Prostate Cancer Patch Classification

Zhang Xinyue

Abstract

Gleason grading system is a method to evaluate the severity of prostate cancer. Under a microscope, it standardized and classified the degree of abnormality into different levels. Traditionally, grading is carried out manually. However, manual classification has the problems of strong subjectivity and is not suitable for large-scale high-throughput analysis. That is also the reason why AI-assisted pathological grading is important. The project aims to develop an automated classification model using deep learning, categorizing prostate histology patches into five classes: benign (Stroma, Normal) and malignant (Gleason Patterns 3, 4, and 5). Based on a dataset of H&E-stained images, we implemented a Transfer Learning approach based on the ResNet-18 architecture. The significant challenge is class imbalance. The number of G5 patterns is far less than that of other graded images. To solve this problem, a Baseline phase is used to inverse frequency weighting, and an Optimization phase is employed to manual penalty weighting and learning rate scheduling. The final model demonstrates improved generalization and sensitivity towards critical malignant classes, providing a robust framework for computer-aided diagnosis.

Introduction

Clinical background

As a tumor representing one of the most commonly diagnosed cancers, prostate cancer is a malignant tumor originating from the epithelial cells of the prostate gland[1]. In 1960s, there was a standard to grading the severity of prostate cancer called Gleason grading system[2]. The pathological sections were stained with Hematoxylin and Eosin (H&E), and then examined microscopically. Through glandular architecture, prostate cancer is classified into different scores[3,4]. In Gleason Pattern 3 (G3), the tumor still retains clear, complete and separated glandular structures, with regular glandular contours and no glandular fusion. The overall

structure is relatively close to normal prostate tissue. In Gleason Pattern 4 (G4), the tumor glandular structure is significantly damaged. G4 presents as glandular fusion, cribriform or poorly formed glandular formation, but some glandular structures can still be seen. The tumor is moderate to high invasiveness.

Gleason Pattern 5 (G5) represents the tumors completely or almost completely lose glandular structure. As solid cells growing in patches, arranged in a cord-like pattern, or accompanied by necrosis (comedonecrosis), it is a symbol of highly aggressive tumors.

Distinguishing between these patterns is critical for treatment planning. At present, the manual classification mainly relied on is highly subjective, and different pathologists may classify it differently. Meanwhile, relying on manual grading of each pattern one by one is relatively slow and it is difficult to handle a large number of patterns simultaneously.

Computational Pathology and Objectives

Convolutional Neural Networks (CNNs) have remarkable success in medical image segmentation and classification tasks[5-6]. They can automatically learn hierarchical feature representations directly from raw image data. By leveraging convolutional filters and deep architectures, CNNs can effectively capture complex spatial patterns, textures, and morphological characteristics. That are critical for interpreting medical images.

In segmentation tasks, CNN-based models enable precise delineation of anatomical structures and pathological regions, while in classification tasks they have achieved performance comparable to, or in some cases surpassing, that of human experts. These advantages have led to the widespread adoption of CNNs across various medical imaging modalities, including histopathology, radiology, and microscopy[7-8].

Methodology

Dataset description and partitioning

The dataset consists of histological patches extracted from Whole Slide Images (WSIs), which is provided by the National University Hospital. They were scanned at 20x magnification (~0.5 μm/pixel). Annotations were performed by pathologists using the A!HistoNotes platform(an annotation platform developed by CDPL team of BII).

The dataset has been divided into 3 subsets, SetA, SetB, Test_NTU_additional.. Each subset has already divided the patterns into training set, validation set and testing set. But the distribution has a clear class imblance. In Stroma, Normal, G3 and G4 dataset, there are 800 patches in each training set. But there are only 400 patches in G5 for training.

To ensure rigorous evaluation, we reorganized the provided subsets (SetA, SetB, Test_NTU_additional) into a logical machine learning split. The training set combined SetA-Train and SetB-Train to maximize feature diversity. The validation set combined SetA-Test and SetB-Test to monitor optimization and perform model selection. The Test_NTU_additional folder was strictly reserved for the final performance report to assess true generalization.

Data preprocessing and augmentation

Histological images are rotation-invariant; the biological meaning of a tissue patch does not change if it is rotated. To prevent overfitting and improve the model's robustness to variations in slide preparation, we applied extensive Data Augmentation during training:

Geometric Transformations: Random Horizontal Flip, Random Vertical Flip, and Random Rotation (90 degrees).

Color Space Transformations: H&E staining can vary significantly between laboratories. We applied Color Jitter (Brightness ±0.15, Contrast ±0.15, Saturation ±0.1, Hue ±0.05) to force the model to learn structural features rather than relying on specific color shades.

Normalization: All images were resized to 224x224 and normalized using ImageNet mean and standard deviation statistics.

Model Architecture: ResNet-18

ResNet-18 is a typical deep convolutional neural network (CNN). Its core innovation is the residual/skip connection. By introducing the identity shortcut, it enables the network to maintain stable training while deepening[9].

For medical images, especially pathological and small sample data, the data volume is usually limited and the annotation cost is high. Moreover, deep models. Like ResNet-50/101 are easier to overfitting. ResNet-18 is an excellent model to solve these problems.

Because in this project the scale of dataset is small, when training deep CNN from scratch, the following problems are prone to occur: overfitting, unstable convergence, and poor generalization ability. Thus we use transfer learning[11]. Utilizing a model pre-trained on a large-scale natural image dataset as a feature extractor and then transfering it to the current medical task.

Following the transfer learning strategy, a convolutional neural network backbone was employed to extract visual features from the input images. The convolutional base was initialized with ImageNet-pretrained weights, enabling the model to leverage generic low-level visual representations such as edges, textures, and basic structural patterns learned from large-scale natural image data.

The pretrained convolutional layers were used as a feature extractor, transforming the input images into high-level feature representations. To adapt the network to the target medical classification task, the original classification head of the pretrained model was removed.

A new fully connected (FC) layer was then introduced as the task-specific classifier. This linear layer outputs five logits, corresponding to the five target classes in our dataset. The newly added classification layer was randomly initialized and trained from scratch, allowing the model to learn discriminative features specific to the medical domain while retaining the general visual representations encoded in the pretrained convolutional base[10].

This architecture enables efficient knowledge transfer from large-scale natural image

datasets to a medical imaging task with limited annotated samples, while maintaining a compact and computationally efficient model design.

Loss Function: Weighted Cross-Entropy

G5 has a different number of patterns from other datasets. To address the problems of imbalanced data, we modified the standard Cross-Entropy Loss. In a standard loss, the model dominates by predicting majority classes. We introduced a Weighted Cross-Entropy Loss:

$$L = -\sum_{c=1}^{m} w_c y_{o,c} \log(p_{o,c})$$

Where w_c is the weight for class c.

Baseline model: Weights were calculated inversely proportional to class frequency

Optimization model: We manually tuned weights to heavily penalize misclassification of hard classes (G3) and the minority class (G5).

Experimental Setup

The experiments were conducted using PyTorch on a GPU-accelerated environment. For choosing optimizer, Adam (Adaptive Moment Estimation) was chosen for its fast convergence properties. Batch Size is 32. Learning Rate Schedule: To improve optimization, we implemented a StepLR Scheduler in the second phase, decaying the learning rate by a factor of 0.1 every 7 epochs. This allows the model to take large steps initially and settle into a precise local minimum towards the end.

Result

Baseline model is trained with standard inverse-frequency weighting for 15 epochs, and optimized model is trained with manual class weighting and learning rate scheduling for 25 epochs. Analyzing the difference between the baseline model and the optimized model.

Baseline Model Performance

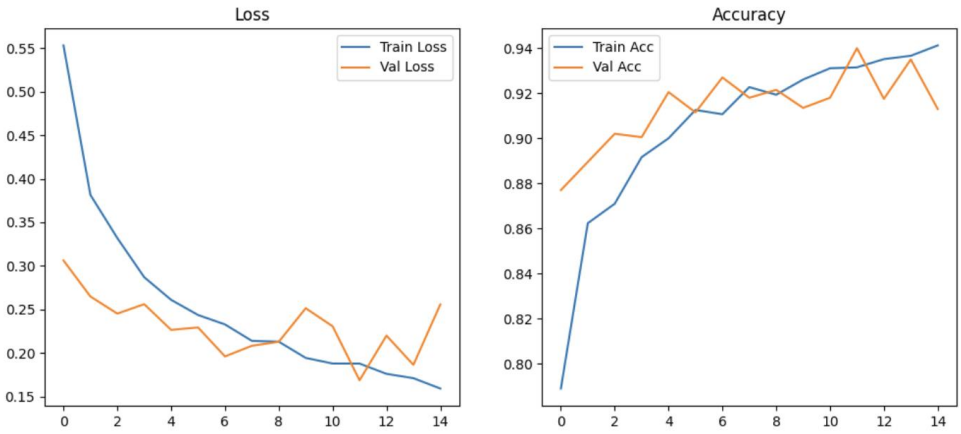

Figure 1: Learning Curves (Loss and Accuracy) of the Baseline ResNet-18 Model.

The baseline model showed stable convergence after transfer learning. The training loss decrease smoothly from 0.56 to 0.16 and the validation loss also showed same trend. The accuracy curve support this observation too. Training accuracy steadily increases and reaches approximately 94%, while validation accuracy remains stable within the range of 91–94% across epochs (Figure 1).

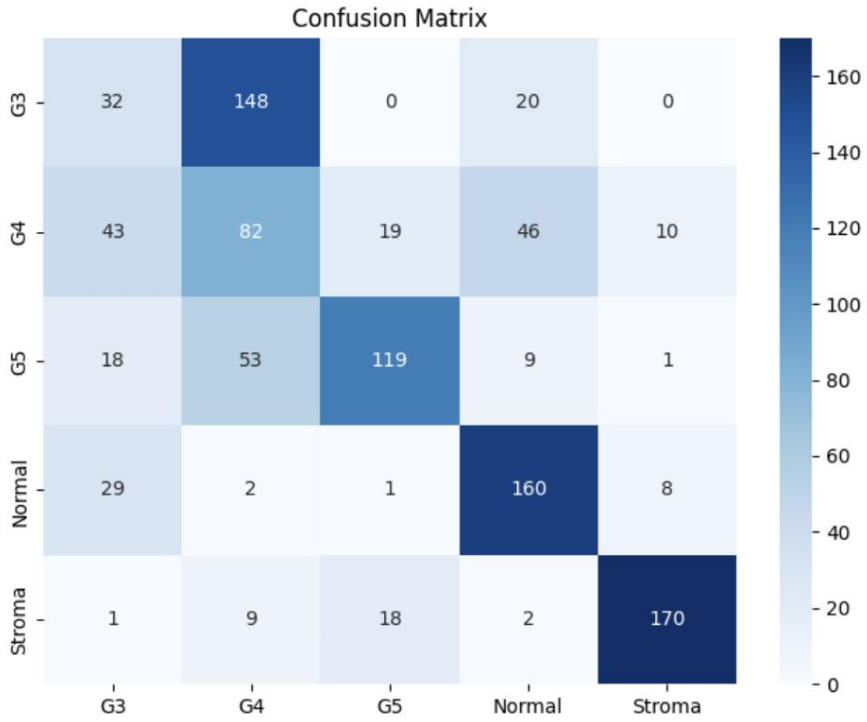

Figure 2: Confusion Matrix of the Baseline Model evaluating on Test_NTU_additional.

Evaluation the test set, there has a misclassification problem in different classes. The main problem is the model cannot recognize G3 and G4 patterns properly. G3 was only recognized correctly in 32 patterns. There are 148 patterns recognized as G4 (Figure 2).

Optimized Model Performance

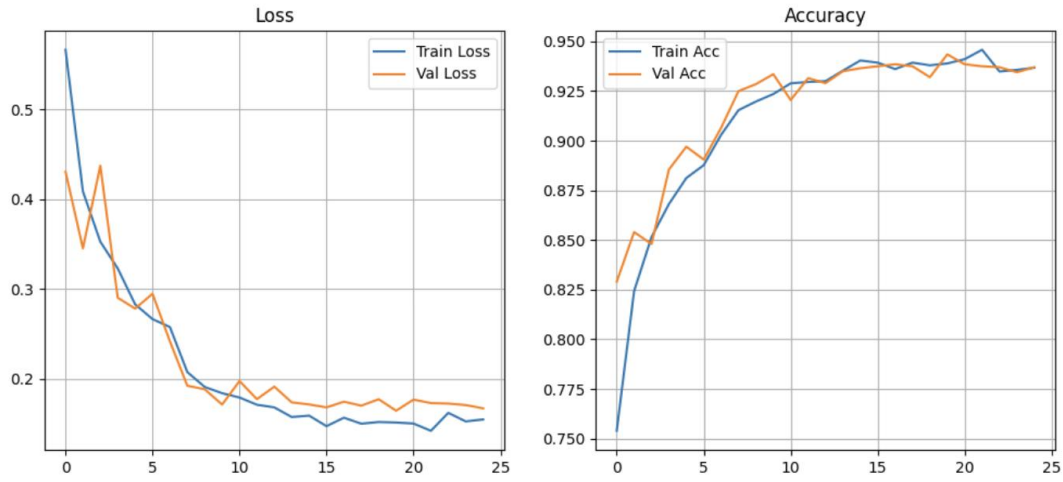

Figure 3: Learning Curves (Loss and Accuracy) of the optimized ResNet-18 Model.

After introducing manual class-specific penalty weighting and learning rate scheduling, the final optimized model exhibits improved training stability and more balanced class-wise behavior. The loss curves show smooth convergence in later epochs, and training and validation accuracy remain closely matched, indicating stable generalization (Figure 3).

	Percision	Recall	F1-score	Support
G3	0.43	0.45	0.44	200
G4	0.33	0.48	0.39	200
G5	0.52	0.46	0.49	200
Normal	0.72	0.82	0.77	200
Stroma	0.86	0.78	0.82	200
Accuracy			0.60	1000
Macro avg	0.57	0.60	0.58	1000

Weighted avg	0.57	0.60	0.58	1000

Table 1: Classification Report

On the final test set, the optimized model achieves more balanced recall across malignant classes. Notably, G3 recall improves substantially from 0.16 to 0.45, representing a major gain in sensitivity for low-grade malignant tissue. G4 recall reaches 0.48, while G5 recall remains at a reasonable level of 0.46, indicating that improved sensitivity to G3 is not achieved at the complete expense of high-grade cancer detection. Performance on benign classes remains strong, with recall of 0.82 for Normal and 0.78 for Stroma (Table 1).

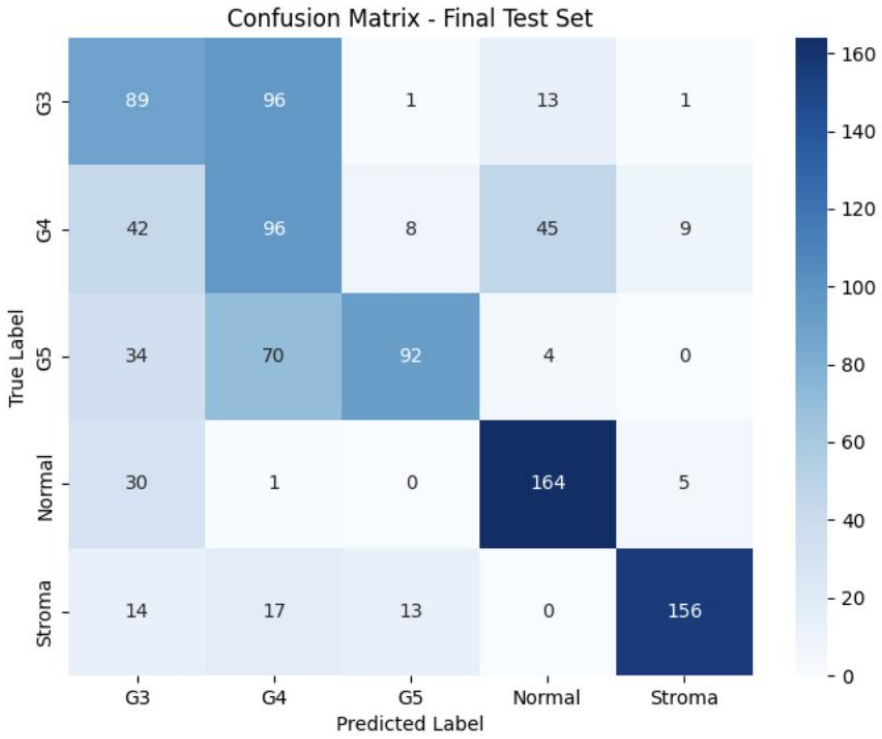

Figure 4: Confusion Matrix of the optimized Model evaluating on Test_NTU_additional.

There has a error pattern in clinical environment according to confusion martix of the optimized model(Figure 4). It is easier to misclassify the near Gleason grades, G3–G4 and G4–G5. But it seldomly classify tumors to benign patterns.

Discussion

This project faces the challenges about how to automated Gleason grading using deep learning. The ResNet-18 model using transfer learning performs well. It can

distinguish benign tumors from malignant ones very well. However, there still has challenges in distinguishing different Gleason grading patterns. This observation underscores the inadequacy of accuracy as a sole evaluation metric in the presence of severe class imbalance.

In fact, the original dataset has severe class imbalance. Although the large amount of benign tissues is easier to distinguish, G3 and G4 patterns have an unclear shape and G5 patterns have risk in clinic. The baseline model seek for the highest accuracy score. The benign tissues is easier to distinguish, and they have the most number of training patterns. So if the model can distinguish the benign tissues correctly, there will be a high accuracy score. This led to G3 being ignored and G5 being conservatively predicted as G4. It is best for mathematics but clinical manifestations are very poor.

We choose to add manual penalty weighting, increasing the cost of wrongly classifying G3 and G5. The model is encouraged to pay more attention to these classes, effectively reshaping the decision boundary toward regions corresponding to malignant tissue. This change improve the ability to recognize G3, reducing the risk of ignoring early malignant changes.

However, increasing penalty weights also amplifies gradient magnitudes and can destabilize training if applied in isolation. To mitigate this effect, a learning rate scheduling strategy was incorporated to control the optimization dynamics. Larger learning rates in early training allow the model to rapidly adapt to the modified loss landscape, while reduced learning rates in later stages enable fine-grained refinement of the decision boundary and promote stable convergence.

By incorporating manual penalty weighting and learning rate scheduling, the optimized model successfully reshapes the decision boundary toward clinically meaningful discrimination. Because the Gleason scores of these pathological sections themselves are controversial, there may be misjudgments between adjacent Gleason grades. So it still has many misjudgments.

Conclusion

In this project, we developed a deep learning–based framework for automated Gleason grading of prostate histopathology using a transfer learning approach with a ResNet-18 backbone. The incorporation of manual penalty weighting and learning rate scheduling enables a deliberate trade-off between sensitivity and specificity, producing a model whose decision-making process is better aligned with clinical diagnostic goals. This highlights the importance of optimization strategies that explicitly account for class imbalance and clinical relevance in computational pathology applications.

Reference

1. Epstein, J. I., Egevad, L., Amin, M. B., Delahunt, B., Srigley, J. R., Humphrey, P. A., & Grading Committee (2016). The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. The American journal of surgical pathology, 40(2), 244–252. https://doi.org/10.1097/PAS.0000000000000530

2. Gleason D. F. (1966). Classification of prostatic carcinomas. Cancer chemotherapy reports, 50(3), 125–128.

3. Epstein, J. I., Walsh, P. C., Carmichael, M., & Brendler, C. B. (1994). Pathologic and clinical findings to predict tumor extent of nonpalpable (stage T1c) prostate cancer. JAMA, 271(5), 368–374.

4. Humphrey P. A. (2004). Gleason grading and prognostic factors in carcinoma of the prostate. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc, 17(3), 292–306. https://doi.org/10.1038/modpathol.3800054

5. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical image analysis, 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005

6. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature,

542(7639), 115–118. https://doi.org/10.1038/nature21056

7. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. MICCAI 2015, LNCS 9351, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

8. Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., & Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine, 25(8), 1301–1309. https://doi.org/10.1038/s41591-019-0508-1

9. He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90

10. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Jianming Liang (2016). Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?. *IEEE transactions on medical imaging*, *35*(5), 1299–1312. https://doi.org/10.1109/TMI.2016.2535302