Computers in Biology and Medicine 152 (2023) 106340

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

# Automatic diagnosis and grading of Prostate Cancer with weakly supervised learning on whole slide images

Jinxi Xiang [a,1], Xiyue Wang [b,1], Xinran Wang [c], Jun Zhang [a,*], Sen Yang [a], Wei Yang [a], Xiao Han [a], Yueping Liu [c,*]

[a] *AI Lab, Tencent, Shenzhen, China*
[b] *College of Computer Science, Sichuan University, Chengdu, China*
[c] *Department of Pathology, The Fourth Hospital of Hebei Medical University, Shijiazhuang, China*

## ARTICLE INFO

*MSC:*
41A05
41A10
65D05
65D17

*Keywords:*
Weakly supervised learning
Prostate cancer
Gleason grading
Whole slide Images
Slide-level labels
Graph convolution network

## ABSTRACT

**Background:** The workflow of prostate cancer diagnosis and grading is cumbersome and the results suffer from substantial inter-observer variability. Recent trials have shown potential in using machine learning to develop automated systems to address this challenge. Most automated deep learning systems for prostate cancer Gleason grading focused on supervised learning requiring demanding fine-grained pixel-level annotations.
**Methods:** A weakly-supervised deep learning model with slide-level labels is presented in this study for the diagnosis and grading of prostate cancer with whole slide image (WSI). WSIs are first cropped into small patches and then processed with a deep learning model to extract patch-level features. A graph convolution network (GCN) is used to aggregate the features for classifications. Throughout the training process, the noisy labels are progressively filtered out to reduce inter-observer variations in clinical reports. Finally, multi-center independent test cohorts with 6,174 slides are collected to evaluate the prostate cancer diagnosis and grading performance of our model.
**Results:** The cancer diagnosis (2-level classification) results on two external test sets ($n$ = 4,675, $n$ = 844) show an area under the receiver operating characteristic curve (AUC) of 0.985 and 0.986. The Gleason grading (6-level classification) results reach 0.931 quadratic weighted kappa on the internal test set ($n$ = 531). It generalizes well on the external test dataset ($n$ = 844) with 0.801 quadratic weighted kappa with the reference standard set independently. The model enables pathological meaningful interpretability by visualizing the most attended lesions which are highly consistent with expert annotations.
**Conclusion:** The proposed model incorporates a graph network in weakly supervised learning with only slide-level reports. A robust learning strategy is also employed to correct the label noise. It is highly accurate (> 0.985 AUC for diagnosis) and also interpretable with intuitive heatmap visualization. It can be unified with a digital pathology pipeline to deliver prostate cancer metrics for a pathology report.

## 1. Introduction

Prostate cancer is the second most frequent malignancy (after lung cancer) in men worldwide, causing a high mortality ratio and also risks of over-diagnosis and over-treatments [1,2]. Gleason grading system of biopsied tissue for prostate cancer provides prognostic markers for patients and is key reference information for treatment planning. In clinical routines, pathologists characterize tumors into different patterns based on the histological architecture of the tumor tissue. Specifically, biopsy specimens are classified into one of five groups based on the distribution of the Gleason pattern, commonly referred to as the International Society of Urological Pathology (ISUP) grade group, ISUP grade, Gleason grade group, or simply grade groups (GGs) 3-6 [3,4]. However, substantial inter-observer and intra-observer variability in grading reduces its usefulness for individual patients. Although specialized urological pathologists have greater concordance, such expertise is not widely available. Prostate cancer diagnostics could thus benefit from robust, reproducible Gleason grading [5]. Therefore, there is strong need for computer-aided assessment of biopsies for patient diagnosis [6–8].

Deep learning has been studied and shown promising use in the diagnosis of several digital pathology application [9,10]. Some earlier studies have applied feature engineering approaches to address Gleason

[*] Corresponding authors.
*E-mail addresses:* junejzhang@tencent.com (J. Zhang), liuyp@hbmu.edu (Y. Liu).
[1] Equal contributor.

https://doi.org/10.1016/j.compbiomed.2022.106340
Received 2 July 2022; Received in revised form 2 November 2022; Accepted 16 November 2022
Available online 21 November 2022

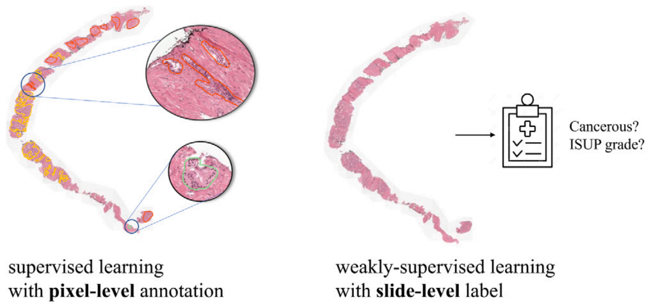

Cancerous?
ISUP grade?
supervised learning
with **pixel-level** annotation
weakly-supervised learning
with **slide-level** label

**Fig. 1.** Supervised learning and the proposed weakly-supervised learning.

grading [11,12]. Later, automated systems powered by deep learning enhance confidence in Gleason grading by improving consistency and providing center-independent clinical-level grading [5]. Recent years have witnessed the field transaction to the application of deep learning for detecting prostate cancer, and later Gleason grading of tissue microarray, and now biopsies [13–15].

A typical workflow to train an automated Gleason grading system usually requires large-scale annotated data performed by experienced pathologists. For instance, Tolkach et al. claimed that the creation of more than 1.5 million patches from 389 slides of the training dataset took longer than a year [15]. Similarly, Nagpal et al. used 112 million pathologist-annotated image patches from 1226 slides and evaluated on an independent validation dataset of 331 slides [16]. However, the lack of large, annotated datasets is severe in computational pathology considering that patch-level annotation in gigapixel WSIs is a tedious, time-consuming, and error-prone process. Further, due to the gigapixel size of WSIs, millions of patch-level annotations are sampled from considerably a small number of slides (tens to hundreds of slides) to extract patches.

Apart from cumbersome annotations, supervised learning could induce bias such as scanner vendors, stain variations, etc. [17]. The deep-learning diagnostic models, when trained using patch-level supervision, have been shown to suffer from performance drop when tested on data from different sources and imaging devices [18–20]. This challenge boosts the application of alternative computational strategies that work with slide-level annotations, as shown in Fig. 1. The transaction from patch-level supervised learning to slide-level weak supervised learning can not only reduce annotations, but also enable us to exploit the large-scale slide-level data that are readily available with clinical reports. Thus, weakly supervised learning can potentially reach clinical-level grading performance as MIL-RNN [20] with more than 10,000 slides.

We aimed to develop a weakly supervised model based on variants of multiple-instance learning (MIL) with clinically acceptable accuracy for prostate cancer diagnosis, localization, and Gleason grading using 10,616 public training datasets and more than 6,174 slides from independent test cohorts. Building on the standard MIL framework, we intend to address some unresolved bottlenecks:

**Feature extraction.** MIL employs the ImageNet pre-trained RestNet50 to extract low-dimensional features from image patches [19,20]. However, the significant domain deviation between WSI and natural images because cells and tissues exhibit a variety of shapes and appearances that reflect multiple states. Transferring natural image encoder to WSI leads to sub-optimal performance [21,22]. To this end, we employ contrastive learning [23–25] to pretrain a pathology-specific feature extractor.

**Feature aggregation.** Based on the assumption of local and independence of instance in MIL, one line of existing methods, e.g. AB-MIL [26] and CLAM [19], uses local attention-based pooling to selectively aggregate features to inform the slide-level diagnosis. However, it would be more beneficial to learn interactions between instances, which may help the model become more context-aware [19,21,27]. Intuitively, this is consistent with the behavior of pathologists considering both the contextual information around a single area and the correlation between different areas when making a diagnostic decision. Inspired by the property of graph convolution in relation-aware representations [28–30], we use GCN as a powerful tool to represent the WSI with context awareness.

**Robust training.** Addressing noisy labels in Gleason grading is a great challenge, considering that the inter-observer consistency among expert pathologists can be lower than 0.60 linear kappa [16,31]. WSIs can also suffer from color variations, imaging blur, contamination, etc. It is risky for the neural network to overfit the noisy data, thus, they may not generalize to unseen data for clinical applications. We intend to use a simple self-ensemble method to remove noisy samples under the threshold of uncertainty.

Our main contributions are summarized as: (1) A weakly-supervised learning framework named GCN-MIL is proposed for prostate cancer detection and grading using only slide-level labels to alleviate the annotation requirements. (2) Frontier techniques, such as self-supervised pertaining for feature extraction, graph convolution for feature aggregation, and robust training for noisy label filtering are employed to effectively enhance the overall diagnosis and grading performance. (3) Extensive experiments have been conducted on large-scale multi-center datasets with quantitative metric evaluation and interpretability heatmap, showing that the proposed model can potentially reach clinical-grade performance.

The following sections are **Material and Methods** that describe the training/testing datasets, and the details about the proposed deep learning model; **Results** that present quantitative and qualitative evaluations on diagnosis and grading performance; **Discussion** that compares the proposed method with other baseline methods, and point out the limitations; **Conclusions** that conclude main findings.

## 2. Materials and methods

Our weakly supervised model GCN-MIL for prostate cancer grading consists of a self-supervised CNN model $f$ to conduct feature extraction followed by a graph convolution network $g$ and attention pooling model $p$ to aggregate features, as shown in Fig. 2.

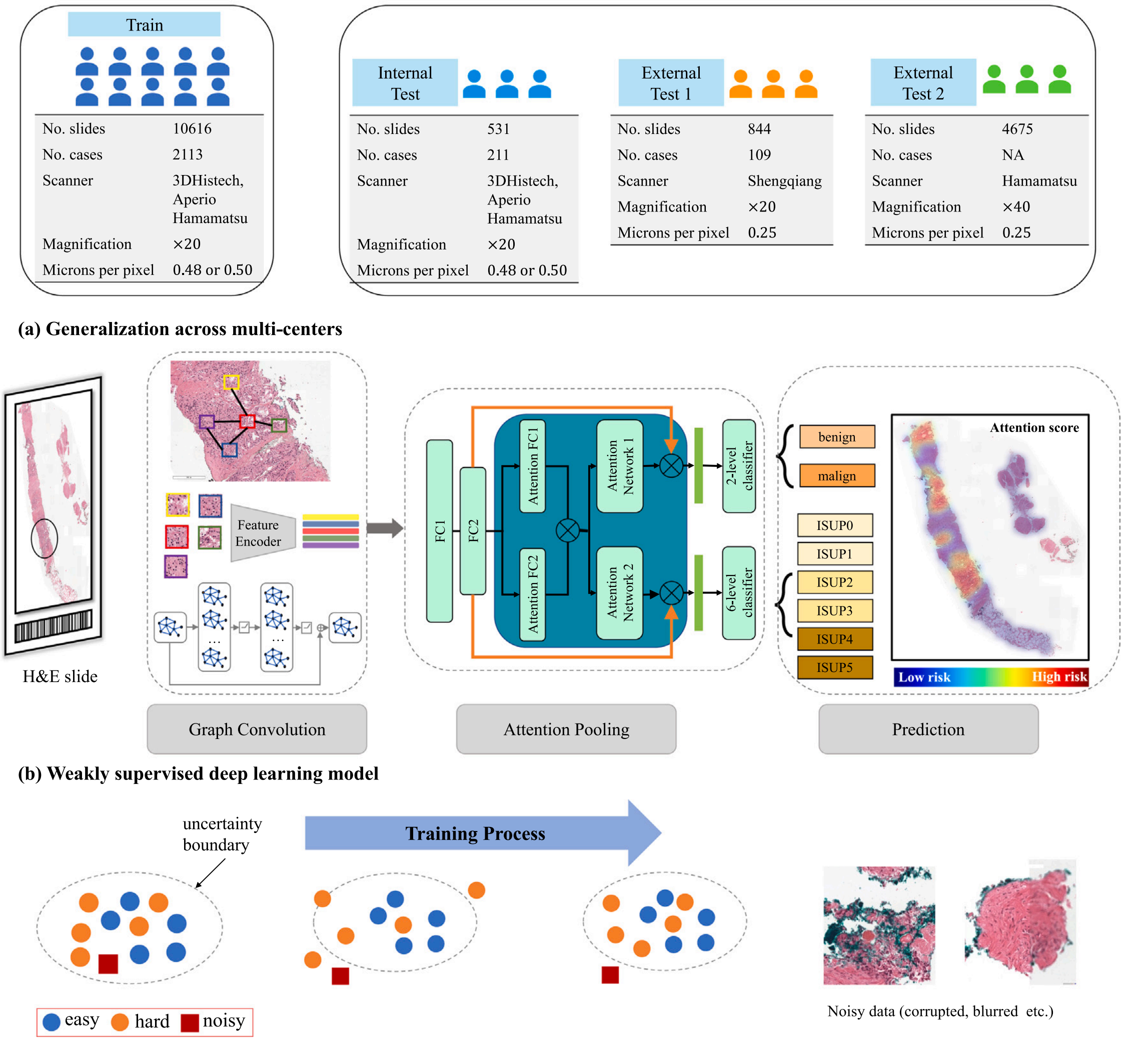
### 2.1. Pre-processing

The data preprocessing aims to segment the tissue regions of a gigapixel WSI (usually $> 10,000 \times 10,000$ resolution), and then crop the tissues into thousands of small image patches. For the training set and test set, we follow the same data preprocessing steps provided by the CLAM opensource tool.[2] Concretely, a gigapixel WSI is loaded into memory at a downscaled level (1.0 μm/pixel) and then automatedly segmented with Otsu's method to exclude non-tissue regions. Image patches of are extracted from tissue regions of the WSI within segmented foreground contours.

### 2.2. Feature extraction

During training and inference, patches are encoded once by a pre-trained CNN ResNet50 into a descriptive representation as 1024-dimension vectors. Using low-dimensional features instead of raw images made it tractable to train the deep learning models with all tiles in a slide (up to 10,000 patches or more in a slide) simultaneously on a GPU. We propose to use self-supervised contrastive learning for learning the feature extractor RestNet50 [24,32]. Specifically, we consider using contrastive learning for WSI [33], a state-of-the-art self-supervised learning framework [24,33] that enables robust representations to be learned without the need for manual labels.

[2] https://github.com/mahmoodlab/CLAM



**Train**

| | |
|---|---|
| No. slides | 10616 |
| No. cases | 2113 |
| Scanner | 3DHistech,<br>Aperio<br>Hamamatsu |
| Magnification | ×20 |
| Microns per pixel | 0.48 or 0.50 |

**Internal Test**

| | |
|---|---|
| No. slides | 531 |
| No. cases | 211 |
| Scanner | 3DHistech,<br>Aperio<br>Hamamatsu |
| Magnification | ×20 |
| Microns per pixel | 0.48 or 0.50 |

**External Test 1**

| | |
|---|---|
| No. slides | 844 |
| No. cases | 109 |
| Scanner | Shengqiang |
| Magnification | ×20 |
| Microns per pixel | 0.25 |

**External Test 2**

| | |
|---|---|
| No. slides | 4675 |
| No. cases | NA |
| Scanner | Hamamatsu |
| Magnification | ×40 |
| Microns per pixel | 0.25 |

**(a) Generalization across multi-centers**

**(b) Weakly supervised deep learning model**

**(c) Filtering out noisy samples progressively**

**Fig. 2.** The workflow of the proposed prostate cancer diagnosis and grading system. The feature extraction model takes tiled images of 256 × 256 (0.5 μm/pixel) as input and produces a 2048-dimensional embedding vector. A graph is constructed based on the embedding vectors and their spatial position in the slide. DeepGCN convolution is conducted on the graph-structure data to pass information among nodes. The classifier makes the final prediction of ISUP grades by pooling over all nodes using learning attention. Since the ISUP labels are imperfect, the training process iteratively filters out noisy labels using uncertainty estimation.

When applying to histopathological images, contrastive learning is performed by extracting small, overlapping patches from each image to create large-scale unlabeled data. The datasets used for self-supervised learning consist of 15,580,262 unlabeled histopathological images extracted from 30,072 WSIs of the TCGA [32], denoted as $\{s_j\}, j = 1, \dots, M$ and $M = 30{,}072$. The contrastive prediction task is to maximize the mutual information between the same image while in the meantime minimizing negative images, i.e.:

$$\mathcal{L}_q = -\log \frac{\exp\left(q \cdot k_+/\tau\right)}{\sum_{i=0}^{K} \exp\left(q \cdot k_i/\tau\right)} \tag{1}$$

where $q = f(s_j)$; $k = f_{\text{EMA}}(s_j)$; $\tau$ is a temperature constant; $f_{\text{EMA}}$ meaning a exponential moving average version of $f$. After self-supervised training, the pre-trained feature extractor $f$ is frozen and used to extract the 1024-dimensional vectors $\mathbf{b} = \{b_i \in \mathbb{R}^{1024}\}, i = 1, ..N$ of images $\mathbf{x} = \{x_i\}, i = 1, ..N$.

### 2.3. Feature aggregation

Adapting MIL for the Gleason grading problem requires considering both instance- and global-level features in the tumor and other tissues for quantifying and assessing the patient risk of cancer development. We use a GCN to capture context-aware information among patches by treating each embedding vector as nodes and their adjacent patches via edges connection.

To construct the global graph representation of a slide, we saved $(c_x, c_y)$ coordinates of cropped patches $\mathbf{b} = \{b_i\}$ in raw images to build the adjacency matrix $A$ for each slide via fast approximate nearest neighbors k-NN (k = 8). A slide could be denoted as CNN features and graph adjacent matrix $(\mathbf{b}, A)$. For each vertex $v$ (that has node feature $\mathbf{b}_v$) and its neighboring vertices $u$ (that has node feature $\mathbf{b}_u$) of $A$:

$$\begin{aligned} \mathbf{m}_v^{(l)} &= \rho^{(l)}\left(\left\{\phi^{(l)}\left(\mathbf{b}_v^{(l)}, \mathbf{b}_u^{(l)}, \mathbf{b}_{e_{vu}}^{(l)}\right) \to \mathbf{m}_{vu}^{(l)}\right\}\right) \\ \mathbf{b}_v^{(l+1)} &= \zeta^{(l)}\left(\mathbf{b}_v^{(l)}, \mathbf{m}_v^{(l)}\right) \end{aligned} \tag{2}$$

where $l$ is the layer number; $\phi$ is a message construction function that calculates a message $\mathbf{m}_{vu}^{(l)}$ between $\mathbf{b}_v$ and its neighbor $\mathbf{b}_u$ (with edge feature $\mathbf{b}_{e_{vu}}$); $\rho$ is a permutation invariant aggregation function that aggregates all messages passed to $\mathbf{b}_v$; $\zeta$ updates the existing node feature at $v$ with the aggregated message $\mathbf{b}_v^{(l)}$. We adapt the implementations of $\phi$, $\rho$, $\zeta$ from DeepGCN [34]. We construct a 4-layer DeepGCN network with a hidden dimension of 128, layer normalization, ReLU activation, and Dropout layer.

As a result, input WSIs are formulated as a graph-based data structure and further processed through message passing, and each node is able to iteratively accumulate feature vectors from its neighboring nodes and generate a new feature vector at the next hidden layer of the network, thus GCNs can learn to represent for each feature in a node.

We get a new bags of instances $\{z_i\}$ after graph convolution of $\{b_i\}$. By pooling over all nodes, we are able to obtain global representation for the entire WSI which serves as the inputs for the classifier, i.e.:

$$
\begin{aligned}
y &= \sum_{i=1}^{N} \alpha_i z_i \\
\alpha_i &= \frac{\exp\left\{\mathbf{w}^{\top}\left(\tanh\left(\mathbf{V}z_i^{\top}\right)\odot \operatorname{sigm}\left(\mathbf{U}z_i^{\top}\right)\right)\right\}}{\sum_{j=1}^{N}\exp\left\{\mathbf{w}^{\top}\left(\tanh\left(\mathbf{V}z_j^{\top}\right)\odot \operatorname{sigm}\left(\mathbf{U}z_j^{\top}\right)\right)\right\}}
\end{aligned}
\tag{3}
$$

where $\mathbf{w} \in \mathbb{R}^{128\times128}, \mathbf{V} \in \mathbb{R}^{128\times128}, \mathbf{U} \in \mathbb{R}^{128\times128}$ are parameters; $\odot$ is an element-wise multiplication and sigm() is the sigmoid non-linearity.

### 2.4. Robust training

The Gleason grade labels are actually noisy because in practical workflow, ground-truth labels can probably be corrupted by annotation inconsistency or image quality, and thus become noisy signals for training. As a result, they inevitably degenerate the robustness of learned models, especially for deep neural networks. To progressively filter out the noisy labels in training data, the model is self-ensembled throughout the training process and the key to learning with noisy labels lies in accurately and directly characterizing the uncertainty of label noise in the data. We identify and exclude noisy samples by making a boundary to distinguish between noisy labels and clean ones through uncertainty, building on the assumption that noisy labels are of large loss and uncertainty compared with clean ones. At each iteration, we only choose samples within the uncertainty boundary to update the GCN-MIL model. As the model converges, it is capable of distinguishing noisy labels better and thus would result in a better model at the next iteration. Depending on the computational budget, a maximal number of iterations for filtering can be set to save time.

### 2.5. Pseudo-code

As shown in Algorithm 1, we summarize the above-mentioned details to describe the input, output, modules, and training loop using the 2-level classification as an example. The input is the training set and the test set containing WSIs and slide-level labels. The target output is the predicted label $\{\hat{\mathbf{y}}_i\}$ of test set $\{\mathbf{x}_i^{\text{test}}\}$. We first preprocess a WSI into a set of small image patches, which are embedded with the feature extractor into vectors. In the training loop, a training sample $\mathbf{b}_i^{\text{train}}, \mathbf{y}_i$ is obtained. From line 8 to 14 in Algorithm 1, we evaluate whether the training sample is noisy or not. We update the GCN-MIL model with clean training samples from line 16 to 19. After training, we use the trained GCN-MIL network to predict the labels $\{\hat{\mathbf{y}}_i\}$ of the test set. For 6-level classification, we use a different linear classifier head as shown in Fig. 2 and filter the noisy_samples recorded during 2-level classification training.

### 2.6. Model complexity

In Table 1, we briefly analyze the model complexity in terms of the number of parameters of modules (Params) and Multiply-Add Cumulation (MACs). The preprocess module mainly involves segmentation and image cropping, both are non-parametric operations with 0 Params. Suppose we obtain 1000 image patches of a WSI which we use to measure the MACs of the following modules. After preprocessing, the self-supervised pretrained encoder RestNet50 is used for feature extraction. At this stage, the dimension of the 2D image is greatly reduced from 256 × 256 to a 1D vector. The Params of RestNet50 is 25 M and MACs 4.134 T. The graph convolution and attention pooling modules operating vectors are relatively lightweight, with 200 k and 795 k parameters, respectively.

**Algorithm 1:** Pseudo-code of GCN-MIL

```
   Input: training test $\{\mathbf{x}_i^{\text{train}}, \mathbf{y}_i\}$ and test set $\{\mathbf{x}_i^{\text{test}}\}$
   Output: predicted label $\{\hat{\mathbf{y}}_i\}$ of test set $\{\mathbf{x}_i^{\text{test}}\}$
 1 Preprocess WSIs into $256 \times 256$ image patches ;   // segment
    then crop
 2 Feature extraction $\{\mathbf{b}_i^{\text{train}}\} \leftarrow \{\mathbf{x}_i^{\text{train}}\}$; $\{\mathbf{b}_i^{\text{test}}\} \leftarrow \{\mathbf{x}_i^{\text{test}}\}$;
 3 noisy_samples = [ ] ;
 4 $T = 100$ ;
   /* training loop                                   */
 5 while $e < T$ do
 6     get training sample $\mathbf{b}_i^{\text{train}}, \mathbf{y}_i$ ;         // batch_size=1
 7     ;
 8     GCN-MIL.eval();                          // no gradients
 9     logits = GCN−MIL($\mathbf{b}_i^{\text{train}}$) ;
10     prob = softmax(logits) ;          // prob of positive
11     $d_r = 0.40 \times (1 - 1/(e+1))$ ;                // boundary
12     upper_bound = $(1.0 - d_r)$ if $(\mathbf{y}_i == 0)$ else 1.0 ;
13     lower_bound = 0.0 if $(\mathbf{y}_i == 0)$ else $d_r$ ;
14     if prob > upper_bound or prob < lower_bound:
        noisy_samples.append(($\mathbf{b}_i^{\text{train}}, \mathbf{y}_i$))
        continue ;                   // skip noisy sample
15     ;
16     GCN-MIL.train() ;                     // with gradients
17     logits = GCN−MIL($\mathbf{b}_i^{\text{train}}$) ;
18     loss = cross_entropy_loss(logits, $\mathbf{y}_i$) ;
19     loss.backward() ;
20     e = e + 1 ;
   /* infer with the trained GCN-MIL                  */
21 logits = GCN−MIL($\mathbf{b}_i^{\text{test}}$) ;
22 $\hat{\mathbf{y}}_i$ = argmax(logits) ;
```

**Table 1**
Model complexity.

| Module | Params | MACs |
|---|---|---|
| Preprocess | 0 | – |
| Feature Extraction | 25 M | 4.134 T |
| Graph Convolution | 200 k | 152 M |
| Attention Pooling | 795 k | 871 M |

**Table 2**
Data characteristics of the training set, internal validation set and external validation set.

| Group | PANDA train | Internal test | External test 1 | External test 2 | SICAP |
|---|---|---|---|---|---|
| Benign | 2892 | – | 220 | – | 36 |
| ISUP1 | 2666 | – | 169 | – | 14 |
| ISUP2 | 1343 | – | 88 | – | 22 |
| ISUP3 | 1242 | – | 108 | – | 23 |
| ISUP4 | 1249 | – | 125 | – | 18 |
| ISUP5 | 1224 | – | 134 | – | 42 |
| Total | 10,616 | 531 | 844 | 4675 | 155 |

### 2.7. Multi-center WSI datasets

To access the robustness and clinical applicability of our system, we collected 5 groups of datasets from 5 different centers, namely, PANDA, internal test set, external test 1 from private center, external test 2 from DiagSet-B, and SICAP, as shown in Table 2. These datasets consist of WSIs that were stained with hematoxylin and eosin (H&E), the most widely used stain in routine histopathology diagnostics, that highlights general tissue morphology such as cell nuclei and cytoplasm. We used PANDA to train our model, and internal test, external test 1, external test 2, and SICAP to evaluate its performance. The training set Prostate cANcer graDe Assessment (PANDA) challenge dataset consists

**Table 3**
Sensitivity and specificity at selected points on the receiver operating characteristic curves for cancer diagnosis.

| Dataset | Operating point 1: Threshold = 0.50 | | | Operating point 2: Threshold = 0.40 | | |
|---|---|---|---|---|---|---|
| | Sensitivity | Specificity | AUC | Sensitivity | Specificity | AUC |
| PANDA ($n$ = 10,616) | 97.9% | 95.3% | 0.992 | 99.2% | 73.5% | 0.992 |
| External test 1 ($n$ = 844) | 97.8% | 86.4% | 0.985 | 99.0% | 74.1% | 0.985 |
| External test 2 ($n$ = 4675) | 98.3% | 88.9% | 0.986 | 99.1% | 73.2% | 0.986 |

of 10,616 WSIs from two centers and each WSI is associated with a single ISUP label. The grading process consisted of finding and classifying cancer tissue into Gleason patterns based on the architectural growth patterns of the tumor [35]. The internal test set of PANDA containing around 531 slides is exclusive to participants and is used to evaluate the final score of the developed algorithm. This dataset is not publicly available but can be used to assess your model blindly via the Kaggle website. The External test set 1 includes 844 slides from 100 subjects (cancer 220, non-cancer 624) from the Fourth Hospital of Hebei Medical University, Shijiazhuang, China. The selection of slides in the external test set was enriched for higher ISUP grades to permit evaluation of predictions for uncommon grades. The External test set 2, also named Diagset-B, are released for prostate cancer diagnosis, consisting of 4675 scans with assigned binary diagnosis with diagnosis given independently by a group of histopathologists [36]. SICAPv2 database includes 155 biopsies from 95 different patients who signed the pertinent informed consent. The slides were analyzed by a group of urogenital pathologists at Hospital Clinico of Valencia, and a combined Gleason score was assigned per biopsy [37].

We use several evaluation metrics for diagnosis, i.e., AUC, accuracy (ACC), sensitivity, and specificity.

True positive (TP) = the number of cases correctly identified as positive. False positive (FP) = the number of cases incorrectly identified as positive. True negative (TN) = the number of cases correctly identified as negative. False negative (FN) = the number of cases incorrectly identified as negative.

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{6}$$

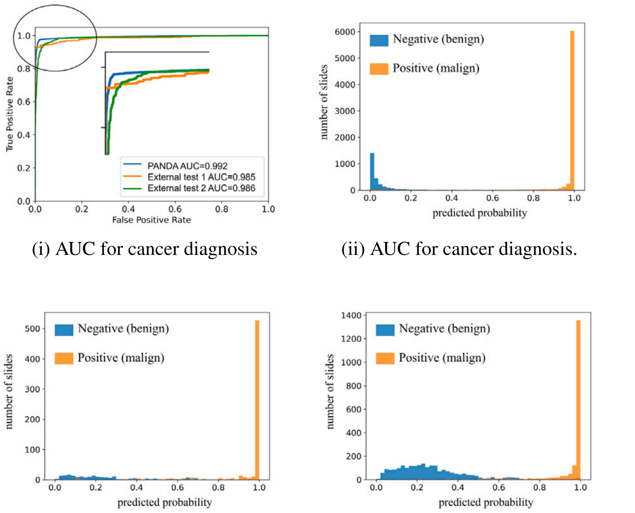
We use linear kappa and quadratic weighted kappa for Gleason grading as suggested by [35].

## 3. Results

### 3.1. Assessment on cancer diagnosis

The diagnosis of prostate cancer is a 2-level classification problem that aims to determine whether a WSI is positive (malign, ISUP 1,2,3,4,5) or negative (benign, ISUP 0). The training set PANDA is split into 5 stratified folds with balanced ISUP labels in each fold and then cleaned with the iterative filtering process in Fig. 2, where samples with large predicted loss and uncertainty outside boundary are considered to be noisy. In total, 849 samples (7.9% of total) are removed from the original dataset. The AUC representing the ability of the AI system to distinguish malignant from benign biopsies was 0.992 (95% CI 0.989–0.994) for cross-validation in clean PANDA set (begin = 2804, malignant = 6963). On large-scale external test set 1, the system achieved AUC of 0.985 (95% CI 0.981–0.989) on external test 2 (benign = 2,090, malignant = 2,585), and 0.986 (95% CI 0.844–0.988) on external test 1 (benign = 100, malignant = 838).

In Fig. 3(i), we present the AUC curve on three datasets, i.e., PANDA ($n$ = 10,616), External test 1 ($n$ = 844), and External test 2 ($n$ = 4,675). Fig. 3(ii), (iii), (iv) are the histograms that show the predicted probability on the test sets. The $x$-axis is the probability, and the $y$-axis is the number of slides. The histograms are explained as follows: given class labels $Y \in \{0, 1\}$, we use the classifier GCN-MIL that produces the probability $P(Y = 1|X = x)$ for each test WSI $x$. To summarize the performance of the classifier on a test set, we plot a histogram of predicted probabilities of WSIs in the same figure. Positive and negative slides are presented in different colors. Ideally, positive slides should have high probabilities, whereas negative slides should have low probabilities.



(i) AUC for cancer diagnosis (ii) AUC for cancer diagnosis.

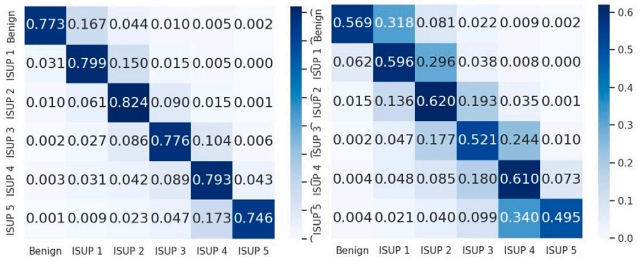
(iii) Predicted probability on external test 1 (n=844)

(iv) Predicted probability on external test 2 (n=4,675)

**Fig. 3.** Clinical-level experimental results for prostate cancer diagnosis. (i) the AUC curve on three test datasets; (ii), (iii), and (iv) is the prediction probability histogram on PANDA, external test 1, and external test 2, respectively.

From the histogram of slide classification prediction of PANDA, external test set 2, and external test set 1, cancerous and non-cancerous slides can be clearly separable. The complete cancer diagnosis results are shown in Table 3. On the external test set 2 and set 1, our system produces 1.7% and 2.2% false-negative slides (the slide score threshold value is 0.50 by default), respectively, and the corresponding false-positive is 11.1% and 13.6%. In clinical practice, it is required to recall more positive slides to avoiding missing cancerous cases. Thus, if the slide score threshold values is reduced to 0.40, only 0.9% false-negative are produced on external test 2 and 1.0% on the external test 1. Meanwhile, our system produces 26.8% false-positive slides on external test 2 and 25.9% on external test 2. An automated diagnosis system is required to be highly sensitive to positive slides for potential clinical use [31]. In the meantime, it should achieve an acceptable specificity. In practical workflow, pathologists can discard benign slides suggested by the AI system for further consideration due to the high sensitivity (>99.0%). The detected malignant slides need further pathological evaluations to avoid false positive.

For different test groups, this ratio value may vary slightly and the results indicate that our WSI analysis system can be applied for pre-screening part completely normal slides and reducing the workload of



(i) PANDA dataset (n=10,613) (ii) External test set 1 (n=844)

**Fig. 4.** ISUP grade confusion matrix between predictions and reference standard.

pathologists. We observed that some false negative slides showed small-sized cancers consisting of only several glands, whereas false-positive slides often exhibited strong stain variations or atrophic glands.

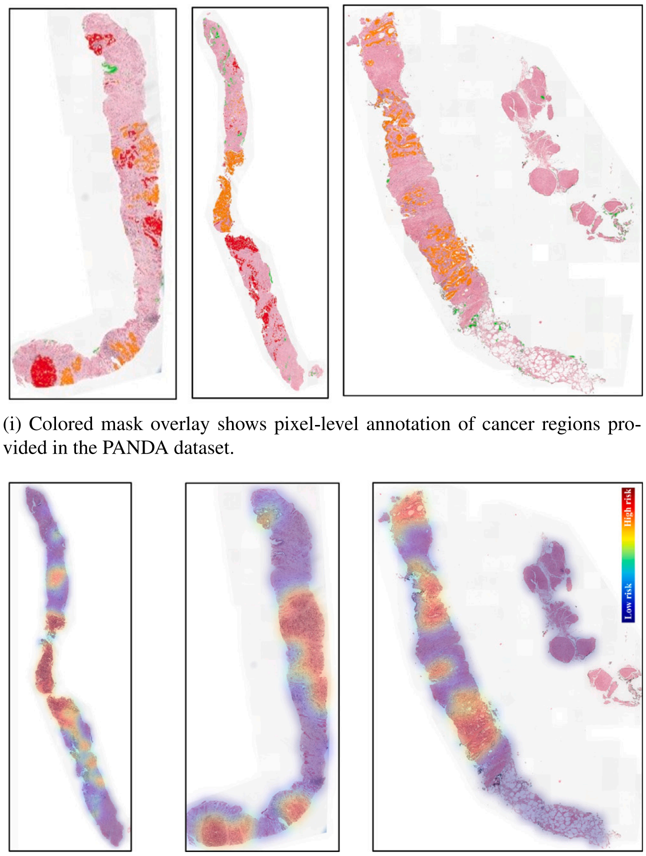
### 3.2. Assessment on Gleason grading

The grading of prostate cancer is a 6-level classification problem that aims to grade a WSI into one of the 6 levels, i.e., ISUP 0,1,2,3,4,5. Results in Table 4 highlight some observations regarding the Gleason grading performance. the proposed GCN-MIL model showed the accuracy of 0.785 (95% CI: 0.775–0.789), linear kappa $\kappa$ of 0.850 (95% CI: 0.847–0.854), and quadratic weighed kappa $\kappa_{quad}$ of 0.919 (95% CI: 0.911–0.924) in the training dataset ($n$ = 10,613) with 5-fold cross-validation split. The model generalize well on the independent test set ($n$ = 531) where most of the test set images were graded by multiple pathologists, resulting in the $\kappa_{quad}$ of 0.931. Accuracy and linear kappa are not evaluated as the test set cannot be accessed. On external test set 1 ($n$ = 844), the proposed model achieved the accuracy of 0.677 (95% CI: 0.665–0.694), linear kappa $\kappa$ of 0.723 (95% CI: 0.712–0.735), and quadratic weighed kappa $\kappa_{quad}$ of 0.805 (95% CI: 0.789–0.814). The prediction performance of the model was degraded when externally validated, which is greatly affected by the inconsistency of annotations from different centers considering that the training set was graded by pathologists from different nations and also might be attributed to domain shift of image quality caused by stain color variations, scanner types. Fig. 4 depicted the confusion matrices of PANDA and external test set. The independent test set is not shown because it cannot be accessed.

The performance comparison result of the proposed model to three baseline methods is presented in Table 4. More specifically, we compare three baseline methods, namely, Attention-MIL [26], CLAM-MB [19], and DSMIL [21], resulting in the $\kappa_{quad}$ of 0.765, 0.781, 0.728 on the independent test set, respectively. For external test set 1, Attention-MIL achieved the $\kappa_{quad}$ of 0.614, CLAM-MB 0.650, DSMIL 0.585. Attention-based MIL and CLAM-MB are built on the assumption that instances from one bag are independent and thus ignore the complex correlations among patches, which might hold true for binary classification problems like cancer diagnosis, but not for Gleason grading. DSMIL models the relations between key instance and other instances but still, it lacks context information of regions as required by grading.

As shown in Table 4, we develop three variants of GCN-MIL, i.e., GCN-MIL-1, GCN-MIL-2, and GCN-MIL-3., with different input features and training strategies. In specific, GCN-MIL-1 is trained with noisy labels from original reports, and the input features are commonly used ImageNet pretrained, the same as other competing methods. GCN-MIL-2 is trained with clean labels by progressively filtering out the noisy labels. GCN-MIL-3 is our full model which combines clean labels with TCGA pretrained features.

The results in Table 4 demonstrate that the ISUP grade prediction $\kappa_{quad}$ of our method increased from 0.780 of CLAM-MB to 0.855 on PANDA, from 0.781 of CLAM-MB to 0.853 on the internal test set, and from 0.650 of CLAM-MB to 0.715 on the external test set. One reason for this improvement is that our proposed GCN-MIL is context-aware with GCN by treating each patch as nodes and modeling the interactions. Further, the ISUP grade prediction $\kappa_{quad}$ improves with GCN-MIL-2 over GCN-MIL-1 by removing noisy samples with high uncertainty in the training dataset. The $\kappa_{quad}$ increases from 0.855 to 0.897 on PANDA, from 0.853 to 0.890 on the internal test set, and from 0.715 to 0.762 on the external test set 1. The best performance is achieved by GCN-MIL3 that integrates TCGA pre-trained ResNet50 to extract features and the robust learning strategy to remove noisy labels, resulting in the $\kappa_{quad}$ of 0.919 on 5-fold cross-validation, 0.931 on the internal test set, and 0.801 on the external test set 1.



(i) Colored mask overlay shows pixel-level annotation of cancer regions provided in the PANDA dataset.

(ii) Predicted heatmap overlay visualization weighted by attention (red indicates higher attention)

**Fig. 5.** Interpretability and visualization. Notice the close agreement between the dense attention scores and the ground truth. In practice, attention scores are computed per patch; here we used a sliding-window approach for dense visualization.

### 3.3. Model interpretability

Moving a step further away from statistical analysis, we made some model interpretability to uncover how deep learning models understand a given slide of prostate cancer. The human-readable interpretability of a well-trained weakly-supervised deep learning classifier can verify that the model's predictive basis is consistent with the well-known morphologies used by pathologists, and can also be used to analyze failure cases.

Additionally, an entire slide-level heatmap is available for AI-assisted human-in-the-loop clinical diagnosis. Our proposed GCN-MIL model makes final slide-level predictions by identifying and aggregating regions of high diagnostic importance (high attention scores) in WSI while ignoring regions of low diagnostic relevance (low attention

**Table 4**
Comparative study with baseline methods on training, internal and external test dataset.

| Methods | Accuracy(%) | Linear kappa $\kappa$ | Quadratic kappa $\kappa_{quad}$ |
|---|---|---|---|
| Cross-validation set ($n$ = 10,616) | | | |
| Attention-MIL [26] (2018) | 0.459(0.444 – 0.483) | 0.611(0.591 – 0.625) | 0.779(0.758 – 0.783) |
| CLAM-MB [19] (2021) | 0.459(0.441 – 0.487) | 0.611(0.589 – 0.626) | 0.780(0.760 – 0.792) |
| DSMIL [21] (2021) | 0.455(0.436 – 0.477) | 0.601(0.585 – 0.615) | 0.769(0.758 – 0.788) |
| GCN-MIL-1 | 0.623(0.616 – 0.631) | 0.736(0.731 – 0.738) | 0.855(0.849 – 0.859) |
| GCN-MIL-2 | 0.732(0.721 – 0.741) | 0.812(0.807 – 0.816) | 0.897(0.887 – 0.905) |
| GCN-MIL-3 | 0.785(0.775 – 0.789) | 0.850(0.847 – 0.854) | 0.919(0.911 – 0.924) |
| Internal Test set ($n$ = 531) | | | |
| Attention-MIL [26] (2018) | NA | NA | 0.765 |
| CLAM-MB [19] (2021) | NA | NA | 0.781 |
| DSMIL [21] (2021) | NA | NA | 0.728 |
| GCN-MIL-1 | NA | NA | 0.853 |
| GCN-MIL-2 | NA | NA | 0.894 |
| GCN-MIL-3 | NA | NA | 0.931 |
| External Test set 1 ($n$ = 844) | | | |
| Attention-MIL [26] (2018) | 0.576(0.553 – 0.591) | 0.559(0.533 – 0.596) | 0.614(0.581 – 0.667) |
| CLAM-MB [19] (2021) | 0.584(0.568 – 0.611) | 0.584(0.559 – 0.637) | 0.650(0.618 – 0.717) |
| DSMIL [21] (2021) | 0.424(0.421 – 0.426) | 0.428(0.398 – 0.447) | 0.585(0.549 – 0.611) |
| GCN-MIL-1 | 0.544(0.532 – 0.559) | 0.602(0.590 – 0.624) | 0.715(0.701 – 0.743) |
| GCN-MIL-2 | 0.617(0.589 – 0.636) | 0.667(0.626 – 0.695) | 0.762(0.719 – 0.792) |
| GCN-MIL-3 | 0.677(0.665 – 0.694) | 0.723(0.712 – 0.735) | 0.801(0.789 – 0.814) |

- GCN-MIL-1: ImageNet pretrained feature + noisy labels (same settings with comparative MIL methods).
- GCN-MIL-2: ImageNet pretrained feature + clean labels.
- GCN-MIL-3: TCGA pretrained feature + clean labels.

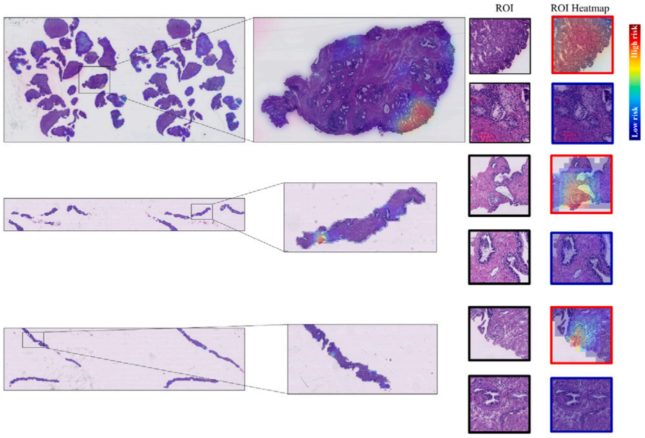
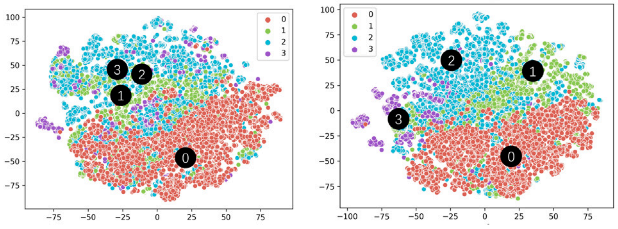


**Fig. 6.** Our model is highly sensitive to even *tiny cancerous regions* while refracting from misjudging non-cancer regions, as heatmaps of external validation show.

scores). To visualize and interpret the relative importance of each region in the WSI, we can generate fine-grained attention heatmaps by using overlapping patches (e.g., 80% overlap) and averaging the attention scores of overlapping regions.

Fig. 5(i) shows the pixel-level annotations overlayed with the original slide in the PANDA dataset, and Fig. 5(ii) is the corresponding heatmap predicted from the attention. Although we did not use any pixel-level or patch-level annotation to train the model explicitly which specific instances in a slide to attend, the results demonstrated that through weakly supervised learning using slide-level labels only, trained GCN-MIL models are generally capable of delineating the boundary between tumor and normal tissue. The attention heatmaps exhibit a high level of agreement with the pathologist annotations of tumour regions when evaluated on predicted slides. Fig. 6 shows our model is highly sensitive to even very tiny cancerous regions while refracting from misjudging non-cancer regions on external validation data.



(i) Features of ImageNet pre-trained ResNet50.)

(ii) Features of TCGA-pretrained ResNet50.

**Fig. 7.** Feature of SICAPv2 dataset [37] with different pre-trained models.

To enhance more interpretability of our model, we further investigated the patch-level feature space learned by our feature extractor ResNet50 compared with the commonly-used ImageNet pre-trained ResNet50. We analyzed the distribution of the features extracted from patches of different Gleason scores by using the SICAPv2 dataset [37] containing 10,340 annotated patches from 155 slides which were down-sampled to 10x resolution and divided into patches of size 512 × 512 and overlap of 50% between them. In specific, there are 4417 patches of benign, 1636 patches of Gleason 3, 3622 patches of Gleason 4, 665 patches of Gleason 5. The dimension-reduced features extracted with ImageNet pre-trained ResNet 50 and from TCGA pre-trained ResNet 50 by t-SNE [38] on SICAPV2 are shown in Fig. 7.

For ImageNet pre-trained model in Fig. 7(a), though the features of benign patches can be separated from cancerous ones, it is hard to distinguish between different Gleason scores as they aligned together without distinctive cluster centroids. In comparison, Fig. 7(b) represents the feature vector distributions extracted with TCGA pre-trained ResNet50 for the patch images containing Gleason patterns 3, 4, or 5 and shows that apparent differences exist among Gleason pattern-wise distributions. Patches of different Gleason score are separated into distinct clusters in the feature space and exhibit morphology characteristic of their respective subtype. Thus, we assume that self-supervised learning enables the ResNet 50 model to assign distinguishable features to different Gleason patterns, making the GCN-MIL model predict the grade groups based on the generated feature map. The results indicate that the learned features represent prostate cancer morphology well and this is one key reason why our system has good generalization for unseen datasets from different centers.

## 4. Discussion

### 4.1. Overall performance

In the present study, we developed an automated system for prostate cancer screening and grading across full-range Gleason scores validated on multi-center datasets. The current study moved a step further on the development of an accurate and interpretable deep learning model by weak supervision using only slide-level labels for prostate cancer screening and full range Gleason grading on the basis of more than 10,000 training slides and 6,000 validation slides. This method is particularly suitable for gigapixel WSIs with only slide-level labels using self-supervised learning to reduce the image dimension and using GCN to enforce context awareness for predictions. According to the analysis results, we suggest that using sufficiently large-scale histopathological images and slide-level annotations, one is capable of developing a highly accurate machine learning system for clinical-grade usages. This machine-learning system could assist clinical prostate cancer screening and grading to reduce inter-observer variability.

Our GCN-MIL model yielded the AUC in the range of 0.985–0.992 for cancer screening. For Gleason grading, it achieved the quadratic weight kappa $\kappa_{quad}$ value of 0.931 and 0.810 for internal and external test set compared to the pathologists-based reference standards, respectively. In literature, the inter-observer Gleason scoring concordance rates measured in linear kappa vary in the range of 0.40–0.50 between general pathologists, and 0.56–0.70 for urologic pathologists [39]. Another recent machine learning study reported $\kappa_{quad}$ 0.918 on the internal test dataset and inter-observer variability on the external test set with reference to two pathologists are 0.723 and 0.707 [5].

The degradation of prediction performance when externally validated might cause by the variation in image quality such as color distribution. The stain color distribution of the external dataset was visually different from the training dataset. We have incorporated color augmentation in self-supervised pre-training model for feature extraction to cope with the stain color variations among institutions, but it might not have been sufficient to deal with the external dataset. It is potential to apply advanced data-preprocessing techniques including normalization, transformation, filtering, and feature extractions to remove unwanted experimental/imaging variation and technical error [40–42].

### 4.2. Related works

Most previous studies using machine learning tools have focused on supervised learning by collecting pixel-level annotations. There exist some earlier trials on semi-supervised or weakly supervised learning to mitigate annotation requirements but there are still limitations. For example, Bulten et al. proposed a semi-automatic labeling technique to circumvent the need for full manual annotation by pathologists. Still, they required pixel-level annotations for the epithelium segmentation model but annotations demands have been alleviated compared with Gleason grading. The seminal work of MIL-RNN [20] probably produced the highest-end result that can be reached by the implementation of a slide-level weakly supervised learning on an enormously large dataset (more than 10,000 slides) without human annotations. Lu et al. [19] further developed an interpretable weakly supervised deep-learning method for data-efficient WSI processing and learning that only requires slide-level labels. While these proposed weakly supervised learning models can reach high cancer diagnosis AUC, their applications for the full range of Gleason grading are still limited. Some weakly supervised work of grading of prostate biopsies have focused solely on small datasets or subsets of Gleason patterns, and they have not analyzed the clinical implications of the introduction of AI-assisted prostate pathology [14,43,44].

In the comparative analysis, our proposed GCN-MIL method outperformed other baseline MIL methods by a large margin. We highlight two strengths of our model: feature representation with self-supervised learning and graph neural network. Using an ImageNet pre-trained model as feature extractor is common for previous MIL models. We discover in our study that by conducting self-supervised training on the TCGA dataset with 30,072 WSIs, convolution neural network can extract prostate cancer-specific histological features into our proposed system leads to better performance. As indicated by the visualization results on the patch-annotated SICAPv2 dataset, we discovered that there were vivid performance changes for Gleason pattern discrimination compared with the ImageNet pre-trained feature extractor. This supports the assumption that our self-supervised learning for feature extraction model might have actually learned the Gleason pattern-specific features. On the other hand, most previous MIL models are based on the assumption that instances are independent of each other, which is not true in the prostate cancer grading problem because it requires quantification of regions. We introduced a graph neural network by treating each instance as a node in the graph and their spatial information as edges connecting nodes to capture region information. As the comparative results in Table 4 show, while the attention-based MIL method is known as a powerful one in the weakly supervised learning setting [19,21], adopting the CLAM or DSMIL model was not effective for Gleason grading.

### 4.3. Visualization

In addition to the statistical analysis of the prediction performance of the proposed model, we attempted to analyze the model mechanism in terms of visual interpretability that is intuitive and tractable to readers. Our GCN-MIL model uses global attention pooling to identify and aggregate all instances in a slide to make final prediction and thus, the attention score can reveals the contribution and significance of each instance to final output. In Fig. 5, we generate an attention heatmap by converting the attention scores for the predicted slides and mapping the normalized scores to their corresponding spatial location in the original slide. In the predicted heatmap of test slides, high-attention regions generally show high correspondence with region-level annotations already established and recognized from the training set for cancerous prostate glands. Nonetheless, this simple and intuitive interpretability and visualization technique can provide researchers insight into the morphological patterns dominating the predictions of the model.

### 4.4. Limitations

Nevertheless, this study has some limitations related to the size and reference standard of the study data. The reference standard was not strongly established. Gleason grading is known to be highly variable among observers. The internal test set ($n$ = 531) was labeled by three pathologists whereas the external test set was evaluated using a weak reference standard derived from either a single pathologist or the original hospital diagnosis. While this research aims to reduce the development cost of AI systems, the clinical utility of the research will be better demonstrated with stronger reference standards. Future research could involve the participation of an increased number of hospitals and pathologists. Moreover, although our study primarily focuses on prostate cancer diagnosis and grading, it would be an interesting extension of GCN-NIL for applications like multi-modal data fusion for drug discovery [45–48], gene mutation prediction [49], etc.

## 5. Conclusions

In this paper, a weakly-supervised deep learning model called MIL-GCN is proposed for the diagnosis and grading of prostate cancer. MIL-GCN employs a self-supervised pretrained encoder to extract features, and then the features are aggregated using GCN to make it context-aware across instances. Finally, the model predicts a label with attention-based MIL sum over all instances. The experimental results suggested that our weakly-supervised deep learning model combined with a graph neural network could provide an accurate and reliable method of screening and grading prostate cancer, thus accelerating the diagnosis and enhancing the consistency of pathologists.

One finding of this paper is that it is potential to develop an automated Gleason grading system using slide-level labels to achieve clinical-grade performance, without pixel-level annotations. Another finding is that by integrating multiple techniques, e.g., weakly supervised learning, graph convolution, and robust training, into the weakly supervised learning framework we can offer a practical workflow for clinical diagnosis and grading of prostate cancer across different datasets collected from different centers. For future research, it would be interesting to explore pathology-specific self-supervised learning beyond contrastive learning which is initially designed for natural images. Various feature aggregation networks, e.g., transformer with self-attention mechanism, can be explored to make the weakly supervised model context aware.

## Availability of data and materials

The private test data during the current study is not publicly available due to restrictions in the ethical permit. PANDA can be accessed through https://www.kaggle.com/c/prostate-cancer-grade-assessment. Diagset can be accessed through https://ai-econsilio.diag.pl/. SICAP can be accessed through https://data.mendeley.com/datasets/9xxm58dvs3/1. The computer code of the deep learning model will be available on publishing.

## CRediT authorship contribution statement

**Jinxi Xiang:** Methodology, Writing. **Xiyue Wang:** Conceptualization, Methodology. **Xinran Wang:** Conceptualization, Investigation. **Jun Zhang:** Writing, Supervision. **Sen Yang:** Visualization, Validation. **Wei Yang:** Validation. **Xiao Han:** Supervision. **Yueping Liu:** Supervision, Reviewing.

## Acknowledgments

All authors approved the version of the manuscript to be published.

## 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.

### *Ethics approval and consent to participate*

This study was approved by the Research Ethics Committee of The Fourth Hospital of Hebei Medical University, China.

### *Consent for publication*

The consent approved by the Institutional Review Board of the Fourth Hospital of Hebei Medical University was signed.

## References

[1] P. Rawla, Epidemiology of prostate cancer, World J. Oncol. 10 (2) (2019) 63.
[2] B.T. Alenezi, M.H. Alsubhi, X. Jin, G. He, Q. Wei, Y. Ke, Global development on causes, epidemiology, aetiology, and risk factors of prostate cancer: An advanced study, 2021.
[3] J.I. Epstein, L. Egevad, M.B. Amin, B. Delahunt, J.R. Srigley, P.A. Humphrey, The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma, Am. J. Surg. Pathol. 40 (2) (2016) 244–252.
[4] J.I. Epstein, M.J. Zelefsky, D.D. Sjoberg, J.B. Nelson, L. Egevad, C. Magi-Galluzzi, A.J. Vickers, A.V. Parwani, V.E. Reuter, S.W. Fine, et al., A contemporary prostate cancer grading system: a validated alternative to the gleason score, Euro. Urol. 69 (3) (2016) 428–435.
[5] W. Bulten, H. Pinckaers, H. van Boven, R. Vink, T. de Bel, B. van Ginneken, J. van der Laak, C. Hulsbergen-van de Kaa, G. Litjens, Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study, Lancet Oncol. 21 (2) (2020) 233–241.
[6] A.H.M. Linkon, M.M. Labib, T. Hasan, M. Hossain, et al., Deep learning in prostate cancer diagnosis and gleason grading in histopathology images: an extensive study, Inf. Med. Unlocked 24 (2021) 100582.
[7] L. Pantanowitz, G.M. Quiroga-Garza, L. Bien, R. Heled, D. Laifenfeld, C. Linhart, J. Sandbank, A.A. Shach, V. Shalev, M. Vecsler, et al., An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study, Lancet Digit. Health 2 (8) (2020) e407–e416.
[8] G. Nir, D. Karimi, S.L. Goldenberg, L. Fazli, B.F. Skinnider, P. Tavassoli, D. Turbin, C.F. Villamil, G. Wang, D.J. Thompson, et al., Comparison of artificial intelligence techniques to evaluate performance of a classifier for automatic grading of prostate cancer from digitized histopathologic images, JAMA Netw. Open 2 (3) (2019) e190442.
[9] A. Janowczyk, A. Madabhushi, Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases, J. Pathol. Inform. 7 (2016).
[10] A. Echle, N.T. Rindtorff, T.J. Brinker, T. Luedde, A.T. Pearson, J.N. Kather, Deep learning in cancer pathology: a new generation of clinical biomarkers, Br. J. Cancer 124 (4) (2021) 686–696.
[11] A. Gertych, N. Ing, Z. Ma, T.J. Fuchs, S. Salman, S. Mohanty, S. Bhele, A. Velásquez-Vacca, M.B. Amin, B.S. Knudsen, Machine learning approaches to analyze histological images of tissues from radical prostatectomies, Comput. Med. Imaging Graph. 46 (2015) 197–208.
[12] T.H. Nguyen, S. Sridharan, V. Macias, A. Kajdacsy-Balla, J. Melamed, M.N. Do, G. Popescu, Automatic gleason grading of prostate cancer using quantitative phase imaging and machine learning, J. Biomed. Opt. 22 (3) (2017) 036015.
[13] A.A. Abbasi, L. Hussain, I.A. Awan, I. Abbasi, A. Majid, M.S.A. Nadeem, Q.-A. Chaudhary, Detecting prostate cancer using deep learning convolution neural network with transfer learning approach, Cogn. Neurodyn. 14 (4) (2020) 523–533.
[14] E. Arvaniti, K.S. Fricker, M. Moret, N. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P.J. Wild, J.H. Rueschoff, M. Claassen, Automated gleason grading of prostate cancer tissue microarrays via deep learning, Sci. Rep. 8 (1) (2018) 1–11.
[15] Y. Tolkach, T. Dohmgörgen, M. Toma, G. Kristiansen, High-accuracy prostate cancer pathology using deep learning, Nat. Mach. Intell. 2 (7) (2020) 411–418.
[16] K. Nagpal, D. Foote, Y. Liu, P.-H.C. Chen, E. Wulczyn, F. Tan, N. Olson, J.L. Smith, A. Mohtashamian, J.H. Wren, et al., Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer, NPJ Digit. Med. 2 (1) (2019) 1–10.
[17] N. Marini, M. Atzori, S. Otálora, S. Marchand-Maillet, H. Müller, H&e-adversarial network: a convolutional neural network to learn stain-invariant features through hematoxylin & eosin regression, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 601–610.
[18] N. Coudray, P.S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A.L. Moreira, N. Razavian, A. Tsirigos, Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning, Nat. Med. 24 (10) (2018) 1559–1567.
[19] M.Y. Lu, D.F. Williamson, T.Y. Chen, R.J. Chen, M. Barbieri, F. Mahmood, Data-efficient and weakly supervised computational pathology on whole-slide images, Nat. Biomed. Eng. 5 (6) (2021) 555–570.
[20] G. Campanella, M.G. Hanna, L. Geneslaw, A. Miraflor, V.W.K. Silva, K.J. Busam, E. Brogi, V.E. Reuter, D.S. Klimstra, T.J. Fuchs, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nat. Med. 25 (8) (2019) 1301–1309.
[21] B. Li, Y. Li, K.W. Eliceiri, 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, 2021, pp. 14318–14328.

[22] M.Y. Lu, R.J. Chen, J. Wang, D. Dillon, F. Mahmood, Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding, 2019, arXiv preprint arXiv:1910.10825.

[23] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.

[24] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.

[25] X. Chen, H. Fan, R. Girshick, K. He, Improved baselines with momentum contrastive learning, 2020, arXiv preprint arXiv:2003.04297.

[26] M. Ilse, J. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: International Conference on Machine Learning, PMLR, 2018, pp. 2127–2136.

[27] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, Y. Zhang, Transmil: Transformer based correlated multiple instance learning for whole slide image classication, 2021, arXiv preprint arXiv:2106.00908.

[28] Y. Guan, J. Zhang, K. Tian, S. Yang, P. Dong, J. Xiang, W. Yang, J. Huang, Y. Zhang, X. Han, 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, 2022, pp. 18813–18823.

[29] M. Rostami, S. Forouzandeh, K. Berahmand, M. Soltani, M. Shahsavari, M. Oussalah, Gene selection for microarray data classification via multi-objective graph theoretic-based method, Artif. Intell. Med. 123 (2022) 102228.

[30] S. Azadifar, M. Rostami, K. Berahmand, P. Moradi, M. Oussalah, Graph-based relevancy-redundancy gene selection method for cancer diagnosis, Comput. Biol. Med. 147 (2022) 105766.

[31] P. Ström, K. Kartasalo, H. Olsson, L. Solorzano, B. Delahunt, D.M. Berney, D.G. Bostwick, A.J. Evans, D.J. Grignon, P.A. Humphrey, et al., Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study, Lancet Oncol. 21 (2) (2020) 222–232.

[32] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, J. Huang, W. Yang, X. Han, Transpath: Transformer-based self-supervised learning for histopathological image classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 186–195.

[33] X. Wang, Y. Du, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, X. Han, RetCCL: clustering-guided contrastive learning for whole-slide image retrieval, Med. Image Anal. (2022) 102645.

[34] G. Li, M. Muller, A. Thabet, B. Ghanem, Deepgcns: Can gcns go as deep as cnns? in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9267–9276.

[35] W. Bulten, K. Kartasalo, P.-H.C. Chen, P. Ström, H. Pinckaers, K. Nagpal, Y. Cai, D.F. Steiner, H. van Boven, R. Vink, et al., Artificial intelligence for diagnosis and gleason grading of prostate cancer: the PANDA challenge, Nat. Med. (2022) 1–10.

[36] M. Koziarski, B. Cyganek, B. Olborski, Z. Antosz, M. Żydak, B. Kwolek, P. Wkasowicz, A. Bukała, J. Swadźba, P. Sitkowski, DiagSet: a dataset for prostate cancer histopathological image classification, 2021, arXiv preprint arXiv:2105.04014.

[37] J. Silva-Rodríguez, A. Colomer, M.A. Sales, R. Molina, V. Naranjo, Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection, Comput. Methods Programs Biomed. 195 (2020) 105637.

[38] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).

[39] T.A. Ozkan, A.T. Eruyar, O.O. Cebeci, O. Memik, L. Ozcan, I. Kuskonmaz, Interobserver variability in gleason histological grading of prostate cancer, Scand. J. Urol. 50 (6) (2016) 420–424.

[40] J. Tang, M. Mou, Y. Wang, Y. Luo, F. Zhu, MetaFS: performance assessment of biomarker discovery in metaproteomics, Brief. Bioinform. 22 (3) (2021) bbaa105.

[41] J. Fu, Y. Zhang, J. Liu, X. Lian, J. Tang, F. Zhu, Pharmacometabonomics: data processing and statistical analysis, Brief. Bioinform. 22 (5) (2021) bbab138.

[42] Q. Yang, J. Hong, Y. Li, W. Xue, S. Li, H. Yang, F. Zhu, A novel bioinformatics approach to identify the consistently well-performing normalization strategy for current metabolomic studies, Brief. Bioinform. 21 (6) (2020) 2142–2152.

[43] M. Lucas, I. Jansen, C.D. Savci-Heijink, S.L. Meijer, O.J. de Boer, T.G. van Leeuwen, D.M. de Bruin, H.A. Marquering, Deep learning for automatic gleason pattern classification for grade group determination of prostate biopsies, Virchows Archiv 475 (1) (2019) 77–83.

[44] W. Huang, R. Randhawa, P. Jain, K.A. Iczkowski, R. Hu, S. Hubbard, J. Eickhoff, H. Basu, R. Roy, Development and validation of an artificial intelligence–powered platform for prostate cancer grading and quantification, JAMA Netw. Open 4 (11) (2021) e2132554.

[45] F. Li, Y. Zhou, Y. Zhang, J. Yin, Y. Qiu, J. Gao, F. Zhu, POSREG: proteomic signature discovered by simultaneously optimizing its reproducibility and generalizability, Brief. Bioinform. 23 (2) (2022).

[46] J. Fu, Y. Zhang, Y. Wang, H. Zhang, J. Liu, J. Tang, Q. Yang, H. Sun, W. Qiu, Y. Ma, et al., Optimization of metabolomic data processing using NOREVA, Nat. Protoc. 17 (1) (2022) 129–151.

[47] J. Tang, J. Fu, Y. Wang, B. Li, Y. Li, Q. Yang, X. Cui, J. Hong, X. Li, Y. Chen, et al., ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies, Brief. Bioinform. 21 (2) (2020) 621–636.

[48] J. Tang, J. Fu, Y. Wang, Y. Luo, Q. Yang, B. Li, G. Tu, J. Hong, X. Cui, Y. Chen, et al., Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains*[S], Mol. Cell. Proteomics 18 (8) (2019) 1683–1699.

[49] Q. Yang, B. Li, J. Tang, X. Cui, Y. Wang, X. Li, J. Hu, Y. Chen, W. Xue, Y. Lou, et al., Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data, Brief. Bioinform. 21 (3) (2020) 1058–1068.