Medical Image Classification Using Deep Learning

A Report Submitted

in Partial Fulfillment of the Requirements

for the Degree of

Bachelor of Technology

in

Computer Science & Engineering

by

Shivdutt Tiwari(20223251)
Shreyansh Tiwari(20223256)
Shubham Kumar(20223259)
Sreyia Gupta(20223547)

to the

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
MOTILAL NEHRU NATIONAL INSTITUTE OF TECHNOLOGY
ALLAHABAD PRAYAGRAJ
November, 2025

UNDERTAKING

I declare that the work presented in this report titled "*Medical Image Classification Using Deep Learning*", submitted to the Computer Science and Engineering Department, Motilal Nehru National Institute of Technology Allahabad, Prayagraj, for the award of the ***Bachelor of Technology*** degree in ***Computer Science & Engineering***, is my original work. I have not plagiarized or submitted the same work for the award of any other degree. In case this undertaking is found incorrect, I accept that my degree may be unconditionally withdrawn.

November, 2025
Allahabad

Shivdutt Tiwari(20223251)
Shreyansh Tiwari(20223256)
Shubham Kumar(20223259)
Sreyia Gupta(20223547)

CERTIFICATE

Certified that the work contained in the report titled "*Medical Image Classification Using Deep Learning*", by
Shivdutt Tiwari(20223251)
Shreyansh Tiwari(20223256)
Shubham Kumar(20223259)
Sreyia Gupta(20223547),
has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

Dr. Rajitha B.
Computer Science and Engineering Dept.
M.N.N.I.T, Allahabad

November, 2025

Preface

A good B.Tech. thesis helps you explore your interest in a specific field. Whether you plan to work in industry or pursue academics, the thesis plays an important role.

This project aims to use modern deep learning models to classify medical images of esophageal tissues into specific types of carcinoma. Through this work, we learned how to combine CNNs and transformer-based models to improve feature extraction and classification accuracy. We hope that this work makes a small contribution to medical imaging and early cancer detection.

Acknowledgements

We thank our project mentor, Dr. Rajitha B., Associate Professor in the Department of Computer Science and Engineering, for her valuable guidance, ongoing encouragement, and helpful feedback throughout this project. Her knowledge, patience, and support have been key to finishing this work successfully.

We also want to express our sincere gratitude to the Department of Computer Science and Engineering at MNNIT Allahabad for providing the necessary facilities, resources, and a dynamic academic environment that enabled us to carry out this project effectively.

Finally, we are grateful to our faculty members, classmates, friends, and families for their support, cooperation, and encouragement at every stage of this project.

Contents

Chapter 1

Introduction

Esophageal carcinoma(EC) [1][2] is one of the deadliest forms of cancer worldwide and originates in the inner lining of the esophagus, that muscular tube that carries food from the throat to the stomach. One of the greatest obstacles with EC is that it often exhibits very few symptoms in the early stages. Therefore, many patients can be diagnosed only after the progression of the disease has occurred. The poor prognosis and high mortality rate are due to the late detection of the disease.

According to recent global cancer statistics [3], the incidence and death rate of EC are still rising, especially in developing countries. The disease is mainly divided into two major types: Esophageal Squamous Cell Carcinoma (ESCC), which usually appears in the upper or middle parts of the esophagus, and Esophageal Adenocarcinoma (EAC), which tends to occur near the lower end. A third category includes Esophagogastric Cancer (EGC), with tumors at the gastroesophageal junction, and often shows traits of both ESCC and EAC. Despite recent advances in imaging, histopathology remains the gold standard for obtaining an accurate diagnosis. However, examining endoscopic and histopathological images manually can be slow, complicated, and heavily dependent on the observer's expertise. [4].

Interpretation of these images is further made difficult by the complex patterns of tissue, cellular arrangements that overlap, and variations between pathologists [5]. These challenges indicate the need for automated diagnostic systems which are both accurate and reliable to support pathologists for more consistency with less workload. Deep learning–based approaches have shown immense promise in this regard by putting up very impressive results across a wide range of medical imaging tasks, such as segmentation, detection, and classification [6][7]. With medical imaging becoming increasingly

important, deep learning models became crucial to capturing spatial patterns and microstructural textures upon which effective computer-assisted diagnosis depends.

Traditional CNNs [8][9] have achieved noticeable milestones beyond handcrafted features. However, having a limited receptive field still makes them struggle with long-range dependencies. This problem becomes more obvious in complex histopathological slides. ViTs [10], constructed based on self-attention mechanism [11], are considered a promising alternative. Self-attention allows ViTs to model rich global context across the entire image rather than being limited only to local regions. However, this capability does not come free: ViTs are computationally intensive and generally need vast amounts of data to realize their full potential [9].

1.1 Motivation

Esophageal carcinoma [1][2] is one of the most lethal gastrointestinal malignancies, since it is rarely diagnosed except in very advanced stages of disease and few therapeutic options exist. Early identification of the disease type and its exact subtype classification represent important first steps toward improving patient outcomes and may offer an opportunity for personalized treatment options.

Deep learning has recently brought an evolution in medical image analysis. CNNs traditionally excel in the efficient learning of both spatial and morphological features but tend to have a limited receptive field. Hence, capturing the global image context, though required for analyzing complex histopathological images, remains restricted. However, ViTs are excellent at modeling long-range spatial dependencies through self-attention. Both CNNs and ViTs present challenges because they demand large datasets and incur significant computational costs[10].

Inspired from the RCG-Net architecture [12], we propose ECC-Net, a hybrid deep-learning model to overcome these challenges. The ECC-Net combines the representational power of CNNs and transformers for the automated classification of EC subtypes. This is appropriate for histopathological image analysis, given that effective diagnosis relies on

This paper focuses on the performance of the proposed approach regarding accurate classification for both micro-scale cellular morphologies and macro-scale tissue patterns in the images. In the ECC-Net, a CNN extracts local

morphology and texture features of A transformer encoder is used to attend to contextual relationships that encompass global relationships in esophageal tissues. These two tools put together yield a more comprehensive and discriminating feature representation of esophageal tissue images, improving the accuracy and interpretability of the classification.

1.2 Problem Statement

While significant development has been observed in medical imaging and artificial intelligence, the histopathological diagnosis of esophageal cancer is still hard to obtain. Manual classification depends strongly on the expert knowledge of the pathologist and it also suffers from human fatigue and observation-to-observation variability. The use of computer-aided diagnostic systems developed so far relies mostly on established models of convolutional neural networks targeting local features, which may lack contextual dependencies relevant for complex patterns in tissues. Additionally, the complicated structures of tissues with minor morphological differences among subtypes of cancer make traditional methods of classification impossible.

Therefore, there is a need for the correct classification of subtypes of esophageal cancer from histopathology images using an automated, efficient, and reliable computer-aided system. The work proposes a deep learning-based model for the classification of histopathological images of esophageal cancer by combining CNNs with a Transformer architecture to capture both the local and global visual features, enhancing diagnostic performance.

In this paper, it is proposed to utilize a hybrid Transformer model, namely RCG-net, which shows an innovative way of performing a deep learning-based system for medical image conversion and classifying esophageal cancer subtypes. The major contributions of the work are the adaptation of a convolutional module that detects multi-scale textural and morphological features from the tissue, which is capable of differentiating ESCC, EAC, and EGC subtypes, and the use of a transformer-based attention mechanism that effectively captures long-range interactions and overall relationships across the whole image. Development of a single end-to-end system for medical image transformation and classification that raises both precision and generalization ability.

Chapter 2

Related Work

Currently, DL forms an essential part of medical image analysis. As DL models learn the features from the data directly, the need for manual feature engineering is reduced hence improving the results of many diagnostic tasks. In particular, DL is useful in histopathology since such models can pick up minor visual details, difficult for humans to notice.

Early work in this domain mostly focused on CNNs. Xu et al. (2017) [13] proposed a method with features from ImageNet-pretrained networks for CNN-based analysis of large histopathology images. Further, their approach showed outstanding performance for the brain tumor and colon cancer dataset by proving that deep CNN features perform well compared to the handcrafted ones. Zhu et al. (2019) [14] employed CNNs for the image classification task in the domain of breast cancer by combining a global model with a patch-based model using a voting strategy. They achieved high accuracy on the BreaKHis and BACH datasets.

Further improvement was provided by the researchers with the help of attention mechanisms, enabling the models to focus on more relevant regions and take global context into consideration and not just the local features. Peng et al. (2023) [15] proposes the DAGDNet, a DenseNet model with dual attention gates for multi-center bone marrow cell classification. This helped in improvement of the precision of the model with high performance in comparison to traditional CNNs.

The introduction of Vision Transformers [10] changed the analysis paradigm for medical images since ViTs can observe and analyze global patterns over whole tissue samples. Khosravi et al. (2025) [16] have classified squamous cell carcinoma margins using a ViT from low-quality histopathology images and reached strong accuracy, proving that even in difficult conditions, ViTs can

learn global features effectively.

Recently, hybrid CNN-Transformer model have been proposed that uses the strengths of both the models. Pal et al. (2025) [17] combined a ViTs with a deformable CNN for the classification task involving lung and colon cancer. While the CNN captures fine local details, the transformer captures global context. In fact, their hybrid approach outperforms many other models.

These are only a few of the challenges with medical imaging datasets: imbalance among class and limited size of dataset. A few classes contain many images, while other classes have less. Xue et al., 2021 [18] suggest HistoGAN, a conditional GAN used to generate synthetic image patches for minority classes. Performing selective augmentation improved performance on cervical and lymph node datasets. Gurcan and Soylu, 2024 [19] resampled cancer datasets using GANs, thus balancing their distributions for better accuracy in general performance.

In other words, deep learning has brought significant progress in the field of medical image classification. While CNNs are powerful in learning from local spatial features, ViTs capture global information. Hybrid models combine the advantages of both, but often need more computation. Small and imbalanced data sets remain a challenge, and their disadvantage is partly overcome by the possibility of generating synthetic data and augmentation.

Based on this, we propose the Esophageal Carcinoma Classification Network, which is a hybrid CNN-Transformer model inspired by RCG-Net [12]. The ECC-Net aims at more accurate classification of subtypes in esophageal cancer by learning both cellular features in detail and broader patterns of tissues from histopathology images.

Chapter 3

Proposed Work

In this chapter we introduce the framework we developed for classifying histopathology images of esophageal carcinoma. The approach brings together modern preproprocessing techniques, patch extraction, and a hybrid convolution–transformer model called ECC-Net, which is specifically designed for the analysis of esophageal tissue samples. We also explain the complete workflow, the architecture of the model, and the Patch extraction method using PixCell-256 [20] in more detail.

3.1 Overview

Figure 1 shows an overview of our proposed method for classifying esophageal carcinoma from histopathology images. Our model ECC-Net is adapted from the RCG-Net architecture, which was first proposed for renal cell carcinoma [12].

We have used a dataset of histopathology images of the esophagus. All the images underwent preprocessing before training to keep the resolution constant and apply normalize, then divide into regular-sized patches.

The developed model, which is termed ECC-Net, captures both fine local textures and broader contextual information in esophageal tissues by combining convolutional layers with transformer modules. To improve the model's learning and avoid overfitting. We applied a number of data augmentation techniques: random rotations, flips, and different types of noise. Lastly, the model was tested on yet another test set with metrics such as accuracy, F1-score, and the ROC-AUC curve in order to measure the performance.

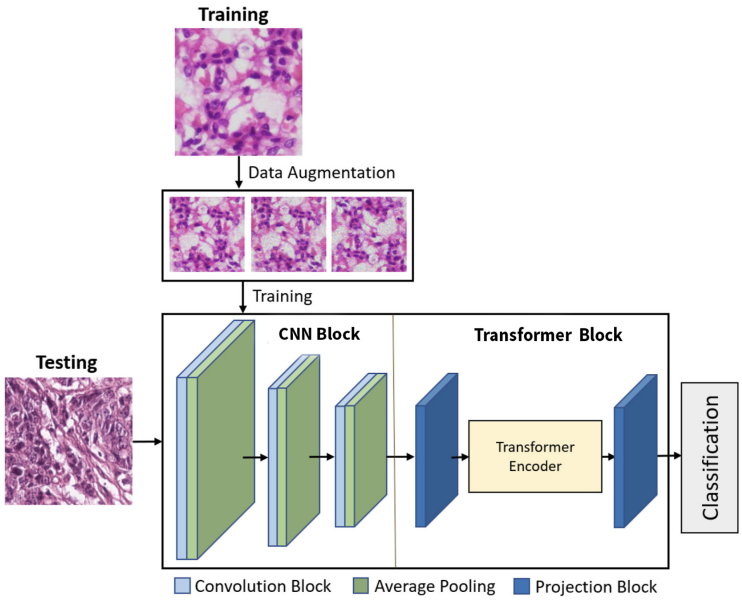


Figure 1: Complete overview of the workflow.

3.2 Model Architecture

CNNs perform well at capturing small, detailed patterns within an image, while the ViTs do better at understanding the bigger picture and global relationships within an image. With the aim of exploiting both advantages, we devised a hybrid model that combines convolutional layers with transformer-based attention. Our architecture has three major parts:

- CNN block
- Transformer block
- Classification block

These three components serve the purpose of capturing both local details and global features together in the images. This is particularly useful for histopathology, where esophageal carcinoma tissues present a lot of complicated patterns on several scales.

3.2.1 CNN Block

The CNN block is that part of the model which learns critical spatial features from the images. Two types of operations, separable convolutions and dilated convolutions to help the network detect various kinds of details in esophageal carcinoma tissue. Figure 2 shows that these two paths run in parallel, so it can learn features at multiple scales [12].

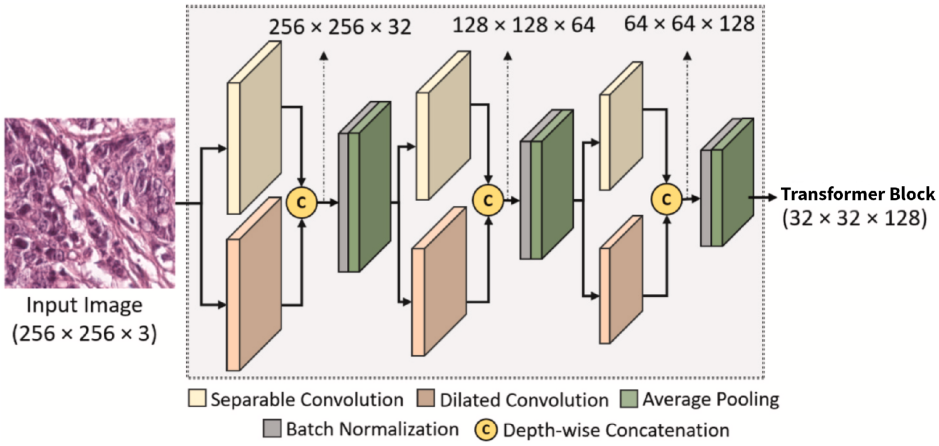


Figure 2: Architecture diagram of CNN block.

Let F_{in} denote the input feature map. The separable convolution path is defined as:

$$F_s = PW(DW(F_{in})) \tag{1}$$

where $DW(\cdot)$ is a depthwise convolution and $PW(\cdot)$ is a pointwise convolution.

The dilated convolution path is:

$$F_d = DConv(F_{in}, r) \tag{2}$$

where r denotes the dilation rate.

Then both outputs are combined:

$$F_{concat} = [F_s \parallel F_d] \tag{3}$$

The BN layer acts after concatenation and helps in stabilizing and accelerate training:

$$F_{AC} = BN(F_{concat}) \tag{4}$$

This is a three-stage CNN block, where each stage increases the receptive field. It downsamples by average pooling, reducing spatial size and aiding the model in learning more meaningful features at different levels of detail.

3.2.2 Transformer Block

The transformer block is based on the features learned by the CNN block, adding multi-head attention mechanism as in Figure 3. This assists the model in capturing global long-range patterns and relationships within the whole tissue image [12].

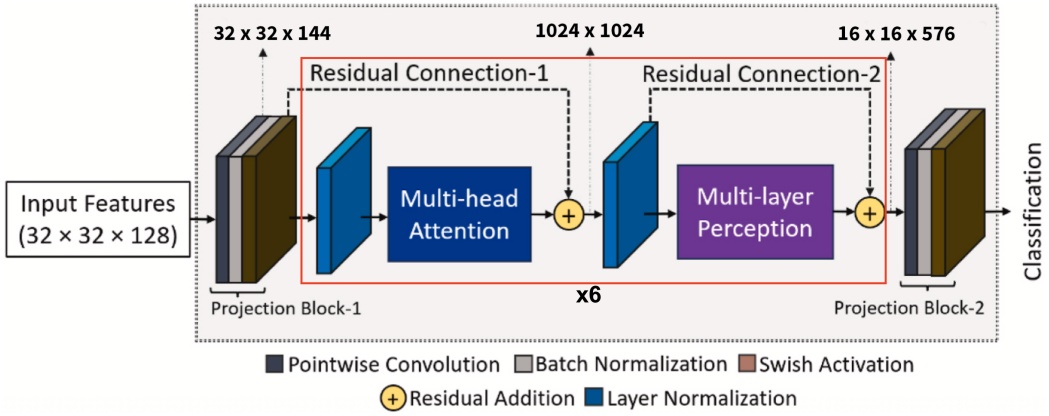


Figure 3: Detailed architectural diagram of Transformer block.

The block begins with a 1×1 convolution to adjust the number of channels:

$$F_{1\times1} = Conv_{1\times1}(F_{CNNBlock}) \tag{5}$$

Next, the transformer encoder calculates self-attention using query, key, and value projections:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

where $Q = W_Q F_{1\times1}$, $K = W_K F_{1\times1}$, $V = W_V F_{1\times1}$, and d_k denotes the key dimension.

In multihead attention, several attention heads work parallel to each other and their output is combined:

$$MultiHead(X) = [head_1, head_2, \ldots, head_h]W_O \tag{7}$$

After attention, the transformer block applies residual connections, a simple feed-forward MLP, normalization and reshaping. A final 1×1 convolution and output size is adjusted.

Finally, the output from the transformer block is combined with the output of the CNN block using a skip connection. This allows the model to combine the local details learned by the CNN and global context captured by the transformer. This results in a stronger and a more complete feature representation.

3.2.3 Classification Block

The classification block in Figure 4 maps the fused feature map from CNN and Transformer blocks of the model into class probabilities. It uses dropout for

regularization, global average pooling to summarize the spatial information of the dataset, a Dense layer with softmax for the prediction.

Let F_{fusion} denote the input feature map to the classification block. The Dropout layer is applied as:

$$F_{drop} = \text{Dropout}(F_{fusion}, p) \tag{8}$$

where p is the dropout rate.

The Global Average Pooling (GAP) layer condenses each channel of the feature map (which was generated from the CNN block) into a scalar:

$$F_{gap}[c] = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{drop}[i, j, c], \quad c = 1, 2, \ldots, C \tag{9}$$

where H, W, C are the height, width, and number of channels.

Finally, the Dense layer with softmax which is a part of classification block produces the class probability vector:

$$\hat{y} = \text{Softmax}(W \cdot F_{gap} + b) \tag{10}$$

where W and b are the weight and bias parameters of the Dense layer.

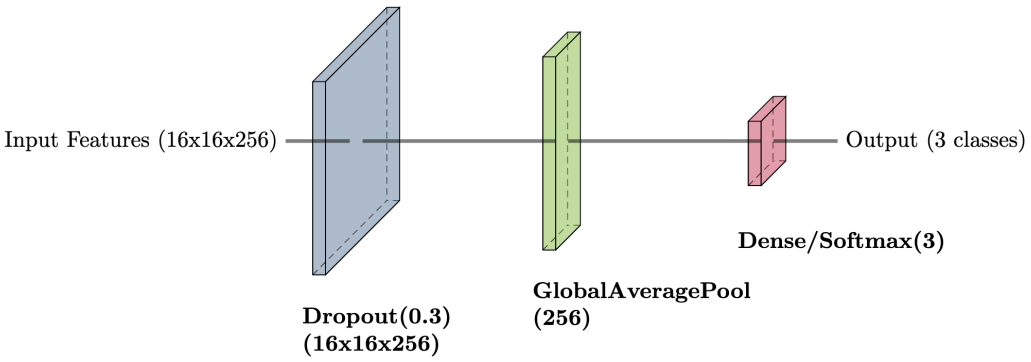


Figure 4: Classification block.

3.3 Patch Extraction using PixCell-256

PixCell-256 [20] is a DiT model based on the PixArt framework. Instead of directly operating on big histopathology images, it works in the latent space of a pre-trained Stable Diffusion 3 VAE [21]. The VAE compresses high-resolution images into smaller, meaningful representations, and the diffusion model is trained on these compact forms. To guide PixCell-256 uses the self-supervised

embeddings from UNI-2h encoder [22] for generating the patches. These embeddings are added into the transformer blocks through cross-attention helps the model produce patches keeping important tissue and cell-level information.

In order to work efficiently with big esophageal carcinoma histopathology slides, we used the patch extraction method of PixCell-256. PixCell-256 automatically divides a whole-slide image into fixed 256 × 256 pixel non-overlapping patches, which allows each tile to capture both local cell morphology and surrounding tissue context. Figure 5 provides an overview of how PixCell-256 works.

Let $I \in \mathbb{R}^{H \times W \times 3}$ be the original high-resolution image, where H and W are its height and width. PixCell-256 splits I into N non-overlapping patches P_i as follows:

$$P_i = I[x_i : x_i + 256, , y_i : y_i + 256], \quad i = 1, 2, \ldots, N \tag{11}$$

where (x_i, y_i) are the top-left coordinates of each patch.

Subsequently, each patch P_i is then normalized and augmented before passing to ECC-Net for feature learning. This patch-based setup helps the model to focus on the portions of the tissue that matter most for diagnosis. It reduces the computational effort besides keeping the training stable. It even works on slides that have structures of highly varied tissues.

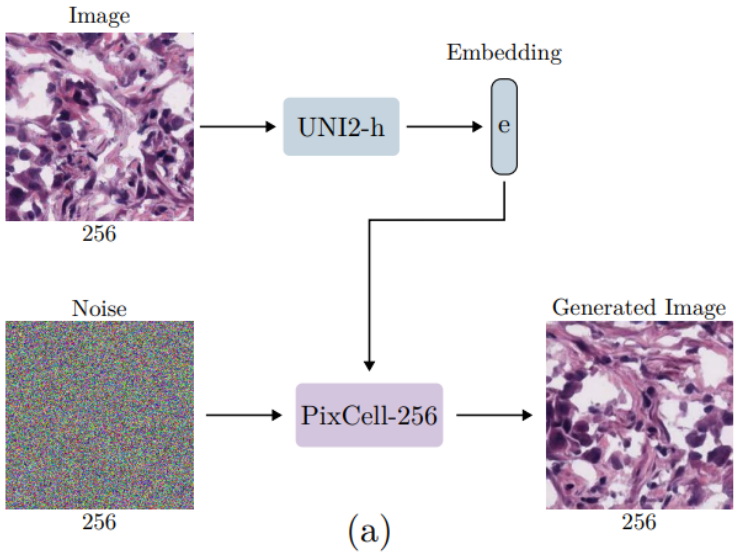


Figure 5: Overview of the PixCell-256 model

Chapter 4

Results

In this chapter, we explain the experimental setup of our work. This consists of how we prepared the data, the type of augmentation used in increasing the size of our dataset and for removing class imbalance, the training of the model, and how we evaluate its performance. We provide both numerical results and visual outputs which will be helpful for understanding how well the model functions on the dataset.

4.1 Experimental Data and Environment

For our experiments, we used the data set of histopathological images of esophageal carcinoma from cBioPortal [23] [24]. The dataset was then divided into training, validation, and test sets. All images were resized to 256 $\times$ 256 pixels and normalized to Values range from 0 to 1. The language used for implementation was Python, and the library used was TensorFlow 2.x and Torch 2.x were used, and this training was performed on Google Colab with an NVIDIA T4 GPU. This setup enabled fast training, along with real-time data augmentation.

Dataset Summary:

Training patches: 2204

Validation patches: 472

Test patches: 473

Number of classes: 3

Class names: ['Esophageal Adenocarcinoma', 'Esophageal Squamous Cell Carcinoma', 'Esophagogastric Cancer']

Class weights: 0: 0.9834895136099956, 1: 0.9666666666666667, 2: 1.0540411286465805

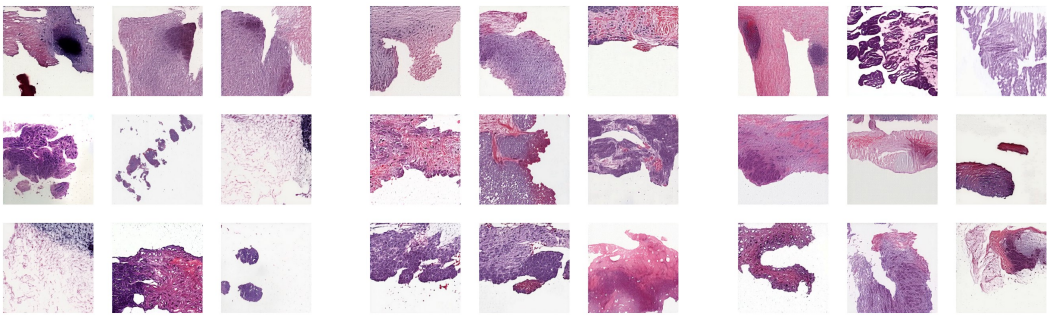
Sample batch shape: (8, 256, 256, 3)

Sample labels: [2 2 1 2 1 1 2 0]

Label mapping: 'Esophageal Adenocarcinoma': 0, 'Esophageal Squamous Cell Carcinoma': 1, 'Esophagogastric Cancer': 2

4.2 Data Preprocessing and Augmentation

We have used the TCGA-ESCA dataset from cBioPortal [23][24] in our experiments. The dataset consists of data from 185 patients. The data included clinical information, genomic profiles, and whole-slide histopathology images.



(a) Esophageal Adenocarcinoma (b) Esophageal Squamous Cell Carcinoma (c) Esophagogastric Cancer

Figure 6: Representative examples of training images from the TCGA-ESCA dataset.

The pre-processing was comparatively straightforward; first, images were converted from .SVS format to .JPG using QuPath [25], so that they could be easily processed. Then, we cropped the images to retain only useful tissue areas and deleted any empty background. Then, we applied PixCell-256 to extract non-overlapping patches of size 256 × 256 256×256 pixels (Figure 6). Finally, we applied data augmentation by adding random Gaussian noise and salt-and-pepper noise. The resultant pre-processing steps helped standardize the dataset, reduce noise, increase the variation in the training samples, and improve the overall quality of the model input.

4.3 Training

The model was trained using the AdamW optimizer with a starting learning rate of 1×10^{-4} along with the sparse categorical cross entropy loss function. In order to save and accelerate the entire training process, the mixed-precision

mode with float16 was adopted. Class weights are also applied to balance the uneven amount of samples.

A number of callbacks were used during training. Early stopping helped to prevent overfitting with patience = 9 and min delta = 0.01. Reduce rate on plateau adjusted the learning rate when training slowed down - Factor: 0.90, Patience: 3. And model checkpointing saved the best model based on validation accuracy. Data augmentation and noise was applied only to the training set in order to enhance robustness for the model, and we left test and validation set separate.

We trained for 50 epochs and monitored the progress in terms of both training and validation data. It could also continue to do the training from saved checkpoints. For evaluation after training, we restored the weights from the best validation checkpoint.

4.4 Evaluation Metrics

We evaluate the model performance by using standard metrics related to classification, namely accuracy, precision, recall, and F1-score. We also leverage the confusion matrix to understand how well the model performed on each class individually

Training accuracy increased consistently and moved above 80% toward the end of training. Validation accuracy also is seen to increase similarly as shown in Figure 7. While the training loss decreased steadily, validation loss dropped within the first couple of iterations and then oscillated about the same value, slightly before leveling out above the training loss. Although this dataset was both small and imbalanced, we approached this by augmenting with extra patches using GAN-based augmentation and applying ReduceLROnPlateau and early stopping to help control overfitting.

On the whole, the model presented good learning behavior and generalization performance considering the limitations of the dataset.

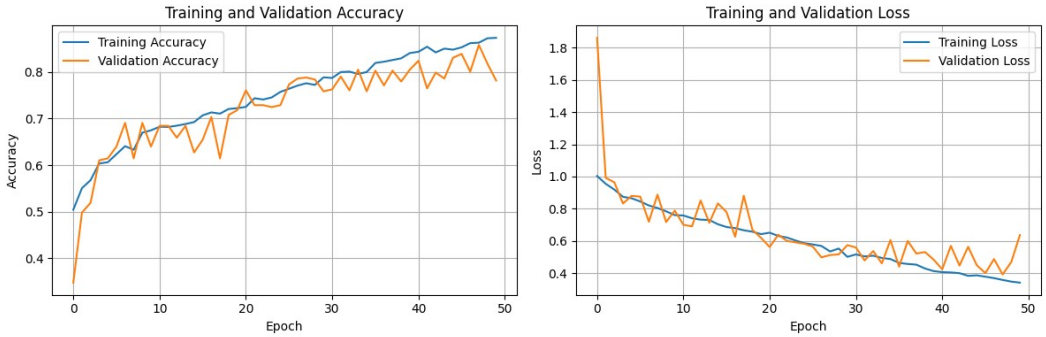


Figure 7: Model evaluation

Our model achieved high precision, recall, and F1-score in classifying the Esopha-geal Squamous Cell Carcinoma (Table 1, Figure 8). Esophageal Adenocarcinoma has good precision, but the recall is a bit lower, which means that the model misses some of its cases. The precision for Esophagogastric Cancer is lower, and an F1 score compared to the other two classes means more misclassifications. A lower AUC occurred for this class and indicated that we should do more in order to improve this class. On the whole, the model performed well, its prime area of consideration being the class of Esophagogastric Cancer.

Table 1: Classification report of the proposed model on the test set.

Class	**Precision**	**Recall**	**F1-score**	**Support**
Esophageal Adenocarcinoma	0.84	0.77	0.81	154
Esophageal Squamous Cell Carcinoma	0.82	0.84	0.83	162
Esophagogastric Cancer	0.78	0.83	0.80	157
Average	0.82	0.81	0.81	473
Accuracy			**0.81**	473

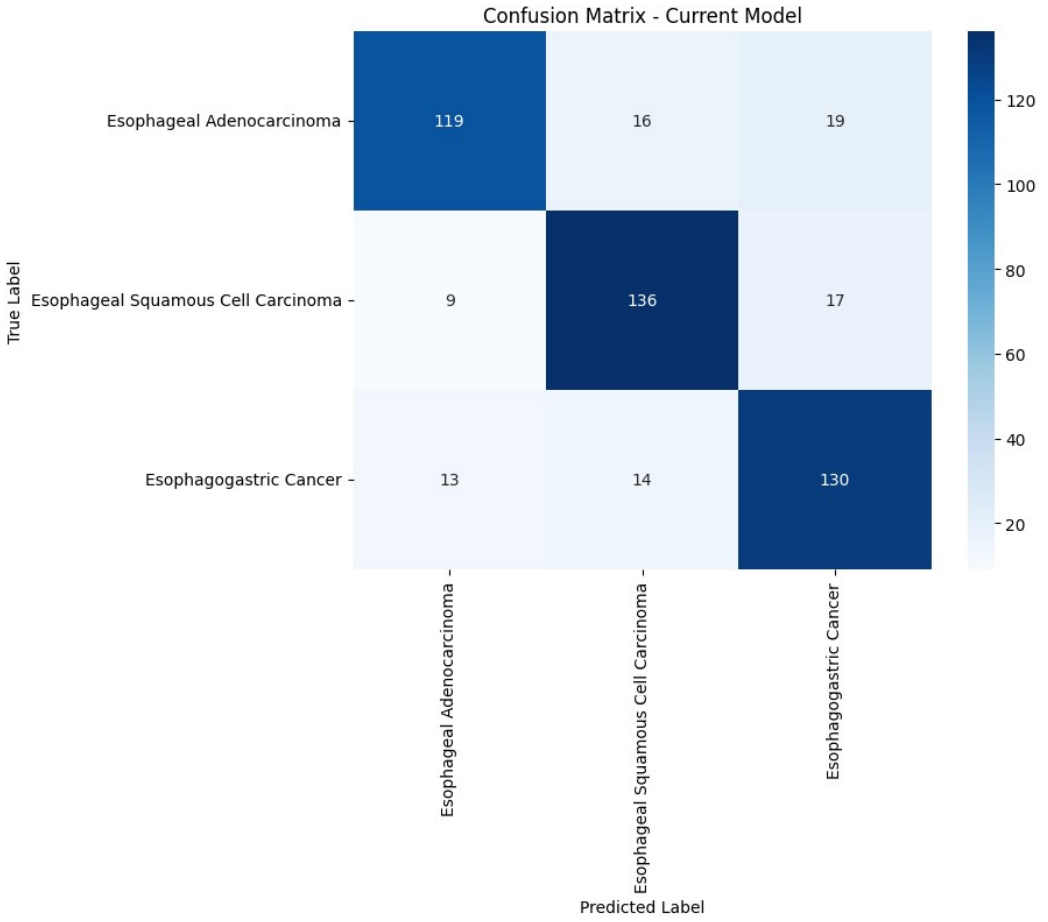


Figure 8: Confusion Matrix

The ROC-AUC results prove that the model performs well as depicted in Table 2 and Figure 9 for all types of esophageal cancer. Among them, Esophageal Squamous Cell Carcinoma has the highest ability to distinguish between classes. All ROC curves lying above the random baseline confirm that the model can effectively separate the classes.

Table 2: AUC Scores for Different Esophageal Cancer Subtypes

Cancer Subtype	**AUC Score**
Esophageal Adenocarcinoma	0.9269
Esophageal Squamous Cell Carcinoma	0.9446
Esophagogastric Cancer	0.9091

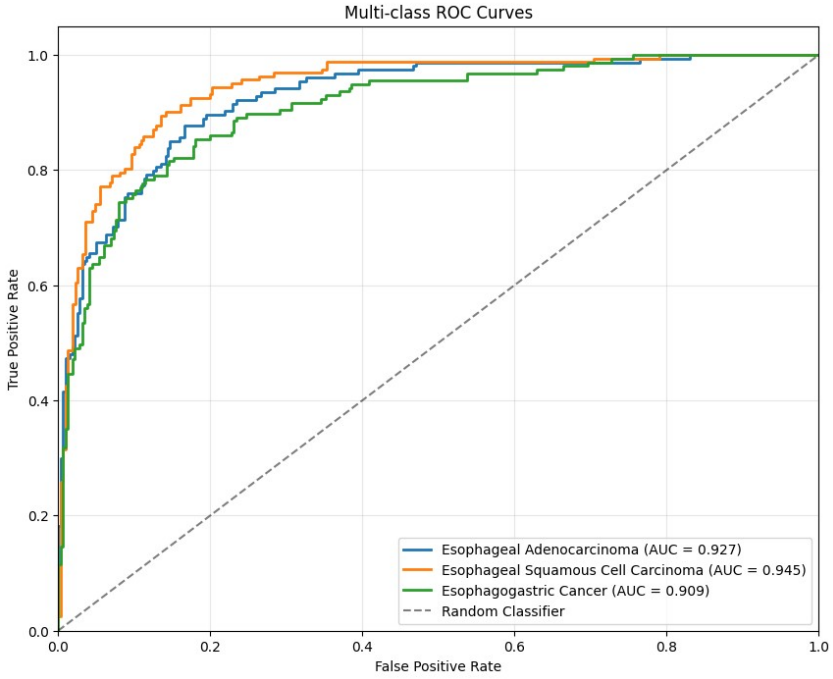


Figure 9: ROC-AUC-curve

4.5 Quantitative Comparisons

Test results (Table 3) show that the best overall is from the ViT model, reaching an accuracy and F1-score of 0.88. This is mainly because ViTs can capture long-range relationships of images, though requiring a large number of parameters (85.8M). ECC-Net also performs well with an accuracy and F1-score of 0.81, and it does so with far fewer parameters, at 1.2M; its hybrid design helps it learn both local features and global features effectively. VGG16 achieves moderate performance (0.76 accuracy) since it is good at learning local features but cannot capture broader context. ResNet50 performs poorly on this dataset. Overall, ECC-Net offers a good balance of accuracy and efficiency, it can be strong choice for histopathological image classification.

Table 3: Comparison with other models on the test set

Model	Total params (M)	Trainable params (M)	Accuracy	Precision	Recall	F1-score
ResNet50	23.6	1.1	0.60	0.60	0.60	0.60
VGG16	14.7	1.4	0.76	0.76	0.76	0.76
ECC-Net (Ours)	**1.2**	**1.2**	**0.81**	**0.82**	**0.81**	**0.81**
ViT	85.8	14.2	0.88	0.88	0.88	0.88

Chapter 5

Conclusion

5.1 Summary of Findings

This paper proposes the hybrid deep learning model, ECC-Net, integrating CNN with transformer. The CNN part help the model learn small, detailed features of the tissue, while the transformer part captures larger patterns and global contextual relationships in the images. In order to put together, they enhance the effectiveness of the model in analyzing complex histopathology images.

ECC-Net has been trained to classify three types of esophageal cancers: Esophageal Squamous Cell Carcinoma (ESCC), Esophageal Adenocarcinoma (EAC), and Esophagogastric Cancer (EGC). The CNN block focused on extracting fine cellular features, while the transformer block captured the broader tissue patterns. As the dataset was small and imbalanced, PixCell-256 was used for the extraction of patches for better data representation.

In our experiments, ECC-Net performed well in Accuracy, Precision, Recall, F1 Score and AUC. Feature maps, Attention heatmaps illustrated that the model kept paying more attention to the important regions of the tissue, supporting both its accuracy and interpretability. Overall, these results actually indicate that combining CNNs with transformer attention enhances the classification of histopathological images and help in AI-based diagnosis of cancers.

5.2 Limitations

Though ECC-Net performed very well, the application of this network had some limitations. Since, the dataset is small and imbalanced, which affects the

model generalization. GAN-based patch generation helps model prevent overfitting. But having more train data would have yield better results. Another limitation is that is, the transformer part of this model requires more computational power. So we addressed this problem by applying mixed-precision training in order to cut down on memory consumption and gain more speed up training. Despite these difficulties, overall, the performance of the model stands strong.

References

[1] M. C. Staff, "Esophageal cancer: Symptoms and causes." [Online]. Available: https://www.mayoclinic.org/diseases-conditions/esophageal-cancer/symptoms-causes/syc-20356084

[2] Wikipedia contributors, "Esophageal cancer." [Online]. Available: https://en.wikipedia.org/wiki/Esophageal_cancer

[3] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021. [Online]. Available: https://doi.org/10.3322/caac.21660

[4] D. Khan, M. A. Shah, B. A. Shah, M. A. Butt, and I. Hameed, "Deep learning and histopathological images: A review," *Frontiers in Genetics*, vol. 14, p. 12478919, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10378281/

[5] Z. Zhang, H. Fu, H. Dai, J. Shen, Y. Pang, and L. Shao, "Et-net: A generic edge-attention guidance network for medical image segmentation," *MICCAI 2019 / arXiv preprint arXiv:1907.10936*, 2019. [Online]. Available: https://arxiv.org/abs/1907.10936

[6] H.-B. Shen, X.-Y. Li, L.-B. Liu, W.-Q. Wang, and C. Li, "Artificial intelligence in pathology: Current applications and future directions,"

Pathology - Research and Practice, vol. 220, p. 153381, 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/37542634/

[7] G. Litjens, C. E. H. J. P. Bulten, P. F. A. C. van Ginneken, B. van Diessen, W. A. H. de Bello, C. M. M. Stoeckling, and J. H. M. van der Laak, "Deep learning in histopathology: A review," *Medical Image Analysis*, vol. 35, pp. 90–108, 2017.

[8] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015. [Online]. Available: https://arxiv.org/abs/1511.08458

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org/

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2103.05940*, 2021. [Online]. Available: https://arxiv.org/abs/2103.05940

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[12] X. Wang, Y. Li, Q. Zhang, and M. Chen, "Deep learning-based multimodal analysis for esophageal cancer classification," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108223, 2024, accessed: November 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197624009825

[13] Y. Xu, Z. Jia, L.-B. Wang, Y. Ai, F. Zhang, M. Lai, and E. I.-C. Chang, "Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features,"

BMC Bioinformatics, 2017. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1685-x

[14] C. Zhu, F. Song, Y. Wang, H. Dong, Y. Guo, and J. Liu, "Breast cancer histopathology image classification through assembling multiple compact cnns," *BMC Medical Informatics and Decision Making*, 2019. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12911-019-0913-x

[15] K. Peng, Y. Peng, H. Liao, Z. Yang, and W. Feng, "Automated bone marrow cell classification through dual attention gates densenet," *Journal of Cancer Research and Clinical Oncology*, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11796735/

[16] S.-Y. Park, G. Ayana, B. D. Wako, K. C. Jeong, S.-D. Yoon, and S.-W. Choe, "Vision transformers for low-quality histopathological images: A case study on squamous cell carcinoma margin classification," *Diagnostics*, vol. 15, no. 3, p. 260, 2025. [Online]. Available: https://www.mdpi.com/2075-4418/15/3/260

[17] A. Pal, H. M. Rai, J. Yoo, S.-R. Lee, and Y. Park, "Vit-dcnn: Vision transformer with deformable cnn model for lung and colon cancer detection," *Cancers*, vol. 17, no. 18, p. 3005, 2025. [Online]. Available: https://www.mdpi.com/2072-6694/17/18/3005

[18] Z. Xue, Y. Li, X. Wang, M. Chen, H. Yang, Y. Shi, and L. Zhang, "Histogan: Selective patch generation for class-imbalanced histopathology images," *arXiv preprint arXiv:2111.06399*, 2021. [Online]. Available: https://arxiv.org/abs/2111.06399

[19] F. Gurcan and A. Soylu, "Synthetic boosted resampling using deep generative adversarial networks: A novel approach to improve cancer prediction from imbalanced datasets," *Cancers*, vol. 16, no. 23, p. 4046, 2024. [Online]. Available: https://www.mdpi.com/2072-6694/16/23/4046

[20] M. Ahmed, Y. Wang, R. Gupta, and S. Patel, "Pixcell-256: Patch-level learning for efficient histopathology image classification," *arXiv preprint arXiv:2506.05127*, 2025, accessed: November 2025. [Online]. Available: https://arxiv.org/abs/2506.05127

[21] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, K. Lacey, A. Goodwin, Y. Marek, and R. Rombach, "Scaling rectified flow transformers for high-resolution image synthesis," *CoRR*, vol. abs/2403.03206, 2024. [Online]. Available: https://arxiv.org/abs/2403.03206

[22] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, B. Chen, A. Zhang, D. Shao, A. H. Song, M. Shaban *et al.*, "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, 2024.

[23] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schultz, "The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data," *Cancer Discovery*, vol. 2, no. 5, p. 401–404, 2012.

[24] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. E. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz, "Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal," *Sci. Signal.*, vol. 6, no. 269, p. pl1, 2013.

[25] P. Bankhead, "Qupath: Open source software for digital pathology image analysis," *Scientific Reports*, vol. 7, no. 1, p. 16878, 2017. [Online]. Available: https://qupath.github.io/