Journal of Imaging Informatics in Medicine (2025) 38:3992–4014
https://doi.org/10.1007/s10278-025-01429-2

ORIGINAL PAPER

Automating Prostate Cancer Grading: A Novel Deep Learning Framework for Automatic Prostate Cancer Grade Assessment using Classification and Segmentation

Saidul Kabir[1] · Rusab Sarmun[1] · Rafif Mahmood Al Saady[2] · Semir Vranic[2] · M. Murugappan[4,5] · Muhammad E. H. Chowdhury[3]

Received: 9 October 2024 / Revised: 23 January 2025 / Accepted: 24 January 2025 / Published online: 6 February 2025

Abstract

Prostate Cancer (PCa) is the second most common cancer in men and affects more than a million people each year. Grading prostate cancer is based on the Gleason grading system, a subjective and labor-intensive method for evaluating prostate tissue samples. The variability in diagnostic approaches underscores the urgent need for more reliable methods. By integrating deep learning technologies and developing automated systems, diagnostic precision can be improved, and human error minimized. The present work introduces a three-stage framework-based innovative deep-learning system for assessing PCa severity using the PANDA challenge dataset. After a meticulous selection process, 2699 usable cases were narrowed down from the initial 5160 cases after extensive data cleaning. There are three stages in the proposed framework: classification of PCa grades using deep neural networks (DNNs), segmentation of PCa grades, and computation of International Society for Urological Pathology (ISUP) grades using machine learning classifiers. Four classes of patches were classified and segmented (benign, Gleason 3, Gleason 4, and Gleason 5). Patch sampling at different sizes (500×500 and 1000×1000 pixels) was used to optimize the classification and segmentation processes. The segmentation performance of the proposed network is enhanced by a Self-organized operational neural network (Self-ONN) based DeepLabV3 architecture. Based on these predictions, the distribution percentages of each cancer grade within the whole slide images (WSI) were calculated. These features were then concatenated into machine learning classifiers to predict the final ISUP PCa grade. EfficientNet_b0 achieved the highest F1-score of 83.83% for classification, while DeepLabV3+ architecture based on self-ONN and EfficientNet encoder achieved the highest Dice Similarity Coefficient (DSC) score of 84.9% for segmentation. Using the RandomForest (RF) classifier, the proposed framework achieved a quadratic weighted kappa (QWK) score of 0.9215. Deep learning frameworks are being developed to grade PCa automatically and have shown promising results. In addition, it provides a prospective approach to a prognostic tool that can produce clinically significant results efficiently and reliably. Further investigations are needed to evaluate the framework's adaptability and effectiveness across various clinical scenarios.

Keywords ISUP grading · Prostate cancer · Artificial intelligence · Deep learning · Classification · Segmentation

✉ M. Murugappan
m.murugappan@kcst.edu.kw

✉ Muhammad E. H. Chowdhury
mchowdhury@qu.edu.qa

[1] Department of Electrical and Electronic Engineering, University of Dhaka, Dhaka 1000, Bangladesh

[2] College of Medicine, QU Health, Qatar University, Doha, Qatar

[3] Department of Electrical Engineering, Qatar University, 2713 Doha, Qatar

[4] Intelligent Signal Processing (ISP) Research Lab, Department of Electronics and Communication Engineering, Kuwait College of Science and Technology, Block 4, Doha, Kuwait

[5] Department of Electronics and Communication Engineering, Vels Institute of Sciences, Technology, and Advanced Studies, Chennai, Tamil Nadu, India

Introduction and Related Works

Prostate cancer (PCa) ranks as the second most prevalent form of cancer among men worldwide [1]. Over 1.4 million individuals are annually affected by it, with the Globocan 2020 report indicating precisely 1,414,259 new cases and 375,304 deaths within a single year. When the disease is detected early in its asymptomatic stage and properly managed, patients can achieve up to a 98% long-term survival rate, as per medical guidelines [2–4]. The Gleason grading system is widely recognized as the most precise and frequently used method for assessing the histopathology of PCa, setting the benchmark for diagnosis, treatment, and prognosis of the condition [5]. The International Society for Urological Pathology (ISUP) revised the system in 2014, which was also acknowledged by the World Health Organization (WHO) [6]. The Gleason grading system defines five distinct grades for PCa [7]. Grade 1 is identified by closely packed, well-differentiated glands forming distinct nodules. In Grade 2, these well-differentiated glands are more spaced out, creating clearer nodules. Grade 3 features these glands as dispersed, separate entities that are still well-differentiated. Grade 4 is marked by poorly differentiated glands, merging, or forming sieve-like patterns, including structures resembling glomeruli. Grade 5 is characterized by the lack of glandular differentiation and the presence of necrosis. As the grade increases, it becomes more severe.

The Gleason grade is calculated by adding the primary (major) component of the tumor and any secondary (minor) component that makes up more than 5% of the tumor. If there is no secondary component present, the Gleason grade is determined by doubling the grade of the primary component. Therefore, the Gleason grade is fundamentally the sum of the primary and secondary pattern grades, such as 7 (3 + 4). In specific scenarios, if a tumor is composed of two hierarchical structures and the smaller portion is less than or equal to 5%, and this minor component has a lower grade, then the Gleason grade is found by adding the primary pattern grade to itself. However, if the minor component has a higher grade, the score is found by the sum of the primary and this higher secondary pattern grade. If the tumor has more than two hierarchical structures, the Gleason score is defined as the sum of the primary pattern grade and the grade of the most severe pattern observed. This procedure of Gleason patterns assessment has been a manual task for pathologists so far, which is not only a labor-intensive process but is also prone to inconsistencies due to relatively high interobserver variability. This issue is particularly severe given the complexities in the grading system.

Subsequently, a biopsy's Gleason score is converted to an ISUP grade on a scale from 1 to 5, based on both the primary and secondary growth patterns observed. By combining the scores of the primary and secondary patterns, a new Gleason grade is calculated, with scores like 3 + 3 being classified as ISUP grade 1, and so on, up to 4 + 5, 5 + 4, and 5 + 5, which are categorized as ISUP grade 5. Table 1 shows the different Gleason grades, based on majority and minority patterns, and their corresponding ISUP grades.

The Gleason grading system plays a crucial role as a prognostic tool in the management of PCa, assisting in the treatment decision-making process for patients. However, this system is challenged by the risk of either missing or over-grading cancer, potentially leading to unnecessary interventions. A notable limitation is the significant variability in grading assessments among pathologists, which could either lead to overtreatment or, in worse cases, the overlooking of a critical diagnosis, thereby affecting its utility for individual patient care [8].

The rising demand for prostate biopsies presents numerous challenges for pathology departments. A significant hurdle is managing the sheer volume of biopsy samples required for accurate diagnosis. Standard practice dictates collecting ten to twelve samples from each patient, leading to pathologists in the USA needing to assess over 10 million tissue samples annually. This immense workload not only escalates labor expenses but also impacts the operational efficiency of pathology departments [9].

With the advancement of technology, it is playing a more vital role in healthcare, particularly in diagnosing and treating diseases. Recently, there has been a significant surge in the adoption of computer-aided diagnosis (CAD) systems to support physicians in making precise decisions. Prompt and early identification is critical in the management and outcome of many diseases like PCa. In recent years, deep learning has experienced extensive adoption throughout the healthcare sector. Employing deep learning techniques in medicine has the potential to significantly

Table 1 Corresponding ISUP grades for gleason score distribution

Gleason score (majority + minority)	ISUP grade
Benign	0
3 + 3	1
3 + 4	2
4 + 3	3
4 + 4	4
3 + 5	4
5 + 3	4
4 + 5	5
5 + 4	5
5 + 5	5

reduce diagnostic inconsistencies. Deep learning models possess the capability to detect intricate patterns and characteristics in large datasets, enabling precise and consistent evaluations [10–13]. This approach can help reduce dependence on individual medical practitioners and decrease the variation in interpretations that are subjective between observers. In recent years, deep learning has seen widespread use across the healthcare industry [14–17].

Recent progress in the field has focused on improving prostate cancer (PCa) diagnosis efficiency through the exploration of various machine-learning-based PCa grading methods [18–21]. Nguyen et al. [22] explored classifying prostate cancer into three main categories: benign, grade 3, and grade 4, by examining prostate tissue through the use of texture-based features and color space analysis. Gorelick et al. [23] developed a method for evaluating similarity by extracting texture features from images and using a nearest neighbor classifier to categorize Gleason levels from 1 to 3. Waliszewski et al. [24] introduced a method based on fractal analysis for classifying adjacent Gleason groups, achieving a sensitivity of 81% and a specificity of 75%. These approaches demonstrate the diverse methodologies employed in leveraging machine learning toward the accurate grading of PCa. Nowadays, biomedical imaging has become crucial in the effective detection and treatment of cancer. Deep learning offers enhancements in Gleason grading by elevating precision and reducing errors attributable to human factors.

Artificial intelligence (AI) technologies have become extensively utilized in numerous histopathological settings, particularly in those addressing PCa [25, 26]. Developing a precise and dependable PCa grading algorithm requires overcoming numerous challenges, such as the lack of detailed labels, the varied morphology of slide images, the presence of high-resolution images with extensive empty spaces, and the introduction of artifacts due to variations in staining techniques [18]. Advances in hardware, datasets, and algorithmic methods have led to deep learning (DL) becoming an essential tool in the detection and categorization of PCa. Unlike models based on manually selected features, DL models can autonomously identify relevant features from whole-slide images and assess the severity of PCa. This eliminates the need for the labor-intensive process of feature selection, offering a more efficient method of accurate PCa grading [19]. Campanella et al. [27] analyzed 44,000 whole slide images (WSIs) from breast, skin, and prostate tissues to develop a deep learning system capable of differentiating cancerous from non-cancerous slides without the need for detailed pixel annotations. This method leverages a multiple-instance learning framework to create a detailed feature representation. It utilized a recurrent neural network (RNN) to synthesize these features and deliver the final diagnosis. Their approach's effectiveness was highlighted by attaining an impressive AUC (area under the curve) score of 0.986 in detecting PCa.

Strom et al. [9] introduced a system comprising two ensembles of convolutional neural networks (CNNs), each consisting of 30 models based on the InceptionV3 architecture pre-trained on the ImageNet dataset. The efficacy of their method was underscored by achieving an outstanding AUC (score of 0.986 in the detection of PCa. The training utilized 6682 WSIs, while an independent set of 1631 WSIs was used for testing, and an additional 330 WSIs were employed for external validation. In the binary classification task, the system showed remarkable precision, achieving an AUC of 0.997 for the independent test and 0.986 for the external validation. In the task of Gleason grading, the system reached a mean pairwise kappa statistic of 0.62, which falls within the inter-observer variability range observed among 23 pathologists (0.60–0.73) indicating human experts level performance. Marrón-Esquivel et al. [28] analyzed inter-observer discrepancies within a dataset of 80 WSIs annotated by five pathologists. They employed convolutional neural network (CNN) architectures to develop models aimed at reducing these discrepancies. The study found significant variability ($\kappa = 0.6946$) in pathologists' interpretations, with models achieving a kappa score of 0.826 ± 0.014 on the test set.

Arvaniti et al. [29] introduced a convolutional neural network (CNN) architecture for classifying tissue microarray (TMA) images into low (GS6-GS7) and high (GS8-GS10) Gleason scores. The dataset consisted of 895 TMA images, with 641 TMAs used for training the CNN. Model performance was assessed on a separate test set of 245 TMAs, which were annotated separately by two experienced pathologists. The authors found that CNN's agreements with the pathologists were 0.75 and 0.71, as measured by Cohen's quadratic kappa score. These levels of agreement were found to be on par with the inter-pathologist agreement of 0.71.

Another approach to identifying the Gleason score involves segmenting the cancerous tissue. Ing et al. [30] gathered 513 high-resolution images of primary PCa to evaluate the effectiveness of four convolutional neural networks (CNNs) in the semantic segmentation of tumors into high and low grades. Among these CNNs, U-Net achieved a precision of 0.885. Bulten et al. [31] explored the use of deep learning for the assessment of Gleason grading of prostate biopsies, addressing the issue of inter-observer variability that limits the Gleason score's utility for individual prostate cancer patients. They utilized 5759 biopsies of 1243 patients from Radboud University Medical Center, employing a semi-automatic labeling technique based on pathologists' reports. Validation against a consensus from expert pathologists on an independent set of 550 biopsies showed the system's high agreement with the reference standard (Cohen's kappa of 0.918) and superior performance in certain

assessments compared to both experienced pathologists and those in training. Singhal et al. [32] introduced a novel training strategy that focuses on learning domain-agnostic features. Utilizing 3741 core needle biopsies (CNBs) from two centers for training, the study evaluates the system's effectiveness across three patient cohorts. The results show high accuracy (89.4%) and excellent agreement (κquad of 0.92) on an internal test set, with similarly strong performance on external datasets, indicating the potential of the system to improve the accuracy and reliability of PCa grading and thereby enhance patient care. Although deep learning has shown potential in prostate cancer research, much of the work has been confined to basic classifications like benign versus malignant or distinguishing between Gleason grades 3 and 4.

Our approach integrates classification and segmentation, further refined by a machine learning-based classifier to derive a final Gleason or ISUP score based on the combined tissue type distributions. A novel segmentation architecture is also proposed combining self-ONN with the DeepLabV3 architecture. It also makes use of cross-validation for a thorough and accurate model evaluation. Recognizing the significance of image detail, we experimented with different patch sizes (500×500 and 1000×1000 pixels), focusing on the architectural rather than cellular patterns of prostate cancer, leading to notably enhanced results.

Major Contributions

The main contributions of this study are as follows:

1 A large dataset of prostate cancer biopsy images was investigated for the performance of patch-based classification and segmentation approaches.
2 A novel segmentation architecture combining DeepLabV3, and self-organizing neural networks (Self-ONN) is proposed for PCa segmentation.
3 An automated framework that integrates classification and segmentation techniques with a machine learning classifier accurately determines ISUP scores of prostate cancer (PCa) images is developed.
4 The study examined the effectiveness of classification and segmentation using patch sizes of 500×500 and 1000×1000 pixels.

The rest of the paper is structured as follows: the "Proposed Methodology" section offers a detailed discussion of the materials and methods utilized in our study. The "Proposed Method" section presents a comprehensive analysis of both the quantitative and qualitative performance of our proposed framework. Finally, the "Numerical Results" section summarizes key findings and concludes the paper by discussing future research.

Proposed Methodology

Dataset Description

The Prostate cancer grade assessment (PANDA) challenge dataset [33] was created in a collaboration between the Computational Pathology Group at Radboud University Medical Center and the Department of Medical Epidemiology and Biostatistics at Karolinska Institute and it remains the most extensive public collection of whole-slide images. It comprises 5160 hematoxylin and eosin (H&E) slides of core needle biopsies from Radboud and an additional 5456 cases from Karolinska hospitals, scanned at a magnification of $20\times$ and stored in TIFF format. To assist in the assessment of the Gleason score and ISUP grade, segmentation masks are also included in the same format as the images, to delineate the areas contributing to the Gleason score and ISUP grade determination. For the Radboud dataset, the segmentation masks included details about the various Gleason patterns present in the images, whereas the Karolinska dataset provided only information distinguishing between benign and malignant cases.

Table 2 shows the details of the dataset regarding the sources, number of cases of each grade, and the nature of the available data. Contrary to prior challenges, WSIs are used instead of small TMAs. Challenges were introduced in this dataset to establish the robustness of the learning process and increase complexity. Segmentation masks for all the cases were not provided and the provided masks were prone to inaccuracies, including both false positives and negatives. Both Gleason score labels and the ISUP grade labels were provided for all cases, with a focus on predicting the ISUP grade. The label accuracy did not reach the gold standard, due to the intrinsic challenges and subjective interpretation involved in examining slides, even for the most skilled experts. This level of inconsistency adds additional hurdles to the training of deep learning models, making it more challenging to develop highly accurate algorithms.

Data Preprocessing

The PANDA challenge dataset presents a unique set of challenges as it is quite a noisy dataset. Training models with such noisy data can impede learning processes and the generalization of models. A crucial initial step in utilizing this dataset effectively involves thorough data cleaning and preprocessing. This helps in selecting appropriate data for training while eliminating any misleading or noisy cases, thereby facilitating the development of a more reliable learning model. Specifically, the Radboud Dataset provides masks that detail the Gleason patterns, in contrast to the Karolinska dataset, which only offers masks that distinguish

Table 2 Details of the Prostate cancer grade assessment (PANDA) challenge dataset

		Radboud University Medical Center (number of cases)	Karolinska Institute (number of cases)
Number of samples	ISUP grade 0	967	1925
	ISUP grade 1	852	1814
	ISUP grade 2	675	668
	ISUP grade 3	925	317
	ISUP grade 4	768	481
	ISUP grade 5	973	251
	Segmentation mask values	0: Background (non-tissue) or unknown, 1: Stroma, 2: Benign, 3: Gleason 3, 4: Gleason 4, 5: Gleason 5	0: Background (non-tissue) or unknown, 1: Benign tissue, 2: Cancerous tissue

between benign and malignant areas. Examples of images and masks from the two sources are shown in Fig. 1.

Consequently, the Karolinska dataset proves inadequate for training that involves patch-based segmentation or classification due to its limited information. As previously mentioned, the Radboud dataset also suffers from the issue of noisy labels, which can adversely affect model training and performance. As a result of the PANDAS challenge, 1153 cases were identified in the Radboud dataset where the Gleason scores/ISUP grades derived from mask images did not match ground truth labels [34]. We aimed to eliminate these cases as well because removing them from their experiments improved their performance. Based on their approach, we checked the ground truth of all cases using the score derived from the mask images using the criteria outlined in Table 1. As a result of our investigation, we found 845 additional cases with incorrect Gleason score representations or incomplete mask data. These cases were also excluded. As a result of the challenge winners' findings, 463 cases had image markings, empty masks, or missing masks entirely, making them unsuitable for patch-level model training and evaluation. In consultation with two clinical experts, we ensured that the dataset processing would yield meaningful results, yield robust performance, and become an effective clinical tool in the future. Figure 2 shows some examples of cases which were excluded from the analysis as well as the types of cases excluded.

The dataset contains images and masks formatted as WSIs, which are high-resolution images of gigapixel scale. The size of these images surpasses the memory capacities of current GPUs and neural networks, making direct

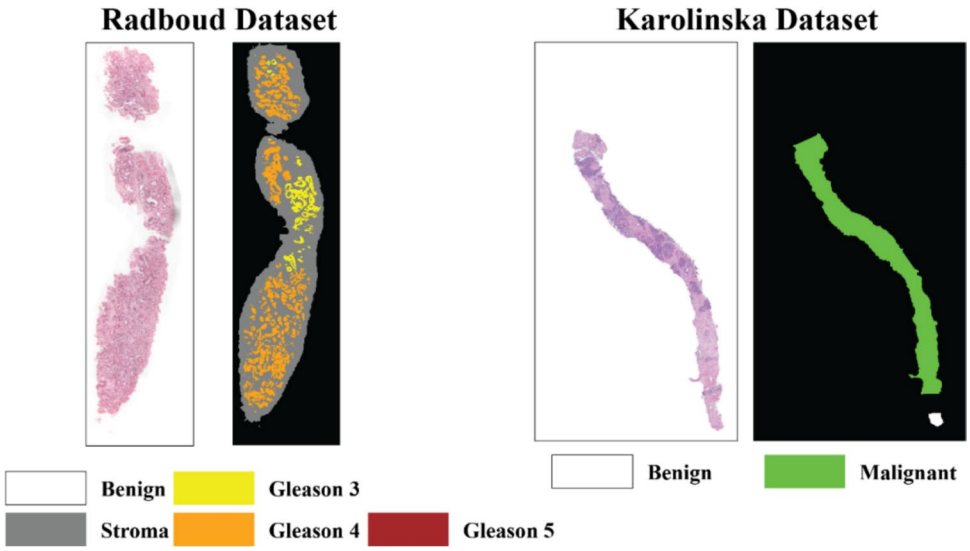

Fig. 1 Example image and corresponding mask from the two sources of the Prostate cancer grade assessment (PANDA) challenge dataset

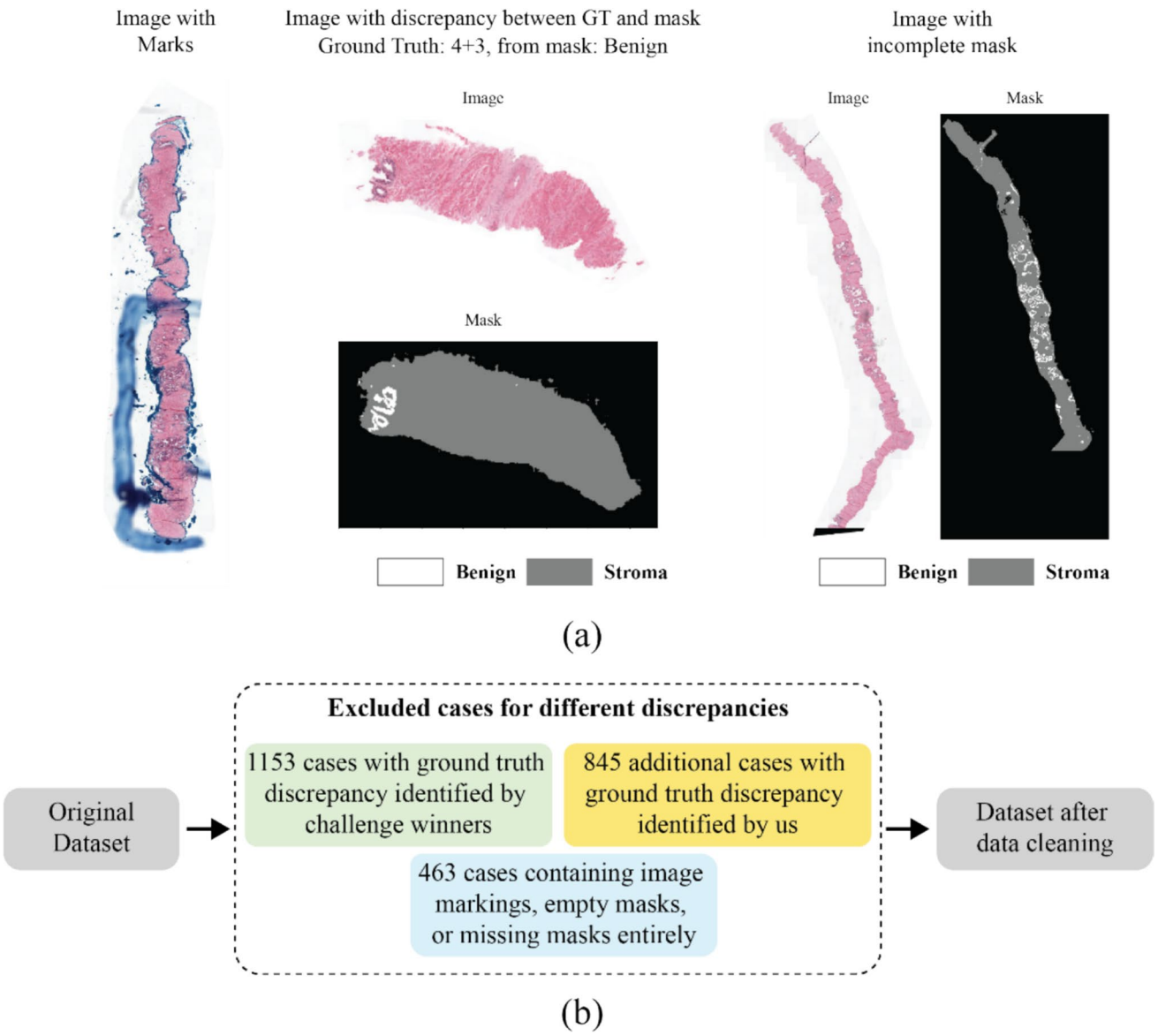

Fig. 2 **a** Examples of cases discarded due to various issues in the images or masks. **b** Flowchart showing the different types of images discarded due to discrepancy

processing infeasible. To address this challenge, a preprocessing step is employed. A well-established strategy to manage this issue involves patch sampling of the WSIs, where smaller sections of the images, known as patches, are extracted from the larger source images. These patches can then be fed into neural networks as input. This approach has become the predominant and most effective method for utilizing WSIs in deep learning applications [27, 29, 35]. Marrón-Esquivel et al. [28] worked with a patch size of 750×750 for classifying PCa patches. In this study, we experimented with both 500×500 and 1000×1000 patch sizes to compare performance and investigate the effect of patch size on pattern identification of PCa. For segmentation, patches were also created from masks in the corresponding sizes. In the classification process, a patch was labeled as benign if 100% of its tumor area was benign. A threshold of 50% overlap with the annotations was set for the extraction of patches from malignant areas, meaning a patch needed to contain at least 50% of a specific Gleason pattern to be classified under that Gleason class. For instance, a patch with over 50% Gleason 3 tumor was categorized as Gleason 3. Figure 3 shows examples of patches with corresponding classification labels and segmentation masks. To evaluate the performance of various models, a subject-wise fivefold cross-validation dataset was compiled using the patches from the selected cases.

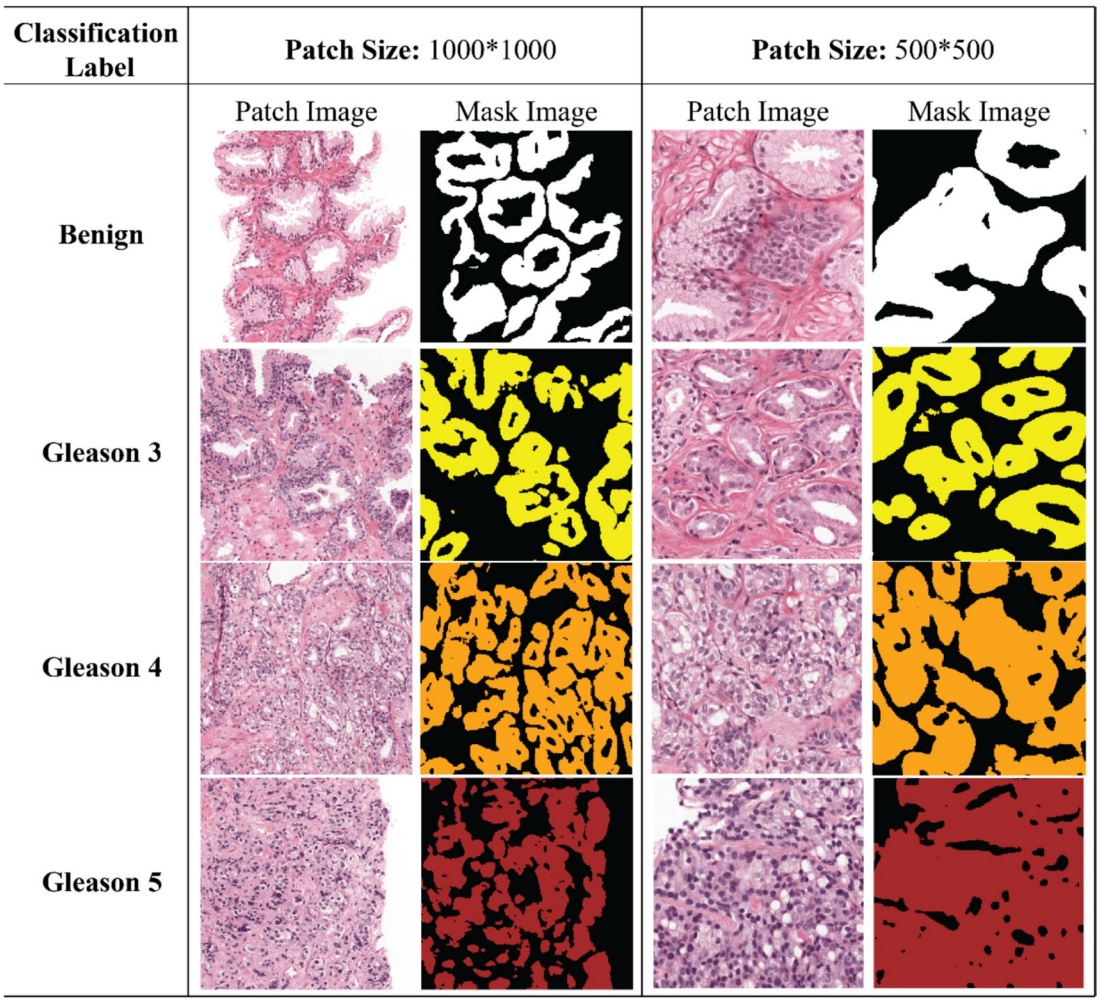

Fig. 3 Patch images and corresponding masks of different Gleason patterns

Data Augmentation

It is important to make sure the number of images is balanced by class when training a deep learning network. Generating augmented images is a common technique to balance datasets and increases the heterogeneity of the dataset. It prevents overfitting of the model and helps in generalization during learning [36, 37]. Different transformations such as horizontal flips, vertical flips, and 90-degree rotations were applied. Figure 4 shows examples of generated synthetic images using image augmentation techniques from authentic images, and Table 3 depicts the number of patch images

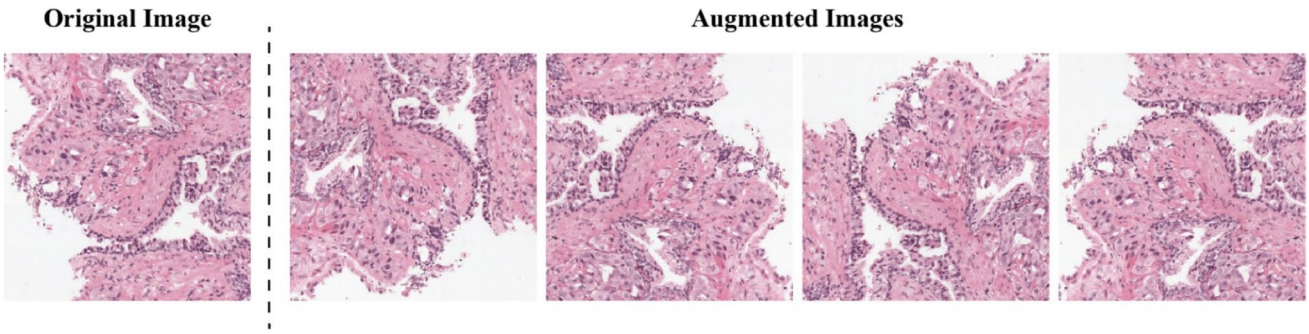

Fig. 4 Examples of augmented patch images generated from an authentic patch image

Table 3 Distribution of train, test and validation set for classification and segmentation with augmentation

Dataset		Number of patches created			
		Training		Validation	Test
		Before augmentation	After augmentation		
Classification (500×500)	Benign	71,534	168,141	9243	12,552
	Gleason 3	108,401	168,141	14,193	23,036
	Gleason 4	168,141	168,141	21,490	37,680
	Gleason 5	40,351	168,141	4392	6367
Classification (1000×1000)	Benign	8460	34,315	794	1225
	Gleason 3	22,072	34,315	2405	4442
	Gleason 4	34,315	34,315	3683	7799
	Gleason 5	7244	34,315	699	1186
Segmentation (500×500)		563,800	749,854	75,707	642,216
Segmentation (1000×1000)		175,249	321,552	18,269	154,650

generated for train, test, and validation set for the classification and segmentation problem.

Proposed Method

A novel automatic framework is introduced in this work for the prediction of ISUP score using hematoxylin–eosin-stained WSIs of PCa cases. The framework is composed of three key stages:

- Determining the distribution of various PCa grades using classification models.
- Determining the distribution of PCa grades via segmentation models.
- Determining the final ISUP grade using machine learning classifiers from a combination of the distributions obtained from both classification and segmentation.

To determine the Gleason score of a biopsy image, it is crucial to identify the majority and minority patterns. Two common approaches to achieve this are classification and segmentation of the patches generated from the samples. Initially, patches are created from the WSIs. In the classification method, models are trained to categorize patches into benign, Gleason 3, Gleason 4, and Gleason 5. After identification of the patches, the percentage of each class is calculated. For the segmentation approach, models are developed to conduct multilabel segmentation across these four classes (benign, Gleason 3, Gleason 4, and Gleason 5), and subsequently, the area occupied by each class in the predicted masks is quantified. Machine learning classifiers are then utilized to estimate the ISUP grade based on the class distribution of the samples. The efficacy of these classifiers is evaluated both on the distributions of classification and segmentation individually and on different combinations of distributions derived from both approaches. An overview of the methodology is shown in Fig. 5.

Classification Networks

The goal of the classification phase is to sort patch images based on whether they predominantly display a specific Gleason pattern. A typical approach for image classification involves fine-tuning pretrained models that have been initially trained on the ImageNet dataset. This process adjusts the model's weights using a smaller dataset to specialize it to a particular task. Using a pre-trained model is efficient because it leverages the model's pre-existing knowledge of general features learned from a broad dataset saving both the training time and resources [38]. To fit the model to the target task, its last fully connected layers are replaced, and it is fine-tuned at a reduced learning rate on the new dataset for gradual weight updates. CNN-based networks are renowned in image classification for their efficiency in detecting spatial patterns within images, thanks to their layered structure and the reuse of weights. This setup enables them to extract relevant features at different abstraction levels, making them ideal for visual recognition. Various standard CNN architectures have set benchmarks in image classification due to these characteristics. However, Vision Transformer models have recently made their mark, utilizing self-attention mechanisms to process both the immediate and broader contextual relationships within images. This quality allows Vision Transformers to surpass the capabilities of traditional CNNs. In our analysis of multiple deep learning models for this classification approach, we highlight the four models that showed comparable top performance, based on their classification efficacy. This study reports on the performance of three CNN-based models (DenseNet121, EfficientNet_b0, Inception_v3) and a Vision Transformer model on the dataset.

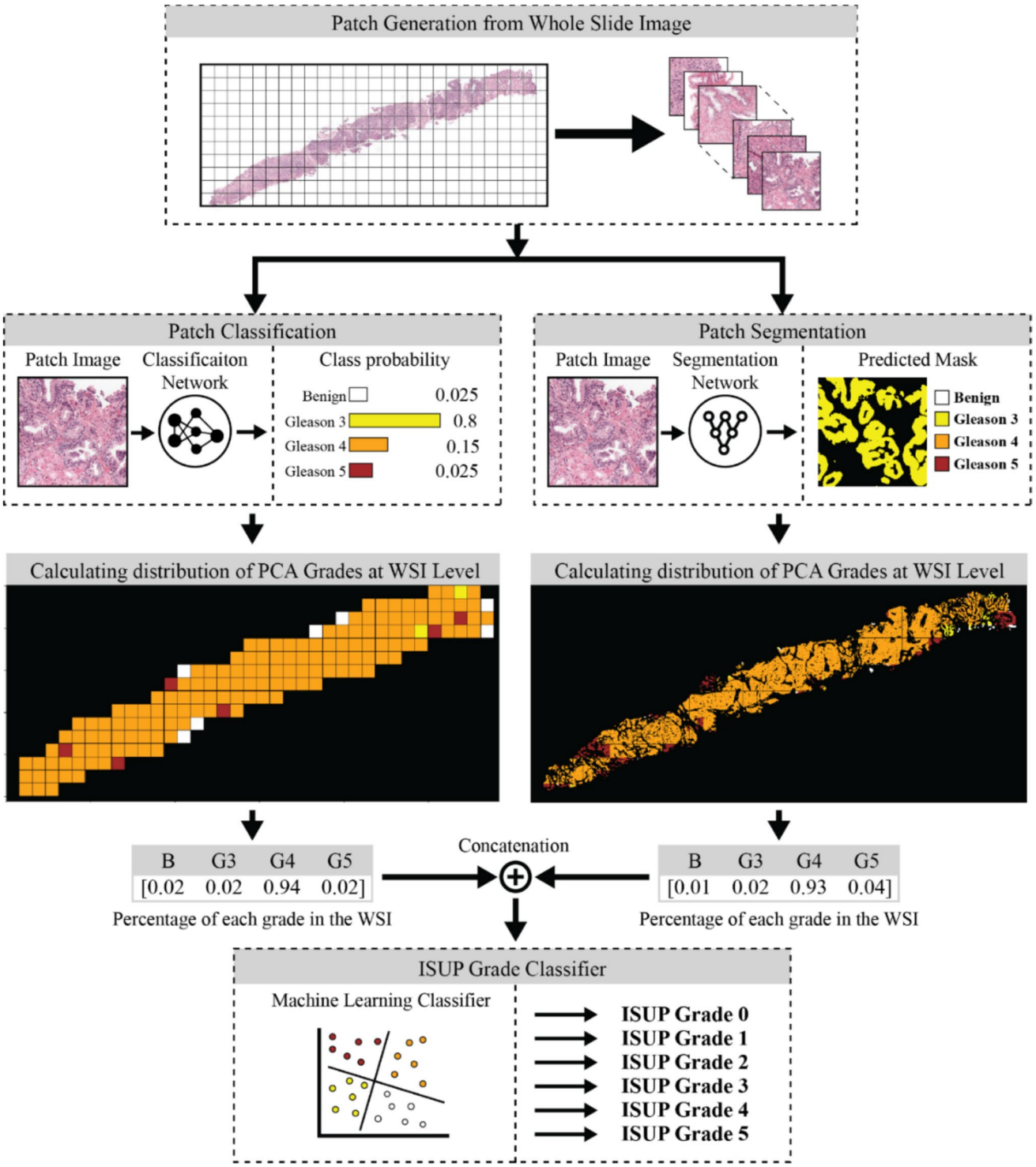

Fig. 5 An overview of the proposed framework for automatic assessment of prostate cancer grades combining both classification and segmentation techniques

DenseNet121

DenseNet121 [39] is a complex deep-learning framework known for its highly interconnected design, featuring numerous dense blocks where each layer is directly connected to every other layer in its block. This architecture was detailed in a study by Huang et al. (2017) in their work on densely connected convolutional networks. Each dense

block within DenseNet121 contains layers that share feature maps with all preceding layers, enhancing feature reuse and streamlining the flow of information. This setup not only improves training efficiency by enhancing gradient flow but also boosts the network's ability to represent and learn intricate patterns and features. The model incorporates bottleneck layers to reduce the number of features, employs batch normalization for stability, and uses ReLU activation functions for non-linearity. Transition layers are placed in between the dense blocks to reduce spatial dimensions, and the network concludes with global average pooling, a fully connected layer adjusted for the number of classes in the task, and a SoftMax activation function.

EfficientNet_b0

EfficientNet-B0 [40] presents a method for enlarging CNNs through compound scaling, which simultaneously optimizes network width, depth, and resolution. As the initial model in the EfficientNet series, it was crafted using neural architecture search techniques, resulting in a design that delivers outstanding accuracy and efficiency. EfficientNet is a family of CNN models designed for scalability and efficiency. The core idea behind EfficientNet is the use of compound scaling, which uniformly scales the network's depth, width, and resolution with a set of fixed scaling coefficients. This approach allows for balanced network growth, optimizing accuracy and efficiency. It employs a baseline network, EfficientNet-B0, developed through a neural architecture search that optimizes both accuracy and FLOPs (floating-point operations).

Inception_v3

Inception v3 [41] is an advanced CNN developed by Google, that evolves from earlier versions by incorporating key enhancements to boost its accuracy and efficiency. It leverages factorized convolutions and augments the inception modules to minimize parameter count without compromising on the model's depth or breadth. Additionally, it integrates label smoothing, a method that mitigates overfitting by tempering the certainty of the labels. These improvements enable Inception v3 to deliver exceptional performance in image classification tasks, proving its adeptness at processing complex visual information. The architecture strikes a deliberate balance between computational efficiency and the model's capabilities, establishing it as an effective instrument for diverse image-processing tasks.

Vision Transformer

The Vision Transformer (ViT) [42] architecture utilizes the transformer framework, traditionally employed in natural language processing, for image classification tasks. ViT employs self-attention mechanisms to grasp global interactions within visual data, offering a novel approach to understanding images. The Vision Transformer (ViT) begins by dividing the input image into a series of fixed-size patches. Each patch is linearly transformed into a token embedding, with positional embeddings added to imbue spatial context. The structure of ViT is based on a sequence of transformer encoder layers. Each layer comprises self-attention mechanisms and feed-forward networks. The self-attention enables the model to focus on different image areas, capturing intricate relationships over vast distances. The feed-forward networks apply nonlinear transformations to enhance these representations. Following the transformer encoders, the token embeddings feed into a classifier head, which uses linear layers and a softmax function to predict class probabilities. ViT leverages transformer technology to achieve notable success in image classification, offering a fresh approach to employing self-attention for visual tasks. However, its effective training typically demands significant computational power and substantial datasets.

Segmentation Networks

This segment aims to distinguish between various regions associated with different grades (benign, Gleason 3, Gleason 4, Gleason 5) using multiclass segmentation. In this approach, regions corresponding to each grade are identified and predicted as different values within a mask, effectively differentiating each grade. The U-Net architecture [43] has been designed for precisely segmenting biomedical images using CNNs and has shown excellent results in this field. Its distinctive U-shaped design features a symmetrical split into two main pathways: the encoder, which compresses, and the decoder, which decompresses. The encoder pathway is composed of multiple convolutional layers, each followed by a ReLU activation function and max pooling to gradually decrease the spatial dimensions of the feature maps while simultaneously increasing their depth or channel count. The architecture contains a bottleneck at the center of it which is a constricted segment outfitted with extra convolutional layers. Proceeding from the bottleneck, the decoder pathway expands the feature maps to their original spatial dimensions. A critical aspect of U-Net is its use of skip connections, which link feature maps from each encoder level directly to corresponding levels in the decoder, ensuring a cohesive fusion of high-level and detailed information into the decoder's output. A widely adopted method to enhance the architecture's effectiveness involves integrating advanced encoder architectures. DenseNet and EfficientNet enhance U-Net for image segmentation by using dense connections and optimized performance under limited resources, respectively. Both

architectures promote feature reuse, mitigate vanishing gradients, and enhance feature propagation, resulting in precise segmentation. In this study, we presented findings for both the standard U-Net and the U-Net framework enhanced by incorporating encoders from DenseNet and EfficientNet.

DeepLabV3 [44] and DeepLabV3 + [45] are sophisticated models for semantic segmentation, specifically designed to enhance the segmentation of objects at various scales and obtain more accurate boundaries. These models are significant expansions of the DeepLab series that employ deep convolutional neural networks for the purpose of high-resolution picture segmentation. Chen et al. [44] released DeepLabV3, which improves upon previous versions by integrating an atrous convolution technique. This strategy effectively expands the range of filters to capture context at several scales without compromising resolution. The model incorporates an atrous spatial pyramid pooling (ASPP) module at the network's end. This module examines a convolutional feature layer with filters at various sampling rates and effective fields-of-views. As a result, it captures objects and image context at multiple scales. DeepLabV3 + enhances the segmentation results of DeepLabV3 by incorporating a decoder module that specifically improves the accuracy of object boundaries. Efficientnet-b0 was used as encoder for this network.

Convolutional neural networks (CNNs) are limited by their homogenous, linear neuron structures, which do not fully capture the diversity of biological neural systems. Generalized operational perceptrons (GOPs) and operational neural networks (ONNs) address these shortcomings by offering heterogeneous and non-linear models. GOPs, inspired by biological processes, excel in complex scenarios where conventional models struggle. ONNs build on this concept by introducing a variety of operators per neuron, enhancing flexibility. They move beyond traditional linear convolutions by incorporating diverse operational units like nodal and pool operators. This innovation maintains core CNN principles of weight sharing and localized connectivity yet broadens the functional scope of the network layers. Figure 6 shows the difference between CNNs and ONNs.

We adapted the DeepLabV3 network by incorporating self organized operational neural networks (self-ONN) [46], which have demonstrated superior performance compared to traditional convolutional neural networks. All the CNN layers in both the DeepLabV3 and DeepLabV3 + networks were replaced by self-ONN layers to create our architecture. Figure 7 depicts the architecture of the self-ONN based DeepLabV3 + model. We also conducted comparisons between the networks to demonstrate the superiority of our self-ONN-based network.

ISUP Grade Classifier

The concluding phase of the framework employs a machine learning classifier to determine the ISUP grade of a biopsy, leveraging the pattern or grade distributions derived from the classification and segmentation of patches. The percentage of all the grades of malignant tissue (benign, Gleason 3, Gleason 4, and Gleason 5) in a WSI were calculated from both the classification and segmentation approach. For the classification approach, the proportion of each class within the total number of patches was determined. For the segmentation approach, the proportion of the area covered by each

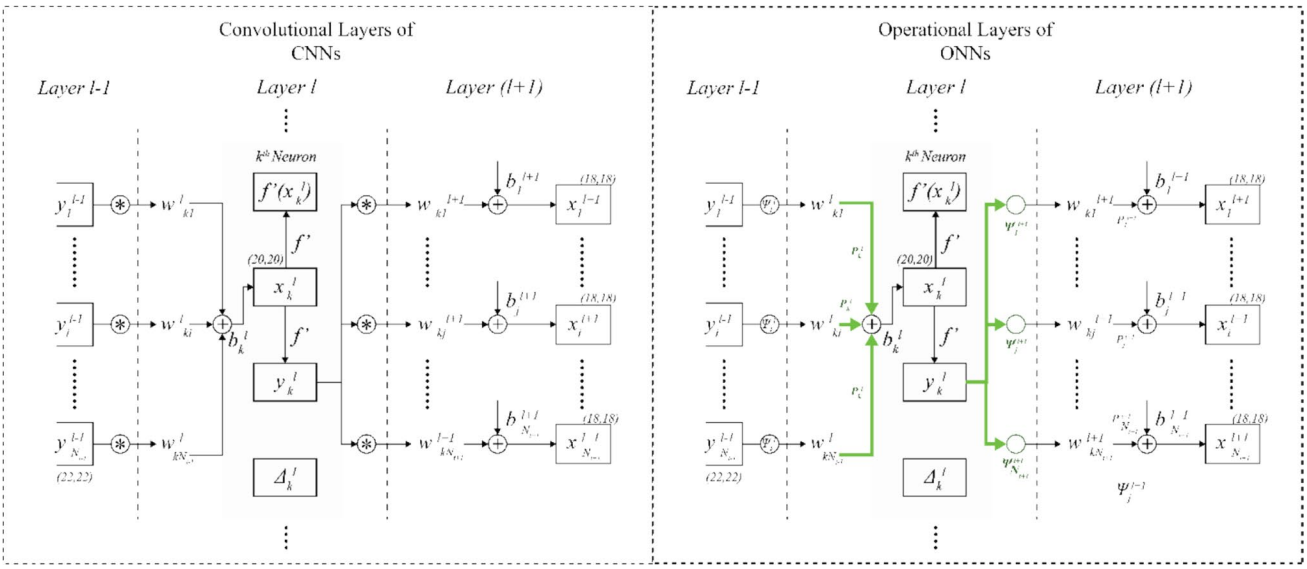

Fig. 6 Comparison of operations of convolutional layers of CNNs and operational layers of ONNs

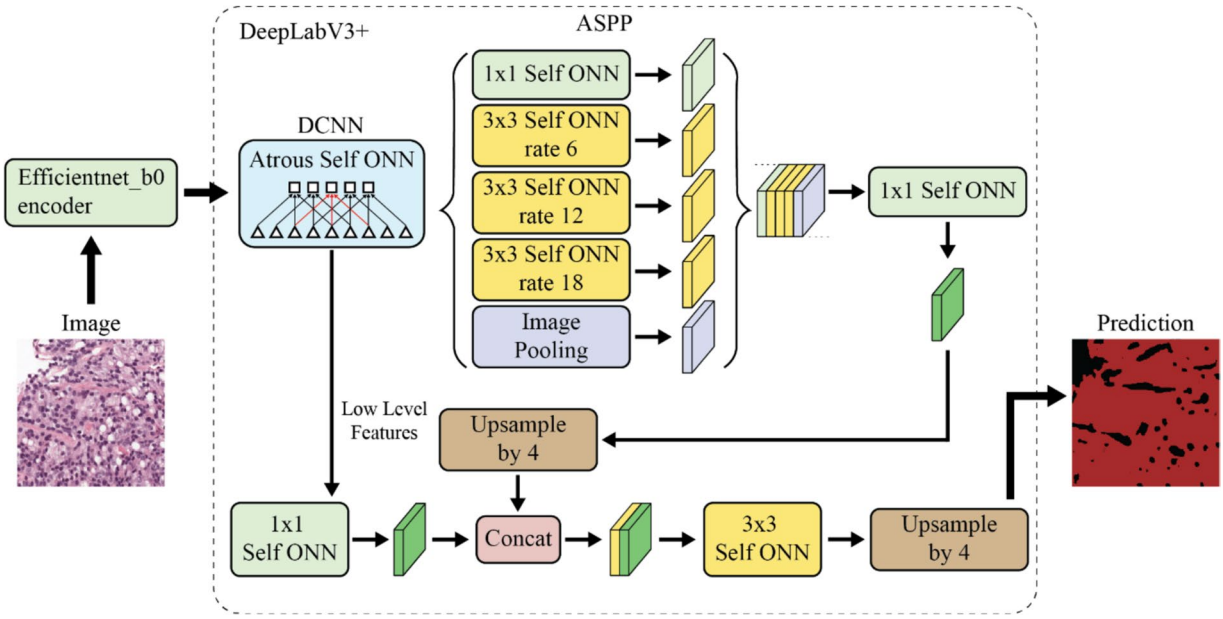

Fig. 7 Architecture of DeepLabV3 + [45] customized with self-ONN layers

class within the total area of malignant tissue was calculated. These proportions served as feature vectors for training machine learning classifiers aimed at predicting the ISUP grade. Various machine learning classifiers were evaluated to assess their effectiveness. The classifiers explored in this research include multilayer perceptron (MLP), RandomForest (RF), linear regression (LR), extraTrees (ET), k-nearest neighbor (KNN), XGBoost, and support vector machines (SVM). These classifiers were subjected to testing with various hyperparameter configurations, with the optimal set of hyperparameters being chosen based on performance outcomes.

Experimental Setup

The dataset was divided into five folds subject-wise to do cross-validation for model evaluation where each fold was used once as a test set while the other folds were used for training and validation. It was repeated for all the folds. Models were set to train for 100 epochs and the best epoch results were saved on the validation set results. The patch images were resized to 224 × 224 pixels for classification models and 256 × 256 for segmentation models before training as it is the input shape required to use ImageNet weights. Segmentation models pytorch (SMP) package [47] was used to implement the segmentation models. Images were also normalized using standard normalization. A combination of Adam optimizer with a learning rate of 0.0001 provided the best results. The hardware configuration used for the experiments was Intel(R) Xeon(R) CPU, NVIDIA GeForce RTX 3090, 64 GB RAM, Python 3.9.16, and Pytorch version 1.13.

Numerical Results

Performance Metrics

The most common metrics employed to evaluate the performance of classification include precision, recall, F1-score, and accuracy. These metrics can be calculated using the following formulas:

$$\text{Precision} = \frac{T_p}{T_p + F_p} \tag{1}$$

$$\text{Recall/Sensitivity} = \frac{T_p}{T_p + F_n} \tag{2}$$

$$\text{Specificity} = \frac{T_n}{T_n + F_p} \tag{3}$$

$$F1\text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

$$\text{Accuracy} = \frac{\sum_{c=1}^{N} T_{p_c}}{N_T} \tag{5}$$

where T_p is true positive, F_p is false positive, F_n is false negative, T_n is true negative, and N_T is the number of classes.

For evaluating segmentation performance, commonly utilized metrics include the intersection over Union (IoU), dice similarity coefficient (DSC), false negative rate (FNR),

false positive rate (FPR), and specificity, all of which were also employed in this work. Formulas used to calculate these metrics are provided below:

$$IoU = \frac{T_p}{T_p + F_p + F_n} \tag{6}$$

$$DSC = \frac{2T_p}{2T_p + F_p + F_n} \tag{7}$$

$$FNR = \frac{F_n}{T_p + F_n} \tag{8}$$

$$FPR = \frac{F_p}{F_p + T_n} \tag{9}$$

where T_p is true positive, F_p is false positive, F_n is false negative, and T_n is true negative, respectively.

In this research, the primary evaluation metric employed for final ISUP grade prediction was the quadratic weighted kappa (QWK) which was also the primary metric used in the PANDA challenge, assessing the degree of agreement between the predicted labels and ground truth labels. The QWK was derived by initially constructing an N-by-N confusion matrix M, where N = 6 (ISUP grades 0–5). For every element $M_{i,j}$, i^{th} row corresponds to the ISUP score that was the ground truth and the j^{th} column to the predicted score. The N-by-N weight matrix w is formulated from the squared differences between the target and predicted scores using the following equation:

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2} \tag{10}$$

Furthermore, a confusion matrix of expected outcomes, E, was calculated under the assumption that there was no correlation between the ground truth and predicted labels. QWK was then computed by multiplying the elements of the weight matrix w with those of the confusion matrix M, and the expected outcomes matrix E, following the formula:

$$QWK = \frac{\sum_i \sum_j w_{i,j} M_{i,j}}{\sum_i \sum_j w_{i,j} E_{i,j}} \tag{11}$$

Classification Performance

The classification performance of deep learning models will be discussed in this section. The classification models were trained separately on patches of size 500 × 500 and 1000 × 1000 to investigate the effect of the area of region available in the patch.

Classification Results on 500 × 500 Sized Patches

EfficientNet_b0 demonstrated the best metrics in terms of F1-score and accuracy with the highest overall accuracy of 90.13% and F1-score of 83.83%, closely followed by Densenet121 and Inception_v3, with accuracies of 89.48% and 89.46%, respectively. Densenet121 had the highest specificity of 92.95% of all the models. Figure 8a shows the confusion matrices for the models and it can be observed that most of the confusion was between Gleason 4 and Gleason 5. Vision Transformer showed a slightly lower performance compared to the other models. Table 4 shows the different classification performance metrics of all the models like F1-score, precision, and sensitivity, on 500 × 500 sized patches.

Classification Results on 1000 × 1000 Sized Patches

Densenet121 showed a slight dip in accuracy to 88.9% but exhibited an improved F1-score of 83.51%. EfficientNet_b0 maintained a robust performance with an accuracy of 89.2% and an F1-score of 83.83%. Inception_v3, however, experienced a notable decrease in performance across all metrics indicating a potential challenge in handling image patches with more tissue region. Figure 8b shows the confusion matrices for the models. Vision Transformer displayed a comparable performance to its 500 × 500 patch size metrics, with a slight improvement in accuracy to 88.94% and precision to 82.68%. Table 5 shows the different classification performance metrics of all the models like F1-score, precision, and sensitivity, on 1000 × 1000 sized patches.

Overall, EfficientNet_b0 consistently outperformed the other models across most metrics for both patch sizes, showcasing its effectiveness with varying input dimensions. DenseNet121 also delivered a strong performance, consistently achieving an F1-score above 80% in both datasets. In contrast, Inception_v3 and Vision Transformer fell short of matching the performance levels of the former two models.

Segmentation Performance

In this section, the segmentation performances of the deep learning models will be discussed. Following the superior performance of DenseNet121 and EfficientNet networks in segmentation, these encoders were investigated and combined with the UNet.

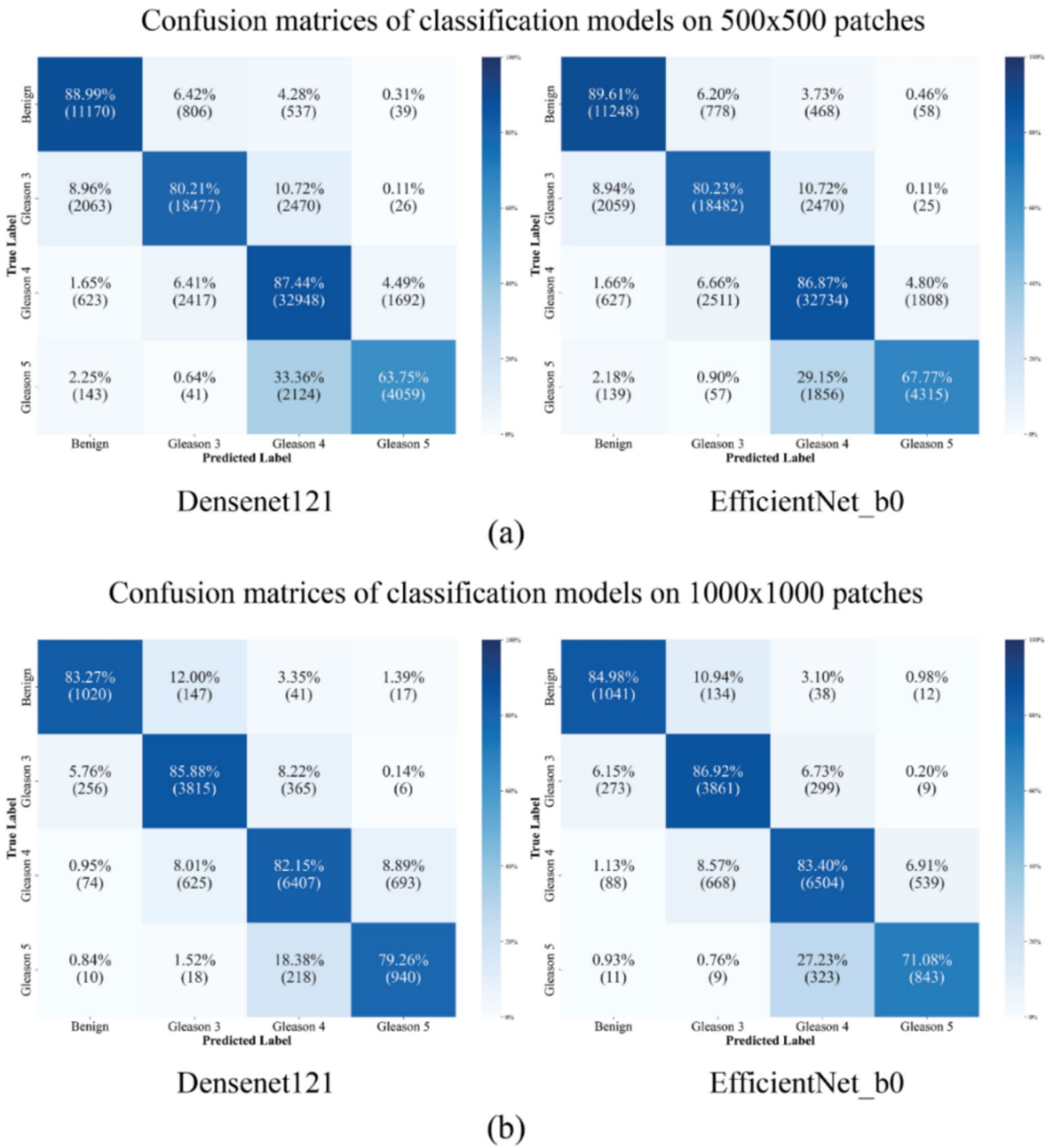

Fig. 8 Confusion matrix for Densenet121 and EfficientNet_b0 on 500×500 patches (**a**) and DenseNet121 and EfficientNet_b0 on 1000×1000 patches (**b**)

Segmentation Results on 500×500 Sized Patches

EfficientNet UNet showed slightly better performance overall compared to DenseNet UNet and significantly outperformed the baseline UNet model. The EfficientNet UNet achieved an accuracy of 55.62%, an IoU of 88.39%, and a Dise score coefficient (DSC) of 89.56%. It also showed superior performance in segmenting benign and various Gleason score categories, with the highest DSC of 91.4% for Gleason 5. DenseNet UNet also demonstrated strong performance

with an accuracy of 55.36%, an IoU of 87.53%, and a DSC of 88.73%. Its performance across different Gleason scores was commendable, particularly with a high DSC of 96.26% for benign cases. The baseline UNet model had significantly lower performance metrics with an accuracy of 55.05%, an IoU of 49.58%, and a DSC of 52.1%. This showed the effectiveness of the ImageNet-based encoders.

The DeepLabV3 and DeepLabV3 + architectures demonstrated excellent performance across all metrics. The DeepLabV3 model with EfficientNet encoder achieved an accuracy of 81.09%, an IoU of 83.14%, and a DSC of 84.51%. It performed particularly well in segmenting Gleason score 5, attaining a DSC of 91.93%. The DeepLabV3 + with EfficientNet encoder model had a slightly lower overall accuracy of 80.92% but comparable IoU (83.27%) and DSC (84.69%) to the DeepLabV3 variant. Its DSC for Gleason 5 was also high at 91.98%. The incorporation of SelfONN boosted the performance further with DeepLabV3 with Self-ONN achieving an 80.96% accuracy, 83.30% IoU, and 84.72% DSC, with its best DSC of 92.33% for Gleason 5. The DeepLabV3 + with SelfONN emerged as the top performer with 81.14% accuracy, 83.47% IoU, and the highest overall DSC of 84.90%. It excelled at segmenting Gleason 5 with a DSC of 93.04% while maintaining competitive performance for other grades. These results highlight the effectiveness of the DeepLabV3 and DeepLabV3 + architectures, especially when combined with SelfONN, in accurately segmenting prostate cancer. Table 6 shows all the performance metrics of the segmentation models on the 500 × 500 sized patches like Dice Score (DSC) for each class and the Intersection over Union (IoU).

Segmentation Results on 1000 × 1000 Sized Patches

DenseNet UNet reported an accuracy of 84.40%, an IoU of 79.00%, and a DSC of 80.94%. EfficientNet UNet maintained competitive performance with an accuracy of 84.25%, an IoU of 76.45%, and a DSC of 78.42%. This model showed a balanced performance across different Gleason scores, with the highest DSC of 85.37% for Gleason 5. The baseline UNet's performance improved significantly on the larger patches, achieving an accuracy of 83.99%, an IoU of 75.62%, and a DSC of 77.85%, but it was still not as effective as the other two models. The DeepLabV3

Table 4 Performance metrics of classification models (in %) trained on 500 × 500 sized patches

Model	Accuracy	Precision	Sensitivity	F1-score	Specificity
Densenet121	89.48	80.68	80.39	80.38	**92.95**
Efficientnet_b0	**90.13**	**83.92**	**83.85**	**83.83**	92.01
Inception_v3	89.46	82.07	81.81	81.83	92.35
Vision Transformer	88.51	80.06	80.21	79.67	90.30

The bold values indicate the highest values in the column

Table 5 Performance metrics of classification models (in %) trained in 1000 × 1000 sized patches

Model	Accuracy	Precision	Sensitivity	F1_score	Specificity
Densenet121	88.90	**84.49**	83.14	83.51	**92.16**
Efficientnet_b0	**89.20**	84.40	**83.60**	**83.83**	91.90
Inception_v3	84.64	75.83	74.99	75.20	88.50
Vision Transformer	88.94	82.68	81.67	81.97	92.14

The bold values indicate the highest values in the column

Table 6 Performance metrics (in %) of segmentation models trained on 500 × 500 sized patches

Model	Accuracy	IoU	DSC	DSC (Benign)	DSC (Gleason 3)	DSC (Gleason 4)	DSC (Gleason 5)
DenseNet Unet	79.65	81.51	82.98	72.86	86.91	82.42	89.72
EfficientNet UNet	80.29	82.34	83.78	73.51	87.55	83.13	90.93
Vanilla UNet	55.05	49.58	52.1	47.69	62.18	55.29	43.24
EfficientNet DeepLabV3	81.09	83.14	84.51	74.35	88.21	83.54	91.93
EfficientNet DeepLabV3 +	80.92	83.27	84.69	74.46	87.86	84.46	91.98
EfficientNet DeepLabV3 (with SelfONN)	80.96	83.30	84.72	74.36	88.37	83.82	92.33
EfficientNet DeepLabV3 + (with SelfONN)	**81.14**	**83.47**	**84.9**	73.91	**88.71**	83.94	**93.04**

The bold values indicate the highest values in the column

model with EfficientNet encoder achieved an accuracy of 83.85%, an IoU of 79.73%, and a DSC of 81.91%. It demonstrated strong performance in segmenting Gleason 5 with a DSC of 92.5%, while also performing reasonably well for Gleason 4 (DSC 82.68%) and Gleason 3 (DSC 84.75%). However, its DSC for benign cases was relatively lower at 67.7%. The DeepLabV3 + model had a slightly lower accuracy of 83.52% compared to the DeepLabV3 variant, but a higher DSC of 81.11%. It excelled at segmenting Gleason 5, achieving the highest DSC of 93.47% among all models. Its performance on Gleason 3 (DSC 84.85%) was also noteworthy. Incorporating SelfONN with the DeepLabV3 model led to a minor improvement in accuracy (83.67%) and DSC (81.99%), while the IoU remained similar (79.71%). It maintained a high DSC of 92.65% for Gleason 5 and showed improved performance for Gleason 4 (DSC 83.96%). The-DeepLabV3 + with SelfONN emerged as the top performer with the highest overall DSC of 83.08% and an accuracy of 83.99%. It achieved the second-highest DSC for Gleason 5 at 93.56% and performed well for Gleason 4 (DSC 84.94%) and Gleason 3 (DSC 85.6%). However, its DSC for benign cases (68.21%) was relatively lower compared to some other models. These results demonstrate the effectiveness of the DeepLabV3 and DeepLabV3 + architectures, particularly when combined with self-ONN, in accurately segmenting prostate cancer across various Gleason scores. Table 7 shows the performance metrics of the segmentation models on the 500 × 500 sized patches. Figure 9 shows a comparison of predictions made by the different segmentation models.

Table 7 Performance metrics (in %) of segmentation models trained on 1000 × 1000 sized patches

Model	Accuracy	IoU	DSC	DSC (Benign)	DSC (Gleason 3)	DSC (Gleason 4)	DSC (Gleason 5)
DenseNet Unet	84.40	79.00	80.94	69.35	85.45	80.15	88.81
EfficientNet UNet	84.25	76.45	78.42	69.15	79.33	79.85	85.37
Vanilla UNet	83.99	75.62	77.85	66.13	82.98	74.04	88.22
EfficientNet DeepLabV3	83.85	79.73	81.91	67.7	84.75	82.68	92.5
EfficientNet DeepLabV3 +	83.52	78.88	81.11	66.28	84.85	79.84	93.47
EfficientNet DeepLabV3 (with SelfONN)	83.67	79.71	81.99	67.08	84.27	83.96	92.65
EfficientNet DeepLabV3 + (with SelfONN)	83.99	80.98	**83.08**	68.21	85.6	84.94	93.56

The bold values indicate the highest values in the column

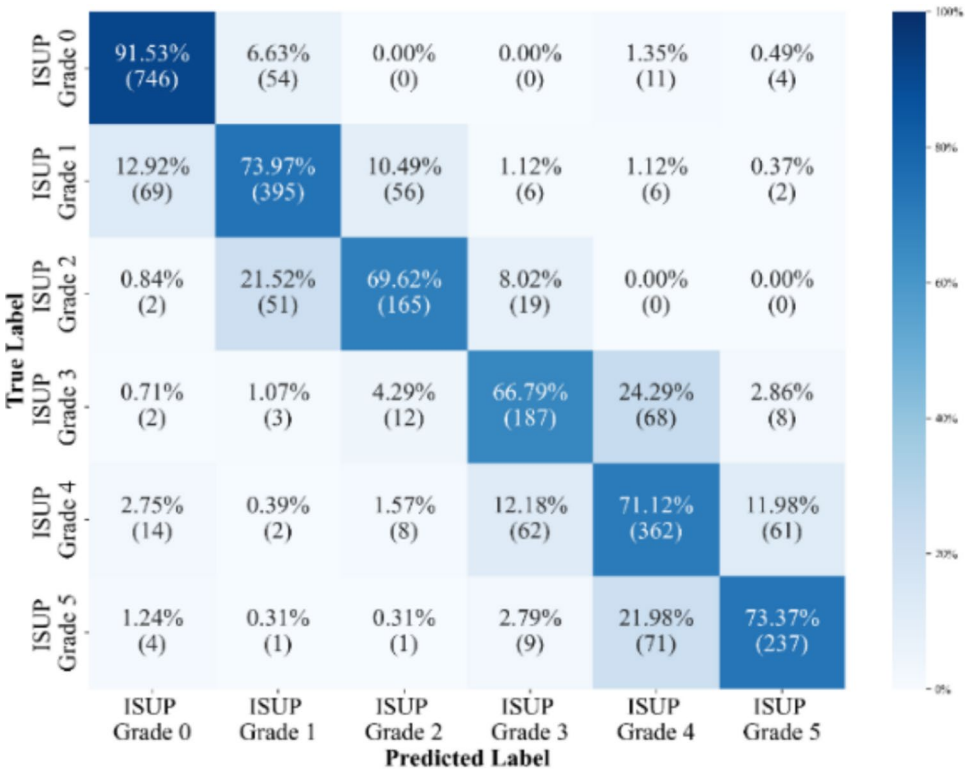

Fig. 9 Confusion matrix of the best performing final grade classifier (RandomForest)

The proposed self-ONN-based DeepLabV3 and DeepLabV3 + models have significantly outperformed other models in segmentation performance. In terms of metrics, the models slightly underperformed on the larger-sized patches in patch-level segmentation. This also depicts the impact of integrating advanced encoders with UNet architecture to boost performance.

Performance of Final Grade Classifiers

The final stage of the framework consists of a machine learning-based classifier to predict the ISUP grade of a WSI from the percentage distribution of each class in the image. First, the percentage of each class (benign, Gleason 3, Gleason 4, and Gleason 5) needs to be calculated for a WSI. These calculations are performed using two methodologies (classification and segmentation) outlined in the preceding sections. For classification, the process begins with determining the number of patches for each category predicted within the WSI. Subsequently, the percentage representation of each category is computed, resulting in a vector that represents the class percentage distribution within the WSI. Similarly, in segmentation, the percentage calculations are done based on the predicted area covered by each class within the WSI, which forms the corresponding vector. These vectors serve as input features for training machine learning models. Initially, vectors derived from deep learning models were utilized independently to train and evaluate classifiers, aiming to assess the performance of each model on its own. Following this, vectors from various deep learning models were concatenated, serving as composite features for the machine learning algorithms to evaluate the efficacy of combining multiple deep learning models.

Models trained on 500 × 500 patches yielded superior features compared to those trained on 1000 × 1000 patches. Specifically, when using Densenet121-derived features, the ExtraTrees classifier achieved the highest performance, with a Quadratic Weighted Kappa (QWK) score of 0.8468. In contrast, features derived from EfficientNet_b0, when used with the XGBoost classifier, reached an even higher QWK score of 0.8587, marking the highest score obtained from models trained on 500 × 500 patches. Conversely, the best QWK score achieved using features from 1000 × 1000 patch-based models was significantly lower, at only 0.6424. This evidence suggests that for classification tasks, the performance of features from models trained on smaller, 500 × 500 patches is significantly better. Table 8 shows QWK scores for all classifiers on features generated from different classification models.

Features generated from the segmentation networks overall showed better performance than the classification models with the highest QWK score of 0.9140 achieved by DeepLabV3 + with self-ONN. Again, features from segmentation models trained on 1000 × 1000 patches outperformed the 500 × 500 patch-based models which only achieved the highest score of 0.84, whereas the models based on 1000 × 1000 patches achieved highest QWKs of 0.8853 and 0.914, respectively. Table 9 shows QWK scores for all classifiers on features generated from different segmentation models. Subsequently, to explore potential enhancements in performance, various feature set combinations were tested. The highest-performing models from both classification and segmentation approaches were selected, and their feature vectors were concatenated. This concatenated feature set was then used to train the classifiers.

Table 10 shows the performance of the classifiers which is the kappa agreement score QWK, on different feature combinations (classification and segmentation). A combination of features has outperformed individual model features in all cases. The combination of features from classification models based on 500 × 500 patches achieved the highest QWK score of 0.86 whereas feature combination of segmentation models based on 1000 × 1000 patches achieved 0.91. The best QWK score of 0.9215 was achieved by the RandomForest classifier trained on a combination of high-performing features from both classification and segmentation. This demonstrated a combination of classification

Table 8 Performance of classifiers on features generated from classification approach

Classifier	fivefold cross-validated QWK			
	Trained on 500 × 500 patch size		Trained on 1000 × 1000 patch size	
	Densenet121	EfficientNet_b0	Densenet121	EfficientNet_b0
MLP	0.842438	0.846698	0.590591	0.63705
RandomForest	0.835312	0.845768	0.552992	0.619265
Linear Regression	0.827112	0.838998	**0.600642**	**0.642488**
ExtraTrees	**0.846861**	0.854062	0.540032	0.626638
KNN	0.84099	0.847315	0.529318	0.599451
XGBoost	0.846367	**0.858738**	0.534543	0.636177
SVM Classifier	0.843084	0.853314	0.547538	0.628727

The bold values indicate the highest values in the column

Table 9 Performance of classifiers on features generated from the segmentation approach

Classifier	fivefold cross-validated QWK			
	Trained on 500 × 500 patch size		Trained on 1000 × 1000 patch size	
	EfficientNet DeepLabV3 (with SelfONN)	EfficientNet DeepLabV3 + (with SelfONN)	EfficientNet DeepLabV3 (with SelfONN)	EfficientNet DeepLabV3 + (with SelfONN)
MLP	0.829008	0.790134	0.883217	0.909623
RandomForest	0.826295	0.786989	0.875651	0.9076
Linear Regression	0.836197	**0.815572**	0.881236	0.902995
ExtraTrees	**0.840013**	0.80784	0.884601	0.908974
KNN	0.818308	0.810445	0.881494	0.907084
XGBoost	0.82366	0.808155	0.879546	0.906277
SVM classifier	0.837823	0.804833	**0.885375**	**0.914088**

The bold values indicate the highest values in the column

Table 10 Performance of classifiers on different feature combinations

Classifier	fivefold cross-validated QWK		
	Classification models features combined	Segmentation models features combined	All combined
MLP	0.857978	0.901554	0.910530
RandomForest	0.854837	**0.911982**	**0.921568**
Linear regression	0.844959	0.899536	0.903913
ExtraTrees	**0.860701**	0.899083	0.917415
KNN	0.848810	0.902129	0.905579
XGBoost	0.858815	0.903579	0.917771
SVM classifier	0.848420	0.902882	0.905025

The bold values indicate the highest values in the column

approach with segmentation as well as different patch sizes have contributed to increased performance. Figure 9 shows the confusion matrix for the best-performing classifier of the framework.

Discussion

In this paper, a novel deep learning-based framework was proposed in combination of classification and segmentation to predict grades of PCa from WSIs. The framework included customized DeepLabV3 networks consisting of self-ONN layers instead of CNNS. In the initial PANDA challenge, there were reports of inconsistencies in labeling. Despite these issues, the gap in accuracy between the development and test sets was not markedly wide. This was attributed to the development set comprising WSIs from two different institutions. In contrast, this study observed a considerably larger discrepancy in performance between the training and test sets, primarily because the development set originated from a single institution [48]. To gauge the extent of label noise, the study assessed the reference standard against labels obtained through a semi-automatic method. The comparison revealed an accuracy of 0.675 (with a kappa score of 0.819) for the Gleason scores and an accuracy of 0.720 (with a kappa score of 0.853) for the grade groups [31]. Extensive data cleaning was performed to tackle this issue and ensure that training and evaluation of the framework were conducted reliably.

Although classification and segmentation are not new approaches for PCa grade assessment, the novelty of the framework lies in the combination of both classification and segmentation techniques trained on different-sized patches as well as the development of a customized segmentation network based on DeepLabV3 and self-ONN. Classification has been used extensively on the PANDA challenge dataset. The winner of the challenge [34] used classification on a representative image created from a WSI by concatenating patches with dark regions, which have a higher probability of containing cancer tissue. A similar approach was taken by Yang et al. [49] and Nishio et al. [48] who performed classification on such preprocessed images generated from

whole slide images using a multi-channel ResNet network and an EfficientNet network combined with a label distribution technique respectively. The main concern with this preprocessing technique is it may ignore vital regions when creating the image thus causing a faulty representation of the WSI. Esquivel et al. [28] performed patch-based classification training on the Radboud dataset and tested it on their custom dataset as well as the Radboud dataset. They achieved a kappa score of 0.74 on their custom dataset and 0.51 on the Radboud dataset respectively. Also, their classification was only limited to Gleason patterns 3, 4, and 5. Nagpal et al. [50] patch-based classification and achieved a 70% accuracy in 331 cases. Praful et. al. [51] have proposed radiomics-based deeply supervised U-Net for PCa segmentation from MR Images. Their proposed method achieved a maximum mean DSC score of 89.58% in a multi-site T2-weighted MRI structured dataset.

Bulten et al. [31] introduced an advanced variant of UNet, incorporating CycleGAN for image normalization and to accommodate variations. They achieved a kappa score of 0.918 on an internal dataset training with the Radboud dataset. Singhal et al. [32] also developed a custom U-net-based architecture and achieved a QWK score of 0.93 although they only tested on 1303 cases. Prabhu et.al [52] analyzed Radboud University Medical Centre images to develop a PCa detection system using five different types of deep learning models, including MobileNet V2, Inception ResNet V2, DenseNet 169, ResNet101 V2, and NASNetMobile. The results showed that InceptionResNet V2 could accurately classify data with a RMSE of 0.631664. Based on the results of this study, the data was only classified using transfer learning approaches and not segmented. In recent work, Liang et.al [53] described an attention-LSTM-based aggregator approach for PCa classification. With a five-fold cross-validation method, their proposed LSTM-aggregator achieved a maximum mean classification rate of 0.745 and a QWK of 0.903. Zhongyi et al. [54] propose a novel Intensive-Sampling Multiple Instance Learning Framework for PCa detection. They have achieved a maximum QWK of 0.860 in classifying six different types of PCa using the PANDA dataset. Yujie et al. [55]have proposed a novel multi-stage fully convolutional neural network for prostate segmentation in ultrasound images and achieved a maximum binary classification accuracy of 99.15% and a mean DSC score of 94.90% in the CCH-TRUSPS dataset. However, they have achieved a maximum binary classification accuracy of 99.72% and a DSC score of 93.65% achieved using a multi-site T2-weighted MRI structured dataset [55]. There are few major challenges in their work: (a) the number of parameters used for PCa classification is ~ 41B (b) the model has been trained and tested with a limited number of samples (c) they have not reported the performance of the model based on QWK in their work.

According to Singh et al. [56], a novel 3D CNN can perform segmentation and classification of PCa based on MRI images from the SPIE-AAPM-NCI prostate dataset. Using binary classification (positive/negative), they achieved a maximum accuracy of 87% and specificity of 85%. Based on the PANDA dataset, Balaha et al. [57] proposed a deep transfer learning-based approach with an optimizer for both PCa segmentation and classification. A maximum segmentation accuracy of 98.46% was reported using the U-Net model, and a classification accuracy of 88.91% was reported using the classification model. They developed a model for segmenting and classifying PCa using deep learning across three different datasets (PANDA, Transverse Plane, and ISUP). Table 11 contains a comparison of our methodology and results with the existing approaches toward PCa grade assessment.

In our study, we integrated features derived from both classification and segmentation techniques to classify PCa images according to ISUP grades, establishing a more robust framework that leverages the strengths of both approaches while minimizing errors. Additionally, we explored the use of two patch sizes, 500×500 and 1000×1000 while most studies used single-sized patches ranging from 299×299 to 911×911.

A critical aspect of our research was the implementation of fivefold cross-validation, which allowed for optimal utilization of the data and enabled comprehensive reporting on the dataset's entirety. This approach significantly contributed to a more dependable assessment of the model's performance. In our study, both the training and testing phases were conducted using data from the same source, without evaluating the model on an external, unseen dataset. While this method guarantees a certain level of generalization and robust performance within the studied dataset, it leaves the model's efficacy on new, untested data open to question. The inability to use the Karolinska dataset for training stemmed from the absence of cancer grade annotations in the masks, coupled with variations in staining colors, which would render predictions on this dataset less effective. Looking ahead, employing generative adversarial networks (GANs) could offer a solution for adapting to stain variations and facilitating model training that is resilient to differences in staining. In the future, enhancing the Karolinska dataset with cancer-grade masks could generate a new dataset that facilitates more generalized training and evaluation of models.

Conclusion

This research introduces a novel framework that integrates classification and segmentation networks for analyzing tissue images to identify the distribution of different cancer cell patterns. The framework uses classification and segmentation networks to calculate the distribution of the percentage of different

Table 11 Comparison with other existing approaches for Prostate cancer grade assessment

Authors	Dataset source	Applications	Patch size	Dataset size	Result
Nagpal et al. [50]	Naval Medical Center, TCGA Dataset, San Diego Marin Medical Laboratories	Gleason scoring of whole slide images	911×911	Training: 1226 slides Validation: 331 slides	Accuracy: 70%
Arvantiti et al. [29]	University Hospital Zurich	6-class Gleason grading	750×750	Training: 641 patients Testing: 245 patients	Agreement score (DL vs pathologist 1): 0.75 Agreement score (DL vs pathologist 2): 0.71
Lucas et al. [58]	Amsterdam University Medical Centers	4-class classification	299×299	96 samples	QWK: 0.70
Bulten et al. [31]	Radboud University Medical Center	6-class Gleason grading	-	5759 samples	Kappa: 0.854
Egevad et al. [59]	Pathology Imagebase dataset hosted on the ISUP Web site	5-class classification	2048×2048 (representation of WSI)	87 samples	Kappa agreement score: 0.63
Singhal et al. [32]	Muljibhai Patel Urological Hospital (MPUH), PANDA challenge dataset (Radboud University Medical Center and Karolinska Institute)	6-class classification	-	Training: 155 biopsies (MPUH), 3586 biopsies (Radboud) Testing: 425 biopsies (MPUH), 1201 (Radboud), 1303 biopsies (Karolinska Institute)	Accuracy: 83.1% Kappa: 0.93
Nishio et al. [48]	Radboud University Medical Center	6-class classification	1536×1536 (representation of WSI)	5160 biopsies (Radboud) 5456 biopsies (Karolinska Institute)	Kappa: 0.85
Prabhu et.al [52]	Radboud University Medical Center	6-class classification	224×224 (representation of WSI)	25% for testing and 75% for training	Accuracy: 84.99% (Inception-ResNet V2)
Balaha et.al [57]	PADA, transverse plane, and ISUP grade-wise dataset	2 Class	512×512	N/A	88.91% (PANDA) 100% (transverse plane) 98.46% (ISUP)
Singh et.al [56]	SPIE-AAPM-NCI Dataset	2 class	N/A	112 lesions–training, 70 lesions for testing	Accuracy: 87% Sensitivity: 89% Specificity: 85%
Proposed method	**Radboud University Medical Center**	**6-class classification**	**500×500 and 1000×1000**	**fivefold cross-validation on 2699 cases**	**F1-score: 83.83%** **DSC: 84.9%** **QWK: 0.9215**

The bold values indicate the highest values in the column

cancer cell patterns in the tissue image and then uses a machine learning classifier to predict the final ISUP grade. The study used the Radboud University Medical Center's dataset from the PANDA challenge, selecting 2699 out of 5160 cases. It employed five-fold subject-wise cross-validation, utilizing 500×500 and 1000×1000 patches in classification and segmentation networks. The classification networks achieved the highest F1-score of 83.83%, and segmentation networks a DSC of 84.9%. Self-ONN improved segmentation performances in DeepLabV3-based models. These networks helped compute cancer pattern percentages in WSIs for machine learning classifiers, including MLP, RandomForest, and SVM. Performance improved by merging features from classification and segmentation models, achieving a QWK score of 0.9215, surpassing many existing methods. Future directions include making the models tolerant to staining intensity variations for broader applications and the noisy data in the dataset can be relabeled with better segmentation models to be made usable for training. This study underscores the promising potential of deep learning in grading PCa, while also highlighting the need for further research to overcome current limitations and assess the framework's clinical applicability.

Author Contribution Saidul Kabir (SK): methodology, software, write—original draft.

Rusab Sarmun (RS): software, write—original draft.

Rafif Mahmood Al Saady (RMAS): data curation, validation.

Semir Vranic: investigation, validation.

M Murugappan: methodology, supervision, visualization, investigation, review and revise the paper.

Muhammad E. H. Chowdhury: conceptualization, methodology, supervision, review and revise the paper.

Data Availability This study uses data from the PANDA challenge dataset [33] at the Radboud University Medical Center and the Karolinska Institute.

Declarations

Ethical Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing Interests The authors declare no competing interests.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians 2018;68:394–424.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians 2021;71:209–49.
3. Sanda MG, Cadeddu JA, Kirkby E, Chen RC, Crispino T, Fontanarosa J, et al. Clinically localized prostate cancer: AUA/ASTRO/SUO guideline. Part II: recommended approaches and details of specific care options. The Journal of Urology 2018;199:990–7.
4. Sanda MG, Cadeddu JA, Kirkby E, Chen RC, Crispino T, Fontanarosa J, et al. Clinically localized prostate cancer: AUA/ASTRO/SUO guideline. Part I: risk stratification, shared decision making, and care options. The Journal of Urology 2018;199:683–90.
5. Samaratunga H, Delahunt B, Yaxley J, Srigley JR, Egevad L. From Gleason to International Society of Urological Pathology (ISUP) grading of prostate cancer. Scandinavian Journal of Urology 2016;50:325–9.
6. Faraj SF, Bezerra SM, Yousefi K, Fedor H, Glavaris S, Han M, et al. Clinical validation of the 2005 ISUP Gleason grading system in a cohort of intermediate and high risk men undergoing radical prostatectomy. PloS One 2016;11:e0146189.
7. Epstein JI. An update of the Gleason grading system. The Journal of Urology 2010;183:433–40.
8. Li Y, Huang M, Zhang Y, Chen J, Xu H, Wang G, et al. Automated gleason grading and gleason pattern region segmentation based on deep learning for pathological images of prostate cancer. IEEE Access 2020;8:117714–25.
9. Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. The Lancet Oncology 2020;21:222–32.
10. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine 2019;25:954–61.
11. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature Medicine 2018;24:1559–67.
12. Murugappan M, Prakash N.B, Jeya R, Mohanarathinam A, Hemalakshmi G.R, Mufti Mahmud, A novel few-shot classification framework for diabetic retinopathy detection and grading, Measurement, 2022:200:111485.
13. Murugappan, M., Bourisly, A.K., Prakash, N.B. et al. Automated semantic lung segmentation in chest CT images using deep neural network. Neural Computing and Applications, 2023:35:15343–15364.
14. Sarmun R, Kabir S, Prithula J, Alqahtani A, Zoghoul SB, Al-Hashimi I, et al. Enhancing intima-media complex segmentation with a multi-stage feature fusion-based novel deep learning framework. Engineering Applications of Artificial Intelligence 2024;133:108050.
15. Bushra F, Chowdhury MEH, Sarmun R, Kabir S, Said M, Zoghoul SB, et al. Deep learning in computed tomography pulmonary angiography imaging: A dual-pronged approach for pulmonary embolism detection. Expert Systems with Applications 2024;245:123029.
16. Khan MM, Chowdhury MEH, Arefin ASMS, Podder KK, Hossain MSA, Alqahtani A, Murugappan M, Khandakar A, Mushtak A, Nahiduzzaman M. A Deep Learning-Based Automatic Segmentation and 3D Visualization Technique for Intracranial Hemorrhage Detection Using Computed Tomography Images. Diagnostics. 2023; 13(15):253.
17. Hemalakshmi, G.R., Murugappan, M., Sikkandar, M.Y. et al. Automated retinal disease classification using hybrid transformer model (SViT) using optical coherence tomography images. Neural Computing and Applications, 2024: 36: 9171–9188

18. Li J, Li W, Gertych A, Knudsen BS, Speier W, Arnold CW. An attention-based multi-resolution model for prostate whole slide image classification and localization. ArXiv Preprint ArXiv:190513208 2019. https://doi.org/10.48550/arXiv.1905.13208.
19. Xu H, Park S, Hwang TH. Computerized classification of prostate cancer gleason scores from whole slide images. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2019;17:1871–82.
20. Xie H, Zhang Y, Wang J, Zhang J, Ma Y, Yang Z. Automated Prostate Cancer Diagnosis Based on Gleason Grading Using Convolutional Neural Network. ArXiv Preprint ArXiv:201114301 2020. https://doi.org/10.48550/arXiv.2011.14301.
21. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. International conference on machine learning, PMLR; 2018, p. 2127–36.
22. Nguyen K, Sabata B, Jain AK. Prostate cancer grading: Gland segmentation and structural features. Pattern Recognition Letters 2012;33:951–61.
23. Gorelick L, Veksler O, Gaed M, Gómez JA, Moussa M, Bauman G, et al. Prostate histopathology: Learning tissue component histograms for cancer detection and classification. IEEE Transactions on Medical Imaging 2013;32:1804–18.
24. Waliszewski P, Wagenlehner F, Gattenloehner S, Weidner W. On the relationship between tumor structure and complexity of the spatial distribution of cancer cell nuclei: a fractal geometrical model of prostate carcinoma. The Prostate 2015;75:399–414.
25. Huang M, Han H, Wang H, Li L, Zhang Y, Bhatti UA. A clinical decision support framework for heterogeneous data sources. IEEE Journal of Biomedical and Health Informatics 2018;22:1824–33.
26. Li Y, Niu S, Ji Z, Fan W, Yuan S, Chen Q. Automated choroidal neovascularization detection for time series SD-OCT images. International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer; 2018, p. 381–8.
27. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine 2019;25:1301–9.
28. Marrón-Esquivel JM, Duran-Lopez L, Linares-Barranco A, Dominguez-Morales JP. A comparative study of the inter-observer variability on Gleason grading against Deep Learning-based approaches for prostate cancer. Computers in Biology and Medicine 2023;159:106856.
29. Arvaniti E, Fricker KS, Moret M, Rupp N, Hermanns T, Fankhauser C, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. Scientific Reports 2018;8:12054.
30. Ing N, Ma Z, Li J, Salemi H, Arnold C, Knudsen BS, et al. Semantic segmentation for prostate cancer grading by convolutional neural networks. Medical Imaging 2018: Digital Pathology, vol. 10581, SPIE; 2018, p. 343–55.
31. Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. The Lancet Oncology 2020;21:233–41.
32. Singhal N, Soni S, Bonthu S, Chattopadhyay N, Samanta P, Joshi U, et al. A deep learning system for prostate cancer diagnosis and grading in whole slide images of core needle biopsies. Scientific Reports 2022;12:3383.
33. Bulten W, Kartasalo K, Chen P-HC, Ström P, Pinckaers H, Nagpal K, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nature Medicine 2022;28:154–63.
34. FAM_TARO. https://www.kaggle.com/competitions/prostate-cancer-grade-assessment/discussion/169143 20/05/2024.
35. Kabir S, Vranic S, Al Saady RM, Khan MS, Sarmun R, Alqahtani A, et al. The utility of a deep learning-based approach in Her-2/neu assessment in breast cancer. Expert Systems with Applications 2024;238:122051.
36. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. Journal of Big Data 2019;6:1–48.
37. Tellez D, Litjens G, Bándi P, Bulten W, Bokhorst J-M, Ciompi F, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Medical Image Analysis 2019;58:101544.
38. Rahman T, Chowdhury MEH, Khandakar A, Islam KR, Islam KF, Mahbub ZB, et al. Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray. Applied Sciences 2020;10:3233.
39. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, p. 4700–8.
40. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. International conference on machine learning, PMLR; 2019, p. 6105–14.
41. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, p. 2818–26.
42. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv Preprint ArXiv:201011929 2020. https://doi.org/10.48550/arXiv.2010.11929.
43. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, Springer; 2015, p. 234–41
44. Chen L-C, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. ArXiv Preprint ArXiv:170605587 2017;5. https://doi.org/10.48550/arXiv.1706.05587.
45. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV), 2018, p. 801–18.
46. Kiranyaz S, Malik J, Abdallah H Ben, Ince T, Iosifidis A, Gabbouj M. Self-organized operational neural networks with generative neurons. Neural Networks 2021;140:294–308.
47. P. Iakubovskii. Segmentation Models Pytorch. https://Github.Com/Qubvel/Segmentation_modelsPytorch 20/05/2024.
48. Nishio M, Matsuo H, Kurata Y, Sugiyama O, Fujimoto K. Label distribution learning for automatic cancer grading of histopathological images of prostate cancer. Cancers (Basel) 2023;15:1535.
49. Yang B, Xiao Z. A multi-channel and multi-spatial attention convolutional neural network for prostate cancer ISUP grading. Applied Sciences 2021;11:4321.
50. Nagpal K, Foote D, Liu Y, Chen P-HC, Wulczyn E, Tan F, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. NPJ Digital Medicine 2019;2:48.
51. Hambarde P, Talbar S, Mahajan A, Chavan S, Thakur M, Sable N. Prostate lesion segmentation in MR images using radiomics based deeply supervised U-Net. Biocybernetics and Biomedical Engineering 2020;40:1421–35.
52. Kanna GP, Kumar SJKJ, Parthasarathi P, Kumar Y. A review on prediction and prognosis of the prostate cancer and gleason grading of prostatic carcinoma using deep transfer learning based approaches. Archives of Computational Methods in Engineering 2023;30:3113–32.

53. Liang M, Hao C, Ming G. Prostate cancer grade using self-supervised learning and novel feature aggregator based on weakly-labeled gbit-pixel pathology images. Applied Intelligence 2024;54:871–85.
54. Yang Z, Wang X, Xiang J, Zhang J, Yang S, Wang X, et al. The devil is in the details: a small-lesion sensitive weakly supervised learning framework for prostate cancer detection and grading. Virchows Archiv 2023;482:525–38.
55. Feng Y, Atabansi CC, Nie J, Liu H, Zhou H, Zhao H, et al. Multistage fully convolutional network for precise prostate segmentation in ultrasound images. Biocybernetics and Biomedical Engineering 2023;43:586–602.
56. Singh SK, Sinha A, Singh H, Mahanti A, Patel A, Mahajan S, et al. A novel deep learning-based technique for detecting prostate cancer in MRI images. Multimedia Tools and Applications 2024;83:14173–87.
57. Balaha HM, Shaban AO, El-Gendy EM, Saafan MM. Prostate cancer grading framework based on deep transfer learning and Aquila optimizer. Neural Computing and Applications 2024:1–26.
58. Lucas M, Jansen I, Savci-Heijink CD, Meijer SL, de Boer OJ, van Leeuwen TG, et al. Deep learning for automatic Gleason pattern classification for grade group determination of prostate biopsies. Virchows Archiv 2019;475:77–83.
59. Egevad L, Swanberg D, Delahunt B, Ström P, Kartasalo K, Olsson H, et al. Identification of areas of grading difficulties in prostate cancer and comparison with artificial intelligence assisted grading. Virchows Archiv 2020;477:777–86.