Revolutionizing Prostate Cancer Diagnosis: An Integrated Approach for Gleason Grade Classification and Explainability

Anil B. Gavade
Dept., of Electronics and Communication Engineering
KLS Gogte Institute of Technology, Belagavi-590008, India
Email: abgavade@git.edu

Rajendra B. Nerli
Dept., of Urology, JN Medical College, KLE Academy of Higher Education and Research, Belagavi 590010, India.
Email: rajendranerli@yahoo.in

Shridhar C. Ghagane
KAHER'S Dr. Prabhakar Kore Basic Science Research Center, JNMC Campus, Belagavi-590010, India
Email: shridhar.kleskf@gmail.com

Les Sztandera*
Dept., Computer Information Systems, Thomas Jefferson University, Philadelphia, PA 19107, USA
Email: Les.Sztandera@jefferson.edu

Abstract—**Accurate grading of Prostate Cancer (PCa) is vital for effective treatment planning and prognosis. This study introduces an advanced framework for Gleason Grade (GG) classification, addressing challenges in accuracy, computational efficiency, and interpretability. Utilizing the SICAPv2 dataset, which contains annotated prostate biopsy Whole Slide Images (WSIs) graded from GG0 to GG5, the framework integrates cutting-edge machine learning and deep learning techniques. Feature extraction is performed using a custom-designed Variational Autoencoder (VAE) with a VGG16 backbone, chosen for its computational efficiency, while dimensionality reduction with Principal Component Analysis (PCA) optimally selects 50 features for classification. The classification pipeline combines machine learning models, including Support Vector Machines (SVM), logistic regression, and random forests, with custom Deep Neural Networks (DNNs). SVM with an Radial Basis Function (RBF) kernel achieved an accuracy of 84% following hyperparameter tuning, while a custom five-layer dense neural network incorporating dropout and batch normalization demonstrated superior performance with an accuracy of 94.6%. Explainable AI (XAI) techniques, such as SHapley Additive exPlanations (SHAP), gradient-weighted class activation mapping (Grad-CAM), and Local Interpretable Model-Agnostic Explanations (LIME), enhance model interpretability by providing insights into feature importance and aligning predictions with clinical expertise. This framework delivers a robust, scalable, and interpretable solution for automated GG classification, bridging the gap between advanced AI techniques and clinical application.**

Keywords- Cancer diagnosis; Dimensionality reduction; Explainable AI; Feature extraction; Gleason grade classification.

I. INTRODUCTION

Prostate cancer remains a significant global health issue, ranking among the leading causes of cancer-related mortality in men. The prostate gland [6], located below the bladder and comparable in size to a walnut, plays a crucial role in male reproductive health by producing seminal fluid. Clinical manifestations often include Lower Urinary Tract Symptoms (LUTS), haematuria, erectile dysfunction, and urinary retention.

Traditional diagnostic methods, such as Digital Rectal Examination (DRE), prostate-specific antigen (PSA) screening, and 12-core Transrectal Ultrasound (TRUS)-guided biopsy, exhibit notable limitations. Over-diagnosis rates can reach up to 45%, while clinically significant cancers may be missed in 30% of cases. These challenges underscore the necessity for advanced diagnostic techniques to effectively distinguish aggressive from non-aggressive cancer types [1-4].

The integration of WSI with Artificial Intelligence (AI) presents transformative potential in prostate cancer diagnostics, particularly for GG. High-resolution digital images of prostate biopsy samples are acquired through WSI scanners and undergo preprocessing steps, such as normalization and artifact removal, to enhance image quality [8]. AI-driven models segment tissue regions and extract significant histopathological features using deep learning techniques, including Convolutional Neural Networks (CNNs) and VAE [5].

Following feature extraction, AI models classify tissue patterns into respective GG, facilitating precise cancer grading. Post-classification validation ensures model robustness, while explainability tools such as SHAP, LIME, Grad-CAM, and Saliency Maps enhance transparency and interpretability. Figure 1 illustrates the developed AI pipeline, addressing critical diagnostic challenges by improving accuracy, efficiency, and consistency in GG assessment.

This framework integrates VAEs, XAI techniques, and preprocessing methods to enhance GG classification precision, support personalized clinical decisions, and improve PCa outcomes. In this paper, Section II covers the related work, Section III delves into the methods and materials, Section IV presents the results and discussions, and Section V provides the conclusion.

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-247-0

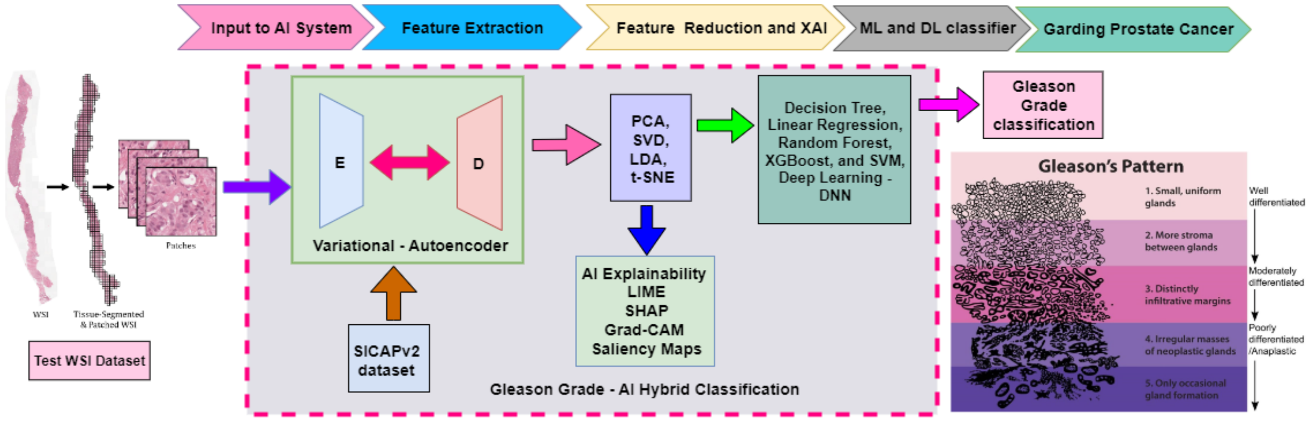

Input to AI System
Feature Extraction
Feature Reduction and XAI
ML and DL classifier
Garding Prostate Cancer
WSI
Tissue-Segmented & Patched WSI
Patches
Test WSI Dataset
E
D
Variational - Autoencoder
SICAPv2 dataset
PCA, SVD, LDA, t-SNE
AI Explainability LIME SHAP Grad-CAM Saliency Maps
Decision Tree, Linear Regression, Random Forest, XGBoost, and SVM, Deep Learning - DNN
Gleason Grade - AI Hybrid Classification
Gleason Grade classification
Gleason's Pattern
1. Small, uniform glands
2. More stroma between glands
3. Distinctly infiltrative margins
4. Irregular masses of neoplastic glands
5. Only occasional gland formation
Well differentiated
Moderately differentiated
Poorly differentiated /Anaplastic

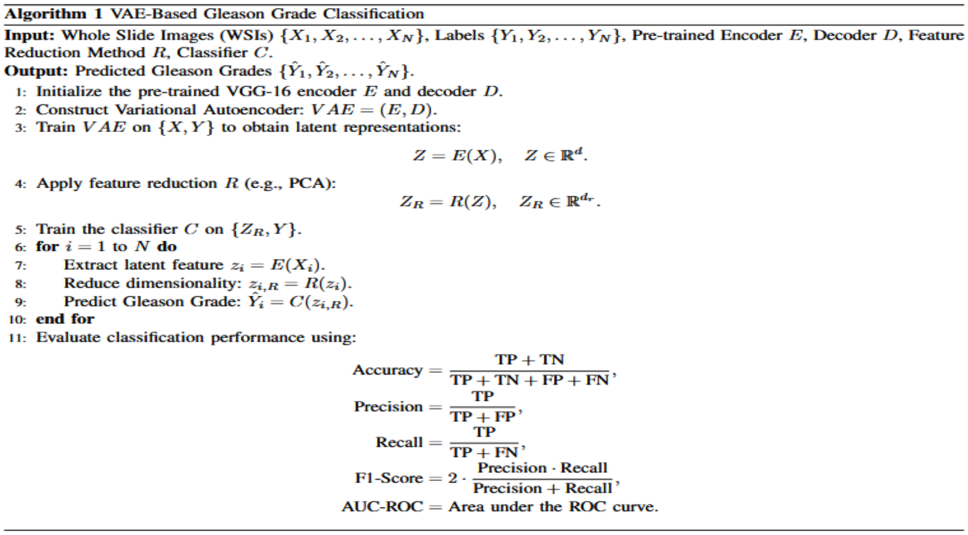
Figure 1. The block diagram shows a hybrid system for GG Group using WSIs, ensuring transparent and accurate PCa diagnosis and treatment with XAI techniques

II. Related Work

The integration of AI into the GG of PCa using WSI has brought about significant advancements in accuracy, consistency, and efficiency [9], [22]. Firjani et al. [10] laid the groundwork by achieving 100% accuracy in classifying prostate tissues into benign and malignant using a k-Nearest Neighbors (KNN) classifier on Diffusion-Weighted Imaging (DWI). Singhal et al. [11] improved segmentation and grading of PCa in WSIs of core needle biopsies with a DL model combining U-Net and Atrous Spatial Pyramid Pooling (ASPP) modules, achieving an accuracy of 89.4% and a quadratic-weighted kappa of 0.92. Azizi et al. [12] leveraged recurrent neural networks (RNN) on temporal enhanced ultrasound (TeUS) data, with Long Short-Term Memory (LSTM) networks achieving an accuracy of 0.93, an AUC of 0.96, a sensitivity of 0.76, and a specificity of 0.98. Bulten et al. [13] developed an automatic DL model for GG, attaining a quadratic Cohen's kappa score of 0.918 using biopsies. Tsuneki et al. [15] employed transfer learning to classify WSIs into adenocarcinoma and benign lesions, achieving a high ROC-AUC of up to 0.9873. Pati et al. [16] introduced WholeSIGHT, a weakly-supervised method for joint segmentation and classification, demonstrating a Dice coefficient of 0.76 on three public PCa WSI datasets. Müller et al. [17] presented DeepGleason, an open-source DNN system for automated GG, achieving a macro-averaged F1-score of 0.806, an AUC of 0.991, and an accuracy of 0.974. Hammouda et al. [18] proposed a multi-stage classification-based DL system for GG, achieving a precision of 0.92, recall of 0.89, and accuracy of 0.93 on 3,080 WSIs. Duenweg et al. [19] highlighted the impact of different WSI scanners on image quality, which significantly affects computational analysis performance, underscoring the need for standardized WSI scanner protocols. Mittmann et al. [20] developed an AI system for interpretable GG that mimics pathologist explanations, achieving a Dice score of 0.713 ± 0.003 using a dataset of 1,015 tissue microarray core images annotated by 54 pathologists. Belinga [11] proposed an AI-assisted system that improved GG accuracy and consistency, with a quadratically weighted Cohen's kappa of 0.872 compared to 0.799 without AI assistance, evaluated on 160 biopsies graded by 14 observers. Collectively, these studies underscore the transformative potential of AI and digital pathology in enhancing the diagnostic accuracy and consistency of GG in PCa.

III. Methods and Materials

Hybrid PCa GG uses a custom VAE with a pre-trained VGG-16 encoder for feature extraction and a two-layer Dense decoder for reconstruction. Trained on SICAPv2 datasets [5], [7], it ensures accurate GG classification and clinical relevance, as shown in Figure 2. To further optimize performance, we apply advanced feature reduction techniques, including PCA, Singular Value Decomposition (SVD), linear discriminant analysis (LDA), and t-distributed Stochastic Neighbor Embedding (t-SNE), ensuring dimensionality reduction while retaining critical data characteristics. The pipeline employs several state-of-the-art classifiers—Decision Tree, Random Forest, XGBoost, and SVM—which are fine-tuned via hyperparameter optimization to improve predictive accuracy. These classifiers are evaluated using performance metrics like accuracy, precision, recall, and F1-Score to ensure robust and reliable results. Furthermore, to enhance model transparency and interpretability, we incorporate XAI techniques. LIME offers local insights into individual predictions, SHAP quantifies global feature contributions, Grad-CAM visualizes critical regions in the images, saliency Maps highlight influential pixels, and feature Maps provide insights into the learning process at various layers. This comprehensive approach not only enhances the precision of GG but also supports transparency, ensuring the AI model is trustworthy for clinical use, and paves the way for more personalized PCa diagnosis and treatment strategies. The SICAPv2 dataset [22] (GG0–GG5) provided a robust benchmark for validating AI-driven PCa models.



```
Algorithm 1 VAE-Based Gleason Grade Classification
Input: Whole Slide Images (WSIs) $\{X_1, X_2, \ldots, X_N\}$, Labels $\{Y_1, Y_2, \ldots, Y_N\}$, Pre-trained Encoder $E$, Decoder $D$, Feature Reduction Method $R$, Classifier $C$.
Output: Predicted Gleason Grades $\{\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_N\}$.
Initialize the pre-trained VGG-16 encoder $E$ and decoder $D$.
Construct Variational Autoencoder: $VAE = (E, D)$.
Train $VAE$ on $\{X, Y\}$ to obtain latent representations:
    $Z = E(X), \quad Z \in \mathbb{R}^d.$
Apply feature reduction $R$ (e.g., PCA):
    $Z_R = R(Z), \quad Z_R \in \mathbb{R}^{d_r}.$
Train the classifier $C$ on $\{Z_R, Y\}$.
for $i = 1$ to $N$ do
    Extract latent feature $z_i = E(X_i)$.
    Reduce dimensionality: $z_{i,R} = R(z_i)$.
    Predict Gleason Grade: $\hat{Y}_i = C(z_{i,R})$.
end for
Evaluate classification performance using:
    $\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$
    $\text{Precision} = \frac{TP}{TP + FP},$
    $\text{Recall} = \frac{TP}{TP + FN},$
    $\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$
    $\text{AUC-ROC} = \text{Area under the ROC curve.}$
```

Figure 2. Implemenation algorithm VAE-Based Hybrid Algorithm for Gleason Grade Classification

IV. Results and Discussion

This section presents the VAE-based hybrid pipeline's performance in GG classification, emphasizing key results, feature reduction, and XAI techniques.

A. *Feature Extraction by VAE and Feature reduction*

As shown in Table I, VGG-16 is the optimal feature extractor for GG classification, balancing quality, efficiency, and interpretability by extracting 512 compact and effective features. It avoids the redundancy seen in ResNet-50 and DenseNet-121, which produce 2048 and 1024 features, respectively. Despite DenseNet-121 being faster, VGG-16's moderate extraction time ensures reliability, minimizing overfitting and making it ideal for medical imaging.

TABLE I. FEATURE EXTRACTION

VAE with CNN as Encoder	**VAE Performance as Feature Extractor**				
	No. of features extracted from Model	***Feature Dimensions Before Flattening***	***Time taken by Model for FE***	***Time taken for feature Decoding***	***Time taken for PCA Transformation***
VGG-16	512	(None, 7, 7, 512)	63.06 sec	0.55 sec	0.05 sec
VGG-19	512	(None, 7, 7, 512)	65.50 sec	0.32 sec	0.01 sec
ResNet-50	2048	(None, 7, 7, 2048)	30.55 sec	0.28 sec	0.04 sec
DenseNet-121	1024	(None, 7, 7, 1024)	33.84 sec	0.32 sec	0.07 sec

Feature reduction performance in Table II indicates that various CNN models used as VAE encoders achieve similar dimensionality reduction to 50 features using PCA, SVD, and t-SNE. VGG-16, VGG-19, and ResNet-50 demonstrate comparable performance in feature reduction, with ResNet-50 extracting the highest number of features at 2048. DenseNet-121, extracting 1024 features, achieves the highest reduction with SVD, reducing features to 137. VGG-16 and VGG-19, with 512 features, consistently and efficiently reduce dimensionality while maintaining feature quality.

TABLE II. FEATURE REDUCTION

CNN Model As VAE Encoder	**Feature Reduction after VAE**				
	No. of features extracted from Model	***PCA***	***SVD***	***LDA***	***t-SNE***
VGG-16	512	50	100	1	2
VGG-19	512	50	93	1	2
ResNet-50	2048	50	103	1	2
DenseNet-121	1024	50	137	1	2

B. *Feature Explainability*

In Table IV, SHAP was the fastest, completing its task in 0.92 seconds while using only 1.47 MB of memory. LIME, although computationally intensive, required the highest memory at 9.97 MB. Grad-CAM stood out for its superior visual explanations, achieving a good balance with a runtime of 1.87 seconds and memory usage of 5.16 MB. Saliency maps provided a well-rounded performance, combining reasonable speed at 1.15 seconds with moderate memory usage of 6.74 MB. Figure 3 illustrates explainability techniques: (a) Significant contributions of 50 features to classification using XAI SHAP, (b) Grad-CAM heatmap for GG2 showing lower activation in cooler colors, indicating a lower likelihood of malignancy, and (c) Grad-CAM heatmap for GG4 displaying higher activation in warmer colors, highlighting regions significant for predicting malignancy.

TABLE III. EXPLAINABILITY OF FEATURE

XAI Technique	***Time (seconds)***	***Peak Memory Usage (MB)***
SHAP	0.9193	1.4740
LIME	1.4421	9.9731
Grad-CAM	1.8705	5.1574
Saliency Map	1.1513	6.7392

C. *Machine Learnning classification*

In Table IV, the performance metrics for various machine learning classification models are as follows: Decision Tree achieved accuracy, precision, recall, and F1-score of 0.47. Linear Regression showed consistent scores of 0.70 across all metrics. Random Forest performed better with scores of 0.78 across all metrics. XGBoost had moderate performance with scores of 0.72. SVM demonstrated the highest performance with accuracy and recall at 0.81, and precision and F1-score at 0.80.

TABLE IV. PERFORMANCE METRICS FOR VARIOUS ML CLASSIFICATION

Metric	**ML Model**				
	Decision Tree	***Linear Regression***	***Random Forest***	***XGBoost***	***SVM***
Accuracy	0.47	0.70	0.78	0.72	0.81
Precision	0.48	0.71	0.78	0.72	0.80
Recall	0.47	0.70	0.78	0.72	0.81
F1-Score	0.47	0.70	0.78	0.72	0.81

D. *Machine Learnning classification with hyperparmeter tuning*

In Table V, hyperparameter tuning significantly improved model performance. SVM showed the highest gains, with accuracy rising from 0.81 to 0.84, precision from 0.80 to 0.85, recall from 0.81 to 0.84, and F1-score from 0.81 to 0.84—improvements of 3.7% in accuracy and recall, and 3.8% in precision and F1-score. Random Forest improved from 0.78 to 0.81 in accuracy and recall, while XGBoost's accuracy rose from 0.72 to 0.76 and recall from 0.72 to 0.75. Decision Tree saw the largest improvement, with an 8.5% boost across all metrics (0.47 to 0.51). Linear Regression maintained consistent performance at 0.72. Overall, hyperparameter tuning enhanced all models, with SVM outperforming others in all metrics.

TABLE V. PERFORMANCE METRICS FOR VARIOUS ML CLASSIFICATION WITH HYPERPARMETER TUNING

Metric	**ML Model hyperparameter**				
	Decision Tree	***Linear Regression***	***Random Forest***	***XGBoost***	***SVM***
Accuracy	0.51	0.72	0.81	0.76	0.84
Precision	0.53	0.72	0.80	0.74	0.85
Recall	0.51	0.72	0.81	0.75	0.84
F1-Score	0.51	0.72	0.81	0.75	0.84

E. *Deep learning – Deep neural network*

In Table VI, we evaluated the performance of DNN models with different architectures. The DNN model with 3 Dense Layers achieved an accuracy of 0.79, precision of 0.81, recall of 0.79, and F1-score of 0.80. The performance improved with the DNN model featuring 5 Dense Layers, which achieved an accuracy of 0.89, precision of 0.90, recall of 0.89, and F1-score of 0.89. The model with 5 Dense Layers combined with Dropout and Batch Normalization demonstrated the best results, with an accuracy of 0.94, precision of 0.96, recall of 0.94, and F1-score of 0.95. The addition of Dropout and Batch Normalization led to a significant improvement, increasing accuracy by 5.6%, precision by 6.7%, recall by 5.6%, and F1-score by 6.7% compared to the model without these techniques. This indicates that Dropout and Batch Normalization played a crucial role in boosting the model's overall performance.

TABLE VI. DEEP LEARNIG – DNN MODEL

Metric	**DL – DNN with 3 Dense Layers**	**DL – DNN with 5 Dense Layers**	**DL – DNN with 5 Dense Layers + Dropout & Batch Normalization**
Accuracy	0.79	0.89	**94.6**
Precision	0.81	0.90	**96**
Recall	0.79	0.89	**94**
F1-Score	0.80	0.89	**0.95**

F. *Comparative Analysis of ML and DL Techniques for Prostate Cancer Diagnostics*

The performance comparison shows ML models like SVM achieving ~0.84 accuracy, while DL models excelled, with a 5-layer DNN using dropout and batch normalization achieving ~0.94; Figure 4 highlights that DL, combined with advanced regularization methods, offers superior accuracy and robustness in PCa GG classification.

V. Conclusion

In this study, a comprehensive framework for GG classification using WSI of PCa was developed by integrating DNN and ML models. VGG-16 was identified as the optimal feature extractor, offering a balance of feature quality and computational efficiency by extracting 512 features in 63.06 seconds. It outperformed DenseNet-121 and ResNet-50 in reducing redundancy and ensuring efficient dimensionality reduction through PCA, SVD, and t-SNE. Hyperparameter tuning enhanced ML performance, with SVM achieving the highest accuracy of 84%, while DL models incorporating dropout and batch normalization demonstrated significant improvements. A five-layer DNN achieved 94.6% accuracy, highlighting the effectiveness of regularization in preventing overfitting. A novel aspect of this research lies in the integration of XAI techniques to improve model interpretability. SHAP provided rapid, memory-efficient insights, while Grad-CAM delivered detailed visualizations, ensuring transparency in decision-making. LIME and Saliency Maps further contributed to understanding model outputs, underscoring the need for transparent AI in clinical settings. Future work will expand this framework to larger datasets and explore advanced neural architectures and XAI methods, aiming to develop scalable, interpretable, and clinically reliable AI models for PCa diagnostics. The implementation, tested on an open-access dataset, could benefit from additional testing on more benchmark and clinical datasets to enhance its clinical utility.

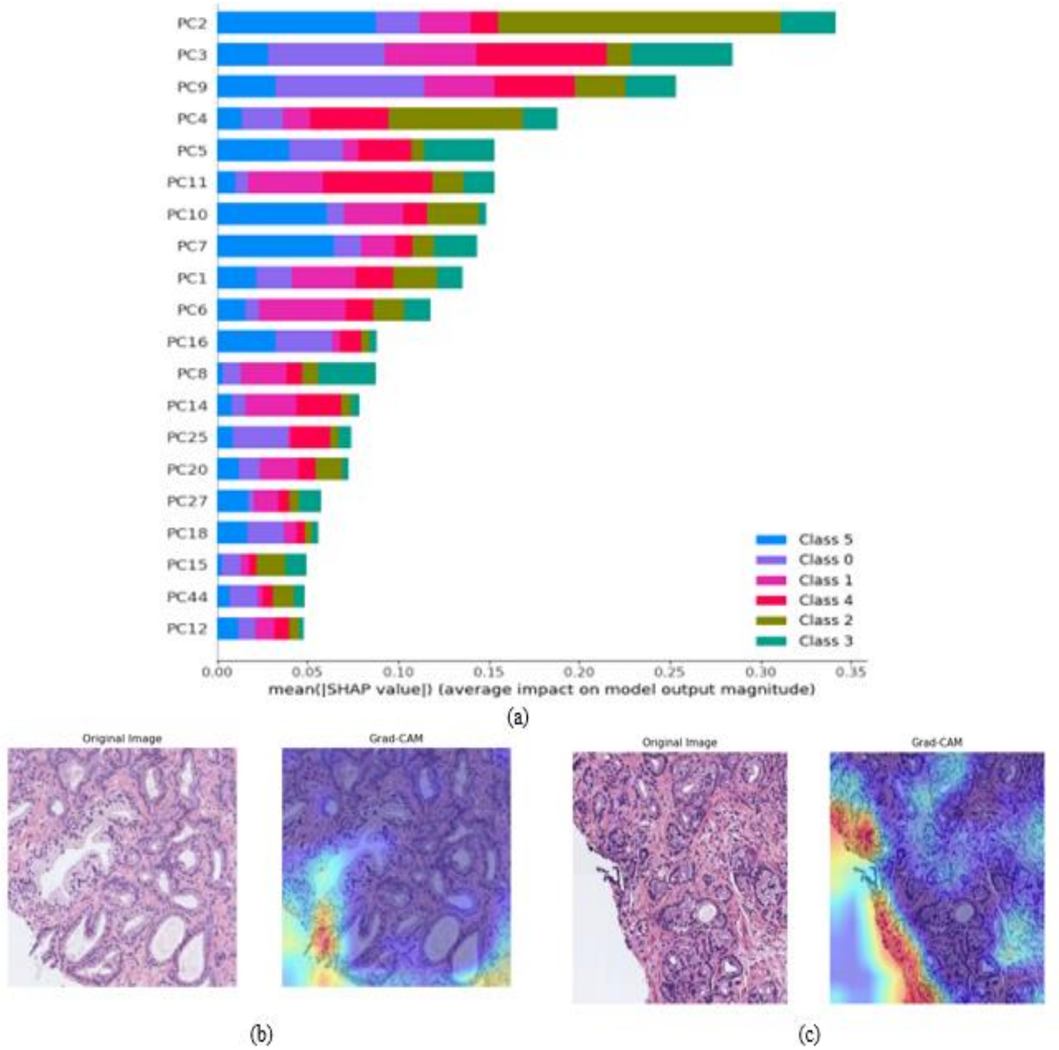


Figure 3. Comparison of Explainability Techniques for Prostate Cancer Gleason Grade Classification (a) Significant contributions of 50 features to classification using XAI SHAP (b) Grad-CAM heatmap for GG2 and (c) Grad-CAM heat map for GG4

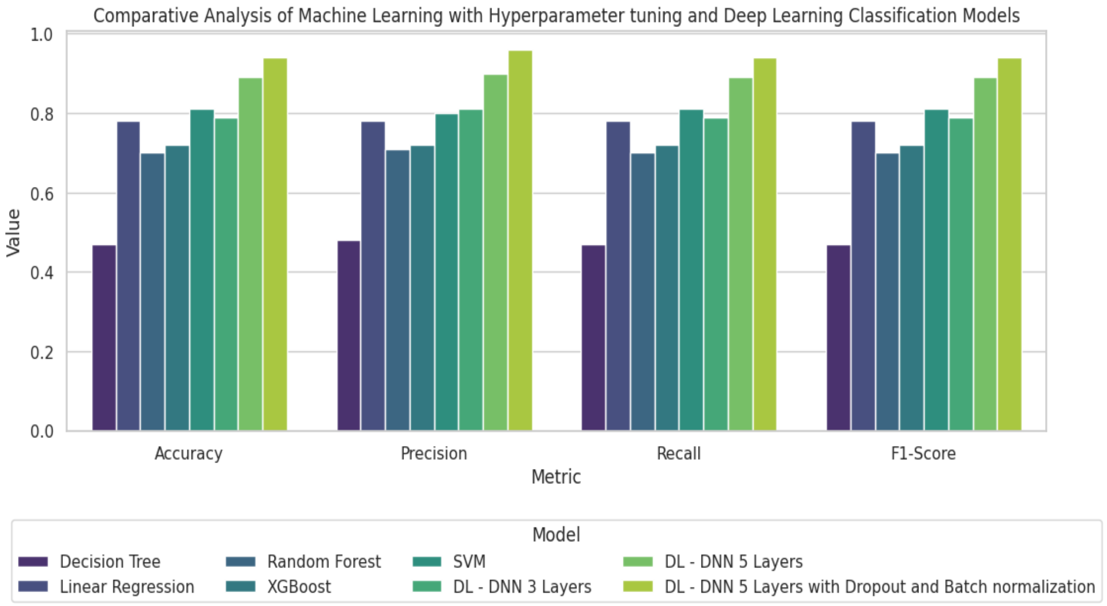


Figure 4. Evaluating ML and DL Models for Prostate Cancer Diagnostics: A Performance Insight

References

[1] American Cancer Society, "Cancer Facts & Statistics," accessed Jan. 15, 2025. https://www.cancer.org/research/cancer-facts-statistics.html.

[2] American Cancer Society, "Cancer Statistics Center," accessed Jan. 15, 2025. https://cancerstatisticscenter.cancer.org/#/.

[3] PathologyOutlines.com, "Prostate WHO Classification," accessed Jan. 15, 2025. https://www.pathologyoutlines.com/topic/prostateWHO.html.

[4] J. G. Kench et al., "WHO Classification of Tumours fifth edition: evolving issues in the classification, diagnosis, and prognostication of prostate cancer," *Histopathology*, vol. 81, no. 4, pp. 447-458, 2022.

[5] A. B. Gavade et al., "Innovative Prostate Cancer Classification: Merging Auto Encoders, PCA, SHAP, and Machine Learning Techniques," presented at *Int. Conf. Adv. Robot. Control Artif. Intell. (ARCAI 2024)*, Perth, Australia, Dec. 9–12, 2024. (unpublished).

[6] 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.

[7] K. A. Gadad et al., "Beyond Single Models: Hybrid Approaches for Multiclass Cancer Identification," in *2024 3rd Int. Conf. Adv. Technol. (ICONAT)*, pp. 1-6, IEEE, 2024.

[8] R. B. Nerli et al., "Artificial Intelligence and Histopathological Diagnosis of Prostate Cancer," *J. Sci. Soc.*, vol. 51, no. 2, pp. 153-156, 2024.

[9] A. S. Balraj et al., "PRADclass: Hybrid Gleason Grade-Informed Computational Strategy Identifies Consensus Biomarker Features Predictive of Aggressive Prostate Adenocarcinoma," *Technol. Cancer Res. Treat.*, vol. 23, p. 15330338231222389, 2024.

[10] A. Firjani et al., "A diffusion-weighted imaging based diagnostic system for early detection of prostate cancer," *J. Biomed. Sci. Eng.*, vol. 6, no. 3, pp. 346, 2013.

[11] N. Singhal et al., "A deep learning system for prostate cancer diagnosis and grading in whole slide images of core needle biopsies," *Sci. Rep.*, vol. 12, no. 1, p. 1-11, 2022.

[12] S. Azizi et al., "Deep recurrent neural networks for prostate cancer detection: analysis of temporal enhanced ultrasound," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2695-2703, 2018.

[13] W. Bulten et al., "Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study," *Lancet Oncol.*, vol. 21, no. 2, pp. 233-241, 2020.

[14] M. Tsuneki, M. Abe, and F. Kanavati, "A deep learning model for prostate adenocarcinoma classification in needle biopsy whole-slide images using transfer learning," *Diagnostics*, vol. 12, no. 3, p. 768, 2022.

[15] P. Pati et al., "Weakly supervised joint whole-slide segmentation and classification in prostate cancer," *Med. Image Anal.*, vol. 89, p. 102915, 2023.

[16] D. Müller et al., "DeepGleason: a System for Automated Gleason Grading of Prostate Cancer using Deep Neural Networks," *arXiv preprint arXiv:2403.16678*, 2024.

[17] K. Hammouda et al., "Multi-Stage Classification-Based Deep Learning for Gleason System Grading Using Histopathological Images," *Cancers*, vol. 14, no. 23, p. 5897, 2022.

[18] S. R. Duenweg et al., "Whole slide imaging (WSI) scanner differences influence optical and computed properties of digitized prostate cancer histology," *J. Pathol. Inform.*, vol. 14, p. 100321, 2023.

[19] G. Mittmann et al., "Pathologist-like explainable AI for interpretable Gleason grading in prostate cancer," *arXiv preprint arXiv:2410.15012*, 2024.

[20] A. Belinga, "AI-Enhanced Gleason Grading: A Comprehensive Approach," *arXiv preprint arXiv:2409.17122*, 2024.

[21] J. Silva-Rodríguez, "SICAPv2 - Prostate Whole Slide Images with Gleason Grades Annotations," *Mendeley Data*, V1, 2020. [Online]. Available: https://doi.org/10.17632/9xxm58dvs3.1 (accessed Dec. 30, 2024).

[22] J. P. Dominguez-Morales et al., "A systematic comparison of deep learning methods for Gleason grading and scoring," *Med. Image Anal.*, vol. 95, 103191, 2024.

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org