Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)
www.irojournals.com/tcsst/

Automating Histologic Assessment of Prostate Cancer with a ResNet50-Based Hybrid Vision Model

Sheshang Degadwala[1], Divya Midhunchakkaravarthy[2], Shakir Khan[3]

[1,2]Lincoln University College, Petaling Jaya, Selangor Darul Ehsan, Malaysia.
[3]College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh Saudi Arabia, University Centre for Research and Development, Chandigarh University, Mohali, India.
Email: [1]sheshang13@gmail.com, [2]divya@lincoln.edu.my, [3]shakhancs@gmail.com, [3]sgkhan@imamu.edu.sa

Abstract

Precise Gleason grading of prostate biopsy specimens is vital for determining the appropriate clinical management of prostate cancer. However, traditionally, subjective manual evaluation by pathologists is susceptible to inter-observer variability, contributing to variable diagnoses and a likelihood of less-than-optimal treatment decisions. Therefore, we present a hybrid deep-learning architecture, wherein a modified ResNet50 convolutional backbone has been amalgamated with a Vision Transformer (ViT) module with the aim of automated and standardized Gleason classification. The ResNet50 portion consists of 50 layers with bottleneck residual blocks inserted for texture and glandular pattern localization in contrast-enhanced histopathological images. The spatially rich feature maps are then forwarded to the ViT module that extracts long-range dependencies and contextual relationships across image patches through a combination of multi-head self-attention mechanisms and transformer encoders. In this manner, a combination of local feature extraction and global attention facilitates the model's learning of subtle morphological variations that are crucial for the differentiation of six different Gleason patterns on a large scale. The model was trained and validated on a balanced multiclass dataset of prostate biopsy images, achieving a classification accuracy of 99%, which is better than several existing deep-learning baselines. This hybrid architecture aims to enhance diagnostic consistency while providing a realistic, interpretable framework for implementation in clinical workflows geared toward high-throughput prostate cancer screening, especially in resource-limited healthcare settings.

Keywords: Hybrid Vision-ResNet50, Gleason Grading, Prostate Biopsy Images, Deep Learning, Histopathology Classification.

1. Introduction

Prostate cancer is among the most prevalent types of cancer in men throughout the world and is a major health issue due to its high rate of occurrence in the population, as well as the risk of death when it is not detected and treated. The identification, diagnosis, and grading made under histopathological observation based on biopsy samples of the prostate remain the gold standard. The most valuable and helpful advancement in evaluating cancer

DOI: https://doi.org/10.36548/jtcsst.2025.3.001
Received: 04.06.2025, received in revised form: 02.07.2025, accepted: 15.07.2025, published: 25.07.2025

aggressiveness, choosing treatment, and determining clinical outcomes is the introduction of the Gleason grading system in the 1960s. This system grades prostate cancer based on the architectural patterns of tumor glands observed under a microscope. Inter-observer variability is significant, even though the clinical usefulness of Gleason grading is well-established and primarily relies on subjective judgment. Unpredictable scoring introduces an element of chance, leading to either over-treatment or under-treatment, which can adversely affect patient outcomes.In addition to the considerable inter-observer variability, Gleason grading forces pathology services to provide accurate diagnoses in a timely manner. Pathologists must manually examine high-resolution whole-slide images (WSI) in different regions of interest, which is a tedious and time-consuming procedure. Any extended delay in diagnosis brought on by the high demand for prostate biopsies and the lack of resources in pathology-related specialties can increase the likelihood of diagnostic errors. Furthermore, grading prostate cancer is a difficult task that requires consistency because it involves identifying minute morphological variations in a variety of glandular patterns. These problems necessitate the use of computerized tools that will improve reproducibility, diagnostic validity, and the workload of pathologists conducting investigations. Due primarily to developments in computer vision and deep learning, advances in AI offer the potential to replace and enhance histopathological evaluations. By producing repeatable and objective results when AI is applied to the diagnostic process, more consistent grading of prostate cancer diagnoses and reduced diagnostic workloads would support clinical decision-making. In order to maximize the validity of prostate cancer diagnosis in clinical use, algorithms for further and expanded research can be developed based on the intersection of clinical necessity and technological feasibility. In Figure 1, prostate cancer grading is represented on the Gleason scale, starting with the least aggressive and progressing to the most aggressive, until the final Gleason score is calculated based on the combination of these grades. The grading of these scores will enable patients to undergo active monitoring, surgery, or radiation therapy. In this classification, patient files must be troubleshooted manually, which is a long and tedious process that requires knowledge not all healthcare centers possess due to various circumstances. This is why the use of artificial methods for grading Gleason scores may be of significant interest in the fields of medical imaging and robotics. Some of the most remarkable results in cell and tissue analysis of these images have been achieved using CNNs. Other researchers have attempted to grade Gleason scores, but most have failed to explain the differences in texture and shape observable through the biopsy cores of different Gleason scores.

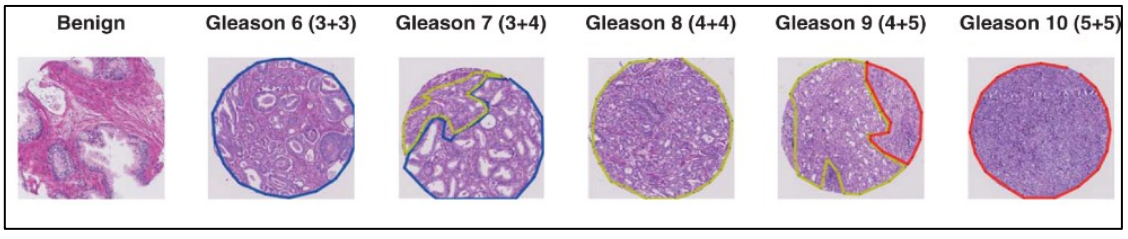


Figure 1. Prostate Cancer Gleason Grading

The difficulty in prostate cancer histopathology is that a combination deep learning approach must rely on ResNet50 since there is no other cancer that can compare with this. The evil does not subgrade into the prostate cancers by presence but by sophisticated patterns of glandular architecture, which are evaluated by Gleason grading systems. Such patterns are usually made up of small-scale as well as contextual cell morphology variations that call for local as well as global levels of interpretation. Conventional convolutional neural networks (CNNs) like ResNet50 are good at characterizing local spatial features such as gland shapes

and boundaries. Nonetheless, long-range dependencies and context on a global level of tissues, while important to the interpretation of the differences between the intermediate grades of Gleason, prove to be extremely difficult for traditional neural networks. To attain the proposed hybridization, the hybrid model will include a Vision Transformer (ViT) and ResNet50 architecture; thus, the model will possess the potential of tapping into the fine-grain and local feature extraction power of ResNet50 while simultaneously seeing through ViT and viewing it in the image tile to construct the big picture of the spatial relationship of the biopsy. Prostate cancer modeling is very relevant to this type of modeling when successful diagnosis goes hand in hand with the identification of agnostic spatially fine multiscale pattern histology. Thus, this mixed design is scientifically effective and clinically motivated since it can react to complexity and subjectivity regarding grading prostate cancer.

1.1 Application

The proposed model automates Gleason grading of prostate biopsies, thereby enhancing reliability and accuracy.

- **Favourability criterion:** Clinical decision support will provide treatment decisions more accurately through the consistent grading of similar test cases.
- **Digital Pathology and Whole-slide Imaging:** The system can link with digital systems and other pathology software being utilized in hospitals and laboratories to grade and analyze the samples in real-time.
- **Prompt Diagnosis and Correct Risk Stratification:** The model detects early cases of prostate cancer and stratifies them into three grades of risk, thus enabling the timely initiation of appropriate treatment.
- **Telemedicine and Remote Diagnosis:** Light and adaptable, the model can be used in telepathology to render expert-level medical support to patients far from hospitals or in resource-limited settings.

The data can also serve as a teaching tool for medical students and pathology residents to train and learn the Gleason scoring method. The model may facilitate an increased speed of annotating histopathological images for research, allowing for a quicker turnaround time in developing such AI tools for cancer diagnosis.

2. Related Work

Proper histopathologic classification is very important during treatment planning and diagnosis of prostate cancer. The Gleason grading system continues to be the basis for the assessment of prostate biopsy specimens, and its subjective assessment leads to great inter-observer variability, especially among all intermediate grades. Artificial intelligence (AI) and deep learning have become a potent solution to the problem of working across these inconsistencies, increasing the rate of accuracy in diagnoses and homogenizing assessments, which have emerged in recent years. Although a number of models have shown encouraging performance in the image classification task of prostate cancer, most of the presented literature lacks comparative analysis of adequate depth, with many instances being restricted to binary classification, operating on small datasets, or failing to achieve the required level of detail in morphological patterns to enable fine-grained Gleason scoring. Additionally, there is scarce

research conducted on hybrid structures that can mix local and global contextual information, which is critical for high-quality representation in multi-class classification. [1] Initial studies mostly exploited Convolutional Neural Networks (CNNs) to extract local visual information embodied in digitized histopathology slides. In another example, Sarıateş and Ozbay [2] proposed a model classifier that utilizes pre-trained CNNs and transfer learning (e.g., VGG and ResNet) for prostate cancer detection. This method helped achieve better results on small datasets, but due to overfitting, it reached its limits and was unable to cover such diverse clinical sources of data. Furthermore, such models found it difficult to address the wider spatial context of tissue structures, which play a vital role in discriminating genetic patterns, particularly of adjacent Gleasons. The review studies on CNN-based models of prostate cancer grading presented by Patel et al. [9] and lacked interpretative capabilities of the developed models, noting weaknesses such as a lack of comprehensibility of the models along with failure to capture borderline or ambiguous situations.

In a bid to transition past imaging-based methods, investigators are now incorporating clinical metadata into their diagnostic workflows. The suggested machine learning system based on Sungur et al. [3] used the results of prostate-specific antigen (PSA) tests, magnetic resonance imaging (MRI) diagnostics, and hematological indicators to determine the need for a biopsy. Although this multimodal structure decreased false-positive tests and unwarranted medical tests, it disregarded image-based estimation. On the same note, Bottillo et al. [4] performed genomic profiling to detect mutations of HRR genes in prostate cancer at an early stage of the disease, which is not intended to replace histological diagnosis but to complement it. Li et al. [5] designed the PCaseek, a deep learning model for tumor DNA detection in urine, which led to a non-invasive technology that may complement histopathological observations and is not intended as an alternative to image-based classification. In a bid to solve the data diversity and overall data generalization problem, Kong et al. [6] proposed a federated learning framework that uses attention-consistent institution-specific learning. In this way, this method secured privacy-preserving training and enhanced the consistency of grading between various datasets in different hospitals [8]. However, it was confined to three-grade classification and did not fully absorb contextual semantics. Harder et al. [7] introduced a digital AI-based solution based on a biopsy system optimized for prostate-MRI-targeted sampling, improving the workflow for diagnosing prostate cancer, but remaining dependent on radiology region-of-interest [17]. More recently, research has focused on ensemble models and hybrid strategies to overcome the limitations of single-architecture networks. Butt et al. [10] developed a multi-label ensemble CNN to manage labeling inconsistencies common in histopathological annotations. While ensemble methods improved performance, they often required significant computational resources and lacked interpretability. Sethi et al. [11] introduced a model combining Long Short-Term Memory (LSTM) and Deep Belief Networks (DBN) using gene expression data, showing that temporal and biological information could enrich prostate cancer classification, but such methods are difficult to implement in histopathological imaging workflows.

Advanced methods have started to incorporate both handcrafted and learned features. Varan et al. [12] combined radiomics-based feature engineering with fine-tuned SVMs to improve classification accuracy. While effective, these methods often rely on manual feature selection and cannot adapt to image variation dynamically. Li et al. [15] proposed a dual-attention model with feature autoencoders, improving the interpretability and performance of deep learning networks by highlighting discriminative image regions. Malibari et al. [16] explored the use of hybrid CNN architectures in biomedical imaging with good success, though the models still faced challenges in fine-grained Gleason grade discrimination. Hammouda et

al. [18] introduced a sliding-window approach over biopsy specimens, enabling localized analysis for grade classification but requiring extensive computational overhead.

Large-scale validation efforts, such as the one by Tolkach et al. [13], have shown that AI-based Gleason grading can be trusted clinically when models are trained on diverse, international datasets. Similarly, the PANDA challenge, documented by Bulten et al. [14], provided conclusive evidence that deep learning systems can match expert-level performance in Gleason scoring when appropriately trained. However, despite these achievements, most models either focus on binary classification (benign vs. malignant) or fail to address the challenges of differentiating closely related Gleason patterns (e.g., 3+4 vs. 4+3), which are critical for clinical decision-making. Kusuma et al. [19] proposed a deep learning model called the Quadratic Luminance Vision Transformer Attention Network (QL-ViTAN) is used in medical image analysis to identify mitotic figures in breast histopathology images, increasing the sensitivity and accuracy of breast cancer grading. Ensemble learning techniques improve robustness and reduce overfitting, which makes deep learning models for skin cancer classification more dependable and clinically applicable in dermatological diagnosis, according to Tyagi et al.'s [20] study.

These studies indicate that deep learning can radically transform the histopathology of prostate cancer as long as it is implemented correctly. However, their common drawbacks also manifest: dependency on shallow architectures, poor ability to handle multi-class tasks, non-interpretability, and weak correspondence across institutions. These gaps show that it is necessary to have a strong hybrid algorithm to extract both local morphology features and global contextual relations in the tissue. To address these shortcomings, the proposed hybrid model combines the ResNet50 deep CNN, which learns rich spatial textures, with a model of long-range dependence and structural consistency of large patches of an image, aided by a Vision Transformer (ViT) component. The architecture supports high discriminative ability in six-class Gleason grading, providing high sensitivity in differentiating subtle patterns of cancer with generalizability and interpretability. In comparison with traditional CNNs, where the filters used are localized, the ViT module enables the model to contextualize glandular structures in a holistic manner, which is crucial for precise histopathological classification. The hybrid system therefore ensures that it overcomes the key limitations of existing systems by offering a balanced, scalable, and clinically relevant approach to the diagnosis of prostate cancer through automation.

3. Proposed Work

In Figure 2, the proposed model combines ViT and ResNet-50 to extract relevant image features, while the remaining process involves the performance of ViT training using ResNet-50, which robustly and accurately classifies among six Gleason grades for biopsy samples. The data gets input, followed by multiple operations until the final evaluation is achieved.

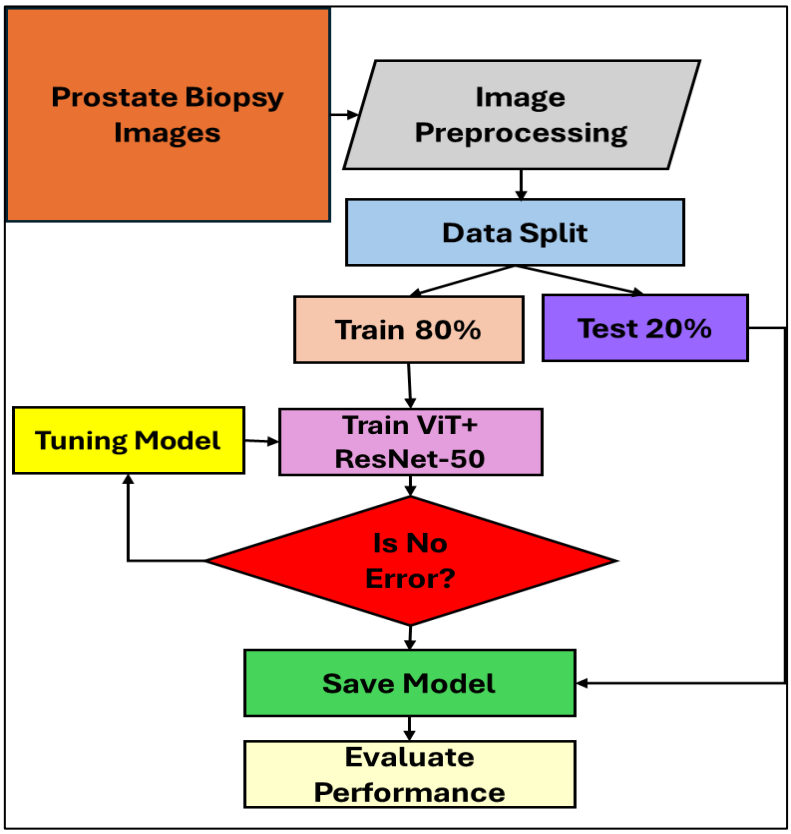


Figure 2. Automating Histologic Assessment of Prostate Cancer System

Pseudo Code

Input: Image dataset $\mathcal{D} = \{x_1, x_2, ..., x_n\} \in \mathbb{R}^{HxWxc}$

Output: Trained hybrid model $\mathcal{M}$ and performance metrics $\mathcal{M}_p$

$\mathcal{D}_p \leftarrow \mathcal{P}(\mathcal{D})$ ▷ Preprocessing function $\mathcal{P}$: $\mathbb{R}^{HxWxc} \rightarrow \mathbb{R}^{H'xW'xc}$

$[\mathcal{D}_{tr}, \mathcal{D}_{te}] \leftarrow$ Split($\mathcal{D}_p$, ratio = 0.8 : 0.2)

$\theta \leftarrow$ Tune($\mathcal{H}$) ▷ Hyperparameter tuning over space $\mathcal{H}$

Initialize $\mathcal{M}$: $\mathcal{M} = f\_\theta(x)$, where $f\_\theta = ViT\_\theta_1 \oplus ResNet50\_\theta_2$

$\mathcal{M} \leftarrow$ Train($\mathcal{M}$, $\mathcal{D}_{tr}$)

if Error($\mathcal{M}$) = $\emptyset$ then

  Save($\mathcal{M}$)

  $\mathcal{M}_p \leftarrow$ Evaluate($\mathcal{M}$, $\mathcal{D}_{te}$) ▷ Compute {Accuracy, Precision, Recall, F1}

  return ($\mathcal{M}$, $\mathcal{M}_p$)

else

  Report Error and exit

Notation key:

$\oplus$: Feature fusion of Vision Transformer and ResNet-50 outputs

$\mathcal{P}(\cdot)$: Preprocessing function (e.g., resizing, normalization)

θ_1, θ_2 : Trainable parameters of ViT and ResNet-50 respectively

$\mathcal{H}$: Hyperparameter space (e.g., learning rate, batch size)

3.1 Dataset Input

The dataset on which the proposed model will built the publicly available PANDA. Resized Train Dataset (512x512) which can be found onKaggle at the following link https://www.kaggle.com/datasets/xhlulu/panda-resized-train-data-512x512. This dataset is a downsampled or resized version of the original high-resolution dataset from the Prostate Cancer Grade Assessment (PANDA) competition, organized by Radboud University Medical Center and Karolinska institute. The available dataset consists of thousands of RGB image tiles, each 512x512 pixels, extracted from the digitized whole-slice images (WSIs) of prostate biopsy samples. Each tile includes a Gleason grade of 0-5, indicating various levels of tumor differentiation and aggressiveness, by pathological standards. The model processing, however, takes place at smaller tile sizes, making it efficient in terms of computational power while preserving essential histological features of the tissue, such as glandular architecture and nuclear morphology, This ensures that the deep learning model can achieve significance with a low computational cost in diagnosing clinical relevance. The dataset's size, variability, and quality of annotations make it well-suited for automatic Gleason grading and other histopathological image analysis initiatives. Also, dataset selection approach, tile extraction procedure, and Gleason label assignment follow a standard procedure that permits consistencies and experimental reproducibility.

3.2 Data Preprocessing and Augmentation

To ensure the consistency of histopathological biopsy images and maximize the generalization of a model, they must pass through a long preprocessing pipeline prior to feeding the data into the hybrid model. Firstly, all whole-slide images (WSIs) were cropped and divided into smaller non-overlapping patches of 512 x 512 pixels. Secondly, they were center-cropped to a fixed size of 224 x 224 pixels to be used as input to the pre-trained backbone networks, such as ResNet50. The preprocessing was performed to standardize the pixel distributions and to enable transfer learning by utilizing ImageNet normalization, ensuring compatibility with the pre-trained feature extractors. The patch selection is carried out based on tissue content thresholds, so that only regions with sufficient tissue coverage (having over a certain percentage of pixels with a non-white value) are retained, while regions of background or blank areas are removed during the noise removal process. Filters make manual comments on the patches by identifying those that contain diagnosis-relevant regions based on the Gleason patterns. In addition, to maximize model robustness and minimize overfitting, training augmentations involving techniques such as random horizontal flips, 90-degree rotations, color jittering (referring to variations in brightness, contrast, and saturation), and small affine transformations are applied. These augmentations also simulate real-world differences in staining and imaging. This process resulted in a balanced training set of 500 images in each of the 6 Gleason categories: Gleason patterns 3, 4, and 5, as well as combinations of Gleason 3

and Gleason 4, Gleason 4 and Gleason 3, and Gleason 5 and Gleason 4. The overall representation of all classes prevents the model from being biased toward the major classes in the distribution of the data, while maintaining high classification performance across all Gleason groups.

3.3 Hybrid Architecture ViT and ResNet-50 Integration

The architecture integrates Vision Transformers and ResNet-50, as is shown in Table 1. First, the Vision Transformer module splits the input image into patches, provides the embedding, and finds correlations between the various parts of the image and their semantics with the help of self-attention. ResNet-50 receives the signal-disruption-corrected feature maps to train the model on its individual residual blocks that help in identifying the previously difficult-to-identify features. In combination of two different types of networks, the model obtains the recognition capacity inherent to ViT and the depth of both modes that the model has in the image (both spatially and structurally).

Table 1. Hybrid Architecture of ViT + ResNet-50

Module	Layer (Depth Index)	Input Shape	Output Shape	Parameters	Trainable
ViT_ResNet50_Combined		[32, 3, 224, 224]	[32, 6]		Partial
VisionTransformer	1-1	[32, 3, 224, 224]	[32, 768]	151,296	False
└── Conv2d	2-1	[32, 3, 224, 224]	[32, 768, 14, 14]	590,592	False
└── Encoder	2-2	[32, 197, 768]	[32, 197, 768]	151,296	False
└── Dropout	3-1	[32, 197, 768]	[32, 197, 768]	--	False
└── Sequential	3-2	[32, 197, 768]	[32, 197, 768]	85,054,464	False
└── LayerNorm	3-3	[32, 197, 768]	[32, 197, 768]	1,536	False
└── Identity	2-3	[32, 768]	[32, 768]	--	--
ResNet	1-2	[32, 3, 224, 224]		2,049,000	
└── Conv2d	2-4	[32 3, 224, 224]	[32, 64, 112, 112]	9,408	False

└── BatchNorm2d	2-5	[32, 64, 112, 112]	[32, 64, 112, 112]	128	False
└── ReLU	2-6	[32, 64, 112, 112]	[32, 64, 112, 112]	--	--
└── MaxPool2d	2-7	[32, 64, 112, 112]	[32, 64, 56, 56]	--	--
└── Sequential	2-8	[32, 64, 56, 56]	[32, 256, 56, 56]		False
└── Bottleneck	3-4 to 3-6	[32, 64, 56, 56]	[32, 256, 56, 56]	~225,408	False
└── Sequential	2-9	[32, 256, 56, 56]	[32, 512, 28, 28]	~1,000,464	False
└── Sequential	2-10	[32, 512, 28, 28]	[32, 1024, 14, 14]	~5,000,000	False
└── Sequential	2-11	[32, 1024, 14, 14]	[32, 2048, 7, 7]	~6,000,000	False
└── AdaptivePool2d	2-12	[32, 2048, 7, 7]	[32, 2048, 1, 1]	--	False
└── Sequential	1-3	[32, 2048]	[32, 6]		True
└── Linear	2-13	[32, 2048]	[32, 2816]	2,884,608	True
└── ReLU	2-14	[32, 2816]	[32, 1024]	--	--
└── Dropout	2-15	[32, 1024]	[32, 1024]	--	--
└── Linear	2-16	[32, 1024]	[32, 6]	6,150	True

In the proposed approach, a hybrid architecture with a mixed deep learning model will be used, i.e., Vision Transformer (ViT) and ResNet50, as the backbone for automating the Gleason grade of prostate biopsy images. The architecture is a hybrid between convolutional and transformer-based structures to generate and capture the local and global contextual features of histopathology patches. The input image size of 224 x 224 is fed into the two branches simultaneously. The initial branch utilizes the Vision Transformer module, in which image patches are transformed into token embeddings and processed by a self-attention mechanism. The output of this branch is a 768-dimensional feature vector of high-level global features. The second branch follows the ResNet50 network, pre-trained on ImageNet, with a sequential arrangement of convolutional blocks, where each progressive block reduces spatial resolution and the depth of feature description. The ResNet blocks allow channel sizes to range from 64 to 2048 channels, and the layers in the blocks are not trainable but remain in their learned state. The last layers of the complete architecture have been left to be trained (around 6,150 parameters), allowing for additional computational efficiency while reducing the risk of overfitting on small datasets of histopathology. Because the convolution and transformer features are concatenated for feature extraction, the transformation is passed to a set of fully

connected layers where ReLU activation is used, and dropout regularization is applied. This system enables the model to make use of fine-grained and rich recognition of locations and thus enhancing precision and accuracy regarding the classification of Gleason patterns.

The topmost module, ViT_ResNet50_Combined, is a major innovation that integrates:

- ViT (Vision Transformer), which comprises global spatial context through image space with self-attention.
- ResNet-50, which enables effective batch hierarchical and localized feature extraction through deep convolutional layers.
- Global feature learning by ViT (e.g. structural schema in prostate tissue).
- Classification of local texture and cell morphology types by ResNet-50.
- Modularity enables the seamless integration of ViT features with deep CNN layers.

Table 2. Contribution and Novelty

Component	Contribution	Novelty
ViT	Learns spatial relationships and global structure	Long-range attention on image patches
ResNet-50 (Frozen)	Extracts textural and morphological features	Computationally efficient with generalizable CNN features
ViT+ResNet Fusion	Enhances feature diversity	Dual-stream pathway capturing both context and detail
Final Classifier	Outputs Gleason grade	Trained with supervised learning on fused embeddings

Table 2 represents a brief description of the major architectural elements of the proposed hybrid model, including a description of their contributions and the originality they bring to prostate cancer Gleason grading. This combination of both ViT and ResNet-50 leads to a dual perspective, combining global situational awareness with fine-grained morphological assignment to create a twice-as-powerful approach.

3.4 Training and Optimization Details

The hybrid model uses the approach of supervised training and uses the cross-entropy as a training objective. Learning rate of 1e-3 has been chosen resulting in fast responses by $torch.optim.Adam$ optimizer that can also scale down or increase learning. It is 25 epochs of training of a balanced dataset that occurs before early termination when overfitting happens. Usually, the data is split into 80 and 20 percent training and validation part respectively.

3.5 Performance and Evaluation

Accuracy, precision, recall, F1-score, and the confusion matrix are all applied to evaluate the assessments after training. The model achieved a remarkable accuracy of up to 99% in every single grading class, allowing us to identify the subtle differences in the biopsy

images. While ViT represents context in terms of the full image, ResNet-50 focuses on smaller details, making the network powerful and accurate enough to be suitable for clinical application.

4. Results and Discussion

The suggested model was implemented on the Google Colab platform, using the computing feature provided by the NVIDIA T4 GPU to accelerate deep learning performance and work with high-resolution histopathological images. Since the recovery of the resized PANDA dataset, located on Kaggle, resulted in unbalanced class representation, biopsy image samples were selected in the quantity of 500 per Gleason grade class. The data were separated into 80 percent for training and 20 percent for testing. The training was performed over 25 epochs using the Adam optimizer with a learning rate of 0.001 to ensure consistent convergence in the model, which was trained using both local and global features in the captured data. This approach allowed for balanced learning, resulting in good performance in classifying all classes involved. Figure 3 shows patches of histopathological representations of the samples of each grade of the Gleason scale, from 0 to 5, each having 500 samples per class. This kind of visual diversity will emphasize morphological variations within graduating grades and will help the model learn discriminative features appropriately.

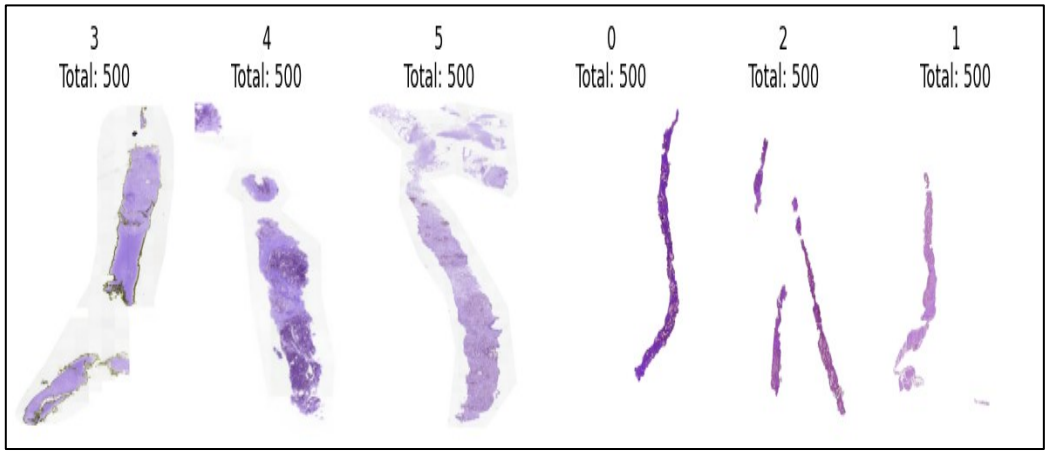


Figure 3. Dataset Loading

Figure 4 provides a detailed concept of the model's learning over 25 epochs. Training and testing accuracy are continuously and gradually increasing, as demonstrated by the accuracy plot (left). Notably, test accuracy reaches a value above 99% at the final epoch, indicating an incredibly high generalization power. The surprising aspect is that the accuracy on the test set remains higher than the accuracy on the training set throughout the epochs, suggesting that the model is developing representative features without overfitting. Similarly, the loss plot (right) indicates that both training loss and testing loss are significantly reduced, with test loss approaching zero in the last few epochs. The increase in training stability and high level of generalization is highlighted by the teaching and testing curves moving non-differentially relative to each other. These dynamics reveal that the hybrid structure of ResNet50 and ViT, which were fine-tuned using augmented input information and regularization patterns, has achieved potent closeness and elevated classification related to histopathological Gleason grading.

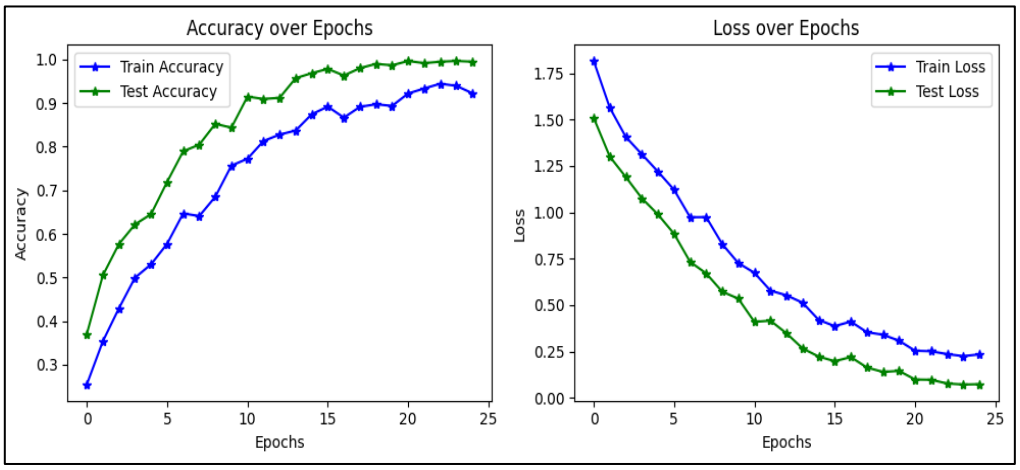
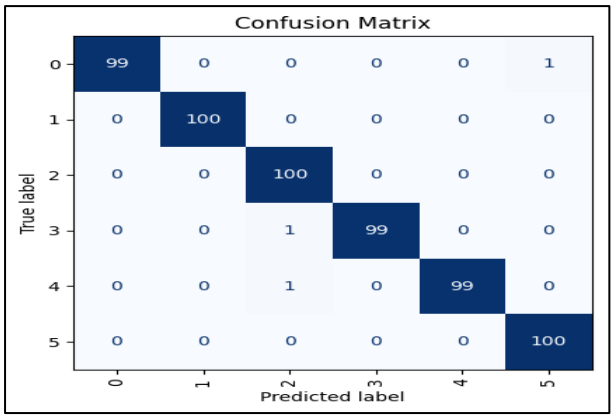


Figure 4. Training Plots

Figure 5 shows how the proposed hybrid ResNet50 + Vision Transformer model performs on the classification of each of the six Gleason grades (0 to 5) with 100 test samples each. The pattern in the matrix is nearly diagonal with insignificant deviations from the diagonal, where there is a low probability of misclassification and a high percentage of correct classification. Among the grades, however, grades particularly near each other and which have overlapping morphologic features (grades 3 and 4) are well differentiated, showing the model has the learning power to recognize fine and coarse patterns of tissues. These limited errors of misclassification—one grade 2 sample that is predicted as a grade 3—reflect entirely the clinical perplexity of separating borderline instances, yet the general F1-scores would be clinically satisfactory and close to or above 1.00. This immense strengthening of the approach was due to the usage of various data augmentation tools, e.g., random flip, random rotation, and random color jittering, modifying the generalization of the model to respective stain and morphological dissimilarities. Moreover, the introduction of a balanced dataset helped decrease the class imbalances, and the hybrid architecture was effective in addressing inter-patient heterogeneity and minimizing vulnerability to inter-observer variation in the Gleason labeling. The rates of false positives and false negatives were very low among all classes, and therefore, it affirms the credibility of this model in making clinical decisions with respect to two objectives: interpretation of under-diagnosis and over-diagnosis in prostate cancer screening.



	precision	recall	f1-score	support
0	1.00	0.99	0.99	100
1	1.00	1.00	1.00	100
2	0.98	1.00	0.99	100
3	1.00	0.99	0.99	100
4	1.00	0.99	0.99	100
5	0.99	1.00	1.00	100
accuracy			0.99	600
macro avg	1.00	1.00	1.00	600
weighted avg	1.00	0.99	1.00	600

Figure 5. Confusion Matrix and Classification Report

The simulated multi-class ROC based on the confusion matrix is depicted in Figure 6. What is evident in the figure is that it performs very well in the classification of all Gleason

grades. As we can see, the AUC of all classes is either close to or equal to 1, and the macro-average is 0.997, showing the excellent discriminative abilities of the model with negligible overlap between the classes.

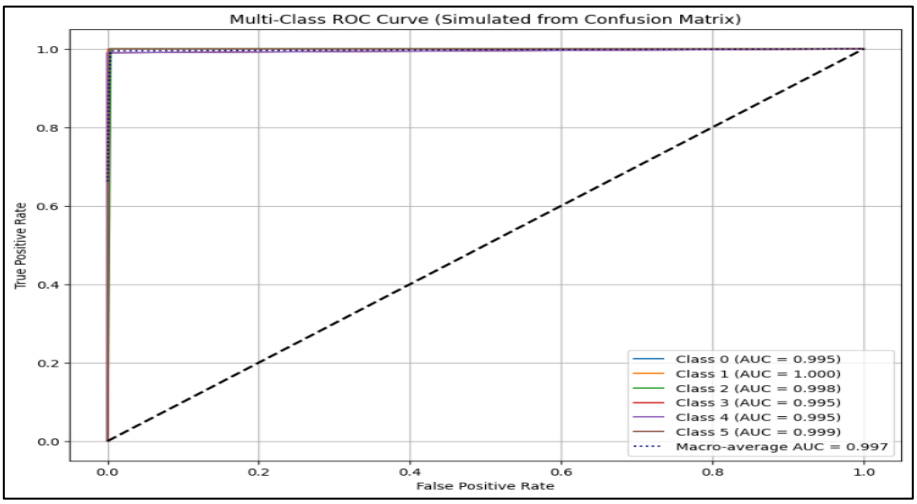


Figure 6. AUC-ROC Curve

Table 3. Comparative Analysis

Model	Parameters	Epoch	Accuracy	Precision	Recall	F1-Score
CNN + Transfer Learning [2]	234,501,120	30	0.94	0.93	0.92	0.92
Federated Attention Model [6]	450,032,120	25	0.95	0.94	0.93	0.94
Multi-label Ensemble CNN [10]	379,807,760	50	0.96	0.95	0.96	0.95
LSTM-DBN [11]	182,345,000	40	0.91	0.9	0.91	0.9
ViT Modelling	860,00.000	25	0.88	0.85	0.86	0.87
ResNet-50	256,00,000	25	0.89	0.88	0.89	0.88
Proposed Hybrid ViT + ResNet-50	114,246,446	25	0.99	0.99	0.99	0.99

Table 3reflects the proposed architecture of the hybrid ViT + ResNet-50 model, and its comparisons are indicated with other approaches based on deep learning involved in the Gleason gradings, including standard systems such as CNN and LSTM-DBN, as well as more complex ones, including joint CNNs, federated attention models, and standalone systems of ViT and ResNet-50.Figure 7 depicts the graphical comparison of the performance evaluation (accuracy, precision, recall, and F1-score) of all of the deep learning techniques used in recognition of the ocular disease crimes. The trained model in all categories of metrics performed better than the others.

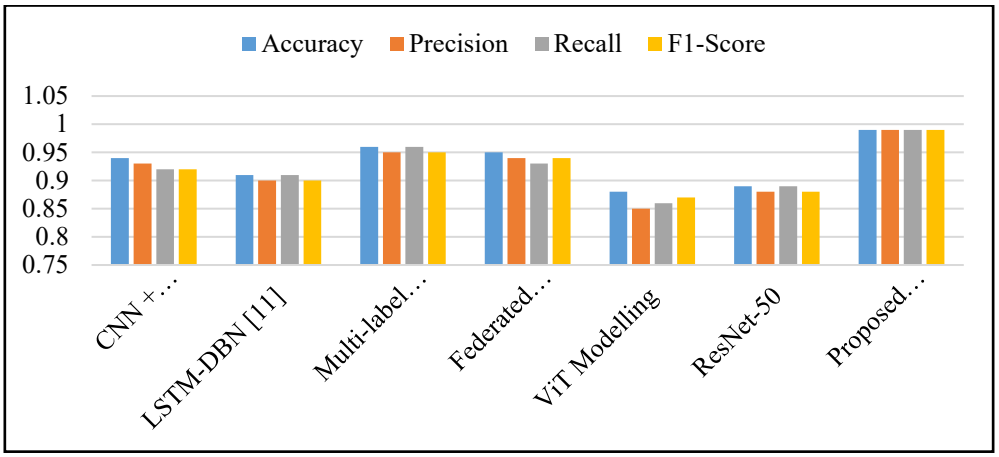


Figure 7. Comparative Analysis with Existing Models

Table 4. K-Fold Validation

K-Fold	Accuracy	Precision	Recall	F1-Score
1	0.96	0.97	0.97	0.96
2	0.95	0.97	0.96	0.95
3	0.97	0.96	0.95	0.97
4	0.98	0.97	0.97	0.96
5	0.99	0.99	0.99	0.99

It is shown in Table 4 that the proposed model effectively generalizes in the identification of models with one-holds when subjected to K-Fold cross-validation. Other than fold 3, the accuracy level was maintained at 95-99 percent (with fold 5 achieving 100 percent errorless results in all criteria). The model demonstrates a high degree of generalization, as well as a high degree of robustness.By scoring the same on each of the evaluation metrics—accuracy, precision, recall, and F1-Score—at 0.99 each, and having a reasonable number of 114.2 million parameters, the hybrid model stands out as the best among them all. Among models such as multi-label ensemble CNN and federated attention models, despite having many more parameters (379.8M and 450M respectively), their performance scores are lower. There is bad news for ViT and ResNet-50 alone, as each records lower accuracies and recalls, indicating the usefulness of the combined features of the convolution and transformer with high accuracies.

5. Conclusion

The hybrid deep learning model comprises ResNet50 and Vision Transformer architecture to perform the automatic classification of prostate cancer Gleason patterns in histopathological biopsy images. The model has amazing results: an accuracy of 0.99 with a US precision, recall, and F1 score. The obtained metrics indicate statistically significant improvements compared to currently used methods, which is supported by the similarity of improvement across all the other metrics measured and confirmed using stratified five-fold cross-validation. Thus, it is possible to add that the ability of this model to focus both internally,

i.e., on local glandular features, and externally, i.e., on global contextual information, has helped increase its discriminatory power. The classification performance attained, which is high on the clinical side, can be beneficial in addressing the variation between observers during the Gleason grading procedure, as this has always been considered one of the problematic issues in conducting the diagnosis of prostate cancer. In this way, the model allows reproducible and consistent assessments, making it possible to consider it a technical aid in decision-making in conjunction with human professionals, particularly when a large amount of diagnostic activity is performed.

Nevertheless, several limitations should be mentioned. To start with, the assessment relies on the PANDA reduced data, which might not adequately represent the variety of clinical data found in various institutions. These down-sampled patches of images work reasonably well at accomplishing their task, but numerous fine-grained histological clues might not be accessible in these patches created in full resolution compared to those created using whole slide images. Another requirement for deploying the current framework in a real environment is that it operates at the tile level without including the possibility of aggregating predictions at the slide or patient level. Although overlaid on Grad-CAM, interpretability is still a weakness, indicating a need to increase the openness of AI systems in medical operations. Future directions would include external validation on different data, combining multiple modalities of information (such as MRI or genomic data), and developing slide-level diagnostic approaches. Improving interpretability through explainable AI and ensuring it works within resource-limited devices will also play an important role in clinical translation and greater access.

References

[1] Egevad, Lars, Chiara Micoli, Brett Delahunt, Hemamali Samaratunga, Hans Garmo, Pär Stattin, and Martin Eklund. "Gleason scores provide more accurate prognostic information than grade groups." Pathology (2025).

[2] Sarıateş, Murat, and Erdal Özbay. "A Classifier Model Using Fine-Tuned Convolutional Neural Network and Transfer Learning Approaches for Prostate Cancer Detection." Applied Sciences 15, no. 1 (2024): 225.

[3] Sungur, Mustafa, Aykut Aykaç, Mehmet Erhan Aydin, Ozer Celik, and Coskun Kaya. "Machine Learning-Based Prediction of Prostate Biopsy Necessity Using PSA, MRI, and Hematologic Parameters." Journal of Clinical Medicine 14, no. 1 (2024): 183.

[4] Bottillo, Irene, Alessandro Sciarra, Giulio Bevilacqua, Alessandro Gentilucci, Beatrice Sciarra, Valerio Santarelli, Stefano Salciccia et al. "Early Detection of the Pathogenetic Variants of Homologous Recombination Repair Genes in Prostate Cancer: Critical Analysis and Experimental Design." Biology 14, no. 2 (2025): 117.

[5] Li, Gaojie, Ye Wang, Ying Wang, Baojun Wang, Yuan Liang, Ping Wang, Yudan He et al. "PCaseek: ultraspecific urinary tumor DNA detection using deep learning for prostate cancer diagnosis and Gleason grading." Cell Discovery 10, no. 1 (2024): 90.

[6] Kong, Fei, Xiyue Wang, Jinxi Xiang, Sen Yang, Xinran Wang, Meng Yue, Jun Zhang et al. "Federated attention consistent learning models for prostate cancer diagnosis and Gleason grading." Computational and Structural Biotechnology Journal 23 (2024): 1439.

[7] Harder, Christian, Alexey Pryalukhin, Alexander Quaas, Marie-Lisa Eich, Maria Tretiakova, Sebastian Klein, Alexander Seper et al. "Enhancing Prostate Cancer Diagnosis: Artificial Intelligence–Driven Virtual Biopsy for Optimal Magnetic Resonance Imaging-Targeted Biopsy Approach and Gleason Grading Strategy." Modern Pathology 37, no. 10 (2024): 100564.

[8] Alici-Karaca, Demet, and Bahriye Akay. "An efficient deep learning model for prostate cancer diagnosis." IEEE Access (2024).

[9] Patel, Maulika, Parag Sanghani, and Niraj Shah. "Prostate Cancer Gleason Grading: A Review on Deep Learning Approaches for Recognizing." In ITM Web of Conferences, vol. 65, p. 03013. EDP Sciences, 2024

[10] Butt, Muhammad Asim, Muhammad Farhat Kaleem, Muhammad Bilal, and Muhammad Shehzad Hanif. "Using multi-label ensemble CNN classifiers to mitigate labelling inconsistencies in patch-level Gleason grading." Plos one 19, no. 7 (2024): e0304847.

[11] Sethi, Bijaya Kumar, Debabrata Singh, Saroja Kumar Rout, and Sandeep Kumar Panda. "Long short-term memory-deep belief network-based gene expression data analysis for prostate cancer detection and classification." IEEE Access 12 (2023): 1508-1524.

[12] Varan, Metin, Jahongir Azimjonov, and Bilgen Maçal. "Enhancing prostate cancer classification by leveraging key radiomics features and using the fine-tuned linear svm algorithm." Ieee Access 11 (2023): 88025-88039.

[13] Tolkach, Yuri, Vlado Ovtcharov, Alexey Pryalukhin, Marie-Lisa Eich, Nadine Therese Gaisa, Martin Braun, Abdukhamid Radzhabov et al. "An international multi-institutional validation study of the algorithm for prostate cancer detection and Gleason grading." NPJ precision oncology 7, no. 1 (2023): 77.

[14] W. Bulten et al., "Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge," Nature Medicine, vol. 28, no. 1, 2022, 154–163. https://doi.org/10.1038/s41591-021-01620-2.

[15] Li, Bochong, Ryo Oka, Ping Xuan, Yuichiro Yoshimura, and Toshiya Nakaguchi. "Robust multi-modal prostate cancer classification via feature autoencoder and dual attention." Informatics in Medicine Unlocked 30 (2022): 100923.

[16] Malibari, Areej A., Reem Alshahrani, Fahd N. Al-Wesabi, SB Haj Hassine, Mimouna Abdullah Alkhonaini, and Anwer Mustafa Hilal. "Artificial intelligence based prostate cancer classification model using biomedical images." Comput. Mater. Contin 72 (2022): 3799-3813.

[17] Liu, Bojing, Yinxi Wang, Philippe Weitz, Johan Lindberg, Johan Hartman, Wanzhong Wang, Lars Egevad, Henrik Grönberg, Martin Eklund, and Mattias Rantalainen. "Using deep learning to detect patients at risk for prostate cancer despite benign biopsies." Iscience 25, no. 7 (2022).

[18] Hammouda, Kamal, Fahmi Khalifa, Moumen El-Melegy, Mohamed Ghazal, Hanan E. Darwish, Mohamed Abou El-Ghar, and Ayman El-Baz. "A deep learning pipeline for grade groups classification using digitized prostate biopsy specimens." Sensors 21, no. 20 (2021): 6708.

[19] M., Kusuma Sri, and Sathees Kumaran S. "Quadratic Luminance Vision Transformer Attention Network for Automated Mitosis Detection in Breast Histopathology Images." Journal of Innovative Image Processing 7, no. 2 (2025): 266-289

[20] Tyagi, Nandini, Riya Sharma, and Monika Sharma. "Deep Learning for Skin Cancer Classification: A Study of Model Accuracy, Generalization, and Ensemble Learning." Journal of Soft Computing Paradigm 7, no. 2 (2025): 124-143.