Efficient Grading of Prostate Cancer WSI with Deep Learning

Riddhasree Bhattacharyya[a], Paromita Roy[b], Sugata Banerji[c], and Sushmita Mitra[a]

[a]Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, West Bengal, India.
[b]Department of Pathology, Tata Medical Center, Kolkata 700160, West Bengal, India
[c]Department of Mathematics and Computer Science, Lake Forest College, Lake Forest, IL 60045, USA.

ABSTRACT

The field of histopathology, which involves visual examination of tissue samples at a microscopic scale, is very important for the diagnosis of cancer. Although this task is currently performed by human experts, the design of computer vision-based systems to assist human experts is an interesting research area. This problem is ideal for the application of computer-based image analysis; especially, with the great success of convolutional neural networks (CNNs) in image segmentation and classification in the last decade. However, applying CNNs to this problem is challenging for a number of reasons, such as excessive high resolution (involving huge computational burden), variations in sample processing, and insufficient annotation. In this current work, we propose a CNN-based approach to tackle the problem of prostate cancer grading from Whole Slide Images (WSIs). We use a patch-based, multi-step training algorithm to address the challenges of large image size, tissue sample variations and partial annotation. Then we propose two novel classification strategies using an ensemble of CNN models to classify tissue slide images into different ISUP grades (1 – 5). We demonstrate the efficacy of our method on the publicly available large scale Prostate cANcer graDe Assessment (PANDA) Challenge dataset. The effectiveness of the technique is measured using Cohen's quadratic kappa score. The results are shown to be highly accurate (kappa score of 0.88) and better than other leading state-of-the art methods.

Keywords: histopathology, prostate cancer, ISUP grading, WSI, computer vision, CNN

1. INTRODUCTION AND BACKGROUND

Prostate cancer is the second most common cancer affecting men worldwide. The Gleason scoring system, used by the pathologists to determine the stage of cancer and the course of a patient's treatment, is greatly prone to intra-observer and inter-observer variability.[1] A new grading system (named as ISUP grading system) was introduced after the 2014 conference by the International Society of Urological Pathology.[2] The ISUP grading system (based on Gleason grading) considers the two most prominent Gleason grades (in terms of proportion and severity) in the malignant prostate biopsies and then maps it to an ISUP grade (1–5). Benign biopsies are given ISUP grade 0. Computer Aided Diagnosis (CAD) systems can reduce the workload of pathologists and alleviate the problem of subjectivity in Gleason grading. But fully-supervised deep learning algorithms are data-hungry. So, huge and diverse data of Whole Slide Images of prostate biopsies (covering different possible cancerous structures), annotated at a deeper level of granularity, are needed by the supervised algorithms to provide accurate results. Considering the large size of Whole Slide Images (WSIs), curating such large datasets with pixel-level annotations from pathologists becomes unfeasible most of the time. In these circumstances, weakly supervised learning algorithms can learn to predict the ISUP grade of a WSI using only the slide-level label. Hence, the target should be to obtain greater degree of accuracy using only the slide-level labels.

The proposed methods in the literature for designing efficient Computer-Aided prostate cancer diagnosis/grading system using digitized prostate biopsies can be categorized based on the following two criteria:

Further author information: (Send correspondence to Sugata Banerji)
Riddhasree Bhattacharyya: E-mail: riddhasreeb@gmail.com
Paromita Roy: E-mail: paromita.roy@tmckolkata.com
Sugata Banerji: E-mail: banerji@lakeforest.edu
Sushmita Mitra: E-mail: sushmita@isical.ac.in

Medical Imaging 2024: Digital and Computational Pathology, edited by John E. Tomaszewski, Aaron D. Ward, Proc. of SPIE Vol. 12933, 1293305 · © 2024
SPIE · 1605-7422 · doi: 10.1117/12.3006835

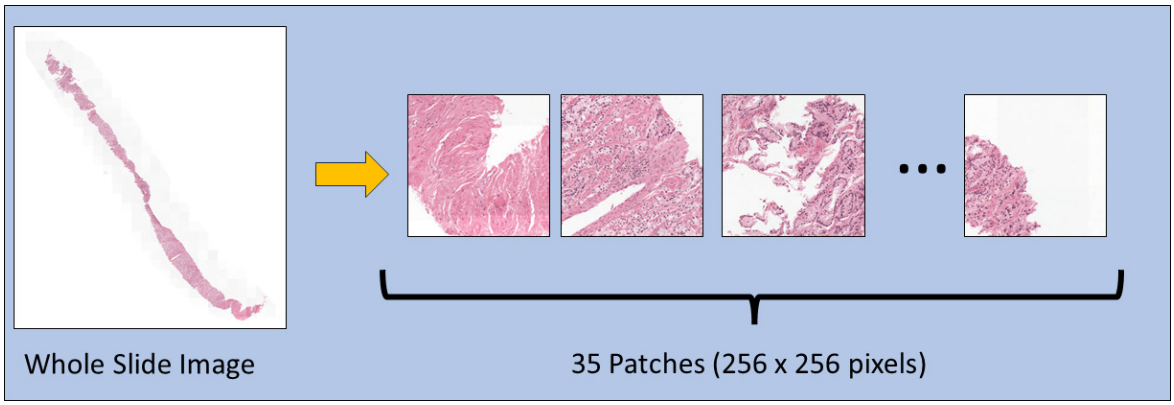

Figure 1. A WSI from the PANDA dataset and sample patches generated from it.

in terms of the learning framework used and in terms of the task accomplished by the proposed frameworks. Considering the learning framework, the proposed methods can be classified into:

Bottom-up frameworks: Here, the patches are first classified into Gleason grades using fully supervised training and the predictions of the patches are then combined to obtain WSI-level prediction. Many initial works in prostate cancer diagnosis and grading[3–5] were based on these frameworks. However, the downside of these frameworks is the amount of images with fine-grained annotations required for training the models.

Top-down frameworks: These frameworks mainly use multiple instance learning (a form of weakly supervised learning) approaches where the models can be trained using only the slide-level labels (readily available from clinical records). Among the proposed methods in the literature that used this framework, the instance-based Multiple Instance Learning (MIL) and the bag-based Multiple Instance Learning (MIL) approaches were commonly witnessed in the prostate histology domain. While some researchers[6] focused on instance-based MIL approaches (where some degree of noise is introduced in the training process as the instances are pseudo-labelled using the slide-level label), several researchers[7,8] also used bag-based approaches (where no noise is introduced during the training process due to the absence of pseudo-labelling) using patch-based image reconstruction strategy.

Considering the task accomplished by the proposed methods in the literature, different researchers have focused on the prostate cancer grading task at different levels of granularity. Nagpal et al.[9] developed a two-stage classification system to classify WSIs of prostate biopsies into four Gleason grade groups (1, 2, 3, 4/5). Li et al.[10] proposed an attention-based multi-resolution model for grading the prostate WSIs as benign, low-grade (Gleason grade 3+3 or 3+4) or high-grade (Gleason grade 4+3 or higher). Bulten et al.[11] designed a deep-learning system to perform 6-class classification based on the Gleason score. A weakly-supervised method based on cross-slide contrastive learning was designed by Wang et al.[12] to perform binary classification (benign vs cancerous) of prostate biopsies among other cancer datasets. Recently, Toledo et al.[13] presented a quantum-inspired deep-probabilistic learning ordinal regression model to assign one among five Gleason scores (6, 7, 8, 9, 10) to the prostate WSIs. Other researchers[8,14–16] used different techniques to predict the ISUP grades (0, 1, 2, 3, 4, 5) of WSIs of prostate biopsies.

In the following sections, we propose two methods to predict the ISUP grades (0, 1, 2, 3, 4, or 5) of WSIs of prostate biopsies. The first method is an end-to-end deep learning approach where we use the Multiple Instance Learning (bag-based approach) paradigm of Weakly Supervised learning to train our backbone model. The second method builds upon the first method using the trained models of the first method for feature extraction from the WSIs. This paper contributes to the existing literature on prostate histology in view of the following points. Here, we attempt to find out the ISUP grades that have the possibility of greater confusion among themselves, analyze the potential causes of confusion, and accordingly adopt a divide and conquer kind of approach by training different models on different subsets of ISUP grades. Training different models for different subsets of ISUP grades helps in learning the features of the grades better and these features can be effectively utilized

by other methods. Also, we demonstrate the effectiveness of our method using the largest publicly available dataset of prostate biopsies: the PANDA[17] dataset.The dataset being relatively new, relatively few works have attempted to perform ISUP grading using this dataset. Among the few works that did experiments with ISUP grading using the PANDA dataset, their focus was on other related issues like grading of local cancerous patterns using only slide-level labels of prostate WSIs,[14] improving the performance on hard cases with small tumour areas,[16] comparing the performance of different loss functions for classifying prostate cancer images[15] etc, with the task of ISUP grading merely being a subpart of many papers.

2. METHOD

Since there is an ordering to the classes or grades in this multi-class classification problem, we pose this as an ordinal regression problem. Contrary to the way a normal classifier penalizes each misclassification equally, ordinal regression models consider the distance between the predicted and actual labels. To achieve this, each label is mapped to a five-element binary vector, where sum of the vector is equal to the corresponding ISUP grade. For example: ISUP grade 0 is mapped to [0, 0, 0, 0, 0] whereas ISUP grade 5 is mapped to [1, 1, 1, 1, 1]. Then, this problem becomes equivalent to a multi-label binary classification problem and binary cross entropy (BCE) loss is used for training the models.

2.1 Proposed method 1: ensemble of CNN models

Here, a WSI is considered as a bag of 35 (256 × 256) patches extracted from the image. Then, each patch is converted to a low-dimensional embedding by passing them through the convolutional part of the EfficientNet B1 model.[18] The low-dimensional embeddings of the patches are then concatenated before passing through the Generalized Mean Pooling (GeM pooling) and fully connected layers.[19] This form of training can work with the absence of patch-level labels, and no noise is introduced in the training process as only the bag-level label is used. The robust and efficient GeM pooling operation has shown promising results in the tasks of image classification. It has a learnable parameter that allows the network to adapt to different tasks automatically. After flattening the feature vector, a Squeeze and Excitation module[20] is used before passing it through a fully-connected layer for final classification. As we know not all features are equally important for classification, the Squeeze and Excitation module learns to weigh each feature according to its relevance.

After training the above model M_1 using the PANDA dataset, the confusion matrix of the validation data, shown in Figure 2(a) showed a large degree of confusion among ISUP grades 3, 4, 5 and among the ISUP grades 1, 2. In particular, a lot of images with ISUP grade 3 (4 + 3), ISUP grade 5 (4 + 5) were misclassified as ISUP grade 4 and ISUP grade 2 (3 + 4) were misclassified as ISUP grade 1(3 + 3) [The brackets following the

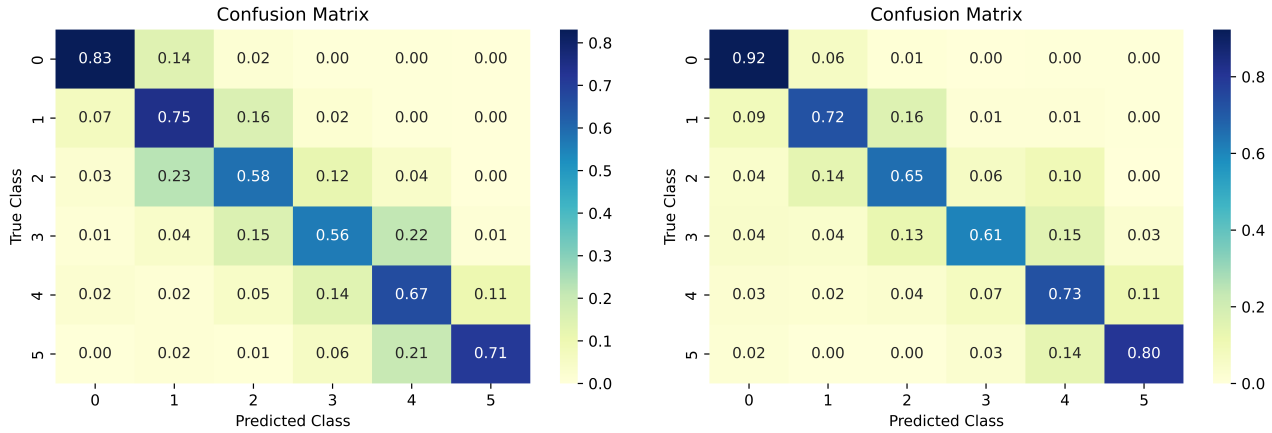

Figure 2. Confusion matrix for prostate cancer grading based on a single CNN model (a) and a multilevel ensemble of CNN models (b).

ISUP grades indicate the Gleason grades corresponding to the ISUP grade with the first value indicating the majority Gleason pattern and the second value indicating the minority Gleason pattern]. From such output, it seems very likely that model M_1 is finding it difficult to capture the features of the minority Gleason pattern 3 and minority Gleason pattern 5 respectively in many ISUP grade 3 (4 + 3) and ISUP grade 5 (4 + 5) images and consequently misclassifying them as ISUP grade 4 (4 + 4). The same argument goes for the confusion among images belonging to ISUP grades 1, 2. On analyzing the misclassified images, it was observed that the minority Gleason pattern was not only very small in amount relative to the majority Gleason pattern but also the minority Gleason patterns were overlapping with the majority Gleason patterns in most of the misclassified images' patches. This makes it difficult for the model M_1 to capture the features of the minority Gleason pattern.

This leads us to use a multilevel ensemble approach. Precisely, we train a family of models (M_2 family) with different patch sizes (192 × 192, 224 × 224, 256 × 256) using only data belonging to ISUP grades 3, 4, 5, a model M_3 using data belonging to ISUP grades 0, 1, 2 and another family of models (M_4 family) with different patch sizes (192 × 192, 224 × 224, 256 × 256) using only data belonging to ISUP grades 1, 2. The intuition for training different models for different subsets of ISUP grades is that the models will be able to capture the features of those particular subsets better, thus resolving the confusion among the corresponding data subsets to a greater extent. Also, different patch sizes are considered for training different models on a subset of ISUP grades with the argument that the features of the minority Gleason pattern will be effectively captured amidst the majority Gleason pattern by varying the receptive field of different models and ensembling them. Finally, the prediction is done using these models in the following way: If the model M_1 classifies an image as grade 0, 1 or 2, the prediction is deferred to model M_3. Now, if the model M_3 classifies an image as ISUP grade 0, the prediction of M_3 is taken. But if M_3 classifies an image as ISUP grade 1 or 2, the M_4 family models (trained on grades 1 and 2) are used for obtaining the final prediction. Again, if the model M_1 classifies an image as grade 3, 4 or 5, the decision is deferred to the M_2 family models (trained on grades 3, 4, 5); with the ensembled decision of the M_2 family being taken as the final prediction. The clear improvement in results can be seen from the ensemble confusion matrix, shown in Figure 2(b).

2.2 Proposed method 2: bag of CNN features

The second method uses the trained models M_1 and a model from the M_2 family (trained on 256 × 256 patches) to obtain a low-dimensional meaningful representation for each patch.[21] To be more specific, the features of the patches belonging to images having ISUP grades 0, 1, or 2 are extracted using the model M_1, and those belonging to images having ISUP grades 3, 4, or 5 are extracted using the selected model from the M_2 family. Next, these representations of the patches are clustered using k-means clustering. Next, a codeword for each WSI or a slide-level representation is obtained by calculating the proportional part of the patches (of a WSI) belonging to each of the clusters. Thereafter, these derived codewords for the images are saved in a tabular format along with the labels. Classifiers (M_5) based on gradient-boosted decision trees like LGBM,[22] and XGBoost[23] are trained using this tabular data. During inference, the features of the image patches are extracted using the model M_1 or the selected model from M_2 family depending on whether the image belongs to ISUP grades 0, 1, 2 or ISUP grades 3, 4, 5 (as predicted by the model M_1). Next, a codeword is obtained for each WSI using the extracted features from the previous step and the k-means model created during training; with the ISUP grade being predicted after passing these codewords through the classifier M_5 used during training.

To improve the classification performance further, one classifier M_6 was also trained with this approach using only data belonging to ISUP grades 3, 4, 5 and another classifier M_7 was trained using only data belonging to ISUP grades 1, 2. Then, a multi-level ensemble was formed with all the models trained using this approach. This means that during prediction, if the M_5 classifier classifies an image as ISUP grade 0, the prediction of M_5 is taken. But, if M_5 classifies an image as one belonging to any of ISUP grades 3, 4, 5, the decision is deferred to model M_6. Similarly, if M_5 classifies an image as one belonging to any of ISUP grades 1, 2, the decision is deferred to model M_7.

After experimenting with different values, the k value was taken as 5 for the k-means clustering models used during training M_5, M_6, M_7. This value also implicitly reaffirms the different broad clusters (background, benign tissue, Gleason patterns 3, 4, 5) to which a patch of a digitized prostate biopsy can belong. Each cluster represents a dominant feature found in the group of patches.

3. RESULTS

3.1 Dataset

To validate the efficacy of our proposed methods, we use the extensive multi-center dataset from the Kaggle Prostate cANcer graDe Assessment (PANDA) Challenge.[17,24] This largest publicly available dataset of WSIs consists of 10,616 Hematoxylin and Eosin stained digitized biopsies from two different centers.The images are available at three different magnifications: 20X, 5X, and 1.6X. From this dataset, we have considered approximately an equal number of intermediate-resolution images in each of the six ISUP grades, and used 80% images for training and 20% for validation and testing purposes.

Since an entire WSI cannot be fed directly into a CNN due to computational limitations, the image is divided into patches first. Then, the patches with the lowest average pixel intensities (least background white areas) are considered from each image for training. The idea is that the average pixel intensity of the patches containing more tissue will be much less compared to the patches covering the background white areas. The number of patches is selected empirically by observing the patches in the sorted order of pixel intensities. Also, the number of patches is selected such that it is able to maintain a good balance between the number of patches and the considered batch size during training of the models.

3.2 Data augmentation

We use runtime data augmentation at the patch-level to increase the robustness of the models and avoid overfitting. Common affine transforms and some other transformations were randomly applied to the patches. Considering the staining variations caused by different centers, stain augmentation was applied to increase the generalization capability of the models. More precisely, each RGB patch is mapped to the HED (Hematoxylin-Eosin-DAB) color space first. Then, the Hematoxylin and Eosin channels are augmented and the image is again converted back to RGB color space.

3.3 Evaluation metric

For evaluating a traditional classification system, accuracy is the metric commonly used. However, for the current problem which is in the medical imaging domain, we consider Cohen's quadratic weighted kappa as the evaluation measure . In medical diagnosis, it is important that misclassifications are as close as possible to the actual grade. Since the kappa score takes this into account and penalizes errors between non-adjacent classes more than errors between adjacent classes, it is considered an important evaluation metric for medical image grading. The kappa scores of the two methods proposed here are compared to other techniques in Table 1.

Table 1. Comparison of our proposed methods with other methods on the PANDA challenge dataset

Method	**Kappa Score**
Self-learning system (Features + Average + MLP)[14]	0.8245
Self-learning system (Features + Average + kNN)[14]	0.7927
GG% + kNN[14]	0.8152
GG% + MLP[14]	0.8229
Multichannel and Multispatial (MCMS) Attention CNN (best: Layer 3 and 4-AVG)[8]	0.8503
Method using squared error loss of ordinal regression[15]	0.8120
CLAM[16]	0.7660
ISMIL[16]	0.8600
Multi-CNN Ensemble (Proposed Method 1)	**0.8810**
Bag-of-words Model + XGBoost classifier (Proposed Method 2)	**0.8707**
Bag-of-words Model + LGBM classifier (Proposed Method 2)	**0.8705**

4. CONCLUSIONS

In this work, we presented two techniques for ISUP grading of prostate cancer from intermediate-resolution WSIs. From the first technique, we observed that training different CNN models for different subsets of ISUP grades (having greater confusion among themselves) significantly improved the grading performance. The performance of the second proposed weakly-supervised learning-based method shows that if the CNN models can extract meaningful features from the patches of the whole slide images, they can be represented in a more concise manner (using clustering) that can be used with other classifiers to further improve the performance. Considering the bag-based MIL approach used to train our CNN models, we did not focus on the ROI localization task here. As the loss functions are defined at the bag-level, it is difficult for these models to delineate the regions responsible for making the corresponding predictions. Another limitation of the first method is that a fixed number of patches must be considered from each image and there is always a tradeoff between the considered number of patches and the batch size. Though the fixed number chosen empirically has proved to be good enough here, an approach that allows considering a variable number of patches from each image would be more desirable. The second method is free from the limitation of allowing a fixed number of patches for each image as there is no need to consider the images in batch during training.

ACKNOWLEDGMENTS

The work presented in this paper is funded by J. C. Bose National Fellowship grant number JCB/2020/000033.

REFERENCES

[1] Burchardt, M., Engers, R., Müller, M., Burchardt, T., Willers, R., Epstein, J., Ackermann, R., Gabbert, H., De La Taille, A., and Rubin, M., "Interobserver reproducibility of Gleason grading: Evaluation using prostate cancer tissue microarrays," *Journal of Cancer Research and Clinical Oncology* **134**, 1071–1078 (Oct. 2008).

[2] Epstein, J. I., "A new contemporary prostate cancer grading system," *Annales de Pathologie* **35**(6), 474–476 (2015).

[3] Nir, G., Hor, S., Karimi, D., Fazli, L., Skinnider, B. F., Tavassoli, P., Turbin, D., Villamil, C. F., Wang, G., Wilson, R. S., et al., "Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts," *Medical image analysis* **50**, 167–180 (2018).

[4] Nir, G., Karimi, D., Goldenberg, S. L., Fazli, L., Skinnider, B. F., Tavassoli, P., Turbin, D., Villamil, C. F., Wang, G., Thompson, D. J., et al., "Comparison of artificial intelligence techniques to evaluate performance of a classifier for automatic grading of prostate cancer from digitized histopathologic images," *JAMA network open* **2**(3), e190442–e190442 (2019).

[5] Silva-Rodríguez, J., Colomer, A., Sales, M. A., Molina, R., and Naranjo, V., "Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection," *Computer methods and programs in biomedicine* **195**, 105637 (2020).

[6] Campanella, G., Silva, V. W. K., and Fuchs, T. J., "Terabyte-scale deep multiple instance learning for classification and localization in pathology," *arXiv preprint arXiv:1805.06983* (2018).

[7] Nishio, M., Matsuo, H., Kurata, Y., Sugiyama, O., and Fujimoto, K., "Label distribution learning for automatic cancer grading of histopathological images of prostate cancer," *Cancers* **15**(5), 1535 (2023).

[8] Yang, B. and Xiao, Z., "A multi-channel and multi-spatial attention convolutional neural network for prostate cancer isup grading," *Applied Sciences* **11**(10) (2021).

[9] Nagpal, K., Foote, D., Liu, Y., Chen, P.-H. C., Wulczyn, E., Tan, F., Olson, N., Smith, J. L., Mohtashamian, A., Wren, J. H., et al., "Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer," *NPJ digital medicine* **2**(1), 48 (2019).

[10] Li, J., Li, W., Gertych, A., Knudsen, B. S., Speier, W., and Arnold, C. W., "An attention-based multi-resolution model for prostate whole slide imageclassification and localization," *arXiv preprint arXiv:1905.13208* (2019).

[11] Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., and Litjens, G., "Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study," *The Lancet Oncology* **21**(2), 233–241 (2020).

[12] Wang, X., Xiang, J., Zhang, J., Yang, S., Yang, Z., Wang, M.-H., Zhang, J., Yang, W., Huang, J., and Han, X., "Scl-wc: Cross-slide contrastive learning for weakly-supervised whole-slide image classification," *Advances in neural information processing systems* **35**, 18009–18021 (2022).

[13] Toledo-Cortés, S., Useche, D. H., Müller, H., and González, F. A., "Grading diabetic retinopathy and prostate cancer diagnostic images with deep quantum ordinal regression," *Computers in Biology and Medicine* **145**, 105472 (2022).

[14] Silva-Rodríguez, J., Colomer, A., Dolz, J., and Naranjo, V., "Self-learning for weakly supervised gleason grading of local patterns," *IEEE Journal of Biomedical and Health Informatics* **25**(8), 3094–3104 (2021).

[15] Nirthika, R., Manivannan, S., and Ramanan, A., "Loss functions for optimizing kappa as the evaluation measure for classifying diabetic retinopathy and prostate cancer images," in [*2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*], 144–149 (2020).

[16] Yang, Z., Wang, X., Xiang, J., Zhang, J., Yang, S., Wang, X., Yang, W., Li, Z., Han, X., and Liu, Y., "The devil is in the details: a small-lesion sensitive weakly supervised learning framework for prostate cancer detection and grading," *Virchows Archiv* **482**, 1–14 (02 2023).

[17] "Prostate cANcer graDe Assessment (PANDA) Challenge." https://www.kaggle.com/c/prostate-cancer-grade-assessment. Accessed on Tue, January 30, 2024.

[18] Tan, M. and Le, Q., "EfficientNet: Rethinking model scaling for convolutional neural networks," *Proceedings of Machine Learning Research PMLR* **97**, 6105–6114 (2019).

[19] Radenović, F., Tolias, G., and Chum, O., "Fine-tuning CNN image retrieval with no human annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(7), 1655–1668 (2019).

[20] Hu, J., Shen, L., and Sun, G., "Squeeze-and-excitation networks," in [*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 7132–7141 (2018).

[21] Panda, S. and Panda, C., "A review on image classification using bag of features approach," *International Journal of Computer Sciences and Engineering* **7**, 538–542 (06 2019).

[22] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y., "LightGBM: A highly efficient gradient boosting decision tree," in [*Proceedings of the 31st International Conference on Neural Information Processing Systems*], 3149–3157 (2017).

[23] Chen, T. and Guestrin, C., "XGBoost: A scalable tree boosting system," in [*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*], *KDD '16*, 785–794, Association for Computing Machinery (2016).

[24] Bulten, W., Kartasalo, K., Chen, P.-H. C., and et al, "Artificial intelligence for diagnosis and gleason grading of prostate cancer: the PANDA challenge," *Nature Medicine* **28**, 154–163 (01 2022).