Deep Learning Methodologies Applied to Digital Pathology in Prostate Cancer: A Systematic Review

Noémie Rabilloud 1,*, Pierre Allaume 2, Oscar Acosta 1, Renaud De Crevoisier 1,3, Raphael Bourgade 4, Delphine Loussouarn 4, Nathalie Rioux-Leclercq 5, Zine-eddine Khene 1,5, Romain Mathieu 5, Karim Bensalah 5, Thierry Pecot * and Solene-Florence Kammerer-Jacquet 1,2

Abstract: Deep learning (DL), often called artificial intelligence (AI), has been increasingly used in pathology thanks to the use of scanners to digitize slides which allow us to visualize them on monitors and process them with AI algorithms. Many articles have focused on DL applied to prostate cancer (PCa). This systematic review explains the DL applications and their performances for PCAs in digital pathology. Article research was performed using PubMed and Embase to collect relevant articles. A Risk of Bias (RoB) was assessed with an adaptation of the QUADAS-2 tool. Out of the 77 included studies, eight focused on pre-processing tasks such as quality assessment or staining normalization. Most articles (n = 53) focused on diagnosis tasks like cancer detection or Gleason grading. Fifteen articles focused on prediction tasks, such as recurrence prediction or genomic correlations. Best performances were reached for cancer detection with an Area Under the Curve (AUC) up to 0.99 with algorithms already available for routine diagnosis. A few biases outlined by the RoB analysis are often found in these articles, such as the lack of external validation. This review was registered on PROSPERO under CRD42023418661.

Keywords: prostate cancer; deep learning; digital pathology; Gleason grading; convolutional neural networks; artificial intelligence

1. Introduction

1.1. Prostate Cancer

Prostate cancer (PCa) is one of the most prevalent cancers among male cancers, especially aging [1]. The gold standard for diagnosing and treating patients is the analysis of H&E histopathology slides [2]. The observation of the biopsy tissue enables pathologists to detect tumor cells and characterize the aggressiveness of the tumor using Gleason grading. This grading is based on gland structure and ranks 1 to 5 according to the differentiation [3]. When more than one pattern is present on the biopsy, the scoring is defined by the most represented pattern (primary) and the highest one (secondary). For instance, a biopsy with most of pattern 3 and some patterns 4 and 5 will be scored $3 + 5 = 8$. To improve the prognosis correlation among these scores, an update of Gleason grading called the ISUP (International Society of Urological Pathology) Grading system, was proposed...
in 2014, which assigns patients into a group depending on the Gleason score (see Table 1) [4]. These groups have varying prognoses, from group 1 corresponding to an indolent tumor to group 5 having a poor prognosis [5].

Table 1. Correspondence between Gleason score and ISUP Gleason group.

<table>
<thead>
<tr>
<th>Gleason ISUP Group</th>
<th>Gleason Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Score = 6</td>
</tr>
<tr>
<td>2</td>
<td>Score 7 (3 + 4)</td>
</tr>
<tr>
<td>3</td>
<td>Score 7 (4 + 3)</td>
</tr>
<tr>
<td>4</td>
<td>Score 8 (4 + 4, 5 + 3, 3 + 5)</td>
</tr>
<tr>
<td>5</td>
<td>Score 9 or 10</td>
</tr>
</tbody>
</table>

1.2. DL Applied to WSI

Glass slides are digitized with scanners to form whole slide images (WSIs). These images can be used to train DL algorithms. The principle is to teach an algorithm to correlate WSIs to a target provided by the user called the ground truth (GT). Once training is over, it is essential to test (or validate) the algorithm on other WSIs to validate it on other images. All studies must therefore have a training and a testing (or validation) cohort. Additionally, to comprehensively evaluate a model, it is common to use cross-validation (cross-val), i.e., splitting the training dataset into \( n \) parts, called folds, to train a model with \( (n - 1)/n \) of all data and evaluate it with \( 1/n \) of all data, \( n \) times. Finally, the algorithm can be evaluated on an external dataset.

Two main types of algorithms are discussed in this paper: segmentation and classification algorithms. Segmentation algorithms aim at precisely delineating regions of interest in WSIs (see Figure 1A). For instance, a common segmentation task is to localize stroma and epithelium. The most popular architecture is U-Net [6]. Classification algorithms intend to estimate labels, called classes, associated with images. In pathomics, WSIs are divided into tiles that are then encoded into features representing a summary of the tiles. A classification process is then learned from the features (see Figure 1B). These features are deep-learned with convolutional neural networks (CNN). The main architectures that are used for classification are Inception [7], ResNet [8], VGG [9], DenseNet [10], and MobileNet [11]. Those classic architectures can be trained and customized, but it is also possible to use models pre-trained on other data and fine-tune these models on the desired data. This is called transfer learning.

A significant difficulty in pathomics is the very large size of WSIs which prevents them from being processed all at once. This requires tiling WSIs in many tiles and dealing with data of different sizes. Most architectures need the same size of input for every image. The number of tiles selected must therefore be the same for each image. The magnification at which the image is tiled is also a key parameter in handling the input data. Additionally, this leads to potentially very large amounts of data to manually annotate to create the GT, a time-consuming task for pathologists. Strategies such as weak supervision, where only one label per patient is assigned from the pathologists’ report, have emerged to speed up this work.
Figure 1. DL applied to WSIs. (A) Segmentation algorithm, (B) Classification algorithm. WSIs are divided into many tiles, and every tile is encoded into features. Tiling can be performed at different magnifications, but an identical number of tiles per WSIs is generally required. The encoding into features can be trained or performed with a pre-trained algorithm. Features are then used to train a classification algorithm.

1.3. Applications and Evaluation of Algorithms

In pathomics applied to PCa, DL algorithms are applied for pre-processing, diagnosis, or prediction. Pre-processing algorithms evaluate input data quality and staining normalization. Diagnosis methods focus either on cancer detection or on Gleason grading. Prediction approaches focus on predicting survival and cancer progression or genomic signatures. Algorithms are evaluated with different metrics that are summarized in Supplementary Table S1. The most used metric is the Area Under the Curve (AUC) which can be used to evaluate diagnosis and prognosis tasks.
1.4. Aim of This Review

This paper reviews the main articles mentioning DL algorithms applied to prostate cancer WSIs until 2022. It highlights current trends, future perspectives, and potential improvements in the field. It aims to be a thorough but comprehensive take on this subject for everyone interested.

2. Materials and Methods

This systematic review followed the PRISMA guidelines, containing advices and checklists to frame systematic reviews [12]. It was registered on PROSPERO, a website identifying all existing and undergoing systematic reviews under CRD42023418661.

PubMed and Embase, two biomedical databases, were used to look for articles until 1 September 2023 with the following keywords:

(‘prostate’ OR prostate OR prostatic) AND (cancer OR carcinoma OR malignant OR lesions) AND (‘artificial intelligence’ OR ‘algorithm’ OR ‘deep learning’ OR ‘machine learning’ OR ‘automated classification’ OR ‘supervised learning’ OR ‘neural network’) AND (‘whole slide image’ OR ‘digital pathology’ OR pathomics OR ‘he’ OR ‘H&E’ OR ‘histological’)

Additionally, selected papers had to:
- be written in English,
- focus on prostate cancer,
- use pathology H&E-stained images,
- rely on deep learning.

The selection of papers was first evaluated with titles only. Abstracts were then reviewed, leading to a collection of 45 selected papers. Finally, 32 articles that did not come through the database search but were referenced in many of the 45 papers were manually added. Additional searches via PubMed and Google were performed to check if conference abstracts led to published articles. Figure 2 illustrates the selection process.
A risk-of-bias study was performed for each paper, using an adaptation of QUADAS-2 [13] (see Table 2) as the most suitable tool for AI papers. QUADAS-AI is not yet available [14]. Remarks made in this paper and criteria in the IJMEDI checklist [15] were used to create a homemade checklist adapted to this review. This allowed us to evaluate all papers on the same criteria to outline the main biases.

Table 2. Checklist used to perform RoB analysis for every paper.

<table>
<thead>
<tr>
<th>Quadas-2 Category</th>
<th>Adaptation</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient selection</td>
<td>Data source</td>
<td>Existence of exclusion criteria</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data publicly available</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reference to demographics and stratification</td>
</tr>
<tr>
<td></td>
<td>Data process</td>
<td>Sufficient amount of training data *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Normalization process properly described</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strategy for data imbalance</td>
</tr>
<tr>
<td>Index Test</td>
<td>Model</td>
<td>Description of model task</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Description of model output</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Description of architecture</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Definition of hyperparameter values</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Code publicly available</td>
</tr>
<tr>
<td>Reference standard</td>
<td>Ground truth</td>
<td>Ground truth defined in accordance with good clinical practice</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ground truth defined by multiple experts</td>
</tr>
<tr>
<td>Flow and timing</td>
<td>Analysis</td>
<td>Description of splitting methods</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Definition of a metric appropriate to the analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Relevance of algorithms with state of the art</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Use of external testing cohort</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Explainability of the model</td>
</tr>
</tbody>
</table>

* For diagnosis and prediction tasks, a number of 200 WSIs was required.

3. Results

Data was collected for every article: first author name, year of publication, aim of the paper, neural network used, number of slides in the training, validation (internal and external) cohorts, sub-aims of the paper and their corresponding performances with an Excel Sheet. All tables can be found in the Supplementary Materials with more detailed information (Supplementary Tables S2–S8).

3.1. Pre-Processing

Out of the 77 selected articles, three proposed methods for WSI quality assessment and 6 attended to correct for staining and/or scanning discrepancies across a set of WSIs (Table 3).

Table 3. Papers focusing on pre-processing tasks.

<table>
<thead>
<tr>
<th>First Author, Year Reference</th>
<th>DL Architecture</th>
<th>Training Cohort</th>
<th>IV Cohort</th>
<th>EV Cohort</th>
<th>Aim</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schömig-Markiefka, 2021 [16]</td>
<td>Inception-ResNetV2</td>
<td>Already trained</td>
<td>Subset of TCGA slides</td>
<td>686 WSIs</td>
<td>Impact of artifacts on tissue classification</td>
<td>Tissue classification performance decreases with the appearance of artifacts</td>
</tr>
<tr>
<td>Haghighat, 2022 [17]</td>
<td>ResNet18 PathProfiler</td>
<td>198 WSIs</td>
<td>3819 WSIs</td>
<td>None</td>
<td>Prediction of tile-level usability</td>
<td>AUC: 0.94 (IV)</td>
</tr>
</tbody>
</table>
3.1.1. Quality Assessment

Two articles proposed to evaluate the quality of a collection of slides by using a purity score (AUC of 0.8) [18] or a binary criterion of usability (AUC of 0.983) [17], attributed by pathologists. A third study proposed to artificially alter a collection of WSIs with artifacts and to evaluate the resulting impact on the tissue classification of tiles, which was perfectly performed without artifacts (F1-score = 1) [16]. The authors demonstrated that the main causes for the drop in performance are defocusing, jpg compression and staining variations. Indeed, the latter topic has been the subject of quite some articles in the field.

3.1.2. Staining Normalization and Tile Selection

WSIs stained in different locations show differences in intensity observed in the three-color channels (RGB). The scanners also influence the acquired WSIs. Many methods were proposed to normalize staining and/or scanning to obtain more generic DL models. Swiderska-Chadaj et al. evaluated the performance of their WSI normalization when classifying patients for benign/malignant tissues [24]. First, they improved classification performance with external datasets when images were scanned with the same scanner as the one used to scan the training cohort. Then, they trained a generative model to normalize WSIs for the external dataset, improving their classification performance (from an AUC of 0.92 to 0.98). Similarly, Rana et al. virtually stained with H&E non-stained WSIs.
with the help of generative models. Then, they stained the same biopsies with H&E and compared virtual and stained slides, reaching a high correlation (PC of 0.96) [22]. Salvi et al. focused not only on staining normalization but also on tile selection. Basing itself on segmentation, it allows for better select tiles to represent the WSI and its tissue’s complexity [25]. The model improvement (gain of 0.3 in specificity, reaching a sensitivity of 1) highlights the need for preprocessing steps, such as stain normalization or segmentation-guided tile selection.

3.2. Diagnosis

Most of the selected papers (n = 53) focused on diagnosis, whether it be about tissue segmentation (n = 5), detection of cancer tissue (n = 21), attribution of a Gleason grading (n = 10), or both detection of cancer tissue and Gleason grading (n = 19 with 2 in segmentation also). Diagnosis can be performed at the pixel level, tile level, WSI level or patient level. Methods processing data at the pixel level are segmentation algorithms trained to localize areas of clinical interest, e.g., malignant and benign epithelium. Methods processing data at tile, WSI or patient level are classification algorithms trained to identify categories of clinical interest associated with the input data.

3.2.1. Segmentation

Many segmentation studies (see Table 4) demonstrated high performance for gland segmentation: a Dice of 0.84 [26], a Dice of 0.9 [27], an F1 score from 0.83 to 0.85 [28] and an AUC of 0.99 [29]. Precisely annotating WSIs is time-consuming, and several approaches focused on strategies to reduce the number of annotations. In [30], the authors trained a model with rough annotations obtained with traditional image processing techniques. Then they fine-tuned this model with a few precise annotations made by pathologists (AUC gain of 0.04). In [28], slides were first stained with immunohistochemistry (IHC) for which DL models were already trained. The slides were stained with H&E, and a new model was trained using the annotations masks obtained from the IHC segmentation model.

Table 4. Papers focusing on segmenting glands or tissue to help future classification tasks.

<table>
<thead>
<tr>
<th>First Author, Year Reference</th>
<th>DL Architecture</th>
<th>Training Cohort</th>
<th>IV Cohort</th>
<th>EV Cohort</th>
<th>Aim</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ren, 2017 [26]</td>
<td>U-Net</td>
<td>22 WSIs</td>
<td>5-fold cross val</td>
<td>None</td>
<td>Gland Segmentation in H&amp;E slides</td>
<td>F1: 0.84</td>
</tr>
<tr>
<td>Li, 2019 [29]</td>
<td>R-CNN</td>
<td>40 patients</td>
<td>5-fold cross val</td>
<td>None</td>
<td>Epithelial cell detection</td>
<td>AUC: 0.998</td>
</tr>
<tr>
<td>Bulten, 2019 [28]</td>
<td>U-Net</td>
<td>20 WSIs</td>
<td>5 WSIs</td>
<td>None</td>
<td>Epithelium segmentation based on IHC</td>
<td>IoU: 0.854 F1: 0.915</td>
</tr>
<tr>
<td>Bukowy, 2020 [30]</td>
<td>SegNet (VGG16)</td>
<td>10 WSIs</td>
<td>6 WSIs</td>
<td>None</td>
<td>Weakly- and strongly-annotated segmentation</td>
<td>$AUC: 0.85$ (strong) $AUC: 0.93$ (weak fine-tuned strong)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>140 WSIs</td>
<td></td>
<td></td>
<td>Epithelium segmentation with a combination of 3 models</td>
<td>$ACC: 0.86$ (DL)</td>
</tr>
<tr>
<td>Salvi, 2021 [27]</td>
<td>U-Net + post-processing</td>
<td>100 patients</td>
<td>50 patients</td>
<td>None</td>
<td>Gland segmentation</td>
<td>Dice: 0.901 U-Net + post-processing Dice: 0.892 U-Net only</td>
</tr>
</tbody>
</table>

IV: Internal Validation. EV: External Validation. CNN: Convolutional Neural Network. AUC: Area Under the Curve. F1: F1-score. IoU: Intersection Over Union. $\text{metric}$ implies the mean of this metric. cross val: cross-validation.
3.2.2. Cancer Detection

Even if segmentation helps select tiles of interest, other processes can be used to improve classification. There is the possibility of using multi-scale embedding to take in an area at different magnifications at the same time (as a pathologist would do) [31]. To select tiles and improve the explainability of the model, one of the most popular approaches for patient classification in pathology is Multiple Instance Learning (MIL) [32–34]). From a WSI divided into a high number of tiles, the goal is to associate to each tile a probability for the considered classification task, e.g., the presence of cancer, and then use the most predictive tiles to diagnose. This DL approach directly addresses the tile selection during the training, allowing us to deal with WSIs of different sizes and identify the tiles used to decide. This paradigm was notably used for cancer detection by Campanella et al. on the largest training and external validation cohorts to date [32] and by Pinckaers et al. [33]. These articles achieve an AUC of 0.99 on internal validation cohorts and above 0.90 on external cohorts. Indeed, many other approaches obtained AUC over 0.95 [31–46]. In 2021, the FDA approved the deployment of PaigeProstate [34], based on the MIL architecture of Campanella [32]. It was evaluated with three different external cohorts [47–49]. It showed a high performance (sensitivity of minimum 0.94 and specificity of minimum 0.93), focusing on a high sensitivity to avoid missing small areas of cancer. Other articles focused on the detection of more specific patterns, such as cribriform patterns with an accuracy of 0.88 and AUC of 0.8 [50–52] or perineural invasion with an AUC of 0.95 [36]. To evaluate the generalization of their model trained on biopsies for which they obtained an AUC of 0.96, Tsuneki et al. applied it to TUR-P (TransUrethral Resection of Prostate) biopsies with an AUC of 0.80. After fine-tuning their model with colon WSIs, performance on TUR-P biopsies increased to 0.85 AUC [53]. Wen fine-tuning their model with the TUR-P training cohort, they obtained an AUC of 0.9 for every testing cohort [54]. All the articles focusing on cancer detection are in Table 5.

Table 5. Papers focusing on cancer detection only.

<table>
<thead>
<tr>
<th>First Author, Year</th>
<th>DL Architecture</th>
<th>Training Cohort</th>
<th>IV Cohort</th>
<th>EV Cohort</th>
<th>Aim</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Litjens, 2016 [43]</td>
<td>Custom CNN</td>
<td>150 patients</td>
<td>75 patients</td>
<td>None</td>
<td>Cancer detection at the pixel level</td>
<td>AUC: 0.99</td>
</tr>
<tr>
<td>Kwak, 2017 [44]</td>
<td>Custom CNN</td>
<td>162 TMAs</td>
<td>491 TMAs</td>
<td>None</td>
<td>Cancer detection at the sample level</td>
<td>AUC: 0.974</td>
</tr>
<tr>
<td>Kwak, 2017 [45]</td>
<td>Custom CNN</td>
<td>162 TMAs</td>
<td>185 TMAs</td>
<td>None</td>
<td>Cancer detection at the sample level</td>
<td>AUC: 0.95</td>
</tr>
<tr>
<td>Campanella, 2018 [46]</td>
<td>ResNet34 and VGG11-BN (MIL)</td>
<td>12610 WSIs</td>
<td>1824 WSIs</td>
<td>None</td>
<td>Cancer detection at the WSI level</td>
<td>AUC: 0.98</td>
</tr>
<tr>
<td>Campanella, 2019 [32]</td>
<td>RNN (MIL) &amp; ResNet34 (MIL)</td>
<td>12 132 WSIs</td>
<td>1784 WSIs</td>
<td>12 727 WSIs</td>
<td>Cancer detection at the WSI level</td>
<td>AUC: 0.991 (IV) &amp; 0.932 (EV)</td>
</tr>
<tr>
<td>García, 2019 [55]</td>
<td>VGG19</td>
<td>6195 glands (from 35 WSIs)</td>
<td>5-fold cross val</td>
<td>None</td>
<td>Malignancy gland classification</td>
<td>AUC: 0.889</td>
</tr>
<tr>
<td>Singh, 2019 [50]</td>
<td>ResNet22</td>
<td>749 WSIs</td>
<td>3-fold cross val</td>
<td>None</td>
<td>Cribriform pattern detection at the tile level</td>
<td>ACC: 0.88</td>
</tr>
<tr>
<td>Jones, 2019 [56]</td>
<td>ResNet50 &amp; SqueezeNet</td>
<td>1000 tiles from 10 WSIs</td>
<td>200 tiles from 10 WSIs</td>
<td>70 tiles from unknown number of WSIs</td>
<td>Malignancy detection at the tile level</td>
<td>ACC: 0.96 (IV) &amp; 0.78 (EV)</td>
</tr>
<tr>
<td>Authors, Year</td>
<td>Model</td>
<td>Dataset</td>
<td>Magnification</td>
<td>TMA Classification</td>
<td>Other Results</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-------</td>
<td>---------</td>
<td>---------------</td>
<td>-------------------</td>
<td>--------------</td>
<td></td>
</tr>
<tr>
<td>Duong, 2019 [31]</td>
<td>ResNet50 and multiscale embedding</td>
<td>602 TMAs</td>
<td>303 TMAs</td>
<td>None</td>
<td>TMA classification using only ×10 magnification: AUC: 0.961; TMA classification with multi-scale embedding: AUC: 0.971</td>
<td></td>
</tr>
<tr>
<td>Raciti, 2020 [34]</td>
<td>PaigeProstate Pre-trained [32]</td>
<td>232 biopsies</td>
<td></td>
<td></td>
<td>Malignancy classification at the WSI level: Sens: 0.96, Spec: 0.98; Improvement of pathologist’s classification: Sens: 0.738 to 0.900, Spec: 0.966 to 0.952</td>
<td></td>
</tr>
<tr>
<td>Han, 2020 [38]</td>
<td>AlexNet</td>
<td>286 WSIs from 68 patients</td>
<td>Leave one out cross val</td>
<td>None</td>
<td>Cancer classification at the WSI level: AUC: 0.98</td>
<td></td>
</tr>
<tr>
<td>Ambrosini, 2020 [51]</td>
<td>Custom CNN</td>
<td>128 WSIs</td>
<td>8-fold cross val</td>
<td>None</td>
<td>Cribriform pattern detection for biopsies: $\frac{\text{AUC}}{\text{F1}}: 1.0$</td>
<td></td>
</tr>
<tr>
<td>Bukhari, 2021 [57]</td>
<td>ResNet50</td>
<td>640 WSIs</td>
<td>162 WSIs</td>
<td>None</td>
<td>Cancer/hyperplasia detection at the tile level: Sens: 0.99, Specificity: 0.995</td>
<td></td>
</tr>
<tr>
<td>Pinckaers, 2021 [33]</td>
<td>ResNet-34 (MIL) Streaming CNN</td>
<td>5209 biopsies</td>
<td>535 biopsies</td>
<td>205 biopsies</td>
<td>Cancer detection for biopsies: AUC: 0.99 (IV) &amp; 0.799 (EV); AUC: 0.992 (IV) &amp; 0.902 (EV)</td>
<td></td>
</tr>
<tr>
<td>Perincheri, 2021 [49]</td>
<td>Paige Prostate Pre-trained [34]</td>
<td>1876 biopsies from 116 patients</td>
<td></td>
<td></td>
<td>Paige classification evaluation: Sens: 0.99, Spec: 0.93; 110/118 patients were correctly classified</td>
<td></td>
</tr>
<tr>
<td>Da Silva, 2021 [48]</td>
<td>PaigeProstate Pre-trained [34]</td>
<td>600 biopsies from 100 patients</td>
<td></td>
<td></td>
<td>Malignancy classification for biopsies: Sens: 0.99, Spec: 0.93; Malignancy classification for patients: Sens: 1, Spec: 0.78</td>
<td></td>
</tr>
<tr>
<td>Raciti, 2022 [47]</td>
<td>PaigeProstate Pre-trained [34]</td>
<td>610 biopsies</td>
<td></td>
<td></td>
<td>Cancer detection for patients: Sens: 0.974, Spec: 0.948, AUC: 0.99</td>
<td></td>
</tr>
<tr>
<td>Krajnansky, 2022 [58]</td>
<td>VGG16-mode</td>
<td>156 biopsies from 262 WSIs</td>
<td>Ten biopsies from 87 WSIs</td>
<td>None</td>
<td>Malignancy detection for biopsies: FROC: 0.944; Malignancy detection for patients: AUC: 1</td>
<td></td>
</tr>
<tr>
<td>Tsuneki, 2022 [53]</td>
<td>EfficientNetB1 pre-trained on colon</td>
<td>1182 needle biopsies</td>
<td>1244 TURP biopsies</td>
<td>500 needle biopsies</td>
<td>Cancer detection for classic and TURP biopsies: AUC: 0.967 (IV) &amp; 0.987 (EV) AUC (TURP): 0.845</td>
<td></td>
</tr>
<tr>
<td>Tsuneki, 2022 [54]</td>
<td>EfficientNetB1 pre-trained on ImageNet</td>
<td>1060 TURP biopsies</td>
<td>500 needle biopsies, 500 TURP</td>
<td>768 WSIs</td>
<td>Cancer detection in classic and TURP biopsies: AUC: 0.885 (IV) TURP</td>
<td></td>
</tr>
</tbody>
</table>

*Note: AUC = Area Under the Curve, Sens = Sensitivity, Spec = Specificity, TMA = Tissue Microarray, WSI = Whole Slide Imaging.*
When estimating Gleason grading, many papers only focused on classifying tiles or small regions like TMAs by taking advantage of classical CNN architectures trained on large datasets of natural images such as ImageNet [60]. In that context, tiles were encoded into features which corresponded to the input data for classification [21,38,61–66]. One of the first papers in the field used a cohort of 641 TMAs, obtaining a quadratic Cohen Kappa of 0.71 [21]. Similarly, Kott et al. obtained an AUC of 0.82 [62]. A few articles directly addressed the localization at the pixel level of Gleason patterns [28,29,41–71]. Performances vary between IoU of 0.48 [70,71] to IoU around 0.7 [28,29,68]. Dice score of 0.74 [69], quadratic Cohen kappa of 0.854 [41], sensitivity of 0.77 and specificity of 0.94 [67] (see Table 6). Adding epithelium detection greatly improved performance when properly segmenting areas depending on Gleason grades (gain of 0.07 in mean IoU) [29]. Many articles used the same pipeline to group Gleason scores according to ISUP recommendations [35,36,52,72,73]. These studies relied on the annotation of glands according to their Gleason pattern. WSIs were then split into tiles, and every single tile was classified according to the majority Gleason pattern (GP). Once all tiles were classified, heatmaps were generated, and a classifier was trained to properly aggregate ISUP grade groups or at least differentiate high from low grade [74]. Two algorithms based on this pipeline are now commercially available: (i) IBEX [36] (AUC of 0.99 for cancer detection, AUC of 0.94 for low/high-grade classification); (ii) DeepDx [72] (quadratic kappa of 0.90). DeepDx was further evaluated on an external cohort by Jung et al. [75] (kappa of 0.65 and quadratic kappa of 0.90). A third algorithm capable of Gleason grading (kappa of 0.77) was commercialized by Aiforia [42]. Another important milestone in the field was the organization and the release of the PANDA challenge focusing on Gleason grading at the WSI level without gland annotations [76]. This is an incredibly large cohort (around 12,000 biopsies of 3500 cases) publicly available, including slides from 2 different locations and external validation from 2 other sites. Best algorithms reached a quadratic Kappa of 0.85 on external validation datasets. One of the goals of Gleason grading algorithms is the potential decrease of inter-observer variability for pathologists using the algorithm. Some algorithms already have better Kappas than a cohort of pathologists compared to the ground truth [21,41,77].
Table 6. Articles focusing on Gleason grading.

<table>
<thead>
<tr>
<th>First Author, Year Reference</th>
<th>DL Architecture</th>
<th>Training Cohort</th>
<th>IV Cohort</th>
<th>EV Cohort</th>
<th>Aim</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Källén, 2016 [63]</td>
<td>OverFeat</td>
<td>TCGA</td>
<td>10-fold cross val</td>
<td>213 WSIs</td>
<td>GP tile classification</td>
<td>ACC: 0.81</td>
</tr>
<tr>
<td>Jimenez Del Toro, 2017 [74]</td>
<td>GoogleNet</td>
<td>141 WSIs</td>
<td>47 WSIs</td>
<td>None</td>
<td>Classify WSIs with a majority GP</td>
<td>ACC: 0.89</td>
</tr>
<tr>
<td>Arvaniti, 2018 [21]</td>
<td>MobileNet + classifier</td>
<td>641 TMA</td>
<td>245 TMA</td>
<td>None</td>
<td>TMA grading</td>
<td>qKappa: 0.71/0.75 (0.71 pathologists)</td>
</tr>
<tr>
<td>Poojitha, 2019 [64]</td>
<td>CNNs</td>
<td>80 samples</td>
<td>20 samples</td>
<td>None</td>
<td>GP estimation at tile level (GP 2 to 5)</td>
<td>F1: 0.97</td>
</tr>
<tr>
<td>Nagpal, 2019 [73]</td>
<td>InceptionV3</td>
<td>1159 WSIs</td>
<td>331 WSIs</td>
<td>None</td>
<td>GG classification</td>
<td>ACC: 0.7</td>
</tr>
<tr>
<td>Silva-Rodriguez, 2020 [52]</td>
<td>Custom CNN</td>
<td>182 WSIs</td>
<td>5-fold cross val</td>
<td>703 tiles from 641 TMA [21]</td>
<td>GG at the tile level</td>
<td>qKappa: 0.81 (0.77 with [21] method)</td>
</tr>
<tr>
<td>Otalora, 2021 [65]</td>
<td>MobileNet-based CNN</td>
<td>641 TMA, 255 WSI</td>
<td>245 TMA, 46 WSI</td>
<td>None</td>
<td>GG classification</td>
<td>wkappa: 0.52</td>
</tr>
<tr>
<td>Hammouda, 2021 [78]</td>
<td>CNNs</td>
<td>712 WSIs</td>
<td>96 WSIs</td>
<td>None</td>
<td>GP at the tile level</td>
<td>F1: 0.76</td>
</tr>
<tr>
<td>Marini, 2021 [66]</td>
<td>Custom CNN</td>
<td>641 TMA, 255 WSI</td>
<td>245 TMA, 46 WSI</td>
<td>None</td>
<td>GS at TMA level</td>
<td>qKappa: 0.81</td>
</tr>
<tr>
<td>Marron-Esquível, 2023 [77]</td>
<td>Dense-Net121</td>
<td>15,020 patches</td>
<td>102,324 patches (PANDA + fine-tuning)</td>
<td>2612 patches</td>
<td>Tile-level GP classification</td>
<td>qKappa: 0.826</td>
</tr>
<tr>
<td>Ryu, 2019 [72]</td>
<td>DeepDx Prostate</td>
<td>1133 WSIs</td>
<td>700 WSIs</td>
<td>None</td>
<td>GG classification</td>
<td>qKappa: 0.907</td>
</tr>
<tr>
<td>Karimi, 2019 [61]</td>
<td>Custom CNN</td>
<td>247 TMA, 86 TMA</td>
<td>None</td>
<td>Malignancy detection at the tile level</td>
<td>Sens: 0.86, Spec: 0.85</td>
<td></td>
</tr>
</tbody>
</table>

Ryu, 2019 [72] Malignancy detection at the tile level
Sens: 0.86, Spec: 0.85
GP 3 vs. 4/5 at tile level
Sens: 0.82, Spec: 0.82
<table>
<thead>
<tr>
<th>Authors, Year</th>
<th>Model</th>
<th>WSIs</th>
<th>Malignancy detection at the WSI level</th>
<th>AUC</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nagpal, 2020 [35]</td>
<td>Xception</td>
<td>524</td>
<td>430</td>
<td>322</td>
<td>AUC: 0.981</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GG1-2 vs. GG3-5</td>
</tr>
<tr>
<td>Pantanowitz, 2020 [36]</td>
<td>IBEX</td>
<td>549</td>
<td>2501</td>
<td>1627</td>
<td>Cancer detection at the WSI level</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low vs. high grade (GS 6 vs. GS 7-10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GP3/4 vs. GP5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Perineural invasion detection</td>
</tr>
<tr>
<td>Ström, 2020 [37]</td>
<td>InceptionV3</td>
<td>6935</td>
<td>1631</td>
<td>330</td>
<td>Malignancy detection at the WSI level</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.986 (EV)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GG classification</td>
</tr>
<tr>
<td>Li, 2021 [39]</td>
<td>Weakly supervised VGG11bn</td>
<td>13,115</td>
<td>7114</td>
<td>79</td>
<td>Malignancy of slides</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low vs. high grade at the WSI level</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Acc: 0.927</td>
</tr>
<tr>
<td>Kott, 2021 [62]</td>
<td>ResNet</td>
<td>85</td>
<td>5-fold cross-val</td>
<td>None</td>
<td>Malignancy detection at the tile level</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ACC: 0.85 for fine-tuned detection</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GP classification at the tile level</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Spec: 0.94</td>
</tr>
<tr>
<td>Marginean, 2021 [79]</td>
<td>CNN</td>
<td>698</td>
<td>37</td>
<td>None</td>
<td>Cancer area detection</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Spec: 0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GG classification</td>
</tr>
<tr>
<td>Jung, 2022 [75]</td>
<td>DeepDx Prostate Pre-trained</td>
<td>252</td>
<td>98</td>
<td>None</td>
<td>Correlation with reference pathologist (pathology report comparison)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>qKappa: 0.904 (0.858)</td>
</tr>
<tr>
<td>Silva-Rodriguez, 2022 [40]</td>
<td>VGG16</td>
<td>252</td>
<td>98</td>
<td>None</td>
<td>Cancer detection at the tile level</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GS at the tile level</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GP at the tile level</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>qKappa: 0.655</td>
</tr>
<tr>
<td>Bulten, 2022 [76]</td>
<td>Evaluation of multiple algorithms (PANDA challenge)</td>
<td>10,616</td>
<td>545</td>
<td>741 patients (EV1) 330 patients (EV2)</td>
<td>GG classification</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>qKappa: 0.862 (EV1)</td>
</tr>
<tr>
<td>Li, 2018 [68]</td>
<td>Multi-scale U-Net</td>
<td>187 tiles from 17 patients</td>
<td>37 tiles from 3 patients</td>
<td>None</td>
<td>Segment Stroma, benign and malignant gland segmentation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Stroma, benign and GP 3/4 gland segmentation</td>
</tr>
<tr>
<td>Li, 2019* [29]</td>
<td>R-CNN</td>
<td>513</td>
<td>5 fold cross-val</td>
<td>None</td>
<td>Stroma, benign, low- and high-grade gland segmentation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Benign vs. GP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F1: 0.893 (IV) &amp; 0.835 (EV)</td>
</tr>
<tr>
<td>Bulten, 2019* [28]</td>
<td>U-Net</td>
<td>62</td>
<td>40</td>
<td>20</td>
<td>Benign vs. GP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F1: 0.893 (IV) &amp; 0.835 (EV)</td>
</tr>
</tbody>
</table>

IoU: Intersection over Union

Kappa: Cohen’s Kappa

Acc: Accuracy

Sens: Sensitivity

Spec: Specificity

F1: F1 Score
Deep learning was also used to predict clinical outcomes such as recurrence status, survival or metastasis \((n = 10\), see Table 7\) or to predict genomic signatures from WSIs \((n = 5\), see Table 8\). This is the most complex task to perform as no visible phenotypes are known by pathologists to make such decisions.

### 3.3.1. Clinical Outcome Prediction

When focusing on recurrence, AUC around 0.8 [80,81] and Hazard Ratios (HR) above 4.8 [42,82,83] were obtained. A couple of articles studied the probability of developing metastasis [84,85]. The first article aimed to study if a patient developed lymph node metastasis after treatment [84] within undisclosed time frame, achieving an AUC of 0.69. The second article focused on distant metastasis, obtaining AUCs of 0.779 and 0.728 for 5- and 10-year metastasis. By combining image features and clinical data, the performance was improved to reach an AUC of 0.837 for 5- and 0.781 for ten years [85]. Liu et al. proposed to detect if benign slides belonged to a man who had no cancer or one who had cancer but on other biopsies and reached 0.74 AUC [86]. In any case, DL allows to establish decent survival models with HR of 1.38 for Nagpal et al. [73], 1.65 for Leo et al. [8] and 7.10 for Ren et al. [73,87,88].
### Table 7. Articles focusing on clinical outcome prediction.

<table>
<thead>
<tr>
<th>First Author, Year Reference</th>
<th>DL Architecture</th>
<th>Training Cohort</th>
<th>IV Cohort</th>
<th>EV Cohort</th>
<th>Aim</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kumar, 2017 [81]</td>
<td>CNNs</td>
<td>160 TMAs</td>
<td>60 TMAs</td>
<td>None</td>
<td>Nucleus detection for tile selection</td>
<td>ACC: 0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Recurrence prediction</td>
<td>AUC: 0.81(DL) &amp; 0.59 (clinical data)</td>
</tr>
<tr>
<td>Ren, 2018 [83]</td>
<td>AlexNet + LSTM</td>
<td>271 patients</td>
<td>68 patients</td>
<td>None</td>
<td>Recurrence-free survival prediction</td>
<td>HR: 5.73</td>
</tr>
<tr>
<td>Ren, 2019 [88]</td>
<td>CNN + LSTM</td>
<td>268 patients</td>
<td>67 patients</td>
<td>None</td>
<td>Survival model</td>
<td>HR: 7.10 when using image features</td>
</tr>
<tr>
<td>Leo, 2021 [87]</td>
<td>Segmentation-based CNNs</td>
<td>70 patients</td>
<td>NA</td>
<td>679 patients</td>
<td>Cribriform pattern recognition</td>
<td>Pixel TPV: 0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Prognosis classification using cribriform area measurements</td>
<td>Pixel TNV: 0.79</td>
</tr>
<tr>
<td>Wessels, 2021 [84]</td>
<td>xse_ResNext34</td>
<td>118 patients</td>
<td>110 patients</td>
<td>None</td>
<td>LNM prediction based on initial RP slides</td>
<td>AUC: 0.69</td>
</tr>
<tr>
<td>Esteva, 2022 [85]</td>
<td>ResNet</td>
<td>4524 patients</td>
<td>1130 patients</td>
<td>None</td>
<td>Distant metastasis at five years (5Y) and ten years (10Y)</td>
<td>AUC: 0.837 (5Y) AUC: 0.781 (10Y)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Prostate cancer-specific survival</td>
<td>AUC: 0.765</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Overall survival at ten years</td>
<td>AUC: 0.652</td>
</tr>
<tr>
<td>Pinckaers, 2022 [82]</td>
<td>ResNet50</td>
<td>503 patients</td>
<td>182 patients</td>
<td>204 patients</td>
<td>Univariate analysis for DL predicted biomarker evaluation</td>
<td>OR: 3.32 (IV)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>HR: 4.79 (EV)</td>
</tr>
<tr>
<td>Liu, 2022 [86]</td>
<td>10-CNN ensemble model</td>
<td>9192 benign biopsies from 1211 patients</td>
<td>2851 benign biopsies from 297 patients</td>
<td>None</td>
<td>Cancer detection at the patient level</td>
<td>AUC: 0.727</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cancer detection at patient level from benign WSIs</td>
<td>AUC: 0.739</td>
</tr>
<tr>
<td>Huang, 2022 [80]</td>
<td>NA</td>
<td>243 patients</td>
<td>None</td>
<td>173 patients</td>
<td>Recurrence prediction at three years</td>
<td>AUC: 0.78</td>
</tr>
<tr>
<td>Sandeman, 2022 [42]</td>
<td>Custom CNN (AIForIA)</td>
<td>331 patients</td>
<td>391 patients</td>
<td>126 patients</td>
<td>Malignant vs. benign</td>
<td>AUC: 0.997</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Grade grouping</td>
<td>ACC: 0.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>wKappa: 0.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Outcome prediction</td>
<td>HR: 5.91</td>
</tr>
</tbody>
</table>


#### 3.3.2. Genomic Signatures Prediction

Three groups started to work on the inference of genomic signatures from WSIs with the assumption that morphological features can predict pathway signatures [89–91]. These exploratory studies found correlations between RNA prediction from H&E images and RNA-seq signature expressions (PC ranging from 0.12 to 0.74). Schaumberg et al. carefully selected tiles containing tumor tissue and abnormal cells to train a classifier to predict SPOP mutations, reaching an AUC of 0.86 [92]. Dadhania et al. also used a tile-
based approach to predict ERG-positive or negative mutational status, reaching around 0.8 AUC [93].

Table 8. Articles focusing on genomic signatures prediction.

<table>
<thead>
<tr>
<th>First Author, Year Reference</th>
<th>DL Architecture</th>
<th>Training Cohort</th>
<th>IV Cohort</th>
<th>EV Cohort</th>
<th>Aim</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schaumberg, 2018 [92]</td>
<td>ResNet50</td>
<td>177 patients</td>
<td>None</td>
<td>152 patients</td>
<td>SPOP mutation prediction</td>
<td>AUC: 0.74 (IV)</td>
</tr>
<tr>
<td>Schmauch, 2020 [89]</td>
<td>HE2RNA</td>
<td>8725 patients (Pan-cancer)</td>
<td>5-fold cross val</td>
<td>None</td>
<td>Prediction of gene signatures specific to prostate cancer</td>
<td>PC: 0.18 (TP63) 0.12 (KRT8 &amp; KRT18)</td>
</tr>
<tr>
<td>Chelebian, 2021 [91]</td>
<td>CNN from [37] fine-tuned</td>
<td>Pre-trained ([37])</td>
<td>7 WSIs</td>
<td>None</td>
<td>Correlation between clusters identified with AI and spatial transcriptomics</td>
<td>No global metric</td>
</tr>
<tr>
<td>Dadhania, 2022 [93]</td>
<td>MobileNetV2</td>
<td>261 patients</td>
<td>131 patients</td>
<td>None</td>
<td>ERG gene rearrangement status prediction</td>
<td>AUC: 0.82 to 0.85 (depending on resolution)</td>
</tr>
<tr>
<td>Weitz, 2022 [90]</td>
<td>NA</td>
<td>278 patients</td>
<td>92 patients</td>
<td>None</td>
<td>CCP gene expression prediction</td>
<td>PC: 0.527</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>BRICD5 expression prediction</td>
<td>PC: 0.749</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SPOPL expression prediction</td>
<td>PC: 0.526</td>
</tr>
</tbody>
</table>


3.4. Risk of Bias Analysis

A Risk of Bias (RoB) analysis was performed for every article. Details are described in Supplementary Table S2. Results are summarized in Figure 3. Missing criteria were categorized as high risk, and partially addressed criteria (e.g., only half the dataset is publicly available) were considered as intermediate risk. Articles that validated existing algorithms or focused on prediction algorithms where ground truth was not defined by people were classified as Not Applicable (NA). This analysis particularly outlines the lack of publicly available code (and hyperparameters) and data. External cohorts for validation are often not addressed. More efforts could be provided for model explainability and dealing with imbalanced data. Indeed, this is a common difficulty in pathomics that biases training and evaluation.
4. Discussion

4.1. Summary of Results

This review article provided a systematic search following PRISMA guidelines using multiple keywords in PubMed and Embase to identify studies related to DL for PCa in pathomics. The databases and the chosen keyword might represent limitations to this systematic review. Along with breast and colorectum, prostate cancer is the organ most explored by AI. It results in a high number of publications increasing every year (see Figure 4). Among them, nine articles focused on pre-processing steps, which are key to having a robust model, namely using normalization, quality assessment and smart tile selection. It is a rather new subject (see Figure 4), and its investigation is still ongoing, with new methods to improve how input data is handled. Among the 77 articles in this review, most algorithms \((n = 50)\) were developed for the diagnosis of cancer and the Gleason grading. There are many different algorithms, and they tend to suffer the same biases (see Supplementary Figure S1D,E). It is, therefore, harder to see what the benefits from the models are compared to the others in their respective tasks. For detection purposes, classification algorithms were more common than segmentation algorithms. The former provided heat maps giving information on the location, while the latter provided precise information at the pixel level. For the diagnosis, the AUC comprised between 0.77 and 0.99 with many papers \((n = 16)\) with AUC above 0.95, favoring the use in routine activity.

The most performing algorithms used the MIL paradigm, allowing some explainability, and were trained on a large number of images. Algorithms for Gleason grading were less performing, with a quadratic Cohen Kappa comprised between 0.60 and 0.90. However, the definition of ground truth in Gleason grading suffers from high inter-observer variability that renders training less reliable. The prediction was less explored, with few articles approaching the prognosis (recurrence or metastasis development, for instance) or genomic alterations, but the interest in these investigations is increasing (see Figure 4). The AUC was comprised between 0.59 and 0.86. Prediction studies aim at correlating morphological patterns to a prediction that could enable the discovery of new patterns and help in patient personalized treatment. However, more robust studies, properly designed and validated, are needed to validate this assumption.
4.2. Biases

To evaluate the biases in our review, we adapted the QUADAS-2 tool to all the studies mentioned. Several biases can be noted in the studies at the image or algorithm level. Indeed, there is no standardization on the staining protocols, which translates to WSIs. This bias could be overcome using normalization such as Vahadane [94] or Macenko methods [95] or GANs (Generative Adversarial Networks) [24] or color augmentation [32,36,61,65,66,73,96], but most articles do not use image normalization to overcome this bias (50 out of 77 do not see Figure 3). In addition, most scanners have their proprietary format, which can impair the generalizability of the algorithms. Also, the brand of the scanner impacts the scanning process, potentially decreasing algorithms’ performance when trained on different brands. It is important to develop a strategy for testing algorithms with different scanners as proposed by the UK National Pathology Imaging Co-operative (NPIC). When working on diagnosis algorithms, ground truth is biased by inter-observer variability, especially in Gleason grading studies [21,41,77]. It is important to have multiple experts provide ground-truth annotation to not follow just one pathologist’s judgement, but it is not always the case (only 32 out of 77 articles do it, Figure 3).

Furthermore, there are a few general biases in the performance evaluation of algorithms. An imbalanced dataset will induce a bias if the issue is not addressed and not all articles consider it (only 41 out of 77, Figure 3). It is possible to use only a balanced dataset or DL techniques to reduce the impact of imbalance. A model should be trained multiple times to ensure its reproducibility. The performance average of all these trainings should be considered as the performance metric of the model. However, very few articles include confidence intervals on their metrics, which are yet key to evaluating the model in the existing state of the art. Less than half the articles include external validation cohorts (24 out of 77, Figure 3), but they are necessary to ensure that the evaluated model(s) is (are) not performing well only on the training WSIs, and that is also where normalization or color augmentation during training becomes crucial.

4.3. Limits of AI

There are limitations inherent to pathomics. Contrary to radiologic images, WSI has to be divided into tiles. Most classification algorithms must have a fixed input size, generally a defined number of tiles. It means that a subset of the slide has to be selected, and
the heterogeneous aspect of the tumor might not be considered. This is also affected by the choice of magnification under which the WSIs is tiled [31]. A few articles focus on new methodologies to handle this type of data [82,97]. There are also articles that suggest a smart tile selection to use more informative data and reduce computational time [25]. The particularity in PCa pathology is that the main type of images are biopsies that contain a low percentage of tissue. It can be interesting to include multiple biopsies of patients to increase the number of tiles available for training. However, the way these different biopsies are given to algorithms has to be considered. At the very least, they have to be split into the same datasets (training or testing). Otherwise, bias could be induced in the study.

Ethical regulation makes access to data difficult [98]. The existence of challenges (e.g., PANDA, Prostate cANcer graDe Assessment) is a very good way to provide data to many researchers. It also facilitates collaborations on model development. It is necessary to be able to reproduce results, which is limited by the lack of publicly available cohorts (21 out of 77 used shared data, Figure 3). However, few publications shared completely the methodology and their code with consequences on the reproducibility of the model, hindering a proper comparison of usefulness and improvement of new algorithms (only 15 out of 77 shared it, Figure 3). The popularity of AI in these last years has also increased the number of models and data to be compared, computed, and stored. This has an economic but also environmental cost that needs to be addressed [99]. Computational costs can be reduced by using more frugal but efficient architectures. Transfer learning can also reduce training time using previously developed and trained models that are fine-tuned to fit the studied data. The downfall is to conform to the input data format. Focusing on more efficient architectures and how to properly share methodology in the field are potential improvements to be found to develop long-term viable solutions. The focus should also be directed towards the explainability of the algorithms. If they are to be implemented in clinical setups, “blackbox” models will not be trusted by pathologists (Figure 3, 40 of 77 attempted some form of explainability).

4.4. Impact of AI in Routine Activity

Nonetheless, several AI tools are now available on the market for the diagnosis in routine activity for prostate cancer from different companies: Ibex, Paige, Deep Bio and AIFORIA, whose algorithms were recently published [36,42,49,75]. They can be approved for first or second-read applications if they are used before as a screening or after as a quality check for the pathologist diagnosis. The Galen™ Prostate from Ibex was the first to obtain CE under IVDR (In Vitro Diagnostic Medical Devices Regulation) in February 2023. The sensitivity and specificity of these products are very high when excluding slides with intermediate classification probability, also called undetermined categories. Indeed, a number of slides with undetermined categories are impacted by many parameters, such as pre-analytic conditions and the format of slides.

Consequently, performances depend on the site where the algorithms are deployed. In addition, their integration into routine activity supposes a digital workflow that is not widely available. Properly integrated into the workflow, it could help save time, but it is difficult to implement due to interoperability issues (integration in the Laboratory Information System (LIS) and the Image Management System (IMS)). An optimized integration supposes at least the automatized assignment of cases, the contextual launch, the cases prioritization, the visualization of heatmaps directly in the IMS and the integration of results directly in the report. Ethical considerations become an additional question when processing patient data, especially if sent to a cloud environment.
4.5. Multimodal Approach for Predictive Algorithms

Like other organs, the prediction of prostate cancer seems to be a more difficult question than detection. Indeed, the underlying assumption is that there exists a morphologic pattern in the images that can predict prognosis or genomic alteration. It is very likely that the answer is multifactorial and could benefit from multimodal approaches such as combining the WSI with radiologic, biological, and molecular data. The main challenge is properly combining all these data of different natures and evaluating the added value when combining them compared to the performance obtained by considering each separately.

5. Conclusions

In conclusion, DL has been widely explored in PCa, resulting in many pre-processing, diagnosis, or prediction publications. This systematic review highlights how DL could be used in this field and what significant improvements it could bring. It also included suggestions to reduce research biases in this field while outlining the inherent limits of these tools. Despite these limitations, PCa was one of the first organs to benefit from reliable AI tools that could already be used in routine activity for diagnosis purposes: cancer detection and Gleason grading... However, for predictive purposes, further studies are needed to improve the robustness of the algorithms, which could lead to more personalized treatment: prognosis, molecular alteration, etc.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics13162676/s1, Figure S1: Representation of Risk of Bias for subsets of selected articles; Table S1: Details of risk of bias for all articles; Table S2: Metrics found in articles of the review and their definition; Tables S3–S8: Extended Tables 3–8 found in this article.

Author Contributions: N.R. and S.-F.K.-J. came up with the idea of this review and conceptualized it. N.R., S.-F.K.-J., T.P. and P.A. then established a proper methodology to write this review. Articles search, data collection, bias analysis and figure creation were made by N.R.; Original drafting was written by N.R., T.P. and S.-F.K.-J.; Writing, reviewing, and editing was performed by O.A., R.D.C., R.B., D.L., N.R.-L., Z.-e.K., R.M. and K.B. All this work was supervised by T.P. and S.-F.K.-J. All authors have read and agreed to the published version of the manuscript.

Funding: N.R. is funded by ARED PERTWIN grant, EraPerMED project. T.P. is funded by a Chan Zuckerberg Initiative DAF grant (2019-198009).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in the Supplementary Materials.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References


**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.