Systematic comparison of deep learning strategies for weakly supervised Gleason grading

Sebastian Otálora[a,b], Manfredo Atzori[b], Amjad Khan[b], Oscar Jimenez-del-Toro[a], Vincent Andrearczyk[b], and
Henning Müller[a,b]

[a]Institute of Information Systems, University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
[b]University of Geneva, Geneva, Switzerland

ABSTRACT

Prostate cancer (PCa) is one of the most frequent cancers in men. Its grading is required before initiating its treatment. The Gleason Score (GS) aims at describing and measuring the regularity in gland patterns observed by a pathologist on the microscopic or digital images of prostate biopsies and prostatectomies. Deep Learning-based (DL) models are the state-of-the-art computer vision techniques for Gleason grading, learning high-level features with high classification power. However, for obtaining robust models with clinical-grade performance, a large number of local annotations are needed. Previous research showed that it is feasible to detect low and high-grade PCa from digitized tissue slides relying only on the less expensive report–level (weakly) supervised labels, thus global rather than local labels. Despite this, few articles focus on classifying the finer-grained GS classes with weakly supervised models. The objective of this paper is to compare weakly supervised strategies for classification of the five classes of the GS from the whole slide image, using the global diagnostic label from the pathology reports as the only source of supervision. We compare different models trained on hand-crafted features, shallow and deep learning representations. The training and evaluation are done on the publicly available TCGA-PRAD dataset, comprising of 341 whole slide images of radical prostatectomies, where small patches are extracted within tissue areas and assigned the global report label as ground truth. Our results show that DL networks and class-wise data augmentation outperform other strategies and their combinations, reaching a kappa score of $\kappa = 0.44$, which could be further improved with a larger dataset or combining both strong and weakly supervised models.

Keywords: Computational pathology, prostate cancer, deep learning, weak supervision

1. INTRODUCTION

Previous research showed that it is feasible to detect low and high-grade PCa from digitized tissue slides relying only on the less expensive report–level (weakly) supervised labels, thus global rather than local labels. Despite this, few articles focus on classifying the finer-grained GS classes with weakly supervised models. The objective of this paper is to compare weakly supervised strategies for classification of the five classes of the GS from the whole slide image, using the global diagnostic label from the pathology reports as the only source of supervision. We compare different models trained on hand-crafted features, shallow and deep learning representations. The training and evaluation are done on the publicly available TCGA-PRAD dataset, comprising of 341 whole slide images of radical prostatectomies, where small patches are extracted within tissue areas and assigned the global report label as ground truth. Our results show that DL networks and class-wise data augmentation outperform other strategies and their combinations, reaching a kappa score of $\kappa = 0.44$, which could be further improved with a larger dataset or combining both strong and weakly supervised models.

Prostate Cancer (PCa) is the fourth most common cancer, and in 2018 there were 1.28 million new diagnoses of it worldwide*. It is a highly heterogeneous disease displaying different tumor types in glands and having

Further author information: (Send correspondence to S.O.)
S.O.: E-mail: juan.otaloramontenegro@hevs.ch

*https://www.who.int/en/news-room/fact-sheets/detail/cancer: Retrieved 1st of January, 2020 high inter-rater variability among pathologists.[1] The Gleason score (GS) system is used in clinical practice as a standard protocol when assessing prostate adenocarcinoma. GS aims to quantify the tumor aggressiveness and disease prognosis to prepare the treatment. GS is the sum of the two most prominent Gleason patterns (GP) observed in the tissue slide, GPs range from 1 to 5, producing a GS that usually ranges from 6 to 10, since GP 1 and 2 are rarely reported, because biopsies are seldom taken in these cases. Computational pathology methods that automatically measure the extent of the GP in the image and estimate the final GS are clinically relevant to obtain more objective (quantitative) and reproducible diagnoses. Such methods could then be implemented in the (digital) pathology laboratories as a decision support system or as an automatic screening, separating cancer from benign slides.[2] The digitized whole slide images (WSI) that can be up to $100,000^2$ pixels pose a computational challenge because of their massive size and consequently, the critical work to annotate the images thoroughly, which leads to a small amount of annotated regions of the total tissue.[3] Studies that automatically predict GP and GS have shown that it is feasible to use machine learning methods to learn from large data sets with annotated regions.[4–10] The machine learning models for computer-aided PCa diagnosis were previously linear classifiers trained with a set of expert predefined image features, also known as hand-crafted features or feature-engineering in machine learning literature. Hand-crafted features for PCa include hematoxylin and eosin (H&E) or immunohistochemistry (IHC) intensity, morphological features of nuclei and glands, texture filter-banks, and graph-related features, among others.[8] In the last decade, computational methods that automatically learn the relevant features directly from the images, mainly Convolutional Neural Networks (CNN), have been applied successfully in medical imaging tasks and particularly in computational pathology.[4,11,12] CNN based models have outperformed hand-crafted features when a large amount of annotated data are available. Despite the potential of CNN models in computational pathology, there is still a significant barrier for training robust models: a costly pathologist-intensive annotation process to obtain these large data sets with their corresponding manual annotations. Even with the lack of detailed annotations in the WSI, weakly supervised learning approaches in computational pathology aim to train models without using costly pixel-wise labels but instead using the global report-level labels as the source of supervision. Those readily available labels account for the global findings of the pathologist in the image, but without any manual delineation of the specific regions analyzed,[6,9,13] e.g., the overall GS without any annotations in the image of where are precisely the two predominant GPs.

A summary of the performance reported for PCa grading and detection of the most recent studies is presented in Table 1. From the reported results in the literature, two main factors arise for the success of automated GP and GS classification DL models: The use of regional annotations and the number of patients, and therefore of WSI included in the training of deep CNN models. In the work of Arvainiti et al,[7] the authors train a deep learning network with a small set of well-curated annotations from tissue micro-arrays of 641 patients, obtaining a $\kappa = 0.75$, comparable with the inter-rater agreement among pathologists. Nagpal et al[14] used 112 million annotated image patches derived from 912 slides to train a Gleason pattern and scoring model, obtaining an accuracy of 0.7 for the five Gleason scores. Burlutskiy et al[15] trained a CNN with glandular tissue without basal cells obtained from immunofluorescence images, achieving a F_1-score of 0.80 for the task of PCa detection in a test set of 63 biopsies. In the work of Ström et al,[16] the authors trained a deep CNN with 6682 needle biopsies from 976 patients. The CNN obtained $\kappa = 0.62$ to classify the five groups stated by the International Society of Urological Pathology (ISUP): GS6, GS7=3+4, GS7=4+3, GS8,[GS9, GS10], obtaining also a performance within the range of the corresponding values for the expert pathologists ($\kappa \in [0.60, 0.73]$).

Jimenez-del-Toro et al[17] trained a CNN model, with pathology report labels only, to classify low vs. high-grade prostatectomies from 235 patients in the TCGA repository obtaining an accuracy of 78%. In the work of Campanella et al,[9] the authors used transfer learning and a massive data set of more than $12,132$ WSI's of prostate biopsies. Using this dataset, they train weakly supervised two-class CNN classifiers obtaining a PCa detection model with an Area Under the ROC Curve (AUC) of 0.986 allowing to ignore more than 75% of the slides while retaining 100% sensitivity. While their results paved the way for automated screening tools in computational pathology (where the pathologist can discard non-cancerous slides), their generalization to fine-grained Gleason pattern classes and clinical scenarios with highly heterogeneous data remains to be confirmed (the latter being an important point not only for PCa but for computational pathology in general). Finally, Bulten et al[13] recently proposed an automatic PCa grading system with DL using weak supervision from the report labels with 1243 biopsies. Their system consisted of multiple stages: first, three CNN models are used to perform cancer detection, discarding non-epithelial tissue, and segmenting each gland pattern. Then, a model is trained and refined using *pure* biopsies (GS6 = 3+3, GS8 = 4+4, GS10 = 5+5). Their model obtained $\kappa = 0.723$ on the external test data set of Arvaniti,[7] remarkably achieving performance in the range of strongly supervised methods.

Table 1: Reported performance for PCa grading and scoring using deep learning models. The first five rows correspond to strongly supervised methods using pixel-wise annotations. The last three rows are weakly supervised methods that use global labels. MC stand for Multi-Center, i.e., if the study involved images from multiple institutions, which increases complexity and requires good generalization performance

Reference	**Classes**	**Results**	**#Patients**	**Annotations**	**MC**
Arvaniti[7]	GS6,GS7,GS8,GS9,GS10	$\kappa = 0.75$	641	Strong	No
Nagpal[14]	GS6,GS7,GS8,GS9,GS10	ACC= 0.70	342	Strong	Yes
Burlutskiy[15]	With/out basal cells	$F_1 = 0.80$	229	Strong	No
Ström[16]	ISUP: 1,2,3,4,5	$\kappa = 0.67$	976	Strong	Yes
Jimenez-del-Toro[17]	[GS6, GS7] vs [GS8, GS9,GS10]	ACC= 0.78	235	Weak	Yes
This paper	GS6, GS7, GS8, GS9, GS10	$\kappa = 0.441$	341	Weak	Yes
Campanella[9]	Benign vs Cancer	AUCs of 0.986	7159	Weak	Yes
Bulten[13]	ISUP: 1,2,3,4,5	$\kappa = 0.723$	1243	Weak	Yes

In this paper, we compare weakly supervised models for the classification of the Gleason score, directly from H&E whole slide images of prostatectomies. The training of the models uses the global diagnostic label from the pathology reports as the only source of supervision. The trained models explore strategies based on a combination of DL features with morphological features extracted from automatically segmented nuclei, data augmentation, and depth of machine learning models. For the first time, we set robust baselines using DL for the task of fine-grained Gleason grading towards a clinically-usable computer-assisted diagnosis system with limited data and annotations. Finally, we also provide insights for better data-augmentation and aggregation strategies in this challenging task.

2. METHODS

2.1 Data set

The data set consists of 341 cases of prostatectomies WSIs from the public resource of The Cancer Genome Atlas repository of prostate adenocarcinoma (TCGA-PRAD)[†]. We pair each WSI with its corresponding GS label from the provided pathology reports. Due to the massive pixel size of a WSI, the features extracted and learned come from small regions in the image. We ensure that these regions (patches) are large enough to capture gland structures. The patch extraction is performed only in the tissue regions, using the Blue-Ratio (BR) mapping described in Chang et al.[18] BR mapping restrains areas without nuclei such as those containing fat, connective tissue, or background. Regions of 500 × 500 pixels are computed at a 20× apparent magnification (0.5 microns per pixel). The central 224 × 224 pixels of these regions are extracted and are the input patches for training all the DL networks. The number of WSIs used and the number of extracted patches are in Table 2, while two example WSIs with a subset of the locations of their patches are displayed in Figure 1.

2.2 Morphological features from the nuclei

The alterations in the nuclei features, such as nuclei size, shape, texture, and spatial architecture, reflect the complex molecular-level changes that occur during the formation of cancer and are a hallmark of PCa.[8] Initially, we evaluate the classification performance of hand-crafted morphological nuclei features. Nuclei features are computed from automatically segmented nuclei from each patch. The mask-RCNN model,[19] a multi-instance based DL segmentation technique, is used to segment each nucleus. The Mask-RCNN segmenter is fine-tuned on a separate data set[‡] with annotated nuclei of the prostate. After each nucleus mask instance is computed, the following features are extracted: area, diameter of a circle with the same area as the nucleus, ratio of pixels inside

[†] https://portal.gdc.cancer.gov/projects/TCGA-PRAD Retrieved 1st of January, 2020

[‡] Available at https://nucleisegmentationbenchmark.weebly.com/: Retrieved 1st of January,2020

Table 2: Number of patches and WSIs in parenthesis, extracted from the TCGA-PRAD data set, that are used to validate the performance of the models.

Class/Set	Training	Validation	Test
GS6	44,758 (23)	28,006 (17)	11,690 (6)
GS7	155,849 (88)	48,310 (27)	33,391(17)
GS8	66,302 (35)	27,654 (15)	21,232 (11)
GS9	95,519 (50)	43,446 (24)	30,868 (16)
GS10	11,665 (6)	3,855 (2)	1,944 (1)
Total	**375,093 (205)**	**151,271 (85)**	**99,125 (51)**

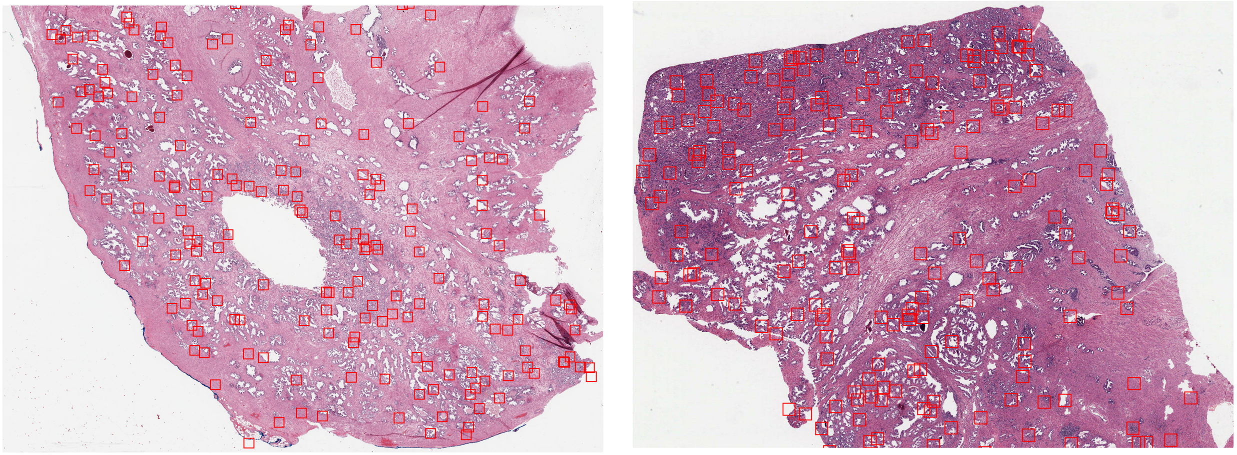

Figure 1: Example of patch extraction for two TCGA-PRAD WSIs with GS 6. The selected patch locations (red bounding boxes) target glands and areas with a high density of nuclei.

the nucleus to pixels in the total bounding box enclosing it, major and minor axis lengths, area of the bounding box enclosing the nucleus, perimeter, eccentricity, convex area, solidity, and orientation. The median feature value, for all nuclei in the patch, of the 11 features, is used as an 11-dimensional feature vector for each patch. An additional feature is added to the feature vector: the number of nuclei in each patch, resulting in a final 12-dimensional feature vector of hand-crafted features. Example patch segmentations are shown in Figure 2.

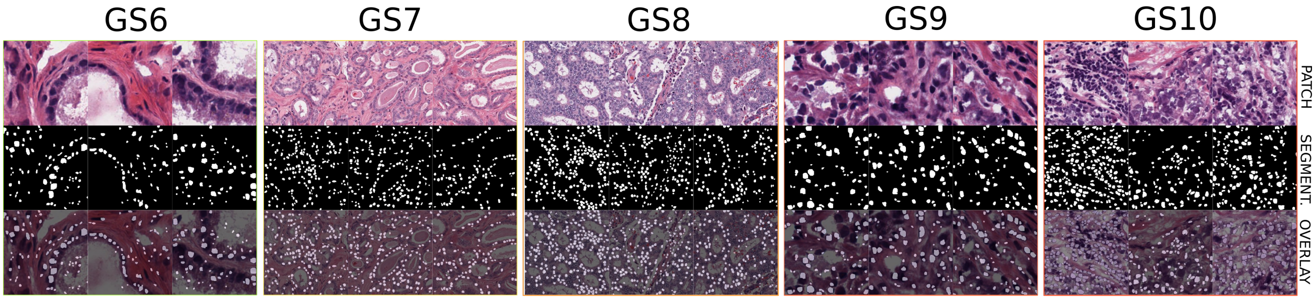

Figure 2: Three examples of the automatic nuclei segmentations for each of the five Gleason Scores. First row of each group are the original patches. The second row are the resulting binary masks after aggregating all the individual nuclei masks in the patch. The third column is the overlay of the segmentations over the original patches.

2.3 CNN architectures

Prostate cancer (PCa) is one of the most frequent cancers in men. Its grading is required before initiating its treatment. The Gleason Score (GS) aims at describing and measuring the regularity in gland patterns observed by a pathologist on the microscopic or digital images of prostate biopsies and prostatectomies. Deep Learning-based (DL) models are the state-of-the-art computer vision techniques for Gleason grading, learning high-level features with high classification power. However, for obtaining robust models with clinical-grade performance, a large number of local annotations are needed.

Introduction Previous research showed that it is feasible to detect low and high-grade PCa from digitized tissue slides relying only on the less expensive report–level (weakly) supervised labels, thus global rather than local labels. Despite this, few articles focus on classifying the finer-grained GS classes with weakly supervised models. The objective of this paper is to compare weakly supervised strategies for classification of the five classes of the GS from the whole slide image, using the global diagnostic label from the pathology reports as the only source of supervision. We compare different models trained on hand-crafted features, shallow and deep learning representations. The training and evaluation are done on the publicly available TCGA-PRAD dataset, comprising of 341 whole slide images of radical prostatectomies, where small patches are extracted within tissue areas and assigned the global report label as ground truth. Our results show that DL networks and class-wise data augmentation outperform other strategies and their combinations, reaching a kappa score of $\kappa = 0.44$, which could be further improved with a larger dataset or combining both strong and weakly supervised models.

Prostate Cancer (PCa) is the fourth most common cancer, and in 2018 there were 1.28 million new diagnoses of it worldwide[§]. It is a highly heterogeneous disease displaying different tumor types in glands and having high inter-rater variability among pathologists.[1] The Gleason score (GS) system is used in clinical practice as a standard protocol when assessing prostate adenocarcinoma. GS aims to quantify the tumor aggressiveness and disease prognosis to prepare the treatment. GS is the sum of the two most prominent Gleason patterns (GP) observed in the tissue slide, GPs range from 1 to 5, producing a GS that usually ranges from 6 to 10, since GP 1 and 2 are rarely reported, because biopsies are seldom taken in these cases. Computational pathology methods that automatically measure the extent of the GP in the image and estimate the final GS are clinically relevant to obtain more objective (quantitative) and reproducible diagnoses. Such methods could then be implemented in the (digital) pathology laboratories as a decision support system or as an automatic screening, separating cancer from benign slides.[2] The digitized whole slide images (WSI) that can be up to $100,000^2$ pixels pose a computational challenge because of their massive size and consequently, the critical work to annotate the images thoroughly, which leads to a small amount of annotated regions of the total tissue.[3] Studies that automatically predict GP and GS have shown that it is feasible to use machine learning methods to learn from large data sets with annotated regions.[4–10] The machine learning models for computer-aided PCa diagnosis were previously linear classifiers trained with a set of expert predefined image features, also known as hand-crafted features or feature-engineering in machine learning literature. Hand-crafted features for PCa include hematoxylin and eosin (H&E) or immunohistochemistry (IHC) intensity, morphological features of nuclei and glands, texture filter-banks, and graph-related features, among others.[8] In the last decade, computational methods that automatically learn the relevant features directly from the images, mainly Convolutional Neural Networks (CNN), have been applied successfully in medical imaging tasks and particularly in computational pathology.[4, 11, 12] CNN based models have outperformed hand-crafted features when a large amount of annotated data are available. Despite the potential of CNN models in computational pathology, there is still a significant barrier for training robust models: a costly pathologist-intensive annotation process to obtain these large data sets with their corresponding manual annotations. Even with the lack of detailed annotations in the WSI, weakly supervised learning approaches in computational pathology aim to train models without using costly pixel-wise labels but instead using the global report-level labels as the source of supervision. Those readily available labels account for the global findings of the pathologist in the image, but without any manual delineation of the specific regions analyzed,[6, 9, 13] e.g., the overall GS without any annotations in the image of where are precisely the two predominant GPs. A summary of the performance reported for PCa grading and detection of the most recent studies is presented in Table 1. From the reported results in the literature, two main factors arise for the success of automated GP and GS classification DL models: The use of regional annotations and the number of patients, and therefore of WSI included in the training of deep CNN models. In the work of Arvainiti et al,[7] the authors train a deep learning network with a small set of well-curated annotations from tissue micro-arrays of 641 patients, obtaining a $\kappa = 0.75$, comparable with the inter-rater agreement among pathologists. Nagpal et al[14] used 112 million annotated image patches derived from 912 slides to train a Gleason pattern and scoring model, obtaining an accuracy of 0.7 for the five Gleason scores. Burlutskiy et al[15] trained a CNN with glandular tissue without basal cells obtained from

[§] https://www.who.int/en/news-room/fact-sheets/detail/cancer: Retrieved 1st of January, 2020 immunofluorescence images, achieving a F_1-score of 0.80 for the task of PCa detection in a test set of 63 biopsies. In the work of Ström et al,[16] the authors trained a deep CNN with 6682 needle biopsies from 976 patients. The CNN obtained $\kappa = 0.62$ to classify the five groups stated by the International Society of Urological Pathology (ISUP): GS6, GS7=3+4, GS7=4+3, GS8,[GS9, GS10], obtaining also a performance within the range of the corresponding values for the expert pathologists ($\kappa \in [0.60, 0.73]$).

Jimenez-del-Toro et al[17] trained a CNN model, with pathology report labels only, to classify low vs. high-grade prostatectomies from 235 patients in the TCGA repository obtaining an accuracy of 78%. In the work of Campanella et al,[9] the authors used transfer learning and a massive data set of more than $12,132$ WSI's of prostate biopsies. Using this dataset, they train weakly supervised two-class CNN classifiers obtaining a PCa detection model with an Area Under the ROC Curve (AUC) of 0.986 allowing to ignore more than 75% of the slides while retaining 100% sensitivity. While their results paved the way for automated screening tools in computational pathology (where the pathologist can discard non-cancerous slides), their generalization to fine-grained Gleason pattern classes and clinical scenarios with highly heterogeneous data remains to be confirmed (the latter being an important point not only for PCa but for computational pathology in general). Finally, Bulten et al[13] recently proposed an automatic PCa grading system with DL using weak supervision from the report labels with 1243 biopsies. Their system consisted of multiple stages: first, three CNN models are used to perform cancer detection, discarding non-epithelial tissue, and segmenting each gland pattern. Then, a model is trained and refined using *pure* biopsies (GS6 = 3+3, GS8 = 4+4, GS10 = 5+5). Their model obtained $\kappa = 0.723$ on the external test data set of Arvaniti,[7] remarkably achieving performance in the range of strongly supervised methods.

Reported performance for PCa grading and scoring using deep learning models. The first five rows correspond to strongly supervised methods using pixel-wise annotations. The last three rows are weakly supervised methods that use global labels. MC stand for Multi-Center, i.e., if the study involved images from multiple institutions, which increases complexity and requires good generalization performance.

In this paper, we compare weakly supervised models for the classification of the Gleason score, directly from H&E whole slide images of prostatectomies. The training of the models uses the global diagnostic label from the pathology reports as the only source of supervision. The trained models explore strategies based on a combination of DL features with morphological features extracted from automatically segmented nuclei, data augmentation, and depth of machine learning models. For the first time, we set robust baselines using DL for the task of fine-grained Gleason grading towards a clinically-usable computer-assisted diagnosis system with limited data and annotations. Finally, we also provide insights for better data-augmentation and aggregation strategies in this challenging task.

The data set consists of 341 cases of prostatectomies WSIs from the public resource of The Cancer Genome Atlas repository of prostate adenocarcinoma (TCGA-PRAD)[¶]. We pair each WSI with its corresponding GS label from the provided pathology reports. Due to the massive pixel size of a WSI, the features extracted and learned come from small regions in the image. We ensure that these regions (patches) are large enough to capture gland structures. The patch extraction is performed only in the tissue regions, using the Blue-Ratio (BR) mapping described in Chang et al.[18] BR mapping restrains areas without nuclei such as those containing fat, connective tissue, or background. Regions of 500 × 500 pixels are computed at a 20× apparent magnification (0.5 microns per pixel). The central 224 × 224 pixels of these regions are extracted and are the input patches for training all the DL networks. The number of WSIs used and the number of extracted patches are in Table 2, while two example WSIs with a subset of the locations of their patches are displayed in Figure 1.

The alterations in the nuclei features, such as nuclei size, shape, texture, and spatial architecture, reflect the complex molecular-level changes that occur during the formation of cancer and are a hallmark of PCa.[8] Initially, we evaluate the classification performance of hand-crafted morphological nuclei features. Nuclei features are computed from automatically segmented nuclei from each patch. The mask-RCNN model,[19] a multi-instance based DL segmentation technique, is used to segment each nucleus. The Mask-RCNN segmenter is fine-tuned on a separate data set[‖] with annotated nuclei of the prostate. After each nucleus mask instance is computed, the

[¶] https://portal.gdc.cancer.gov/projects/TCGA-PRAD Retrieved 1st of January, 2020

[‖] Available at https://nucleisegmentationbenchmark.weebly.com/: Retrieved 1st of January,2020 following features are extracted: area, diameter of a circle with the same area as the nucleus, ratio of pixels inside the nucleus to pixels in the total bounding box enclosing it, major and minor axis lengths, area of the bounding box enclosing the nucleus, perimeter, eccentricity, convex area, solidity, and orientation. The median feature value, for all nuclei in the patch, of the 11 features, is used as an 11-dimensional feature vector for each patch. An additional feature is added to the feature vector: the number of nuclei in each patch, resulting in a final 12-dimensional feature vector of hand-crafted features. Example patch segmentations are shown in Figure 2.

In the experimental evaluation using CNN models, we first evaluate the impact on the performance of the CNN architecture. Similar to what happens in many natural image classification tasks using deep learning models, the choice of the architecture plays an essential role in the performance. In our case, the main drivers of performance change were the capacity of the model (number of trainable parameters) and the connection pattern, as seen in the results in Section 3. The detail of the evaluated architectures are as follows:

- **Shallow CNN (S-CNN)**: A CNN architecture with four layers. Each of the layers consists of 32 3× three convolutional kernels, with batch normalization, ReLU activation, dropout of 0.25, and max-pooling of a 2×2 neighborhood. The last layer is a fully–connected dense layer with a softmax activation to predict the five GS classes. The total number of trainable parameters is 3.7 million. The learning rate was set to 0.001.
- **VGG-19**: This corresponds to the widely known VGG CNN[20] architecture with 19 layers and with a total of 143.6 million parameters. The best learning rate found for this architecture was 0.001.
- **DenseNet-121 (DenseNet)**: The 121-layer variation of the DenseNet architecture,[21] with 8 million parameters was used. The learning rate was set to 0.0001.

The Adam optimizer was used for all the CNN architectures and each learning rate was explored in the validation partition using the values in the set $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$.

Second, we explored whether complementary information was obtained by combining morphological features of nuclei. In this case, the answer was negative in our test set, probably due to a large number of irrelevant nuclei segmented and the mix of different cell types.

2.4 Implementation details

All DL models were trained on a Titan Xp GPU using the Keras framework with a Tensorflow backend. The DenseNet and VGG19 models were initialized with ImageNet pre-trained weights. Training the architectures from scratch was also tested, but it resulted in a small decrease in performance, and the convergence was slower. Because the convolution operation is not rotation invariant and the patterns that appear in each GS can be present at any arbitrary rotation, for all the architectures, augmented patches were used by flipping among the x and y axes and rotating by 90, 180 and 270 degrees. To increase the stain variability in our dataset, a stain-augmentation method was employed using the Vahadane stain augmentor from the StainTools library[**]. Finally, a further performance increase was obtained using augmentations weighted by the number of patches in each class (CWDA), ensuring that only underrepresented classes get the most of the augmentations without risk of overfitting on the already populated classes. In CWDA, the number of augmentations applied to a patch in a batch is inversely proportional to the number of patches of its class.

3. EXPERIMENTAL RESULTS

The model performance is measured as the inter-rater agreement. The raters can be either the pathologist that annotated the dataset or the prediction model. A performance measure that is usually used[7,16,22] is Cohen's kappa that is defined as:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}, w_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

[**] https://github.com/Peter554/StainTools: Retrieved 1st of July, 2019

Table 3: Performance for the DL models evaluated.

Model	Kappa score	F1	Accuracy
S-CNN + MV	0.3326	0.5228	0.52941
Morph. + MLP	0.1904	0.3444	0.4509
VGG19 + MV	0.3714	0.5044	0.5686
Densenet + MV	0.4066	0.5552	0.5889
Densenet + MV + CWDA	**0.4414**	**0.5872**	**0.6078**

Table 4: Confusion matrix for the TCGA test set using the DenseNet model with class-wise data augmentation.

	GS6	GS7	GS8	GS9	GS10
GS6	2	1	1	2	0
GS7	0	14	1	2	0
GS8	0	0	3	7	1
GS9	0	0	4	12	0
GS10	0	0	0	1	0

Where i, j are the ordered scores, $N = 5$ is the total number of Gleason scores. $O_{i,j}$, is the number of images that were classified with a score i by the first rater and j by the second. $E_{i,j}$ denotes the expected number of images receiving rating i by the first expert and rating j by the second. The quadratic term $w_{i,j}$ penalizes the ratings that are not close. When the predicted Gleason score is far from the ground-truth class, $w_{i,j}$ gets closer to 1.

For each model, we aggregate the individual patch predictions and obtain a global GS label for the test WSI following a majority voting scheme (MV). Formally, a model f, outputs the set of probabilities: $f_k(x) = \{p(y = k|x)\}$ for $k \in \{1, 2, 3, 4, 5\}$. For obtaining the final label for the WSI I, a summation over its P patch predictions is done and then the class that has the largest value is selected: $\text{GS}(I) = \arg\max_k \{\sum_{j=1}^{P} f_k(x_j))\}$. We report the performance of each model in Table 3. From the results, it is clear that the choice of the CNN architecture has an important impact on the performance. Particularly, the use of pre-trained deep architectures is beneficial to the performance of the model with respect to the randomly–initialized shallow CNN architecture. Other efforts to improve the CNN models by combining hand-crafted and CNN features were also tested but did not improve the results significantly. Finally, the best result ($\kappa = 0.44$) was obtained using the class-wise data augmented DenseNet architecture.

In Table 3 the performance of all approaches is displayed.

4. DISCUSSION

We systematically evaluated the performance obtained with several machine learning strategies in the challenging task of Gleason grading, relying only on weakly supervised WSIs, i.e. with only global labels and no manual region annotation. We thoroughly assessed both handcrafted and deep learning models with a heterogeneous and large data set of radical prostatectomies, obtained from several health centers with different scanners. The best model turned out to be a DenseNet architecture. Although the best results are still far from the inter-pathologist agreement of $\kappa \approx 0.7$, we identified salient strategies and drawbacks from the tested approaches. The best strategy achieved a $\kappa = 0.44$ using a densely connected CNN and an strategy to overcome class-imbalance. Even though we aimed to have a balanced data set for both training and validating the proposed algorithms, some of the Gleason scores (e.g. GS6 and GS7) had fewer examples than the other scores. This resulted in a poorer classification performance overall with the evaluated methods.

However, our experiments open the possibility for a further benchmark of designs and optimizations of DL models. We also foresee several potential improvements: combining strongly and weakly supervised models and also training with larger data sets. As recent studies show, this can improve the performance of weakly supervised models, even though in this example, only a straightforward two-class classification had good performance.[9]

5. CONCLUSIONS

Deep learning architectures trained only with weak labels and a medium-sized data set of WSIs (N=341) achieve a moderate agreement with a pathologist in the task of Gleason grading. As shown recently in other studies,[23] the role of data augmentation and patches with more context (multi-scale) are key to obtain a good performance of the models. In this work, we show that there is still a gap in performance from weakly trained models to strongly supervised ones when using a moderate amount of images. In future work, we plan to examine what is the minimum amount of weakly supervised data needed to reach a performance plateau minimizing time spent on unnecessary data set labeling for Gleason score classification.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825292 (ExaMode, htttp://www.examode.eu/).

REFERENCES

[1] Trpkov, K., "Contemporary gleason grading system," in [*Genitourinary Pathology*], Magi-Galluzzi, C. and Przybycin, C. G., eds., 13–32, Springer (2015).

[2] Fraggetta, F., "Clinical-grade computational pathology: Alea iacta est," *Journal of Pathology Informatics* **10** (2019).

[3] Pawlowski, N., Bhooshan, S., Ballas, N., Ciompi, F., Glocker, B., and Drozdzal, M., "Needles in haystacks: On classifying tiny objects in large images," *arXiv preprint arXiv:1908.06037* **1908** (2019).

[4] Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-Van De Kaa, C., Bult, P., Van Ginneken, B., and Van Der Laak, J., "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific reports* **6**, 26286 (2016).

[5] Doyle, S., Madabhushi, A., Feldman, M., and Tomaszeweski, J., "A boosting cascade for automated detection of prostate cancer from digitized histology," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 504–511, Springer (2006).

[6] Jimenez-del Toro, O., Atzori, M., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., Rönnquist, P., and Müller, H., "Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score," in [*Medical Imaging 2017: Digital Pathology*], **10140**, 101400O, International Society for Optics and Photonics (2017).

[7] Arvaniti, E., Fricker, K. S., Moret, M., Rupp, N. J., Hermanns, T., Fankhauser, C., Wey, N., Wild, P. J., Rueschoff, J. H., and Claassen, M., "Automated gleason grading of prostate cancer tissue microarrays via deep learning," *Scientific Reports* **8** (2018).

[8] Carleton, N. M., Lee, G., Madabhushi, A., and Veltri, R. W., "Advances in the computational and molecular understanding of the prostate cancer cell nucleus," *Journal of Cellular Biochemistry* **119 (9)** (2018).

[9] Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Silva, V. W. K., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., and Fuchs, T. J., "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature medicine* **25 (8)**, 8 (2019).

[10] Parwani, A. V. et al., "Commentary: Automated diagnosis and gleason grading of prostate cancer–are artificial intelligence systems ready for prime time?," *Journal of Pathology Informatics* **10**(1), 41 (2019).

[11] Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N. N., Tomaszewski, J., González, F. A., and Madabhushi, A., "Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent," *Scientific reports* **7**, 46450 (2017).

[12] Jimenez-del Toro, O., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., Rousson, M., Müller, H., and Atzori, M., "Analysis of histopathology images: From traditional machine learning to deep learning," in [*Biomedical Texture Analysis*], Depeursinge, A., Al-Kadi, O. S., and Mitchell, J., eds., 281–314, Elsevier (2018).

[13] 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 Gleason Grading of Prostate Biopsies using Deep Learning,” *The Lancet Oncology* (Jan 2020).

[14] 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).

[15] Burlutskiy, N., Pinchaud, N., Gu, F., Hägg, D., Andersson, M., Björk, L., Eurén, K., Svensson, C., Wilén, L. K., and Hedlund, M., “Segmenting potentially cancerous areas in prostate biopsies using semi-automatically annotated data,” in [*Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*], Cardoso, M. J., Feragen, A., Glocker, B., Konukoglu, E., Oguz, I., Unal, G., and Vercauteren, T., eds., *Proceedings of Machine Learning Research* **102**, 92–108, PMLR, London, United Kingdom (08–10 Jul 2019).

[16] Ström, P., Kartasalo, K., Olsson, H., Solorzano, L., Delahunt, B., Berney, D. M., Bostwick, D. G., Evans, A. J., Grignon, D. J., Humphrey, P. A., Iczkowski, K. A., Kench, J. G., Kristiansen, G., van der Kwast, T. H., Leite, K. R. M., McKenney, J. K., Oxley, J., Pan, C.-C., Samaratunga, H., Srigley, J. R., Takahashi, H., Tsuzuki, T., Varma, M., Zhou, M., Lindberg, J., Lindskog, C., Ruusuvuori, P., Whlby, C., Grnberg, H., Rantalainen, M., Egevad, L., and Eklund, M., “Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study,” *The Lancet Oncology* (2020).

[17] Jimenez-del Toro, O., Atzori, M., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., Rönnquist, P., and Müller, H., “Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score,” in [*Medical Imaging 2017: Digital Pathology*], **10140**, 101400O, International Society for Optics and Photonics (2017).

[18] Chang, H., Loss, L. A., and Parvin, B., “Nuclear segmentation in h&e sections via multi-reference graph cut (mrgc),” in [*International symposium biomedical imaging*], (2012).

[19] He, K., Gkioxari, G., Dollár, P., and Girshick, R., “Mask r-cnn,” in [*Computer Vision (ICCV), 2017 IEEE International Conference on*], 2980–2988, IEEE (2017).

[20] Simonyan, K. and Zisserman, A., “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556* (2014).

[21] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q., “Densely connected convolutional networks.,” in [*CVPR*], *1*(2), 3 (2017).

[22] Arvaniti, E. and Claassen, M., “Coupling weak and strong supervision for classification of prostate cancer histopathology images,” *Medical Imaging meets NeurIPS Workshop* (2018).

[23] Karimi, D., Nir, G., Fazli, L., Black, P. C., Goldenberg, L., and Salcudean, S. E., “Deep learning-based gleason grading of prostate cancer from histopathology images-role of multiscale decision aggregation and data augmentation,” *IEEE journal of biomedical and health informatics* (2019).