pp. 167–179.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization.

Kohlberger, T., Liu, Y., Moran, M., Chen, P.-H.C., Brown, T., Hipp, J.D., Mermel, C.H., Stumpe, M.C., 2019. Whole-slide image focus quality: Automatic assessment and impact on ai cancer detection. J. Pathol. Inform. 10.

Kothari, S., Phan, J.H., Wang, M.D., 2013. Eliminating tissue-fold artifacts in histopathological whole-slide images for improved image-based prediction of cancer grade. J. Pathol. Inform. 4 (1), 22.

Kwon, Y., Won, J.-H., Kim, B.J., Paik, M.C., 2020. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. Comput. Statist. Data Anal. 142, 106816.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.

Loftus, T.J., Shickel, B., Ruppert, M.M., Balch, J.A., Ozrazgat-Baslanti, T., Tighe, P.J., Efron, P.A., Hogan, W.R., Rashidi, P., Upchurch, Jr., G.R., et al., 2022. Uncertainty-aware deep learning in healthcare: a scoping review. PLoS Digit. Health 1 (8).

López-Pérez, M., Amgad, M., Morales-Álvarez, P., Ruiz, P., Cooper, L.A., Molina, R., Katsaggelos, A.K., 2021. Learning from crowds in digital pathology using scalable variational Gaussian processes. Sci. Rep. 11 (1), 1–9.

Lopez-Perez, M., Garcia, L., Benitez, C., Molina, R., 2021. A contribution to deep learning approaches for automatic classification of volcano-seismic events: Deep Gaussian processes. IEEE Trans. Geosci. Remote Sens. 59 (5), 3875–3890.

López-Pérez, M., Morales-Álvarez, P., Cooper, L.A.D., Molina, R., Katsaggelos, A.K., 2023. Deep Gaussian processes for classification with multiple noisy annotators. Application to breast cancer tissue classification. IEEE Access 11, 6922–6934.

López-Pérez, M., Schmidt, A., Wu, Y., Molina, R., Katsaggelos, A.K., 2022. Deep Gaussian processes for multiple instance learning: Application to CT intracranial hemorrhage detection. Comput. Methods Programs Biomed. 219, 106783.

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

Morales, S., Engan, K., Naranjo, V., 2021. Artificial intelligence in computational pathology–challenges and future directions. Digit. Signal Process. 119, 103196.

National Cancer Institute, 2015. Environmental carcinogens and cancer risk. https://www.cancer.gov/about-cancer/causes-prevention/risk/substances/carcinogens. Accessed on August 31, 2023.

Nguyen, A., Yosinski, J., Clune, J., 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 427–436.

Olsson, H., Kartasalo, K., Mulliqi, N., Capuccini, M., Ruusuvuori, P., Samaratunga, H., Delahunt, B., Lindskog, C., Janssen, E.A., Blilie, A., et al., 2022. Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. Nat. Commun. 13 (1), 1–10.

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

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Vol. 32. Curran Associates, Inc., pp. 8024–8035.

Priego-Torres, B.M., Sanchez-Morillo, D., Fernandez-Granero, M.A., Garcia-Rojo, M., 2020. Automatic segmentation of whole-slide H&E stained breast histopathology images using a deep convolutional neural network architecture. Expert Syst. Appl. 151, 113387.

Rasmussen, C.E., 2003. Gaussian processes in machine learning. In: Summer School on Machine Learning. Springer, pp. 63–71.

Rastogi, V., Puri, N., Arora, S., Kaur, G., Yadav, L., Sharma, R., 2013. Artefacts: a diagnostic dilemma–a review. J. Clin. Diagn. Res.: JCDR 7 (10), 2408.

Rolls, G.O., Farmer, N.J., Hall, J.B., 2008. Artifacts in Histological and Cytological Preparations. Leica Microsystems.

Salvi, M., Acharya, U.R., Molinari, F., Meiburger, K.M., 2021. The impact of pre-and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis. Comput. Biol. Med. 128, 104129.

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

Senaras, C., Niazi, M.K.K., Lozanski, G., Gurcan, M.N., 2018. DeepFocus: detection of out-of-focus regions in whole slide digital images using deep learning. PLoS One 13 (10), e0205387.

Shakhawat, H.M., Nakamura, T., Kimura, F., Yagi, Y., Yamaguchi, M., 2020. Automatic quality evaluation of Whole Slide Images for the practical use of whole slide imaging scanner. ITE Trans. Media Technol. Appl. 8 (4), 252–268.

Shrestha, P., Kneepkens, R., Vrijnsen, J., Vossen, D., Abels, E., Hulsken, B., 2016. A quantitative approach to evaluate image quality of whole slide imaging scanners. J. Pathol. Inform. 7 (1), 56.

Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A., 2022. Cancer statistics, 2022. CA: Cancer J. Clin. 72 (1), 7–33.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (Eds.), ICLR.

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

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826.

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

Taqi, S.A., Sami, S.A., Sami, L.B., Zaki, S.A., 2018. A review of artifacts in histopathology. J. Oral Maxillofac. Pathol.: JOMFP 22 (2), 279.

Toledo-Cortés, S., de la Pava, M., Perdomo, O., González, F.A., 2020. Hybrid deep learning Gaussian process for diabetic retinopathy diagnosis and uncertainty quantification. In: Fu, H., Garvin, M.K., MacGillivray, T., Xu, Y., Zheng, Y. (Eds.), Ophthalmic Medical Image Analysis. Springer International Publishing, Cham, pp. 206–215.

Tomasetti, L., Engan, K., Khanmohammadi, M., Kurz, K.D., 2020. Cnn based segmentation of infarcted regions in acute cerebral stroke patients from computed tomography perfusion imaging. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. Association for Computing Machinery, pp. 1–8.

Urdal, J., Engan, K., Kvikstad, V.r., Janssen, E.A., 2017. Prognostic prediction of histopathological images by local binary patterns and rUSboost. In: 2017 25th European Signal Processing Conference (EUSIPCO). IEEE, pp. 2349–2353.

Wang, Z., Hosseini, M.S., Miles, A., Plataniotis, K.N., Wang, Z., 2020. FocusLiteNN: High efficiency focus quality assessment for digital pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 403–413.

Wetteland, R., Kvikstad, V., Eftestøl, T., Tøssebro, E., Lillesand, M., Janssen, E.A.M., Engan, K., 2021. Automatic diagnostic tool for predicting cancer grade in bladder cancer patients using deep learning. IEEE Access 9, 115813–115825.

William F, L., Mittu, R., A. Sofge, D., Shortell, T., A. McDermott, T., 2021. Systems Engineering and Artificial Intelligence (Chapter 19). Springer.

Williams, C.K., Rasmussen, C.E., 2006. Gaussian Processes for Machine Learning, Vol. 2. MIT Press, Cambridge, MA.

Wilson, A.G., Hu, Z., Salakhutdinov, R., Xing, E.P., 2016a. Deep kernel learning. In: Artificial Intelligence and Statistics. PMLR, pp. 370–378.

Wilson, A.G., Hu, Z., Salakhutdinov, R.R., Xing, E.P., 2016b. Stochastic variational deep kernel learning. Adv. Neural Inf. Process. Syst. 29.

Wright, A.I., Dunn, C.M., Hale, M., Hutchins, G.G., Treanor, D.E., 2020. The effect of quality control on accuracy of digital pathology image analysis. IEEE J. Biomed. Health Inf. 25 (2), 307–314.

Wu, H., Phan, J.H., Bhatia, A.K., Cundiff, C.A., Shehata, B.M., Wang, M.D., 2015. Detection of blur artifacts in histopathological whole-slide images of endomyocardial biopsies. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp. 727–730.

Wu, Y., Schmidt, A., Hernández-Sánchez, E., Molina, R., Katsaggelos, A.K., 2021a. Combining attention-based multiple instance learning and gaussian processes for CT hemorrhage detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 582–591.

Wu, Z., Yang, Y., Fashing, P.A., Tresp, V., 2021b. Uncertainty-aware time-to-event prediction using deep kernel accelerated failure time models. In: Machine Learning for Healthcare Conference. PMLR, pp. 54–79.

Wu, Z., Yang, Y., Gu, J., Tresp, V., 2021c. Quantifying predictive uncertainty in medical image analysis with deep kernel learning. In: 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI). IEEE, pp. 63–72.