




Article

Classifying Malignancy in Prostate Glandular Structures from Biopsy Scans with Deep Learning

Ryan Fogarty ^{1,2} , Dmitry Goldgof ² , Lawrence Hall ², Alex Lopez ³, Joseph Johnson ⁴, Manoj Gadara ^{5,6}, Radka Stoyanova ⁷, Sanoj Punnen ⁸, Alan Pollack ⁷, Julio Pow-Sang ⁹ and Yoganand Balagurunathan ^{1,*} 

- ¹ Department of Machine Learning, H. Lee Moffitt Cancer Center, Tampa, FL 33612, USA
² Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33620, USA
³ Tissue Core Facility, H. Lee Moffitt Cancer Center, Tampa, FL 33612, USA
⁴ Analytic Microscopy Core Facility, H. Lee Moffitt Cancer Center, Tampa, FL 33612, USA
⁵ Anatomic Pathology Division, H. Lee Moffitt Cancer Center, Tampa, FL 33612, USA
⁶ Quest Diagnostics, Tampa, FL 33612, USA
⁷ Department of Radiation Oncology, University of Miami Miller School of Medicine, Miami, FL 33136, USA
⁸ Desai Sethi Urology Institute, University of Miami Miller School of Medicine, Miami, FL 33136, USA
⁹ Genitourinary Cancers, H. Lee Moffitt Cancer Center, Tampa, FL 33612, USA
* Correspondence: yoganand.balagurunathan@moffitt.org

Simple Summary: In recent years, the prostate cancer histopathological description proposed by Gleason has emerged as a universal standard used for disease diagnosis and progression. Recently, a grading scheme on a point scale is based on Gleason patterns. Current scores are highly dependent on the expert urinary histopathologist and show a high level of variability among experts. To aid the clinician, we have developed deep learning models that provide a decision aid in identifying the primary cancer grade (dominant Gleason pattern).

Abstract: Histopathological classification in prostate cancer remains a challenge with high dependence on the expert practitioner. We develop a deep learning (DL) model to identify the most prominent Gleason pattern in a highly curated data cohort and validate it on an independent dataset. The histology images are partitioned in tiles (14,509) and are curated by an expert to identify individual glandular structures with assigned primary Gleason pattern grades. We use transfer learning and fine-tuning approaches to compare several deep neural network architectures that are trained on a corpus of camera images (ImageNet) and tuned with histology examples to be context appropriate for histopathological discrimination with small samples. In our study, the best DL network is able to discriminate cancer grade (GS3/4) from benign with an accuracy of 91%, F₁-score of 0.91 and AUC 0.96 in a baseline test (52 patients), while the cancer grade discrimination of the GS3 from GS4 had an accuracy of 68% and AUC of 0.71 (40 patients).

Keywords: prostate; Gleason cancer grading; pathology; uropathology; whole-slide image; ISUP grade; Gleason score; deep learning; convolutional neural network; transfer learning



Citation: Fogarty, R.; Goldgof, D.; Hall, L.; Lopez, A.; Johnson, J.; Gadara, M.; Stoyanova, R.; Punnen, S.; Pollack, A.; Pow-Sang, J.; et al. Classifying Malignancy in Prostate Glandular Structures from Biopsy Scans with Deep Learning. *Cancers* **2023**, *15*, 2335. <https://doi.org/10.3390/cancers15082335>

Academic Editor: Kentaro Inamura

Received: 13 February 2023

Revised: 7 April 2023

Accepted: 12 April 2023

Published: 17 April 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Prostate cancer is a neoplasm in the prostate gland, most often epithelial in origin, with over 95% of adenocarcinoma subtype. The neoplasms are classified from different grades of aggressiveness using Gleason patterns 1–5, then combined into a Gleason score (dominant + subdominant Gleason patterns) detailed below [1,2]. Standard diagnosis requires a fine needle biopsy of the gland where the histology is assessed from hematoxylin and eosin (H&E)-stained tissue sections by an expert genitourinary pathologist [1]. The prostate adenocarcinomas histopathology displays an abnormal architectural glandular pattern with a very high degree of benign epithelial–stromal relationships. Most widely used Gleason scoring patterns were adopted by the International Society for Urological

Pathology (ISUP) proposed in 2014, later adopted by the World Health Organization (WHO) in 2016 [3]. The patterns are described by a modified Gleason grading that shows five distinct patterns with direct relationships to cancer invasiveness, which were conceived purely based on clinical outcomes [4]. The pattern spans from single, separated well-formed glands in Gleason pattern 1, ISUP grade group 1 (GS 3 + 3), to stromal infiltration in Gleason patterns 4 to 5 (ISUP grade groups 2 to 5) [4]. The cancer types have relied more on the epithelial–stromal architecture than any other clinical grade based classification to describe disease aggressiveness with direct relation to the clinical outcome [5]. Use of the ISUP scoring scheme has helped to reduce the scoring range, but this expert-based standard has significant intra- and inter-variability among genitourinary pathologists and clinical centers and results in care differences among patients [6–8]. In a recent report, concordance rates between two observers for primary and secondary Gleason patterns were 63.96% ($\kappa = 0.34$) and 63.45% ($\kappa = 0.37$), respectively, while Gleason grades was at 57.9% ($\kappa = 0.39$) [9]. This does not get better with diagnoses around the globe; concordance ranged from 0.44 to 0.49, while urological pathologists showed moderate improvement to 0.68 [10].

Development in the last decade has seen promise in using machine learning (ML) and deep learning (DL) tools to improve diagnostic variability and provide a decision support system (DSS) to aid the pathologist and improve quality of care or treatment response [11–16]. A potential implementation of a DSS is shown in Figure 1; this study concentrates on the decision classifier. A complete implementation of a DSS will include many preprocessing steps such as region extraction, which were supported through preprocessing tiles as detailed in Section 2. Feature extraction and composition through explicit means (not implicitly derived through CNN layers) such as radiomic features and feature composition is a fertile area for improvement [17,18] but not a focus of this study. Our minimalist approach is covered in the proceeding. Recent deep learning techniques with convolutional neural networks (CNN) have shown tremendous promise in extracting non-human visible salient features from diagnostic pathological images [19–24]. Our results demonstrate that these diagnostic clues are available in local glandular structures or small patches of prostate biopsy whole-slide images (WSI). Due to the subtle nature of the features in histopathological images coupled with limited sample sizes, generalization of the model continues to be a challenge across medical centers and sources [25].

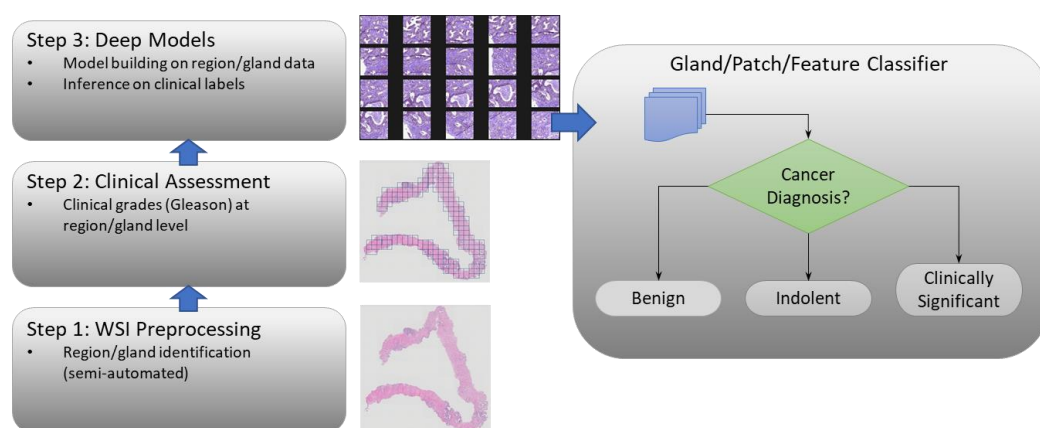


Figure 1. Prostate Cancer Decision Support System.

Recently, others have shown use of deep network’s ability to identify cancer grades, progression and outcome using medical imaging (CT/MRI) utilizing various data augmentation methods and transfer learning (TL) approaches [14,15,26–29]. Improving classification by transfer learning from large data sources such as ImageNet has been a standard approach for many years for many medical imaging modalities [30]. However, more recent studies have shown that TL may be limited to learning patterns in small sample sets [31,32].

Our research goal is to develop a deep learning model to discriminate Gleason grades in whole-slide H&E images. In this study, our focus is to accurately identify aggressive primary Gleason patterns in annotated image patches. To enable proper network model training, the WSIs of prostate needle biopsies were manually annotated following the pathologist-assigned Gleason pattern scored at the gland level. We focused on evaluating several deep learning model architectures based on CNNs (EfficientNet [33], ResNet [34]), and the Visual Geometry Group (VGG-16 and VGG-19 [35]) networks. Each has advantages, but in this domain, VGGs showed the best performance by a significant margin. Each of these networks were trained using transfer learning, with a degree of fine tuning on feature weights to adapt the network to the histopathological classification as previously used in other studies [36].

There have been few attempts in the past to discern the gland patterns. This study, to our knowledge, is one of the first to create a large cohort of manual gland level scoring (over 14k glands, see Table 1) and use deep learning (DL)-based models to discriminate the Gleason patterns. We believe these results form a baseline comparison for the primary patterns at this level of granularity. In comparison, most other studies have used DL-based classification algorithms to discern Gleason grades or scores (primary + secondary) at the whole-slide level, containing multiple glands that represent primary and secondary patterns, most often in an unbalanced proportion [16,21]. Ström et al. created an ensemble of 30 Inception V3 nets to score individual patches as benign or a Gleason pattern and pass the results through a boosted decision tree to inform an overall Gleason (ISUP) grade [22]. Pinckaers et al. demonstrated a novel streaming CNN (based on ResNet-34) to process entire biopsy scans' centered, cropped patches to detect malignancy status (cancer from benign) at the whole-slide level [21]; this method was compared against two baseline approaches from Companella et al. and Bulten et al. Companella et al. applied multi-instance learning (MIL) with a recurrent neural network (RNN) to predict a combined score from patches classified with a ResNet-34 [37]. Bulten et al. used an extended U-Net to predict Gleason patterns at the pixel level and to determine the proportions of malignancy to inform an overall grade decision [19]. It is to be noted that most prior work has focused on cancer status discrimination at the slide level assessing multiple pattern (primary and secondary) scores. In contrast, our proposed work focuses on a discriminating Gleason pattern (primary) at the patch level, limiting the reader variability. Hence, our performance can be qualitatively compared with prior research. In this work, we demonstrate the use of a VGG CNN model for classifying small patches (or individual, variably sized glands) in a WSI obtained from a core needle prostate biopsy. Additionally, we contrast the model performance on two diverse datasets independently obtained.

Table 1. University of Miami/Moffitt Cancer Center Cohort.

	Total	Benign	GS3	GS4
Subjects	52	23	38	32
Whole-Slide Images	150	72	72	60
Labeled Glands	14509	6882	5143	2484

In the following Section 2, the studied datasets are discussed, as well as the techniques used in this study, details on the models (DL architectures), training and tuning techniques and data processing. In Section 3, we show the results for our findings on Gleason pattern discrimination (benign versus cancer (all grades); GS3 versus GS4) at the prostate gland level. Section 4 provides further discussion of the results and use of a decision support system to improve clinical diagnosis. Finally, in Section 5, we summarize our findings.

2. Materials and Methods

2.1. Data Cohorts

Our study used two retrospectively curated data cohorts of prostate cancer patients' biopsies obtained at two different cancer centers. Patients' data were retrospectively obtained using the respective clinical centers' research protocols. The data were de-identified for research use, obtained after an Institutional Review Board (IRB) review of our research project, and the patients waived their informed consent rights for retrospective research usage. The first cohort is a dataset obtained from the University of Miami (UM) and curated at the H Lee Moffitt Cancer Center (MCC) and will be referred to as the UM/MCC data cohort. The second cohort was derived from the Kaggle PANDA histopathology open challenge; the de-identified patient data with Gleason grades were provided by Radboud University, Nijmegen, the Netherlands as part of their effort to promote open science—available on the National Institute of Health's Cancer Imaging Archive website (<https://www.cancerimagingarchive.net/>) (accessed 24 February 2023). The patient data from both sources were completely anonymized with no treatment details or outcomes provided.

2.1.1. Gland Level Patient Data Cohort

We obtained digitized whole-slide histopathology with 20× magnification scanned on an Olympus VS120 scanner (Olympus Life Sciences, Inc., Tokyo, Japan). These images were imported into the Visiopharm[®] digital pathology software, and gland regions were manually delineated and annotated by our research urinary pathologists (AL and MG) with over 15 years and 9 years of clinical experience in prostate histopathology scoring, respectively. The glands were scored on benign, GS3, GS4 or GS5 levels.

The study used 52 patients, 150 WSIs and 14,509 glands; the data will be referenced as UM/MCC data (see Table 1). An independent cohort was assembled from the Kaggle PANDA's challenge used as a training/validation cohort (at a 90/10 ratio), with 24,800 patches with glands having the same primary pattern (see Table 2).

Table 2. Kaggle PANDA Radboud Synthesized Cohort.

	Total	Benign	GS3	GS4	GS5
Biopsy scans	1240	310	310	310	310
Patches	24,800	6200	6200	6200	6200

To train or test on WSIs, (relatively) small image patches were created for each of the labeled glands (dominant Gleason pattern). These specific glands were identified and converted from vector files in the Visiopharm[®] MLD format into segmentation masks. These masks were then used to extract individual image patches using several bounding box techniques from the WSIs, including fixed-size bounding boxes, squared bounding boxes and tight bounding boxes. The latter two techniques resulted in image patches of various sizes, which required subsequent resizing or resampling techniques to be used by the CNN ML (details deferred to Section 2.2.1).

Figure 2 shows several patches with tight bounding boxes for the classification levels benign, GS3, or GS4, respectively. Note that the images of glands shown below are within 15% of 120 × 120, 240 × 240, 360 × 360 and 720 × 720 pixels. The aspect ratios of the sample set's width and height has much higher variability than the samples shown.

Due to limited dataset size, we estimated the discriminators' performance using the Monte Carlo cross-validation (MCCV) technique and reported the ensemble statistics [38], which will be shown in Section 3. Due to extensive time required for network training on the computational resource, retraining of the network was avoided and re-sampling of the outcome was resorted to.

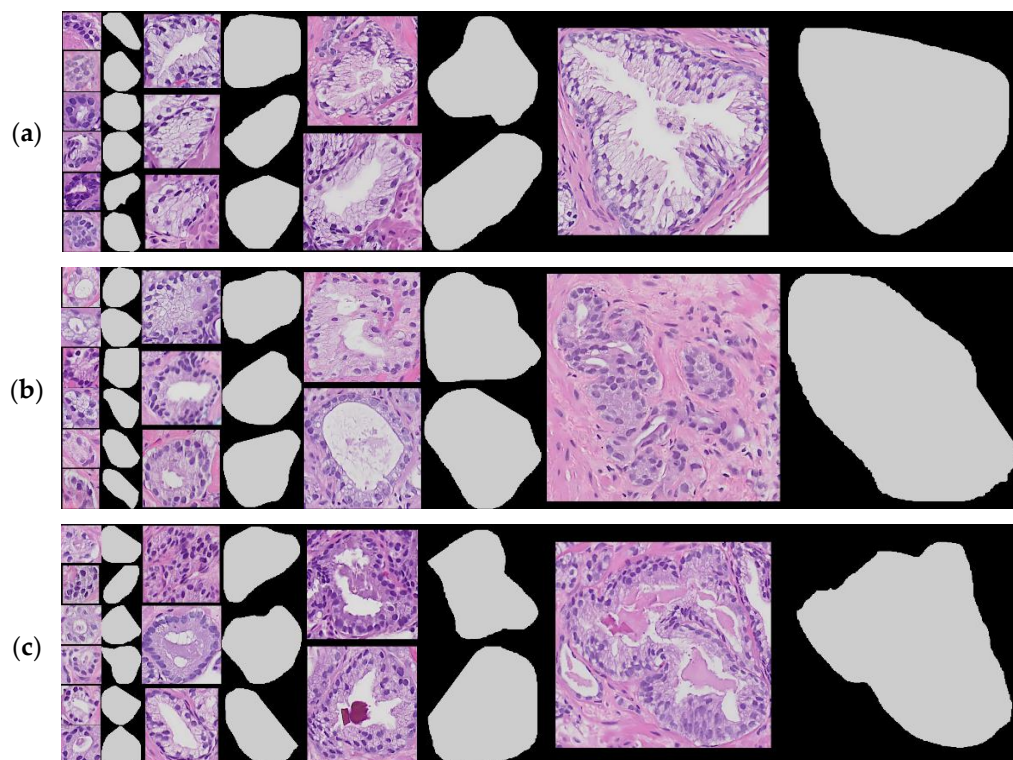


Figure 2. Sample selection of UM/MCC gland-level patches using the tight bounding-box technique paired with corresponding mask layer (on the right): (a) samples of benign, (b) samples of GS3, and (c) samples of GS4.

2.1.2. PANDA Radboud Data Cohort

We curated the large open-source prostate pathology data cohort shared as a part of the prostate cancer grade assessment (PANDA) challenge organized through the Kaggle open challenge platform [16]. The PANDA training set had expert-provided annotations; it was composed of two separate patient sources: the Karolinska Institute and Radboud University. In our study, we used images from the Radboud collection because of the fidelity of the labeled segmentation masks that overlapped with the cases in the UM/MCC dataset.

The PANDA Radboud dataset, scanned from needle biopsy slides, was synthesized into a set of data patches from the Kaggle-provided data. The partial WSI images were downsampled by $2\times$ then Otsu binarized to isolate foreground and background. Fixed-sized patches were extracted from the foreground (biopsy image) area by sliding a window (400×400) over the label mask (where foreground was identified) and accepting patch areas that contained an appropriate density of segmented (labeled) data for some target Gleason score. The degree of window overlap p_w was adjusted until each WSI sample produced at least 20 patches (overlap p_w starting from 0.5 and adjusted as high as 0.8 if enough samples were not generated per image). Several thresholds were tuned to generate an approximately equal set of patches for each Gleason pattern level by testing each label mask pixel (x_i) for the ratio of Gleason–label mask p_α (nominally ≥ 0.1) that was not identified as background or stroma and the purity of label p_β (0.95) at the targeted Gleason level. The cut points for the quality filters were heuristically fixed at these levels.

$$\begin{aligned}
 p_\alpha &= 0.1 \\
 p_\beta &= 0.95 \\
 N &= \text{width} * \text{height} \\
 l_{\text{epithelium}} &= 2 \\
 M &= \sum_{i=1}^N x_i \geq l_{\text{epithelium}} \\
 l_{\text{target}} &= GS; GS \in \{3, 4, \text{or } 5\} \\
 T &= \sum_{i=1}^N x_i \equiv l_{\text{target}} \\
 \text{accept} &= \frac{M}{N} \geq p_\alpha \ \& \ \frac{T}{N} \geq p_\beta
 \end{aligned}$$

Once these patches were generated for each Radboud image file, the patches were rank sorted by the proportion of epithelial/malignant label mask coverage M/N , and the 20 patches with highest ratio of label were added to a synthesized set of patches for Gleason levels benign, GS3, GS4 and GS5. If, after adjusting the sliding window overlap as high as $p_\omega = 0.8$, an image was still unable to produce 20 patches, it was excluded from the training set. The PANDA Radboud dataset may include multiple glands per patch, which is different than the UM/MCC data but is sufficient as a pretraining dataset for distinguishing Gleason patterns. In Figure 3, patches and their corresponding Radboud masks are shown for Gleason levels benign, GS3, GS4 and GS5.

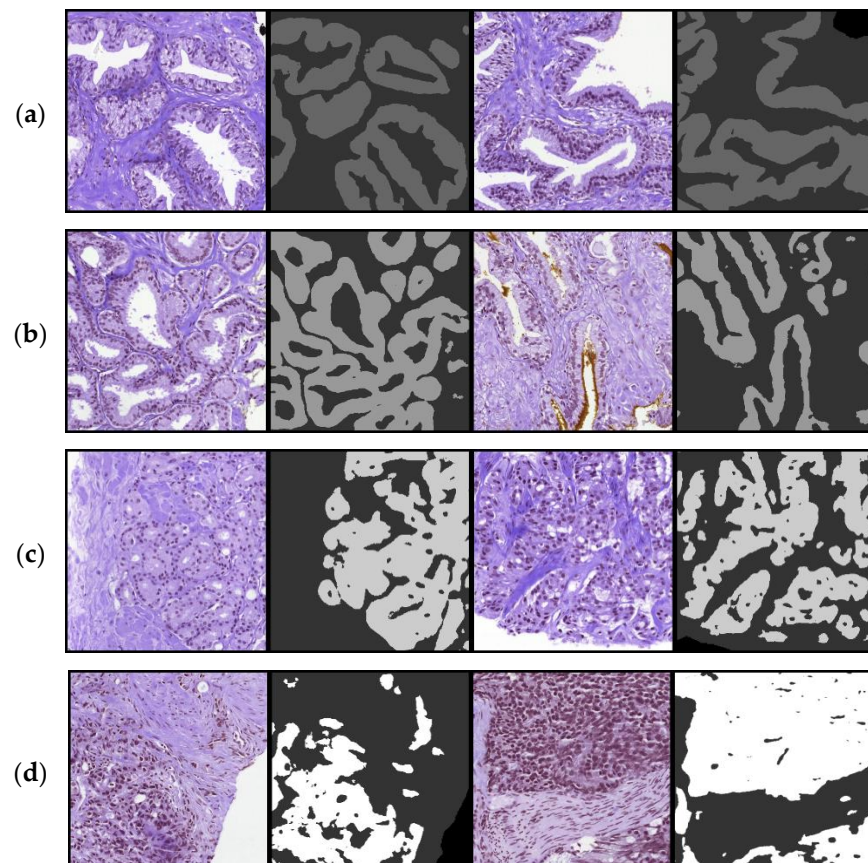


Figure 3. Sample selection of PANDA Radboud patches using the fixed-box technique paired alongside corresponding mask layer (on the right): (a) two samples of benign, (b) two samples of GS3, (c) two samples of GS4 and (d) two samples of GS5.

In generating this dataset, we made the training and validation/test cohorts as uniform as possible for the Gleason pattern; patches for benign were only pulled from images with clinical diagnosis benign, ISUP grade 1 (3 + 3) for GS3 pattern and grade 4 (4 + 4) for

GS4 pattern. For Gleason pattern 5, very few images were graded as 5 + 5; therefore, patches were drawn from 4 + 5, 5 + 4 and 5 + 5 samples. The patch quality metrics, epithelial/glandular density p_α (nominally ≥ 0.1) and purity of label $p_\beta = 0.95$ ensured that each patch was appropriate for the primary Gleason grade. Note that this procedure was not required for the UM/MCC dataset since patches were extracted from each labeled gland. Table 2 shows the resultant dataset with 310 images and the corresponding 6200 patches per each Gleason level. Most of the clinical data such as subject identification was excluded from the Kaggle PANDA collection. The patient data was completely anonymized with no outcome or treatment data provided.

2.2. Image Preprocessing

2.2.1. Sample-Mix

A necessary step in preprocessing the data was to ensure the image patches were identical in size prior to being processed through the DL CNN. There have been many resizing techniques that were tried previously such as in [39,40], cropping to fixed size areas (or loose bounding boxes) around the areas of interest, such as the glands. In our study, we adopted a *sample-mix* approach, which was inspired by other image mixing techniques such as in [41–43]. The approach preserves the scale and the aspect ratio of textural features; see sample in Figure 4. The tiled approach allows smaller and larger images to be adjusted to the same size, preserving their original textural characteristics, immaterial of the gland size (small or large). It is possible to construct sample-mixed patches for any dimension and rank, where rank is the number of tiles sampled along each axis. The examples show a target dimension of 300×300 pixels and rank of three tiles (along the horizontal and vertical), requiring the sampled tiles to be 100×100 (or $1/9$ of the target size).

The sample-mix methodology, a resize strategy, was used when image data sources were variably sized and needed to be resized to train the network model. We used this strategy in the UM/MCC training cohort, where the extracted patches with tight bounding boxes were of different sizes. For images that were smaller in the horizontal or vertical direction than the sample tile (100×100), those samples were simply removed from the dataset as a preprocessing step. When data elimination is not desired, the algorithm can automatically generate a sample mix with smaller sample tiles.

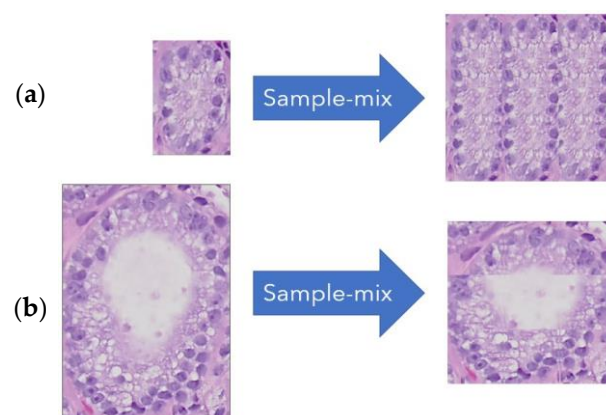


Figure 4. Sample-mix rescale examples: (a) $126 \times 188 \rightarrow 300 \times 300$, (b) $290 \times 417 \rightarrow 300 \times 300$.

2.2.2. Standardization

The technique of staining hematoxylin and eosin (H&E), respectively, provides pathologists with functional and morphological details at the cellular level. It is well-recognized that even after over a century of its usage, there are many variables such as the stain protocol, dye quality and dye age that are uncontrolled factors causing inter-laboratory variability resulting in visual differences in slide appearance (color and intensity) [44,45].

The staining of images across sources is often inconsistent, and varying degrees of chemical application may result in significantly different color saturation [46]. In Figure 5, three partial views of WSIs with clearly varied colors are shown; the first two are from the same UM/MCC cohort, and the third on the right is from PANDA Radboud data. It is possible that these shade differences between the cohorts dampens generalization of the DL models.

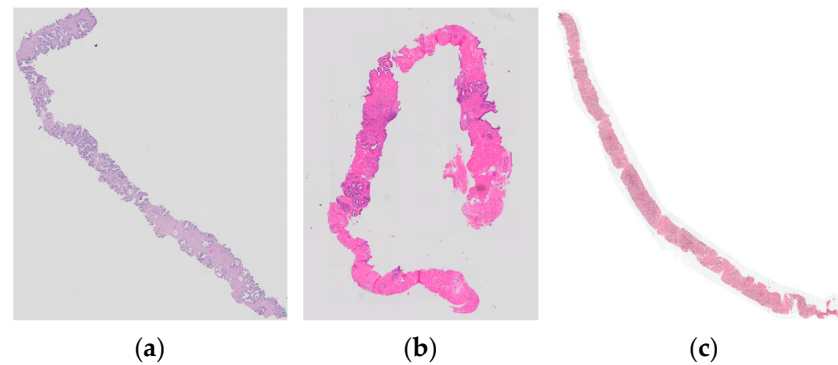


Figure 5. Examples to illustrate stain contrast differences among the samples in cohort: (a,b) UM/MCC dataset, (c) PANDA Radboud.

To mitigate these wide variations, the image patches were normalized using z-score standardization, where color channels are re-centered to zero mean and unit standard deviation. This technique is the most often used approach for training on large image datasets such as ImageNet [27]. As a result of applying transfer learning starting with ImageNet weights, standardization is a necessary preprocessing step to ensure proper feature extraction through the CNN layers. Unfortunately, this transformation step may not always improve a model's reproducibility in histopathology images. It has been reported that most H&E-stained image intensity follows a bi- or tri-modal distribution [47–49]. The standardization followed in most learning methods uses linear scaling that may not compensate for the distributional spread, which would be an additional source of alteration in the model training.

2.2.3. Eliminating Outliers

Outlier detection (OD) is an important step in maximizing performance of an ML algorithm [50]. To remove outliers in the dataset, we performed the random sampling and consensus (RANSAC) method as has been applied in regression problems [51]. In a classification problem, we remove samples that never or very seldom classify correctly after an initial training of the DL models. The technique is similar to a histogram-based OD in which outliers are removed based on a threshold rule, classically as a distance away from the 25% and 75% quartile, normalized by inter-quartile range or IQR. The RANSAC technique requires multiple models to be trained on the data (as is a natural approach when performing CV experiments); inference is then performed on the entire sample set for each model, and a consensus of each sample's performance is determined. While performing the test, the samples that never classify correctly were removed from the training set. To preserve the integrity of the performance metrics, RANSACed outliers are removed from future training sets, but validation data are not altered.

In determining the outliers, the consensus scores are drawn by inferencing the training data with the models derived using the CV folds. As a result, the sample models will have seen the data many times, and hence, low consensus scores of no correct classification or one correct classification out of multiple models applied (e.g., 0/20 and 1/20) imply likely outliers. This technique was applied at least once for the binary DL classifiers shown in Section 3.

In Figure 6, the RANSAC consensus scores are shown as stacked histograms when training GS3 versus GS4 DL models. The colors in the vertical bars and the legend along the bottom of the chart represent different consensus scores in 20 models; thus, the worst outliers score 0/20 times, and the best performers score 20/20 times. The stacked histograms add up to the total number of subjects in our training set, and as can be seen, with each iteration, several low-performing samples are eliminated from the set. Additionally, this plot demonstrates how with each iteration, the remaining training samples see a gradual improvement in the consensus score. Other unsupervised or semi-supervised OD techniques are under consideration for future work [52,53].

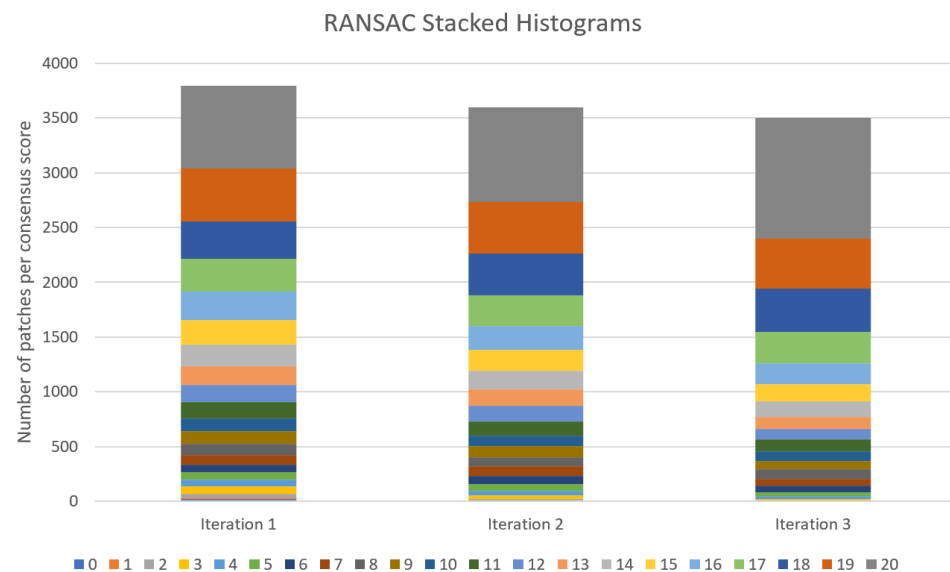


Figure 6. Three iterations of RANSAC histograms training GS3 vs. GS4 DL models.

2.2.4. Balancing Data

The GS3 versus GS4 classification experiment was largely unbalanced; in the UM/MCC cohort, the GS3 majority class was twice as large as the GS4 minority class. To ensure that the machine learners did not simply prefer the majority class, in all experiments, the data was balanced. To train DL networks, we used bootstrapping to dynamically balance both training and validation sets [54]. A custom TensorFlow iterator was created to ensure data was balanced on every batch training update. Performance statistics were estimated using the bootstrap technique.

2.3. Deep Learning

A convolutional neural network (CNN) utilizing transfer learning from very large datasets such as ImageNet shows promise for classification problems. Networks with lower inductive bias are expected to outperform CNN architectures as more and larger datasets become available (such as those from Kaggle PANDA), through knowledge distillation and improved architectures that optimize generalized learning [55–57].

Most DL models that are studied with CNN feature layers combine with a binary or multiclass dense classification layer. Common techniques such as dropout, pooling and batch normalization were used between layers to improve performance [58–60]. The fully connected classification layer has a 32-node layer to aggregate features from the CNN, followed by as many neurons as are required for the classification task (1 for binary classification, or 3+ for multi-classifier), as shown in Figure 7.

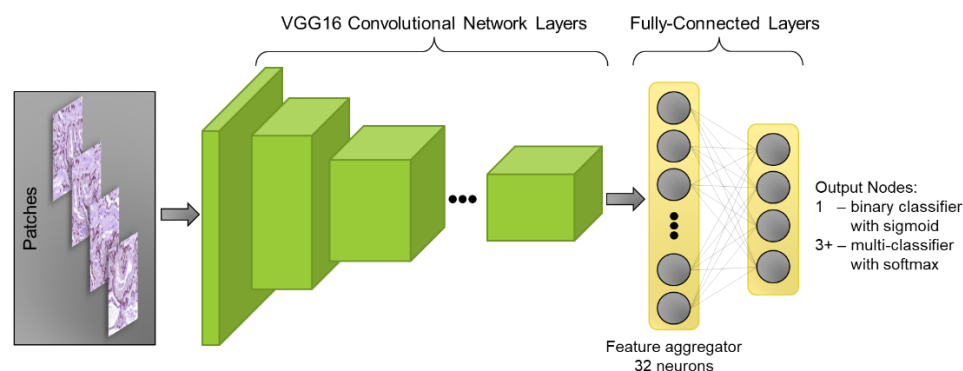


Figure 7. CNN feature layers, with dense classification layer.

2.3.1. Optimization Technique

Research has shown that adjusting the learning rate in a cyclic fashion can help to escape local minima and saddle points [61]. The learning rate may be “shocked” or annealed by jumping back to a maximum on a periodic basis [62]. Our optimization strategy leveraged the cosine annealing technique. In this technique, the learning rate is adjusted from a maximum rate to a lower rate (perhaps one or two orders of magnitude smaller) and updated at each batch (partial training of epoch). Cosine-annealed training was used in this study, cycled every 29 epochs in our experiments. Additional hyper parameter values are shown in the Supplementary Document Table S1.

2.3.2. Transfer Learning

Transfer learning is very effective at jumpstarting NN training, especially when data are limited. The UM/MCC cohort is relatively small; hence, we began training by initializing a VGG-16 network on ImageNet feature weights, a popular approach shortly following the original AlexNet [27,63]. We started by coarse tuning the fully connected (FC) classification layers with the UM/MCC data and followed by fine tuning the CNN feature layers and FC layers. We improved the results by first training on a larger Kaggle PANDA Radboud dataset (both a coarse tune followed by a fine tune to train the CNN feature weights), then fine tuning with data in our own UM/MCC cohort. The NN learner goes through four stages of learning, as shown in Figure 8. The technique of pretraining on one dataset and then tuning on another is a common and effective transfer-learning approach. A recent study corroborates this method when used for prostate pathology grading [64].

To create a more generalized model, we combined both our PANDA and UM/MCC datasets into one large training set and then trained our CNN to classify both sources. The combined dataset is also trained starting with ImageNet weights, first on the FC layers and eventually fine-tuning all weights in the FC and CNN layers.

2.4. Measuring Performance

Accurate measurement of ML performance poses a challenge when there are a limited number of subjects for model training. In the study, a randomized Monte Carlo cross-validation (MCCV) technique was used to estimate the network-based discriminators’ performance [38]. When applying MCCV, we performed a minimum of 20 folds, as recommended to ensure most data were tested since folds were sampled with replacements.

The following metrics were computed to evaluate deep networks’ classification performance: accuracy, sensitivity, specificity, precision, negative predicted value (NPV), F_1 -score and area under the receiver operating characteristic curve (AUC) [65]. Statistical metrics were computed using the Python scikit-learn metrics package (sklearn.metrics) for DL experiments. These metrics were calculated across the folds following the recommendations described by Forman and Scholz [66]. The performance metrics are provided with 95% confidence intervals determined by the bootstrap estimation procedure [67].

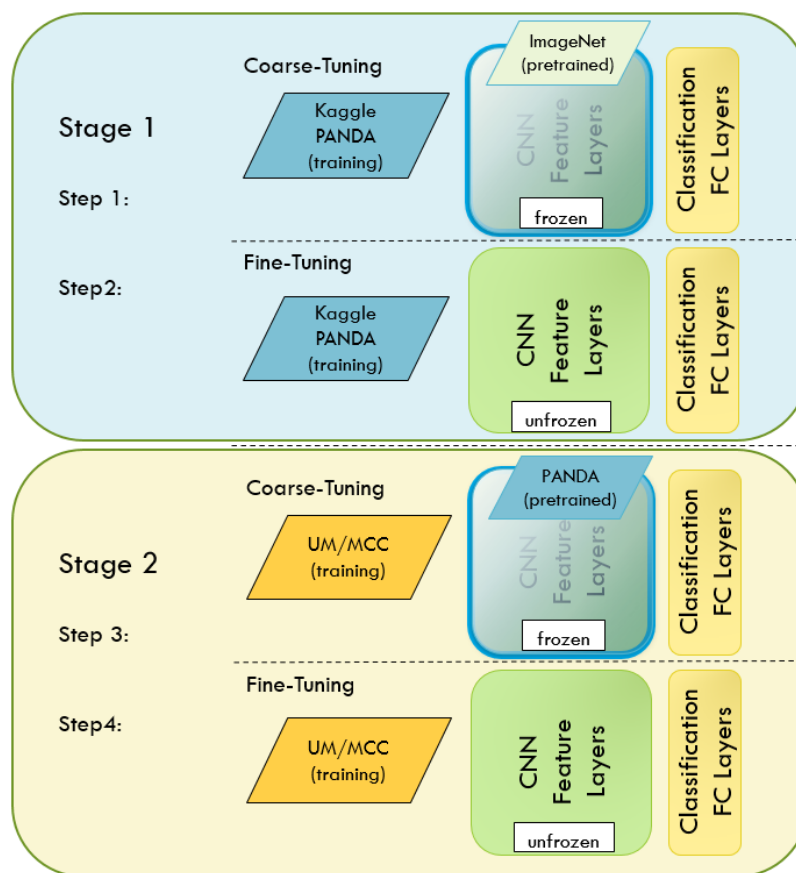


Figure 8. Four steps, transfer learning and fine tuning of NN.

2.5. Implementation Challenges

There were two main challenges for this study. The first major challenge was sample size and curation of a data cohort with gland-level labels in prostate histopathology. This work involved a clinical expert-driven manual gland scoring (semi-automatic) to create a pure cohort of about 14,000 labeled glands. The second challenge in this study was investigating the many state-of-the-art DL architectures, finding hyperparameters and tuning methods that resulted in optimized training with the diverse size of the gland-level patches. The computational resources required for model building in a timely manner posed a challenge.

3. Results

We used deep networks to perform binary classification tests to discriminate various grades of primary Gleason patterns at the glandular and small-tile level. These experiments compare benign versus GS3/4 (malignant) and finer grade levels, GS3 versus GS4 levels, as some may have clinical significance on the decision boundary of cancer progression and treatment.

We evaluated several types of deep networks for prostate histopathology classification, given the constraints of small sample sets. We found CNNs are well-studied with small sample dataset constraints, and multiple prior studies have shown stable performance with these constraints [68,69]. Several popular deep CNN architectures and their initial performances are detailed in the Supplementary Document Table S2. The VGG-16 and sample-mix technique proved to be the top performer and most practical, so all remaining tests include this combination. Table S3 in the Supplementary Document shows performance for several alternative resizing techniques.

Deep Network Performance

Results using the VGG16 CNN network trained on the PANDA Radboud, UM/MCC and combined dataset (PANDA + UM/MCC) follow. The models were built and trained in Python 3.8.10, Tensorflow/Keras 2.9.1 on the NVIDIA® DGX™/A100 platform. Model files and samples of Python code are available for download at <https://github.com/rfogarty/glandLevelGleasonClassification.git> (accessed on 13 February 2023).

We found the best DL network trained on PANDA's data (tile size of 400×400) showed exceptional results in discriminating cancer from benign with an AUC of 0.981 and AUC of 0.997 for discriminating cancer grades (GS3 vs. GS4), with other metrics shown in Table 3. The reported scores are cross-validation scores, and optimistic versus a holdout test set—the GS3 versus GS4 results are especially optimistic.

Table 3. PANDA Radboud classifier scores.

	Trained on PANDA Radboud	
	PANDA Radboud Benign vs. GS3/4/5	PANDA Radboud GS3 vs. GS4
Accuracy	0.941 (0.88, 0.98)	0.979 (0.95, 1.0)
Sensitivity	0.964 (0.92, 0.99)	0.980 (0.94, 1.0)
Specificity	0.920 (0.80, 0.98)	0.979(0.93, 1.0)
Precision	0.927 (0.83, 0.98)	0.979 (0.94, 1.0)
NPV	0.959 (0.90, 0.99)	0.980 (0.94, 1.0)
F ₁ -score	0.944 (0.88, 0.98)	0.980 (0.95, 1.0)
AUC	0.981 (0.93, 1.0)	0.997 (0.99, 1.0)

Table 4 summarizes performance of our VGG-16 DL architecture and establishes a baseline of performance for our UM/MCC dataset for the two binary problems studied. The networks were pretrained on ImageNet only (light blue columns) and pretrained on ImageNet followed by a pretraining on PANDA Radboud (darker blue columns). In this case, the benign versus GS3/4 binary classifier performed much better than the GS3 versus GS4, measuring better than 20% in almost all metrics. Although in both cases we scored better after a PANDA Radboud pretraining, the results of pretraining on PANDA only showed a marginal improvement. The scores for benign versus GS3/4 were measured using a conventional 10-fold CV, while the scores in GS3 versus GS4 classification were measured using a 20-fold Monte Carlo CV (because of very limited data size). Scores include 95% confidence intervals computed using the bootstrap method in parentheses.

Table 4. UM/MCC performance 1-stage versus 2-stage (ImageNet + PANDA) training.

	Trained on UM/MCC			
	Benign vs. GS3/4 (1-Stage ImageNet Transfer-Learning)	Benign vs. GS3/4 (2-Stage ImageNet Plus PANDA Transfer-Learn)	GS3 vs. GS4 (1-Stage ImageNet Transfer-Learning)	GS3 vs. GS4 (2-Stage ImageNet Plus PANDA Transfer-Learn)
Accuracy	0.901 (0.79, 0.98)	0.911 (0.81, 0.97)	0.669 (0.53, 0.84)	0.680 (0.54, 0.84)
Sensitivity	0.898 (0.75, 0.97)	0.897 (0.71, 0.97)	0.732 (0.37, 0.93)	0.753 (0.47, 0.90)
Specificity	0.898 (0.75, 0.97)	0.897 (0.71, 0.97)	0.606 (0.26, 0.87)	0.606 (0.19, 0.92)
Precision	0.912 (0.75, 1.0)	0.923 (0.76, 0.99)	0.660 (0.52, 0.86)	0.670 (0.52, 0.90)
NPV	0.912 (0.75, 1.0)	0.923 (0.76, 0.99)	0.699 (0.54, 0.86)	0.712 (0.55, 0.83)
F ₁ -score	0.903 (0.78, 0.98)	0.908 (0.77, 0.98)	0.686 (0.47, 0.83)	0.702 (0.51, 0.84)
AUC	0.955 (0.87, 0.99)	0.955 (0.87, 0.99)	0.706 (0.43, 0.90)	0.714 (0.44, 0.90)

Tests of cross-source generalization on our best PANDA-trained models and UM/MCC-trained models were poor, which demonstrates that the datasets are quite distinct. The best DL discriminator trained on PANDAs and tested on UM/MCC was able to differentiate cancer from benign (benign vs. GS3/4) with an AUC of 0.738. While the AUC drops to worse than random guessing (<0.5) for GS3 vs. GS4. In the alternate case, there was marginal performance for DL networks trained on UM/MCC and tested on PANDAs with an AUC of 0.522 for benign vs. GS3/4/5 and AUC of 0.692 for GS3 vs. GS4. Details of this experiment are shown in Table S4 of the Supplementary Document. As shown below, we will significantly improve these results by training on a combined PANDA + UM/MCC dataset.

Finally, in Table 5, results are summarized for two networks that were simultaneously trained on a combined PANDA Radboud plus UM/MCC dataset. The columns in light green on the left are derived from the network that classifies benign versus malignancy (GS3/4), while the columns in dark green show results from a network that classifies a GS3 versus GS4 rating. The DL networks were configured to classify both source and Gleason pattern simultaneously, such as PANDA-GS3 or UMMCC-GS4, so both networks trained on four classes. These classes were then reduced to just a Gleason score for the inference decision (and comparison to patch label). Our AUC results on PANDA Radboud data, 0.988 for benign versus GS3/4 and 0.996 for GS3 versus GS4, demonstrate performance that equals the network trained solely on PANDA Radboud data. For our UM/MCC data, AUC is estimated at 0.963 for benign versus GS3/4 and 0.710 for GS3 versus GS4.

Table 5. PANDA + UMMCC training.

	Trained on Combined PANDA Radboud + UM/MCC			
	PANDA Radboud Benign vs. GS3/4	UM/MCC Benign vs. GS3/4	PANDA Radboud GS3 vs. GS4	UM/MCC GS3 vs. GS4
Accuracy	0.961 (0.93, 0.99)	0.915 (0.80, 0.97)	0.970 (0.94, 1.0)	0.668 (0.53, 0.84)
Sensitivity	0.944 (0.89, 0.99)	0.902 (0.75, 0.99)	0.971 (0.92, 1.0)	0.647 (0.36, 0.84)
Specificity	0.978 (0.95, 1.0)	0.928 (0.82, 0.98)	0.968 (0.88, 1.0)	0.689 (0.24, 0.87)
Precision	0.977 (0.95, 1.0)	0.927 (0.83, 0.98)	0.970 (0.89, 1.0)	0.687 (0.52, 0.85)
NPV	0.946 (0.90, 0.99)	0.908 (0.78, 0.99)	0.972 (0.92, 1.0)	0.665 (0.55, 0.83)
F ₁ -score	0.960 (0.92, 0.99)	0.913 (0.79, 0.97)	0.970 (0.94, 1.0)	0.656 (0.47, 0.84)
AUC	0.988 (0.96, 1.0)	0.963 (0.86, 0.99)	0.996 (0.99, 1.0)	0.710 (0.52, 0.90)

4. Discussion

DL methods have proven to be more effective in discriminating objects from different categories, exceeding human perception in the recent decade [27]. It is known that the DL methods' performance drops in discriminating subjects that are sparse in their occurrence in the training sets. In our case, various quality factors affect our performance, which include inconsistent lighting conditions or stain quality, stain differences and generalizing across sources [70,71]. State-of-the-art approaches have ushered in techniques for much more complicated classification tasks, including Gleason scoring (GS) or ISUP Gleason grading of prostate pathology [72–76]. Classifying indolent from cancer grade based on H&E pathology with multiple glands is empirically a difficult problem—there is a subtle distinction between any neighboring patterns with a fuzzy discrimination boundary between the pattern scoring levels [8,77].

In our study, discrimination from benign/indolent versus cancer (GS3/4) grades shows excellent performance at the glandular level (AUC 0.96, Table 5). Grades of cancer discrimination GS3 versus GS4 performance was lower for the UM/MCC dataset (AUC 0.71, Table 5) but excellent for Kaggle PANDA Radboud data (whose patches may include multiple glands). Our results demonstrate that classifying individual glands at a Gleason level has acceptable performance and provides a basis to develop overall slide-level Gleason

pattern scores. Grading individual glandular features has the advantage of increased fidelity in the decision process and provides supporting evidence for pathologists. Other studies have employed different approaches to the problem [78,79]. It is challenging to make a direct comparison with other findings; nevertheless, a comparison with recently published works follows. Singhal et al. show an accuracy of discriminating benign from malignancy of 0.85 on PANDA Radboud biopsy images [80], while our work shows an accuracy of 0.92 at the glandular level on our UMMCC data cohort and 0.96 on the PANDA Radboud data. Comparing with PANDA's challenge data, in Bulten (2022) [16], a representative algorithm had validation metrics on sensitivity and specificity of tumor detection at 99.7% (98.1–99.7) and 92.9% (91.9–96.7), respectively, while our benign-versus-malignant classifier demonstrated Radboud scores of 94.4% (89.0–99.0) and 97.8% (95–1.0). In each of these comparisons, the results of our tests were tallied on single patches or glands, while the other studies compute results over the entire biopsy. See Table 5 for a summary of our results.

We used a cyclic learning rate to improve gradient descent, but the technique may also be used to choose an effective ensemble set [62]. Generally, we can train an ensemble set using a variety of methods to improve generalization and to lower variance [81]. The high variance reported on the VGG-16 DL network was contributed to by the small validation sets but also by not leveraging an ensemble technique.

The generalization of consistent ISUP grading or Gleason scoring across sources proves to be difficult if the network has not trained on that source or if the network is retrained on a new dataset (forgetting what it has previously learned). The Supplementary Document Table S4 shows an example of catastrophic forgetting when testing PANDA data on a network fine-tuned on UM/MCC data, a common problem with machine learners [82]. Model training can be improved by continual learning, when integrating new sources, to ensure good cross-source generalization [83,84]. Alternatively, since datasets are relatively small within this domain, we successfully demonstrated combined-datasets training that integrated previously unseen sources, as shown in Table 5.

An orchestrated solution classifying glandular features and detection of additional features (such as density of nuclei and other recognized pathologic features) would exemplify a DSS that provides trust. Our contention is that ML and classifying entire WSIs is a useful aid, but DL CNNs, in particular, may make decisions that are not consistent with human observation and perception [12].

Using modern graphics processing units such as the NVIDIA® A100, training on our combined prostate pathology cohorts require less than 1 h of compute time per DL model (roughly 17 h for 20 models of a 20-fold CV). A single inference decision on an image patch takes fractions of a second to process on any high-end commercial and consumer grade GPU device (a GPU accelerator is not needed for inference when deployed in the field), supporting rapid response and interrogation of data for histopathologists.

Limitations and Future Improvements

A significant limitation to the approach is to classify higher grade patterns caused by limited samples in these grades. It is well-understood that a higher-grade Gleason pattern has progressively receding luminal regions and shows distinct morphological characteristics. Small sample size across the grades and varying glandular patterns make it difficult for the model to train and discriminate the patterns. We used the public cohort (PANDA Radboud) to find patches with Gleason pattern examples, but they may unavoidably contain multiple glands.

It is recognized that wide confidence bounds for some of our performance metrics could be attributed to smaller sample size. We believe using an ensemble technique, training on larger datasets and consensus scoring on WSIs will minimize the complexity of the DL model and improve the confidence bounds [85–87].

Future improvements to consider for generalization are consistent image preprocessing across sources, such as stain correction [14,88–90] and image resizing [91]. In future experiments, we intend to show improved generalization on unseen sources by first reduc-

ing discrepancies among data sources. Additionally, cross-source generalization could be improved by training an ensemble across many sources while ensuring that confounding factors are not contributing to shortcut learning—a generalized solution must learn from the wisdom of the masses to apply as a reference standard [92].

5. Conclusions

The baseline scores presented in this paper focus on discriminating primary Gleason patterns on individual glands or small patches of a prostate WSI. Our study demonstrates that a CNN DL model discriminates malignant patterns from benign tissue with a high level of accuracy. Furthermore, we were able to show validation of the findings in an independent larger-sized cohort (Kaggle PANDA Radboud data). Our work shows that classification of an indolent Gleason pattern from a clinically significant Gleason pattern shows impressive discrimination. We would need a larger sample cohort from diverse multi-centers to improve discrimination at the glandular level. Increasing the fidelity of an automated Gleason scoring scheme will provide a decision aid for clinical judgement.

It is well-recognized that clinical translation of pathological findings in the clinic would require improved region targeting. There are improved biopsy methods that show promise of improving tumor detection [93].

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/cancers15082335/s1>: Table S1: Hyper Parameters; Table S2: Comparison of deep learning networks on GS3 versus GS4 classification; Table S3: Comparison of Image Patch Resizing Techniques; Table S4: Cross-source inference demonstrates poor performance.

Author Contributions: The research article had several contributions that are captured in the following goals for the study. Conceptualization and methodology, R.F., L.H., D.G., Y.B.; c: Clinical samples: (UM): A.P., S.P., R.S.; (HLM/MCC): Y.B., J.P.-S.; (PANDAS): <https://doi.org/10.1038/s41591-021-01620-2>; software development, R.F., J.J.; clinical overread: A.L., M.G., formal analysis, investigation R.F., L.H., D.G., Y.B.; writing—original draft preparation, Y.B., R.F., L.H., D.G.; writing—review and editing, R.F., D.G., L.H., J.J., A.L., R.S., M.G., S.P., A.P., J.P.-S., Y.B.; funding acquisition, R.S., S.P., A.P., Y.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Cancer Institute (1R01CA190105, R01CA189295, U01CA200464, P30CA240139 and U01CA239141).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of the University of South Florida/Moffitt Cancer Center (MCC 18104, continued renewal 03/2022). The retrospective research study waives the requirement for approval.

Informed Consent Statement: Patients' informed consent was waived for the retrospective non-invasive study collected by an umbrella research protocol.

Data Availability Statement: Available and subject to institutional compliance (data transfer agreement).

Acknowledgments: We express sincere gratitude to our late mentor and friend Robert Gillies, who had been instrumental in initiating this direction of work with a goal to use imaging technologies to improve care in oncology. We are thankful for several inspiring discussions with members of Radiomics researchers at the H Lee Moffitt Cancer Center and the University of South Florida, Department of Computer Science. We convey our acknowledgements of several technicians, research staff who helped to curate the data used for the study.

Conflicts of Interest: The authors declare no conflict of interest related to this work.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area Under (ROC) Curve
CNN	Convolutional Neural Network
CV	Cross Validation

DL	Deep Learning
DSS	Decision Support System
FC	Fully Connected
GS	Gleason Score
ISUP	International Society of Urological Pathology
MCCV	Monte Carlo Cross-Validation
MCC	Moffitt Cancer Center
MIL	Multi-Instance Learning
ML	Machine Learning
NN	Neural Network
OD	Outlier Detection
PANDA	Prostate cANcer graDe Assessment
RANSAC	Random Sampling and Consensus
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SGD	Stochastic Gradient Descent
UM	University of Miami
WSI	Whole-Slide Image

References

- Humphrey, P.A. Histopathology of Prostate Cancer. *Cold Spring Harb. Perspect. Med.* **2017**, *7*, a030411. [[CrossRef](#)]
- Marini, N.; Otolara, S.; Wodzinski, M.; Tomassini, S.; Dragoni, A.F.; Marchand-Maillet, S.; Morales, J.P.D.; Duran-Lopez, L.; Vatrano, S.; Müller, H.; et al. Data-driven color augmentation for H&E stained images in computational pathology. *J. Pathol. Inform.* **2023**, *14*, 100183. [[CrossRef](#)] [[PubMed](#)]
- Humphrey, P.A.; Moch, H.; Cubilla, A.L.; Ulbright, T.M.; Reuter, V.E. The 2016 WHO Classification of Tumours of the Urinary System and Male Genital Organs-Part B: Prostate and Bladder Tumours. *Eur. Urol.* **2016**, *70*, 106–119. [[CrossRef](#)] [[PubMed](#)]
- Epstein, J.I.; Zelefsky, M.J.; Sjöberg, D.D.; Nelson, J.B.; Egevad, L.; Magi-Galluzzi, C.; Vickers, A.J.; Parwani, A.V.; Reuter, V.E.; Fine, S.W.; et al. A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score. *Eur. Urol.* **2016**, *69*, 428–435. [[CrossRef](#)] [[PubMed](#)]
- Sehn, J.K. Prostate Cancer Pathology: Recent Updates and Controversies. *MO Med.* **2018**, *115*, 151–155. [[PubMed](#)]
- Allsbrook, W.C.; Mangold, K.A.; Johnson, M.H.; Lane, R.B.; Lane, C.G.; Amin, M.B.; Bostwick, D.G.; Humphrey, P.A.; Jones, E.C.; Reuter, V.E.; et al. Interobserver reproducibility of Gleason grading of prostatic carcinoma: Urologic pathologists. *Hum. Pathol.* **2001**, *32*, 74–80. [[CrossRef](#)] [[PubMed](#)]
- Allsbrook, W.C.; Mangold, K.A.; Johnson, M.H.; Lane, R.B.; Lane, C.G.; Epstein, J.I. Interobserver reproducibility of Gleason grading of prostatic carcinoma: General pathologist. *Hum. Pathol.* **2001**, *32*, 81–88. [[CrossRef](#)] [[PubMed](#)]
- Egevad, L.; Ahmad, A.S.; Algaba, F.; Berney, D.M.; Boccon-Gibod, L.; Compérat, E.; Evans, A.J.; Griffiths, D.; Grobholz, R.; Kristiansen, G.; et al. Standardization of Gleason grading among 337 European pathologists: Gleason grading in Europe. *Histopathology* **2013**, *62*, 247–256. [[CrossRef](#)]
- Ozkan, T.A.; Eruyar, A.T.; Cebeci, O.O.; Memik, O.; Ozcan, L.; Kuskonmaz, I. Interobserver variability in Gleason histological grading of prostate cancer. *Scand. J. Urol.* **2016**, *50*, 420–424. [[CrossRef](#)]
- Oyama, T.; Allsbrook, W.C., Jr.; Kurokawa, K.; Matsuda, H.; Segawa, A.; Sano, T.; Suzuki, K.; Epstein, J.I. A comparison of interobserver reproducibility of Gleason grading of prostatic carcinoma in Japan and the United States. *Arch. Pathol. Lab. Med.* **2005**, *129*, 1004–1010. [[CrossRef](#)]
- Zarella, M.D.; Bowman, D.; Aeffner, F.; Farahani, N.; Xthona, A.; Absar, S.F.; Parwani, A.; Bui, M.; Hartman, D.J. A Practical Guide to Whole Slide Imaging: A White Paper From the Digital Pathology Association. *Arch. Pathol. Lab. Med.* **2019**, *143*, 222–234. [[CrossRef](#)] [[PubMed](#)]
- Baker, N.; Elder, J.H. Deep learning models fail to capture the configural nature of human shape perception. *iScience* **2022**, *25*, 104913. [[CrossRef](#)] [[PubMed](#)]
- Chan, H.P.; Samala, R.K.; Hadjiiski, L.M.; Zhou, C. Deep Learning in Medical Image Analysis. *Adv. Exp. Med. Biol.* **2020**, *1213*, 3–21. [[CrossRef](#)]
- De Haan, K.; Zhang, Y.; Zuckerman, J.E.; Liu, T.; Sisk, A.E.; Diaz, M.F.P.; Jen, K.-Y.; Nobori, A.; Liou, S.; Zhang, S.; et al. Deep learning-based transformation of H&E stained tissues into special stains. *Nat. Commun.* **2021**, *12*, 4884. [[CrossRef](#)]
- Deng, S.; Zhang, X.; Yan, W.; Chang, E.I.; Fan, Y.; Lai, M.; Xu, Y. Deep learning in digital pathology image analysis: A survey. *Front. Med.* **2020**, *14*, 470–487. [[CrossRef](#)] [[PubMed](#)]
- Bulten, W.; Kartasalo, K.; Chen, P.-H.C.; Ström, P.; Pinckaers, H.; Nagpal, K.; Cai, Y.; Steiner, D.F.; van Boven, H.; Vink, R.; et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: The PANDA challenge. *Nat. Med.* **2022**, *28*, 154–163. [[CrossRef](#)]

17. Paul, R.; Hawkins, S.; Balagurunathan, Y.; Schabath, M.; Gillies, R.; Hall, L.; Goldgof, D. Deep Feature Transfer Learning in Combination with Traditional Features Predicts Survival among Patients with Lung Adenocarcinoma. *Tomography* **2016**, *2*, 388–395. [[CrossRef](#)] [[PubMed](#)]
18. Paul, R.; Kariev, S.; Cherezov, D.; Schabath, M.; Gillies, R.; Hall, L.; Goldgof, D.; Drukker, K.; Mazurowski, M.A. Deep radiomics: Deep learning on radiomics texture images. In *Medical Imaging 2021: Computer-Aided Diagnosis*; SPIE: Bellingham, WA, USA, 2021. [[CrossRef](#)]
19. Bulten, W.; Pinckaers, H.; van Boven, H.; Vink, R.; de Bel, T.; van Ginneken, B.; van der Laak, J.; Hulsbergen-van de Kaa, C.; Litjens, G. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: A diagnostic study. *Lancet Oncol.* **2020**, *21*, 233–241. [[CrossRef](#)] [[PubMed](#)]
20. Bulten, W.; Balkenhol, M.; Belinga, J.-J.A.; Brillhante, A.; Çakır, A.; Egevad, L.; Eklund, M.; Farré, X.; Geronatsiou, K.; Molinié, V.; et al. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Mod. Pathol.* **2021**, *34*, 660–671. [[CrossRef](#)]
21. Pinckaers, H.; Bulten, W.; van der Laak, J.; Litjens, G. Detection of Prostate Cancer in Whole-Slide Images Through End-to-End Training With Image-Level Labels. *IEEE Trans. Med. Imaging* **2021**, *40*, 1817–1826. [[CrossRef](#)]
22. 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.; et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: A population-based, diagnostic study. *Lancet Oncol.* **2020**, *21*, 222–232. [[CrossRef](#)] [[PubMed](#)]
23. 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 Digit. Med.* **2019**, *2*, 48. [[CrossRef](#)] [[PubMed](#)]
24. Rana, A.; Lowe, A.; Lithgow, M.; Horback, K.; Janovitz, T.; Da Silva, A.; Tsai, H.; Shanmugam, V.; Bayat, A.; Shah, P. Use of Deep Learning to Develop and Analyze Computational Hematoxylin and Eosin Staining of Prostate Core Biopsy Images for Tumor Diagnosis. *JAMA Netw. Open* **2020**, *3*, e205111. [[CrossRef](#)] [[PubMed](#)]
25. Varoquaux, G.; Cheplygina, V. Machine learning for medical imaging: Methodological failures and recommendations for the future. *Npj Digit. Med.* **2022**, *5*, 48. [[CrossRef](#)] [[PubMed](#)]
26. Anwar, S.M.; Majid, M.; Qayyum, A.; Awais, M.; Alnowami, M.; Khan, M.K. Medical Image Analysis using Convolutional Neural Networks: A Review. *J. Med. Syst.* **2018**, *42*, 226. [[CrossRef](#)] [[PubMed](#)]
27. Krizhevshy, A.; Sutskever, I.; Hilton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [[CrossRef](#)]
28. Thrall, J.H.; Li, X.; Li, Q.; Cruz, C.; Do, S.; Dreyer, K.; Brink, J. Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success. *J. Am. Coll. Radiol.* **2018**, *15*, 504–508. [[CrossRef](#)]
29. Zhu, Y.; Wei, R.; Gao, G.; Ding, L.; Zhang, X.; Wang, X.; Zhang, J. Fully automatic segmentation on prostate MR images based on cascaded fully convolution network. *J. Magn. Reson. Imaging* **2019**, *49*, 1149–1156. [[CrossRef](#)]
30. Morid, M.A.; Borjali, A.; Del Fiol, G. A scoping review of transfer learning research on medical image analysis using ImageNet. *Comput. Biol. Med.* **2021**, *128*, 104115. [[CrossRef](#)]
31. Raghu, M.; Zhang, C.; Kleinberg, J.; Bengio, S. Transfusion: Understanding Transfer Learning for Medical Imaging. In Proceedings of the NeurIPS, Vancouver, BC, Canada, 8–14 December 2019.
32. Alzubaidi, L.; Duan, Y.; Al-Dujaili, A.; Ibraheem, I.K.; Alkenani, A.H.; Santamaria, J.; Fadhel, M.A.; Al-Shamma, O.; Zhang, J. Deepening into the suitability of using pre-trained models of ImageNet against a lightweight convolutional neural network in medical imaging: An experimental study. *PeerJ Comput. Sci.* **2021**, *7*, e715. [[CrossRef](#)]
33. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 6105–6114.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: Las Vegas, NV, USA, 2016; pp. 770–778.
35. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
36. Linkon, A.H.M.; Labib, M.M.; Hasan, T.; Hossain, M.; Jannat, M.-E. Deep learning in prostate cancer diagnosis and Gleason grading in histopathology images: An extensive study. *Inform. Med. Unlocked* **2021**, *24*, 100582. [[CrossRef](#)]
37. Campanella, G.; Hanna, M.G.; Geneslaw, L.; Mirafior, A.; Werneck Krauss Silva, V.; Busam, K.J.; Brogi, E.; Reuter, V.E.; Klimstra, D.S.; Fuchs, T.J. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **2019**, *25*, 1301–1309. [[CrossRef](#)]
38. Xu, Q.-S.; Liang, Y.-Z. Monte Carlo cross validation. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 1–11. [[CrossRef](#)]
39. Talebi, H.; Milanfar, P. Learning to Resize Images for Computer Vision Tasks. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; IEEE: Montreal, QC, Canada, 2021; pp. 487–496.
40. Hashemi, M. Enlarging smaller images before inputting into convolutional neural network: Zero-padding vs. interpolation. *J. Big Data* **2019**, *6*, 98. [[CrossRef](#)]

41. Kim, J.; Jang, J.; Seo, S.; Jeong, J.; Na, J.; Kwak, N. MUM: Mix Image Tiles and UnMix Feature Tiles for Semi-Supervised Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14512–14521.
42. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. *arXiv* **2019**, arXiv:1905.04899.
43. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond Empirical Risk Minimization. *arXiv* **2017**, arXiv:1710.09412.
44. Kleczek, P.; Jaworek-Korjakowska, J.; Gorgon, M. A novel method for tissue segmentation in high-resolution H&E-stained histopathological whole-slide images. *Comput. Med. Imaging Graph.* **2020**, *79*, 101686. [[CrossRef](#)]
45. Chan, J.K. The wonderful colors of the hematoxylin-eosin stain in diagnostic surgical pathology. *Int. J. Surg. Pathol.* **2014**, *22*, 12–32. [[CrossRef](#)]
46. Cooper, L.A.; Demicco, E.G.; Saltz, J.H.; Powell, R.T.; Rao, A.; Lazar, A.J. PanCancer insights from The Cancer Genome Atlas: The pathologist’s perspective. *J. Pathol.* **2018**, *244*, 512–524. [[CrossRef](#)]
47. Hutchison, D.; Kanade, T.; Kittler, J.; Kleinberg, J.M.; Mattern, F.; Mitchell, J.C.; Naor, M.; Nierstrasz, O.; Pandu Rangan, C.; Steffen, B.; et al. Image Segmentation with Implicit Color Standardization Using Spatially Constrained Expectation Maximization: Detection of Nuclei. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2012*; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7510, pp. 365–372.
48. Hoffman, R.A.; Kothari, S.; Wang, M.D. Comparison of normalization algorithms for cross-batch color segmentation of histopathological images. In Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, 26–30 August 2014; IEEE: Chicago, IL, USA, 2014; pp. 194–197. [[CrossRef](#)]
49. Magee, D.; Treanor, D.; Crellin, D.; Shires, M.; Mohee, K.; Quirke, P. Colour Normalisation in Digital Histopathology Images. In *Optical Tissue Image Analysis in Microscopy, Histopathology and Endoscopy*; Daniel Elson: London, UK, 2009; Volume 100, pp. 100–111.
50. Fernández, Á.; Bella, J.; Dorronsoro, J.R. Supervised outlier detection for classification and regression. *Neurocomputing* **2022**, *486*, 77–92. [[CrossRef](#)]
51. Fischler, M.A.; Bolles, R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **1981**, *24*, 381–395. [[CrossRef](#)]
52. Ur Rehman, A.; Belhaouari, S.B. Unsupervised outlier detection in multidimensional data. *J. Big Data* **2021**, *8*, 80. [[CrossRef](#)]
53. Li, Y.; Pei, W.; He, Z. SSORN: Self-Supervised Outlier Removal Network for Robust Homography Estimation. *arXiv* **2022**, arXiv:2208.14093.
54. Davison, A.C.; Hinkley, D.V.; Schechtman, E. Efficient Bootstrap Simulation. *Biometrika* **1986**, *73*, 555–566. [[CrossRef](#)]
55. D’Ascoli, S.; Touvron, H.; Leavitt, M.; Morcos, A.; Biroli, G.; Sagun, L. ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. *arXiv* **2021**, arXiv:2103.10697. [[CrossRef](#)]
56. Abnar, S.; Dehghani, M.; Zuidema, W. Transferring Inductive Biases through Knowledge Distillation. *arXiv* **2020**, arXiv:2006.00555.
57. Feng, H.; Yang, B.; Wang, J.; Liu, M.; Yin, L.; Zheng, W.; Yin, Z.; Liu, C. Identifying Malignant Breast Ultrasound Images Using ViT-Patch. *Appl. Sci.* **2023**, *13*, 3489. [[CrossRef](#)]
58. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
59. Zhou, Y.T.; Chellappa, R. Computation of optical flow using a neural network. In Proceedings of the IEEE International Conference on Neural Networks, Anchorage, AK, USA, 4–9 May 1998; IEEE: San Diego, CA, USA, 1998; Volume 2, pp. 71–78.
60. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *PMLR* **2015**, *37*, 448–456.
61. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv* **2017**, arXiv:1608.03983.
62. Huang, G.; Li, Y.; Pleiss, G.; Liu, Z.; Hopcroft, J.E.; Weinberger, K.Q. Snapshot Ensembles: Train 1, get M for free. *arXiv* **2017**, arXiv:1704.00109.
63. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 17. [[CrossRef](#)] [[PubMed](#)]
64. Qu, J.; Hiruta, N.; Terai, K.; Nosato, H.; Murakawa, M.; Sakanashi, H. Gastric Pathology Image Classification Using Stepwise Fine-Tuning for Deep Neural Networks. *J. Healthc. Eng.* **2018**, *2018*, 8961781. [[CrossRef](#)]
65. Trevethan, R. Sensitivity, Specificity, and Predictive Values: Foundations, Plabilities, and Pitfalls in Research and Practice. *Front. Public Health* **2017**, *5*, 307. [[CrossRef](#)] [[PubMed](#)]
66. Forman, G.; Scholz, M. Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement. *Assoc. Comput. Mach.* **2010**, *12*, 49–57. [[CrossRef](#)]
67. DiCiccio, T.J.; Efron, B. Bootstrap Confidence Intervals. *Stat. Sci.* **1996**, *11*, 40. [[CrossRef](#)]
68. Hesamian, M.H.; Jia, W.; He, X.; Kennedy, P. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *J. Digit. Imaging* **2019**, *32*, 582–596. [[CrossRef](#)]
69. Gao, J.; Jiang, Q.; Zhou, B.; Chen, D. Convolutional neural networks for computer-aided detection or diagnosis in medical image analysis: An overview. *Math. Biosci. Eng.* **2019**, *16*, 6536–6561. [[CrossRef](#)]
70. Pujar, A.; Pereira, T.; Tamgadge, A.; Bhalerao, S.; Tamgadge, S. Comparing The Efficacy of Hematoxylin and Eosin, Periodic Acid Schiff and Fluorescent Periodic Acid Schiff-Acridine Techniques for Demonstration of Basement Membrane in Oral Lichen Planus: A Histochemical Study. *Indian J. Dermatol.* **2015**, *60*, 450–456. [[CrossRef](#)]

71. Azevedo Tosta, T.A.; de Faria, P.R.; Neves, L.A.; do Nascimento, M.Z. Computational normalization of H&E-stained histological images: Progress, challenges and future potential. *Artif. Intell. Med.* **2019**, *95*, 118–132. [[CrossRef](#)] [[PubMed](#)]
72. Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F.A.; Brendel, W. ImageNet-trained CNNs are biased towards texture; Increasing Shape Bias Improves Accuracy and Robustness. *arXiv* **2019**, arXiv:1811.12231.
73. Ciga, O.; Xu, T.; Martel, A.L. Self supervised contrastive learning for digital histopathology. *Mach. Learn. Appl.* **2022**, *7*, 100198. [[CrossRef](#)]
74. Yang, P.; Yin, X.; Lu, H.; Hu, Z.; Zhang, X.; Jiang, R.; Lv, H. CS-CO: A Hybrid Self-Supervised Visual Representation Learning Method for H&E-stained Histopathological Images. *Med. Image Anal.* **2022**, *81*, 102539. [[CrossRef](#)]
75. Wang, X.; Yang, S.; Zhang, J.; Wang, M.; Zhang, J.; Yang, W.; Huang, J.; Han, X. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **2022**, *81*, 102559. [[CrossRef](#)]
76. Chen, H.; Li, C.; Wang, G.; Li, X.; Mamunur Rahaman, M.; Sun, H.; Hu, W.; Li, Y.; Liu, W.; Sun, C.; et al. GasHis-Transformer: A multi-scale visual transformer approach for gastric histopathological image detection. *Pattern Recognit.* **2022**, *130*, 108827. [[CrossRef](#)]
77. Egevad, L.; Swanberg, D.; Delahunt, B.; Ström, P.; Kartasalo, K.; Olsson, H.; Berney, D.M.; Bostwick, D.G.; Evans, A.J.; Humphrey, P.A.; et al. Identification of areas of grading difficulties in prostate cancer and comparison with artificial intelligence assisted grading. *Virchows Arch.* **2020**, *477*, 777–786. [[CrossRef](#)]
78. Salvi, M.; Bosco, M.; Molinaro, L.; Gambella, A.; Papotti, M.; Acharya, U.R.; Molinari, F. A hybrid deep learning approach for gland segmentation in prostate histopathological images. *Artif. Intell. Med.* **2021**, *115*, 102076. [[CrossRef](#)]
79. Bulten, W.; Bándi, P.; Hoven, J.; Loo, R.v.d.; Lotz, J.; Weiss, N.; Laak, J.V.d.; Ginneken, B.V.; Hulsbergen-van de Kaa, C.; Litjens, G. Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard. *Sci. Rep.* **2019**, *9*, 864. [[CrossRef](#)]
80. Singhal, N.; Soni, S.; Bonthu, S.; Chattopadhyay, N.; Samanta, P.; Joshi, U.; Jojera, A.; Chharchhodawala, T.; Agarwal, A.; Desai, M.; et al. A deep learning system for prostate cancer diagnosis and grading in whole slide images of core needle biopsies. *Sci. Rep.* **2022**, *12*, 3383. [[CrossRef](#)]
81. Paul, R.; Schabath, M.; Gillies, R.; Hall, L.; Goldgof, D. Convolutional Neural Network ensembles for accurate lung nodule malignancy prediction 2 years in the future. *Comput. Biol. Med.* **2020**, *122*, 103882. [[CrossRef](#)] [[PubMed](#)]
82. McCloskey, M.; Cohen, N.J. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *Psychol. Learn. Motiv.* **1989**, *24*, 57. [[CrossRef](#)]
83. Li, Z.; Hoiem, D. Learning without Forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2935–2947. [[CrossRef](#)]
84. Rusu, A.A.; Rabinowitz, N.C.; Desjardins, G.; Soyer, H.; Kirkpatrick, J.; Kavukcuoglu, K.; Pascanu, R.; Hadsell, R. Progressive Neural Networks. *arXiv* **2022**, arXiv:1606.04671.
85. Alahmari, S.S.; Goldgof, D.B.; Mouton, P.R.; Hall, L.O. Challenges for the Repeatability of Deep Learning Models. *IEEE Access* **2020**, *8*, 211860–211868. [[CrossRef](#)]
86. Langford, J. Tutorial On Practical Prediction Theory For Classification. *J. Mach. Learn. Res.* **2005**, *6*, 273–306.
87. Hosen, M.A.; Khosravi, A.; Nahavandi, S.; Creighton, D. Improving the Quality of Prediction Intervals Through Optimal Aggregation. *IEEE Trans. Ind. Electron.* **2015**, *62*, 4420–4429. [[CrossRef](#)]
88. Khan, A.M.; Rajpoot, N.; Treanor, D.; Magee, D. A Nonlinear Mapping Approach to Stain Normalization in Digital Histopathology Images Using Image-Specific Color Deconvolution. *IEEE Trans. Biomed. Eng.* **2014**, *61*, 1729–1738. [[CrossRef](#)]
89. Dave, P.; Alahmari, S.; Goldgof, D.; Hall, L.O.; Morera, H.; Mouton, P.R. An adaptive digital stain separation method for deep learning-based automatic cell profile counts. *J. Neurosci. Methods* **2021**, *354*, 109102. [[CrossRef](#)]
90. Ruifrok, A.C.; Johnston, D.A. Quantification of histochemical staining by color deconvolution. *Anal. Quant. Cytol. Histol.* **2001**, *23*, 291–299.
91. Danon, D.; Arar, M.; Cohen-Or, D.; Shamir, A. Image resizing by reconstruction from deep features. *Comput. Vis. Media* **2021**, *7*, 453–466. [[CrossRef](#)]
92. Ben Ahmed, K.; Hall, L.O.; Goldgof, D.B.; Fogarty, R. Achieving Multisite Generalization for CNN-Based Disease Diagnosis Models by Mitigating Shortcut Learning. *IEEE Access* **2022**, *10*, 78726–78738. [[CrossRef](#)]
93. Cauni, V.; Stanescu, D.; Tanase, F.; Mihai, B.; Persu, C. Magnetic Resonance/Ultrasound Fusion Targeted Biopsy of the Prostate Can Be Improved By Adding Systematic Biopsy. *Med. Ultrason.* **2021**, *23*, 277–282. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.