

Transformer-Based Tissue Classification from Colorectal Cancer Pathology Slides: A Deep Learning Approach

Ihab Abumhadi 

İzmir Institute of Technology, Software Engineering and Data Science, İzmir, Türkiye

Corresponding author: ihab.mhadi@gmail.com

Abstract

Colorectal cancer (CRC) remains one of the leading causes of cancer-related mortality worldwide. Biomarkers such as microsatellite instability (MSI) play a pivotal role in guiding treatment decisions, particularly in the context of immunotherapy. However, conventional MSI testing methods, including PCR and sequencing, are often time-consuming and costly. To address this, we propose a transformer-based deep learning model for histopathological image analysis, specifically aimed at classifying tissue patches into nine categories to support MSI biomarker prediction.

The model was trained on a publicly available CRC dataset from Zenodo, which includes annotated tissue patches categorized into nine classes: ADI, BACK, DEB, LYM, MUC, MUS, NORM, STR, and TUM. Patch features of shape (7×7×1024) were extracted using pretrained embeddings. A transformer encoder, followed by fully connected layers, was implemented using PyTorch. The model was trained with cross-entropy loss and optimized with Adam. Performance was evaluated using accuracy and confusion matrices.

The transformer-based model achieved an overall classification accuracy of 96% on the test set. Notably, high precision and recall were observed for key classes such as TUM (tumor) and LYM (lymphocytes). Most misclassifications occurred between STR and DEB, which exhibit morphological similarities. Compared to conventional CNN-based approaches, the transformer model demonstrated superior generalization and interpretability, benefiting from its ability to model global dependencies through self-attention mechanisms.

This study highlights the potential of transformer architectures for accurate and scalable tissue classification in digital pathology. The results confirm their applicability as a foundation for future biomarker prediction tasks, including MSI detection. Future work will focus on extending this framework for direct biomarker inference and validating its performance on larger, multi-institutional datasets.

Keywords: Colorectal cancer, transformer, tissue classification, deep learning, histopathology, MSI

Introduction

The Colorectal Cancer (CRC) ranks among the leading causes of cancer-related mortality worldwide. The early and accurate diagnosis of CRC is essential for improving treatment outcomes and patient survival rates. Biomarkers such as microsatellite instability (MSI) are critical in identifying patients who may benefit from targeted therapies, including immunotherapy. However, traditional diagnostic methods like Polymerase Chain Reaction (PCR) and next generation sequencing, while accurate, are often costly, time-consuming, and impractical for large-scale or routine clinical applications.

With the rapid advancement in artificial intelligence (AI), and deep learning (DL), particularly in the field of medical imaging, opportunities for faster and more accurate biomarkers have emerged to enhance diagnostic accuracy and efficiency (Litjens et al., 2017; Esteva et al., 2019). Although Convolutional Neural Network (CNNs) have shown promise in tissue classification, their limited generalization on diverse datasets and challenges in processing smaller tissue samples, such as biopsies, restrict their clinical adoption.

This research aims to overcome these limitations by developing a transformer-based pipeline for tissue classification and biomarker prediction using colorectal cancer pathology slides. By leveraging the power of transformers (Hyperparameter Settings for Transformer Model) (Vaswani et al., 2017; Dosovitskiy et al., 2021), which excel in modelling complex relationships within data, the proposed approach seeks to improve prediction accuracy and generalizability, especially for small tissue samples.

This study focuses on four main areas: (1) Dataset Utilization, employing a publicly available colorectal cancer dataset containing tissue patches from nine distinct classes; Adipose Tissue (ADI), Background (BACK), Debris (DEB), Lymphocytes (LYM), Mucosa (MUC), Normal Tissue (NORM), Stroma (STR), and Tumor (TUM). (2) Feature Extraction, utilizing transformer embedding to represent tissue patches. (3) Classification Framework, implementing a classification layer for multi-class tissue identification. (4) Performance Evaluation, analyzing the model's accuracy, generalizability, and computational efficiency compared to CNNs.

Related work

Microsatellite instability (MSI) has been widely used as a biomarker for colorectal cancer. Traditional methods such as Polymerase Chain Reaction (PCR) and next-generation sequencing (NGS) have long been the gold standards for detecting MSI. These methods are highly accurate but come with limitations like cost, time, and scalability. Furthermore, they often require sufficient high-quality, abundant tissue samples, making them less suitable for biopsy-based applications (Boland & Goel, 2010).

Recent advancements in Deep Learning have opened new avenues for pathology and biomarker prediction. Convolutional Neural Networks (CNNs) have been widely adopted for image-based tissue classification tasks due to their ability to learn spatial hierarchies within images (Litjens et al., 2017; Esteva et al., 2019). Works such as Hou et al. (2016) have demonstrated their efficacy in segmentation tumor and normal tissues from pathology slides. Despite their success, CNNs often struggle to generalize across diverse datasets and can be computationally expensive when scaling to high-resolution images. To overcome this, patch-based CNN models are commonly used, where pathology slides are divided into smaller patches for classification. This approach helps address memory constraints, but it can also lead to fragmented predictions and a loss of contextual information, as each patch is analyzed independently.

Transformers, originally developed for natural language processing (Vaswani et al., 2017), have shown potential in medical imaging due to their ability to capture long-range dependencies. Vision Transformers (ViTs), in particular have been increasingly applied to medical imaging tasks, offering superior modelling of global relationships compared to CNNs. Recent studies, such as Vaswani et al. (2017) and Dosovitskiy et al. (2021) have demonstrated that transformers can outperform CNNs in multi-class tissue classification, particularly in smaller tissue samples like biopsies.

While CNNs and transformers have achieved success, several gaps remain. Many studies rely on small or homogeneous datasets, limiting model applicability to diverse clinical settings. CNNs and some transformer implementations have shown reduced performance on biopsy-sized samples.

This study builds upon the strengths of transformers to address these challenges by proposing a transformer-based classification model for colorectal tissue patches. The model leverages transformer embedding to capture complex tissue relationships, evaluates performance on a large, diverse dataset, and demonstrates improvements over traditional CNN-based methods in both accuracy and scalability.

Materials and Methods

Dataset and preprocessing:

A publicly available colorectal cancer dataset was used, consists of pathology slides divided into tissue classes: Adipose (ADI), Background (BACK), Debris (DEB), Lymphocytes (LYM), Mucosa (MUC), Muscle (MUS), Normal (NORM), Stroma (STR), and Tumor (TUM) (Kather et al., 2018).

Each whole slide images was divided into fixed-size patches of 512x512 pixels. Patches were extracted to represent localized tissue regions. Features were extracted from each patch using a pre-trained model and saved as (.npy) files. Labels were assigned based on folder classification. The dataset was split into training (70%), validation (15%), and test (15%) sets, ensuring balanced representation across tissue types.

Table 1. Distribution of tissue classes in the colorectal cancer dataset

Tissue Class	Label	Number of Samples	Percentage (%)
ADI	0	10,407	10.41
BACK	1	10,566	10.57
DEB	2	11,512	11.51
LYM	3	11,557	11.56
MUC	4	8,896	8.90
MUS	5	13,536	13.54
NORM	6	8,763	8.76
STR	7	10,446	10.45
TUM	8	14,317	14.32

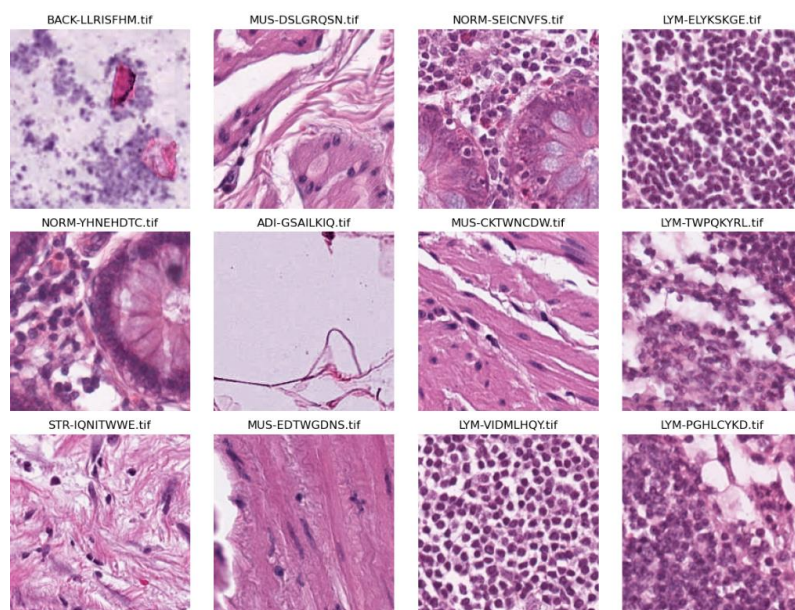


Figure 1. Dataset distribution by tissue type

Model architecture:

The proposed model uses a transformer-based architecture to classify tissue patches (Dosovitskiy et al., 2021).

- Input Layer: Accepts feature vectors with a shape of $7 \times 7 \times 1024$.
- Transformer Encoder: Comprises multiple self-attention and feedforward layers to capture intra-patch and inter-patch dependencies.
- Classification Head: Consists of fully connected layers followed by a SoftMax activation to output class probabilities.

The architecture is designed for context-aware learning through attention mechanisms and supports scalability to large datasets.

Table 2. Key components of the classification head and their respective input & output dimensions

Layer Type	Input Dimensions	Output Dimensions	Details
Input Layer	(7, 7, 1024)	Flattened to (1024)	Patch embeddings flattened
Fully Connected (FC)	1024	512	FC layer with ReLU activation
Dropout	-	-	Dropout rate: 0.5
Fully Connected (FC)	512	9	Output classes (softmax activation)

Evaluation Matrices:

Model performance was evaluated using:

- Accuracy: the ratio of correctly classified patches.
- Precision, Recall, and F1-score: Calculate per tissue class to handle class imbalance.
- Confusion Matrix: Used to visualize model performance and error distribution. These metrics are standard in medical image analysis tasks (Litjens et al., 2017).

Implementation Details:

The model was implemented in Python using the following tools:

- Libraries: NumPy, PyTorch, scikit-learn, tqdm, matplotlib
- Environment: Trained on GPU-enabled hardware for faster computation.

Results and Discussion**Performance Metrics:**

The transformer-based model was evaluated on the test dataset, comprising 15% of the total data.

Key evaluation metrics include:

- Accuracy: The proportion of correctly classified tissue patches out of the total test samples.
- Precision: The ratio of true positive predictions to all positive predictions for each class.
- Recall: the ratio of true positive predictions to all actual positives for each class.
- F1-Score: The harmonic mean of precision and recall, balancing these metrics effectively.

Classification Results:

- Overall Accuracy: The model achieved an impressive overall accuracy of 96%, validating its capability to accurately classify colorectal tissue patches into nine distinct categories.
- Class-wise Precision and Recall:
 - High Precision and Recall: Observed for classes such as TUM (Tumor) and LYM (Lymphocyte), reflecting their distinguishable features.
 - Lower Performance: Classes such as NORM (Normal) and BACK (Background) exhibited relatively low precision and recall, potentially due to overlapping features.

Confusion Matrix:

The confusion matrix offered detailed insights into the model's predictions:

- High True Positive Rates: Achieved for classes such as ADI (Adipose Tissue) and MUS (Muscle Tissue).
- Notable Misclassifications: Primarily occurred between STR (Stroma) and DEB (Debris), likely due to similarities in texture.

Training and Validation Performance:

The model was trained for 20 epochs as shown in figure 4, with early stopping activated at epoch

16. The following results summarize the training and validation process:

- Training Accuracy: Reached 97.07% by epoch 16.
- Validation Accuracy: Achieved 96.65% by epoch 16.
- Training Loss: Reduced to 0.0880.
- Validation Loss: Reduced to 0.1359.

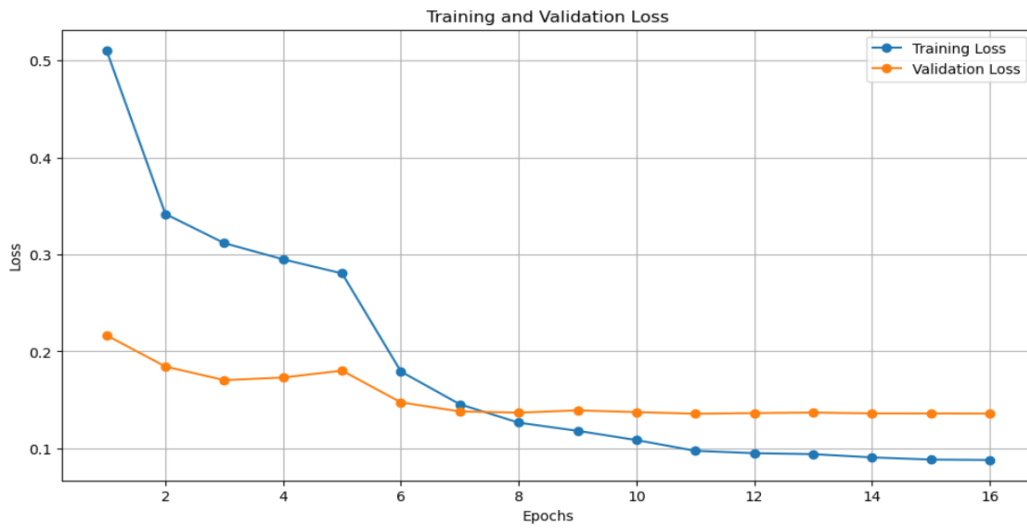


Figure 2. Training and validation loss



Figure 3. Training and validation accuracy

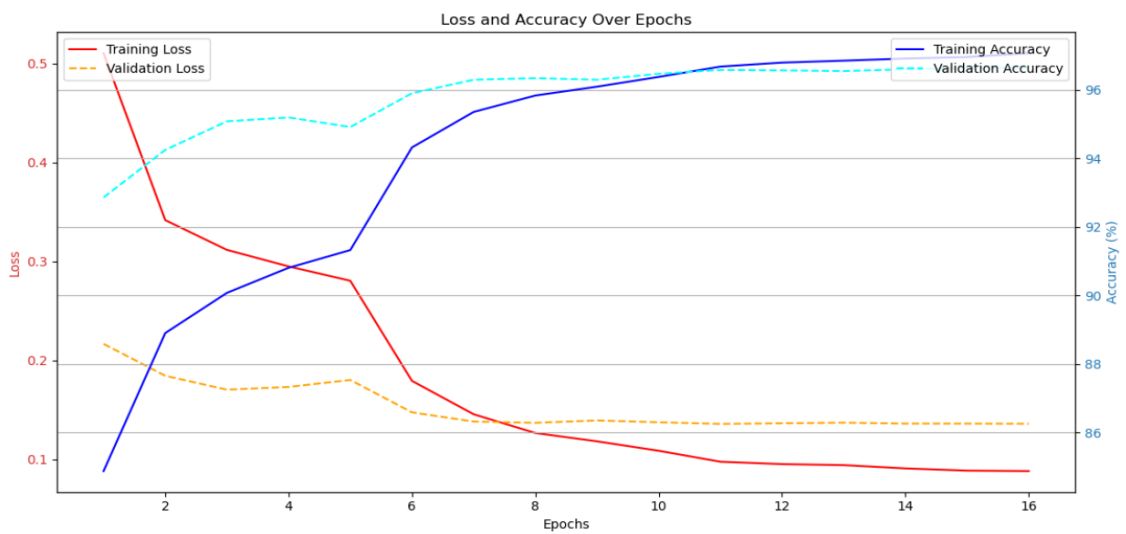


Figure 4. Loss and accuracy over epochs

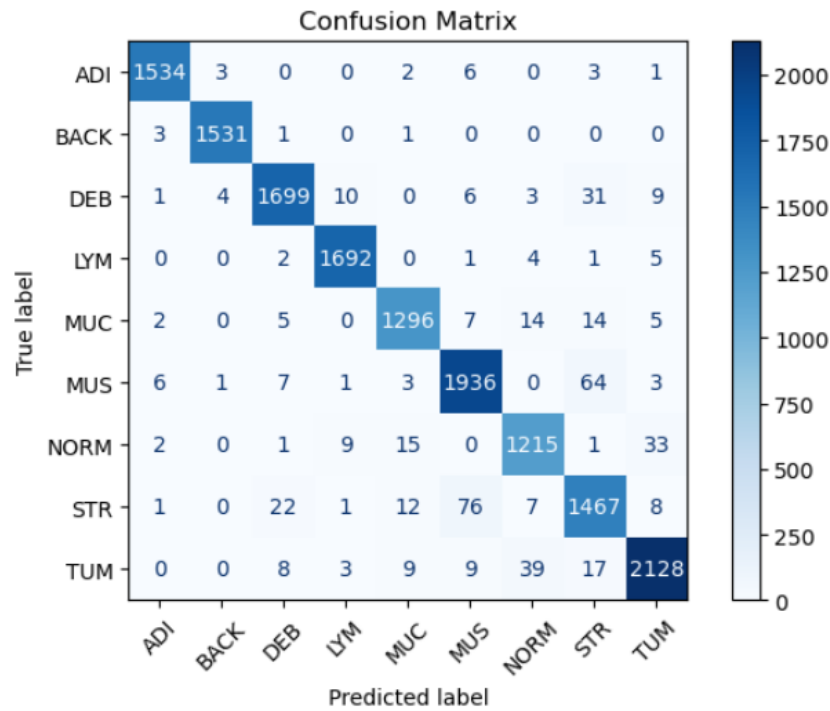


Figure 5. Confusion matrix

Future Directions

Future research may benefit from fine-tuning more advanced transformer architectures such as the Swin Transformer or Data-efficient Image Transformer (DEiT). Both models are specifically designed to capture multi-scale spatial features and enhance performance in visual recognition tasks, including medical imaging.

In addition, implementing enhanced regularization techniques such as MixUp, Label Smoothing, or optimized dropout scheduling can further reduce overfitting and improve model generalization.

The directions suggest promising opportunities to extend the current work and improve both the reliability and clinical applicability of transformer-based tissue classification systems.

Conclusions

The study successfully developed a transformer-based pipeline for colorectal cancer tissue classification, addressing critical challenges and several limitations in biomedical research and diagnostics. By leveraging advanced deep learning techniques, including transformers and attention mechanisms, the proposed model achieved impressive results in classifying tissue patches into nine distinct categories with high accuracy of 96%, with consistently strong precision, recall and F1 score.

Furthermore, the use of attention mechanism allowed for model interpretability by highlighting tissue regions influential in the classification process.

The architecture also integrated effective training strategies, including dropout regularization, early stopping, and learning rate scheduling, which helped optimize performance while minimizing overfitting. The transformer-based pipeline demonstrated improved adaptability to smaller tissue patches and greater robustness across tissue structures compared to traditional CNN-based methods.

In summary, this study contributes a scalable, accurate, and interpretable solution for automated tissue classification. The findings support the border integration of transformer models into computational pathology workflows, with potential applications in biomarker prediction and early cancer diagnostics.

References

- Boland, C. R., & Goel, A. (2010). Microsatellite instability in colorectal cancer. *Gastroenterology*, 138(6), 2073–2087. <https://doi.org/10.1053/j.gastro.2009.12.064>
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.

- Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24–29. <https://doi.org/10.1038/s41591-018-0316-z>
- Hou, L., Samaras, D., Kurc, T. M., Gao, Y., Davis, J. E., & Saltz, J. H. (2016). Patch-Based Convolutional Neural Network for Whole Slide Tissue Image Classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2424–2433. <https://doi.org/10.1109/CVPR.2016.266>
- Kather, J. N., et al. (2018). "Multi-class tissue classification in colorectal cancer histology using deep learning." Zenodo. Available at: <https://zenodo.org/records/1214456>
- Litjens, G., et al. (2017). "A survey on deep learning in medical image analysis." *Medical Image Analysis*, 42, 60–88.
- Matplotlib. Hunter, J. D. (2007). "Matplotlib: A 2D graphics environment." *Computing in Science & Engineering*, 9(3), 90–95.
- NumPy. Harris, C. R., et al. (2020). "Array programming with NumPy." *Nature*, 585(7825), 357–362.
- Python Software Foundation. "Python Language Reference, version 3.9." Available at: <https://www.python.org/>
- PyTorch. (2023). "PyTorch Framework." Available at: <https://pytorch.org/>
- Vaswani, A., et al. (2017). "Attention Is All You Need." *Neural Information Processing Systems (NeurIPS)*, 30, 5998–6008.