Automated Prostate Cancer Diagnosis Based on Gleason Grading Using Convolutional Neural Network

Haotian Xie¹, Yong Zhang², Jun Wang³*, Jingjing Zhang⁴, Yifan Ma⁴, Zhaogang Yang⁵*

¹Department of Mathematics, The Ohio State University, Columbus, OH, 43210, United States
²Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, 43210, United States
³Department of Informatics, King’s College London, London, WC2R 2LS, UK
⁴Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, OH, United States.
⁵Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX, 75390, United States

* Correspondence:
Zhaogang Yang
Zhaogang.Yang@UTSouthwestern.edu

Keywords: Diagnosis, Prostate cancer, Gleason score, CNN, Image classification, Segmentation, Loss function
Abstract

Prostate cancer (PCa) is the most common deadly cancer in the United States. The Gleason grading system using histological images is the most powerful diagnostic and prognostic predictor of PCa and is essential in treatment planning for patients. The current standard inspection is evaluating Gleason H&E-stained histopathology images by pathologists. However, it is complicated, time-consuming, and subject to observers. Hence, an automatic classification system is necessary to reduce variations and improve clinical outcomes accuracy. Deep learning (DL) based-methods that automatically learn image features and achieve higher generalization ability have attracted significant attention. However, challenges remain especially using DL to train the whole slide image (WSI), a predominant clinical source in the current diagnostic setting, containing billions of pixels, morphological heterogeneity, and artifacts. To make DL become a reality in the clinical practice, the above difficulties need to be addressed. Hence, we proposed a convolutional neural network (CNN)-based automatic classification method for accurate grading of PCa using whole slide histopathology images. In this paper, a data augmentation method named Patch-Based Image Reconstruction (PBIR) was proposed to reduce the high resolution and increase the diversity of WSIs. In addition, a distribution correction (DC) module was developed to enhance the adaption of pretrained model to the target dataset by adjusting the data distribution. Besides, a Quadratic Weighted Mean Square Error (QWMSE) function was presented to reduce the misdiagnosis caused by equal Euclidean distances. These strategies enabled our method to take advantage of the tremendous amount of clinical information in large and small patches in the labeled dataset. We studied their effects on the classification performance empirically. Our experiments indicated the combination of PBIR, DC, and QWMSE function was necessary for achieving superior expert-level performance, leading to the best results (0.8885 quadratic-weighted kappa coefficient). We expect this system to help pathologists reduce the probability of misdiagnosis and support prostate cancer treatment via an automated, reproducible, and accurate method with consistent standards.
1 Introduction

Prostate cancer (PCa) is the most prevalent type of death-associated cancer in men\textsuperscript{1,2}. Prostate-Specific Antigen test and Digital Rectal Examination are traditional screening methods but suffer from low accuracy\textsuperscript{3}. Novel diagnostic methods such as liquid biopsy-based on extracellular vesicles are far away from the clinical setting\textsuperscript{4-8}. Until now, medical imaging inspection of biopsies stained with hematoxylin and eosin (H&E) is a powerful diagnostic tool\textsuperscript{9}. PCa is a heterogeneous disease described using diverse histopathologic patterns with the Gleason grade system, which is one of the most reliable methods to assess the stages and aggressiveness of the PCa\textsuperscript{10-12}. The Gleason score has also been used in evaluating clinical outcomes such as surveillance and metastasis\textsuperscript{13,14}. The final Gleason score is the sum of the most two primary pattern scores (ranging from 1 to 5) in the tissue specimen producing a final grade of 2 to 10\textsuperscript{15,16}. In the current clinical setting, the lowest Gleason score assigned is 3 + 3\textsuperscript{17}. However, current histopathology techniques based on the assessment of Gleason score have several limitations. Except for time-consuming and labor-intensive, they rely on evaluating the pathologist and suffer from high variability, subjective interpretation, and very low reproducibility\textsuperscript{18-20}. Therefore, developing a more efficient and automated system to differentiate Gleason scores objectively and increase the accuracy and sensitivity is necessary.

In recent years, deep learning (DL) has been proposed as a potential for generating an objective, accurate, and reproducible Gleason score that facilitated pathologists in their evaluations\textsuperscript{21-24}. Fundamentally, DL performs feature learning and classification in a single framework without handmade features widely used in traditional ML\textsuperscript{25-27}. Besides, increased computational power offers DL an opportunity to train more advanced algorithms, which improves their performance on more complex tasks further\textsuperscript{28}. Convolutional neural networks (CNNs) are a kind of DL neural networks that allow learning hierarchical representation of the data, which learn features in at least one of their layers and extract information with inherent translation invariance\textsuperscript{29-31}. Comparing with other methods, CNNs often obtain better results since they can represent the subtle patterns embedded in images precisely\textsuperscript{32,33}. Ronneberger et al. presented a U-shaped neural network named U-Net, which enabled to capture context using a contracting path and consisted of a symmetric expanding path for precise localization\textsuperscript{34}. More recently, Li et al. have proposed a semi-supervised learning method using the expectation maximization in a deep learning framework for prostate cancer grading\textsuperscript{35}. Overall, these studies have shown that CNNs could achieve good classification accuracies of image analysis.

However, challenges remain when utilizing CNNs for biomedical image classification due to the complexity of the clinical image source, which is limiting the applicability of CNNs. Taking the whole slide image (WSI) as an example, it is an essential source of information containing region and cellular-level features, which has significant implications for diagnosis\textsuperscript{36-38}. The large size WSI (100 k × 100 k) bring challenges when compared with other processed imaging modalities, because the super-high-resolution cannot be directly processed by the CNN model, even though pyramid representation decreases the size of WSI, the size of one single downsampled image is still too large for the acceptable computation\textsuperscript{39,40}. Also, the computational resources required for convolution when scanning granular images have increased dramatically\textsuperscript{41,42}. In addition, the heterogeneity of color information (H&E staining) presented in WSI is problematic. WSI contains massive blank pixels with minimum clinical information but brings the biased distribution of features during the classification, causing under-fitting problems when applying the migration learning to improve model performances in the CNN tasks\textsuperscript{43}. Besides, color aberration or location error of tissues or cells that affect
the contrast and resolution of imaging may result in misdiagnosis\textsuperscript{44,45}. Another challenge in biomedical image classification is to ensure high sensitivity and specificity of calculated clinical outcomes. For Gleason grading, scanned tissue within the same class has large variations, making the classification tasks hard due to the intraclass variances are larger than interclass variances. Moreover, datasets used for training depend on the evaluation of pathologists, which may suffer from subjective interpretations and cause low reproducibility or misdiagnosis\textsuperscript{46-48}.

In this work, we proposed an accurate and automated framework for Gleason grading. EfficientNet was chosen as the DL algorithm backbone for training\textsuperscript{49}. The main contributions can be summarized as follows:

- A patch-level regroup method named Patch-Based Image Reconstruction (PBIR) was proposed to segment original super-high-resolution images into multiple small patches, which were grouped as new images with different aspect ratios. PBIR avoided blank pixels but focused on H&E-stained pixels and coped with difficulty in training a CNN model in super-high-resolution datasets.
- A Distribution Correction (DC) Module was applied to the trained model by building on an extra layer of the CNN, which corrected the biased distribution presented on WSIs, allowing the enhancement of the data adaptation ability of the trained model without increasing the number of parameters and calculations.
- Fine-tuned the model based on a self-proposed loss function named Quadratic Weighted Mean Square Error (QWMSE), which improved the classification accuracy.

2 Materials and Methods

2.1 PANDA Dataset

The datasets sponsored by Radboud University Medical Center and Karolinska Institute contain around 11,000 digitized H&E-stained biopsies WSIs, hold by the Kaggle competition named Prostate cANcer graDe Assessment (PANDA) Challenge-Prostate cancer diagnosis using the Gleason grading system. Hematoxylin is a basic dye that stains the nuclei with purple color, while other parts are stained into different shades of pink. Representative WSIs with varying sizes in the PANDA datasets were shown in Figure 1. WSIs in the PANDA datasets only contain a small portion of tumor-associated H&E-stained patterns, surrounded by massive blank pixels that are irrelative to classification tasks. Based on the morphologic pattern of tumours, five patterns are assigned with numbers from 1 to 5 with declining differentiated. Pattern 1 corresponds to well differentiate tissue with small and uniform glands presenting the highest similarity to benign tissues, followed Pattern 2 and 3 correspond to moderately differentiated. Pattern 4 with irregular masses of neoplastic glands and Pattern 5 with rare gland formation correspond to poorly differentiated. Patterns 1 and 2 are excluded when evaluating the Gleason score since they are rare and do not present malignant features. The final Gleason score is reported as the sum of the two most apparent histological specimen patterns. In current clinical practice, Gleason score ranges from 6 (3+3) to 10 (5+5). In this paper, once the biopsy was assigned with Gleason scores, an ISUP grade on a 1-5 scale was transformed based on Gleason scores. Biopsies not containing cancer were represented by 0 ISUP grade, which is uncommon in other studies. We have to include this parameter because of the evaluation system offered by Kaggle.
Figure 1. Whole-slide images with different sizes in the PANDA dataset.

2.2 Network

This paper developed an automated grading method of PCa based on EfficientNet. Figure 2. showed the schematic of our method. Briefly, this classification method could be divided into two parts. In the first part, an image preprocessing step called PBIR was conducted. The targeted WSI was segmented into multiple 224×224 patches that were sorted, selected, and regrouped into new images for processing later. Subsequently, acquired new images were imported into EfficientNet-B2 model to extract image features and then perform classification. This paper utilized the traditional Cross-Entropy (CE) loss function for training EfficientNet-B2 model. Besides, a distribution correction module was applied by adding an additional convolutional layer on EfficientNet-B2, in which the distribution of images was adjusted, thus improving the adaptation ability of our model towards handling the PANDA dataset. Next, we fine-tuned the trained model by the QWMSE to further enhance model performance on Gleason score classification, which classified new images into the ISUP score of 0, 1, 2, 3, 4, 5 based on Gleason score (i.e., background (no tissue), stroma, healthy or benign epithelium, Gleason 3, Gleason 4, and Gleason 5).
Figure 2. The schematic of the convolutional network used for Gleason grading of PCa. (a). An original WSI image in the training dataset. (b). Image preprocessing by the following mentioned method named PBIR. The WSI was firstly segmented into 224 * 224 patches. Patches were selected (n = 36) and formed a new image with different aspect ratios after sortation. Other data augmentation methods were also applied during the procedure. (c). The output of step (b) was then used as the input to train EfficientNet-B2 model. And a data correction module was introduced as an extra convolutional layer to the trained model. (d). The output of the pre-trained model was employed as an input of the same model with the QWMSE loss function to fine-tune the model. (e). An original WSI image in the test dataset. (f). Images were classified into ISUP score 0, 1, 2, 3, 4, 5 after training and fine-tuning with QWMSE based on EfficientNet-B2 model.

2.3 Patch-based Image Reconstruction (PBIR)

The schematic of the PBIR is shown in Figure 3. First, the original WSIs were segmented into multiple non-overlapping 224 * 224 patches. To guarantee that the number of possible segmented patched were integer, images were padded with blank pixels if the height or width of images were not the multiples of 224. Then, patches were sorted according to the proportion of stained tissues and cells, and those with larger proportions were selected and reorganized with different aspect ratios (i.e., 1×36, 3×12, 4×9, 6×6) to form a new training image. Other data augmentation methods were also applied in the image preprocessing, such as random cropping, random rotations, flipping, and color jittering. PBIR enabled the CNN model to ignore the surrounding blank patterns but focused on the stained pathological patterns associated with the severity-level classification. Additionally, through segmenting and regenerating the training image, image resolution was primarily reduced (e.g., from 100k
* 100k to 224k * 8k) while retaining the most informative patches in images, enabling the training of CNN models in the super-high-resolution datasets.

**Figure 3.** Schematic of the PBIR. WSIs were segmented into multiple 224 * 224 patches, which were sorted and then formed new images with different aspect ratios (i.e., 1×36, 3×12, 4×9, 6×6).

### 2.4 Distribution Correction Module

In this paper, we adopted EfficientNet pretrained in ImageNet datasets as our baseline network. However, data distribution in ImageNet and the PANDA dataset is different ([Figure 4](#)). Images in the ImageNet have obvious objectives while images in the PANDA dataset contain more blank pixels instead of valuable information. In order to mitigate the problem of data distribution difference during the transfer learning, an additional convolutional layer with the kernel size of 1 × 1 was added before the first convolutional layer started training the model to modify some of the image distribution in the PANDA datasets. The newly added convolutional layer was adaptively adjusted accordingly because parameters of the convolutional layer in EfficientNet were updated in the process of backpropagation, implicitly narrowing the difference between our dataset and the ImageNet. The output of this layer was calculated as follows:

\[
H(i,j) = \sum_{\tau=1}^{S} f_\tau(i,j) \quad \text{Equation 1}
\]

Where H is an output feature map, and i and j are the coordinates of a pixel in both the input feature maps and output feature map. S denotes the number of channels in the image. \( f_\tau \) is the convolution operation in channel \( \tau \).
2.5 Quadratic Weighted Mean Square Error (QWMSE)

A self-defined loss function named QWMSE was proposed to fine-tune the model since class label distance was meaningful and different in this task, which was not conceded in traditional CE loss function. QWMSE was proposed based on the QWK, an evaluation metric in this paper described below:

\[
L = \frac{1}{n} \sum (y_m - \hat{y})^2
\]

Equation 2

Similarly, \( y_m \) is the ground truth category, \( \hat{y} \) is the predicted category, which is a scaler transformed from a one-hot code as follows:

\[
\hat{y} = a \ast \hat{y}_{net}
\]

Equation 5
Where \( \hat{y}_{\text{net}} \) is the predicted vector of the network output, and a vector \{0, 1, 2, 3, 4, 5\} used to weight the network output \( \hat{y}_{\text{net}} \). The highest probability of matched vector factor should determine the ISUP value (b).

\[
a \ast \hat{y}_{\text{net}} = b \tag{Equation 6}
\]

For example, if \( b = 2 \), the most accuracy result should be \( 0*0 + 1*0 + 2*1 + 3*0 + 4*0 + 5*0 = 2 \). However, this equation has multiple solutions, and the solution may fall into the local optimum. For example, \( 0*0 + 1*0.5 + 2*0 + 3*0.5 + 4*0 + 5*0 = 2 \).

To alleviate this problem, we decided to use \( a^2 \) when calculating the ISUP value (b):

\[
\sqrt{a^2 \ast \hat{y}_{\text{net}}} = b \tag{Equation 7}
\]

Accordingly, our QWMSE was proposed as follows:

\[
\text{Loss} = \text{MSE}(y_m - \sqrt{\hat{y}}) \tag{Equation 8}
\]

Where the MSE is the Euclidean distance.

However, due to the presence of vector factor 0, we assigned vector factor 0 a small relaxation factor (\( \varepsilon \)), which offered 0 a small value, letting vector factor 0 have a value of \( \varepsilon \) to participate in the calculation instead of always being a 0. This modification could avoid the results coming into a catastrophic solution. In this study, we had to include ISUP 0 because of the evaluation system provided by Kaggle, so a small relaxation factor (\( \varepsilon \)) was necessary. In other Gleason classification tasks or other evaluation system, which only consider ISUP from 1 to 5, this relaxation factor can be eliminated.

### 2.6 Evaluation Metrics

Our method was evaluated in the PANDA challenge provided by Kaggle. Since the testing set was not provided, we split the original training set into a training set (2/3) and a validation set (1/3) and utilized the validation set to verify the performance of models. We employed two indicators, namely CE and QWK to assess model performance. The specific calculation processes were shown in Equation (2) to (3). Later we used the above-mentioned QWMSE to fine-tune the trained modal.

CE is the most common and widely used loss function for multi-classes classification tasks, where the predicted input is a probability value between 0 and 1 (Equation 9).

\[
CE = -\frac{1}{s_n} \sum_{m=1}^{s_n} \sum_{n=1}^{N} y_{mn} \log (p(y_{mn})) \tag{Equation 9}
\]

Where \( m \) is the number of the samples, and \( N \) is the number of the classifications. \( y_{mn} \) is the predicted result of classification, such as 0, and 1. \( p(y_{mn}) \) is the predicted probability of \( y_{mn} \).

In classification problems in military or medicine fields, classification errors can lead to very serious consequences\(^5\). In such error-sensitive scenarios, the use of QWK as an evaluation metric can compensate for chance-errors in the classification. QWK measures the agreement between predict results with truth results used in multi-valued classification questions\(^5\).
In Equation 3, for example, \((i - j)^2 = 0\) indicates that there is no difference between the real and predicted values, whereas \((i - j)^2 = 25\) indicates that the difference between the actual and predicted values is at its maximum. From \(W_{ij}\) we observed that QWK gave a higher weight when the difference between the predicted value and the real value was considerable. For instance, the weight \(W_{ij}\) was 0.0625 when the Gleason score 2 was predicted to the Gleason score 3, while the weight \(W_{ij}\) is 0.25 when the Gleason score 2 was predicted to be Gleason score 4. Hence, the lower weight, the higher QWK.

### 2.7 Implementation Details

All experiments in this paper were based on Python and Pytorch deep learning framework. The device used to accelerate the training of the CNN model was a single V100 GPU. The Adam optimization algorithm optimized the model with an initial learning rate of 3E-4 and a cosine annealing strategy for learning rate recession. To mitigate the early overfitting of the model to the mini-batch in the initial phase and to maintain the smoothness of the distribution, this experiment adopted the warmup strategy in epoch 1, 2, 5 and 6, thus preserving the stability in the deeper layers of the model. The number of selected segmented patches to reform the image was set to 36. The experiment was divided into two parts. The first part was the comparison between the results of different grouping methods when segmenting WSIs using PBIR. The second part compared the model using only the data distribution correction module, and the model using the data distribution correction module after the optimal solution in the experiment of the first part was selected. Besides, the results after QWMSE fine-tuning were also compared.

### 3 Results

#### 3.1 The performance of the PBIR

Experiments were conducted on different patches with specified aspect ratios (i.g., 1x36, 3x12, 4x9, 6x6) to evaluate their effectiveness. Here we compared the influences of different aspect ratios. As Figure 5. demonstrated, patches with a 6 * 6 aspect ratio achieved the lowest training CE loss with a value of 0.4518 and 0.7128 in testing CE loss, a concordance to other dimensions. 3 * 12 ratio was the second lowest with 0.4547 train CE loss and 0.7177 test CE loss. However, the 3 * 12 ratio demonstrated the highest QWK score (0.8801) compared to other settings. The detailed results were listed in Table 1. Hence, 3 * 12 was selected as input data in the following experiment after considering its superior performance. In this study, the PBIR method sliced the complete WSI into patches suitable for training and completed data augmentation by random reconstruction and grouping, thus alleviating the overfitting problem. Through this method, training the CNN based approach in super-high-resolution datasets became feasible since the size was mostly reduced.
Figure 5. CE loss and QWK comparisons among reconstructed images using patches with different lengths and widths.

Table 1. Comparison of different aspect ratios

| Aspect Ratio | Train CE Loss | Test CE Loss | QWK     |
|--------------|---------------|--------------|---------|
| 1x36         | 0.4578        | 0.7202       | 0.8764  |
| 3x12         | 0.4547        | 0.7177       | 0.8801  |
| 4x9          | 0.4586        | 0.7013       | 0.8683  |
| 6x6          | 0.4518        | 0.7128       | 0.8664  |

3.2 The Performance of Distribution Correction (DC)

Subsequently, we examined the performance of models after involving DC. Figure 6A. showed the original CE loss of training and testing after applying PBIR with a 3 * 12 aspect ratio. After introducing the DC, a lower CE loss, from 0.4692 to 0.4547 in the training set, and from 0.7177 to 0.6996 in the test set, were achieved (Figure 6B).

Our proposed loss function also validated the effectiveness of DC. We first calculated the CE loss of the model after conducting PBIR with the 3 * 12 aspect ratio and QWMSE (Figure 6C). A CE loss was then tested after involving the DC adjustment, and the trend of training loss and validation loss was improved by reducing the value from 0.3282 to 0.2954 and 0.5563 to 0.5410, respectively (Figure 6D).

QWK was also calculated to prove the efficacy of the DC module (Figure 7). It can also be seen the QWK showed a slight increase by 0.25% (from 0.8801 to 0.8823) after DC adjustment at the training stage. Besides, the QWK showed a small increase of 0.95% (from
0.8857 to 0.8885) compared with the model without DC adjustment during the second training stage using QWMSE.

These comparisons demonstrated the efficacy of the DC module. When the DC module was combined with PBIR, DC scaled the distribution of the dataset and reduced the distribution difference between the experimental dataset employed in this paper, therefore further reduced the impact of blank pixels in the background on EfficientNet model and improved the performance of our model.

Figure 6. A and B were evaluated using CE loss in the training stage. A was the CE loss of the model only applied the PBIR. B was the CE loss of the model after combining the PBIR and DC. C and D were evaluated using the QWMSE loss function in the second-training stage. C was the loss of the model when combined PBIR and QWMSE modules. D was the loss of the model when combined PBIR, DC, and QWMSE modules. Train loss was in blue and validation loss was in yellow.

3.3 The Performance of QWMSE

A suitable loss function in an experiment can orient the results of the optimization target. We designed a QWMSE loss function to fine-tune the model, allowing the model to avoid misdiagnosis. After train loss and value loss flattened out in the training stage, a lower loss was observed after fine-tuning the model using the QWMSE loss function (Figure 6A and 6C). And a similar trend could be observed in Figure 6B and 6D. One possible reason was that the curve retrained using the QWMSE function inherited the CE loss function results from the PBIR method.

A clear improvement of the QWK, ranging from 0.8764 to 0.8857, was showed when compared with only utilizing the PBIR after boosting our model with QWMSE (Figure 7). Similarly, an upgrade could be found when compared methods that applied PBIR + DC and PBIR + DC + QWMSE, ranging from 0.8823 to 0.8885. Detailed information was exhibited in Table 2. This was probably because the second weighting of QWMSE focused on reducing the occurrence of more serious categorization errors (e.g., misclassifying category 0 into category 5), allowing better results on the error-sensitive QWK metric.
Importantly, we have demonstrated the effectiveness of using PBIR, DC, and QWMSE separately or some of their combination. The best-performance method was combined with these three strategies, leading to the highest QWK.

![Figure 7. QWK scores of the model when different strategies were applied.](image)

## 4 Discussion

In this work, we trained CNNs to classify Gleason scores in biopsy WSIs. EfficientNet-B2 was selected as the baseline model, which scaled each network dimensions uniformly by offering a fixed set of scaling coefficients, allowing super pass efficiency. Three strategies, including the PBIR, DC, and QWMSE, were proposed to improve the classification performance of EfficientNet-B2 on the PANDA dataset. Our model was trained and evaluated on WSIs obtained from combined datasets provided by the Radboud University Medical Center and Karolinska Institute. The outputs of the models were evaluated by a set of QWK values. We believe the results of this study are important, since to the best of our knowledge, there have been no previously reported studies that were validated on large datasets that consist of 11,000 WSIs, especially for Gleason grading of PCa.

Automated grading of Gleason score from WSIs is promising for the diagnosis and prognosis of PCa. However, one of the main challenges is the extreme size of the WSI. The input resolution of CNN-based models usually is less than 1000 (i.e., 224 * 224), but the image resolution of PANDA dataset is much larger (100k * 100k). Training on such a super-high-resolution image usually requires more in-depth network architecture and huge computation resources, making the real-world application unrealistic. Besides, it is time-consuming to train WSIs with heterogeneous information. WSIs normally contain massive blank pixels that are irrelevant to classification. Hence, in this study, we proposed a PBIR method to segment the WSI into patches, sort the patches, and select sorted patches to reorganize into a suitable structure.

Moreover, transfer learning is generally adopted for training CNN models due to its prominent role in improving model performance and overcoming limited data size. When
conducting transfer learning, parameters trained in the original dataset may not fit well into
the new dataset due to differences of dataset distribution. Most of the targets that need to be
identified or classified in a task usually occupy a high proportion of pixels in the image (e.g.,
ImageNet), which allows the CNN model to obtain better results than a dataset with targets
that occupy a deficient proportion of pixels, like in our PANDA dataset, which most of pixels
in the WSIs of cell and tissue samples used were blank pixels or a single-color stain. Obviously, these blank pixels were unhelpful in classifying the severity of prostate cancer. To
mitigate this problem, we proposed a DC module to reduce the influence of background
factors on the model. The method can be adapted to all trained models by adjusting the
coefficients in the convolutional layer of trainable CNNs with virtually no increase in the
parameters. Its setting is similar to other known normalization methods, where input data was
scaled appropriately to achieve better results than unscaled. Simultaneously, the joint use of
the convolutional layers and the pooling layers reduced the CNNs parameters and
computation cost. Finally, the features extracted by the convolutional layers and the pooling
layers were passed to the fully connected layer for the final classification.

Furthermore, the classification in this task was quite different from traditional
classification tasks because it involved additional regression indexes due to the ambiguous
differences among ISUP scores. In traditional classification, categories are usually encoded
using one-hot coding, resulting in equal Euclidean distances between categories. In this
paper, there was a clear unequal Euclidean distance between categories, thus instead of
using CE loss function that was often used for solving classification tasks that have explicit
differences among categories, finding an alternative loss function was essential for Gleason
grading. Hence, we defined a novel loss function named QWMSE based on the quadratic
weighted kappa (QWK) capable of evaluating the loss to further fine-tuning the model,
obtaining higher sensitivity. In fact, defining a unique solution when classifying subjects
present in more than two categories is difficult. Therefore, to increase the accuracy, we first
trained the model using the cross-entropy loss function, obtaining an ideal initial value before
the fine-tuning. Later, we increased the interval of weights during fine-tuning with a low
learning rate, so the amount of error increase was more considerable when the value was
shifted towards the error category.

In the future, we intend to utilize our boosted model to classify more datasets, which is
essential in developing practical assistance for surgical pathologists and predicting patient
outcomes. We would like to expect a bright future where artificial intelligence and clinicians
cooperate and facilitate medical care.

5 Conclusion

Due to the high death rate and occurrence of PCa and the limited reproducibility of
Gleason grading among manual examination by pathologists, there is a need to develop an
automated classification algorithm for Gleason grading. This paper presented an automated
Gleason scoring system based on convolutional neural networks that achieved start-of-the-art
performance in PCa diagnosis based on histological WSIs. We adopted EfficientNet as
the network backbone to help the network focus on regions with high interests. In addition, a
combination of data segmentation, augmentation, and correction of data distribution boosted
the performance of our method. We also employed QWMSE as a loss function to compensate
artifacts and improved the capability of our system. Extensive experiments were conducted
to validate the efficacy of each module in our model. We eventually proved our method would
help the pathologist to make a consistent and reproducible diagnosis in the near future by
showing a QWK of 0.8885. We also expect this model will support therapeutic and prognostic studies on PCa. This model can also be useful in other biomedical image classification tasks.

Table 2. Comparison of classification performance between different strategies apply to EfficientNet

| EfficientNet | PBIR | DC | QWMSE | QWK |
|--------------|------|----|-------|-----|
|              | √    |    | √     | 0.8801 |
|              | √    | √  |       | 0.8823 |
|              | √    |    | √     | 0.8857 |
|              | √    | √  | √     | 0.8885 |

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Conflicts of Interest

The authors declare that they have no competing interests.

Authors’ Contributions

HT. Xie and ZG. Yang conceived and designed the experiments. HT Xie, Y. Zhang, and J Wang analyzed and extracted data. J. Zhang and Y. Ma constructed figures. Y. Zhang participated in table construction. All authors participated in the writing, reading, and revising of the manuscript and approved the final version of the manuscript.

Acknowledgement

N/A

Reference

1. Lee, F. et al. Prostate cancer: comparison of transrectal US and digital rectal examination for screening. 168, 389-394 (1988).
2. Grönberg, H. J. T. L. Prostate cancer epidemiology. 361, 859-864 (2003).
3. Catalona, W. J. et al. Comparison of digital rectal examination and serum prostate specific antigen in the early detection of prostate cancer: results of a multicenter clinical trial of 6,630 men. 151, 1283-1290 (1994).
4. Ma, C. et al. Isolation and Detection Technologies of Extracellular Vesicles and Application on Cancer Diagnostic. Dose-response: a publication of International Hormesis Society 17, 1559325819891004-1559325819891004, doi:10.1177/1559325819891004 (2019).
Zhang, J. et al. Immunomagnetic Sequential Ultrafiltration (iSUF) Platform for Enrichment and Purification of Extracellular Vesicles from Biofluids. bioRxiv, 2020.05.20.2005.2013.089573, doi:10.1101/2020.05.13.089573 (2020).

Walters, N., Nguyen, L. T. H., Zhang, J., Shankaran, A. & Reátegui, E. Extracellular vesicles as mediators of in vitro neutrophil swarming on a large-scale microparticle array. Lab on a Chip 19, 2874-2884, doi:10.1039/C9LC00483A (2019).

Liu, Y., Ma, Y., Zhang, J., Yuan, Y. & Wang, J. Exosomes: A Novel Therapeutic Agent for Cartilage and Bone Tissue Regeneration. Dose-Response 17, 1559325819892702, doi:10.1177/1559325819892702 (2019).

Shi, C., Xie, H., Ma, Y., Yang, Z. & Zhang, J. Nanoscale Technologies in Highly Sensitive Diagnosis of Cardiovascular Diseases. 8, doi:10.3389/fbioe.2020.00531 (2020).

Pepe, P. & Aragona, F. J. U. Saturation prostate needle biopsy and prostate cancer detection at initial and repeat evaluation. 70, 1131-1135 (2007).

Humphrey, P. A. J. M. p. Gleason grading and prognostic factors in carcinoma of the prostate. 17, 292-306 (2004).

Epstein, J. I. et al. A contemporary prostate cancer grading system: a validated alternative to the Gleason score. 69, 428-435 (2016).

Arora, R. et al. Heterogeneity of Gleason grade in multifocal adenocarcinoma of the prostate. 100, 2362-2366 (2004).

Masic, S. et al. Effects of initial Gleason grade on outcomes during active surveillance for prostate cancer. 1, 386-394 (2018).

Liu, J.-J. et al. Nationwide prevalence of lymph node metastases in Gleason score 3+ 3= 6 prostate cancer. 46, 306-310 (2014).

Stark, J. R. et al. Gleason score and lethal prostate cancer: does 3+ 4= 4+ 3? 27, 3459 (2009).

Egevad, L., Granfors, T., Karlberg, L., Bergh, A. & Stattin, P. J. B. i. Prognostic value of the Gleason score in prostate cancer. 89, 538-542 (2002).

Epstein, J. I. (LWW, 2000).

Nagpal, K. et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. 2, 1-10 (2019).

Bulten, W. et al. Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard. 9, 1-10 (2019).

Madabhushi, A. & Lee, G. (2016).

LeCun, Y., Bengio, Y. & Hinton, G. J. n. Deep learning. 521, 436-444 (2015).

Ngiam, J. et al. in ICML.

Schmidhuber, J. Deep learning in neural networks: An overview. Neural Networks 61, 85-117, doi:https://doi.org/10.1016/j.neunet.2014.09.003 (2015).

Deng, L., Yu, D. J. F. & processing, t. i. s. Deep learning: methods and applications. 7, 197-387 (2014).
25 Xin, Y. et al. Machine learning and deep learning methods for cybersecurity. 6, 35365-35381 (2018).
26 Zhang, L., Tan, J., Han, D. & Zhu, H. J. D. d. t. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. 22, 1680-1685 (2017).
27 Nguyen, G. et al. Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. 52, 77-124 (2019).
28 Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. J. N. R. G. Deep learning: new computational modelling techniques for genomics. 20, 389-403 (2019).
29 Salamon, J. & Bello, J. P. J. I. S. P. L. Deep convolutional neural networks and data augmentation for environmental sound classification. 24, 279-283 (2017).
30 Defferrard, M., Bresson, X. & Vandergheynst, P. in Advances in neural information processing systems. 3844-3852.
31 He, K. & Sun, J. in Proceedings of the IEEE conference on computer vision and pattern recognition. 5353-5360.
32 Oquab, M., Bottou, L., Laptev, I. & Sivic, J. in Proceedings of the IEEE conference on computer vision and pattern recognition. 1717-1724.
33 Lim, L. A. & Keles, H. Y. J. P. R. L. Foreground segmentation using convolutional neural networks for multiscale feature encoding. 112, 256-262 (2018).
34 Ronneberger, O., Fischer, P. & Brox, T. in International Conference on Medical image computing and computer-assisted intervention. 234-241 (Springer).
35 Li, J. et al. An EM-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies. 69, 125-133 (2018).
36 Gilbertson, J. R. et al. Primary histologic diagnosis using automated whole slide imaging: a validation study. 6, 4 (2006).
37 Fine, J. L. et al. Evaluation of whole slide image immunohistochemistry interpretation in challenging prostate needle biopsies. 39, 564-572 (2008).
38 Bauer, T. W. et al. Validation of whole slide imaging for primary diagnosis in surgical pathology. 137, 518-524 (2013).
39 Zhang, Y. et al. in International Conference on Medical Image Computing and Computer-Assisted Intervention. 360-368 (Springer).
40 Peng, C. et al. in 2019 Digital Image Computing: Techniques and Applications (DICTA). 1-8 (IEEE).
41 Sharma, A., Bautista, P. & Yagi, Y. J. A. C. P. Balancing image quality and compression factor for special stains whole slide images. 35, 101-106 (2012).
42 Zarella, M. D. et al. A practical guide to whole slide imaging: a white paper from the digital pathology association. 143, 222-234 (2019).
43 Guo, Q. et al. in 2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE). 810-822 (IEEE).
44 Cai, J., Yang, Q., Xu, Z., Chen, H. & Zhu, X. in *Multiphoton Microscopy in the Biomedical Sciences XX*. 112442A (International Society for Optics and Photonics).

45 Hong, Z. *et al.* Optical diagnosis of gallbladder cancers via two-photon excited fluorescence imaging of unstained histological sections. **30**, 225-233 (2015).

46 Arvaniti, E. *et al.* Automated Gleason grading of prostate cancer tissue microarrays via deep learning. **8**, 1-11 (2018).

47 Lee, M., Ban, J.-J., Im, W. & Kim, M. Influence of storage condition on exosome recovery. *Biotechnology and Bioprocess Engineering* **21**, 299-304, doi:10.1007/s12257-015-0781-x (2016).

48 Egevad, L. *et al.* Identification of areas of grading difficulties in prostate cancer and comparison with artificial intelligence assisted grading. 1-10 (2020).

49 Tan, M. & Le, Q. V. J. a. p. a. Efficientnet: Rethinking model scaling for convolutional neural networks. (2019).

50 Ben-David, A. J. E. S. w. A. Comparison of classification accuracy using Cohen’s Weighted Kappa. **34**, 825-832 (2008).

51 Cohen, J. J. E. & measurement, p. A coefficient of agreement for nominal scales. **20**, 37-46 (1960).