Original Article

Transfer Learning with Pretrained Convolutional Neural Network for Automated Gleason Grading of Prostate Cancer Tissue Microarrays

Abstract

Background: The Gleason grading system has been the most effective prediction for prostate cancer patients. This grading system provides this possibility to assess prostate cancer's aggressiveness and then constitutes an important factor for stratification and therapeutic decisions. However, determining Gleason grade requires highly-trained pathologists and is time-consuming and tedious, and suffers from inter-pathologist variability. To remedy these limitations, this paper introduces an automatic methodology based on transfer learning with pretrained convolutional neural networks (CNNs) for automatic Gleason grading of prostate cancer tissue microarray (TMA). **Methods:** Fifteen pretrained (CNNs): Efficient Nets (B0-B5), NasNetLarge, NasNetMobile, InceptionV3, ResNet-50, SeResnet 50, Xception, DenseNet121, ResNext50, and inception_resnet_v2 were fine-tuned on a dataset of prostate carcinoma TMA images. Six pathologists separately identified benign and cancerous areas for each prostate TMA image by allocating benign, 3, 4, or 5 Gleason grade for 244 patients. The dataset was labeled by these pathologists and majority vote was applied on pixel-wise annotations to obtain a unified label. **Results:** Results showed the NasnetLarge architecture is the best model among them in the classification of prostate TMA images of 244 patients with accuracy of 0.93 and area under the curve of 0.98. **Conclusion:** Our study can act as a highly trained pathologist to categorize the prostate cancer stages with more objective and reproducible results.

Keywords: *Convolutional neural network, Gleason grading, prostate cancer, transfer learning*

Submitted: 02-Jul-2022 **Revised:** 20-Dec-2022 **Accepted:** 22-Mar-2023 **Published:** 14-Feb-2024

Parisa Gifani[1], Ahmad Shalbaf[2,3]

[1]Department of Biomedical Engineering, Science and Research Branch, Islamic Azad University, [2]Cancer Research Center, Shahid Beheshti University of Medical Sciences, [3]Department of Biomedical Engineering and Medical Physics, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Address for correspondence:
Dr. Ahmad Shalbaf, Cancer Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Department of Biomedical Engineering and Medical Physics, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran. E-mail: shalbaf@sbmu.ac.ir

Access this article online

Website: www.jmssjournal.net

DOI: 10.4103/jmss.jmss_42_22

Quick Response Code:

Introduction

Prostate cancer is the second leading cause of cancer death in men.[1] While some forms of prostate cancer are slow to grow and may need limited or even no care, other types are aggressive and can spread rapidly. Early diagnosis of prostate cancer has a better chance of being treated successfully. A commonly used method of diagnosis is based on histopathological data obtained from the tumor found in the biopsy of the prostate. Prostatic carcinomas are graded in accordance with the Gleason scoring system which Gleason and Mellinger first developed in 1966 known as the Gleason score.[2] The Gleason scoring system is acknowledged by the World Health Organization (WHO) and the International Society of Urological Pathology (ISUP).[3] The histological Gleason scoring system-based assessment on the architectural pattern of the tumor is the most powerful prognostic tool in the clinical diagnosis of prostate cancer.[4] The Gleason method segregates prostate carcinoma schemas into five groups based on different histological patterns, ranging from 1 (low risk) to 5 (high risk). The final Gleason score is reported as the sum of the most prominent and second most prominent patterns. As an example 4 + 3 means the two most predominant Grade is 4 and 3, and the Gleason score is 7.[5] Pathologists use several examination methodologies based on architectural patterns to identify the complex histology of the prostate tumor in a qualitative manner.

The final Gleason score determination is based on microscopy-based evaluation of nontrivial cellular and morphological patterns and is dependent on the histological assessment of the respective pathologist. However, this work is time-consuming and often has a high degree of inter-observer variability that

How to cite this article: Gifani P, Shalbaf A. Transfer learning with pretrained convolutional neural network for automated gleason grading of prostate cancer tissue microarrays. J Med Sign Sens 2024;14:4.

results in grading uncertainty.[6-9] Hence, prognosis and therapeutic options based on the Gleason score recorded may be disagreeable.[10] For prostate cancer in particular, intermediate-risk Gleason patterns 3 and 4 can be very difficult to assign unambiguously. Specialized pathologists have been shown to have higher rates of conformity,[11] but such highly-trained specialists are not commonly accessible. Thus to overcome the above-mentioned limitations and to prevent unsuitable therapies,[12,13] an automatic, reliable and reproducible approach using advanced machine learning method like this paper for Gleason score determination on digital pathology images is required. This system can overcome these limitations and can be utilized everywhere with no need to highly-trained pathologists.

Earlier machine learning methods developed for prostate carcinomas are based on hand-made feature extraction, perform feature selection, and finally employ conventional classification methods.[14-19] In recent years, deep learning systems relying on multi-layered neural networks have emerged as a disruptive alternative to the aforementioned feature-based methods. Deep learning methods, instead of using handcrafted features are able to extract and learn increasingly complex, task-related features directly from the data. These methods imitate the workings of the human brain in data processing and generating patterns of decision-making usage. Recent developments in neural network architecture design and training have enabled researchers to solve previously intractable learning tasks of deep learning methods. As a result, several researches in recent years have focused on the application of deep learning as the state-of-the-art in machine learning especially convolutional neural networks (CNN) in a wide range of biomedical image analysis tasks with very success,[20-23] especially in skin cancer,[24] lung cancer,[25] cardiovascular,[26] ophthalmology,[27] and musculoskeletal.[28] Some studies have been recently accomplished in histopathological images using deep learning to detect malignancies.[29-31]

One of the first studies on automatic Gleason grading assessment based on deep learning is the work of Källén *et al*.[32] Tissue slides with homogeneous Gleason grading ignoring heterogeneous Gleason pattern regions are the main restriction of this work. In another study, Zhou *et al*. worked on an intermediate Gleason score of 7.[33] They tried to distinguish Gleason 3 + 4 from Gleason 4 + 3 on whole slide images. del Toro *et al*. developed a binary classifier to differentiate low (7 or lower) versus high (8 or higher) Gleason score from whole slide images.[34] Other recent works on Gleason grading are to train a patch-based classifier and differentiating patches into benign and Gleason Grades 3, 4, 5.[35-37] In Leyh-Bannurah *et al*.,[38] a custom limited pretrained CNN model was used to determine only Benign vs. prostate cancer tissue in a tissue microarray (TMA) images and in Abbasi *et al*.,[39] a deep learning CNN is employed to determine cancer or noncancer in MRI database. Some authors have used the different versions of UNet for the segmentation of nuclei in histopathological images for the Gleason grading of prostate cancer.[40] Bulten *et al*.[41] proposed an extended version of UNet named CycleGan. Ren *et al*.,[42] their CNN architecture consists of an encoding and decoding network. The above studies suggest that automated Gleason grading via deep learning is a feasible task.

In this study, we try to develop a new patch-based classifier using deep transfer learning outputs based on a more recent and powerful set of pretrained CNNs architectures without the need for image segmentation and predict the label of each patch to determine the Gleason score of the TMA images into three classes: Benign, Gleason Grade 3 and Gleason Grade 4, 5. The rest of the paper is organized as follows. The material and method are presented in the next section including database explanation, patch creation, CNN introduction, transfer learning concept and special architectures, and finally evaluation metrics. In the following, the results of this study are presented. Finally, the discussion and conclusion are presented.

Materials and Methods

Database

Data used in this study consists of a set of prostate cancer TMA images of 244 patients.[43,44] The TMAs were prepared at the Vancouver Prostate Centre in Vancouver, Canada. The study was approved by the institutional Clinical Research Ethics Board (CREB No. H15-01064). The prostate TMA spots were annotated by each pathologist by delineating cancerous regions and assigning a Gleason score of 1 (low risk) to 5 (high risk) based on observable histological patterns according to the WHO/ISUP criteria. TMA spots without any cancerous pattern and containing only benign tissue were marked as benign. The schematic of these five groups were illustrated in Figure 1. Gleason pattern 3 describes well-formed and separated glands. Gleason pattern 4 includes fused glands and poorly formed glands. Gleason pattern 5 involves

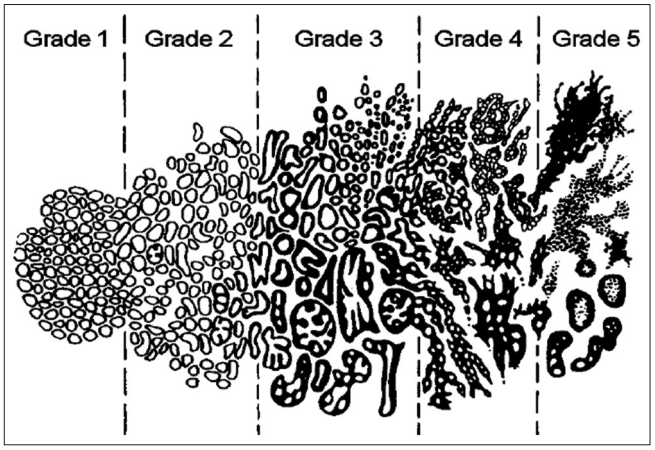

Figure 1: Schematic of Gleason Grades 1 to 5[45]

poorly differentiated individual cells, cords, and linear arrays.

The prostate TMA sots are annotated by six expert pathologists separately on their knowledge and experience and then the majority voting on six ground truth labels was done for further process as the final label. Using the expertise of six specialists in the automated classification of prostate cancer, we can develop a system to be more accurate in the test data and not dependent on a special pathologist. It should be noted, with regard to even number of the pathologists, the majority voting may be equal and not reach a single label. In this situation, the decision of the first pathologist considered the winner.

The final Gleason score for each TMA image is reported as the sum of the most prominent and second most prominent patterns, for example, a Gleason score of 4 + 3 will have a tissue with the most prominent pattern of Gleason Grade of 4 and the second most prominent pattern of Gleason Grade of 3. In our clinical dataset, the pathologist-annotated do not contain 1 and 2 grades and therefore the score is between 6 (3 + 3) and 10 (5 + 5). The subimages contain Grade of 3, 4, 5, and benign labels are illustrated in Figure 2.

Patch creation

The original image resolution of TMA was 5120 × 5120 pixels. For better model training, we need small image regions. Thus, we divided the original images into smaller ones called patches. Image patches with size 750 × 750 pixels were created from each full image using a step size of half of patch size, i.e., 375 pixels. Hence, each original image turns into 169 image patches considering step size equal to half of the patch size. Each patch was labeled according to the annotation in its central 250 × 250 region. Patches containing no or multiple annotations in the central region were discarded. Accordingly, 23,901 image patches are determined. The number of image patches which have benign (11,806), 3 (4703), 4 (7245), and 5 (147) grades were explicitly determined in Table 1. The collected images, for each of the pathological conditions, have different numbers. This difference can cause a weight bias for a class with more members in the process of learning the network, which, as a result, causes the test set of other classes to face errors. Therefore, it is necessary to use augmentation algorithms. Data augmentation methods actually refer to algorithms that create artificial additional data based on real members of classes, without imposing new information on the model. In other words, for grading classes with less data, a new data with the general characteristics of the main members are reproduced to make the data of that class competitive with other classes. Different algorithms have been proposed to implement this approach, which can be done in the simplest case by changing the color, rotation angle, cropping, noise level, spatial shift, etc., In this work, for greater simplicity, we only used the increase of the number of members of the classes by rotation and shift methods. Data augmentation was used during training with up to 10° rotation and 10% height shifts. The number of members of classes with more members was considered as a reference and the number of members of other classes was increased according to the difference with that class. Hence, the image patches are prepared to feed into the models in subsequent processing.

Convolutional neural networks

Deep learning algorithms compared with other conventional machine learning methods have become particularly popular for the identification and diagnosis of diseases in medical imaging with considerable performance improvements. One of the most popular deep-learning methods in the field of medical imaging is CNN.[46] It is the state-of-the-art deep learning methodology consisting of many stacked convolutional layers. The CNN structure comprises a convolutional layer, a maximum or average pooling layer, a nonlinear layer, batch normalization, fully connected (FC) layers, and finally a softmax layer. Pooling layers are frequently used among convolutional layers to boost translational invariance and lessen feature map extent. Nonlinear layers (mostly ReLU function) are used to strengthen the network for solving nonlinear problems. Finally, FC layers prepare extracted features to be classified by the softmax layer.

Pretrained convolutional neural network and transfer learning

The number of parameters in the model increase as networks gets deeper for improved learning efficiency. The deeper networks lead to the more complicated computations and the more demanding training data. We have only

Table 1: The number of benign, Grade 3, Grade 4 and Grade 5 patches in our clinical dataset

	Benign	Grade 3	Grade 4, 5
Number of patches	11,806	4703	7392

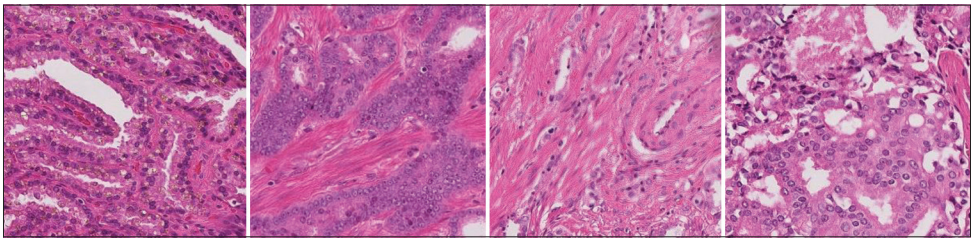

Figure 2: Benign and Grades 3, 4 and 5 from left to right

244 patients with 23,901 image patches to train, validate and test data. It seems too small to train a deep CNN and we need the transfer learning concept. Transfer learning employs take advantage of a pretrained model (CNN model) on a huge database.[47] The pretrained model is then fine-tuned on our new dataset with lesser number of training images as compared to formerly trained datasets. It means we transfer the information (the learned parameters) to our problem to train own model with an insufficient database.[48] In other words, pretrained CNN structures are modified to suit our task. This procedure is usually much faster than the conventional training of the CNN model with random weights. To train enormous parameters of a CNN model adequately, too much data is needed. We use transfer learning to compensate lack of much dataset and achieving better outcomes. The several special CNN architectures are trained on very large amounts of images with many categories and then named pretrained CNNs model. There are trained on ImageNet contains 14 million images of 1000 different categories from animals (dogs, cats, lions,…) to objects (desks, pens, chairs,…).[49] EfficientNets (B0-B5),[50] NasNetLarge,[51] NasNetMobile,[51] InceptionV3,[52] ResNet-50,[53] SeResnet50,[54] xception,[55] DenseNet121,[56] ResNext50,[57] and inception_resnet_v2[58] are popular pretrained CNNs. These networks have benefits for researchers such as lower training time, weaker and cheaper hardware requirement, lower computational load, and fewer images for training.

Multiple pretrained CNN models are utilized in this work and each one looks at input images in its special way and the way of processing information in hierarchical layers of these models makes some differences in extracted features. VGG16 model uses a simple cascaded architecture made of multiple convolutional layers, and can selectively extract optimal classificatory patterns from input images. The Xception model used channel-wise separable convolutions besides spatially separable convolutions for defragmentation of input images to best discriminative features in a stack of layers. InceptionResNetV2 combines the idea of a split, transform and merge within layer from inception modules and the idea of residual networks which believe deeper networks can provide better results. DenseNet121 uses multilayer connectivity to overcome problems due to overfitting by feeding each convolutional layer with multiple abstraction features obtained in previous layers. EfficientNet models utilize multiple tricks including model scaling and limitation over the shape of the input image to gain superior efficiency by better classification accuracy in the company of a lower number of parameters and fewer computational demands. All of these advantages made these models a proper choice for transfer learning applications.

Training details

We evaluated pretrained CNNs by fine-tuning them on our clinical dataset, separately. For this reason, we adopted 15 pretrained CNNs: EfficientNets (B0-B5), NasNetLarge, NasNetMobile, InceptionV3, ResNet-50, SeResnet50, xception, DenseNet121, ResNext50, and inception_resnet_v2. In order to fine-tune all networks, we only used the convolutional part of each model's architecture, removing all fully-connected layers. On top of the last convolutional layer, we added a global average pooling layer, followed by the final classification layer that uses softmax nonlinearity. For fine-tuning the networks, all models were fine-tuned for 50 epochs using Stochastic gradient descent (SGD) optimization with learning rate 0.0001, Nesterov momentum 0.9 and the batch size equal to 32. In all cases, the categorical cross-entropy loss was used as a minimization objective function. It should be noted that the input of each network is of a different size. Hence, in the first step of data preparation, according to different sizes of model inputs, all images were resized to proper sizes and stored in separate folders. Table 2 shows the input size of each pretrained CNN model. These models were trained using the same initialization and learning rate policies.

Evaluation metrics

Independently, 15 versions of the pretrained CNNs model were fine-tuned and prediction was used to score the probability of three classes: Benign, Gleason Grade 3, and Gleason Grade 4, 5 on the test set. Five-fold cross-validation is used to show a better vision of our network's capabilities. Common classification metrics named: Accuracy and area under the curve (AUC) were also used for the evaluation of proposed method. AUC/receiver operator characteristic (ROC) curve is a performance measurement for classification problems at various thresholds settings. ROC is a probability curve and AUC represents the degree of separability and tells how much model is capable of distinguishing between classes.

Results

Independently, 15 versions of pretrained CNNs model: EfficientNets (B0-B5), NasNetLarge, NasNetMobile, InceptionV3, ResNet-50, SeResnet50, Xception, DenseNet121, ResNext50 and inception_resnet_v2 were fine-tuned on a well-annotated dataset of prostate

Table 2: Different input size of pretrained convolutional neural networks

Model name	Input shape
EfficientNetB0, DenseNet121, NasNetMobile, ResNet50, ResNext50, Seresnet50 and Xception	224×224
EfficientNetB1	240×240
EfficientNetB2	260×260
inception_resnet_v2, and inception_v3	299×299
EfficientNetB3	300×300
NasnetLarge	331×331
EfficientNetB4	380×380
EfficientNetB5	456×456

carcinoma TMA images for 244 patients with 23,901 image patches. Deep learning was performed using Python version 3.5 programming language (Python Software Foundation, Beaverton, Oregon) with Keras version 2.1.5 software (GitHub, San Francisco, California) using a graphics processing unit (GeForce GTX 1080 Ti, NVIDIA, Santa Clara, California). Figures 3 and 4 show the value of the accuracy and loss function in the training and validation sets for fine-tuning of different pretrained CNN models respectively. As shown in these figures, the model converges in the training process after 50th epoch, and the data distribution ranges were narrow. Hence, after 50 steps of training, diagnostic accuracy was calculated using the test set. In Table 3, common classification metrics, accuracy per classes and AUC on test dataset in five folds are illustrated. The NasnetLarge framework gives higher accuracy (0.93) and AUC (0.98) in the classification of prostate TMA images as compared to other architectures. The deep learning architecture with the largest AUC was NasnetLarge (AUC: 0.98), inception_resnet_v2 (AUC: 0.96) and Xception (AUC: 0.95). Finally, the confusion matrix for the NASNetLarge was calculated and presented in Tables 4-8 for the classification of the prostate TMA images in three classes (benign, Grade 3, Grade 4, 5) on the test data set in five folds.

Discussion

In this research, we demonstrated the feasibility of 15 state-of-the-art pretrained CNNs to perform as Gleason-grade predictor on a well-annotated dataset of prostate carcinoma TMA images. To choose a suitable classifier, we explored different pretrained CNN architectures which have shown excellent performance on the ImageNet dataset, namely EfficientNets (B0-B5), NasNetLarge, NasNetMobile, InceptionV3, ResNet-50, SeResnet50, Xception, DenseNet121, ResNext50, and inception_resnet_v2. We observe that in terms of Accuracy

Table 3: Average classification metrics on test dataset using pretrained convolutional neural networks models in five folds

Model	Accuracy of three classes	Accuracy of benign	Accuracy of Grade 3	Accuracy of Grade 4, 5	AUC
EfficientNetB0	0.81	0.85	0.73	0.79	0.83
EfficientNetB1	0.79	0.83	0.71	0.77	0.82
EfficientNetB2	0.78	0.81	0.70	0.76	0.80
EfficientNetB3	0.66	0.69	0.59	0.64	0.68
EfficientNetB4	0.71	0.75	0.63	0.70	0.70
EfficientNetB5	0.68	0.72	0.60	0.68	0.69
inception_ resnet_v2	0.91	0.96	0.84	0.91	0.96
InceptionV3	0.78	0.81	0.70	0.76	0.84
NASNetLarge	0.93	0.95	0.89	0.92	0.98
NASNetMobile	0.80	0.85	0.73	0.76	0.86
ResNet50	0.89	0.91	0.85	0.88	0.90
Xception	0.90	0.95	0.85	0.90	0.95
DenseNet121	0.79	0.83	0.71	0.77	0.82
SeResnet50	0.80	0.85	0.73	0.76	0.86
ResNext50	0.79	0.83	0.71	0.77	0.82

AUC – Area under the curve

Table 4: Confusion matrix obtained for the NasNetLarge for classification of the prostate tissue microarray images in three classes (benign, Grade 3, Grade 4, 5) on the test data set in fold 1

Reference	Estimated labels by the proposed method		
	Benign	Grade 3	Grade 4, 5
Benign	2241	66	54
Grade 3	22	842	76
Grade 4, 5	24	88	1367

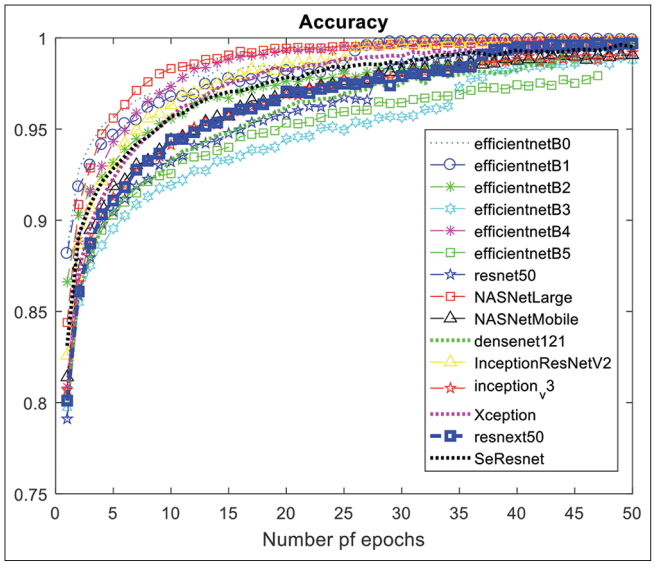

Figure 3: The train accuracy for different pretrained CNN models. CNN: Convolutional neural network

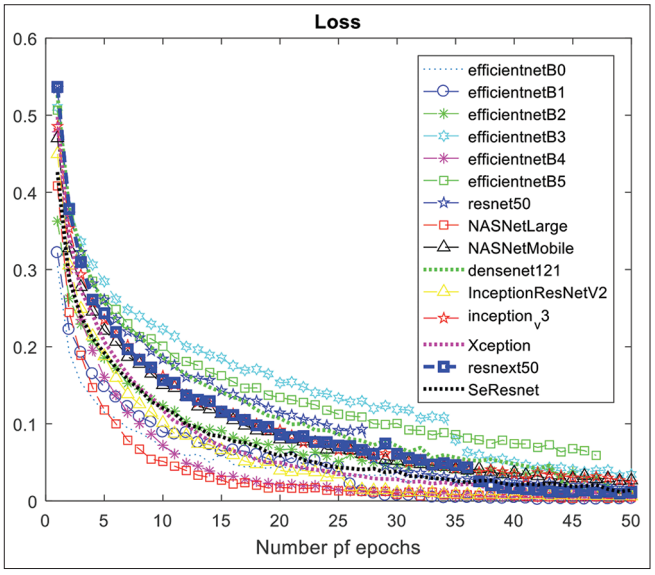

Figure 4: The cross-entropy loss function for different pretrained CNN models. CNN: Convolutional neural network

Table 5: Confusion matrix obtained for the NasNetLarge for classification of the prostate tissue microarray images in three classes (benign, Grade 3, Grade 4, 5) on the test data set in fold 2

Reference	Estimated labels by the proposed method		
	Benign	Grade 3	Grade 4, 5
Benign	2237	68	56
Grade 3	26	837	78
Grade 4, 5	26	92	1360

Table 6: Confusion matrix obtained for the NASNetLarge for classification of the prostate tissue microarray images in three classes (benign, grade 3, grade 4, 5) on the test data set in fold 3

Reference	Estimated labels by the proposed method		
	Benign	Grade 3	Grade 4, 5
Benign	2248	62	51
Grade 3	20	846	75
Grade 4, 5	21	84	1373

Table 7: Confusion matrix obtained for the NasNetLarge for classification of the prostate tissue microarray images in three classes (benign, Grade 3, Grade 4, 5) on the test data set in fold 4

Reference	Estimated labels by the proposed method		
	Benign	Grade 3	Grade 4, 5
Benign	2243	65	54
Grade 3	20	844	76
Grade 4, 5	20	85	1374

Table 8: Confusion matrix obtained for the NasNetLarge for classification of the prostate tissue microarray images in three classes (benign, Grade 3, Grade 4, 5) on the test data set in fold 5

Reference	Estimated labels by the proposed method		
	Benign	Grade 3	Grade 4, 5
Benign	2233	69	59
Grade 3	18	849	74
Grade 4, 5	20	92	1366

Table 9: Comparison of classification results of our work with other studies in classification of the prostate tissue microarray images in four classes

Study	Methods or features	Results
Karimi *et al*.[43]	CNNs combined using a logistic regression model	Low-versus high grade (Grade 3 vs. Grades 4, 5) Accuracy=86%
Nir *et al*.[44]	Random forest classifier	Low-versus high grade (Grade 3 vs. Grades 4, 5) Accuracy=79.4%
Our work	Patch-based classifier using deep transfer learning by NasNetLarge	Three classes (benign, Grade 3, Grade 4, 5) Accuracy=93%

CNNs – Convolutional neural networks

and AUC metrics, the NasnetLarge architecture is the best model among them. Table 9 compares the results of our work with other studies[43,44] which use the same database in the classification of the prostate TMA images in three classes.

It is possible to minimize prostate cancer fatality when patients are diagnosed and treated early. Screening is momentous for early diagnosed. However, the lack of specialists in some areas and high expenses are some restrictions on screening. Artificial intelligence methods like this paper can overcome this limitation and can be utilized everywhere with no need to highly-trained pathologist.

The histopathologic images of this paper were labeled by six pathologists and majority vote was applied on pixel-wise annotations to obtain a unified label. In other words, the trained model made a profit of experience and expertise blend of all six pathologists and compact the knowledge of them. Some researches on Gleason grading have used a single expert dataset to train and test their models overlooking many reports of reproducibility obstacle.[16-19]

There are two main reasons why image patches (169 image patches from each image) are generated from the original TMA image. First, as previously stated, there are 5120 × 5120 pixels in the TMA images and this is high resolution. Consequently, different regions of image may have different labels and cannot be viewed as a single label. Another evidence for patch formation is that the entire image cannot be inserted into the pretrained CNN models even if we disregard the first mentioned point. For example, ResNet50 model requires 224 × 224-pixel images as input. If we compress the 5120 × 5120 image to 224 × 224, most of image information will be lost and the classification will not be efficient.

Conclusion

We actually use a method based on transfer learning with CNNs model as a highly trained pathologist to categorize the prostate cancer stages based on the Gleason grading system. This system distinguishes the histological structures by allocating the Grades from 3, Grades from 4, 5, and benign. The NasnetLarge framework gives higher accuracy (0.93) and AUC (0.98) in the classification of prostate TMA images as compared to other architectures.

Acknowledgments

This research is supported by the Cancer Research Center of Shahid Beheshti University of Medical Sciences (Grant No 29560).

Financial support and sponsorship

None.

Conflicts of interest

There are no conflicts of interest.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin 2015;65:5-29.
2. Gleason DF, Mellinger GT. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. J Urol 1974;111:58-64.
3. Faraj SF, Bezerra SM, Yousefi K, Fedor H, Glavaris S, Han M, *et al.* Clinical validation of the 2005 ISUP Gleason grading system in a cohort of intermediate and high risk men undergoing radical prostatectomy. PLoS One 2016;11:e0146189.
4. Gordetsky J, Epstein J. Grading of prostatic adenocarcinoma: Current state and prognostic implications. Diagn Pathol 2016;11:25.
5. Epstein JI. Prostate cancer grading: A decade after the 2005 modified system. Mod Pathol 2018;31:S47-63.
6. Allsbrook WC Jr., Mangold KA, Johnson MH, Lane RB, Lane CG, Amin MB, *et al.* Interobserver reproducibility of Gleason grading of prostatic carcinoma: Urologic pathologists. Hum Pathol 2001;32:74-80.
7. Allsbrook WC Jr., Mangold KA, Johnson MH, Lane RB, Lane CG, Epstein JI. Interobserver reproducibility of Gleason grading of prostatic carcinoma: General pathologist. Hum Pathol 2001;32:81-8.
8. Salmo EN. An audit of inter-observer variability in Gleason grading of prostate cancer biopsies: The experience of central pathology review in the NorthWest of England. Integr Cancer Sci Ther 2015;2:104-6.
9. Egevad L, Ahmad AS, Algaba F, Berney DM, Boccon-Gibod L, Compérat E, *et al.* Standardization of Gleason grading among 337 European pathologists. Histopathology 2013;62:247-56.
10. McLean M, Srigley J, Banerjee D, Warde P, Hao Y. Interobserver variation in prostate cancer Gleason scoring: Are there implications for the design of clinical trials and treatment strategies? Clin Oncol (R Coll Radiol) 1997;9:222-5.
11. Allsbrook WC Jr., Mangold KA, Johnson MH, Lane RB, Lane CG, Amin MB, *et al.* Interobserver reproducibility of Gleason grading of prostatic carcinoma: Urologic pathologists. Hum Pathol 2001;32:74-80.
12. Epstein JI, Zelefsky MJ, Sjoberg DD, Nelson JB, Egevad L, Magi-Galluzzi C, *et al*. A contemporary prostate cancer grading system: A validated alternative to the Gleason score. Eur Urol 2016;69:428-35.
13. Evans SM, Patabendi Bandarage V, Kronborg C, Earnest A, Millar J, Clouston D. Gleason group concordance between biopsy and radical prostatectomy specimens: A cohort study from Prostate cancer outcome registry – Victoria. Prostate Int 2016;4:145-51.
14. Gertych A, Ing N, Ma Z, Fuchs TJ, Salman S, Mohanty S, *et al.* Machine learning approaches to analyze histological images of tissues from radical prostatectomies. Comput Med Imaging Graph 2015;46 Pt 2:197-208.
15. Nguyen K, Sabata B, Jain AK. Prostate cancer grading: Gland segmentation and structural features. Pattern Recognit Lett 2012;33:951-61.
16. Doyle S, Feldman MD, Shih N, Tomaszewski J, Madabhushi A. Cascaded discrimination of normal, abnormal, and confounder classes in histopathology: Gleason grading of prostate cancer. BMC Bioinformatics 2012;13:282.
17. Doyle S, Feldman M, Tomaszewski J, Madabhushi A. A boosted Bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies. IEEE Trans Biomed Eng 2012;59:1205-18.
18. Gorelick L, Veksler O, Gaed M, Gomez JA, Moussa M, Bauman G, *et al.* Prostate histopathology: Learning tissue component histograms for cancer detection and classification. IEEE Trans Med Imaging 2013;32:1804-18.
19. Nguyen K, Sarkar A, Jain AK. Prostate cancer grading: Use of graph cut and spatial arrangement of nuclei. IEEE Trans Med Imaging 2014;33:2254-70.
20. Litjens G, Kooi T, Bejnordi BE, Setio AA, Ciompi F, Ghafoorian M, *et al*. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60-88.
21. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J Pathol Inform 2016;7:29.
22. Greenspan H, van Ginneken B, Summers RM. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. IEEE Trans Med Imaging 2016;35:1153-9.
23. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018;2:158-64.
24. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115-8.
25. Lakshmanaprabu SK, Mohanty SN, Shankar K, Arunkumar N, Ramirez G. Optimal deep learning model for classification of lung cancer on CT images. Future Gener Comput Syst 2019;92:374-82.
26. Litjens G, Ciompi F, Wolterink JM, de Vos BD, Leiner T, Teuwen J, *et al.* State-of-the-Art deep learning in cardiovascular image analysis. JACC Cardiovasc Imaging 2019;12:1549-65.
27. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018;24:1342-50.
28. Chea P, Mandell JC. Current applications and future directions of deep learning in musculoskeletal radiology. Skeletal Radiol 2020;49:183-97.
29. Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. Med Image Comput Comput Assist Interv 2013;16:411-8.
30. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6:26286.
31. Liu Y, Gadepalli K, Norouzi M, Dahl GE, Kohlberger T, Boyko A, *et al*. Detecting Cancer Metastases on Gigapixel Pathology Images arXiv:1703.02442 (2017). doi: 10.48550/arXiv.1703.02442.
32. Källén H, Molin J, Heyden A, Lundström C, Åström K. Towards Grading Gleason Score using Generically Trained Deep Convolutional Neural Networks. IEEE 13th International Symposium on Biomedical Imaging (ISBI); 2016. p. 1163-67.
33. Zhou N, Fedorov A, Fennessy F, Kikinis R, Gao Y. Large scale digital prostate pathology image analysis combining feature extraction and deep neural network. arXiv:1705.02678 (2017). doi: 10.48550/arXiv.1705.02678.
34. Oscar TD, Manfredo A, Sebastian O, Mats A, Kristian E, Martin H, *et al*. Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade Gleason score. Medical Imaging 2017: Digital Pathology. International Society for Optics and Photonics 2017.

35. Arvaniti E, Fricker KS, Moret M, Rupp N, Hermanns T, Fankhauser C, *et al.* Automated Gleason grading of prostate cancer tissue microarrays via deep learning. Sci Rep 2018;8:12054.
36. Li W, Li J, Sarma KV, Ho KC, Shen S, Knudsen BS, *et al.* Path R-CNN for prostate cancer diagnosis and gleason grading of histological images. IEEE Trans Med Imaging 2019;38:945-54.
37. Linkon AH, Labib M, Hasan T, Hossain M. Deep learning in prostate cancer diagnosis and Gleason grading in histopathology images: An extensive study. Inform Med Unlocked 2021;24:100582.
38. Leyh-Bannurah SR, Wolffgang U, Schmitz J, Ouellet V, Azzi F, Tian Z, *et al.* MP19-20 state-of-the-art weakly supervised automated classification of prostate cancer tissue microarrays via deep learning: Can sufficient accuracy be achieved without manual patch level annotation? J Urol 2020;203 Suppl 4:e306.
39. Abbasi AA, Hussain L, Awan IA, Abbasi I, Majid A, Nadeem MS, *et al.* Detecting prostate cancer using deep learning convolution neural network with transfer learning approach. Cogn Neurodyn 2020;14:523-33.
40. Zeng Z, Xie W, Zhang Y, Lu Y. RIC-unet: An improved neural network based on unet for nuclei segmentation in histology images. IEEE Access 2019;7:21420-8.
41. Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B, *et al.* Automated deep-learning system for Gleason grading of prostate cancer using biopsies: A diagnostic study. Lancet Oncol 2020;21:233-41.
42. Ren J, Sadimin E, Foran DJ, Qi X. Computer aided analysis of prostate histopathology images to support a refined Gleason grading system. Proc SPIE Int Soc Opt Eng 2017;10133:101331V.
43. Karimi D, Nir G, Fazli L, Black PC, Goldenberg L, Salcudean SE. Deep learning-based gleason grading of prostate cancer from histopathology images-role of multiscale decision aggregation and data augmentation. IEEE J Biomed Health Inform 2020;24:1413-26.
44. Nir G, Hor S, Karimi D, Fazli L, Skinnider BF, Tavassoli P, *et al.* Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. Med Image Anal 2018;50:167-80.
45. O'Dowd GJ, Veltri RW, Miller MC, Strum SB. The Gleason score: A significant biologic manifestation of prostate cancer aggressiveness on biopsy. PCRI Insights 2001;4:1-5.
46. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60:84-90.
47. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, *et al.* Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans Med Imaging 2016;35:1299-312.
48. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A Survey on Deep Transfer Learning. In: Kůrková V, Manolopoulos Y, Hammer B, Iliadis L, Maglogiannis I, editors. Artificial Neural Networks and Machine Learning – ICANN 2018. Lecture Notes in Computer Science. Vol. 11141. Cham: Springer; 2018.
49. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A Large-Scale Hierarchical Image Database. In: CVPR; 2009.
50. Tan M, Le QV. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv:1905.11946. doi: org/10.48550/arXiv.1905.11946.
51. Zoph B, Vasudevan V, Shlens J, Le QV. Learning transferable architectures for scalable image recognition. arXiv:1707.07012. doi: 10.48550/arXiv.1707.07012.
52. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. "Rethinking the Inception Architecture for Computer Vision". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 2818-26.
53. He K, Zhang X, Ren S, Sun J. "Deep Residual Learning for Image Recognition". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 770-8.
54. Hu J, Shen L, Sun G. Squeeze -and -excitation networks. arXiv:1709.01507. doi: 10.48550/arXiv.1709.01507.
55. Chollet F. Xception: Deep learning with depthwise separable convolutions. arXiv:1610.02357. doi: 10.48550/arXiv.1610.02357.
56. Huang G, Liu Z, van der Maaten L, Weinberger KQ. "Densely Connected Convolutional Networks". In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2017.
57. Saining X, Ross G, Piotr D, Zhuowen T, Kaiming H. Aggregated Residual Transformations for Deep Neural Networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. p. 1492-500.
58. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. "Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning". In: AAAI; 2017. p. 4278-84.