Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.065230

ARTICLE

Automated Gleason Grading of Prostate Cancer from Low-Resolution Histopathology Images Using an Ensemble Network of CNN and Transformer Models

Md Shakhawat Hossain[1,2,*,#], Md Sahilur Rahman[2,#], Munim Ahmed[2], Anowar Hussen[3], Zahid Ullah[4] and Mona Jamjoom[5]

[1]School of Informatics, Kochi University of Technology, Kami, 782-8502, Japan
[2]RIoT Center, Independent University, Bangladesh, Dhaka, 1229, Bangladesh
[3]Department of Histopathology, Armed Forces Institute of Pathology, Dhaka, 1216, Bangladesh
[4]Information Systems Department, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, 11432, Saudi Arabia
[5]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, 11671, Saudi Arabia

*Corresponding Author: Md Shakhawat Hossain. Email: md.shakhawat@kochi-tech.ac.jp

[#]These authors contributed equally to this work

Received: 07 March 2025; Accepted: 16 May 2025; Published: 03 July 2025

ABSTRACT: One in every eight men in the US is diagnosed with prostate cancer, making it the most common cancer in men. Gleason grading is one of the most essential diagnostic and prognostic factors for planning the treatment of prostate cancer patients. Traditionally, urological pathologists perform the grading by scoring the morphological pattern, known as the Gleason pattern, in histopathology images. However, this manual grading is highly subjective, suffers intra- and inter-pathologist variability and lacks reproducibility. An automated grading system could be more efficient, with no subjectivity and higher accuracy and reproducibility. Automated methods presented previously failed to achieve sufficient accuracy, lacked reproducibility and depended on high-resolution images such as 40×. This paper proposes an automated Gleason grading method, ProGENET, to accurately predict the grade using low-resolution images such as 10×. This method first divides the patient's histopathology whole slide image (WSI) into patches. Then, it detects artifacts and tissue-less regions and predicts the patch-wise grade using an ensemble network of CNN and transformer models. The proposed method adapted the International Society of Urological Pathology (ISUP) grading system and achieved 90.8% accuracy in classifying the patches into healthy and Gleason grades 1 through 5 using 10× WSI, outperforming the state-of-the-art accuracy by 27%. Finally, the patient's grade was determined by combining the patch-wise results. The method was also demonstrated for $4-class$ grading and binary classification of prostate cancer, achieving 93.0% and 99.6% accuracy, respectively. The reproducibility was over 90%. Since the proposed method determined the grades with higher accuracy and reproducibility using low-resolution images, it is more reliable and effective than existing methods and can potentially improve subsequent therapy decisions.

KEYWORDS: Gleason grading; prostate cancer; whole slide image; ensemble learning; digital pathology

1 Introduction

A prostate is a small, walnut-shaped soft organ in men whose primary function is to produce seminal fluids. Healthy prostate tissue consists of non-glandular stroma and stroma-surrounding glands. The normal

glands consist of a lumen and epithelial cells. Cancer of the prostate causes uncontrolled replication of epithelial cells, which disrupts the regular arrangement of glands. The uncontrolled replication of epithelial cells replaces the stroma and lumen in high-grade cancers. Most cases of prostate cancer are found on the gland's periphery. Prostate cancer is the second-deadliest cancer in men in the United States [1,2]. The American Cancer Society estimates it will result in about 35,250 new cases and 299,010 new cases in 2024 worldwide [3].

The Gleason grading, proposed by Donald Gleason in 1974, is the most reliable and widely used diagnosis to estimate the aggressiveness of prostate cancer [4]. The International Society for Urological Pathology (ISUP) later revised this grading system in 2005 and 2014 [5]. The Gleason grading involves identifying the tissue pattern related to the tumor's architectural growth pattern and scoring it. The two most common patterns, primary and secondary, are identified. The most common pattern is the primary or major pattern, and the second most common pattern is the secondary or minor pattern. Each pattern is scored between 1 and 5 depending on its tissue morphology, according to Fig. 1. The pattern is scored one if the tissue cells or glands are dense, uniform, small and well-differentiated. If the pattern shows well-differentiated but loosely arranged glands with more spaces, then it is scored as 2. The Gleason pattern with a score of 2 shows more stroma. The pattern of distinct interpretation of cells from glands at the margins is scored as 3. The Gleason pattern, with a score of four, is characterized by poorly differentiated glands and abnormal masses of cells in the glands. Finally, the pattern that shows irregular glands or lacks glandular differentiation is scored 5. After scoring the primary and secondary pattern, they are summed up to determine the final Gleason grade, as shown in Table 1. The Gleason grade is defined as twice the primary pattern if the secondary pattern is missing.

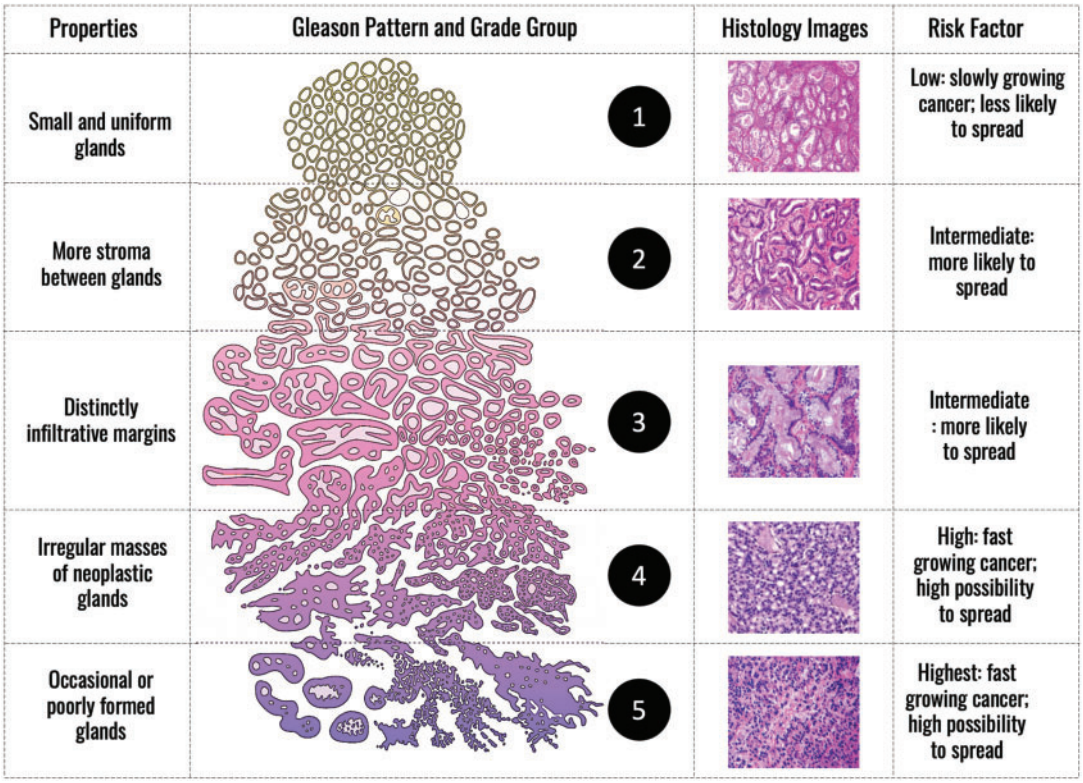

Figure 1: Gleason patterns and their corresponding grades and properties

Table 1: Grading criteria of prostate cancer based on the Gleason pattern and score

Primary Gleason pattern score	Secondary Gleason pattern score	Gleason score	ISUP grade group
3	3	6	1
3	4	7	2
4	3	7	3
4	4	8	4
3	5	8	4
5	3	8	4
4	5	9	5
5	4	9	5
5	5	10	5

Traditionally, pathologists determine the grade by assessing and scoring the tumor structural growth pattern in a Hematoxylin and Eosin (H&E) stained tissue prepared from the patient's biopsy. Previously, pathologists examined the tissue specimen using a microscope by zooming and panning the entire slide. With the advent of the whole slide imaging (WSI) technique, it is now possible to convert the entire tissue specimen into high-resolution digital images, called WSI. WSI can be observed on a computer screen and controlled using a computer mouse, reducing the labor and hustle for manual naked-eye observation of specimens. Nevertheless, the manual examination is still time-consuming, subjective and lacks reproducibility. An image-based automated system would be more practical and efficient due to its ability to significantly reduce the analysis time and labor, eliminate inter and intra-observer variability, and improve reproducibility.

This paper presents an artificial intelligence (AI)-guided image-based method for automated Gleason grading of prostate cancer patients from their H&E WSIs. The proposed method relies on an ensemble network of selected deep-learning models and utilizes the entire WSI instead of pattern-segmented regions of the specimen to predict the Gleason grade of the patient. In this paper, the proposed prostate grading ensemble network is termed ProGENET. Combining multiple deep learning models was found more effective in achieving high accuracy and reproducibility for prostate cancer grading [2] using low-resolution images. This study combined multiple deep learning models of the two most successful image classification architectures: convolutional neural network (CNN) and image transformer.

Most existing methods predicted primary and secondary Gleason scores using deep learning models and then combined the predicted scores to determine the grades. However, in this study, we trained the models to directly predict the grades from the raw pixel data of the image patches extracted from the WSI, taking advantage of deep learning techniques to map raw pixels of images to the desired outputs directly. However, training the models on vast pixel data is highly time-consuming. Therefore, in this study, we utilized a low-resolution WSI of 10× and allowed the deep learning models to process the entire image area instead of some selected regions based on the patterns. This design enabled the method to achieve high accuracy using low-resolution images. We also demonstrated the reproducibility of the proposed "ProGENET" method. In this study, we comprehensively compared CNN and image-based transformer models for Gleason grade prediction. Firstly, individual CNN and transformer models were trained and compared. Then, the ensemble of the best models was compared with the individual models to select the most suitable network. Therefore, the major contribution of this study includes: 1) the development of a highly accurate automated method for predicting the Gleason Grades from the low-resolution (10×) H&E

histopathology images, 2) a comprehensive comparison of CNN and transformer models for Gleason grading and 3) reproducibility assessment of the method to ensure its reliability.

2 Related Works

Several methods have been proposed in the last twenty years to facilitate the autonomous Gleason grading of prostate cancer from histopathology images. While many of these methods leverage deep learning architectures for automated feature extraction, some rely on traditional machine learning techniques. These automated or semi-automated prostate cancer grading methods can be broadly classified into three categories: (1) traditional machine learning-based methods, (2) deep learning-based methods, and (3) hybrid approaches that integrate both traditional and deep learning models.

Traditional machine learning models such as Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), Naive Bayes (NB) Logistic Regression (LR) and Decision Trees (DT) rely on handcrafted features extracted from the images, as opposed to the automatic feature extraction typical in deep learning models. These models require domain expertise to identify relevant features and usually perform well with structured data or smaller datasets. They are computationally efficient, easy to interpret and suitable for training the network using a comparatively small dataset. One of the very first methods for automated Gleason grading was proposed by Jafari-Khouzani and Soltanian-Zadeh in 2003 [6] and was based on a traditional machine learning model. This method utilized multiple wavelets to extract features from the images. Then, these features were processed using a KNN classifier to determine the Gleason pattern score for each image with 97% accuracy. Jafari-Khouzani reported that using multiple wavelet functions enhances feature extraction. Unlike traditional wavelets, multi-wavelets can simultaneously offer orthogonality, symmetry, and better edge handling, leading to more accurate image texture classification. This method extracted features from the entire image instead of focusing on individual cells or gland structures. This comprehensive feature extraction could contribute to the high classification accuracy of the proposed method. Although this method achieved high accuracy, their dependency on the $100X$ images makes the system highly time-consuming and unfit for routine use.

Later, in 2007, Tabesh et al. trained an SVM model to classify histopathology images into low and high-risk Gleason grades based on the images' color, texture, and morphometric features [7]. They utilized 20× images captured by mounting the digital camera on the microscope. This method used a sequential forward feature selection method to identify the sub-optimal features for the SVM classifier and achieved 81.0% classification accuracy. Another SVM-based grading was proposed by Alexandratou et al. [8]. They relied only on the gray level co-occurrence matrix (GLCM) for extracting the texture features and utilized them for investigating the performance of 16 traditional machine learning classifiers. The features for training the classifiers were selected using a feature selection method. This study also reported the SVM most suitable among the traditional machine learning models for grading the histopathology images when trained with optimally selected GLCM features. Shakhawat et al. [9] reported similar findings like the Tabesh [7] and Alexandratou [8] for a different medical image analysis application. Alexandratou et al. utilized 20× histopathology images, which were captured using the digital camera mounted on the microscope and achieved 77.8% accuracy for 4 – $class$ Gleason grading. This method utilized 20× images, which is suitable for practical use. However, the accuracy is not sufficient.

Similarly, Khurd et al. proposed another SVM-based method utilizing Spatial Pyramid Match Kernel and texture features [10]. They achieved 88.8% accuracy but considered only Gleason grades 3 and 4. Xu et al. also proposed an SVM-based classification of prostate texture features [11]. This method utilized a 2.5× WSI, significantly reducing the computation time; however, it failed to produce adequate accuracy (77.1%). Wang

proposed another SVM-based grading method utilizing the bag of local structural features but limited by the accuracy [12].

Deep learning-based models can automatically learn relevant features from raw images, reducing the need for manual feature extraction and domain expertise. These networks, such as CNN and transformers, can learn and model complex, non-linear relations within the images that may be challenging for traditional models. Gummeson et al. [13] proposed a CNN-based approach to directly predict the Gleason grade for the 40× image patches of the prostate tissue. They utilized small convolutional filters to replace the traditional feature extraction with automatic learning of principal features from the images. Despite the limited dataset of 213 images, the study achieved a 7.3% error rate through four-fold cross-validation, demonstrating the potential of the CNN model for accurate Gleason grade prediction. This method provides a coarse image segmentation; each segmentation belongs to one of the four classes: benign and Gleason grades 3 through 5. Another entirely deep learning-based method was proposed by Arvaniti et al. [14]. They proposed a MobileNet-based CNN model that utilizes depthwise separable convolutions instead of standard convolutions. In depthwise separable convolution, a single filter is applied per input channel, unlike the standard convolution, where filters are applied for all channels simultaneously, making it computationally expensive. They applied the MobileNet model on the tissue microarray (TMA) specimen to predict the Gleason grades, which achieved only 65.0% accuracy. TMA allows access to multiple tissue blocks in a single slide, but it is costly and takes a long time to prepare compared to the standard H&E histopathology specimens.

Moreover, such a large population of tissue specimens is not always available. Therefore, TMA-based Gleason grading is not suitable for routine practical use. Strom et al. [15] proposed another deep learning-based method in which they developed an ensemble network of Inception-based CNN models. This Inception model ensemble was first tested to classify the images into benign and malignant classes. Later, the malignant classes were sub-classified into Gleason grades 3 through 5. This model was trained using 10× images and achieved 99.9% and 62% accuracy for $2-class$ and $4-class$ classification, respectively. Another two-stage Gleason grading method utilizing the deep learning technology was proposed by Bulten et al. [16]. This method first segmented the tumor areas using a U-Net model, and then the segmented areas were classified using another deep learning-based classifier. This method considered only Gleason grades 3 through 5 and excluded Gleason grades 1 and 2. Yet, this method struggled to differentiate between adjacent Gleason grades (e.g., grades 3 and 4), leading to occasional misclassification. Li et al. [2] proposed another Gleason grading method utilizing the TMA specimen like the Arvaniti et al. [14]. This method focused on pixel-level classification of prostate TMA for segmenting Gleason pattern regions and then assigning Gleason grades. For this purpose, they developed a multiscale parallel branch CNN architecture that combined atrous spatial pyramid pooling and multiscale standard convolution for the dual focus on pattern region segmentation and Gleason grade prediction. However, the method's accuracy was limited to 77.2% for predicting the Gleason grades.

Silva-Rodriguez proposed a weakly-supervised CNN method for the semantic segmentation of prostate histology images based on the Gleason scores [17]. This method utilized global Gleason scores given by pathologists for training, eliminating the need for pixel-level annotations. The weakly supervised network performs multi-class segmentation through a global aggregation layer, providing localized cancerous pattern detection at the pixel level. Although this reduced the burden of massive data annotation, it achieved a Cohen's quadratic kappa of only 0.67. Li et al. proposed a region-based CNN model to differentiate between low-grade and high-grade prostate cancers [18]. This model leverages a multi-task approach with two main components: an Epithelial Network Head and a Grading Network Head, each optimized for different subtasks within prostate cancer diagnosis. The Epithelial Network Head focuses on detecting cancerous epithelial cells, while the Grading Network Head performs Gleason grading by examining glandular patterns to classify

cancer aggressiveness. One of the primary advantages of this multi-task R-CNN is its ability to incorporate contextual information from both network heads, significantly enhancing model performance compared to single-task models. This approach achieved an overall pixel accuracy of 89.4% for grading prostate cancer.

Hybrid networks are those that combine traditional and deep learning models. Källén et al. [19] proposed another method realizing the suitability of SVM for classifying the Gleason patterns. However, their method relied on the CNN architecture for automatic feature extraction. Then, they classified the CNN-extracted features using the SVM classifier. The SVM is traditionally a binary classifier. Therefore, they trained four SVM classifiers and combined them to perform a $4 - class$ grading of prostate cancer. This method achieved 81.1% accuracy in predicting the Gleason grades for the $10X$ image patches extracted from the WSI. Although this method predicted the scores from comparatively low-resolution images, further improving its accuracy and usability is necessary to determine Gleason grades. Ren et al. proposed another hybrid method. They first segmented the glandular regions from the prostate tissue and then classified them using an RF classifier to predict the Gleason grades [20]. This method utilized 20× images to predict the grades with 83.0% accuracy.

Nagpal et al. [21] proposed a two-stage Gleason grading method, combining deep learning technology with traditional machine learning. In the first stage, they predicted the Gleason scores for the image patches using a CNN model. Then, the patch-wise scores were aggregated to determine the slide-level Gleason grading using a KNN model. This method utilized 20× WSI and achieved 71.2% accuracy. Karimi et al. proposed another TMA-based Gleason grading method [22]. This method combined the predictions from multiple CNNs, each analyzing different image patch sizes, with a logistic regression model for the final prediction. New data augmentation techniques were also proposed to improve accuracy, achieving 92% in distinguishing cancerous from benign tissue and 86.0% in grading low vs. high Gleason scores. Very recently, in 2024, Koziarski et al. [23] proposed another hybrid model in which they combined multiple CNN models for grading. Their ensemble network was evaluated for $2 - class$, $4 - class$, and $8 - class$ grading of prostate cancer using 10× images and it achieved 91.8%, 83.3%, and 63.2% accuracy, respectively. This study also investigated the suitability of image-based transformer models for Gleason grading. This method considered all of the Gleason grades $R1$ through $R5$. More importantly, it was designed to process the entire histopathology WSI and handle tissue-less areas and artifacts for grading the image patches. This is important for an automated Gleason grading method and aligns with the objective of our study. Therefore, in our study, we considered this method the state-of-the-art (SOTA).

The review of the existing image-driven automated Gleason grading methods reveals that traditional machine learning-based methods relied heavily on handcrafted features. These methods performed well with optimally selected features for small datasets. Although these methods achieved reasonable accuracy (e.g., 81%–88%) when trained on high-resolution images (e.g., 20× or higher), they lacked robustness across all Gleason grades. For large and complex datasets, the accuracy of these methods dropped. Traditional machine learning-based models tend to be computationally efficient and interpretable. In contrast, deep learning-based methods achieved higher accuracy than traditional methods when trained on large datasets due to their ability to learn suitable features for binary classification of prostate cancer automatically. However, their performance decreased as the number of Gleason grades increased. Most deep learning-based methods depended on high-resolution images (20×, 40×, or 100×) for acceptable accuracy. The hybrid methods aimed to combine deep learning's automated feature learning capability and traditional machine learning models' simplicity and interpretability. Hybrid methods achieved moderate to high performance (up to 92%) in grading prostate cancer, depending on the architecture and dataset. This review also reveals that the emphasis is shifting toward methods that work on low-resolution images, which are more practical and resource-efficient for clinical use.

Although several AI and image-guided methods were proposed for automating the Gleason grading system, most failed to achieve sufficient accuracy. Some of the methods predicted the Gleason pattern scores or excluded low-risk grades such as grade groups 1 and 2; therefore, evaluating the suitability of these methods in determining the comprehensive Gleason grade for the histopathology image is necessary. The transformer-based models were utilized on a single occasion; therefore, further investigation of their suitability for Gleason grading is necessary. Multiple studies have reported the superiority of transformer models compared to CNN models for diverse medical image classification tasks. Another major problem with the existing methods is their dependency on high-magnification images to achieve sufficient accuracy. Methods proposed by Jafari-Khouzani and Soltanian-Zadeh [6], Gummeson et al. [13] and Bulten et al. [16] achieved sufficient accuracy when trained on 100×, 40× and 20× images, respectively. On the other hand, methods that were trained using low-resolution images, such as 10×, produced lower accuracy, such as 80%. Again, these methods still need to report reproducibility, which is significant in ensuring consistency and reliability. Methods based on the TMA specimen are not suitable for routine use. Alternatively, methods that utilize a single biopsy histopathology image are more practical for routine use. The review of the existing methods revealed that deep learning-based methods tend to achieve higher accuracy than traditional machine learning models when trained using the same resolution of images. All the existing deep learning methods are based on convolution-based architecture. None of the studies investigated the suitability of transformer-based deep learning architectures except the Koziarski et al. [23]. In our study, we trained popular CNN and transformer models and compared them to evaluate their suitability for automated Gleason grading from histopathology images. Among the traditional machine learning models, RF and SVM were reported to achieve better accuracy by multiple studies [8,19,20]. In this study, we trained the CNN and transformer models to predict the Gleason grades individually and ranked them based on their performances. Then, these models were combined based on their rank to form an ensemble network in which a traditional machine-learning model predicted the final grade of the input image.

3 Materials and Methodology

3.1 Dataset

In this study, we have utilized the DiagSet-A https://github.com/michalkoziarski/DiagSet (accessed on 15 May 2025) [23] dataset created by Koziarski et al., which consists of 238 WSIs annotated by expert histopathologists. This dataset contained patch-level Gleason grades for different resolutions. This is one of the significant advantages of this dataset compared to other prostate cancer datasets such as PANDA challenge https://www.nature.com/articles/s41591-021-01620-2#Sec19 (accessed on 15 May 2025) [24] and TCGA dataset https://www.cancerimagingarchive.net/collection/tcga-prad/ (accessed on 15 May 2025) [25]. We used the 10X version of this dataset. We converted the WSIs into image patches based on the annotations of histopathologists to categorize them into background (W), healthy tissue (N), artifacts (A), Gleason grade 1 ($R1$), Gleason grade 2 ($R2$), Gleason grade 3 ($R3$), Gleason grade 4 ($R4$) and Gleason grade 5 ($R5$). However, in this study, we proposed a different approach for the Gleason grading of the patient. We utilized only the image patches annotated as healthy tissue or assigned a Gleason grade for training the deep learning classifiers. The image patches that belonged to the background or artifacts were detected during the pre-processing step before applying the deep learning-based classifier on the WSI. We used 7800 images, including 1300 images belonging to each healthy and Gleason grade class $R1$ through $R5$. Among these images, 6480 were used for training, 720 for validation and 600 for testing the models. The distribution of each class image in the training, validation and test sets were equal.

3.2 Overview of the Proposed Method

The proposed ProGENET method determined the Gleason grades of the patients from their H&E stained WSI specimens. The Algorithm 1 shows the detailed algorithm of the proposed method and Fig. 2 shows the simplified flow chart of the method. This method utilized different resolutions of WSI for different operations to achieve sufficient Gleason grading accuracy in optimal time. Due to its vast size, processing the WSI is highly time-consuming, particularly at higher magnification. The WSI utilizes a pyramidal structure which contains different magnification of the images, starting from 1× to 40× or 60× magnification. The resolution of the images increases with the magnification and the dimension. Consequently, processing a high-resolution WSI such as 20× or 40× is significantly time-consuming compared to a low-resolution such as 1× or 2×. The proposed method utilized 1× WSI to detect and eliminate useless areas of the specimen, such as tissueless and artifact-affected areas. After that, this method detected the healthy patches and predicted the grades $R1$ through $R5$ for individual image patches at 10× magnification. Finally, the patch-wise Gleason grades were combined to determine the patient's grade based on the distribution of the patch-wise Grades. This architecture reduced the computation time significantly yet achieved higher accuracy compared to the previous method [23]. In the previous method, the authors proposed an $8-class$ classification method to classify each image patch as tissue area, healthy tissue area N, tissue artifacts A and Gleason grades $R1$ through $R5$. They have used the same resolution of the WSI for all classes. In our study, we utilized the lowest resolution WSI (1× WSI) for background and artifact detection as such analysis does not require observing detailed pixel information [26]. For grading the image patches and detecting the healthy areas, we utilized 10× WSI, which takes less computation time than the previous method that relied on 20× or 40× WSI. This study also separated the background and artifact detection from the Gleason grading, which helped reduce the classifier model's complexity and achieve better accuracy using the same image resolution compared to the Koziarski et al. [23].

Algorithm 1: Gleason grading of patient from H&E WSI

Input: $WSI_{1\times}, WSI_{10\times}, W_{th}, DoubleU-Net_{TF}, M, P, G, f$
$WSI_{1\times}$: 1× WSI, $WSI_{10\times}$: 10× WSI, W_{th}: threshold to eliminate tissue less patches
$DoubleU-Net_{TF}$: Trained DoubleU-net for tissue fold
B_n: Base models upto rank n with parameters $\{p_1, p_2, p_3, ..., p_n\}$
M: Meta model with parameters $\{g_1, g_2, g_3, ..., g_n\}$
Initialisation:
$Patch_{1\times}$ = Divide $WSI_{1\times}$ into 256 × 256 pixels blocks
while $Patch_{1\times}$! = NIL **do**
    $R, G, B \leftarrow$ Color channels of $Patch_{1\times}$
    I_{Gray}= 0.299 × R + 0.587 × G + 0.114 × B
    $W_{Pixels} \leftarrow$ Percentage of pixels ≥ 200 in I_{Gray}
    if $W_{Pixels} \leq W_{th}$ **then**
        $A = DoubleU-Net_{TF}(I_{sRGB})$
        while A! = $TRUE$ **do**
            $Patch_{10\times} \leftarrow$ Get 10× patch from $WSI_{10\times}$ using the coordinates of $Patch_{1\times}$
            for $C = R, G, B$ channel of $Patch_{10\times}$ **do**
                if $C_{Linear} \leq 0.0031308$ **then**
                    $C_{sRGB} = 12.92 \times C_{Linear}$
                else
                    $C_{sRGB} = 1.0552 \times C_{Linear}^{\frac{1}{2.4}}$

(Continued)

Algorithm 1 (continued)

end if
$I_{sRGB} = C_{sRGB}$
end for
while $i \leq n$ **do**
$B_i \leftarrow$ Load i^{th} base model with p_i
$f_i = B_i(I_{sRGB})$
$i = i + 1$
end while
$y = M(\mathbf{f_i}^T)$
MAX=*maximum(y)*
if $y(1)$==MAX **then** $\psi \leftarrow$R1
end if
if $y(2)$==MAX **then** $\psi \leftarrow$ R2
end if
if $y(3)$==MAX **then** $\psi \leftarrow$ R3
end if
if $y(4)$==MAX **then** $\psi \leftarrow$ R4
end if
if $y(5)$==MAX **then** $\psi \leftarrow$ R5
end if
if $y(6)$==MAX **then** $\psi \leftarrow$ N
end if
end while
end if
end while
$Grade = max - vote(\psi)$
return *Grade*

The proposed ProGENET method first divided the WSI into non-overlapping image patches of 256 pixels × 256 pixels. Then, we detected the tissue-less image patches or backgrounds,W, using 1× resolution. For that purpose, we counted the number of pixels in the image having an intensity value higher than 200. We did this for the gray-scale version of the image. If the percentage of such pixels exceeds 50 in a patch, the proposed method eliminates it from the Gleason grading as it does not have enough tissue. Several methods are available for detecting the tissue-less glass patches, such as based on the images' saturation or the pixels' optical density values. However, we utilized the pixel intensity-based method in this study due to its simplicity and low computation time.

After eliminating the tissue-less patches, the proposed method processed the 1× patches to detect the tissue fold artifact-affected patches, A. Tissue fold is a common artifact in histopathology images produced during the glass slide preparation. Tissue folds are multiple layers of tissue and show texture and color features similar to the cancer regions, which often confuses computerized algorithms. Pathologists ignore such areas during the analysis and diagnosis. Therefore, we detected and eliminated such patches from Gleason grading. In this study, we utilized the DoubleUnet-based tissue fold segmentation method proposed by [26]. The artifact segmentation method proposed by Rubina et al. utilized 1× image patches and resulted in very low false positives. Therefore, we adopted this method for our study. Their study first segmented the tissue folds

using a DoubleUnet model and then determined their severity using a CNN classifier to exclude highly severe artifacts. However, we adopted the DoubleUnet model for the artifact segmentation method in our method, except for the severity classifier. Our study eliminated all the artifact-affected patches regardless of their severity.

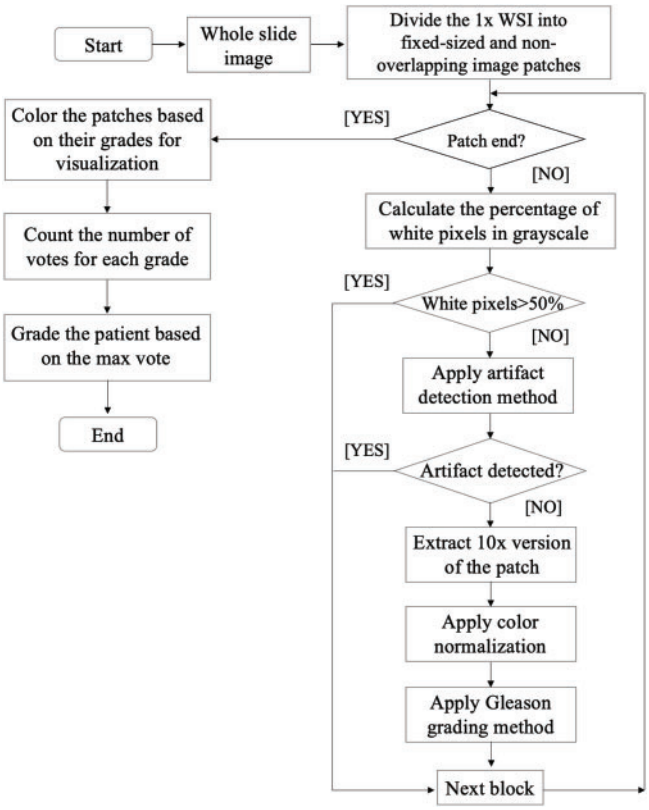

Figure 2: Flowchart of the proposed Gleason grading method

After detecting and eliminating the useless patches, we graded the rest of the image patches, T, where $T = 1 - \{W \cup A\}$ at 10× magnification. For this purpose, firstly, we normalized the image's color values by converting the RGB values to $sRGB$. Then, the image patches, T, were processed using the ensemble classifier to predict their Gleason grade. Finally, the patch-level grades of the WSI were used to determine the patient's grade based on the maximum voting. If the WSI has mostly $R3$ grades at the patch level, the WSI is graded as $R3$. The proposed method directly predicted the Gleason grade per image patches of the histopathology specimen using an ensemble of deep learning networks, unlike the traditional method in which the primary and secondary Gleason patterns are detected first. Then, the grade is derived based on the pattern scores. The proposed method does not rely on handcrafted feature selection but instead uses features automatically learned through the ensemble of CNN and Transformer models.

Further, in this study, we applied the proposed ensembled method for $4 - class$, which included healthy patch N, artifact A, Gleason grade $R1$ and Gleason grade $R2$ and $2 - class$ grading, which included cancerous patch CN and others (tissue less patch W, healthy patch N and artifact A). It allowed us to comprehensively compare the proposed method and the SOTA [23]. We also tested the reproducibility of the proposed method. For that purpose, we tested the same set of images using the proposed method three times and checked the method's consistency. The architecture of the proposed method is shown in Fig. 3.

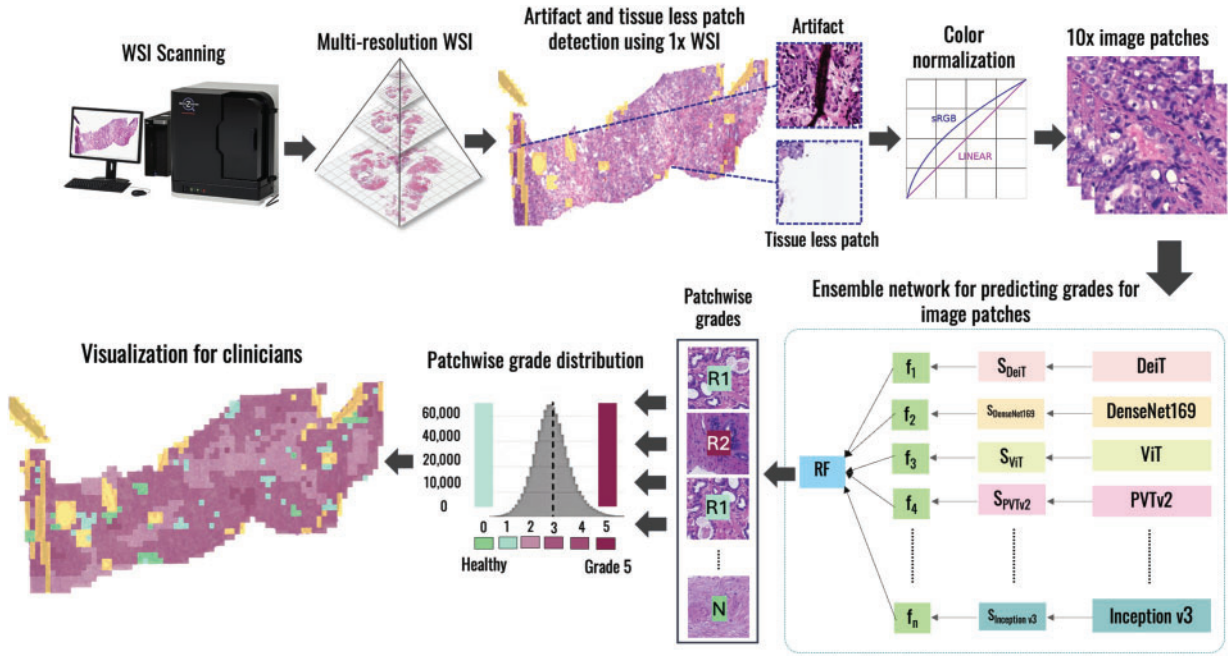

Figure 3: Architecture of the proposed Gleason grading system leveraging digital pathology and ensembled AI

3.3 Ensemble Classifier Development

In this study, we experimented with seven individual deep-learning models and their ensemble networks to predict the patch-wise grades. The seven models included three image-based transformer models and four popular CNN models, as shown in Table 2. This table also shows the optimization space for fine-tuning each model. In our experiment, we explored different hyperparameter values through grid search to find the best combination of hyperparameters that results in the best accuracy on the test set. We utilized only 8 and 12 attention heads for the transformer models to make the model fast and less memory-consuming. A higher number of heads allows us to learn more complex patterns in the image, often leading to higher classification accuracy. The individual models were then ranked based on their test accuracy attained by their best-fine-tuned networks.

Table 2: Hyperparameter values tested to find the best fine-tuned network for each individual models

Hyper-parameters	CNN optimization space	Transformer optimization space
Models	[VGG16, ResNet152, InceptionNetV3, DenseNet169]	[ViT, DeiT, PVTv2]
Epochs	[50, 100, 200]	[50, 100, 200]
Batch sizes	[8, 16, 32]	[16, 32, 64]
Patch sizes	–	[16, 32]
Optimizers	[SGD, Adamax, AdamW, RMSProp]	[SGD, AdamW]
Loss functions	[Categorical Cross Entropy, Binary Cross Entropy]	[Categorical Cross Entropy]
Learning rates	[0.01, 0.001, 0.0001]	[0.01, 0.001, 0.0001]
Dropouts	[0.5, 0.6, 0.7, 0.8]	–

(Continued)

Table 2 (continued)

Hyper-parameters	CNN optimization space	Transformer optimization space
Transformer layers	–	[8, 16, 32, 64]
Attention heads	–	[8, 12]
Embedding dimension	–	[768]

After that, we created ensemble networks by combining the individual networks. In the ensemble network, each network worked as a base model and processed the input image individually to provide a class prediction value. The class prediction values of the base models were then combined to form a feature vector, which is further processed using another simple classifier to predict the final grade for the image patch. We selected the top two, three, four and five-ranked models to create four combinations of base models for the ensemble network. On top of that, we experimented with six different meta classifiers, which included LR, SVM, RF, KNN, DT and Extreme Learning Machine (ELM) model. The best meta-classifiers were selected based on their accuracy on the test data. After that, we compared the individual networks with their ensemble networks to choose the best network for the proposed method. Then, the proposed method was compared with the state-of-the-art method.

4 Results

In this study, we predicted the grades of the image patches using an ensemble network of CNN and image-based transformer models, which were finally combined to determine the patient's grade. Firstly, the best-fine-tuned version of all models was compared based on their accuracy on the test dataset. Table 3 shows the training, validation and test accuracy of the best-fine-tuned version of all models. This table also compares the CNN and transformer models when trained and tested using the same dataset under similar conditions. Fig. 4 shows their box plot comparison. DeiT model achieved the highest test accuracy of 88.0%. The validation accuracy was also highest for DeiT (99.4%) with comparatively low validation loss (2.5%). Therefore, the DeiT model was ranked as the best model. DenseNet169 yielded the second-highest test accuracy of 86.9%. The ViT model achieved similar accuracy to DenseNet169 with lower validation loss and ranked third.

Table 3: Comparison between the best fine-tuned network of each model for predicting the patch-wise grades using 10× images (TrA = Train Accuracy, TL = Train Loss, VA = Validation Accuracy, VL = Validation Loss, TeA = Test Accuracy, sec = Convergence time in seconds)

Rank	Model	TA	TL	VA	VL	TeA	sec
1	DeiT	0.992	0.023	0.994	0.025	0.880	6510
2	DenseNet169	0.991	0.029	0.988	0.039	0.869	5046
3	ViT	0.991	0.027	0.988	0.021	0.865	6090
4	PVTv2	0.989	0.036	0.987	0.030	0.830	5633
5	ResNet152	0.963	0.110	0.952	0.129	0.828	3720
6	VGG16	0.960	0.119	0.950	0.150	0.756	1960
7	InceptionV3	0.947	0.162	0.937	0.181	0.696	1485

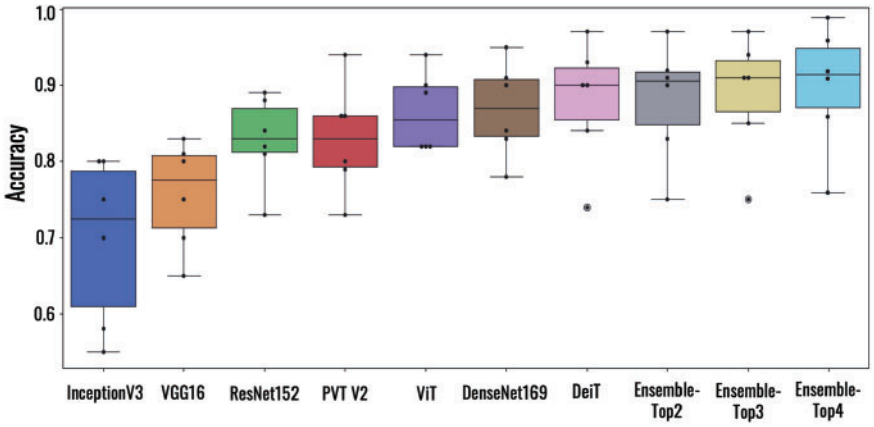

Figure 4: Comparison of the individual and ensemble models using Boxplot

Table 3 also shows the convergence time of the models. Although ViT and DenseNet169 had similar accuracies, ViT had a higher convergence time than DenseNet169. All the transformer models took a comparatively longer time to converge than CNNs when trained using the same data and computational resources; however, they had marginal differences in hyperparameters. DeiT had the highest convergence time. DenseNet169 also had a high convergence time, which could be attributed to its complex architecture and higher number of parameters. The convergence time and the accuracy for the InceptionV3 were the lowest. Figs. 5 and 6 show the validation, training accuracy, and loss curves for the top six models. However, the validation loss was comparatively lower for the CNN models than for the transformers. This indicates the data-hungry nature of transformer models, except for the DeiT model.

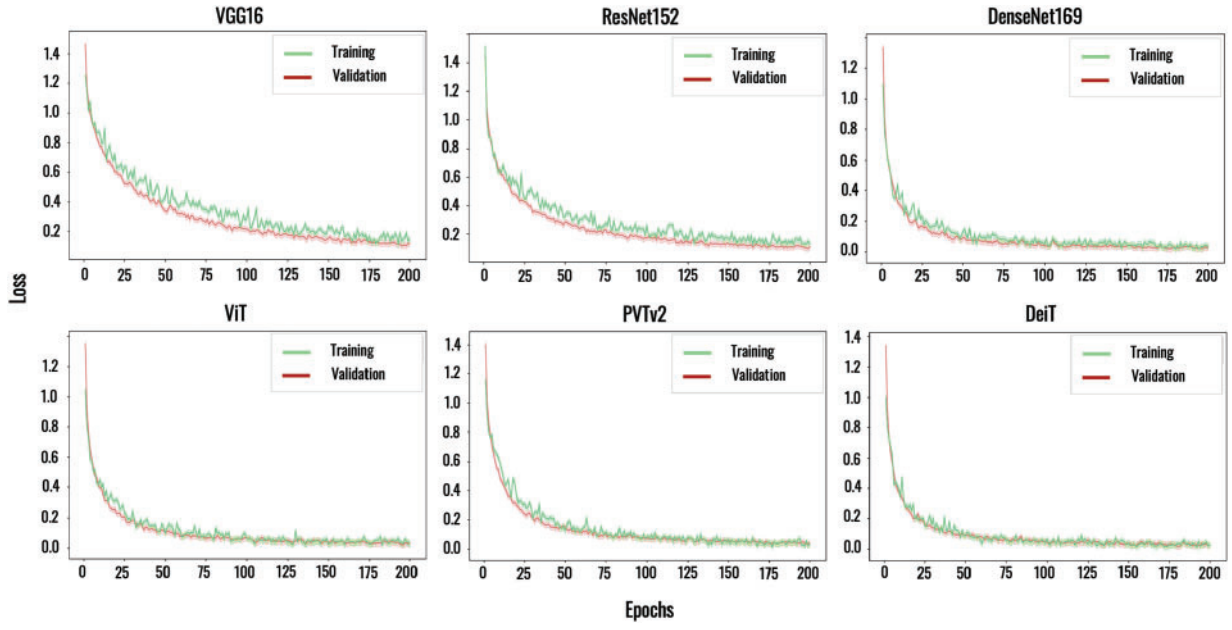

Figure 5: Training and validation loss of the models

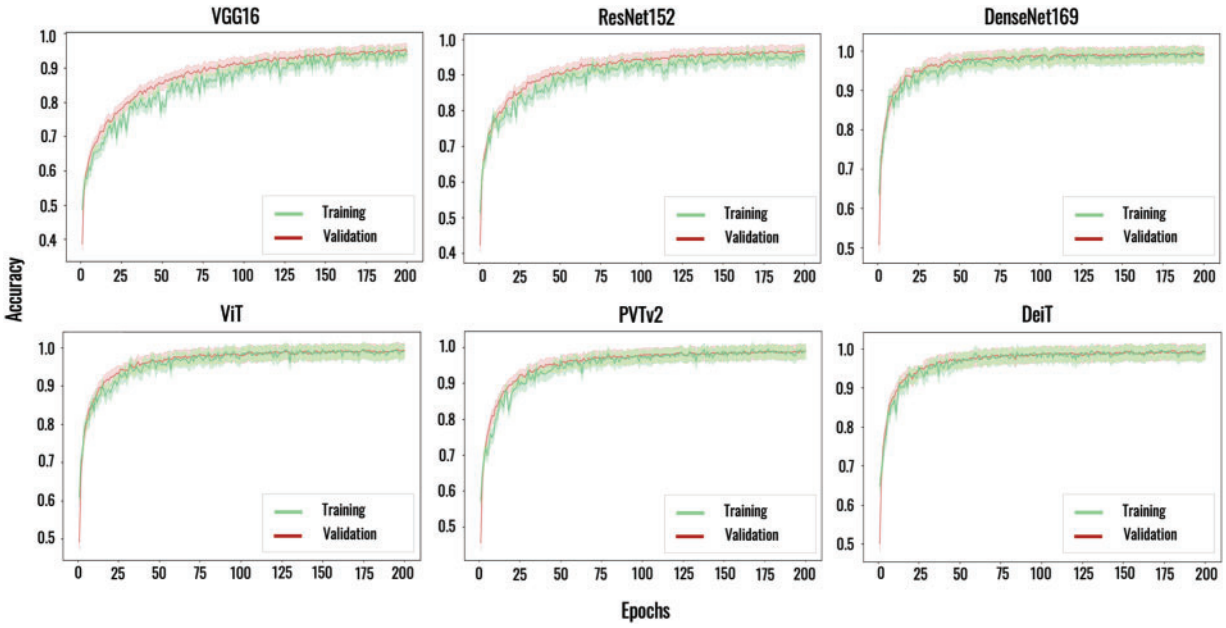

Figure 6: Training and validation accuracy of the models

After that, we tested the accuracy of the ensemble networks for which we experimented with different combinations of base models and meta models, as shown in Table 4. Then, we also compared the test accuracy of the individual models with the ensemble models. Table 5 compares the accuracy, precision, recall, F1 score and area under the curve (AUC) on the test dataset. Table 4 shows that the test accuracy increased with the number of base models used in the ensemble network, although the increment is insignificant. The accuracy of the top 4 base models and top 5 base models is indifferentiable, regardless of the meta classifier. The table also shows that the accuracy minimally changed for the different meta-classifiers, irrespective of the combination of base models. However, we selected the RF model as the meta-classifier for the proposed ensemble network as it yielded the highest accuracy of 90.8%. We also selected the combination of the top 4 ranked models as the base model, as it produced similar results to the top 5-ranked models' ensemble with fewer models.

Table 4: Gleason grading of prostate cancer using the ensemble of top base models incorporated with different meta classifiers (EOT = Ensemble of Top)

Meta classifiers	EOT 2	EOT 3	EOT 4	EOT 5
LR	0.878	0.885	0.890	0.890
SVM	0.870	0.880	0.882	0.880
RF	0.888	0.890	0.908	0.908
KNN	0.870	0.881	0.882	0.884
DT	0.878	0.885	0.890	0.890
ELM	0.853	0.855	0.860	0.855

Table 5: Comparison between the individual and ensemble models of proposed approach in 8-class classification for prostate cancer grading m (Acc. = Accuracy, AP = Average precision, AR = Average Recall, AF1 = Avgerage F1 score, MAA = Macro Avg. AUC, EOT = Ensemble of Top)

Model	Acc.	AP	AR	AF1	MAA
VGG16	0.756	0.757	0.756	0.755	0.950
ResNet152	0.828	0.829	0.828	0.828	0.980
InceptionV3	0.697	0.708	0.697	0.696	0.902
DenseNet169	0.869	0.869	0.868	0.868	0.980
ViT	0.865	0.865	0.865	0.865	0.980
PVTv2	0.830	0.832	0.830	0.829	0.970
DeiT	0.880	0.882	0.880	0.879	0.990
EOT 2	0.830	0.832	0.830	0.829	0.989
EOT 3	0.865	0.865	0.865	0.865	0.990
EOT 4	0.908	0.907	0.902	0.901	0.999
EOT 5	0.908	0.907	0.901	0.901	0.999

Figs. 7 and 8 show the receiver operating characteristic (ROC) curves and confusion matrices of the best individual models and best-ensembled models, accordingly. The results shown in these figures, along with Fig. 4 and Table 5, indicated that ensembling the models improves the performance of the proposed method. However, the ensemble model of the top $2-ranked$ models produced similar results to the individual models. Then, we modified these models for $4-class$ grading and $2-class$ grading of prostate cancer, as shown in Table 6. The accuracy of all the models significantly improved compared to the 8-class grading. The ensemble models produced higher accuracy than the individual models in both $4-class$ and $2-class$ grading. This further confirmed the suitability of ensemble models for prostate cancer grading regardless of the number of classes.

Then, we compared the results of our experiments with the existing methods, shown in Table 7. Firstly, we compared the proposed method with the SOTA method [23]. The best CNN-based model, best transformer-based models and best-ensembled models produced in our study significantly outperformed the corresponding best models of [23] using the same dataset and the same magnification of images. This justifies the design of the proposed method, which separates artifact detection and tissue-less area detection from Gleason grading. Finally, we compared the proposed method with the previously proposed AI and image-based Gleason grading methods, as shown in Table 8. This table shows that the proposed method achieved the highest accuracy of 90.8% for Gleason grading, considering all the Gleason grades and utilizing a low-resolution (10×) standard H&E histopathology image. We also investigated the reproducibility of the proposed method, shown in Fig. 9. For this purpose, we utilized the ensemble of the top $4-ranked$ models using the RF meta classifier, the best model selected as the proposed method. Fig. 9 shows that the proposed method remained highly consistent in three trials on the same dataset, indicating reliability and reproducibility.

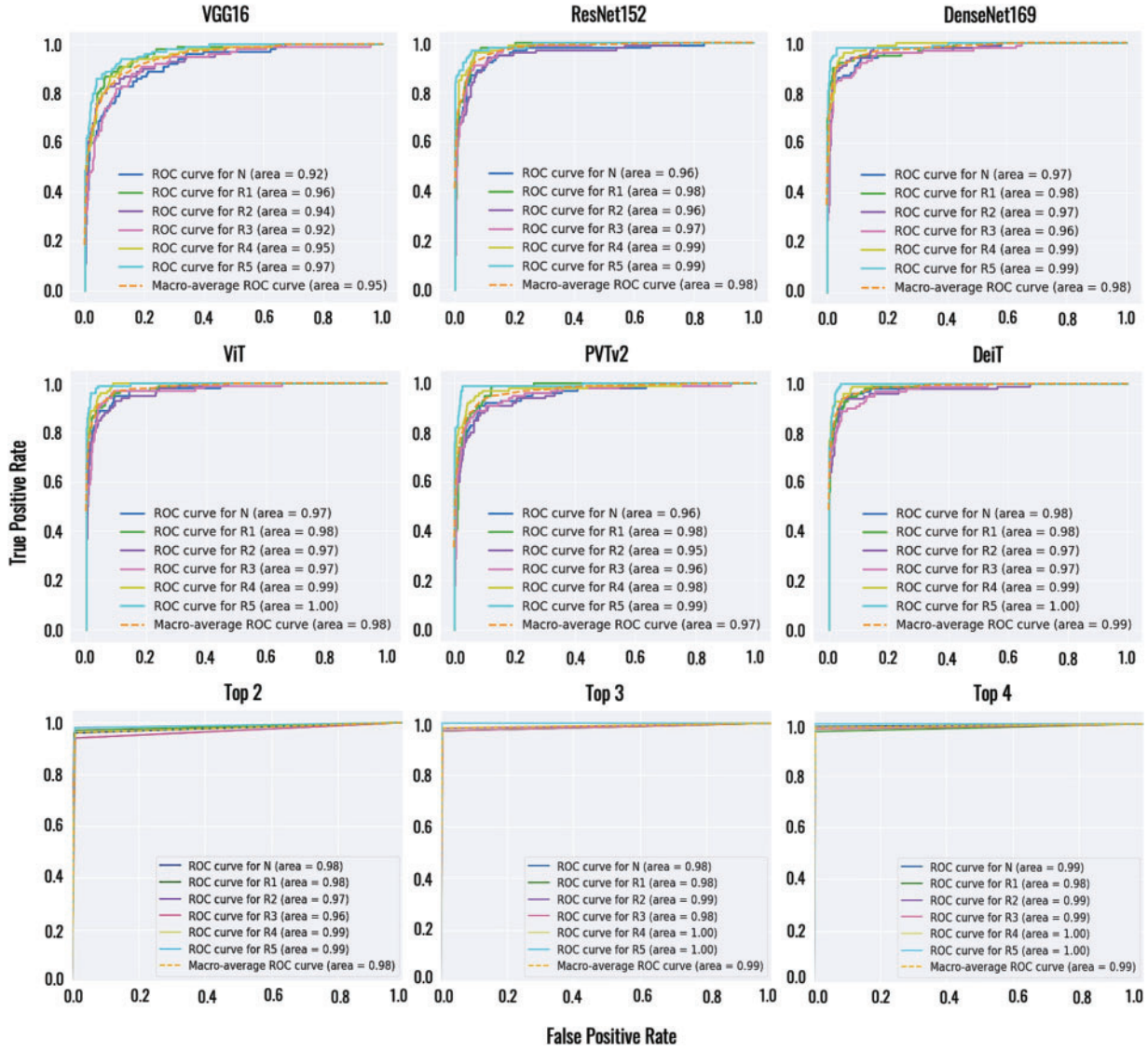

Figure 7: ROC curves of the best individual and ensemble models

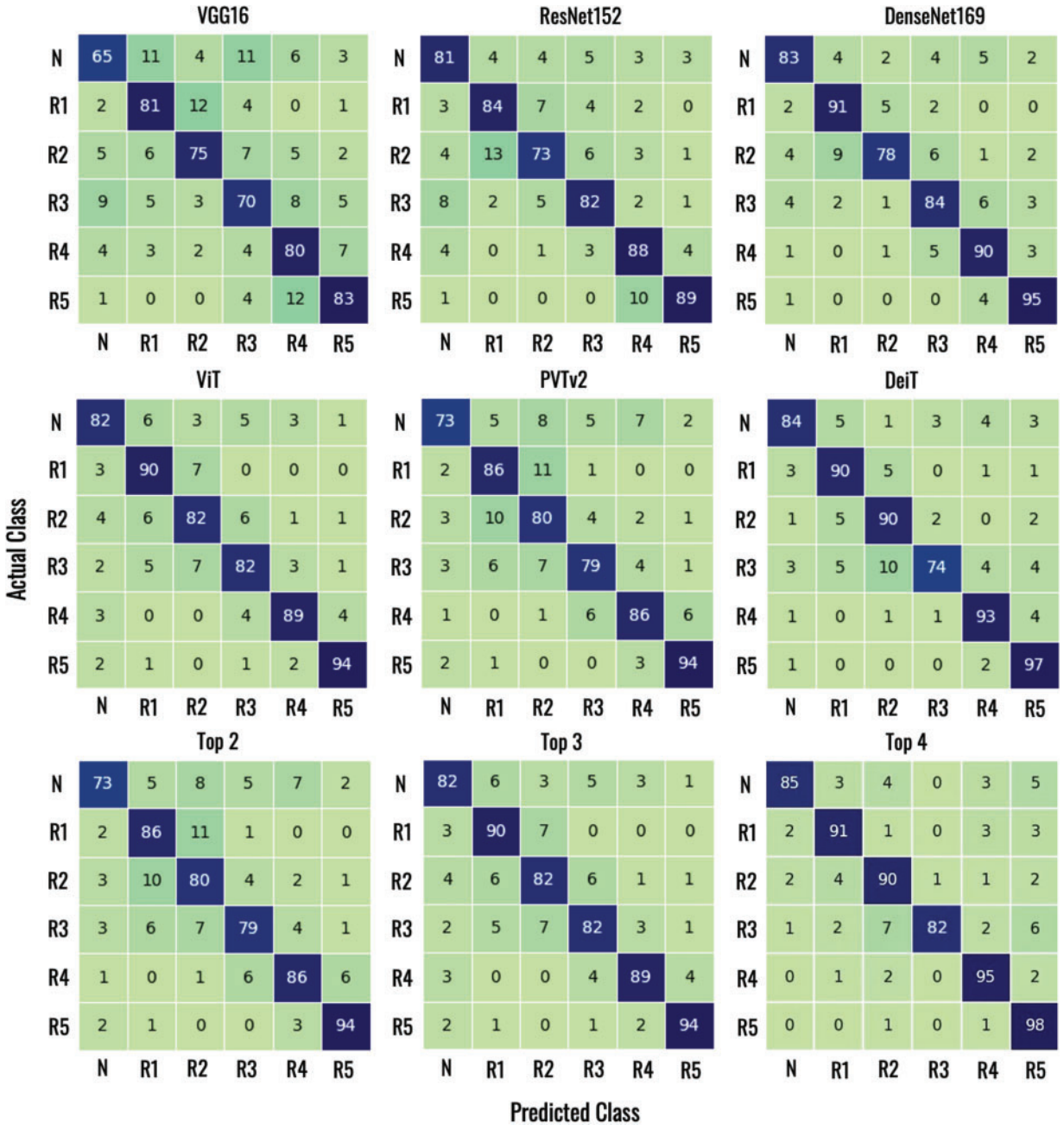

Figure 8: Confusion matrices of the best individual and ensemble models

Table 6: Test accuracy of the proposed approach for 4-class and 2-class classification of prostate cancer (4-CC = 4-Class Classification, 2-CC = 2-Class Classification, EOT = Ensemble of Top)

Methods	4-CC	2-CC
VGG16	0.906	0.913
ResNet152	0.911	0.984
InceptionV3	0.902	0.910

(Continued)

Table 6 (continued)

Methods	4-CC	2-CC
DenseNet169	0.915	0.985
ViT	0.922	0.915
PVTv2	0.910	0.902
DeiT	0.924	0.985
EOT 2	0.925	0.994
EOT 3	0.929	0.994
EOT 4	0.930	0.996
EOT 5	0.930	0.997

Table 7: Comparison between the proposed method and state-of-the-art (SOTA) method (8-CC = 8-Class Classification, 4-CC = 4-Class Classification, 2-CC = 2-Class Classification)

Methods	8-CC	4-CC	2-CC
SOTA [23] CNN Model (VGG19)	0.632	0.838	0.918
SOTA [23] Transformer Model (ViT-B/32)	0.624	0.764	0.894
Best CNN by proposed method (DenseNet169)	0.869	0.915	0.985
Best transformer by proposed method (DeiT)	0.880	0.924	0.985
Best ensembled model by proposed method	0.908	0.930	0.996

Table 8: Comparison between the proposed and existing AI and image based Gleason grading methods

Ref.	Dataset	Classes	Technology	Accuracy
[6]	100× digital histopathology images	Gleason scores 2 – 5	Traditional: Multi-wavelet based feature extraction with KNN classifier	97%
[7]	20× digital camera mounted microscopic image	Low-risk and high-risk grades	Traditional: Color, texture and morphometric features with SVM classifier	81.0%
[8]	20× digital camera mounted microscopic image	Gleason grades 1 – 4	Traditional: SVM	77.8%
[10]	10× histopathology images	Gleason grades 3 – 4	Traditional: Texture-based features with SVM classifier	88.8%
[12]	20× WSI from TCGA dataset	Gleason grades 3 – 5	Traditional: Local structure based Bag-of-features with SVM-based classification	77.4%
[19]	10× WSI from TCGA Dataset	Benign, Gleason grades 3 – 5	Hybrid: CNN-based feature extraction with SVM-based Classification	81.1%
[13]	40× WSI from the private dataset of PathXL in Belfast and Beaumont Hospital in Dublin	Benign and Gleason grades 3 – 5	Deep learning: Small CNN	92.7%
[20]	20× WSI from private dataset	Gleason grades 3 – 5	Hybrid: U-Net for glandular region segmentation and RF-based classification of Gleason pattern scores	83.0%
[14]	Tissue micro-array from private dataset	Benign and Gleason grades 3 – 5	Deep learning: MobileNet based CNN	65.0%
[18]	20× WSI from private dataset	Stroma, low-grade, high-grade. benign	Deep learning: multi-task R-CNN based segmentation	89.4%
[15]	10× WSI images	2 – *class*: benign and malignant; 4 – *class*: benign and Gleason grades 3 – 5	Deep learning: Ensemble of Inception-based CNN models	2 – *class*: 99.9%; 4 – *class*: 62.0%
[21]	20× WSI from TCGA dataset	Benign and Gleason grades 3 – 5	Hybrid: CNN classifier for predicting Gleason patterns and KNN for Gleason grading	70.0%

(Continued)

Table 8 (continued)

Ref.	Dataset	Classes	Technology	Accuracy
[11]	2.5× WSI (down-sampled from 20×) from TCGA dataset	Gleason pattern scores $6, 7\ and\ \geq 8$	Traditional: Local binary pattern based features with SVM classifier	77.1%
[22]	40× Tissue micro-array from private dataset	$2-class$: benign and malignant; $4-class$: benign and Gleason grades 3 – 5	Hybrid: Ensemble of CNN models	$2-class$: 92.0%; $4-class$: 86.0%
[16]	20× WSI	Benign and Gleason grades 3 – 5	Deep learning: U-Net for tumor segmentation and deep learning based classifier for Gleason grades 3 – 5.	91.8%
[2]	Tissue micro-array	Benign and Gleason grades 3 – 5	Deep learning: Multiscale parallel branch convolutional neural network	77.2%
[17]	40× Tissue micro-array	Background, benign and Gleason grades 3 – 5	Deep learning: Weakly-supervised method for semantic segmentation of images based on Gleason scores	67.0%
[23]	10× WSI from DiagSet Dataset	$8-classes$: Tissue less area, artifact, healthy, Gleason grade $R1-R5$; $4-classes$: Tissue less area, healthy, artifacts, Gleason grade $R1-R2$; $2-classes$: cancer and non-cancer	Hybrid: Ensemble of CNN models	$8-class$: 63.2%; $4-class$: 83.8%; $2-class$: 91.8%
ProGENET	**1×WSI for tissue less area and artifact detection and 10× WSI for healthy and Gleason grades from DiagSet Dataset**	$8-classes$: **Tissue less area, artifact, healthy, Gleason grade $R1-R5$; $4-classes$: Tissue less area, healthy, artifacts, Gleason grade $R1-R2$; $2-classes$: cancer and non-cancer**	**Hybrid: Ensemble of CNN and transformer models**	**$8-class$: 90.8%; $4-class$: 93.0%; $2-class$: 99.6%**

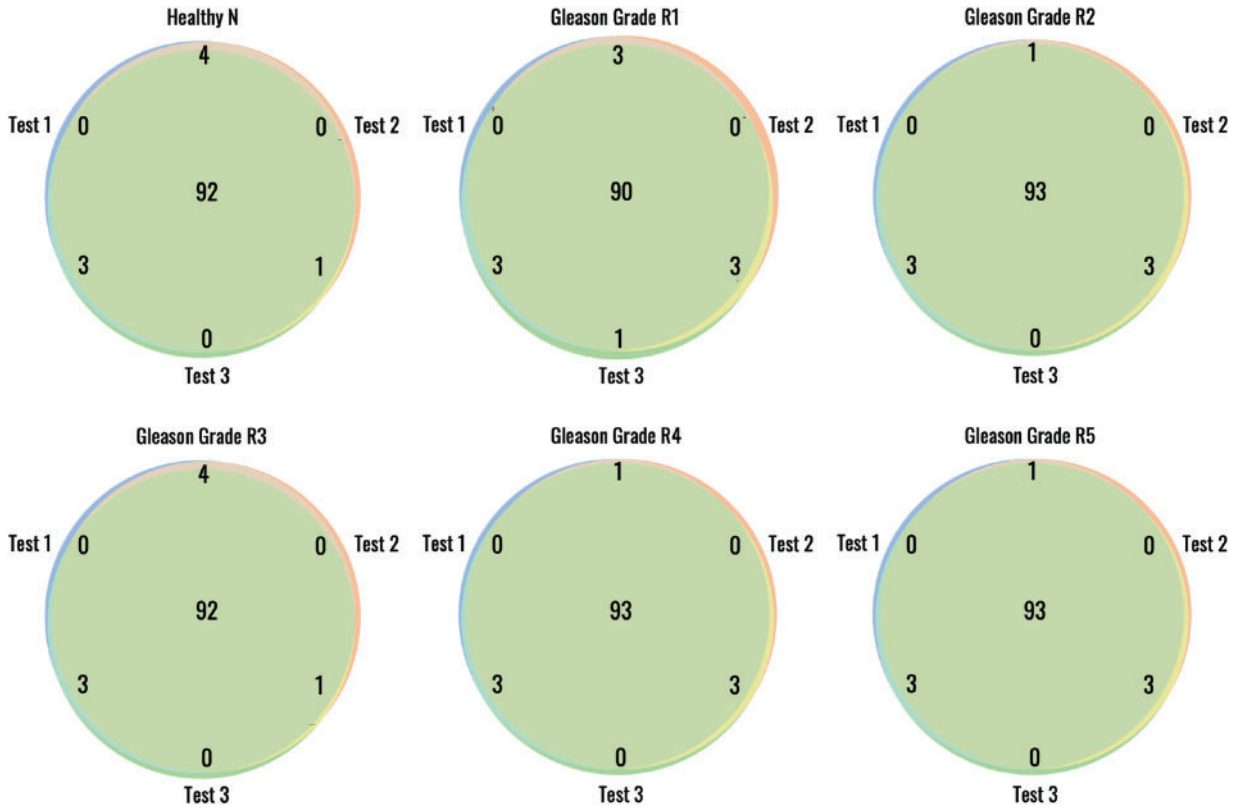

Figure 9: Reproducibility of the proposed method for patch wise Gleason grading from 10× histopathology images

5 Discussion

Gleason grading is routinely performed to diagnose and assess the prognosis of prostate cancer patients. It evaluates the morphological pattern of prostate tissue under the microscope and assigns a score that predicts how aggressive the tumor is. However, pathologists' traditional method of manual Gleason grading has several limitations related to subjectivity, accuracy and reproducibility. Additionally, this manual assessment is highly laborious and time-consuming and with an increasing number of patients, pathologists often need more time to handle heavy workloads, which may impact consistency. Automated AI-guided Gleason grading methods can address these challenges by offering more accurate, consistent, objective and efficient grading, reducing grading time and manual efforts. However, the existing AI-driven methods produced limited accuracy and required high-resolution biopsy images, which affected the grading time.

In this paper, we proposed an ensemble-based AI method that integrated an optimally selected CNN and transformer model to determine the Gleason grade of the patient from a low-resolution image of 10× WSI. The proposed method separated the tissue-less area and artifact detection from the Gleason grading, which yielded high accuracy. This method utilized multi-resolution WSI and detected the tissue less and artifact image patches from 1× WSI. Using 1× images for artifact and tissueless area detection and 10× images for Gleason grading significantly reduced the total evaluation time. In the demonstration, the proposed method achieved a high accuracy of 90.8% and consistency over 90% regardless of the grades. The proposed method was also demonstrated for 4-class grading and binary classification to separate prostate cancer from non-cancerous regions and it outperformed the existing method.

The suitability of transformer models for Gleason grading remained unexplored except for the study of Koziarski et al. [23], which experimented with ViT only. In this study, we conducted a detailed survey of transformer models, including the most popular image-based transformers. It also comprehensively compared CNN and transformer models for predicting the patch-wise Gleason grades. The results of this study report that the transformer-based model achieves better accuracy than most CNN models when trained using the same dataset. However, the transformer-based models took a longer time to converge. Occasionally, transformer models are over-fitted except for the DeiT model. This study also compared the performance of individual models with the ensemble of multiple models. This study finds that the ensemble model produces higher accuracy and more consistent results than the individual models, as demonstrated by the ROC curves, confusion matrices and boxplot.

Most of the existing automated Gleason grading methods excluded grades 1 and 2 because these grades represent low-risk, non-aggressive forms of prostate cancer that are rarely diagnosed in clinical practice. However, to develop a fully automated Gleason grading system, the grading method should include all the grades regardless of their low association with aggressive cancer; it is also necessary to integrate automated background and artifact elimination methods with the grading. The proposed method incorporated background and artifact area detection and considered all grades guided by the ISUP grading scheme. This ensures the efficacy of the proposed method for automated Gleason grading for routine use.

One of the major limitations of the ProGENET method is its high computation time for training. Training an ensemble of multiple deep-learning models is time-consuming and requires high computing resources. However, ensemble models' prediction time is not significantly higher than that of individual models. The individual transformer models were occasionally over-fitted; however, as we have utilized the ensemble, the model over-fitting is not an issue for the proposed method. Another limitation of the current study is that the proposed method was evaluated only on the DiagSet-A dataset [23]. While the results demonstrate high accuracy, further validation on external datasets such as PANDA and TCGA is necessary to assess the model's generalizability and domain transferability. We are currently developing a clinical

application of the system for hospital use, where it will undergo further evaluation in real-world deployment scenarios. This external validation and clinical testing remain a part of our future work.

To evaluate the feasibility of the proposed system for clinical use, we are currently implementing the proposed system using a standard GPU setup. In our experiments, the method required less than 10 s for a WSI, including artifact detection, tissue-less patch filtering, color normalization and patch-wise grading time. This provides a preliminary estimate of the system's overall processing time per case. However, once the system is deployed in a clinical setting, pathologists will demonstrate it, allowing us to accurately measure the actual inference time per patient case. This will offer a more practical assessment of the system's usability in routine diagnostic workflows. The proposed system utilizes a modular framework that can be parallelized for scalability, making it adaptable to high-throughput hospital systems. While the current implementation is optimized for research-grade GPUs, future work could explore lightweight variants of our model and techniques like tensor decomposition could be utilized. Another future work of this study includes investigating the impact of the proposed method on the effective selection of patients for therapy.

6 Conclusion

This paper presented an automated method for grading prostate cancer patients. This method can predict the grades from low-resolution WSI of 10× and achieve high accuracy and consistency in the demonstration. Therefore, the proposed method can improve prostate cancer patients' Gleason grading accuracy and reliability and eliminate inter- and intra-observer variability, eventually improving the subsequent therapy decisions.

Acknowledgement: The authors express their gratitude to Dr. Sahria Bakar of Jessore Medical College for their assistance in preparing the image dataset and evaluating the results of our experiments.

Funding Statement: This work was supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R104), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author Contributions: Conceptualization, Md Shakhawat Hossain and Anowar Hussen; methodology, Md Shakhawat Hossain, Md Sahilur Rahman and Munim Ahmed; software, Md Sahilur Rahman and Munim Ahmed; validation, Anowar Hussen, Zahid Ullah and Mona Jamjoom; formal analysis, Md Shakhawat Hossain; investigation, Md Sahilur Rahman; resources, Zahid Ullah; data curation, Anowar Hussen; writing—original draft preparation, Md Shakhawat Hossain and Md Sahilur Rahman; writing—review and editing, Zahid Ullah and Mona Jamjoom; supervision, Md Shakhawat Hossain; project administration, Anowar Hussen; funding acquisition, Zahid Ullah. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used for this study are available in the Dataset link (accessed on 15 May 2025). The code used in this manuscript is available in this Code link (accessed on 15 May 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. National Cancer Institute. Cancer stat facts: common cancer sites. SEER [Online]. [cited 2025 May 15]. Available from: https://seer.cancer.gov/statfacts/html/common.html.
2. 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. doi:10.1109/access.2020.3005180.

3. American Cancer Society. Key statistics for prostate cancer [Online]. [cited 2025 May 15]. Available from: https://www.cancer.org/cancer/types/prostate-cancer/about/key-statistics.html.
4. Gleason D, Mellinger G. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. J Urology. 1974;111:58–64. doi:10.1016/s0022-5347(17)59889-4.
5. Faraj S, Bezerra S, 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. doi:10.1371/journal.pone.0146189.
6. Jafari-Khouzani K, Soltanian-Zadeh H. Multiwavelet grading of pathological images of prostate. IEEE Transact Biomed Eng. 2003;50:697–704. doi:10.1109/TBME.2003.812194.
7. Tabesh A, Teverovskiy M, Pang H, Kumar V, Verbel D, Kotsianti A, et al. Multifeature prostate cancer diagnosis and Gleason grading of histological images. IEEE Transact Med Imag. 2007;26:1366–78. doi:10.1109/TMI.2007.898536.
8. Alexandratou E, Atlamazoglou V, Thireou T, Agrogiannis G, Togas D, Kavantzas N, et al. Evaluation of machine learning techniques for prostate cancer diagnosis and Gleason grading. Int J Comput Intellig Bioinform Syst Biol. 2010;1:297–315. doi:10.1504/IJCIBSB.2010.031392.
9. Shakhawat H, Nakamura T, Kimura F, Yagi Y, Yamaguchi M. Automatic quality evaluation of whole slide images for the practical use of whole slide imaging scanner. ITE Transact Media Technol Applicat. 2020;8(4):252–68. doi:10.3169/mta.8.252.
10. Khurd P, Bahlmann C, Maday P, Kamen A, Gibbs-Strauss S, Genega E, et al. Computer-aided Gleason grading of prostate cancer histopathological images using texton forests. In: 2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2010; Rotterdam, The Netherlands. p. 636–9.
11. Xu H, Park S, Hwang T. Computerized classification of prostate cancer gleason scores from whole slide images. IEEE/ACM Transact Computat Biol Bioinform. 2019;17:1871–82. doi:10.1109/TCBB.2019.2941195.
12. Wang D, Foran D, Ren J, Zhong H, Kim I, Qi X. Exploring automatic prostate histopathology image gleason grading via local structure modeling. In: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2015; Milan, Italy. p. 2649–52.
13. Gummeson A, Arvidsson I, Ohlsson M, Overgaard N, Krzyzanowska A, Heyden A, et al. Automatic Gleason grading of H and E stained microscopic prostate images using deep convolutional neural networks. In: Medical Imaging 2017: Digital Pathology; 2017; Orlando, FL, USA. Vol. 10140, p. 196–202.
14. Arvaniti E, Fricker K, Moret M, Rupp N, Hermanns T, Fankhauser C, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. Scient Rep. 2018;8:12054. doi:10.1038/s41598-018-30535-1.
15. Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney D, et al. Pathologist-level grading of prostate biopsies with artificial intelligence. arXiv:1907.01368. 2019.
16. Bulten W, Pinckaers H, Boven H, Vink R, Bel T, Ginneken B, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol. 2020;21:233–41. doi:10.1016/S1470-2045(19)30739-9.
17. Silva-Rodriguez J, Colomer A, Naranjo V. WeGleNet: a weakly-supervised convolutional neural network for the semantic segmentation of Gleason grades in prostate histology images. Comput Med Imaging Graph. 2021;88(1):101846. doi:10.1016/j.compmedimag.2020.101846.
18. Li W, Li J, Sarma K, Ho K, Shen S, Knudsen B, et al. Path R-CNN for prostate cancer diagnosis and gleason grading of histological images. IEEE Transact Med Imag. 2018;38:945–54. doi:10.1109/TMI.2018.2875868.
19. Källén H, Molin J, Heyden A, Lundström C, Astrom K. Towards grading gleason score using generically trained deep convolutional neural networks. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI); 2016; Prague, Czech Republic. p. 1163–7.
20. Ren J, Sadimin E, Foran D, Qi X. Computer aided analysis of prostate histopathology images to support a refined Gleason grading system. In: Medical Imaging 2017: Image Processing. 2017; Orlando, FL, USA. Vol. 10133, p. 532–9.
21. Nagpal K, Foote D, Liu Y, Chen P, 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. doi:10.1038/s41746-019-0112-2.

22. Karimi D, Nir G, Fazli L, Black P, Goldenberg L, Salcudean S. Deep learning-based gleason grading of prostate cancer from histopathology images—role of multiscale decision aggregation and data augmentation. IEEE J Biomed Health Inform. 2019;24(5):1413–26. doi:10.1109/jbhi.2019.2944643.
23. Koziarski M, Cyganek B, Niedziela P, Olborski B, Antosz Z, żydak M, et al. DiagSet: a dataset for prostate cancer histopathological image classification. Scient Rep. 2024;14:6780. doi:10.1038/s41598-024-52183-4.
24. Bulten W, Kartasalo K, Chen P, Ström P, Pinckaers H, Nagpal K, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nature Med. 2022;28:154–63. doi:10.1038/s41591-021-01620-2.
25. Zuley ML, Jarosz R, Drake BF, Rancilio D, Klim A, Rieger-Christ K, et al. The cancer genome atlas prostate adenocarcinoma collection (TCGA-PRAD) (Version 4) [Data set]. The Cancer Imaging Archive; 2016. doi:10.7937/K9/TCIA.2016.YXOGLM4Y.
26. Hossain M, Shahriar G, Syeed M, Uddin M, Hasan M, Hossain M, et al. Tissue artifact segmentation and severity assessment for automatic analysis using wsi. IEEE Access. 2023;11:21977–91. doi:10.1109/ACCESS.2023.3250556.