Intelligent Health Systems – From Technology to Data and Knowledge
E. Andrikopoulou et al. (Eds.)

doi:10.3233/SHTI250270

Graph Neural Networks for Gleason Grading in Prostate Histopathology Images

Hafsa AKEBLI[a,1], Kevin ROITERO[a], and Vincenzo DELLA MEA[a]
[a] *University of Udine, Italy*
ORCiD ID: Hafsa Akebli https://orcid.org/0009-0002-3879-876X, Kevin Roitero https://orcid.org/0000-0002-9191-3280, Vincenzo Della Mea https://orcid.org/0000-0002-0144-3802

Abstract. Prostate cancer is a leading cause of cancer-related deaths, with Gleason grading being key for assessing tumor aggressiveness. We propose a Graph Neural Network-based approach to automate Gleason grading using the Automated Gleason Grading Challenge 2022 dataset. Patch-level graphs constructed from Hematoxylin and Eosin-stained Whole-Slide Images were classified into Gleason grades. Our results show that Graph Neural Networks, specifically Graph Attention Networks and Graph Convolutional Networks, effectively distinguish between grades despite class imbalance. Focal Loss improves the classification of the minority Gleason Grade 5, which is crucial for detecting aggressive prostate cancer. Our models outperform state-of-the-art methods, achieving higher F1-scores without scanner generalization techniques.

Keywords. Gleason Grading - Graph Neural Networks - Histopathological Images - Prostate Cancer

1. Introduction

Prostate Cancer (PCa) is the second most common cancer worldwide and ranked as the fifth leading cause of cancer-related deaths among men in 2022 [1]. The Gleason grading system, which was introduced in 1966 by Donald Gleason and updated in 2010 [2], is one of the most reliable methods for evaluating PCa aggressiveness. Gleason grades reflect the growth patterns of prostate adenocarcinoma, and they are related to disease severity. The system scales PCa into five grades based on the differentiation of glandular patterns, ranging from 1 (excellent prognosis) to 5 (poor prognosis). Pathologists determine these grades by examining Hematoxylin and Eosin (H&E)-stained Whole-Slide Images (WSIs).

Research has demonstrated that computational pathology and deep learning techniques have the potential to achieve diagnostic accuracy comparable to expert pathologists [3]. Many studies have shown that deep learning is a powerful approach for automating cancer detection, including in prostate tissues, and for predicting the severity of the cancer stage (Gleason Grade) [4,5,6]. A recent systematic review [7] highlighted the efficacy of AI-based algorithms for PCa diagnosis and grading, especially in identifying morphologic features with prognostic significance. Among deep learning models, Graph Neural Networks (GNNs) [8] have gained significant attention due to

[1] Corresponding Author: Hafsa Akebli; E-mail: akebli.hafsa@spes.uniud.it.

their ability to work on non-Euclidean data and capture complex spatial relationships within images. GNNs have demonstrated particular effectiveness in processing medical images [9], especially in the analysis of histopathological images [10]. Several papers address the Gleason grading of prostate WSIs using GNNs; for example, studies such as [11,12] have used Graph Convolutional Networks (GCNs) to classify PCa.

In this paper, we investigate the application of Graph Attention Networks (GATs) and GCNs for Gleason grading of H&E-stained images from the Automated Gleason Grading Challenge 2022 (AGGC22).

2. Methods

In this study, we address a node classification problem of PCa histopathological images. The objective is to classify different patches of a WSI into Gleason grades: Normal tissue, Stroma, G3, G4, and G5. In this section, we describe the steps taken for graph construction, the application of GNNs, and the dataset used.

2.1. Graph Construction

For each prostate histopathological image, we divided the WSI into patches of 500×500 pixels using a sliding window approach, treating each patch as an individual node in the graph. The feature vectors for each patch were extracted using the "prostate medium" variant of the HistoEncoder model [13], which is a pretrained model on prostate tissue data. We then used the K-nearest neighbors (KNN) algorithm with K=5, connecting each node to its five most similar neighbors across the entire WSI, forming the graph's edges. Cosine similarity was used to initialize edge weights.

2.2. Graph Neural Networks

GNNs rely on an iterative process of aggregating information from neighboring nodes, allowing them to capture both structural and feature-based information from their nodes' neighborhoods. Simultaneously, node representations are combined to learn new features for the target nodes. In this study we utilized both GCNs first introduced in 2017 [14] and GATs first introduced in 2018 [15]. GCNs extend traditional convolutional operations to graph data by aggregating node features from their neighborhood. GATs attention mechanisms, assigning different importance to each neighboring node during feature aggregation. This mechanism allows the model to focus on important regions in histopathological images, improving classification outcomes.

2.3. Dataset

The dataset used in this study is from the AGGC22 organized during the MICCAI conference. It includes three subsets of annotated H&E-stained images. Subset 2 consists of biopsy images, while Subset 1 and 3 contain WSIs of prostatectomy specimens. Subset 1 and 2 images were scanned exclusively by Akoya Biosciences, whereas Subset 3 includes images scanned by multiple scanners (Akoya Biosciences, Olympus, Zeiss, etc.). We used all 286 WSIs available in the training data of this challenge. The dataset is class imbalanced. To highlight this, we calculated the imbalance ratio, which is defined as:

$Imbalance\ Ratio = \frac{\min(|C_i|)}{\max(|C_i|)}, i \in \{1, \ldots, \mathrm{m}\}$, where $|C_i|$ represents the number of samples in class C_i, and m is the total number of classes. From Table 1, we can conclude that G4 is the major class and G5 is the minor class, with an imbalance ratio of 3.70%. To handle class imbalance, we incorporated class weighting into the Focal Loss function. The weights were inversely proportional to the class frequencies, ensuring that the minority classes, particularly G5, received greater emphasis during training

Table 1. Number of patches (nodes) per class and corresponding imbalance ratio.

Class	Total Patches	Imbalance Ratio (%)
Normal	201216	31.03
Stroma	265510	40.96
G3	392647	60.54
G4	648443	100.00
G5	24006	3.70

3. Results

3.1. Implementation Details

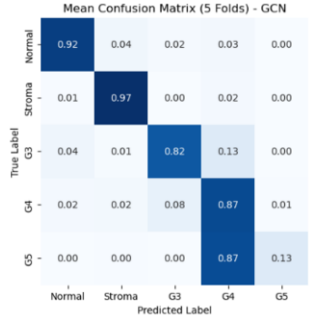
Patches were extracted using a sliding window of 500×500 pixels with a 250-pixel step, retaining only those containing over 90% tissue for Subsets 1 and 3, and over 75% for biopsy images in Subset 2 due to smaller G5 class masks. The GAT model had 3 layers and 16 attention heads, while both GCN and GAT used 64 hidden units per layer. These hyperparameters were selected after experiments testing various configurations. Model evaluation used 5-fold cross-validation, ensuring patches from the same WSI were not shared across folds. Metrics were averaged across folds. Training ran for 200 epochs on an NVIDIA RTX A6000 with a batch size of 16 and a learning rate of 0.005. The Cross-Entropy loss function and Adam optimizer with a weight decay of $1 \times 10{-}4$ were applied, while AdamW with a learning rate scheduler handled Focal Loss to address class imbalance.

3.2. Experimental Results

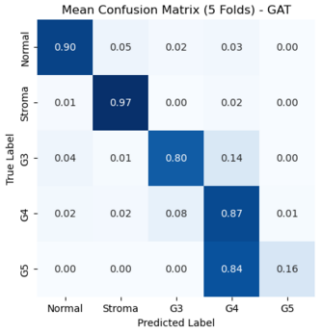
Due to the dataset's imbalance, traditional metrics like accuracy are insufficient. Instead, we focused on F1-score, precision, recall, and confusion matrices. As summarized in Table 2 both GCN and GAT models achieved higher mean F1-scores and recall compared to their counterparts trained with Focal Loss. However, the confusion matrices in Figure 1 reveal a clearer picture of model performance across the five Gleason grades. While the mean metrics for GCN and GAT without Focal Loss appear better, both models struggled with imbalanced classes, especially G5 class, which is critical for aggressive cancer detection. The confusion matrices offer class-wise insights, highlighting that overall metrics may overlook minority class performance. In the GAT model without Focal Loss 1b, G5 patches were almost entirely misclassified. After applying Focal Loss 1d, the model successfully predicted G5 with an 80% recall, indicating that Focal Loss enhances the classification of minority classes without oversampling. The GCN models 1a and 1c showed similar improvements in G5 classification after Focal Loss, though overall performance slightly lagged behind GAT.

Table 2. Mean node classification performance for Gleason Grading over 5-fold cross-validation.

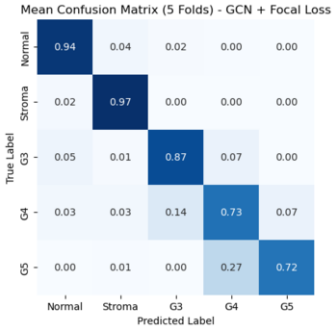
Model	F1 score	Precision	Recall
GAT (3 layers)	0.859	0.865	0.862
GCN (3 layers)	0.866	0.871	0.870
GAT (3 layers) + Focal Loss	0.830	0.858	0.823
GCN (3 layers) + Focal Loss	0.839	0.861	0.834



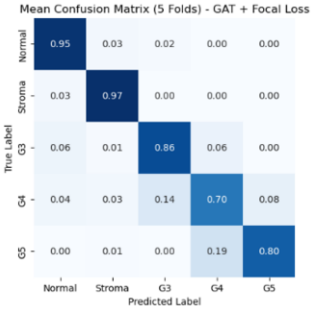
(a) Confusion Matrix of GCN.



(b) Confusion Matrix of GAT.



(c) Confusion Matrix of GCN with Focal Loss.



(d) Confusion Matrix of GAT with Focal Loss.

Figure 1. Confusion matrices.

4. Discussion

Both GCN and GAT models performed well in classifying prostate histopathology images, with GCN achieving a slightly higher F1-score (0.866) than GAT (0.859). Although Focal Loss slightly lowered overall performance, it significantly improved classification of the minority class G5, critical for detecting aggressive PCa, demonstrating its effectiveness for imbalanced datasets. Compared to the state-of-the-art method from the AGGC22 [16], which reported an F1-score of 0.73 on multi-scanner images, our models achieved better baseline results (0.859 with GAT, 0.866 with GCN) without generalization techniques. The AGGC22 model improved to 0.88 with generalization, suggesting that adopting similar techniques could further enhance our performance, particularly on mixed-scanner datasets.

5. Conclusions

In this study, we demonstrated the effectiveness of GCN and GAT models for Gleason grading in prostate histopathology images, with GCN slightly outperforming GAT in overall metrics. Focal Loss played a crucial role in improving the classification of the minority G5 class, which is critical in clinical settings. Our models surpassed the baseline performance reported by the AGGC22 state-of-the-art method, achieving higher F1-scores without the use of scanner generalization techniques. Future work could focus on incorporating such generalization methods and exploring graph-based oversampling like GraphSMOTE to further enhance classification, particularly in imbalanced datasets.

Acknowledgments

This work was partially supported by Next-Generation EU (PNRR: Piano Nazionale di Ripresa e Resilienza --- Missione 4 Componente 2, D.M. 118/2023).

References

[1] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians. 2024;74(3):229-63.
[2] Epstein JI. An update of the Gleason grading system. The Journal of Urology. 2010;183(2):433-40.
[3] Litjens G, Sanchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency ´ of histopathological diagnosis. Scientific Reports. 2016;6(1):26286.
[4] Linkon AHM, Labib MM, Hasan T, Hossain M, Marium-E-Jannat. Deep learning in prostate cancer diagnosis and Gleason grading in histopathology images: An extensive study. Informatics in Medicine Unlocked. 2021;24:100582.
[5] Li Y, Huang M, Zhang Y, Chen J, Xu H, Wang G, et al. Automated Gleason Grading and Gleason Pattern Region Segmentation Based on Deep Learning for Pathological Images of Prostate Cancer. IEEE Access. 2020;8:117714-25.
[6] Singhal N, Soni S, Bonthu S, Chattopadhyay N, Samanta P, Joshi U, et al. A deep learning system for prostate cancer diagnosis and grading in whole slide images of core needle biopsies. Scientific Reports. 2022;12(1):3383.
[7] Marletta S, Eccher A, Martelli FM, Santonicco N, Girolami I, Scarpa A, et al. Artificial intelligence–based algorithms for the diagnosis of prostate cancer: A systematic review. American Journal of Clinical Pathology. 2024 02;161(6):526-34. Available from: https: //doi.org/10.1093/ajcp/aqad182.
[8] Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The Graph Neural Network Model. IEEE Transactions on Neural Networks. 2009;20(1):61-80.
[9] Ahmedt-Aristizabal D, Armin MA, Denman S, Fookes C, Petersson L. Graph-Based Deep Learning for Medical Diagnosis and Analysis: Past, Present and Future. Sensors. 2021;21(14).
[10] Ahmedt-Aristizabal D, Armin MA, Denman S, Fookes C, Petersson L. A survey on graph-based deep learning for computational histopathology. Computerized Medical Imaging and Graphics. 2022;95:102027.
[11] Behzadi MM, Madani M, Wang H, Bai J, Bhardwaj A, Tarakanova A, et al. Weakly-supervised deep learning model for prostate cancer diagnosis and Gleason grading of histopathology images. Biomedical Signal Processing and Control. 2024;95:106351.
[12] Wang J, Chen RJ, Lu MY, Baras A, Mahmood F. Weakly Supervised Prostate Tma Classification Via Graph Convolutional Networks. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI); 2020. p. 239-43.
[13] Pohjonen J. HistoEncoder: Foundation models for digital pathology. GitHub; 2023. https://github.com/jopo666/HistoEncoder.
[14] Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks; 2017. Available from: https://arxiv.org/ abs/1609.02907.
[15] Velickovi ˇ c P, Cucurull G, Casanova A, Romero A, Li ´ o P, Bengio Y. Graph Attention Networks; 2018. Available from: ` https://arxiv. org/abs/1710.10903.
[16] Huo X, Ong KH, Lau KW, Gole L, Young DM, Tan CL, et al. A comprehensive AI model development framework for consistent Gleason grading. Communications Medicine. 2024;4(1):84.