Medical Image Analysis 88 (2023) 102885

Contents lists available at ScienceDirect

Medical Image Analysis

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

An aggregation of aggregation methods in computational pathology

Mohsin Bilal [a,d,1], Robert Jewsbury [a,1], Ruoyu Wang [a], Hammam M. AlGhamdi [a], Amina Asif [a], Mark Eastwood [a], Nasir Rajpoot [a,b,c,*]

[a] Tissue Image Analytics Centre, Department of Computer Science, University of Warwick, UK
[b] The Alan Turing Institute, UK
[c] Department of Pathology, University Hospitals Coventry and Warwickshire, UK
[d] School of Computing, National University of Computer and Emerging Sciences, Islamabad, Pakistan

ARTICLE INFO

Keywords:
Computational pathology
Whole slide image analysis
Aggregation of predictions
Machine learning

ABSTRACT

Image analysis and machine learning algorithms operating on multi-gigapixel whole-slide images (WSIs) often process a large number of tiles (sub-images) and require *aggregating* predictions from the tiles in order to predict WSI-level labels. In this paper, we present a review of existing literature on various types of aggregation methods with a view to help guide future research in the area of computational pathology (CPath). We propose a general CPath workflow with three pathways that consider multiple levels and types of data and the nature of computation to analyse WSIs for predictive modelling. We categorize aggregation methods according to the context and representation of the data, features of computational modules and CPath use cases. We compare and contrast different methods based on the principle of multiple instance learning, perhaps the most commonly used aggregation method, covering a wide range of CPath literature. To provide a fair comparison, we consider a specific WSI-level prediction task and compare various aggregation methods for that task. Finally, we conclude with a list of objectives and desirable attributes of aggregation methods in general, pros and cons of the various approaches, some recommendations and possible future directions.

1. Introduction

The emerging area of computational pathology (CPath) involves a broad range of computational methods to analyse digitized images of tissue slides for a wide variety of downstream applications such as clinical decision-making and biomarker analysis (Abels et al., 2019). A high-resolution scan of a routine histology slide of a tissue specimen generates a whole slide image (WSI), often containing several billions of pixels. A WSI is a multi-gigapixel image containing large amount of information-rich pixel data at various levels of details, for instance a large number of various types of cells and glands, tissue phenotypes, and regions of interest to analyse for WSI-level predictions. However, a WSI can often not be processed entirely in graphic processing units (GPUs) for training or inference, presenting a computational challenge in itself. As a remedy, it is common to divide a WSI into multiple image tiles (or patches), perform the analysis on individual patches (or small groups of patches) and *aggregate* the results of inference from various levels into a WSI-level inference. While localization and recognition of objects like cells benefits from low-level details of the WSI data, prediction at the level of image tiles themselves provides slightly better context at the cost of missing low-level details of the cellular objects, a manifestation of the classical position-class uncertainty trade-off (Wilson & Knutsson, 1988). In this survey, we focus on computational approaches to WSI analysis for various different WSI-level tasks including prediction of diagnostic and molecular labels and survival analysis.

There are multiple levels of aggregation in CPath. For example, aggregation of object (say, nuclei) level predictions into an image patch level prediction and aggregation of patch-level predictions into larger tile level prediction, as illustrated in Fig. 1a. In this paper, we focus on methods that aggregate predictions gathered from objects, image patches or tiles into WSI-level predictions. These methods follow the pipeline shown in Fig. 1b. As part of our comprehensive review, we aim to cover all aspects of CPath literature, including data in a WSI, computational approaches, use cases, evaluation, comparison, and recommendations for possible future directions.

Generally speaking, a CPath model's WSI-level inference uses an

[*] Corresponding author at: Tissue Image Analytics Centre, Department of Computer Science, University of Warwick, Coventry, CV47AL, UK.
E-mail address: N.M.Rajpoot@warwick.ac.uk (N. Rajpoot).
[1] Joint first authors.

https://doi.org/10.1016/j.media.2023.102885
Received 30 August 2022; Received in revised form 2 May 2023; Accepted 28 June 2023
Available online 29 June 2023

aggregated score to label a WSI with a correct diagnostic, prognostic or molecular category, making *aggregation* an essential module in several CPath analytical pipelines. We would like to note that our definition of "aggregation" in this paper is not restricted to aggregated scores only. It also refers to aggregating features obtained from various levels, objects and parts of a WSI for predictive modelling. For instance, there can be more than one WSIs for a patient to model CPath solutions, which require aggregating scores from multiple WSIs per case (Chang et al., 2021).

2. Computational pathology workflow

Fig. 1 illustrates whole slide image analysis workflow to approach the predictive modelling solution by processing the WSIs in three different ways. A CPath model obtains image tiles by dividing WSIs and WSI-level scores by aggregating the results of inference on the tiles. The predictive modelling in a CPath pipeline may follow one of the three approaches: bottom-up inference, top-down inference or tissue phenotypic representation based inference. In the bottom-up approach, image patches or tiles are used to detect, segment and classify various tissue objects like cells and glands as the primary units of information that are then subject to aggregation to represent a WSI, see for instance (Diao et al., 2021; Ho et al., 2022; Lu et al., 2020; Park et al., 2022). The WSI representations for predictive modelling offer options from classical machine learning to graph learning (GL) where graph convolutional neural networks (GCNNs) learn and aggregate all information into a single score for clinical decision-making.

The top-down approach begins with analysing WSIs with tile-level or region-level predictions. In its simpler form, it does not require any annotations for specific objects and regions in a WSI, instead a WSI-level label may be used to weakly label the small image tiles. Tile-level scores or deep features can then be used to model predictive analysis, for instance in a multiple instance learning (MIL) setting. Results of inference on tiles are then aggregated into a WSI-level score (Bilal et al., 2021; Coudray et al., 2018; Kather, Pearson, et al., 2019).

There is a third approach in CPath workflows, in which we first learn to distinguish different tiles as tissue phenotypes or regions of interest (labelled) within a WSI to incorporate *apriori* domain knowledge in predictive modelling (Park et al., 2022; Su et al., 2022; Wang et al., 2021; Yamashita et al., 2021), leading to a tissue phenotypic representation of the WSI.

Fig. 2 and Table 1 present a summary of the aggregation methods found in the CPath literature considering three related aspects of the scientific literature: use cases, the type of input data and aggregation methods.

In this paper, we focus on WSI image analysis workflows for the diagnostic, molecular and prognostic survival predictions at the WSI level as three main histopathology use cases. We have also added an aggregation method which used multiple WSIs and generated an aggregated output at the case level. Table 1 provides more details on the methods for these use cases with additional information of cancer type, clinical problem, aggregation method and workflows, and datasets used whether public or private. Diagnostic tasks include primary cancer screening, cancer subtype classification, cancer grade prediction and metastases detection. In the molecular prediction tasks, we include workflows for gene expression prediction, molecular pathways/subtype prediction, mutation prediction and treatment response prediction (Xie et al., 2022). In the prognostic use case, we include survival and risk prediction workflows. In the next section, we describe methods of aggregation in detail.

3. Methods of aggregation

Most CPath approaches process WSI tiles and aggregate the tile-level predicted labels, scores or probabilities to predict the slide-level label. We refer to both the small image patches and large image tiles as *tiles* from hereon. The problem can be formulated as follows: given a WSI Xcomposed of a set or a *bag* of tiles $X = \{x_1, x_2, \ldots, x_n\}$ and their corresponding predictions $y = \{y_1, y_2, \ldots, y_n\}$, the output prediction y_{wsi} is obtained as follows,

$$y_{wsi} = g_\varphi(\{y_i = f_\theta(x_i), \ \forall i\}) \tag{1}$$

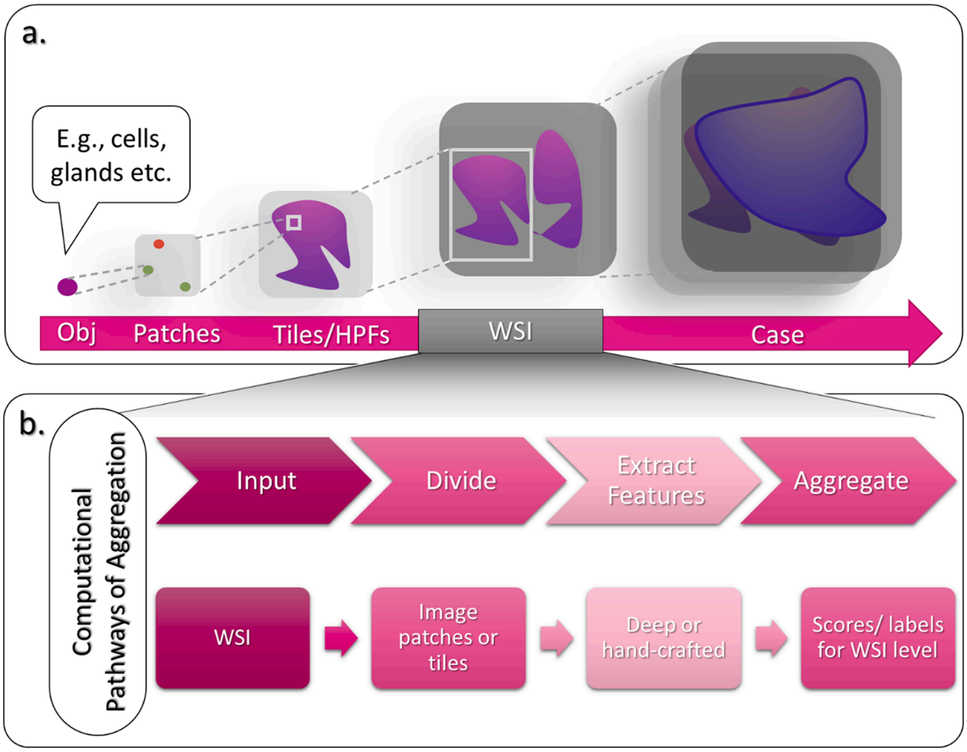

Fig. 1. A general overview of computational pathology data and workflows, a. A WSI contains many cells and glands as objects (Obj.), image patches, tiles or high power fields (HPF). A case may have multiple WSIs. b. Whole slide image analysis workflow here, considers a WSI as input, divided into several image tiles (patches or HPF) for machine learning to compute an aggregated output at WSI-level. Image tiles from a WSI can be processed by machine learning model to infer features; i.e., scores or labels or deep features, to be aggregated by different aggregation methods.

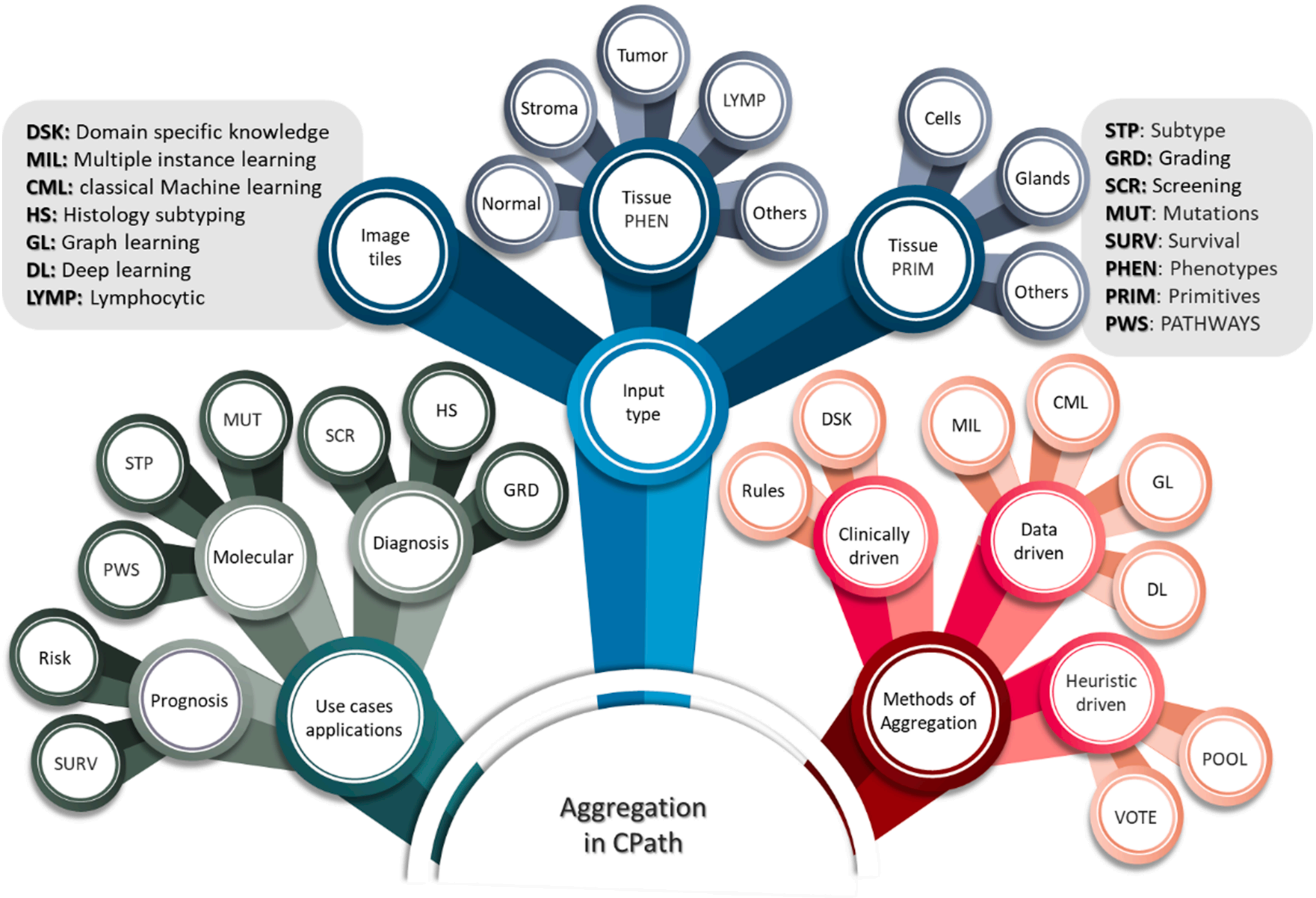

Fig. 2. Aggregating the aggregation literature: use cases, input type and methods of aggregation.

In Eq. (1), f represents a function (usually a neural network) that converts a patch x_n into an instance level probability, score or prediction and is parameterised by θ. The function g is the aggregation function that takes the predictions of $f_\theta(x_i)$, $\forall i$ and combines them into a slide-level prediction. It can have learnable parameters of its own, denoted by φ , or it can be non-parameterised.

Broadly speaking, three types of approaches for defining the function g can be found in the literature, as described in the remainder of this section.

3.1. Heuristic/statistical aggregation

Heuristic approaches also known as global pooling approaches require no machine learning and are not data dependent. Each instance is processed independently and then the scores are aggregated using a fixed formula. All the pooling approaches condense a set of tile-level scores into a slide-level score in some way. There are many different formulations for $g(\cdot)$ but the most commonly used ones are:

- Mean: $g(y) = \frac{1}{n}\sum_{i=1}^{n} y_i$ where n is the number of tiles in the WSI.
- Max: $g(y) = \max(y_i); i = \{1, 2, \ldots, n\}$

Others include Majority voting (mode), Generalised mean (Xu et al., 2014), Noisy-OR (Kraus et al., 2016) and Noisy-AND (Skrede et al., 2020).

3.2. Data-driven aggregation

In data-driven approaches, the function g has a set of learnable parameters φ . The most commonly used approach is a form of attention aggregation proposed by Ilse et al. (2018), which takes a set of feature vectors z extracted from the input bag of instances X using a neural network f parameterised by θ

$$z = f_\theta(X) = \{z_1, z_2, \ldots, z_n\} \tag{2}$$

The bag/slide label is obtained via the following formulation,

$$y_{wsi} = g(z) = \sum_{i=1}^{n} a_i z_i \tag{3}$$

where

$$a_i = \frac{\exp\left\{\underline{w}^T \tanh\left(\underline{V} z_i^T\right)\right\}}{\sum_{j=1}^{n} \exp\left\{\underline{w}^T \tanh\left(\underline{V} z_j^T\right)\right\}} \tag{4}$$

and $\underline{w} \in \mathbb{R}^{L\times\ 1}$and $\underline{V} \in \mathbb{R}^{L\times N}$are learnable parameters. It is worth noting that this formulation assumes all instances are independent.

There are many other formulations that have been proposed including extensions of this attention MIL (Lipkova et al., 2022; Lu et al., 2021), variations that use positional encodings in the attention structure (Shao et al., 2021), ones that use a Recurrent Neural Network (RNN) (Campanella et al., 2019) or a Transformer (Chen et al., 2022) as the aggregation function instead of the attention mechanism. Gildenblat et al. (2021) proposed a data-driven mechanism of certainty pooling in the MIL framework showing better performance than attention mechanism and heuristic aggregation like mean and max pooling.

Other approaches for the data-driven aggregation includes graph learning (Lu et al., 2020) and the application of classification machine learning like random forest, logistic regression, support vector and gradient boosting machines with the handcrafted features of WSI

Table 1
Summary of aggregation methods in CPath.

Authors	Cancer type (s)	Clinical Problem(s)	Aggregation Method
Non-data driven methods			
Schmauch et al. (2020)	28 cancer types	Gene expression	Weighted average pooling
Gildenblat et al. (2021)	Breast cancer	Metastases detection	Certainty Pooling
Skrede et al. (2020)	Colorectal cancer	Cancer specific survival	Noisy And Pooling
Bilal et al. (2021)	Colorectal cancer	Molecular pathways / Mutation prediction	Iterative draw and rank sampling (IDARS), average probability aggregation
Yamashita et al. (2021)	Colorectal cancer	Microsatellite instability prediction	Tissue phenotype classification and average probability aggregation
Bilal et al (2022)	Colorectal cancer	Cancer screening	IDARS, Average (of tiles with probability greater than median) probability aggregation
Kather et al. (2019)	Colorectal / Gastric cancer	Molecular subtypes / Mutation prediction	Naive MIL, Proportion of positive predicted tiles
Su et al. (2022)	Gastric cancer	Microsatellite instability recognition	Majority voting
Kanavati et al. (2020)	Lung cancer	Sub-type classification / Metastases detection	Max pooling
Parameterised methods			
Diao et al. (2021)	5 cancer types	Molecular phenotype prediction	Classical machine learning based aggregation
Chen et al. (2022)	8 cancer types	Sub-type classification / Survival prediction	Hierarchical transformer aggregation
Lipkova et al. (2022)	Allograft rejection	Rejection conditions / Grade prediction	Multitask attention MIL
Hashimoto et al. (2020)	Blood cancer	Sub-type classification	Domain adversarial attention MIL
Lu et al. (2021)	Brain cancer	Sub-type classification	Contrastive and sparse-attention based MIL
Sharma et al. (2021)	Breast / Gastric cancer	Metastasis detection / Celiac prediction	Clustering and Attention MIL
Lu et al. (2021)	Breast / Kidney / Lung cancer	Sub-type classification / Metastases detection	Attention MIL and Instance-level clustering
Li et al. (2021)	Breast / Lung cancer	Sub-type classification / Metastases detection	Dual instance and bag aggregation
Campanella et al. (2019)	Breast / Prostate / Skin cancer	Metastases detection / Grade prediction	Top-k learning, RNN aggregation
Lu et al. (2022)	Breast cancer	Mutation prediction (HER2)	GNN aggregation
Naik et al. (2020)	Breast cancer	Mutation prediction	Attention MIL
Tellez et al. (2019)	Breast cancer	Metastases detection / Tumour proliferation speed	Neural Image Compression
Schirris et al. (2022)	Breast cancer/ Colorectal cancer	Mutation prediction	SSL encoder and VarMIL
Shao et al. (2021)	Breast/ Kidney / Lung cancer	Sub-type classification / Metastases detection	Transformer aggregation
Park et al. (2022)	Colorectal / Gastric cancer	Microsatellite instability prediction	Mean aggregation & light gradient boosting machine
Saillard et al. (2021)	Colorectal / Gastric cancer	Microsatellite instability prediction	SSL encoder, top and bottom scores, and Chowder
Reisenbüchler et al. (2022)	Colorectal / Stomach cancer	Mutation prediction	Local attention graph transformer
Ho et al. (2022)	Colorectal cancer	Cancer detection	Gland segmentation and aggregation by gradient-boosted decision tree
Tomita et al. (2019)	Esophageal cancer	Sub-type classification	Attention MIL
Xie et al. (2022)	Lung cancer	ICI treatment response prediction	End-to-end part-learning GNN
Chang et al. (2021)	Lung cancer	Survival analysis	Hybrid aggregation network
Pinckaers et al. (2020)	Prostate cancer	Grade prediction	Streaming CNN
Zhang et al. (2022)	Rectal cancer	Chemoradiotherapy efficacy prediction	Multi-scale CNN bilinear attention MIL

obtained from the tiles, cells, and glands (Diao et al., 2021; Ho et al., 2022). Handcrafted features take the tile-level prediction scores or probabilities in the same way as the prior two methods but instead of aggregating these to the prediction for the WSI level label directly they extract features from the tile scores such as a histogram of tile prediction scores. Other examples include morphological or statistical features extracted from the prediction maps.

3.3. Clinically-driven aggregation

The final class of aggregation methods use clinical formulae developed by pathologists and currently used in clinical tasks. For example, in breast cancer to categorise the amount of HER2 receptor protein on the surface of cells in the sample, the Royal College of Pathologists (RCPath) guidelines suggest specific criteria as defined in RCPath report, p.99 (Ellis et al., 2016). Other examples of aggregation using domain specific knowledge include mismatch repair assessment to determine microsatellite instability (MSI) status of the sample using the expression of immunohistochemistry (IHC) with 4 antibodies (MMR-IHC: MLH1, MSH2, MSH6 and PMS2) (Awan et al., 2022) and assessment of PD-L1 protein expression using tumour proportion score (Pagni et al., 2020). The latter approach detects and quantifies tumour cells in a WSI in terms of the amount of IHC stain that has been absorbed and then compute the percentage of tumour cells that are strongly stained and compare it with these criteria.

4. Multiple instance learning

Multiple instance learning or MIL is a paradigm of supervised machine learning that deals with incomplete and ambiguous information of labels in training data. The learner receives a set of labelled bags where each bag has multiple unlabelled instances. Dietterich et al. (1997) first coined the term. In its basic form, the MIL problem restricts to a binary classification of bags (Babenko, 2008), but other forms of multiple instance regression (Ray, 2001) and multi-instance multi-label learning (Zhou & Zhang, 2007) can also be found in the machine learning literature. The basic MIL assumption is that every positively labelled bag contains at least one positive "witness" or "key" instance (Babenko, 2008). It implies that all instances in a negatively labelled bag are negative instances. Babenko (2008) presents a thorough review of different MIL methods from classical machine learning for further reading. We have choices of modelling the MIL problem as an instance classifier, a bag classifier (Babenko, 2008) or a combination of instance and bag classifiers.

Predictive modelling in CPath is analogous to the MIL problem if it considers the WSI as a bag with a single label for the purposes of predictive modelling, which is often the case in CPath workflows. Having multiple classes and several labels for each WSI (or bag) is also possible in CPath problems, where the basic assumption of the MIL problem may violate. For some CPath problems, bags may also have instances irrelevant to the prediction task. In other words, both the positive and negative bags may have noisy samples, which are not related to the given label of the WSI. An alternate term of weakly-supervised learning might be more appropriate in such scenarios.

4.1. Multiple instance learning in CPath

In this section, we review recent well-known MIL methods in CPath. Fig. 3 illustrates six recently published MIL methods, which consider a WSI as a bag of patches. We group these methods according to their representation, learning and aggregation modules.

A naive MIL approach fine-tunes all or a few of the last layers of a convolutional neural network (CNN), InceptionV3 in (Coudray et al., 2018) and ResNet18 in (Kather, Pearson, et al., 2019), pre-trained on ImageNet. Each patch gets the same label as the bag (or WSI) during the model's fine-tuning, which means all patches in the positive bag get the positive label. It adds a potential noise in the model training by forcing the network to learn irrelevant or negative instances as positive. The model training follows an instance classifier approach. The model testing applies an average or a majority voting scheme to aggregate probabilities of image patches into scores of a bag or WSI. The majority vote measures a ratio of positively predicted instances over all instances in the bag. Several publications have used naive MIL approach to predict diagnostic and molecular labels of WSIs.

Campanella et al. (2019) proposed an advanced MIL approach for clinical-grade diagnosis of prostate cancer, basal cell carcinoma and breast cancer metastasis. During training, they fine-tune a pretrained CNN (ResNet34) before the CNN is used to predict a score for each tile and then only the top tiles of positive class in each WSI are used for *training*. In the first stage, this pair of *inference* followed by *training* recurs for maximum number of iteration (e.g. 100) to obtain a trained CNN. In the second stage, the trained CNN predicts a few top (e.g. 20) positive tiles from each slide for training an aggregation network. As an aggregation network, they train recurrent neural network on last layer features of selected top tiles and compare it with a random forest trained on the hand-crafted feature of top twenty tiles. This was shown to achieve clinical-grade performance for binary diagnostic tasks when a large number of WSIs were used for training the model.

Clustering-constrained attention multiple instance learning (CLAM) (Lu et al., 2021) is another MIL based aggregation method that represents each WSI as a fixed size bag and each input image patch using fixed features of ResNet50 pre-trained on ImageNet. It performs data-driven aggregation, which involves multi-class attention-based learning to identify associated diagnostic subregions to accurately classify WSI and instance-level clustering over the identified representative regions to constrain and refine the feature space. Several recent studies have employed this method for WSI label predictions.

Bilal et al. (2021) proposed iterative draw and rank sampling (IDaRS) for fine-tuning a CNN (ResNet34) on two smaller subset (random (*r)* and top (*k)*) of tiles from each WSI. After initiating the training with random tiles (50 or 11%), they obtain top *k (5 or 1%)* positive tiles in each subsequent iteration to combine with random (*r*) tiles from each WSI for training. For aggregation, they experiment with several pooling methods like average of positive probability of all tiles or selective tiles (top few (5 or 10), top half (50%)), and weighted average. They report better AUC-ROCs and average precision of PR-curves with average and top half aggregation for molecular (Bilal et al., 2021) and diagnostic (Bilal et al., 2022) labels predictions.

Recently, self-supervised learning based (Saillard et al., 2021), (Schirris et al., 2022) have been proposed for various histopathology tasks. Saillard et al. (2021) use self-supervised learning to fine-tune a pretrained CNN (ResNet50). For each tile, they use an autoencoder to reduce the last layer features of trained ResNet50 to a 256-dimensional vector. They predict microsatellite instability with three different MIL frameworks for aggregation. A better performing aggregation network consists of two multilayer perceptron (MLP) networks. The first MLP processes features to score each tile for the selection of *R* (*R*=10, 25, or 100) top and bottom scores per WSI for the training of second MLP. The second trained MLP processes concatenated top and bottom scores to infer a final aggregated score. In first stage of the DeepSMILE (Schirris et al., 2022), authors use self-supervised learning for fine-tuning a pre-trained CNN (ResNet18 and ShuffleNetV2). In second stage, DeepSMILE proposes MLP-based aggregation network. The aggregation network uses the last layer features of the trained CNN and learns attention followed by a classification module to get an aggregated output for the prediction of MSI and homologous recombination deficiency (HRD).

5. Context in aggregation

In visual processing systems, data-driven modelling requires visual context without losing an appropriate level of finer details. In CPath, processing bags of image tiles without spatial information compromises

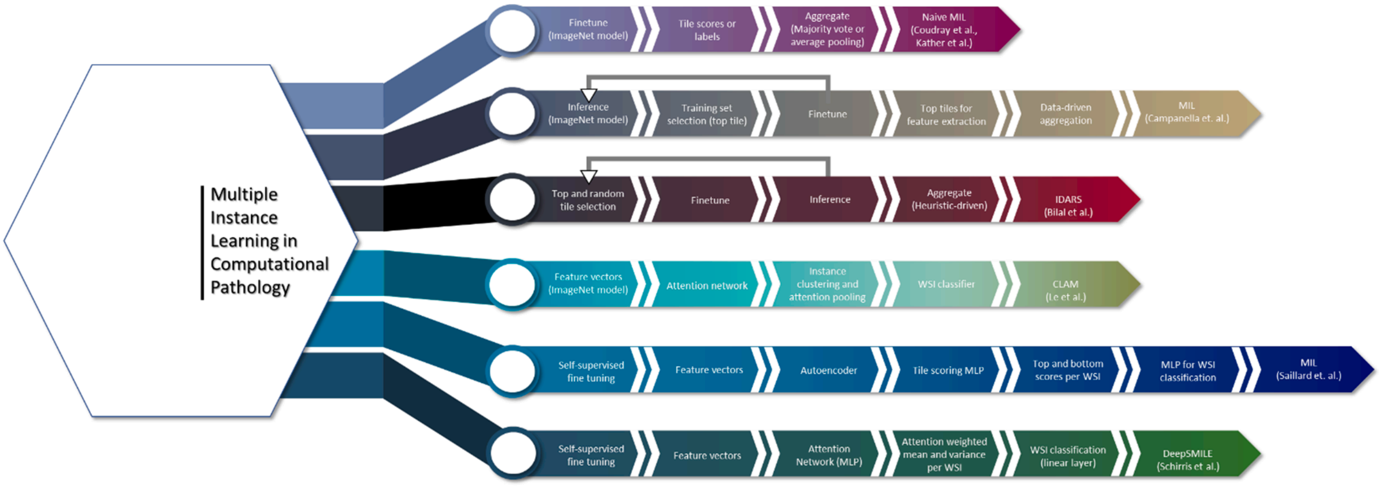
Multiple Instance Learning in Computational Pathology
Finetune (ImageNet model)
Tile scores or labels
Aggregate (Majority vote or average pooling)
Naive MIL (Coudray et al., Kather et al.)
Inference (ImageNet model)
Training set selection (top tile)
Finetune
Top tiles for feature extraction
Data-driven aggregation
MIL (Campanella et. al.)
Top and random tile selection
Finetune
Inference
Aggregate (Heuristic-driven)
IDARS (Bilal et al.)
Feature vectors (ImageNet model)
Attention network
Instance clustering and attention pooling
WSI classifier
CLAM (Le et al.)
Self-supervised fine tuning
Feature vectors
Autoencoder
Tile scoring MLP
Top and bottom scores per WSI
MLP for WSI classification
MIL (Saillard et. al.)
Self-supervised fine tuning
Feature vectors
Attention Network (MLP)
Attention weighted mean and variance per WSI
WSI classification (linear layer)
DeepSMILE (Schirris et al.)

Fig. 3. Popular multiple instance learning pipelines in computational pathology.

the wider visual context. Capturing visual context with constrained computational resources results in losing the finer details and raises a trade-off between visual details and context. If we retain a fuller context at the lowest level of resolution, we will lose essential details of the WSI that impedes predictive modelling. In its trivial form, MIL-based approaches described in Section 4.1 considered patching at an appropriate resolution but with a limited context and without retaining the spatial information of image patches. The aggregation attempts to recover the lost context from disjoint patches in the bag. Yet, this may be a suboptimal solution until we can process an entire WSI as a single sample or retain deep features besides spatial location and awareness of their role in the final prediction. Another approach often found in the literature analyses patches at multiple resolutions, which can be beneficial to capture the heterogeneity of data from multiple regions (Zhang et al., 2022). Next, we describe the global context and graph-based aggregation approaches attempting WSI image analysis with a global context and spatial interaction.

5.1. Global context

Global context aggregation approaches attempt to address the limitations of local context approaches where the narrow field of view considered in the tile aggregation approaches limits the incorporation of global features. The typical tile size in computational pathology problems is 256 × 256 or 512 × 512 pixels. Particularly at higher magnifications, e.g. 20 or 40x, this results in a bag of instances with no consideration for the spatial relationship between the different instances. Efforts to address this problem outside of graph structures are few but varied.

5.1.1. Transformers

The Transformer architecture (Vaswani et al., 2017) has quickly become the state-of-the-art for many language tasks and recently with the introduction of the Vision Transformer (Dosovitskiy et al., 2021) has been applied extensively for computer vision tasks as well. It uses a multi-head self-attention mechanism which unlike RNNs does not consider the order or relative position of tokens in an input sequence. To address this, the transformer uses a positional encoding with each input token adding in the relative position of each component in the sequence. This allows the Transformer to have greater awareness of longer-range dependencies in the input data compared to other models like a CNN's kernels that have a fixed window size. As such it lends itself to address the lack of context problem associated with the local aggregation methods.

To aggregate image tile instances for WSI classification, several methods using Transformers have been used such as SMILE (Lu et al., 2021), TransMIL (Shao et al., 2021), GTP (Zheng et al., 2021) and HIPT (Chen et al., 2022). They operate on a similar paradigm to the local context approaches but attempt to encode in the feature vectors an aspect of the given feature vector's spatial positioning with respect to other feature vectors. The construction of a bag of instances is usually handled in the same way as the local context methods, ie a tissue mask is used to extract only tissue tiles which are then fed through a pre-trained or a fine-tuned encoder, usually a ResNet50. It is after this step that the Transformer approaches differ. SMILE uses a SAM (Sparse-Attention) module to extract the top-N feature embeddings from the bag of instances which is then fed forward into a transformer module. TransMIL uses a Pyramid Position Encoding Generator (PPEG) combined with an alternative Transformer architecture which approximates the self-attention mechanism the Nyströmformer (Xiong et al., 2021). GTP posits that combining ViTs along with graphs can lead to a more efficient approach. They build a graph from the extracted feature vectors and then passes the graph through a GCN and pooling layer before passing this to a transformer layer with the associated positional encodings.

Transformer methods have been shown in ablation studies to have a positive impact on model performance for a variety of tasks including glioma subtyping (Lu et al., 2021), metastasis detection (Shao et al., 2021), lung cancer subtyping (Shao et al., 2021; Zheng et al., 2021) and kidney cancer subtyping (Shao et al., 2021). Although some works (Shao et al., 2021) alter the structure of the positional encoding the majority of Transformer based approaches assume that the existing sinusoidal encoding approach proposed in (Dosovitskiy et al., 2021) is sufficient to include the required global context. While the Transformer proposals show they can outperform other existing approaches such as CLAM (Lu et al., 2021) unfortunately there is not currently a comparison between the different transformer methods on the same tasks(s).

5.1.2. Context-aware methods

To remedy the lack of context present in the bag of instance approaches described above several different methods outside of the Transformers have attempted to consider the spatial arrangement of instances as part of their pipeline.

Neural Image Compression (NIC) (Tellez et al., 2019) uses a representation learning approach. By training a tile encoder with a GAN and re-arranging the extracted feature vectors with the same spatial arrangement as the original tiles in the WSI they are able to compress the original WSI into a format which a CNN can hold in memory. Once extracted for each WSI this compressed image representation is then used to train a standard CNN architecture. This process assumes that by passing the instances through the GAN, the spatial relationships between them are preserved in the deep feature space.

Context-Aware CNN (Shaban et al., 2020) uses a similar idea but instead of just spatially re-arranging the instances in the same way they also use an attention block with the deep feature cube to include an encoding of the spatial context. This feature cube, with the spatial context encoded, is then passed to a classification CNN as in NIC. While Context-Aware CNN was only tested on HPFs of order 10ˆ3 × 10ˆ3 pixels it is very similar to NIC in terms of the overall pipeline and there's nothing preventing the approach from being applied at the WSI level.

Another approach to the same idea is Streaming CNN (Pinckaers et al., 2020). Here the authors attempt to train a CNN with arbitrary input image size by streaming the input image through the model in a series of large tiles and using gradient checkpointing to reduce the memory required to store the activations. By aggregating the gradients over large tile sizes e.g., 4096 or 8192 pixels they are able to train an end-to-end model for WSI classification for regressing to the PAM50 score (a measure for tumour growth) and for classifying metastases in breast cancer without aggregating a bag of instances. Streaming CNNs (Huang et al., 2022) have also shown improved lymph node metastatic detection in gastric cancer.

Sparse Convolutional Context-Aware MIL (SparseConvMIL) by Lerousseau et al. (2021) is a fully differentiable context-aware multiple instance learning paradigm. SparseConvMIL uses a sparse map from tiles embeddings and sparse-input CNN for WSI classification to exploit the spatial relationship of tiles in MIL approach for WSI classification.

The final paradigm is to not just incorporate an awareness of the spatial arrangement of instances but also of their hierarchical relationship as well into the overall pipeline. WSIs and other very high-resolution image formats use an image pyramid structure to store the image at different resolutions so they can display the region required at different magnifications. Similar to multi-resolution approaches one proposal (Jewsbury et al., 2021) splits HPF regions with a quadtree approach to create a bag of instances at different magnifications.

CellMaps (AlGhamdi et al., 2021) is another method for representation of histology images, which uses the cellular density of given image to represent the entire WSI. It can represent various cell types, each of which is corresponded in an image layer. The size of the CellMaps can be compressed to the desired level that the model/algorithm can handle, while all the relevant information is kept intact while reducing the image size. Besides, this representation captures the spatial information of cellular level details from the original image. AlGhamdi et al. (2021) show that the prediction performance is improved when model is trained with the CellMaps representation, comparing with the raw H&E images.

5.2. Graph aggregation

Most of the aggregation approaches we have looked at in other sections, treat instances as entities with no specific relationship beyond belonging to the same slide, ignoring the spatial relationship of patches, cells, or histological entities in the sample. This results in loss of useful information from the spatial relationships and does not account for the importance of context in pathology diagnosis. Pathologists often place a lot of importance on the context in which biological features appear when trying to understand and diagnose a sample.

Graph neural network approaches preserve these spatial relationships, by modelling the tissue as a graph of instances (in Graph terminology, the individual instances are referred to as nodes). This allows context from local neighbourhoods to be used when learning instance scores or representations, and can be used to inform global aggregation by identifying important nodes based on graph structure (Lu et al., 2022).

There are several excellent review articles covering graph neural networks in general (Wu et al., 2021) and GNNs in computational pathology (Ahmedt-Aristizabal et al., 2022), which we refer the reader to for an in-depth survey of GNN techniques. Here, we will present an overview of graph neural networks from the viewpoint of aggregation, focussing on the aspects most relevant in that context.

5.2.1. Graph representation

Let $G = (V, E)$ denote a graph, where V and E are the sets of nodes and edges respectively. Each node $v \in V$ is associated with a feature vector F_v. In the context of CPath, each node v is often a histological entity such as a cell (Wang et al., 2019). It may also be a representation of a tissue region, such as a patch, or a cluster of patches or cells [4,45]. Some methods have also used pixel-based clustering methods such as SLIC to generate the nodes (Pati et al., 2020).The features F_v will describe characteristics of the cell or tissue region. In general, graphs often also have associated edge features F_e, though this is less common in the context of CPath.

In learning instance (node) representations, GNNs aggregate information from a local neighbourhood, as illustrated in Fig. 4a. The way in which this is done differs depending on the type of GNN it is.

A node v in a graph has a local neighbourhood N_v which is the set of all nodes to which it is connected by edges. Aggregation at a representational level usually occurs through messages passed to a node from the nodes in its neighbourhood.

The most general form of a spatial convolutional GNN is:

$$h_v^{(k)} = f_{\theta_k}\left(h_v^{(k-1)}, h_u^{(k-1)} | u \in N_v\right) \tag{5}$$

A typical CGNN will have a small number (usually <10) of such layers, where at each layer the representation at a node aggregates information from a steadily larger region of the graph, as illustrated in Fig. 4b. Depending on the connectivity of the graph and size of the individual instances, this can end up aggregating information from quite a large region of a WSI.

There has been a proliferation of suggestions in the literature for the form of Eq. (5), each trying to incorporate some specific intuition or satisfy some mathematical requirement on how a graph convolutional layer should be.

The EdgeConv graph convolution was introduced in Wang et al. (2019), and focusses on differences between the central node and its neighbours. It has been used for example in HER2 status prediction in [45].

In Xu et al. (2019) the authors show that there are some graph structures that popular GNN variants cannot distinguish between, and propose the Graph Isomorphism network (GIN) to address this. This form allows for a weighting between the importance we place on the information in the central node, compared to the information aggregated from surrounding nodes in its neighbourhood N. It has been applied to breast cancer subtyping in Pati et al. (2020).

Furthering the theme of investigating the expressiveness and representational learning power of graph neural networks, in Corso et al. (2020) the authors identify that commonly used functions to aggregate the messages from a node's neighbourhood fail to distinguish between different message sets. Their proposal to solve this, Principal Neighbourhood Aggregation (PNA) involves using multiple different aggregators, as well as some degree-dependent scaling, to define a message aggregation function which can distinguish between a far larger variety of message sets.

Another popular method is GraphSage [51], a distinguishing feature of which is a sampling of a fixed size set of neighbours during aggregation. This can be especially useful to make learning a highly connected network computationally tractable. GraphSage has been used for cancer grading in Wang et al. (2019); Zhou et al. (2019).

5.2.2. Global graph aggregation

The second level at which a GNN aggregates information occurs at a global level, in a way much more directly analogous to the WSI aggregation covered in other sections. Given a GNN that outputs node-level scores, a 'graph readout' function is used to provide a graph level prediction. As the goal is simply to calculate a single score from a set of instance scores/representations, here we are back in the context of many of the other aggregation methods covered in this survey. We can take the mean, max, use attention, and so on.

A key concept in aggregation is the notion of node importance. The choice of aggregation method reflects our assumptions on which nodes are important to the aggregated prediction. The simplest (and

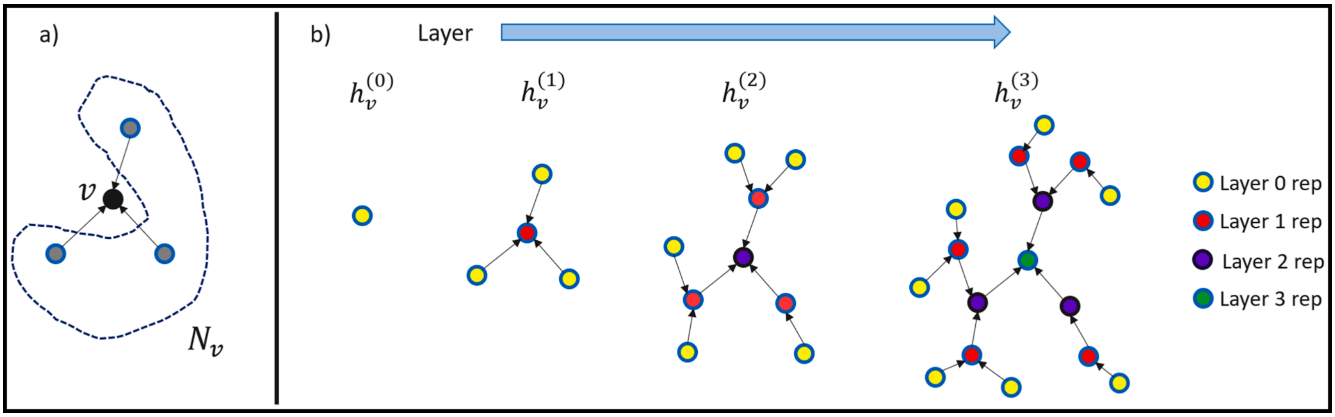

Fig. 4. (a) Local neighbourhood of a node. (b) Aggregation of information over a progressively larger region in successive layers of a spatial GNN.

surprisingly often used) aggregation method is to take the mean of the instance scores. In this case, we are implicitly assuming all instances are equally important. Many global aggregation strategies incorporate an estimate of node importance in the form of weighting or selection on the instances, either based on the instance score (e.g max/top N aggregation), or based on some representation of the instance (e.g attention based methods).

The spatial structure and relationships built into a graph representation can help us with this, by allowing us to define topological measures of importance on the graph structure. There has been much work in the literature on metrics to quantify the structural importance of a node in a graph, and for an in-depth review we refer the reader to Lalou et al. (2018); Landherr et al. (2010). The most common approaches to global aggregation in CPath thus far have been simple aggregators such as the mean, as used in [45], [4], or attention-based approaches such as that in [48]. Another example of a learnable global aggregation is in [49], where node level representations output by a GNN are simply concatenated, and an MLP with learnable weights produces the global graph score.

5.2.3. *Graphs in computational pathology*

Here we will look in brief at a few examples of graph based approaches in CPath which illustrate the information aggregation aspects of graph-based methods particularly well. For a more general review of graph-based methods in CPath, we refer readers to [47].

Graph approaches in CPath usually consist of the following steps. 1. Entity detection/definition, 2. Feature embedding, 3. Graph construction, 4. GNN model training and prediction, 5. Model interpretation (for example using GNNExplainer (Ying et al., 2019)).

A good example of the different levels at which a graph based approach can aggregate information can be found in the SlideGraph+ [45] approach to building graphs for HER2 status prediction on WSIs. In this approach, the basic entities are image patches, which are represented by resnet50 (imagenet pretrained) features. Patches are clustered based on proximity in both position and feature space to form the graph nodes, which have the mean position/feature representation of the patches in the cluster. A further stage of aggregation then occurs during the learning of the graph representation of the node, and depending on if problem is a node or whole graph prediction task this could undergo a final stage of global aggregation such as a mean of node scores. The information in the slide is aggregated in stages: patches -> patch clusters -> graph aggregation over neighbouring patch clusters (nodes) -> global aggregation of node scores into a slide score.

A similar approach is Hact-Net, applied to both breast cancer subtyping (Pati et al., 2020), and gleason grading (Anklin et al., 2021). In this method, tissue regions are generated using a SLIC super-pixel approach. Texture features in these regions are combined with features from a cell-graph embedding of the cells in each region. Then a larger tissue graph is constructed using the tissue regions as the nodes.

Again, this is an excellent example of aggregation at multiple levels: Cells and the surrounding tissue pixels are aggregated into superpixels, in which one level of graph aggregation takes place. The superpixels are in turn used to build a graph, in which aggregation at a higher level occurs.

In other examples of graph-based approaches in CPath, in Wang et al. (2019) cell graphs using morphological and local textural features are used for scoring of prostate cancer Tissue Micro-array (TMA) cores. Another grading application on, colorectal cancer, can be found in Zhou et al. (2019). A cell graph built on larger images is achieved by subsampling the detected cells in an image in a way that produces a representative sampling of nuclei across the image, from which a cell graph is then built.

6. Evaluation and comparative analysis

There are some challenges in CPath that pose the generalizability and verifiability crises. These challenges make it difficult for a single aggregation method to outperform the other methods in all CPath problems. Below we describe the two problems:

(1) Generalizability: In routine clinical practice, tissue block is sectioned at multiple levels which are then mounted onto multiple slides. However, researchers in CPath are usually provided with one or two WSIs for each case. A major challenge is whether the data being provided (i.e., 1 or 2 images per case) contains the information needed for the downstream analysis. It is likely that the amount of information varies when providing different cohort for the same problem, due to the fact that the provided slide (or two) for each case does not necessarily always comprise the same information. The likelihood of such data variability is even higher when the tumour size is large, as the number of slides that can be sectioned is larger.

(2) Verifiability: One of the objectives of the utilisation of ML techniques is to find novel biological insights, which has led some researchers to provide heatmaps showing some tissue regions playing significant roles in the downstream analysis. However, verifying whether the network/algorithm always picks the regions that are actually the causation of the problem being studied remains a challenge. It is possible that the “hot” tissue regions are just meaningless signals that correlated with the trained ground truth. One may argue that it is easy to identify whether the “hot” regions picked by the network are meaningful signals that correlated with the ground truth. This is true with problems like tumour region classification. In contrast, it is challenging when the aim of the study is to predict survival or molecular subtypes. In the absence of localized ground truth (i.e., particularly which tissue regions are directly the actual signals to some CPath problems), verifiability remains a challenge in this domain.

These challenges make it difficult to suggest a standardized solution for the CPath problems. Each problem with its own data can be analyzed with a specific computational solution, including the aggregation method. In the literature, there are a few studies that presented a comparison between different aggregation methods. For example, Laleh et al. (2022) compare the performance of several aggregation methods, including MIL-based methods and simple weakly-supervised (WS) methods. These methods are tested/evaluated against six different CPath problems. Their results show that the simple patch-based methods outperform the MIL-based methods.

In contrast, Zeng et al. (2022) present an opposite finding. The simple patch-based methods show the worse performance, comparing with the MIL-based methods. Such a contradiction proves the generalizability crisis in the CPath, where it is challenging for a single method to generally outperform in all problems and experimental settings. It is likely that for each CPath problem, a different aggregation method is more suited, based on its assumption and experimental setups (e.g., amount of available data). The top method for that problem might fail when the data or the problem is a different one.

Primarily, the experimental analysis evaluates the overall performance of the prediction accuracy, the generalization, and verification of the prediction for the downstream prediction tasks, as performed by Laleh et al. (2022). We identify aggregation as an essential part of CPath applications. The goal of the aggregation method is to combine all the processed information available in a WSI into a final score or category. We argue that factors like aggregation, the type, nature, and amount of input data, features and the underlying machine learning approach impact the overall performance of slide-level predictions. Laleh et al. (2022) partially considered this part of the comparative evaluation in their benchmarking study. However, they did not make a fair comparative analysis of different aggregation methods in the MIL-based setting. Therefore, we conduct a case study of head-to-head comparative analysis of various benchmark aggregation methods.

6.1. Case study: Fair comparative analysis of popular aggregation methods

It is crucial to conduct a fair head-to-head comparison to evaluate the performance of different aggregation methods by fixing all the components within the pipeline apart from the aggregation method they choose. Laleh et al. (2022) conducted comprehensive experiments for benchmarking different end-to-end CPath pipeline. However, their results might not be appropriate for comparing different aggregation methods. In their paper, ResNet (He et al., 2016), EfficientNet (Tan & Le, 2021) and ViT (Dosovitskiy et al., 2021) were all pre-trained on ImageNet and then fine-tuned on the target datasets, while MIL (Campanella et al., 2019; Dietterich et al., 1997), Attention MIL (Ilse et al., 2018) and CLAM (Lu et al., 2021) used a ResNet feature extractor, which was only pre-trained on ImageNet. Secondly, different methods in their work used different backbone networks. The ResNet approach in their paper used ResNet-18 structure, EfficientNet used efficientnet-b7 structure, ViT used vision transformer and the MIL, Attention MIL and CLAM used ResNet-50 structure. Moreover, based on the published code for CLAM, feature extraction uses the features extracted after the third ResNet block, whereas the standard procedure is to use the features generated before the final classification layer. There are so many variables in existing approaches that had not been controlled and may introduce bias into the comparisons. Therefore, we conduct a case study where we attempt to fairly compare different aggregation methods. Moreover, our case study shows that there are many components within an end-to-end framework which can impact the overall performance of the algorithm, sometimes even more than the aggregation method does.

6.1.1. HPV infection prediction in head and neck cancers

We chose the problem of Human papillomavirus (HPV) infection status prediction in head and neck cancers as a case study for comparing some popular aggregation methods. HPV infection status is an important biomarker in head and neck cancers which can affect the prognosis, survival and the treatment selection (Fig. 5).

While immunohistochemistry and PCR are the gold standard for HPV infection diagnosis in the clinical practice, there have recently been a few attempts in the CPath community to solve this problem from analysing digital H&E slides (Kather, Schulte, et al., 2019; Klein et al., 2021).

We believe this problem is ideal as a case study for comparing aggregation methods for several reasons. First of all, like other CPath problems, it requires aggregating predictions from tiles extracted from WSIs. Secondly, clinicians have categorised many histological differences between HPV+ and HPV- H&E slides (Westra, 2012). However, none of these are distinct enough to become a gold standard for human pathologists to reach a diagnosis. It is, therefore, quite possible that the data we provide to the algorithm contains information to enable discrimination between two types of carcinomas, as well as posing some challenges.

6.1.2. Experimental Settings

The dataset we used was retrieved from the Head and Neck Squamous Cell Carcinoma cohort of The Cancer Genome Atlas (TCGA) project. The study by Campbell et al. (2018) provides us with the HPV infection status for these patients. We extracted patches of size 256 × 256 at 10 × magnification from tissue regions for each slide. In total there are 412 cases with 364 HPV- and 48 HPV+, this corresponds to 1, 028,288 negative patches and 120,685 positive patches for a total of 1, 148,973 patches; the fact that this problem is highly imbalanced adds an additional challenge to the task for evaluating the different methods. Three-fold cross validation experiments were conducted where the dataset was split into 3 folds of equal size randomly while stratifying so the class distribution was the same in each fold. The folds were created at the patient level so if a patient had multiple slides they were not included in different folds. The random split was saved and used for each experiment to ensure a fair comparison. In each fold models were fitted on the training set, while the validation set was used for early stopping and model selection. The mean and standard deviation of the AUC-ROC and AUC-PR were reported over the test set experiments as the metrics for evaluation.

We chose 7 different aggregation methods for our case study, which are majority voting, mean pooling, max pooling, median pooling, attention MIL (Ilse et al., 2018), CLAM (Lu et al., 2021) and a context aware GNN based approach (Lu et al., 2022). We selected four different backbone networks for feature extraction, ResNet50 (He et al., 2016), DenseNet161 (Huang et al., 2017), EfficientNet-b5 (Tan & Le, 2019), ConvNeXt-tiny (Liu et al., 2022) These were chosen as they represent the changes and progression of CNN architectures over the last 7 years and have approximately the same number of trainable parameters. For our experiments, we used each network pre-trained on ImageNet as a feature extractor and also trained each one in a weakly supervised manner on the training set from the corresponding fold to learn domain specific, finetuned features. All weakly supervised networks were trained for 50 epochs with a batch size of 64, learning rate of 3e-3, stochastic gradient descent with momentum=0.9, weight decay=1e-4. The AUC-PR on the validation set was used to select the final, finetuned network.

For majority voting, mean, max and median pooling, each of these pooling methods was tested for generating a WSI-level score from tile-level features, a model was trained using ranking loss (Wang et al., 2022) to output a probability for the positive class for each tile for the given case and then these probabilities were pooled using the given heuristics. Gated attention module was used for attention MIL and single-attention-branch model was used for CLAM. For the GNN based approach from each slide we created a slide level graph according to the scheme defined in (Lu et al., 2022) and used each feature extractor to generate the node features from the corresponding patch in the slide. We used a 4-layer GNN consisting of a linear embedding layer, 2 edge convolution layers and then a linear layer to classify the concatenated features from the other layers to compute the final slide level score. For all aggregation methods we set the internal embedding dimension to be equal to half of the input feature size for consistency.

Seventeen patients in the cohort had multiple slides. For these patients in the GNN approach we created a graph where each slide was represented by a subgraph in the overall graph, I.e., if a patient had 2 slides each with 1 piece of tissue, then their corresponding graph had 2 subgraphs representing each slide. For all other methods we created one overall bag of instances containing the extracted features from all patches found in that case's slides.

We used the following hyperparameters for the trainable components of all aggregation methods we tested in this study. We used stochastic gradient descent with a loss rate of 2e-4, weight decay was set to 1e-5 and momentum=0.9, gamma=0.1. All models were trained for a maximum of 200 epochs with early stopping if the validation loss did not improve for 20 epochs after the 50th epoch. These were chosen as they were the original CLAM paper's hyperparameters and in our experiments they resulted in all models converging to at least a good local optimum. All other hyperparameters used by CLAM were kept the same as the original paper's implementation.

6.1.3. Benchmark performance metrics

The area under the receiver operating characteristic (AUC-ROC) and area under the precision recall curve (AUC-PR) were chosen to be the quantitative metrics for evaluating the performance of different aggregation methods. They are both widely used in machine learning research to evaluate the performance of classifiers. The ROC curve is plotted with the true-positive rates and the false-positive rates generated by varying the classification threshold while the precision-recall curve is plotted with precision and recall values. The AUC-ROC is a quantitative metric for describing the ROC curve with a higher AUC-ROC score indicating better classification performance. Similarly, the AUC-PR is a quantitative metric for describing the performance of an algorithm in terms of

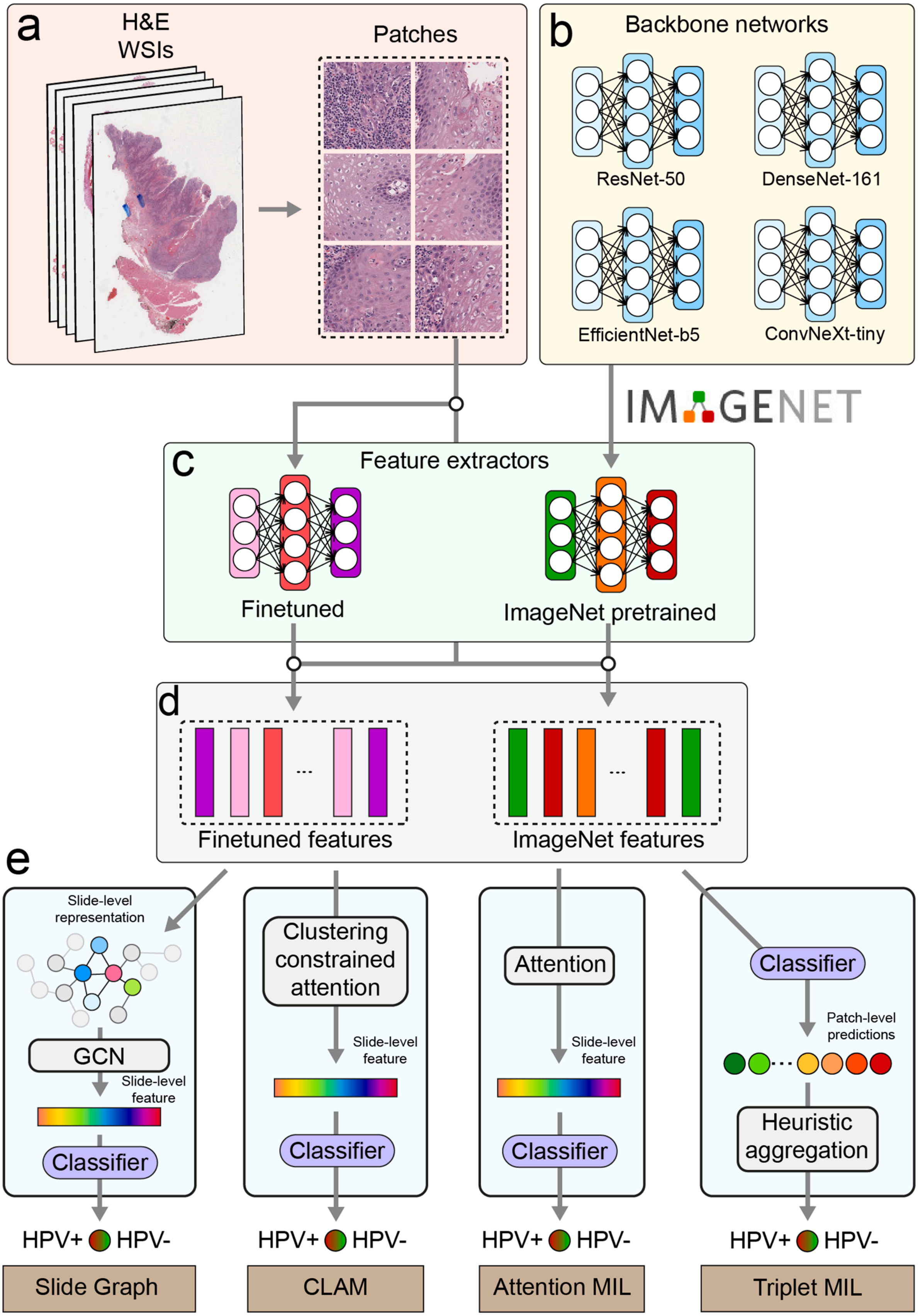

Fig. 5. Overview of the pipeline used for our case study into HPV prediction. (a) Tissue patches and their spatial coordinates are extracted from the tissue regions of a WSI. (b) Backbone network architectures are selected for evaluation. (c) Feature extractors are defined. chosen network architectures are finetuned using weakly supervised learning on extracted patches from training data in addition to using imagenet-pretrained weights of the same architectures. (d) Patches are encoded into feature representations. (e) Corresponding aggregation methods are trained, and slide-level performance is assessed.

precision and recall with a higher AUC-PR indicating better performance. The ROC curve is used to analyse the clinical sensitivity and specificity of a proposed algorithm, while the PR curve is particularly useful for evaluating algorithms in scenarios where there is a severe class imbalance, as is the case in the problem we have chosen here. Both are widely used in medical research to assess the diagnostic accuracy of proposed machine learning algorithms.

6.1.4. Results

To illustrate the performance effect of different aggregation methods for this problem, Table 2 lists the 7 different aggregation methods and their corresponding results in terms of AUROC and AUC-PR values. Additionally, we include the inference time for each method and the number of trainable parameters to illustrate the computational complexity and memory requirement of each method. They show that in terms of AUC-ROC and AUC-PR the aggregation methods have similar levels of performance with the parameterised methods having slight performance gains in terms of AUC-PR on average at the cost of additional inference time and model complexity. This trend was also observed for the other 3 backbone networks tested and also with ImageNet features instead of finetuned features.

To show the effect of the backbone network architecture, Fig. 6 shows the AUC-ROC and AUC-PR curves for all 4 network architectures using average pooling in Fig. 6a (a heuristic approach) and CLAM (a parameterised approach) in Fig. 6b. It shows that all network architectures have approximately similar performance. While there are slight differences for a given network across aggregation approaches, e.g., the ConvNeXt features performs the best with CLAM while for average pooling it is the DenseNet features, there is no single architecture that universally performs best across all the models. We observed that with ImageNet features some aggregation approaches would have outlier backbone networks with significantly reduced performance compared to the others; however, this was not seen for any approach when finetuned features were used.

Finally, Table 3 presents a comparison with respect to the features used by the aggregation approaches. We see that for all aggregation methods finetuned features lead to improved performance, up to 0.22 AUC-PR. This was observed for all methods for all backbone architectures. This provides evidence that in terms of performance the instance level features used are the most important aspect of the overall CPath pipeline. While ImageNet features can still achieve good performance, particularly with CLAM, this indicates that even weakly supervised finetuning allows the backbone networks to extract domain specific features that are useful for this downstream task of HPV status prediction.

Table 2
Results comparison between different aggregation approaches using finetuned ResNet50 features. Inference time is the mean for a single fold ± standard deviation. Bold values indicate best performance or lowest complexity (in time and parameter count), underlined values indicate second best.

	Aggregation	AUC-ROC	AUC-PR	Inference time (s)	Parameters
Heuristic	Majority vote	0.824 ±0.038	0.489 ±0.043	9.683 ±1.334	**2,099,201**
	Average pooling	**0.872 ±0.030**	0.521 ±0.065	9.672 ±1.334	**2,099,201**
	Max pooling	0.758 ±0.034	0.294 ±0.059	**9.670 ±1.334**	**2,099,201**
	Median pooling	0.870 ±0.033	0.533 ±0.064	9.687 ±1.342	**2,099,201**
Parameterised	Attention **MIL**	0.870 ±0.035	0.522 ±0.063	14.377 ±1.108	4,199,426
	CLAM	0.871 ±0.033	**0.544 ±0.034**	22.501 ±11.711	4,204,551
	Graph	0.863 ±0.012	0.504 ±0.039	17.269 ±3.125	9,448,451

Our experiments show that some components within the CPath pipeline other than the aggregation method can have a greater impact on the final performance in this problem. Therefore, we believe researchers should be extremely careful when reaching a conclusion that one aggregation method is better than another. For this problem the choice of aggregation method should be tailored towards what downstream capabilities are required. The instance level features have been shown to be by far the most important aspect with respect to downstream performance for this problem.

7. Discussion

We analysed a significant variety of computational pipelines for predictive modelling in CPath and grouped them in terms of data and computational frameworks. CPath offers data with different levels of details and contextual information, from pixel to patch and cell or gland to tissue phenotypes. Consequently, the computational frameworks originated in relation to the scheme of representing different levels and types of data and the contextual information, e.g., from individual patches or cells to their connectivity through graphs. Besides, the output aggregation methodologies came in as simple pooling, data-driven like machine learning, and clinical rules. In addition, the computational resources, time, and related costs have their impacts on the modelling.

Each of these components of computation pipelines contributes to the success within a given problem frame, but none of it comes without challenges or trade-offs and can be called a single best solution for all CPath problems. It's critical to identify the input and the main goal, design the best solution for the given problem, and define metrics for the success. The significant most metrics are the predictive performance and interpretability or explainability of the predictions. In terms of predictive performance, e.g., in the diagnostic applications the ultimate goal may be a performance comparable to existing systems i.e., clinical practice, and even further improvement through an objective analysis. Both these, the predictability and explainability, are open research problems so far and offer multidisciplinary contributions to impactful solutions and novel insights.

The case study conducted in this paper validates the above argument. Our case study explicitly shows that there are at least two aspects in a CPath pipeline other than the aggregation method. These are different backbone networks and pretraining approaches for feature extractor, which can have an impact on the overall performance of the aggregation workflow. In a fair comparative analysis of aggregation method, it is essential to control all the variables of experiments other than the aggregation method chosen. For those who want to choose the best aggregation method for their research, it is important to consider the pipeline, the clinical explainability and the problem in question. This is because the performance of different aggregation methods varies depending on many aspects, making it challenging to determine the most optimal one. More importantly, we believe all CPath researchers should also bear in mind that the quantitative metrics are not the only thing to pursue. Instead, the generalizability, verifiability of the CPath approach, the biological interpretability and explainability, the ability to generate localised predictions and novel insights into the aetiology of diseases might be more important than a high accuracy. Asif et al. (2021), in their review article, urged the need of rigorous testing of AI model as one of CPath challenges and limitations associated with the AI development lifecycle.

Our case study shows (both in Tables 2 and 3) that fine-tuning resulted in better performance than the ImageNet pre-trained features. It allowed both pooling and data-driven aggregation approaches to achieve broadly the same performance. CLAM (Lu et al., 2021) relies only on transfer learning and data-driven aggregation. The authors choose to build a data- and resource-efficient pipeline by excluding fine-tuning and using less than 100% data for training efficiency in some experiments. Though, the feature extraction also has additional costs in extracting and storing feature vectors of all the patches. Their

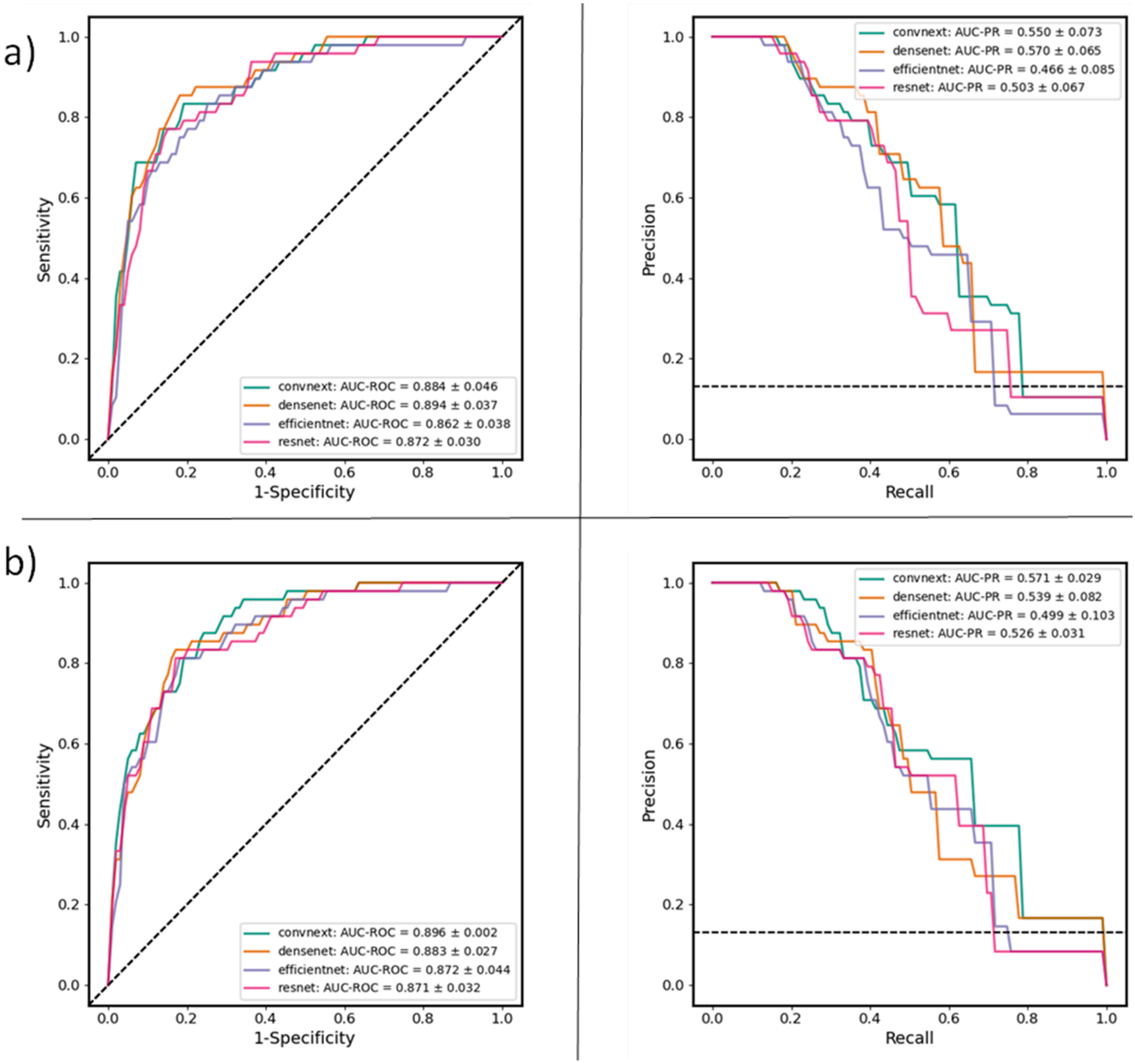

Fig. 6. ROC and PR curves for all the backbone networks assessed. (a) ROC and PR curves for the average pooling aggregation method (heuristic). (b) Curves for the CLAM aggregation approach (parameterised). Dotted black lines indicate performance of a random classifier. All curves shown are using the corresponding network with finetuned features.

Table 3
Results comparison between aggregation approaches using ImageNet pre-trained and finetuned DenseNet161 features. Bold values indicate best performance, underlined values indicate second best.

	AUC-ROC			AUC-PR		
	ImageNet	Finetuned	Difference	ImageNet	Finetuned	Difference
Majority vote	0.827±0.037	0.839±0.053	0.0120	0.444±0.037	0.539±0.111	0.095
Average pooling	0.814±0.039	**0.894±0.037**	0.0802	0.453±0.023	0.584±0.063	0.131
Max pooling	0.780±0.067	0.821±0.038	0.0418	0.422±0.144	0.521±0.076	0.099
Median pooling	0.787±0.024	0.890±0.043	**0.1033**	0.417±0.008	**0.588±0.081**	0.171
Attention MIL	0.790±0.034	0.888±0.034	0.0977	0.341±0.04	0.561±0.051	**0.221**
CLAM	**0.852±0.029**	0.883±0.027	0.0309	**0.504±0.089**	0.558±0.072	0.054
Graph	0.818±0.018	0.878±0.033	0.0612	0.387±0.063	0.562±0.047	0.175

aggregation method has the least difference between using fine-tuned features and pre-trained features. However, it requires further improvement to make the fine-tuning non-essential for aggregation in CPath as compared to RankMIL workflow(Wang et al., 2021) as shown in our case study and a naïve MIL workflow for six different benchmark problems in (Laleh et al., 2022).

Several bottom-up workflows for WSI level prediction have been developed recently. A major advantage of bottom-up approach over the top-down approach is its better explainability and interpretability. Predictive modelling of such methods requires detailed cell and region-level annotations but often work well with small amount of data as compared to top-down modelling. Diao et al. (2021) have demonstrated use of human interpretable features to predict diverse molecular signatures (AUC-ROC 0.601–0.864), including expression of four immune checkpoint proteins and homologous recombination deficiency, with performance comparable to but not better than top-down approaches as found in our comparative analysis with slideGraph. Ho et al. (2022) got similar findings with their gland segmentation based bottom-up approach for the screening of colorectal cancer. Yamashita et al. (2021) has demonstrated better prediction of MSI status than naïve MIL top-down approach (Kather et al., 2019). For slide-level MSI prediction, Park et al. (2022) have compared a top-down approach with a bottom-up approach in which they have combined features from multiple objects and levels of inputs including tissue phenotypic, cells, and glands. The top-down approach produced better AUC-ROC scores whereas the bottom-up produced explainable features to verify differentiating features with expert knowledge.

AI algorithms are prone to biases particularly introducing positive bias when developed and validated in siloed manners that results in deteriorated performances on external cohorts revealing generalisation deficiencies. This commonly occurs as developers have control on establishing validation cohorts and readout experiments. Therefore, it is crucial to evaluate the generalisation of AI algorithm independently across different patient populations, pathology labs, digital pathology scanners, reference standards derived from global panel (Bulten et al., 2022). Data bias, quality and reproducibility of the results are also key challenges in the AI development life cycle (Asif et al., 2021).

Global AI competitions have been an effective approach to overcome the pitfalls of soiled development by crowd sourcing the development of the performant algorithms. These competitions can also overcome generalisation issues if they implement an independent evaluation appropriately, such as in a recent Prostate cANcer graDe Assessment (PANDA) competition (Bulten et al., 2022), which is a single largest competition in pathology to date. They were able to fully-reproduce top-performing 15 algorithms and externally validated their generalisation to independent US and EU cohorts and compared them with the reviews pf pathologists.

In PANDA challenge, a total of 1,010 teams, consisting of 1,290 developers from 65 countries participated and submitted at-least one algorithm of total 34,262 versions. The winner and the most leading teams, adopted an aggregation approach in which a sample of smaller tiles are processed by CNNs and predictions are concatenated in the final classification at WSI-level, without requiring any detailed region/pixel-level annotations. Two of the exciting findings include the similar to and higher statistically significant agreement of algorithm with the uropathologists and higher sensitivities for tumour identification than representative pathologists on external validation subsets of both EU and US, respectively.

All three workflows alongside different aggregation approaches have advantages and challenges associated with them. The bottom-up approaches proposed in (Diao et al., 2021; Park et al., 2022) combined tissue phenotype based workflow (third workflow in Fig. 2) with objects (cells and glans) level workflows, which is likely to be explored further in subsequent studies as its potential and comparative advantages are unconclusive. The emerging trends in CPath combine MIL, attention/-transformer mechanisms, graph representation, and learning for better accuracy, generalization, and interpretation of data for various clinical applications (Bilal et al., 2022; Bilal et al., 2022; Chen et al., 2022; Guan et al., 2022; Javed et al., 2022; Kosaraju et al., 9AD; Zheng et al., 2022). Next-generation CPath workflows could have priority research problems concerning transparency and interpretability of predictions in MIL/top-down workflows, versus clinical-grade/better predictability in interpretable bottom-up workflows.

Other factors may include data efficiency, which assesses the amount of data and the labelled data needed for robust and efficient machine learning. The efficiency of computational resources, the hardware, and the turnaround time required to reach the final WSI-level prediction becomes important when implemented in a real-world setting. To accelerate predictive modelling for CPath solutions to the next level may also require considering the notion of learning paradigm, e.g., end-to-end learning and self/unsupervised learning and new ways of modelling attention mechanism and data loading pipeline for selection of most relevant image tiles for training. It is, however, expected that future research and development will produce more data and allow a broader evaluation to rank approaches for different CPath use cases and performance criteria.

8. Conclusions

In conclusion, there is no one-size-fits-all solution when it comes to choosing an appropriate aggregation method. The choice of method depends on several factors including the input data, problem-solving goal, success metrics, and computational efficiency. Our case study showed that feature choice via fine-tuning had a greater impact on performance than aggregation method for the task of HPV status prediction. All aggregation methods with all backbone networks tested proved to be more effective with fine-tuned features compared to ImageNet features. However, this result may not hold for other predictive tasks in CPath. Further research is required to understand the impact of different factors on the performance of aggregation methods and to make predictions more generalizable, verifiable, biologically interpretable, and explainable. Bottom-up approaches provide interpretability. It is important to remember that while interpretability is important, limiting the use of data-driven analytics to human interpretability only may hinder its potential to deliver novel and discovery-driven results, as top-down approaches have been shown to perform similarly or even outperform human experts in some cases, such as with tumor identification. Data biases and global competition play a crucial role in the generalization and reproducibility of AI algorithms and must be considered in the clinical evaluation of any AI solution.

CRediT authorship contribution statement

Mohsin Bilal: Conceptualization, Writing – original draft, Investigation, Writing – review & editing, Visualization, Project administration. **Robert Jewsbury:** Writing – original draft, Investigation, Visualization, Methodology, Software, Validation, Writing – review & editing. **Ruoyu Wang:** Methodology, Software, Validation, Writing – original draft, Visualization, Writing – review & editing. **Hammam M. AlGhamdi:** Writing – original draft, Investigation, Visualization. **Amina Asif:** Investigation, Writing – review & editing. **Mark Eastwood:** Writing – original draft, Investigation, Writing – review & editing, Visualization. **Nasir Rajpoot:** Supervision, Conceptualization, Investigation, Writing – review & editing, Funding acquisition.

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.

Acknowledgment

Authors are grateful to members of the Tissue Image Analytics (TIA) centre for their feedback on some of the initial ideas for this paper.

References

Abels, E., Pantanowitz, L., Aeffner, F., Zarella, M.D., Laak, J., Bui, M.M., Vemuri, V.N., Parwani, A.V., Gibbs, J., Agosto-Arroyo, E., Beck, A.H., Kozlowski, C., 2019. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. J. Pathol. 249 (3), 286–294. https://doi.org/10.1002/path.5331.

Ahmedt-Aristizabal, D., Armin, M.A., Denman, S., Fookes, C., Petersson, L., 2022. A survey on graph-based deep learning for computational histopathology. Comput. Med. Imaging Graph. 95, 102027 https://doi.org/10.1016/j.compmedimag.2021.102027.

AlGhamdi, H.M., Koohbanani, N.A., Rajpoot, N., Raza, S.E.A., 2021. A novel cell map representation for weakly supervised prediction of ER & PR status from H&E WSIs. Proc. MICCAI Workshop Comput. Pathol. 156, 10.

Anklin, V., Pati, P., Jaume, G., Bozorgtabar, B., Foncubierta-Rodriguez, A., Thiran, J.P., Sibony, M., Maria, G., Goksel, O., de Bruijne, M., 2021. Learning whole-slide segmentation from inexact and incomplete labels using tissue graphs. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 12902. Springer, pp. 636–646.

Asif, A., Rajpoot, K., Snead, D., Minhas, F., Rajpoot, N., 2021. Towards Launching AI Algorithms for Cellular Pathology into Clinical & Pharmaceutical Orbits. arXiv. http://arxiv.org/abs/2112.09496.

Awan, R., Nimir, M., Raza, S.E.A., Bilal, M., Lotz, J., Snead, D., Robinson, A., Rajpoot, N., 2022. Deep Learning based Prediction of MSI Using MMR Markers in Colorectal Cancer. arXiv. http://arxiv.org/abs/2203.00449.

Babenko, Boris, 2008. Multiple Instance Learning: Algorithms and Applications.

Bilal, M., Raza, S.E.A., Azam, A., Graham, S., Ilyas, M., Cree, I.A., Snead, D., Minhas, F., Rajpoot, N.M., 2021. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. Lancet Digit. Health 3 (12), e763–e772. https://doi.org/10.1016/S2589-7500(21)00180-1.

M. Bilal, Y.W. Tsang, M. Ali, S. Graham, E. Hero, N. Wahab, K. Dodd, H. Sahota, S. Wu, W. Lu, M. Jahanifar, A. Robinson, A. Azam, K. Benes, M. Nimir, K. Hewitt, A. Bhalerao, H. Eldaly, S.E. Ahmed Raza, N. Rajpoot (2022). Development and validation of AI-based pre-screening of large bowel biopsies [Preprint]. Pathology. 10.1101/2022.11.30.22282859.

Bilal, M., Nimir, M., Snead, D., Taylor, G.S., Rajpoot, N., 2022. Role of AI and digital pathology for colorectal immuno-oncology. Br. J. Cancer. https://doi.org/10.1038/s41416-022-01986-1.

Bilal, M., Tsang, Y.W., Ali, M., Graham, S., Hero, E., Wahab, N., Dodd, K., Sahota, H., Lu, W., Jahanifar, M., Robinson, A., Azam, A., Benes, K., Nimir, M., Bhalerao, A., Eldaly, H., Ahmed Raza, S.E., Gopalakrishnan, K., Minhas, F., Rajpoot, N., 2022. AI Based Pre-Screening of Large Bowel Cancer via Weakly Supervised Learning of Colorectal Biopsy Histology Images. MedRxiv. https://doi.org/10.1101/2022.02.28.22271565 (2022.02.28.22271565) [Preprint]. Pathology.

Bulten, W., Kartasalo, K., Chen, P.H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., van Boven, H., Vink, R., Hulsbergen-van de Kaa, C., van der Laak, J., Amin, M.B., Evans, A.J., van der Kwast, T., Allan, R., Humphrey, P.A., Grönberg, H., Samaratunga, H., Park, J., 2022. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nat. Med. 28 (1), 154–163. https://doi.org/10.1038/s41591-021-01620-2.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. 25 (8), 1301–1309. https://doi.org/10.1038/s41591-019-0508-1.

Campbell, J.D., Yau, C., Bowlby, R., Liu, Y., Brennan, K., Fan, H., Taylor, A.M., Wang, C., Walter, V., Akbani, R., Byers, L.A., Creighton, C.J., Coarfa, C., Shih, J., Cherniack, A. D., Gevaert, O., Prunello, M., Shen, H., Anur, P., Mariamidze, A., 2018. Genomic, pathway network, and immunologic features distinguishing squamous carcinomas. Cell Rep. 23 (1), e6 https://doi.org/10.1016/j.celrep.2018.03.063.

Chang, J.R., Lee, C.Y., Chen, C.C., Reischl, J., Qaiser, T., Yeh, C.Y., de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C., 2021. Hybrid aggregation network for survival analysis from whole slide histopathological images. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, 12905. Springer International Publishing, pp. 731–740. https://doi.org/10.1007/978-3-030-87240-3_70.

Chen, H., Li, C., Wang, G., Li, X., Mamunur Rahaman, M., Sun, H., Hu, W., Li, Y., Liu, W., Sun, C., Ai, S., Grzegorzek, M., 2022. GasHis-transformer: a multi-scale visual transformer approach for gastric histopathological image detection. Pattern Recognit. 130, 108827 https://doi.org/10.1016/j.patcog.2022.108827.

Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F., 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16144–16155.

G. Corso, L. Cavalleri, D. Beaini, P. Liò, & P. Veličković (2020). Principal neighbourhood aggregation for graph nets. arXiv:http://arxiv.org/abs/2004.05718 [Cs, Stat].

Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A., 2018. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat. Med. 24 (10), 1559–1567. https://doi.org/10.1038/s41591-018-0177-5.

Diao, J.A., Wang, J.K., Chui, W.F., Mountain, V., Gullapally, S.C., Srinivasan, R., Mitchell, R.N., Glass, B., Hoffman, S., Rao, S.K., Maheshwari, C., Lahiri, A., Prakash, A., McLoughlin, R., Kerner, J.K., Resnick, M.B., Montalto, M.C., Khosla, A., Wapinski, I.N., Taylor-Weiner, A., 2021. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. Nat. Commun. 12 (1), 1613. https://doi.org/10.1038/s41467-021-21896-9.

Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T., 1997. Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. 89 (1–2), 31–71. https://doi.org/10.1016/S0004-3702(96)00034-3.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, & N. Houlsby (2021). An image is worth 16x16 words: transformers for image recognition at scale. arXiv:http://arxiv.org/abs/2010.11929.

Ellis, I., Al-Sam, S., Anderson, N., Carder, P., Deb, R., Girling, A., Hales, S., Hanby, A., Ibrahim, M., Lee, A., Liebmann, R., Mallon, E., Pinder, S., Provenzano, E., Quinn, C., Rakha, E., Rowlands, D., Stephenson, T., Wells, C., 2016. Guidelines working group of the UK national coordinating committee for breast pathology G148 HR. Pathology Reporting of Breast Disease in Surgical Excision Specimens Incorporating the Dataset for Histological Reporting of Breast Cancer. The Royal College of Pathologists, pp. 1–160. https://www.rcpath.org/uploads/assets/7763be1c-d330-40e8-95d08f955752792a/G148_BreastDataset-hires-Jun16.pdf.

Gildenblat, J., Ben-Shaul, I., Lapp, Z., Klaiman, E., Del Bimbo, A., Cucchiara, R., Sclaroff, S., Farinella, G.M., Mei, T., Bertini, M., Escalante, H.J., Vezzani, R., 2021. Certainty pooling for multiple instance learning. In: Pattern Recognition. ICPR International Workshops and Challenges, 12661. Springer International Publishing, pp. 141–153. https://doi.org/10.1007/978-3-030-68763-2_11.

Guan, Y., Zhang, J., Tian, K., Yang, S., Dong, P., Xiang, J., Yang, W., Huang, J., Zhang, Y., Han, X., 2022. Node-aligned graph convolutional network for whole-slide image representation and classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18813–18823.

Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., Takeuchi, I., 2020. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3851–3860. https://doi.org/10.1109/CVPR42600.2020.00391.

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 (CVPR), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

Ho, C., Zhao, Z., Chen, X.F., Sauer, J., Saraf, S.A., Jialdasani, R., Taghipour, K., Sathe, A., Khor, L.Y., Lim, K.H., Leow, W.Q., 2022. A promising deep learning-assistive algorithm for histopathological screening of colorectal cancer. Sci. Rep. 12 (1), 2222. https://doi.org/10.1038/s41598-022-06264-x.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243.

Huang, S.C., Chen, C.C., Lan, J., Hsieh, T.Y., Chuang, H.C., Chien, M.Y., Ou, T.S., Chen, K.H., Wu, R.C., Liu, Y.J., Cheng, C.T., Huang, Y.J., Tao, L.W., Hwu, A.F., Lin, I. C., Hung, S.H., Yeh, C.Y., Chen, T.C., 2022. Deep neural network trained on gigapixel images improves lymph node metastasis detection in clinical settings. Nat. Commun. 13 (1), 3347. https://doi.org/10.1038/s41467-022-30746-1.

M. Ilse, J.M. Tomczak, & M. Welling (2018). Attention-based Deep Multiple Instance Learning. 10.48550/ARXIV.1802.04712.

Javed, S.A., Juyal, D., Padigela, H., Taylor-Weiner, A., Yu, L., Prakash, A., 2022. Additive MIL: Intrinsic Interpretability for Pathology. arXiv. http://arxiv.org/abs/2206.01794.

Jewsbury, Robert, Abhir Bhalerao, Nasir, M.Rajpoot, 2021. A QuadTree Image Representation for Computational Pathology. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 648–656.

Kanavati, F., Toyokawa, G., Momosaki, S., Rambeau, M., Kozuma, Y., Shoji, F., Yamazaki, K., Takeo, S., Iizuka, O., Tsuneki, M., 2020. Weakly-supervised learning for lung carcinoma classification using deep learning. Sci. Rep. 10 (1), 1 https://doi.org/10.1038/s41598-020-66333-x. Article.

J.N. Kather, J. Schulte, H.I. Grabsch, C. Loeffler, H. Muti, J. Dolezal, A. Srisuwananukorn, N. Agrawal, S. Kochanny, S. Stillfried, P. Boor, T. Yoshikawa, D. Jaeger, C. Trautwein, P. Bankhead, N.A. Cipriani, T. Luedde, & A.T. Pearson (2019). Deep Learning Detects Virus Presence in Cancer Histology [Preprint]. Cancer Biology. 10.1101/690206.

Kather, J.N., Pearson, A.T., Halama, N., Jäger, D., Krause, J., Loosen, S.H., Marx, A., Boor, P., Tacke, F., Neumann, U.P., Grabsch, H.I., Yoshikawa, T., Brenner, H., Chang-Claude, J., Hoffmeister, M., Trautwein, C., Luedde, T., 2019. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat. Med. 25 (7), 1054–1056. https://doi.org/10.1038/s41591-019-0462-y.

Klein, S., Quaas, A., Quantius, J., Löser, H., Meinel, J., Peifer, M., Wagner, S., Gattenlöhner, S., Wittekindt, C., von Knebel Doeberitz, M., Prigge, E.S., Langer, C., Noh, K.W., Maltseva, M., Reinhardt, H.C., Büttner, R., Klussmann, J.P., Wuerdemann, N., 2021. Deep learning predicts HPV association in oropharyngeal squamous cell carcinomas and identifies patients with a favorable prognosis using regular H&E stains. Clin. Cancer Res. 27 (4), 1131–1138. https://doi.org/10.1158/1078-0432.CCR-20-3596.

Kosaraju S., Park J., Lee H., Yang J.W., Kang M. Deep learning-based framework for slide-based histopathological image analysis. Sci. Rep. 2022 9;12 (1): 19075, doi: 10.1038/s41598-022-23166-0.

Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, & Y. Zhang (2021). TransMIL: transformer based correlated multiple instance learning for whole slide image classification. arXiv:http://arxiv.org/abs/2106.00908 [Cs].

Kraus, O.Z., Ba, L.J., Frey, B., 2016. Classifying and segmenting microscopy images using convolutional multiple instance learning. Bioinformatics 32 (12), i52–i59. https://doi.org/10.1093/bioinformatics/btw252.

Laleh, N.G., Muti, H.S., Loeffler, C.M.L., Echle, A., Saldanha, O.L., Mahmood, F., Lu, M. Y., Trautwein, C., Langer, R., Dislich, B., Buelow, R.D., Grabsch, H.I., Brenner, H., Chang-Claude, J., Alwers, E., Brinker, T.J., Khader, F., Truhn, D., Gaisa, N.T., Kather, 2022. Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. Med. Image Anal. https://doi.org/10.1016/j.media.2022.102474, 102474.

Lalou, M., Tahraoui, M.A., Kheddouci, H., 2018. The critical node detection problem in networks: a survey. Comput. Sci. Rev. 28, 92–117. https://doi.org/10.1016/j.cosrev.2018.02.002.

Landherr, A., Friedl, B., Heidemann, J., 2010. A critical review of centrality measures in social networks. Bus. Inf. Syst. Eng. 2 (6), 371–385. https://doi.org/10.1007/s12599-010-0127-3.

Lerousseau, M., Vakalopoulou, M., Deutsch, E., Paragios, N., 2021. SparseConvMIL: Sparse Convolutional Context-Aware Multiple Instance Learning for Whole Slide Image Classification. arXiv. http://arxiv.org/abs/2105.02726.

Li, B., Li, Y., Eliceiri, K.W., 2021. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14313–14323. https://doi.org/10.1109/CVPR46437.2021.01409.

Lipkova, J., Chen, T.Y., Lu, M.Y., Chen, R.J., Shady, M., Williams, M., Wang, J., Noor, Z., Mitchell, R.N., Turan, M., Coskun, G., Yilmaz, F., Demir, D., Nart, D., Basak, K., Turhan, N., Ozkara, S., Banz, Y., Odening, K.E., Mahmood, F., 2022. Deep learning-enabled assessment of cardiac allograft rejection from endomyocardial biopsies. Nat. Med. 28 (3), 3 https://doi.org/10.1038/s41591-022-01709-2. Article.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A ConvNet for the 2020s. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11966–11976.

Lu, W., Graham, S., Bilal, M., Rajpoot, N., Minhas, F., 2020. Capturing cellular topology in multi-gigapixel pathology images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1049–1058. https://doi.org/10.1109/CVPRW50498.2020.00138.

Lu, M., Pan, Y., Nie, D., Liu, F., Shi, F., Xia, Y., Shen, D., 2021. SMILE: sparse-attention based multiple instance contrastive learning for glioma sub-type classification using pathological images. In: Proceedings of the MICCAI Workshop on Computational Pathology, pp. 159–169. In: https://proceedings.mlr.press/v156/lu21a.html.

Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F., 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. 5 (6), 555–570. https://doi.org/10.1038/s41551-020-00682-w.

Lu, W., Toss, M., Dawood, M., Rakha, E., Rajpoot, N., Minhas, F., 2022. SlideGraph +: whole slide image level graphs to predict HER2 status in breast cancer. Med. Image Anal. 80, 102486 https://doi.org/10.1016/j.media.2022.102486.

Naik, N., Madani, A., Esteva, A., Keskar, N.S., Press, M.F., Ruderman, D., Agus, D.B., Socher, R., 2020. Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains. Nat. Commun. 11 (1), 1 https://doi.org/10.1038/s41467-020-19334-3. Article.

Pagni, F., Malapelle, U., Doglioni, C., Fontanini, G., Fraggetta, F., Graziano, P., Marchetti, A., Guerini Rocco, E., Pisapia, P., Vigliar, E.V., Buttitta, F., Jaconi, M., Fusco, N., Barberis, M., Troncone, G., 2020. Digital pathology and PD-L1 testing in non small cell lung cancer: a workshop record. Cancers 12 (7), 1800. https://doi.org/10.3390/cancers12071800.

Park J., Chung Y.R., Nose A., 2022. Comparative analysis of high- and low-level deep learning approaches in microsatellite instability prediction. Sci. Rep. 18;12 (1): 12218. doi: 10.1038/s41598-022-16283-3.

Pati, P., Jaume, G., Fernandes, L.A., Foncubierta-Rodríguez, A., Feroce, F., Anniciello, A. M., Scognamiglio, G., Brancati, N., Riccio, D., Di Bonito, M., De Pietro, G., Botti, G., Goksel, O., Thiran, J.P., Frucci, M., Gabrani, M., Sudre, C.H., Fehri, H., Arbel, T., Baumgartner, C.F., Dalca, A., Tanno, R., Van Leemput, K., Wells, W.M., Sotiras, A., Papiez, B., Ferrante, E., Parisot, S., 2020. HACT-Net: a hierarchical cell-to-tissue graph neural network for histopathological image classification. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis, 12443. Springer International Publishing, pp. 208–219. https://doi.org/10.1007/978-3-030-60365-6_20.

Pinckaers, H., van Ginneken, B., Litjens, G., 2020. Streaming convolutional neural networks for end-to-end learning with multi-megapixel images. IEEE Trans. Pattern Anal. Mach. Intell. https://doi.org/10.1109/TPAMI.2020.3019563, 1–1.

Ray, S., 2001. Multiple Instance Regression International Conference on Machine Learning.

Reisenbüchler, D., Wagner, S.J., Boxberg, M., Peng, T., 2022. Local Attention Graph-based Transformer for Multi-target Genetic Alteration Prediction. arXiv. http://arxiv.org/abs/2205.06672.

Saillard, C., Dehaene, O., Marchand, T., Moindrot, O., Kamoun, A., Schmauch, B., Jegou, S., 2021. Self-supervised learning improves dMMR/MSI detection from histology slides across multiple cancers. Proc. Mach. Learn. Res. 156, 16, 2021. https://openreview.net/pdf?id=DRnNTPsjTCQ.

Schirris, Y., Gavves, E., Nederlof, I., Horlings, H.M., Teuwen, J., 2022. DeepSMILE: Contrastive self-supervised pre-training benefits MSI and HRD classification directly from H&E whole-slide images in colorectal and breast cancer. Med. Image Anal., 102464 https://doi.org/10.1016/j.media.2022.102464.

Schmauch, B., Romagnoni, A., Pronier, E., Saillard, C., Maillé, P., Calderaro, J., Kamoun, A., Sefta, M., Toldo, S., Zaslavskiy, M., Clozel, T., Moarii, M., Courtiol, P., Wainrib, G., 2020. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nat. Commun. 11 (1), 1 https://doi.org/10.1038/s41467-020-17678-4. Article.

Shaban, M., Awan, R., Fraz, M.M., Azam, A., Tsang, Y.W., Snead, D., Rajpoot, N.M., 2020. Context-aware convolutional neural network for grading of colorectal cancer histology images. IEEE Trans. Med. Imaging 39 (7), 2395–2405. https://doi.org/10.1109/TMI.2020.2971006.

Y. Sharma, A. Shrivastava, L. Ehsan, C.A. Moskaluk, S. Syed, & D.E. Brown (2021). Cluster-to-conquer: a framework for end-to-end multi-instance learning for whole slide image classification. arXiv:http://arxiv.org/abs/2103.10626 [Cs, Eess].

Skrede, O.J., De Raedt, S., Kleppe, A., Hveem, T.S., Liestøl, K., Maddison, J., Askautrud, H.A., Pradhan, M., Nesheim, J.A., Albregtsen, F., Farstad, I.N., Domingo, E., Church, D.N., Nesbakken, A., Shepherd, N.A., Tomlinson, I., Kerr, R., Novelli, M., Kerr, D.J., Danielsen, H.E., 2020. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. Lancet North Am. Ed. 395 (10221), 350–360. https://doi.org/10.1016/S0140-6736(19)32998-8.

Su, F., Li, J., Zhao, X., Wang, B., Hu, Y., Sun, Y., Ji, J., 2022. Interpretable tumor differentiation grade and microsatellite instability recognition in gastric cancer using deep learning. Lab. Invest. 102 (6), 641–649. https://doi.org/10.1038/s41374-022-00742-6.

Tan, M., Le, Q., 2019. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, pp. 6105–6114. In: https://proceedings.mlr.press/v97/tan19a.html.

Tan, M., Le, Q.V., 2021. EfficientNetV2: Smaller Models and Faster Training. arXiv. http://arxiv.org/abs/2104.00298.

Tellez, D., Litjens, G., van der Laak, J., Ciompi, F., 2019. Neural image compression for gigapixel histopathology image analysis. IEEE Trans. Pattern Anal. Mach. Intell. https://doi.org/10.1109/TPAMI.2019.2936841, 1–1.

Tomita, N., Abdollahi, B., Wei, J., Ren, B., Suriawinata, A., Hassanpour, S., 2019. Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. JAMA Netw. Open 2 (11), e1914645. https://doi.org/10.1001/jamanetworkopen.2019.14645.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, & I. Polosukhin (2017). Attention Is all you need. arXiv:http://arxiv.org/abs/1706.03762 [Cs].

Y. Wang, Y. Sun, Z. Liu, S.E. Sarma, M.M. Bronstein, & J.M. Solomon (2019). Dynamic graph CNN for learning on point clouds. arXiv:http://arxiv.org/abs/1801.07829 [Cs].

Wang, J., Chen, R.J., Lu, M.Y., Baras, A., Mahmood, F., 2019. Weakly Supervised Prostate TMA Classification via Graph Convolutional Networks. arXiv. http://arxiv.org/abs/1910.13328.

Wang, K.S., Yu, G., Xu, C., Meng, X.H., Zhou, J., Zheng, C., Deng, Z., Shang, L., Liu, R., Su, S., Zhou, X., Li, Q., Li, J., Wang, J., Ma, K., Qi, J., Hu, Z., Tang, P., Deng, J., Deng, H.W., 2021. Accurate diagnosis of colorectal cancer based on histopathology images using artificial intelligence. BMC Med. 19 (1), 76. https://doi.org/10.1186/s12916-021-01942-5.

Wang, R., Asif, A., Bashir, S., Young, L., Khurram, S.A., 2021. Ranking loss based weakly supervised model for prediction of HPV infection status from multi-gigapixel histology images. In: Proceedings of the Med-NeurIPS Workshop 2021, p. 5.

Wang, R., Khurram, S.A., Asif, A., Young, L., Rajpoot, N., 2022. Rank the Triplets: A Ranking-Based Multiple Instance Learning Framework for Detecting HPV Infection in Head and Neck Cancers Using Routine H&E Images. arXiv. http://arxiv.org/abs/2206.08275.

Westra, W.H., 2012. The morphologic profile of HPV-related head and neck squamous carcinoma: implications for diagnosis, prognosis, and clinical management. Head Neck Pathol. 6 (S1), 48–54. https://doi.org/10.1007/s12105-012-0371-6.

Wilson, R., Knutsson, H., 1988. Uncertainty and inference in the visual system. IEEE Trans. Syst. Man Cybern. 18 (2), 305–312. https://doi.org/10.1109/21.3468.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S., 2021. A comprehensive survey on graph neural networks. IEEE Trans. Neural Netw. Learn. Syst. 32 (1), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386.

C. Xie, C. Vanderbilt, C. Feng, D. Ho, G. Campanella, J. Egger, A. Plodkowski, J. Girshman, P. Sawan, K. Arbour, M. Hellmann, & T. Fuchs (2022). Computational Biomarker Predicts Lung ICI Response via Deep Learning-Driven Hierarchical Spatial Modelling from H&E [Preprint]. In Review. 10.21203/rs.3.rs-1251762/v1.

Y. Xiong, Z. Zeng, R. Chakraborty, M. Tan, G. Fung, Y. Li, & V. Singh. (2021). Nystr \"omformer: A Nystr\"om-based algorithm for approximating self-attention. arXiv: http://arxiv.org/abs/2102.03902 [Cs].

Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M., Chang, E.I.C., 2014. Deep learning of feature representation with multiple instance learning for medical image analysis. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1626–1630. https://doi.org/10.1109/ICASSP.2014.6853873.

K. Xu, W. Hu, J. Leskovec, & S. Jegelka (2019). How powerful are graph neural networks? arXiv:http://arxiv.org/abs/1810.00826 [Cs, Stat].

Yamashita, R., Long, J., Longacre, T., Peng, L., Berry, G., Martin, B., Higgins, J., Rubin, D. L., Shen, J., 2021. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. Lancet Oncol. 22 (1), 132–141. https://doi.org/10.1016/S1470-2045(20)30535-0.

Ying, R., Bourgeois, D., You, J., Zitnik, M., Leskovec, J., 2019. GNNExplainer: Generating Explanations for Graph Neural Networks. arXiv. http://arxiv.org/abs/1903.03894.

Zeng, Q., Klein, C., Caruso, S., Maille, P., Laleh, N.G., Sommacale, D., Laurent, A., Amaddeo, G., Gentien, D., Rapinat, A., Regnault, H., Charpy, C., Nguyen, C.T., Tournigand, C., Brustia, R., Pawlotsky, J.M., Kather, J.N., Maiuri, M.C., Loménie, N., Calderaro, J., 2022. Artificial intelligence predicts immune and inflammatory gene signatures directly from hepatocellular carcinoma histology. J. Hepatol. https://doi.org/10.1016/j.jhep.2022.01.018. S0168827822000319.

Zhang, D., Duan, Y., Guo, J., Wang, Y., Yang, Y., Li, Z., Wang, K., Wu, L., Yu, M., 2022. Using multi-scale convolutional neural network based on multi-instance learning to predict the efficacy of neoadjuvant chemoradiotherapy for rectal cancer. IEEE J. Transl. Eng. Health Med. 10, 4300108 https://doi.org/10.1109/JTEHM.2022.3156851.

Zheng, Y., Gindra, R., Betke, M., Beane, J.E., Kolachalama, V.B., 2021. A Deep Learning Based Graph-Transformer for Whole Slide Image Classification. medRxiv. https://doi.org/10.1101/2021.10.15.21265060 (p. 2021.10.15.21265060).

Zheng, Y., Gindra, R.H., Green, E.J., Burks, E.J., Betke, M., Beane, J.E., Kolachalama, V. B., 2022. A graph-transformer for whole slide image classification. IEEE Trans. Med. Imaging. https://doi.org/10.1109/TMI.2022.3176598, 1–1.

Zhou, Z.H., Zhang, M.L., Schölkopf, B., Platt, J., Hofmann, T., 2007. Multi-instance multi-label learning with application to scene classification. Advances in Neural Information Processing Systems 19. The MIT Press. https://doi.org/10.7551/mitpress/7503.003.0206.

Zhou, Y., Graham, S., Alemi Koohbanani, N., Shaban, M., Heng, P.A., Rajpoot, N., 2019. CGC-Net: cell graph convolutional network for grading of colorectal cancer histology images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 388–398. https://doi.org/10.1109/ICCVW.2019.00050.