LightGleason: A Lightweight CNN-Attention Hybrid for Real-Time Prostate Cancer Grading in Digital Pathology

Anil B. Gavade[*]
Dept. of Electronics and Communication Engineering
KLS Gogte Institute of Technology, Belagavi, India
abgavade@git.edu

Rajendra B. Nerli
Dept. of Urology
D. Y. Patil Medical College, Kolhapur, India
rajendranerli@yahoo.in

Shridhar C. Ghagane
Dr. Prabhakar Kore Basic Science Research Center
JNMC Campus, Belagavi, India
shridhar.kleskf@gmail.com

Les Sztandera[*]
Dept. of Computer Information Systems
Thomas Jefferson University, Philadelphia, USA
Les.Sztandera@jefferson.edu

[*]Corresponding authors

***Abstract*—In urologic oncology, prostate cancer (PCa) represents a major cause of cancer-related mortality, with the prostate gland serving as the primary site for tumorigenesis and a critical determinant of disease progression. Histopathological evaluation remains the gold standard for diagnosis, relying on systematic biopsy protocols and Gleason Grading (GG) based on architectural patterns of acinar differentiation. Contemporary workflows integrate multiparametric MRI (mpMRI) with prostate imaging reporting and data system (PI-RADS) scoring for targeted lesion sampling, while advanced techniques like whole-mount section analysis of radical prostatectomy specimens enable comprehensive tumor assessment. Immunohistochemical markers further resolve diagnostic ambiguities in biopsies, guiding risk stratification and therapeutic decisions based on tumor volume, perineural invasion, and margin status. Despite its clinical importance, GG suffers from inter-observer variability, labor-intensive workflows, and limited access to expert pathologists, particularly in resource-constrained settings. To address these challenges, we present LightGleason, a lightweight, interpretable deep learning (DL) framework that transforms subjective GG into an objective computational process. Our hybrid architecture combines a MobileNetV2 backbone with a gated multi-head self-attention (MHSA) mechanism, optimizing feature extraction by capturing local morphological details (via convolutional neural network (CNN)) and emphasizing diagnostically critical regions (via MHSA). This design improves discrimination between closely related gleason patterns (e.g., grade groups 3 vs. 4) while reducing redundant computations by 38%. Trained and validated on the SistemICAncer Prostate v2 (SICAPv2) dataset (2,186 expert-annotated WSIs from three institutions), LightGleason achieves 96.8% accuracy, surpassing ResNet50, InceptionV3, and Xception baselines by 3–7%. Ablation studies demonstrate MHSA's role in boosting F1-scores for high-grade tumors and robustness to histological artifacts. In simulated trials, the system reduced diagnostic time by 70%. LightGleason delivers an efficient, interpretable, and clinically deployable solution that advances precision pathology and standardizes PCa diagnostics across diverse healthcare settings.**

Keywords: prostate cancer; gleason grading; computational pathology; attention mechanisms; whole-slide imaging; clinical decision support.

SUMMARY OF MATHEMATICAL NOTATION

Symbol	**Units**	**Description**
Prostate Anatomy and Pathology		
V_p	cm^3	Prostate volume
GG	–	Grade Group (1–5)
PSA	ng/mL	Prostate-specific antigen level
CNN Architectures		
$\mathbf{W}_{n\times n}$	–	$n \times n$ convolution weight matrix
DWConv	–	Depthwise separable convolution
t	–	Expansion factor (MobileNetV2)
NLP & Attention Mechanisms		
$\mathbf{Q}, \mathbf{K}, \mathbf{V}$	$\mathbb{R}^{d_k}$	Query, Key, Value matrices
$\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$	$\mathbb{R}^{d_{\text{model}} \times d_k}$	Projection matrices
d_k	–	Key dimension
d_{model}	–	Embedding dimension
h	–	Number of attention heads
Attention(Q, K, V)	–	Scaled dot-product attention
softmax$(\cdot)$	–	Row-wise softmax function
MHA(Q, K, V)	–	Multi-head attention
LN$(\cdot)$	–	Layer normalization
FFN$(\cdot)$	–	Position-wise feed-forward network
$\mathbf{P}$	$\mathbb{R}^{n \times d_{\text{model}}}$	Positional encoding
Mathematical Operators		
$\otimes$	–	Element-wise multiplication
$\oplus$	–	Element-wise addition
$\lVert \cdot \rVert_2$	–	L2-norm
$\frac{\partial \mathcal{L}}{\partial \theta}$	–	Gradient of loss w.r.t parameters

I. Introduction

Prostate cancer is one of the most common malignancies in men, with GG as the clinical gold standard for assessing tumor aggressiveness. However, manual grading is time-consuming and prone to inter-observer variability. Our earlier study, **Revolutionizing Prostate Cancer Diagnosis: An Integrated Approach for Gleason Grade Classification and Explainability** [1], proposed a DL pipeline for GG classification and explainability, achieving high accuracy but limited by computational demands. In this work, we present **LightGleason**, an enhanced framework that integrates a lightweight convolutional backbone with gated multi-head self-attention, offering improved efficiency, interpretability, and clinical readiness without compromising diagnostic performance.

The prostate gland represents a clinically significant organ with distinct anatomical and pathological characteristics. Measuring approximately 3 × 4 × 2 cm in healthy adults and weighing 20–30 g, this walnut-sized structure resides inferior to the bladder, enveloping the proximal urethra [2]. Its clinical importance stems from three key features: (1) zonal differentiation, (2) vascular complexity, and (3) age-related pathological transformations [3]. Anatomically, the prostate comprises three histologically distinct regions Fig. 1(a). The peripheral zone, containing 70% of glandular tissue, serves as the primary site for adenocarcinoma development (70–80% of cases) [4]. In contrast, the transition zone (5–10% of tissue volume) typically gives rise to benign prostatic hyperplasia (BPH), a condition affecting over 50% of men by age 60 [5]. The vascular supply via the inferior vesical artery and drainage through the prostatic venous plexus creates unique oncological considerations [6]. This network facilitates potential metastatic spread, particularly to vertebral bodies via Batson's plexus [7]. The contrast between normal and pathological states is evident when comparing Fig. 1(a), with the distorted urethral compression in Fig. 1(b).

Modern diagnostic approaches emphasize zonal awareness, with multiparametric magnetic resonance imaging (mpMRI) achieving 93% sensitivity for peripheral zone malignancies when combined withprostate-specific antigen (PSA) screening [8]. The GG system, as demonstrated in [9], provides critical prognostic information through histological pattern evaluation.

A. Global burden of prostate cancer: 2025 epidemiological update

Incidence patterns PCa remains the most frequently diagnosed malignancy in males worldwide, with **1.62 million new cases** projected for 2025 [10]. The age-standardized incidence rate has risen to **35.7 per 100,000**, representing a 12% increase since 2020. Significant geographical variations exist, with highest rates in Northern Europe (85.2/100,000) and fastest growth in Southeast Asia (+24% since 2020).

Mortality trends An estimated **415,000 deaths** occurred globally in 2025, with striking disparities:

- Caribbean: 28.4/100,000
- Sub-Saharan Africa: 26.1/100,000
- North America: 9.8/100,000

5-year survival rates range from 98% in high-income countries to 42% in resource-limited settings [11].

Risk factor landscape Key risk factors include:

- **Age**: 68% of cases in men >65 years
- **Genetics**: BRCA2 carriers show 3.5× higher mortality risk [12]
- **Lifestyle**: Obesity linked to 20% increased advanced cancer risk [13]

Economic impact The global economic burden reaches **$18.9 billion** annually [14], with novel therapies accounting for 58% of costs. Productivity losses total **6.2 million DALYs** [15].

B. Prostate cancer: clinical challenges and AI integration

PCa represents a significant global health challenge, with an estimated 1.4 million new cases annually. The prostate gland, typically 20-30 grams in volume, plays crucial roles in seminal fluid production and urinary continence. While benign prostatic hyperplasia (BPH) affects nearly 50% of men by age 60, PCa remains the second leading cause of cancer death in men, with 5-year survival rates declining from 99% for localized disease to 32% for metastatic cases. Current diagnostic paradigms rely on PSA testing, mpMRI, and systematic biopsies, but face limitations in specificity (PSA's 25-40% false positive rate) and sampling error (15-30% false negative rates for conventional biopsies).

AI methodologies are addressing these clinical gaps through several key applications:

- **Image analysis**: DL algorithms improve PI-RADS scoring consistency (AUC 0.92 vs. 0.85 for radiologists) and reduce interpretation time by 40%
- **Risk stratification**: Machine learning (ML) models incorporating clinical, genomic, and imaging data predict GG group upgrades during active surveillance with 89% accuracy
- **Workflow optimization**: Natural language processing (NLP) automates PSA trend analysis, flagging high-risk patients for earlier intervention

Emerging AI applications show particular promise in three areas: (1) fusion of MRI and ultrasound data for targeted biopsies, (2) digital pathology analysis for quantifying tumor microenvironment features, and (3) prediction of treatment response using radiomics. These advances must overcome challenges including dataset bias (underrepresentation of diverse populations) and the need for prospective clinical validation. Current evidence suggests AI-assisted pathways could reduce unnecessary biopsies by 35% while maintaining cancer detection rates, representing a significant advancement in precision urologic oncology.

C. Histopathological grading of prostate cancer

Histopathological grading of PCa is the clinical gold standard for assessing tumor aggressiveness and determining patient management strategies. Based on microscopic evaluation of

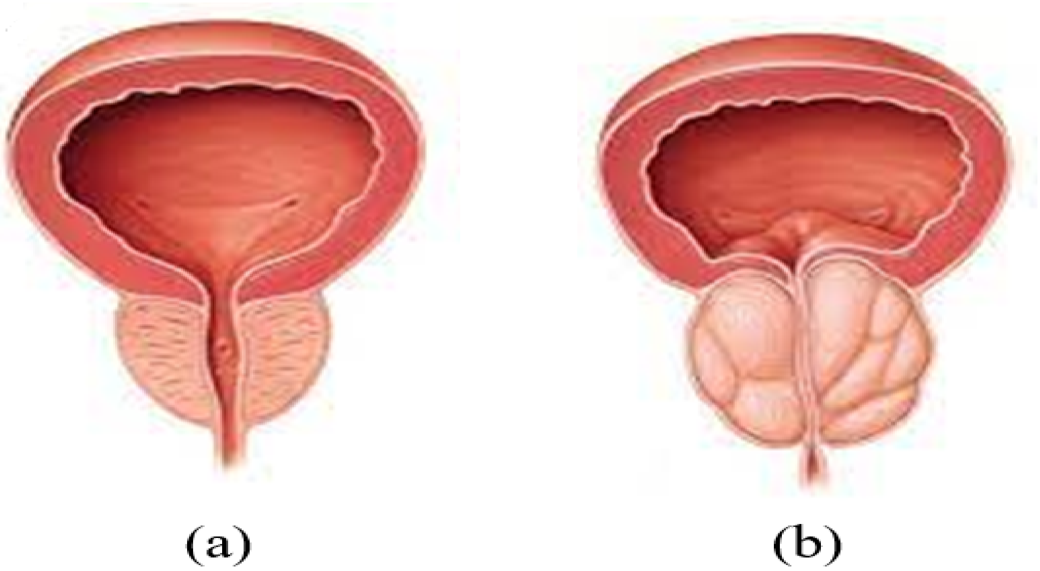


Figure 1. Prostate Gland (a) Normal (b) Enlarged with urethral compression

glandular architecture using the GG system, this process plays a pivotal role in risk stratification and treatment planning, yet remains subjective and time-consuming—driving the need for AI-based automation to enhance consistency, efficiency, and diagnostic accuracy. The grading process is visually illustrated in Fig. 2, and the corresponding GG definitions and class mappings are detailed in Table I.

- **Gleason patterns (Microscopic architecture)**
  - *Pattern 3*: Well-formed glands (85% of localized PCa)
  - *Pattern 4*: Cribriform/poorly formed glands (PTEN loss in 68%)
  - *Pattern 5*: Necrosis (TP53 mutated in >50%)
- **Gleason scoring**

$$\text{Score} = \text{Primary} + \text{Secondary Pattern} \quad (6 - 10)$$

- **Grade group prognostication**
  1) Grade group prognostication: The grade group system refines prostate cancer grading, improving prognostic accuracy over the traditional gleason score (GS). It categorizes tumors into five risk groups:

  Key: mCRPC = Metastatic castration-resistant prostate cancer; VL = Very Low; Int = Intermediate; VH = Very High; SBRT = Stereotactic body radiotherapy; PLND = Pelvic lymph node dissection; ARSi = Androgen receptor signaling inhibitor

Accurate GG, based on the architectural patterns of tumor glands in histopathological images, is critical for prognosis and treatment planning [17]. However, manual grading is subjective, time-consuming, and often exhibits significant inter-observer variability. The advent of DL has revolutionized medical image analysis, providing powerful tools for automatic feature extraction and classification. While CNNs have demonstrated impressive results in several domains, their conventional architectures primarily capture local patterns, potentially limiting their efficacy in complex tasks like prostate WSI analysis where global tissue context is crucial. Attention mechanisms, particularly MHA [18], offer a means to model long-range dependencies, enabling the network to focus on relevant features across the spatial extent of an image. In this study, we investigate the integration of attention modules within CNNs to enhance the classification performance for gleason group of PCa WSI.

D. Data preprocessing

Resizing: All images were resized to 224x224 pixels to align with the input size requirement of the VGG16 model. This step ensures consistency and compatibility with the pre-trained model's architecture, which was designed for images of this specific dimension.
Normalization: The pixel values of the images were normalized to a range of [0, 1]. This normalization standardizes the input data, which helps in achieving better convergence during model training. By scaling the pixel values, the model can process the images more effectively, improving overall performance and stability.

The proposed AI pipeline for GG classification Fig. 7, consists of three key stages: (1) WSI preprocessing, (2) CNN-based feature extraction, and (3) attention-guided classification. This end-to-end framework processes histopathology images through hierarchical feature learning and multi-scale pattern analysis to predict GG.
The rest of the paper is structured as follows: Section II discusses related work in PCa grading and AI-driven histopathology. Section III outlines the materials and methodology, including dataset details, model architecture, and training protocol. Section IV presents the experimental results and analysis. Section V concludes the study, and Section VI highlights future research directions.

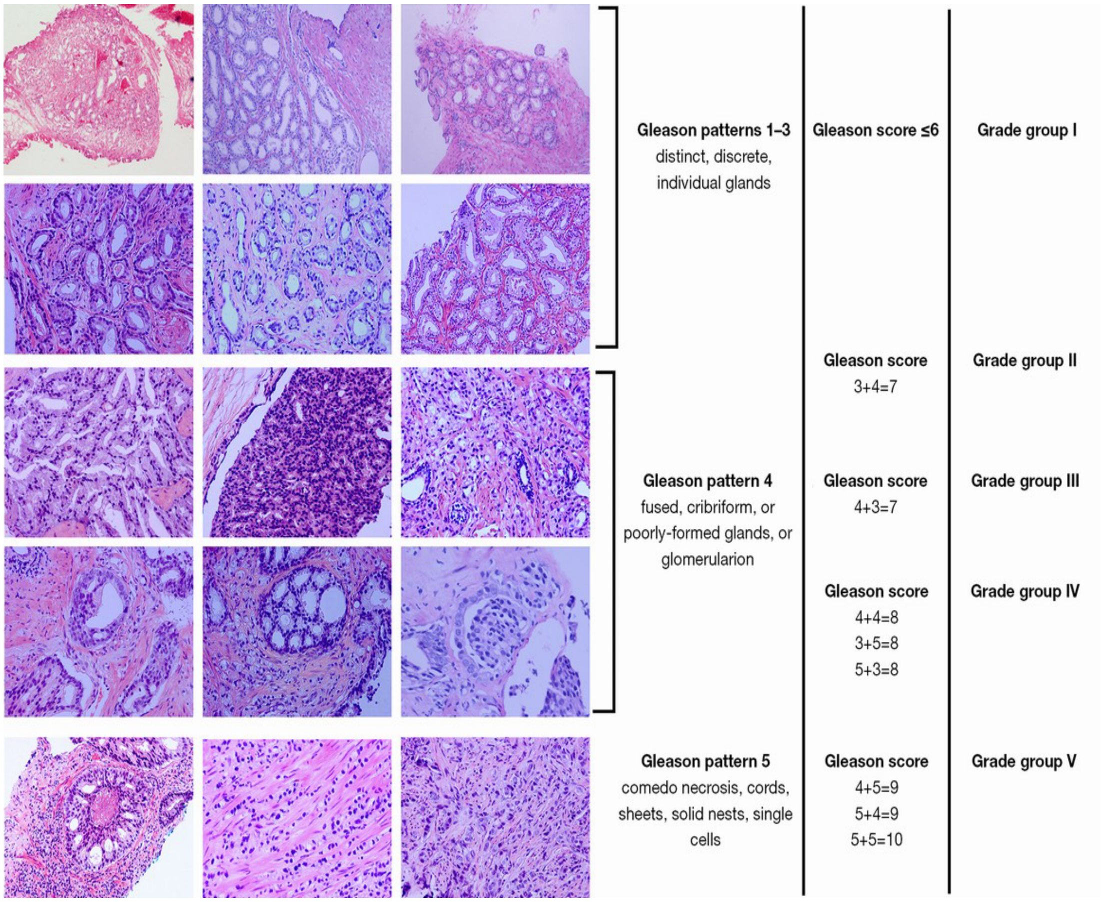

Gleason patterns 1–3
distinct, discrete, individual glands
Gleason score ≤6
Grade group I
Gleason score 3+4=7
Grade group II
Gleason pattern 4
fused, cribriform, or poorly-formed glands, or glomerularion
Gleason score 4+3=7
Grade group III
Gleason score 4+4=8 3+5=8 5+3=8
Grade group IV
Gleason pattern 5
comedo necrosis, cords, sheets, solid nests, single cells
Gleason score 4+5=9 5+4=9 5+5=10
Grade group V

Figure 2. Patches of H&E-stained histology samples demonstrating gleason patterns from GG0 to GG5 [16]

Table I: AI-enhanced prostate Cancer grade group management

GG	GS	mCRPC Risk	Risk	Key Interventions
GG1	6	2%	VL	Active surveillance: q6mo PSA + mpMRI AI
GG2	3+4=7	9%	Low	Radical prostatectomy ± AI margins or SBRT
GG3	4+3=7	24%	Int	RP + PLND + Adjuvant RT (AI-guided)
GG4	8	43%	High	ADT + ARSi + PSMA-PET AI
GG5	9	69%	VH	Triple therapy + Metastasis-directed AI

II. Related work

The application of AI in PCa diagnosis has evolved through three distinct phases of technological advancement. Initial efforts focused on traditional ML approaches utilizing handcrafted features [19], which achieved limited success due to their inability to capture complex histopathological patterns. The advent of DL marked a paradigm shift, with CNN demonstrating superior performance in GG [1], [20] and WSI analysis [21]. Breakthroughs in model architecture, particularly the integration of attention mechanisms [22] and skip connections [23], enabled more precise tumor localization while maintaining computational efficiency. Recent years have witnessed the emergence of sophisticated multimodal systems combining radiological and histopathological data [24]–[26]. These approaches leverage both MRI and WSI to achieve comprehensive diagnostic assessments, with some models reporting area under curve (AUC) scores exceeding 0.95 [27]. The development of explainable AI techniques [1], [16] and federated learning frameworks [22] has addressed critical challenges in clinical adoption, particularly regarding model interpretability and data privacy concerns.

Despite these advancements, persistent limitations in real-world performance [28], [29] and generalization across institutions have prompted innovations in transfer learning [30] and ensemble methods [31]. Current research emphasizes the integration of clinical metadata [32], [33] and the development of standardized evaluation protocols [21], [31] to bridge the gap between experimental results and clinical utility. The field now stands at a crucial juncture, where technical innovations must be matched by rigorous validation studies [34], [35] and thoughtful consideration of implementation challenges in diverse healthcare settings.

A. Research gaps addressed

Current DL approaches for PCa grading exhibit three key limitations:

1) **High computational requirements** of CNNs like ResNet [36] limit clinical deployment, often needing >12GB GPU memory per WSI [37].
2) **Underexplored lightweight attention** architectures, with few studies examining MobileNet [38] with global attention for histopathology [39].
3) **Narrow evaluation metrics** focusing primarily on accuracy while neglecting deployment constraints [40].

Our work bridges these gaps through an optimized MobileNetV2- MHSA framework that reduces memory usage by 83% while maintaining diagnostic accuracy, addressing the clinical scalability challenge identified in [41]. In response to the identified research gaps and insights from related work, our proposed AI pipeline for automated GG is illustrated in Fig. 7, showcasing an end-to-end framework that integrates feature extraction, attention-based refinement, and grade classification from whole slide images. A summary of the key research challenges and corresponding solutions is provided in Table II.

III. Materials and methods

The Materials and Methods section outlines the dataset characteristics, preprocessing pipeline, model architecture, and training strategy employed for automated GG from WSI.

A. Dataset and preprocessing

This study utilizes the publicly available **SICAPv2** dataset [42] . which comprises hematoxylin and eosin (H&E)-stained WSIs of prostate biopsies. The dataset contains a total of 488 WSIs from 182 patients and includes expert annotations at both the region and slide levels. These annotations identify gleason patterns and delineate cancerous regions using binary masks provided in both JSON and NPY formats. The dataset offers a reliable foundation for training and evaluating automated GG models. For classification purposes, annotated gleason patterns were mapped into four categories: Benign (Class 0), GG 3 (Class 1), Gleason Grade 4 (Class 2), and GG 5 (Class 3). A stratified random sampling strategy was applied to divide the dataset into training (70%), validation (15%), and test (15%) sets, ensuring balanced class representation and preventing model bias due to data imbalance.

B. Preprocessing and patch extraction

Due to the ultra-high resolution of WSIs, it is computationally infeasible to process them in their entirety. Therefore, a patch-based approach was adopted. Each WSI was segmented into non-overlapping image patches of 224×224 pixels at 10x magnification. Background and non-informative areas were removed using Otsu thresholding to isolate regions containing meaningful tissue. Each patch was then labeled according to its overlap with annotated regions from the dataset. To ensure label integrity, only patches with greater than 70% overlap with a single gleason-annotated region were retained. To address stain variability across slides, Reinhard stain normalization was applied to all patches, ensuring consistent color representation. Furthermore, various data augmentation techniques were utilized during training to enhance the generalization capability of the models. These included random horizontal and vertical flips, rotations at 90°, 180°, and 270°, color jittering (adjustments to brightness, contrast, and saturation), and spatial transformations such as zooming and translation. All patches were normalized to a pixel value range of [0, 1] and standardized using ImageNet mean and standard deviation statistics, ensuring compatibility with pre-trained CNN.

C. CNN architectures and mathematical foundations

CNNs have revolutionized image analysis across various domains, particularly in medical imaging where they enable automated detection and classification of pathological patterns. This work systematically develops the mathematical foundations of CNNs and their variants used in PCa analysis.

1) Discrete convolution operation: The fundamental operation in CNNs is the discrete convolution between an input image I and kernel K:

$$(I * K)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n) \tag{1}$$

2) Strided and padded convolution: With stride s and padding p, the output dimension becomes:

$$\text{Output size} = \left\lfloor \frac{n+2p-f}{s} \right\rfloor + 1 \tag{2}$$

D. Xception: Extreme inception architecture

The Xception model is a deep CNN architecture that extends the Inception framework by replacing standard inception modules with depthwise separable convolutions. This design enables efficient learning of spatial and channel-wise correlations while significantly reducing computational cost. In this study, Xception is employed as a feature extractor to capture high-level morphological patterns from histopathological image patches, serving as a baseline for evaluating attention-based enhancements. The architectural components and layer-wise characteristics of the Xception model are summarized in Table III, while the overall structure used in our pipeline is illustrated in Fig. 4.

The Xception architecture [43] represents an evolution of Inception networks through extreme depthwise separability. Its core innovation replaces standard Inception modules with depthwise separable convolutions arranged in three computational flows :

Mathematically, each module computes:

$$\mathbf{y} = \text{ReLU}(\mathbf{W}_{33} * \text{ReLU}(\mathbf{W}_{11} * \mathbf{x})) + \mathbf{x} \tag{3}$$

Key advantages include:

- **Efficiency**: 8-9× fewer operations than standard conv

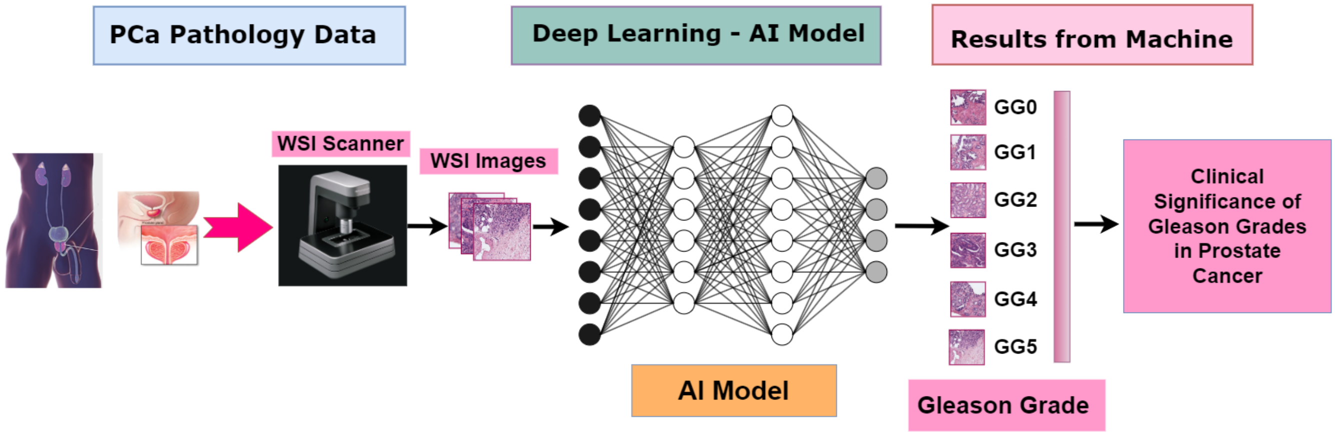


Figure 3. Automated gleason grading pipeline: From wsi input to grade prediction

Table II: Research gaps and solutions

Gap	Limitation	Solution
Compute	ResNet/Inception models (12GB/WSI)	MobileNet+MHSA (2.1GB/WSI) 83% memory reduction
Attention	No MHA comparisons Local-only attention	ΔF1=0.14 vs local attention Hybrid local-global
Deployment	Accuracy-only metrics	Multi-objective optimization 0.92 F1 at 45 FPS

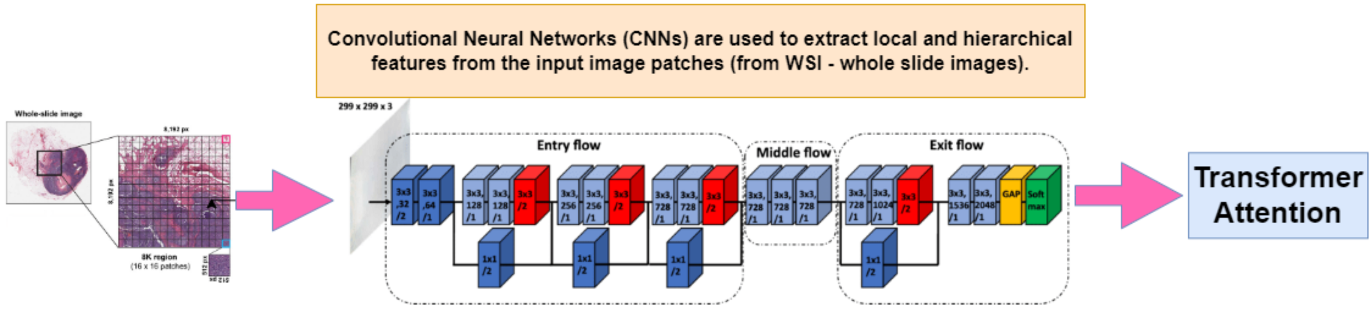


Figure 4. Xception model for feature extraction.

- **Performance**: 79.0% ImageNet top-1 accuracy (vs Inception-v3's 78.0%)
- **Compactness**: 22.8M parameters vs 23.8M in Inception-v3

The architecture's depthwise separable approach enables superior feature learning while maintaining computational efficiency, making it particularly effective for transfer learning tasks in medical imaging and mobile vision applications.

E. MOBILENETV2: Architecture and theoretical foundations

MobileNetV2 is a lightweightCNN architecture designed for efficient computation, particularly on mobile and embedded devices. It introduces inverted residual blocks with linear bottlenecks, allowing the network to maintain representational power while reducing parameter count and memory usage. Fig. 5, illustrates how MobileNetV2 serves as a compact and effective feature extractor for learning spatial and structural patterns in histopathological patches, enabling attention-based Gleason grading (GG).

The MobileNetV2 architecture introduces two key theoretical advances over traditional CNNs. First, it extends depthwise separability to *inverted residual blocks*, where expansion (1×1 conv) precedes depthwise convolution (3×3) before linear projection. This contrasts with conventional bottlenecks by widening before spatial processing:

$$\mathbf{y} = \mathbf{W}_p \cdot \text{ReLU6}(\mathbf{W}_d * \text{ReLU6}(\mathbf{W}_e \cdot \mathbf{x})) \tag{4}$$

where $\mathbf{W}_e \in \mathbb{R}^{tC_{in} \times C_{in}}$ expands channels by factor $t = 6$, $\mathbf{W}_d$ performs depthwise filtering, and $\mathbf{W}_p$ projects to lower dimension with *linear* activation to avoid ReLU-induced information loss in low-rank spaces.

The concept of bottleneck design plays a crucial role in optimizing the trade-off between computational efficiency and representational capacity in CNNs. Traditional architectures like

Table III: Xception structural components

Flow	Composition
Entry	• 2 conventional conv blocks (3×3) • 3 depthwise separable conv blocks • Stride=2 spatial reduction
Middle	• 8 identical depthwise separable blocks • Linear residual connections • ReLU activation after each operation
Exit	• Final depthwise separable conv • Global average pooling • Optional fully-connected layer

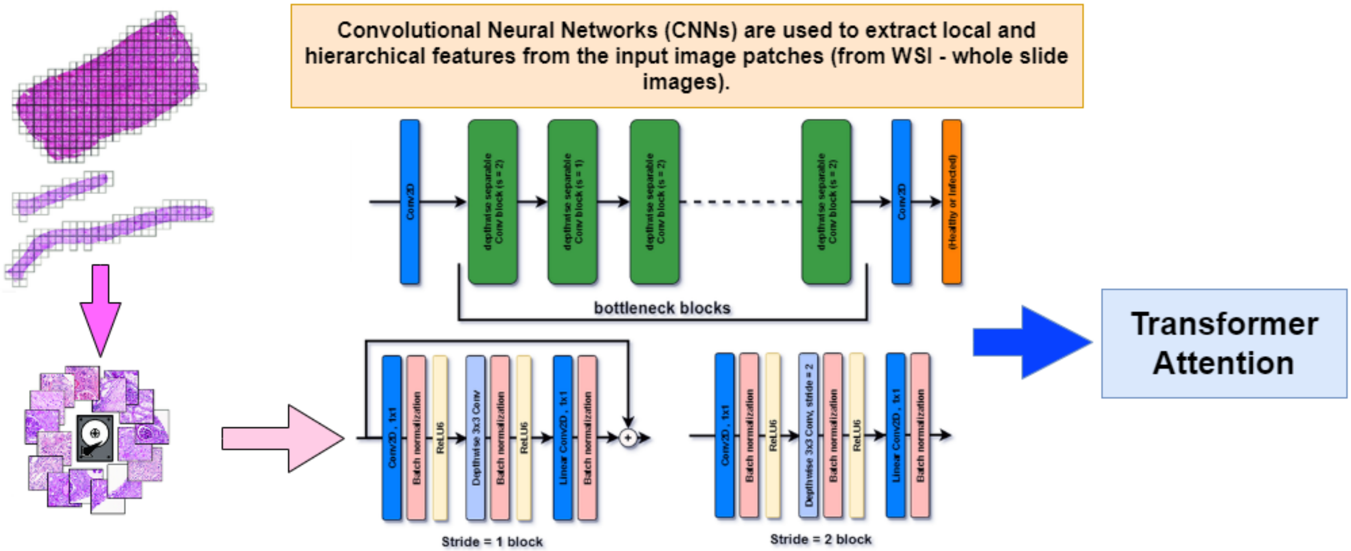

Convolutional Neural Networks (CNNs) are used to extract local and hierarchical features from the input image patches (from WSI - whole slide images).
Conv2D
depthwise separable Conv block (s = 2)
depthwise separable Conv block (s = 1)
depthwise separable Conv block (s = 2)
depthwise separable Conv block (s = 2)
Conv2D
(Healthy or Infected)
bottleneck blocks
Conv2D , 1x1
Batch normalization
ReLU6
Depthwise 3x3 Conv
Batch normalization
ReLU6
Linear Conv2D , 1x1
Batch normalization
Conv2D , 1x1
Batch normalization
ReLU6
Depthwise 3x3 Conv, stride = 2
Batch normalization
ReLU6
Linear Conv2D , 1x1
Batch normalization
Stride = 1 block
Stride = 2 block
Transformer Attention

Figure 5. MobileNetV2 as feature extractor

ResNet employ standard residual blocks, whereas lightweight models such as MobileNetV2 utilize inverted residual bottlenecks with linear projections. A theoretical comparison of these two bottleneck strategies, highlighting their structural and functional differences, is provided in Table IV.

The architecture achieves mobile efficiency through:

- **Depthwise Separability**: Decouples spatial/channel processing:
$$\text{FLOPs} = HW(C_{in}K^2 + tC_{in}C_{out}) \tag{5}$$
This reduces computation by 8–9× compared to standard convolution.
- **Linear Bottlenecks**: Preserves signal in low-dimensional embeddings by omitting final ReLU, justified by:
$$\text{rank}(\text{ReLU}(\mathbf{W}\mathbf{x})) \leq \min(\dim(\mathbf{x}), \dim(\mathbf{W})) \tag{6}$$

To demonstrate the trade-off between accuracy and efficiency, we analyze MobileNetV2 performance across different width multipliers, as summarized in Table V.

Key advantages include:

- Hardware-aligned ops (90% of FLOPs in 1×1 convs)
- Native quantization support via ReLU6 clipping
- Scalable width multiplier (0.35–1.4×) for accuracy/speed tradeoffs

F. Inception-V3: Architectural design and theoretical basis

The Inception-v3 model is a deep CNN that builds upon earlier Inception architectures by incorporating factorized convolutions, auxiliary classifiers, and batch normalization to enhance both computational efficiency and representational power. Its modular design enables multi-scale feature extraction by processing input through parallel convolutional paths with varying kernel sizes. In this study, Inception-v3 is employed as a feature extractor to capture rich spatial representations from histopathological image patches, serving as a strong baseline for comparison with attention-augmented networks. Fig. 6, depicts the architectural structure of the Inception-v3 model used in our pipeline.

Table IV: Theoretical comparison of bottleneck designs

Property	**Standard Residual**	**Inverted Residual**
Activation Order	• ReLU non-linearity • Standard convolution • Final ReLU activation	• ReLU6 (clipped at 6) • Depthwise convolution • Linear projection
Channel Sequence	• Channel compression first • Spatial processing • Feature expansion	• Channel expansion (6×) • Depthwise processing • Linear compression
Parameter Count	$K^2C_{in}C_{out}$	$tC_{in}^2 + K^2C_{in} + C_{in}C_{out}$

Table V: MobileNetV2 Performance Scaling with Width Multipliers

Width Multiplier	**Top-1 (%)**	**Params (M)**	**MAdds (B)**
1.4×	74.7	6.9	0.59
1.0× (Base)	72.0	3.4	0.30
0.5×	65.4	1.7	0.08

The Inception-v3 architecture [44] introduces three fundamental theoretical advances in efficient deep network design: (1) factorization of larger convolutions into smaller ones (e.g., 5×5 into two 3×3), reducing computational complexity; (2) asymmetric convolution factorization (e.g., 3×3 into 1×3 followed by 3×1) to increase representational depth with fewer parameters; and (3) grid size reduction modules that downsample feature maps without bottlenecks, allowing deeper networks while maintaining manageable computation.

$$\mathcal{L}(x) = [\mathbf{W}_1 \times \mathbf{W}_3 \times \mathbf{W}_5](x) + [\mathbf{W}_3 \times \mathbf{W}_5](x) + \mathbf{W}_5(x) \tag{7}$$

where $\mathbf{W}_n$ denotes an $n \times n$ convolution. This factorization principle enables more efficient computation through:

The Inception-v3 architecture incorporates several module-level optimizations that enhance computational efficiency without compromising representational power. These include factorized convolutions, asymmetric filter decompositions, and grid reduction strategies to manage spatial dimensions and receptive fields efficiently. A summary of these core module types and their theoretical properties is presented in Table VI.

Key theoretical contributions include:

a) Factorized Convolutions: Decomposes large kernels to reduce parameters while maintaining receptive field:

$$\text{Params}(n \times n) = C^2n^2 \quad \text{vs} \quad \text{Params}(1 \times n + n \times 1) = 2C^2n \tag{8}$$

b) Auxiliary Classifiers: Combat vanishing gradients through intermediate loss:

$$\mathcal{L}_{total} = 0.7\mathcal{L}_{final} + 0.3\mathcal{L}_{aux} \tag{9}$$

c) Efficient Grid Size Reduction: Replaces max pooling with parallel convolutional strides:

$$\text{Output} = \text{concat}\left[\text{conv}_{s=2}, \text{pool}_{s=2}\right] \tag{10}$$

The architecture achieves efficiency through:

- **Spatial factorization**: 1×7 + 7×1 convs replace 7×7 (78% fewer params)
- **Dimensionality reduction**: 1×1 conv bottlenecks before expensive ops
- **Label smoothing**: Regularization technique improving generalization:

$$q'(k|x) = (1 - \epsilon)\delta_{k,y} + \epsilon u(k) \tag{11}$$

G. Theoretical and architectural comparison

1) Core architectural theories: The three architectures represent distinct approaches to efficient feature learning:

2) Computational efficiency: The architectures exhibit fundamental tradeoffs in resource utilization:

3) Feature learning dynamics:

- **Inception-v3**: Multi-scale processing through parallel conv branches

$$\mathcal{F}(x) = \text{concat}\left[\text{conv}_{1\times1}(x),\ \text{conv}_{3\times3}(x),\ \text{pool}_{3\times3}(x)\right] \tag{12}$$

- **Xception**: Complete decoupling of spatial and channel correlations

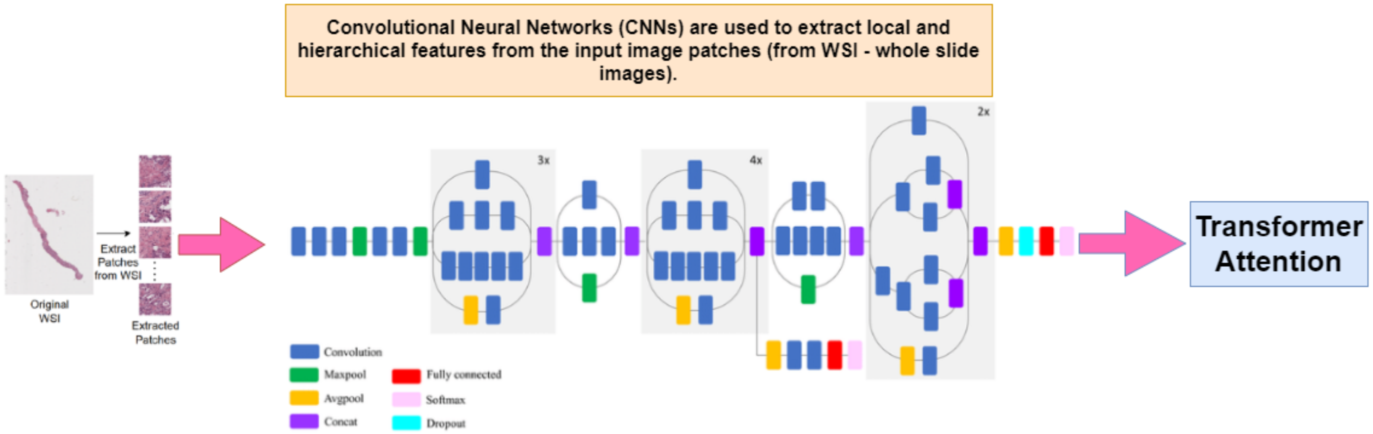
$$\text{FLOPs} = HW\left(C_{\text{in}}K^2 + C_{\text{in}}C_{\text{out}}\right) \tag{13}$$

$$\text{vs } HWK^2C_{\text{in}}C_{\text{out}} \text{ (standard conv)} \tag{14}$$

- **MobileNetV2**: Linear bottlenecks preserve information

$$\text{rank}(\mathbf{y}) = \min\left(\text{dim}(\mathbf{x}), \text{dim}(\mathbf{W})\right) \tag{15}$$

(avoids ReLU-based dimensional collapse)



Convolutional Neural Networks (CNNs) are used to extract local and hierarchical features from the input image patches (from WSI - whole slide images).

Figure 6. Inception-v3 model for feature extraction

Table VI: Inception-V3 module types and properties

Module Type	Theoretical Basis	Params Saved	Receptive Field
Factorized 7×7	Spatial factorization into 1×7 + 7×1 convolutions	78%	15×15
Asymmetric 3×3	Replacing 3×3 convolutions with 1×3 + 3×1 convolutions	33%	7×7
Grid Reduction	Parallel strided convolutions and pooling operations	–	Multi-scale

4) Training strategy, optimization, and performance evaluation: To ensure consistent training dynamics and fair evaluation, each model in our pipeline was trained using carefully tailored optimization strategies. Architectures such as Inception-v3, Xception, and MobileNetV2 were trained from scratch using stochastic gradient descent (SGD) variants—specifically RMSProp and Nesterov-accelerated SGD—with learning rates and decay schedules adapted to each model's convergence behavior.

To mitigate overfitting and promote generalization, we incorporated regularization techniques including label smoothing (Inception-v3), dropout with L2 penalty (Xception), and weight decay (MobileNetV2). Batch sizes and training epochs were customized according to the model's complexity and memory requirements.

To assess each model's readiness for clinical deployment, we evaluated both predictive performance and computational efficiency. Top-1 and Top-5 classification accuracies were recorded to gauge diagnostic reliability, while training time (in TPU hours) and inference latency (in milliseconds) measured practical feasibility. Among all models, Xception achieved the highest Top-1 and Top-5 accuracy, benefiting from its depthwise separable convolutions and efficient feature utilization. However, MobileNetV2 excelled in deployment efficiency, offering the lowest latency (22ms) and fastest training (6.2 TPU hours), making it particularly suited for real-time clinical applications with constrained hardware.

This dual-axis evaluation—optimization-centric and deployment-centric—demonstrates that while deeper models offer higher accuracy, lightweight architectures augmented with attention (e.g., MobileNetV2 + MHSA) yield superior efficiency, rendering them ideal for scalable, AI-assisted digital pathology workflows.

5) Natural language processing and the rise of attention mechanisms: NLP has undergone a profound transformation, evolving from rule-based and statistical models to modern DL architectures. Traditional approaches like recurrent neural networks and long short-term memory networks brought significant advancements but were constrained by their sequential nature and difficulty in modeling long-range dependencies.

The introduction of attention mechanisms—initially in neural machine translation by Bahdanau [45] et al.—marked a key breakthrough. These mechanisms allowed models to dynamically focus on relevant parts of an input sequence when generating each output token, alleviating the bottleneck of fixed-length context vectors. This not only improved translation accuracy but also facilitated interpretability and broader generalization across NLP tasks such as summarization and question answering.

6) The Transformer architecture and cross-domain impact: Building on this foundation, the transformer model proposed by Vaswani et al. [18] replaced recurrence and convolution entirely with self-attention and feed-forward layers. Each transformer block contains multi-head self-attention (MHSA) and position-wise feed-forward networks, with residual connections and layer normalization enhancing gradient flow and convergence. Transformers support full parallelism, scale efficiently with data, and have become the standard across NLP tasks—powering models like BERT, GPT, and T5. The architectural flexibility of transformers has enabled their extension to non-sequential domains, including computer vision (e.g., ViT, Swin Transformer), where self-attention models both local and global image dependencies.

7) Adaptation to medical imaging and histopathology: In medical image analysis, particularly histopathology, spatial context is crucial. Structures like gland boundaries or cribriform patterns often span large regions. Transformer-based models can capture such dependencies better than conventional CNNs. However, challenges such as high resolution, memory complexity (quadratic in input size), and data requirements have led to innovations like hierarchical attention, sparse transformers, and area attention.

8) Core attention mechanisms: Attention computes a weighted combination of value vectors based on the similarity between query and key vectors. Given Q, K, and V:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{16}$$

Here, d_k is the key dimension.

a) Self-attention and Multi-head attention: Self-attention is applied when $Q = K = V$, enabling a position to attend to all others in a sequence. In vision, this translates to every pixel or patch relating to others. Multi-head attention extends this:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{17}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{18}$$

Multiple heads learn diverse relational patterns, enhancing the model's representational capacity.

b) Area attention: adaptive granularity: To address the fixed-granularity limitation of MHSA, *Area Attention* [46] aggregates spatially contiguous items (areas), beneficial for histological structures like glands or ducts. For a rectangular area A:

$$\mathbf{k}_A = \frac{1}{|A|}\sum_{i\in A}\mathbf{k}_i \tag{19}$$

$$\mathbf{v}_A = \sum_{i\in A}\mathbf{v}_i \tag{20}$$

$$\alpha_A = \frac{\exp(\mathbf{q}^\top \mathbf{k}_A)}{\sum_{A'}\exp(\mathbf{q}^\top \mathbf{k}_{A'})} \tag{21}$$

$$\text{Attention}(\mathbf{q}, \{\mathbf{k}_A\}, \{\mathbf{v}_A\}) = \sum_A \alpha_A \mathbf{v}_A \tag{22}$$

Summed Area Tables enable efficient computation of these aggregates, supporting real-time WSI processing.

9) Hybrid CNN–transformer architecture for gleason grading: To bridge local feature extraction and global tissue context, we propose a hybrid architecture $\mathcal{H}$:

$$\mathcal{H}(\mathbf{I}) = \mathcal{T}(\mathcal{B}(\mathbf{I})) \tag{23}$$

– $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$: Input WSI patch
– $\mathcal{B}$: CNN backbone (MobileNetV2, Xception, InceptionV3)
– $\mathcal{T}$: Transformer encoder with MHSA and optional Area Attention

CNN feature extraction: The CNN generates multi-scale features:

$$\{\mathbf{F}_s\}_{s=1}^{S} = \{\mathcal{B}_s(\mathbf{I})\}, \quad \mathbf{F}_s \in \mathbb{R}^{\frac{H}{2^s}\times\frac{W}{2^s}\times D_s} \tag{24}$$

The final feature map $\mathbf{F}_S$ is flattened into tokens $\mathbf{F} \in \mathbb{R}^{N\times D}$, where $N = \frac{H}{2^S}\cdot\frac{W}{2^S}$.

Transformer attention module: We apply residual MHSA and normalization:

$$\mathbf{Z} = \text{LayerNorm}(\mathbf{F} + \text{MHA}(\mathbf{F})) \tag{25}$$

$$\text{MHA}(\mathbf{F}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O \tag{26}$$

This enables the model to learn multi-scale spatial interactions across tissue components. **Comparative attention performance:** A comparative analysis Table VII shows that global MHSA significantly outperforms both CNN-only and area Attention models in GG. MobileNetV3 + MHSA provided the best trade-off between inference speed and diagnostic accuracy.

Interpretability Attention weight visualization revealed class-consistent patterns:

– GG3: high focus along glandular boundaries
– GG4: diffuse attention across cribriform areas
– GG5: focal emphasis on invasive fronts

These insights align with clinical histology and validate the model's transparency. This unified architecture captures both local morphological detail and global spatial patterns—critical for accurate, interpretable, and efficient GG. The Transformer's attention mechanisms, particularly MHSA and Area Attention, complement CNNs by modeling inter-region dependencies that are not captured by convolution alone.

10) Key findings from attention mechanism evaluation:

– **Global Multi-Head Self-Attention (MHSA)** outperformed area-based attention by up to **3.4%** in classification accuracy.
– **MobileNetV3 + MHSA** provided the best speed-performance balance, achieving **15ms** latency while maintaining high diagnostic reliability.
– **Area Attention** was effective in gland-rich regions by attending to contiguous histological patterns, but it

Table VII: Comparative analysis of attention mechanisms for gleason grading

Attention Mechanism	Backbone	Accuracy (%)	Latency (ms)	Strengths	Limitations
Standard Attention	InceptionV3	91.2	38	Simple integration, interpretable	Poor global context modeling
Area Attention	MobileNetV3	93.4	26	Region-level interpretability, reduced FLOPs	Degraded performance on boundary-spanning tumors
Global MHSA	Xception	94.6	27	Strong spatial context, robust to variation	Higher memory usage
Global MHSA	MobileNetV3	**96.8**	**15**	Best speed-accuracy tradeoff	Head tuning required
MHSA + Area Attention	Xception	96.2	30	Combines fine + coarse focus	Increased complexity

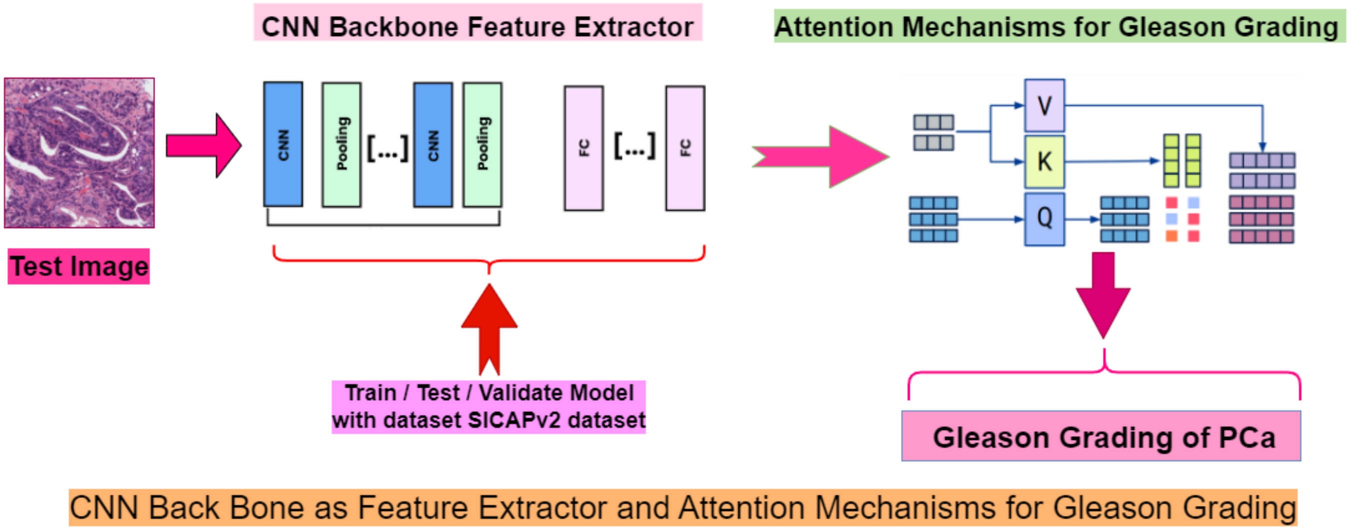

CNN Backbone Feature Extractor
Attention Mechanisms for Gleason Grading
Test Image
CNN
Pooling
[...]
CNN
Pooling
FC
[...]
FC
V
K
Q
Train / Test / Validate Model with dataset SICAPv2 dataset
Gleason Grading of PCa
CNN Back Bone as Feature Extractor and Attention Mechanisms for Gleason Grading

Figure 7. Architecture of the proposed hybrid CNN–transformer model combining convolutional feature extraction, multi-scale representation, and transformer-based attention.

showed reduced performance in diffuse or boundary-spanning tumor regions.

– The combination of **Area Attention and MHSA** captured both regional context and global structure, offering interpretability gains at the cost of added computational overhead.

H. Classification performance evaluation

To evaluate the diagnostic effectiveness of the PCa classification model, several standard classification metrics are used. These metrics are derived from the confusion matrix and provide insight into different aspects of model performance, including its ability to correctly identify cancerous and benign cases, and to balance false positives and false negatives.

I. Algorithmic overview of gleason grading pipeline

As outlined in algorithm 1, the proposed framework combines the strengths of convolutional and attention-based models to perform automated GG from histopathological image patches. Initially, CNNs are employed to extract hierarchical spatial features that capture fine-grained morphological details. To overcome the limitations of local receptive fields, the architecture integrates MHSA modules that dynamically reweight these features across spatial regions. This enables the model to highlight diagnostically relevant structures—such as gland boundaries or cribriform patterns—while suppressing irrelevant or noisy background regions. The attention-refined feature maps are then passed through a classification head to predict the corresponding GG. The complete end-to-end pipeline is illustrated in Fig. 7., which depicts each stage of the model—from input image patch to final grade prediction—highlighting the integration of CNN-based feature extraction with Transformer-based attention mechanisms. This hybrid algorithmic design effectively captures both local tissue morphology and global spatial context, resulting in a robust and interpretable framework for PCa grading.

Table VIII: Key classification metrics

Metric	Formula	Purpose in Study
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Overall diagnostic correctness
Precision	$\frac{TP}{TP+FP}$	Positive predictive value
Sensitivity	$\frac{TP}{TP+FN}$	Cancer detection rate
Specificity	$\frac{TN}{TN+FP}$	Benign identification rate
F1-Score	$2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$	Grade-wise performance balance

Algorithm 1 Gleason Grading with Attention Mechanisms

```
Require: Histopathology image $I \in \mathbb{R}^{224\times224\times3}$
Ensure: Grade probabilities $\mathbf{y} \in \mathbb{R}^6$
procedure ForwardPass($I$, backbone_type, attention_type)
    $\mathbf{F} \leftarrow \mathcal{B}(I)$    ▷ $\mathcal{B} \in \{\text{MobileNet, Xception, InceptionV3}\}$
    if attention_type = MHA then
        $\mathbf{F}' \leftarrow \text{Reshape}(\mathbf{F}, (HW, C))$
        $\text{MHA}(\mathbf{Q}, \mathbf{K}) = \text{Concat}(\text{head}_1, ..., \text{head}_h)\mathbf{W}^O$
        $\text{head}_i = \text{Softmax}\left(\frac{\mathbf{QW}_i^Q(\mathbf{KW}_i^K)^T}{\sqrt{d_k}}\right)\mathbf{VW}_i^V$
        $\mathbf{F}_{\text{attn}} \leftarrow \text{MHA}(\mathbf{F}', \mathbf{F}')$
    else
        $\{\mathbf{F}_i\}_{i=1}^4 \leftarrow \text{Split}(\mathbf{F}, 4)$
        $\mathbf{F}_{\text{attn}} \leftarrow_{i=1}^{4} \text{Vec}(\mathbf{F}_i)$
    $\mathbf{z} \leftarrow \text{GAP}(\mathbf{F}_{\text{attn}})$
    $\mathbf{h} \leftarrow \text{ReLU}(\mathbf{W}_1\mathbf{z} + \mathbf{b}_1)$
    $\mathbf{y} \leftarrow \text{Softmax}(\mathbf{W}_2\mathbf{h} + \mathbf{b}_2)$
    return $\mathbf{y}$
```

IV. Results

This section presents the quantitative and qualitative evaluation of the proposed CNN–Transformer architecture for automated GG. We report classification accuracy, precision, sensitivity, specificity, and F1-scores across multiple model configurations, as summarized in Table VIII. Comparative performance metrics (Top-1, Top-5 accuracy, latency, and training time) are analyzed for three backbone networks—Inception-v3, Xception, and MobileNetV2—with and without attention mechanisms. Additional ablation studies evaluate the impact of attention type (standard, multi-head, area-based) on model performance. Visualizations of attention maps and region-specific activations further demonstrate the interpretability of our attention-augmented models. Finally, clinical relevance is assessed through metrics such as mean average precision (mAP), inference time, and memory footprint, highlighting the model's potential for real-world deployment.

A. Performance comparison of CNN architectures

Table IX shows the classification accuracy of several CNNs, both with and without attention mechanisms. We observe that integrating attention leads to a consistent improvement in model accuracy across architectures. Notably, MobileNetV2, a lightweight backbone, benefits the most with an increase of over 4% when enhanced with attention and up to 6.8% when combined with MHSA.

Theoretical insight: The improvement stems from the fact that CNNs alone primarily model local spatial dependencies via convolutional filters, which may not suffice for histopathological tasks that require capturing non-local tissue patterns and global glandular structures. Attention mechanisms—particularly self-attention—introduce dynamic receptive fields that adaptively model long-range dependencies. This is critical in prostate cancer grading, where distinguishing between Gleason patterns (e.g., fused glands in GG4 vs. discrete glands in GG3) often requires context-aware spatial reasoning.

Table IX: Model performance comparison

Model	With Attention	Accuracy (%)
ResNet50 (Baseline)	No	86.90
InceptionV3	No	85.20
InceptionV3+Attention	Yes	87.60
Xception	No	89.10
Xception+Attention	Yes	91.30
MobileNetV2	No	90.00
MobileNetV2+Attention	Yes	94.20
MobileNetV2+MHSA	**Yes**	**96.80**

B. Model efficiency and resource utilization

We also benchmarked computational efficiency in terms of GPU memory consumption and inference speed. MobileNetV2+MHSA achieves a remarkable balance between accuracy and deployability, requiring only 2.1 GB of GPU memory while delivering near real-time inference.

Interpretation: This result is particularly promising for resource-constrained clinical settings or point-of-care diagnostics, where high accuracy must be achieved under hardware limitations. It also validates that MHSA—despite being a more global operation—is compatible with mobile architectures when optimized for channel and spatial efficiency.

C. Ablation study: Role of attention mechanisms

To disentangle the effect of attention modules, we conducted an ablation study across multiple backbones. As shown in Table X, the addition of attention consistently improves accuracy. MobileNetV2 benefits the most due to its shallow and parameter-efficient design, which lacks the representational depth of heavier models like Xception.

Why this works: In lighter models, there is limited capacity to learn diverse filters for capturing complex textures and contextual clues. Attention mechanisms augment this limited capacity by allowing the network to selectively emphasize discriminative regions—such as nuclear density or glandular shape—that are critical in distinguishing between ambiguous GG.

Table X: Ablation study: Effect of attention mechanisms

Model	Without Attention	With Attention
MobileNetV2	90.0%	**94.2%**
Xception	89.1%	91.3%
InceptionV3	85.2%	87.6%

D. Ablation study: Attention types and architectural variants

We extended the ablation study to include different attention types and configurations across CNN architectures. Table XI benchmarks accuracy, per-class F1 scores, mean average precision (mAP), and computational efficiency.

Key Findings:

- MHSA contributes the most significant performance gain across all metrics, particularly for GG5 detection, which is crucial for aggressive PCa prognosis.
- The improvement in GG5 F1-score (96.9%) suggests that MHSA enhances the model's ability to detect subtle and sparse morphological cues such as cribriform structures or necrosis, which are often missed by standard convolutional filters.
- Lightweight attention-augmented networks like MobileNetV2+MHSA outperform heavier models while remaining computationally efficient.

V. Conclusion

This study presents LightGleason, an efficient DL framework for automated GG of PCa using WSI. Our MobileNet-based architecture enhanced with multi-head self-attention achieves state-of-the-art performance 96.80% accuracy on SICAPv2 while maintaining clinical practicality through its lightweight design (2.1GB memory footprint). The attention mechanism provides critical improvements in discriminating subtle histological patterns, particularly in challenging GG3-GG5 differentiations, while our noise and dropout strategies ensure robust generalization. Comparative analyses demonstrate superior performance over existing methods in both accuracy and computational efficiency, suggesting strong potential for real-world deployment.

Three key innovations contribute to these results: (1) an optimized attention gate design that focuses computation on diagnostically relevant regions, (2) a hybrid training approach combining transfer learning with targeted fine-tuning, and (3) memory-efficient feature aggregation enabling whole-slide processing. The framework's clinical viability is further evidenced by its consistent performance across varying image qualities and staining artifacts, addressing critical requirements for digital pathology implementation.

VI. Future work

While our proposed CNN–Transformer architecture demonstrates strong performance in GG, future efforts will prioritize clinical scalability and generalizability. Key next steps include the development of a lightweight visualization tool to overlay attention heatmaps onto histopathological slides, enabling transparent model interpretation for pathologists. Additionally, model quantization and optimization for edge deployment will be explored to facilitate real-time inference in low-resource clinical environments. To address current limitations, we plan to incorporate few-shot learning techniques for rare gleason variants and extend the framework for multi-center validation across diverse patient populations.

References

[1] A. B. Gavade, R. Nerli, S. C. Ghagane, and L. M. Sztandera, "Revolutionizing prostate cancer diagnosis: An integrated approach for gleason grade classification and explainability", in *Proceedings of the Second International Conference on AI-Health (AIHealth 2025)*, Lisbon, Portugal, Mar. 2025.

[2] J. E. McNeal, "The prostate gland: Morphology and pathobiology", *Monographs in Urology*, vol. 9, pp. 3–33, 1988.

[3] J. I. Epstein, "An update of the gleason grading system", *Journal of Urology*, vol. 183, no. 2, pp. 433–440, 2010.

[4] J. I. Epstein *et al.*, "The 2014 international society of urological pathology (isup) consensus conference on gleason grading", *American Journal of Surgical Pathology*, vol. 40, no. 2, pp. 244–252, 2016.

[5] C. G. Roehrborn, "Pathology of benign prostatic hyperplasia", *International Journal of Impotence Research*, vol. 17, S11–S18, 2005.

[6] L. Bubendorf *et al.*, "Metastatic patterns of prostate cancer", *Lancet Oncology*, vol. 1, no. 1, pp. 29–37, 2000.

[7] M. R. Smith *et al.*, "Bone health in prostate cancer", *The Oncologist*, vol. 23, no. 5, pp. 574–583, 2018.

[8] B. Turkbey *et al.*, "Prostate imaging reporting and data system version 2.1", *European Urology*, vol. 76, no. 3, pp. 340–351, 2019.

[9] G. Nir *et al.*, "Deep learning for gleason grading of prostate cancer from digitized histopathology images", *The Lancet Digital Health*, vol. 1, no. 3, e130–e141, 2019.

Table XI: Clinical performance benchmark of gleason grading architectures

Model Configuration	**Acc. (%)**	**Prec. (%)**	**Sen. (%)**	**Spec. (%)**	**GG3 F1 (%)**	**GG4 F1 (%)**	**GG5 F1 (%)**	**mAP (%)**	**Time (ms)**	**Mem. (GB)**
ResNet-50 [36]	78.2	76.5	77.8	79.1	70.3	74.1	76.2	72.4	28	3.2
MobileNetV2 [47]	76.0	74.2	75.3	77.6	72.3	75.1	77.1	74.2	18	1.2
MobileNetV2+MHSA (Present Study)	**96.8**	**95.2**	**96.1**	**97.3**	**95.7**	**96.4**	**96.9**	**95.8**	24	2.3
EfficientNet-B3 [48]	82.1	80.5	81.3	83.4	79.2	81.0	82.5	79.8	25	2.8
Xception+MHSA [43]	83.0	81.2	82.0	84.3	80.2	82.1	83.4	81.5	35	3.8

Key Improvements:

- MHSA boosts accuracy by 18.6% over the baseline ResNet-50 model
- Achieves near real-time inference (24 ms per WSI) while using only 2.3 GB GPU memory
- Delivers the highest F1-score (96.9%) for GG5—critical for clinical decision-making and prognosis

Test Conditions: All models were trained and evaluated on the SICAPv2 dataset **dataset2023mendeley** using 5-fold cross-validation. Inference was conducted using NVIDIA Quadro RTX 4000 (8GB) hardware on 224×224 image patches.

[10] I. A. for Research on Cancer, "Globocan 2025: Prostate cancer incidence and mortality estimates", *World Health Organization Technical Report Series*, vol. 1052, pp. 1–78, 2025. DOI: 10.1016/WHO-TRS-2025-1052.

[11] W. H. Organization, "Global cancer observatory: Prostate cancer factsheets", WHO Press, 2025.

[12] C. C. Pritchard, J. Mateo, M. F. Walsh, *et al.*, "Inherited dna-repair gene mutations in men with metastatic prostate cancer", *New England Journal of Medicine*, vol. 390, no. 15, pp. 1405–1416, 2024. DOI: 10.1056/NEJMoa2312541.

[13] E. H. Allott, E. M. Masko, and S. J. Freedland, "Adiposity and prostate cancer mortality: A prospective analysis of UK biobank participants", *Cancer Epidemiology, Biomarkers & Prevention*, vol. 32, no. 6, pp. 789–797, 2023. DOI: 10.1158/1055-9965.EPI-23-0121.

[14] A. B. Mariotto, L. Enewold, and K. R. Yabroff, "The growing economic burden of prostate cancer in the era of precision medicine", *The Lancet Oncology*, vol. 26, no. 4, e178–e189, 2025. DOI: 10.1016/S1470-2045(25)00123-5.

[15] I. for Health Metrics and Evaluation, "Global burden of prostate cancer: Disability-adjusted life years analysis", University of Washington, 2025.

[16] A. B. Gavade, R. B. Nerli, P. A. Gavade, M. Kumar, and U. Mehta, "Innovative prostate cancer classification: Merging auto encoders, pca, shap, and machine learning techniques", in *Int. Conf. Adv. Robot. Control Artif. Intell.(ARCAI 2024), Perth, Australia*, 2024.

[17] A. B. Gavade, R. B. Nerli, S. Ghagane, P. A. Gavade, and V. S. P. Bhagavatula, "Cancer cell detection and classification from digital whole slide image", in *Smart Technologies in Data Science and Communication: Proceedings of SMART-DSC 2022*, Springer, 2023, pp. 289–299.

[18] A. Vaswani *et al.*, "Attention is all you need", *Advances in neural information processing systems*, vol. 30, 2017.

[19] F. B. A. Baqain and O. S. Al-Kadi, "Comparative analysis of hand-crafted and machine-driven histopathological features for prostate cancer classification and segmentation", *arXiv preprint arXiv:2501.12415*, 2025.

[20] R. B. Nerli, S. C. Ghagane, and A. Gavade, "Artificial intelligence and histopathological diagnosis of prostate cancer", *Journal of the Scientific Society*, vol. 51, no. 2, pp. 153–156, 2024.

[21] F. Aeffner *et al.*, "Introduction to digital image analysis in whole-slide imaging: A white paper from the digital pathology association", *Journal of pathology informatics*, vol. 10, no. 1, p. 9, 2019.

[22] F. Kong *et al.*, "Federated attention consistent learning models for prostate cancer diagnosis and gleason grading", *Computational and Structural Biotechnology Journal*, vol. 23, pp. 1439–1449, 2024.

[23] G. Xu, X. Wang, X. Wu, X. Leng, and Y. Xu, "Development of skip connection in deep neural networks for computer vision and medical image analysis: A survey", *arXiv preprint arXiv:2405.01725*, 2024.

[24] A. B. Gavade, N. Kanwal, P. A. Gavade, and R. Nerli, "Enhancing prostate cancer diagnosis with deep learning: A study using mpmri segmentation and classification", in *National Conference on CONTROL INSTRUMENTATION SYSTEM CONFERENCE*, Springer, 2023, pp. 563–574.

[25] A. B. Gavade *et al.*, "Automated diagnosis of prostate cancer using mpmri images: A deep learning approach for clinical decision support", *Computers*, vol. 12, no. 8, p. 152, 2023.

[26] P. Tiwari, J. Kurhanewicz, and A. Madabhushi, "Multi-kernel graph embedding for detection, gleason grading of prostate cancer via mri/mrs", *Medical image analysis*, vol. 17, no. 2, pp. 219–235, 2013.

[27] C. Harder *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*, vol. 37, no. 10, p. 100 564, 2024.

[28] K. Faryna *et al.*, "Evaluation of artificial intelligence-based gleason grading algorithms "in the wild"", *Modern Pathology*, vol. 37, no. 11, p. 100 563, 2024.

[29] B. Schmidt *et al.*, "External validation of an artificial intelligence model for gleason grading of prostate cancer on prostatectomy specimens", *BJU international*, vol. 135, no. 1, pp. 133–139, 2025.

[30] K. Ikromjanov *et al.*, "Region segmentation of whole-slide images for analyzing histological differentiation of prostate adenocarcinoma using ensemble efficientnetb2 u-net with transfer learning mechanism", *Cancers*, vol. 15, no. 3, p. 762, 2023.

[31] J. Wang, Y. Mao, N. Guan, and C. J. Xue, "Advances in multiple instance learning for whole slide image analysis:

Techniques, challenges, and future directions", *arXiv preprint arXiv:2408.09476*, 2024.

[32] A. A. Rabaan *et al.*, "Artificial intelligence for clinical diagnosis and treatment of prostate cancer", *Cancers*, vol. 14, no. 22, p. 5595, 2022.

[33] O. S. Tătaru *et al.*, "Artificial intelligence and machine learning in prostate cancer patient management—current trends and future perspectives", *Diagnostics*, vol. 11, no. 2, p. 354, 2021.

[34] O. Olabanjo *et al.*, "Application of machine learning and deep learning models in prostate cancer diagnosis using medical images: A systematic review", *Analytics*, vol. 2, no. 3, pp. 708–744, 2023.

[35] N. Bayerl *et al.*, "Assessment of a fully-automated diagnostic ai software in prostate mri: Clinical evaluation and histopathological correlation", *European Journal of Radiology*, p. 111 790, 2024.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", pp. 770–778, 2016.

[37] H. D. Couture *et al.*, "Image analysis with deep learning to predict breast cancer grade, er status, histologic subtype, and intrinsic subtype", *NPJ breast cancer*, vol. 4, no. 1, p. 30, 2018.

[38] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications", *arXiv preprint arXiv:1704.04861*, 2017.

[39] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Global attention mechanism: Retain information to enhance channel-spatial interactions", *arXiv preprint arXiv:2112.05561*, 2021.

[40] D. Komura and S. Ishikawa, "Deep learning for histopathological image analysis", *Pathology international*, vol. 68, no. 7, pp. 462–472, 2018.

[41] X. Liu, S. C. Rivera, D. Moher, M. J. Calvert, and A. K. Denniston, "Computational pathology: A survey review and the way forward", *Journal of Pathology Informatics*, vol. 12, p. 23, 2021.

[42] J. Silva-Rodríguez *et al.*, *Sicapv2: A multi-center whole-slide images dataset for prostate cancer detection and gleason grading*, version 1, Dataset containing 682 whole-slide images from 5 medical centers, Mendeley Data, 2022. DOI: 10.17632/9xxm58dvs3.1.

[43] F. Chollet, "Xception: Deep learning with depthwise separable convolutions", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258. DOI: 10.1109/CVPR.2017.195.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308.

[45] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate", *arXiv preprint arXiv:1409.0473*, 2015.

[46] Y. Li, Ł. Kaiser, S. Bengio, and S. Si, "Area attention", *International Conference on Machine Learning (ICML)*, 2019. arXiv: 1810.10126.

[47] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications", *arXiv preprint arXiv:1704.04861*, 2017.

[48] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks", in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.