Development of an Interpretable and Uncertainty-Aware Deep Learning Model for Gastric Cancer Histopathological Image Classification

Aikaterini Martakou Galiatsatou
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
kmartgal99@gmail.com

Maria Athanasiou
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
mathanasiou@biosim.ntua.gr

Konstantina S. Nikita
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
knikita@ece.ntua.gr

2025 IEEE 25th International Conference on Bioinformatics and Bioengineering (BIBE) | 979-8-3315-5899-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/BIBE66822.2025.00088

***Abstract*— Histopathological image analysis is the gold standard for cancer diagnosis but is time-consuming, requires a high level of expertise, and is subject to inter-observer variability. The advancement of Digital Pathology enables the application of deep learning models for automating and enhancing diagnostic accuracy. In particular, Convolutional Neural Networks (CNNs) have emerged as a powerful tool for identifying morphological features in histopathological images, achieving accuracy comparable to that of medical experts in specific tasks such as tissue classification. Within the framework of this study, an interpretable CNN model is developed and evaluated for classifying gastric histopathological images as benign or malignant using the GasHisSDB dataset. The Monte Carlo Dropout method is applied to estimate the uncertainty of the model's predictions, while the Gradient-weighted Class Activation Mapping (Grad-CAM) method is combined with quantitative image feature analysis towards enabling the spatial localization of image regions that influence the model's predictions. The proposed approach achieved an AUC score of 98.6% while maintaining low computational cost and architectural simplicity. The generated interpretations yielded useful insights into spatially important image regions and their associated nuclear morphological characteristics.**

Keywords*— *Deep Learning, Convolutional Neural Networks, Gastric Cancer, Histopathology, Image Classification, Uncertainty Quantification, Interpretability, Grad-CAM, Monte Carlo Dropout

I. Introduction

Histopathology, which involves the microscopic examination of tissues and cells, remains the gold standard for diagnosing a wide range of diseases, including cancer [1]. Accurate classification of histopathological images is crucial for effective diagnosis, treatment planning, and prognosis. Traditionally, this task has relied on visual assessment of digital microscopy slides by expert pathologists. However, this process is time-consuming, subjective, and susceptible to inter-observer variability.

Advancements in artificial intelligence have enabled the increasing application of machine learning (ML) and deep learning (DL) methods, such as Convolutional Neural Networks (CNNs), in histopathological image classification [2]. These models are trained to recognize morphological features and patterns associated with either normal or pathological conditions [3], thereby reducing diagnostic subjectivity and accelerating the diagnostic process through the detection of subtle differences in cellular or histological characteristics, which may escape human observation.

Gastrointestinal (GI) cancer is a common form of malignancy, often asymptomatic in its early stages, and constitutes the fifth most frequently diagnosed type of cancer, accounting for 18% of total cancer-related deaths [4]. Histopathological examination remains essential for GI cancer diagnosis, typically involving the microscopic analysis of tissue sections from suspicious lesions, conducted by experienced pathologists [5].

The successful integration of deep learning (DL) methods into histopathological workflows has demonstrated its potential to achieve expert-level diagnostic performance while improving time and cost efficiency and scalability [6]. In the domain of GI cancer, a wide range of DL techniques have been explored for the classification of histopathological images. Most studies adopt pretrained CNNs such as VGG16, ResNet50, and DenseNet121, which are fine-tuned on medical image datasets to extract complex morphological features. Beyond these, hybrid and ensemble models have also been developed to boost performance through methods such as attention mechanisms, including the Multi-Channel Attention Mechanism (MCAM) for gastric cancer [7] and stacking ensembles combining multiple CNNs [8]. In parallel, Vision Transformers (ViTs) and CNN-Transformer hybrids, such as GasHis-Transformer [9] and CCDNet, have emerged as promising alternatives, particularly due to their ability to capture global spatial relationships in complex tissue structures. However, despite their impressive accuracy, these approaches often come with substantial computational demands, high model complexity, lack of interpretability, and limited suitability for deployment in resource-constrained environments or real-time clinical settings.

To address these challenges, the present study proposes an interpretable CNN model, based on a lightweight architecture, for the classification of gastric cancer histopathology images into benign and malignant categories, using the GasHisSDB dataset [10]. To enhance model transparency and user trust, uncertainty quantification is performed using Monte Carlo Dropout [11], enabling assessment of the confidence of the model's predictions, while the Gradient-weighted Class Activation Mapping (Grad-CAM) method [12] is employed to visualize the model's focus on specific spatial regions during inference. Furthermore, a nuclei-level interpretability analysis is conducted, combining Grad-CAM activation heatmaps with quantitative cellular features extracted via HistomicsTK [13], with the aim of providing deeper insights into the biological relevance of the model's decision-making process. An overview of the proposed approach is presented in Figure 1. By focusing on architectural simplicity, reliable performance, and transparency, this work aims to address some of the key challenges that remain open in the current landscape of DL-based histopathology [14].

2471-7819/25/$31.00 ©2025 IEEE
DOI 10.1109/BIBE66822.2025.00088

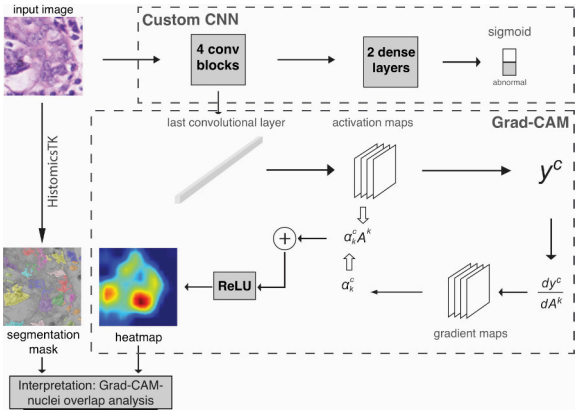

Fig.1. Overview of the proposed pipeline for the classification of gastric histopathology images and the generation of interpretations.

II. METHODS

A. Dataset

The development and evaluation of the proposed model was based on the use of the GasHisSDB dataset, which contains 245,196 histopathology image patches, categorized as Abnormal (97,076 patches) and Normal (148,120 patches) [10]. The dataset was constructed through a two-stage process. Initially, 600 high-resolution (2048×2048) gastric tissue images were annotated as Normal or Abnormal by four pathologists at Longhua Hospital. These images were then partitioned into sub-images of varying resolutions by biomedical researchers at Northeastern University and further verified by two expert pathologists from Liaoning Cancer Hospital, ensuring label consistency and diagnostic accuracy. Within the framework of the present study, the 120×120 pixel resolution was selected, corresponding to a subset of 65,261 image patches, with the aim to strike a balance between preserving sufficient morphological information and ensuring computational efficiency.

B. Proposed CNN Model

1) Model Architecture: The proposed model's architecture was based on a custom 8-layer Convolutional Neural Network (CNN), consisting of four convolutional blocks. Each block contained two Conv2D layers with ReLU activation and 3×3 kernels, followed by Batch Normalization, and a MaxPooling2D operation. The number of filters doubled with each block (from N to 8N), enabling progressive feature abstraction. During training, Dropout was applied after each block with progressively increasing rates (0.25 to 0.4) to prevent overfitting as the network deepened.

The resulting feature maps were flattened and passed through two fully connected (Dense) layers with 512 and 256 neurons, each followed by Batch Normalization and Dropout (applied during training). The final classification layer was a single neuron with sigmoid activation, producing a probability for classifying an image patch as abnormal. This architecture was lightweight yet expressive, leveraging hierarchical feature extraction to capture cancer-relevant patterns. To ensure convergence and generalization, the model integrated regularization techniques (Dropout, Batch Normalization) and was trained using the Adam optimizer.

2) Hyperparameter Tuning: Hyperparameter optimization was performed using Grid Search over a predefined parameter space. The grid featured the number of filters [32, 64], kernel size [(3,3), (5,5)], dropout rate [0.3, 0.5], and learning rate [0.001, 0.0001]. Each configuration was evaluated on a fixed validation set.

C. Uncertainty Estimation

The Monte Carlo Dropout (MC Dropout) method was applied for quantifying prediction uncertainty. After training, each test image was passed through the model 50 times with dropout enabled during inference, producing a distribution of stochastic outputs. The mean of the predicted probabilities from these forward passes was used as the final classification output:

$$\mathbb{E}_{q(y^*|x^*)}(y^*) \approx \frac{1}{T}\sum_{t=1}^{T} \hat{y}^*(x^*, W_1^{(t)}, \ldots, W_L^{(t)}) \qquad (1)$$

where $\mathbb{E}_{q(y^*|x^*)}(y^*)$ is the expected prediction for the input x^* under the predictive distribution $q(y^* \mid x^*)$, representing the mean prediction, T is the number of stochastic forward passes through the model, $\hat{y}^*(x^*, \mathbf{W}_1^{(t)}, \ldots, \mathbf{W}_L^{(t)})$ is the model's predicted output during the t-th pass, given input x^* and a specific set of sampled weights.

The standard deviation of predictions across forward passes served as a measure of uncertainty:

$$\mathrm{Var}(y^*) \approx \frac{1}{T}\sum_{t=1}^{T}\left(\hat{y}_t^*\right)^2 - \left(\frac{1}{T}\sum_{t=1}^{T}\hat{y}_t^*\right)^2 \qquad (2)$$

where $Var(y)$ represents the predictive variance of the model for a new input x, quantifying the uncertainty in its predictions, and $\hat{y}_t^*$ corresponds to the scalar prediction generated at each pass.

D. Interpretability Using Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM), a post-hoc visualization method widely used for explaining the spatial reasoning of CNNs, was deployed for generating interpretations of the model's predictions [12]. Grad-CAM identifies the spatial regions of an input image that contribute most significantly to a model's decision, offering visual explanations without requiring architectural modifications or retraining. The method operates by computing the gradients of the predicted class score y^c with respect to the feature maps A^k from a chosen network layer. These gradients are then globally averaged to produce importance weights α_k^c for class c and channel k:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \qquad (3)$$

where Z is the number of spatial locations. The final Grad-CAM heatmap is computed by taking a weighted combination of the feature maps, followed by a ReLU activation:

$$L_{\text{Grad-CAM}}^c = \mathrm{ReLU}\left(\sum_k \alpha_k^c A^k\right) \qquad (4)$$

This heatmap is then resized to the original image dimensions and overlaid on the input image to visualize which spatial regions most influenced the model's output.

In the present study, Grad-CAM allowed for qualitative validation of the CNN's focus areas in histopathology images, verifying whether the network based its decisions

clinically relevant regions. It should be noted that Grad-AM's effectiveness is influenced by the choice of nvolutional layer, as early layers capture lower-level ıtures (e.g., texture, structural) offering less semantic ormation, while deeper layers may sacrifice spatial olution but provide more semantic meaning. In this ıdy, the last convolutional layer was selected with the n of focusing on high-level semantic patterns associated th tissue-level abnormalities.

Feature Analysis with HistomicsTK

To further explore the morphological characteristics of age regions highlighted by Grad-CAM, the HistomicsTK rary [15] was utilized, which is an open-source toolkit processing whole-slide images. After applying color rmalization (Reinhard) and hematoxylin extraction via in deconvolution, nuclear segmentation was performed ing a multi-scale Laplacian of Gaussian (LoG) detector. r each segmented nucleus, morphometric features (area, entricity, perimeter), Fourier Shape Descriptors (FSD), dient-based features (edge sharpness), and Haralick ure features (contrast and correlation from the gray-level occurrence matrix-GLCM) were extracted.

Focusing on the interpretation of Grad-CAM results for positive predictions, nuclei were categorized as vated (located within Grad-CAM activation regions) or -activated (located outside activation regions), and their acted feature distributions were compared. This analysis vided insight into whether the model's focus aligned with logically suspicious or diagnostically relevant regions, ngthening the biological interpretability of proposed del's decisions.

III. Results And Discussion

Evaluation Framework

To reliably assess the model's generalization ability, a tified 5-fold cross-validation scheme was employed, ntaining the original class distribution across all folds. In h iteration, four folds were used for training, and one held out for testing. A further 75%-25% stratified split the training data created a validation set within each ation of the cross-validation scheme. All hyper-ameters were fixed across folds for consistency. After erparameters' optimization and evaluation through cross-dation, a final evaluation was conducted using a 90-10 n-test split. Model performance was assessed using a nprehensive set of metrics: Accuracy (Acc), Sensitivity ns), Specificity (Spec), Precision (Prec), F1-Score, Area der the Receiver Operating Characteristic Curve (AUC), tthews Correlation Coefficient (MCC) [16], and the den index.

Hypeparameters' Tuning

Following Grid Search, the identified optimal figuration included 64 filters, 3×3 kernels, 0.5 dropout , and 0.0001 learning rate, which provided the best le-off between training stability, generalization, and uracy. The use of smaller filters enabled finer feature action, while higher dropout and lower learning rate imized overfitting. Model training was conducted for up 5 epochs, with EarlyStopping based on validation loss. ge preprocessing was performed using ImageData nerator, applying pixel rescaling (1./255) and generator-ed data loading for training and validation. A decision shold of 0.5397 was determined using Youden's J istic on the ROC curve, by selecting the cutoff that ximized the difference between true positive rate and e positive rate on the validation set.

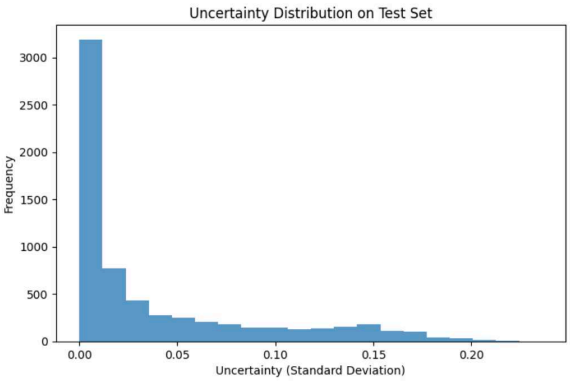

Fig.2. Histogram of Uncertainty (Standard Deviation) on the Test Set.

C. Evaluation of the Model's Performance

Table I summarizes the obtained performance of the proposed model using 5-fold stratified cross-validation. The model achieved balanced and strong discrimination performance, with an average AUC of 98.08%±3.02%, Sensitivity of 89.15%±3.30% and Specificity of 90.65% ±2.96%. The observed MCC of 79.19%±6.28% demonstrated a robust correlation between predicted and actual labels, while the F1-score of 87.25%±3.80% reflected a good balance between precision and recall. These results suggest that the model generalized well across folds and maintained consistent behavior in distinguishing abnormal from normal tissue patches.

TABLE I
PERFORMANCE OF THE PROPOSED CNN MODEL DURING 5-FOLD CROSS-VALIDATION (METRICS REPORTED AS AVERAGE ± STANDARD DEVIATION ACROSS FOLDS).

Acc	*Sens*	*Spec*	*Prec*	*F1*	*AUC*	*MCC*
90.08% ±3.02%	89.15% ±3.30%	90.65% ±2.96%	85.44% ±4.41%	87.25% ±3.80%	96.05% ±1.96%	79.19% ±6.28%

Table II presents the final evaluation results obtained after training the model on 90% of the dataset and testing on the remaining 10%, using the optimal configuration identified through grid search. The model yielded high discrimination performance, with an AUC of 98.60% , MCC of 0.8757, and F1-score of 92.33%. These findings further confirmed the model's effectiveness when deployed on unseen data. Furthermore, when compared to state-of-the-art models trained and evaluated on the GasHisSDB dataset, including MCAM [6], a Gaussian-mixture-based Naïve Bayes framework [17], and a hybrid model combining DenseNet-201 deep features with a Random Forest classifier [18], which achieved accuracy scores of 99.60%, 98.47%, and 91.93%, respectively, the proposed CNN model demonstrated competitive performance, yielding a comparable accuracy of 94.12% despite its simpler architecture. Additionally, when evaluated against the GasHis-Transformer [9], which was developed and assessed on a different gastric histopathological image dataset and achieved an accuracy of 97.97%, the proposed CNN also exhibited comparable performance.

TABLE II
PERFORMANCE OF THE PROPOSED CNN MODEL DURING 90-10 SPLIT.

Acc	*Sens*	*Spec*	*Prec*	*F1*	*AUC*	*MCC*
94.12%	93.15%	94.71%	91.52%	92.33%	98.60%	87.57%

D. Uncertainty Estimation

To visually represent the uncertainty of the model on the test set, a histogram of standard deviation values from MC Dropout predictions was generated (Figure 2). The distribution revealed that in the vast majority of cases, the model exhibited low uncertainty, which indicated high confidence in its predictions. As expected, predictions with high uncertainty were associated with output probabilities near the decision threshold of 0.5397, reflecting the model's lack of confidence in class assignment. Capturing these cases of high-uncertainty predictions could serve as an alert mechanism for human review, ensuring that critical or ambiguous cases are not overlooked.

E. Interpretability

1) Grad-CAM Visualization: Within the context of the Grad-CAM analysis, two different visualisations were generated for each test image with the aim of shedding light on specific regions influencing the model's decisions: (i) the normalized Grad-CAM heatmap, representing the raw activations of the final convolutional layer and (ii) a color-mapped overlay, where the heatmap was projected onto the original input image using a semi-transparent JET colormap.

Figure 3 depicts two examples of selected true positive (TP) cases alongside their corresponding heatmaps and color-mapped overlays. Visual inspection of the obtained heatmaps shows that Grad-CAM activations were systematically concentrated on biologically meaningful cellular structures, particularly dark-stained, morphologically atypical nuclei or densely packed nuclei.

In contrast, regions of background or healthy tissue exhibited minimal activation, demonstrating that the model effectively filtered out irrelevant regions. Overall, the obtained results indicated the model's ability to focus on morphological features commonly associated with abnormal tissue regions, such as nuclear atypia, increased cellular density, and nuclear hyperchromasia.

2) Nuclei-Level Interpretation with HistomicsTK: To quantitatively link Grad-CAM activation regions with biological entities and validate the model's decision-making process through measurable nuclear associations, activation maps were integrated with nuclear segmentation masks generated using HistomicsTK. This enabled not only visual comparison but also precise spatial correlation between activation maps and segmented nuclei, allowing for more detailed biological interpretation. Each Grad-CAM heatmap was converted into a binary activation mask using a threshold of 0.5 and each nucleus (uniquely labeled in the segmentation mask) was evaluated for overlap with Grad-CAM activated regions. Nuclei intersecting with activated regions were labeled as "activated."

Table III summarizes the total number of segmented nuclei, the number of activated nuclei, and the percentage of nuclei overlapping with Grad-CAM activations. The proportion of activated nuclei ranged from 30% to 60%, with the TP case "Abnormal-12012" exhibiting the highest overlap (60%). This variation suggested heterogeneity in how consistently the model's attention aligned with nuclear structures across different cases.

Figure 4 depicts representative visualizations for the two examples of selected TP cases, including the original image, the segmented nuclei, and the Grad-CAM heatmap overlay. For both samples, Grad-CAM heatmaps showed concentrated activations in regions densely populated with morphologically atypical nuclei. The segmentation masks confirmed that these activated regions corresponded to areas of potential diagnostic relevance, such as irregularly shaped, hyperchromatic nuclei. The fact that for the two examined cases, nearly half or more of the nuclei, as shown in Table III, fell within activated regions, strengthened confidence in the interpretability and clinical relevance of the CNN's learned representations. Thus, the combination of heatmaps and nuclei masks demonstrated that the model's decision-making was not arbitrarily distributed across the image but rather focused on biologically meaningful regions.

TABLE III
NUMBER OF TOTAL AND ACTIVATED NUCLEI. THE TWO EXAMINED TRUE POSITIVE CASES ARE HIGHLIGHTED IN BOLD.

Image	*Total Nuclei*	*Active Nuclei*	*Percent in Grad-CAM (%)*
Abnormal-01902	**20**	**9**	**45.0**
Abnormal-23815	11	6	54.55
Abnormal-12012	**15**	**9**	**60.0**
Abnormal-16643	18	7	38.89
Abnormal-12287	20	6	30.0

3) Feature Analysis with HistomicsTK: To further explore the influence of activated nuclei on the model's decisions, quantitative features extracted for each nucleus using HistomicsTK were combined with Grad-CAM activation data towards examining whether activated nuclei exhibited distinct characteristics compared to non-activated ones. For each segmented nucleus, features from the following categories were considered: (i) morphometric, (ii)

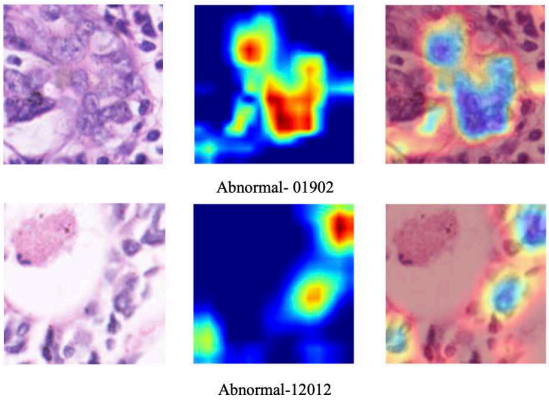

Fig.3. Original image samples (left), Grad-CAM heatmaps (middle), and color-mapped overlays (right) for the true positive samples 01902 and 12012.

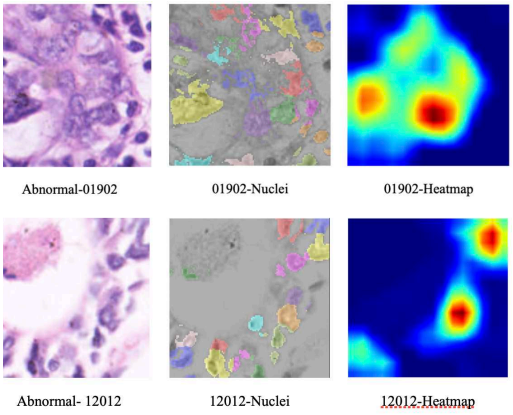

Fig.4. Original image samples (left), segmentation masks (middle), and Grad-CAM heatmaps (right) for the true positive samples 01902 and 12012.

Fourier Shape Descriptors (FSD), (iii) gradient-based, and (iv) Haralick texture. To compare feature distributions between activated and non-activated nuclei, the Mann-Whitney U test was applied independently to each feature [19] and the obtained p-values were recorded.

Table IV presents representative features from each category that exhibited statistically significant differences (p-value < 0.05) between Grad-CAM activated and non-activated nuclei. Among the morphometric features, both area and perimeter were significantly larger in activated nuclei (e.g., Abnormal-12012) suggesting that the model tended to focus on larger nuclei structures. FSD features (e.g., FSD4 and FSD5 in Abnormal-16643) also showed statistically significant separation, highlighting the model's sensitivity to subtle variations in nuclear boundary complexity. Regarding gradient-based features, standard deviation and kurtosis of the gradient magnitude (e.g., Abnormal-01902) both presented statistically significant differences between activated and non-activated nuclei, indicating the model's focus on sharper transitions and localized structural variation. Texture-based Haralick features, including contrast range and correlation mean, highlighted heterogeneity in chromatin texture and spatial relationships within the nucleus as discriminative characteristics. These findings provided evidence regarding the alignment between Grad-CAM activations and observed differences in nuclear morphology and texture.

F. Limitations and Future Work

Potential limitations of the proposed approach include the patch-level data splitting inherent to the GasHisSDB dataset, which consists of 120×120 image patches extracted from whole-slide images. This design may have introduced data leakage, potentially inflating the reported performance metrics. Moreover, the dataset represents a limited subset of gastric histopathology collected under specific staining and imaging conditions, which may restrict generalizability to data from other institutions or protocols. Future work will focus on large-scale validation across more diverse gastric cancer subtypes and imaging conditions, as well as on independent, multi-institutional cohorts to assess real-world applicability. Extending the approach to other histopathological domains (e.g., colorectal, lung, or breast cancer) may further demonstrate its adaptability and generalizability.

TABLE IV
REPRESENTATIVE FEATURES FROM EACH CATEGORY SHOWING STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN ACTIVATED AND NON-ACTIVATED NUCLEI.

Morphometric Features		
Image	***Feature***	***p-value***
Abnormal-12012	Size.Area	0.038995
Abnormal-12012	Size.Perimeter	0.017582
FSD		
Abnormal-16643	Shape.FSD4	0.008296
Abnormal-16643	Shape.FSD5	0.000440
Gradient Based Features		
Abnormal-01902	Gradient.Mag.Std	0.006237
Abnormal-01902	Gradient.Mag.Kurtosis	0.040239
Haralick Features		
Abnormal-01902	Haralick.Contrast.Range	0.003047
Abnormal-01902	Haralick.Correlation.Mean	0.004938

IV. CONCLUSION

This study demonstrated the effectiveness of a relatively lightweight CNN model in the classification of gastric cancer histopathological images. Despite the absence of very deep or pretrained architectures, the model achieved performance levels comparable to state-of-the-art approaches and demonstrated its ability to provide predictions with high confidence. The combination of Grad-CAM and HistomicsTK enabled both visual localization of discriminative image regions and quantitative analysis of nuclear morphology, and provided evidence regarding the biological relevance of the model's decisions. This integrative approach not only improves transparency and trust in deep learning systems but also brings AI one step closer to clinical applicability in digital pathology.

REFERENCES

[1] I. Mallick, "What Is Histopathology?," Verywell Health, Feb. 21, 2022. https://www.verywellhealth.com/histopathology-2252152

[2] A. S. Panayides et al., "AI in Medical Imaging Informatics: Current Challenges and Future Directions," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 7, pp. 1837–1857, Jul. 2020, doi: https://doi.org/10.1109/JBHI.2020.2991043

[3] S. Indolia, A. K. Goswami, S. P. Mishra, and P. Asopa, "Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach," Procedia Computer Science, vol. 132, pp. 679–688, 2018, doi: https://doi.org/10.1016/j.procs.2018.05.069

[4] GIS, "Stomach Cancer," Gastrointestinal Society. https://badgut.org/information-centre/a-z-digestive-topics/stomach-cancer/

[5] F.-H. Zhu et al., "The Histopathological Types and Distribution Characteristics of Gastric Mixed Tumors," Frontiers in Oncology, vol. 12, Jun. 2022, doi: https://doi.org/10.3389/fonc.2022.873005

[6] L. Fass, "Imaging and cancer: A review," Molecular Oncology, vol. 2, no. 2, pp. 115–152, May 2008, doi: https://doi.org/10.1016/j.molonc.2008.04.001

[7] M. Zubair et al., "An interpretable framework for gastric cancer classification using multi-channel attention mechanisms and transfer learning approach on histopathology images," Scientific Reports, vol. 15, no. 1, Apr. 2025, doi: https://doi.org/10.1038/s41598-025-97256-0

[8] Lubna Abdelkareim Gabralla et al., "Automated Diagnosis for Colon Cancer Diseases Using Stacking Transformer Models and Explainable Artificial Intelligence," Diagnostics, vol. 13, no. 18, pp. 2939–2939, Sep. 2023, doi: https://doi.org/10.3390/diagnostics13182939

[9] H. Chen et al., "GasHis-Transformer: A multi-scale visual transformer approach for gastric histopathological image detection," Pattern Recognition, vol. 130, p. 108827, Oct. 2022, doi: https://doi.org/10.1016/j.patcog.2022.108827

[10] W. Hu et al., "GasHisSDB: A new gastric histopathology image dataset for computer aided diagnosis of gastric cancer," Computers in Biology and Medicine, vol. 142, pp. 105207–105207, Mar. 2022, doi: https://doi.org/10.1016/j.compbiomed.2021.105207

[11] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Appendix," arXiv (Cornell University), Jan. 2015, doi: https://doi.org/10.48550/arxiv.1506.02157

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626, Oct. 2017, doi: https://doi.org/10.1109/iccv.2017.74

[13] "Navigation — HistomicsTK documentation," Github.io, 2025. https://digitalslidearchive.github.io/HistomicsTK/# (accessed Jul 01, 2025).

[14] A. Martakou-Galiatsatou, "Development of Interpretable Deep Learning Models for Histopathological Image Classification in Gastrointestinal Cancer", Diploma thesis, School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece, 2025

[15] DigitalSlideArchive, "GitHub - DigitalSlideArchive/HistomicsTK: A Python toolkit for pathology image analysis algorithms.," GitHub, Feb. 07, 2025. https://github.com/DigitalSlideArchive/HistomicsTK

[16] Anishnama, "Matthews Correlation Coefficient(MCC) one of the best metric when 2 classes are imbalanced," *Medium*, Jan. 16, 2023. https://medium.com/@anishnama20/matthews-correlation-coefficient-mcc-one-of-the-best-metric-when-2-classes-are-imbalanced-c0318ac68c21

[17]M. Zubair, M. Owais, T. Mahmood, S. Iqbal, S. M. Usman, and I. Hussain, "Enhanced gastric cancer classification and quantification

interpretable framework using digital histopathology images," Scientific Reports, vol. 14, no. 1, Sep. 2024, doi: https://doi.org/10.1038/s41598-024-73823-9

[18]M. Usai, A. Loddo, A. Perniciano, M. Atzori, and D. Ruberto, "A Comparative Analysis of Image Descriptors for Histopathological Classification of Gastric Cancer," arXiv.org, 2025, doi: https://doi.org/10.48550/arXiv.2503.17105

[19] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," The Annals of Mathematical Statistics, vol. 18, no. 1, pp. 50–60, Mar. 1947, doi: https://doi.org/10.1214/aoms/1177730491