IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

Colon Cancer Detection using Vision Transformers and Explainable AI

Aneesh Jayan Prabhu

Abstract: *Colon cancer is a type of cancer in the large intestine. It usually starts from noncancerous growths called polyps. Symptoms include changes in bowel habits, blood in stool, and stomach pain. Histopathology is the field focused on diagnosing and studying tissue-related diseases by analyzing tissues and cells through a microscope. This paper introduces a method of identifying colon cancer from histopathology images through Vision Transformers (ViT) and highlight the cancer regions through Gradient-weighted Class Activation Mapping(GradCAM).Vision Transformers, a cutting-edge approach harnessing the self-attention mechanism initially designed for transformers in Natural Language Processing (NLP) tasks, are applied for image classification in this study. ViTs involve usage of self attention mechanism that allows model to focus on relevant regions and features, this is essential incase of histopathology images for understanding complex pattern in images. ViTs are more suitable for histopathology image classification because it captures global features effectively by understanding relationship between all image pixels. This method is compared with 2D Convolutional Neural Network . This method is highly useful for detecting colon cancer cells in the tissue.*

Keywords: Colon cancer; Vision Transformers; GradCAM; 2D Convolutional Neural Network; Natural Language Processing.

I. INTRODUCTION

Colon cancer typically occurs with growth of cancer cells in large intestine, it typically affects older adults. The cancer begins as small clumps of cells called polyps, these are generally non cancerous, but some can turn into colon cancer over time. Symptoms of colon cancer include change in bowel habits, rectal bleeding, discomfort in belly area, body weakness and weight loss. Treatments include surgery, radiation therapy and medicines, such as chemotherapy, targeted therapy and immunotherapy. When initially trained on extensive datasets and applied to various mid-sized or smaller image recognition benchmarks such as ImageNet, CIFAR-100, VTAB, etc., the Vision Transformer (ViT) demonstrates outstanding performance compared to leading convolutional networks. Moreover, ViT achieves these results with significantly reduced computational resources needed for training [1].ViT outperforms, achieving 28.10% top-1 accuracy on ImageNet-A with fewer parameters than a comparable Big-Transfer variant. Analyses on image masking and spectral properties reveal ViT's enhanced robustness [2].In response to the MIA-COV19 challenge, this study aims to classify COVID-19 from non-COVID cases using CT lung images. Employing both Vision Transformer (ViT) and DenseNet, initial evaluations on validation datasets indicate that ViT outperforms DenseNet, achieving F1 scores of 0.76 and 0.72 [3].

MIST, a novel deep learning model utilizing Multiple Instance learning network and based on the Swin Transformer, demonstrates high accuracy in classifying colorectal adenomas through whole slide images (WSIs). Trained on 666 WSIs from colorectal adenoma patients, the model achieves an impressive 0.784 accuracy in external validation, surpassing existing methods and approaching the accuracy of local pathologists (0.806). MIST's interpretability aligns with pathologists' assessments, making it a promising and practical tool for colorectal cancer screening, with potential implications for reducing mortality through improved clinical decision support [4]. The ViT-based approach for lung disease detection surpasses the CNN-based VDSNet, demonstrating enhanced performance even with additional parameters. Notably, ViTs exhibit promising potential, achieving a substantial 70.24% accuracy compared to VDSNet's 69.86% on chest X-ray analysis [5]. A thorough examination of state-of-the-art vision transformers in histopathological image analysis, covering classification, segmentation, and survival risk regression, is presented along with discussions

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

on challenges, opportunities, and future directions in computational histopathology [6]. Introducing a novel autoencoder network to transform features extracted by Inception_ResNet_V2 for clustering analysis, this study outperforms existing methods in both classification and clustering tasks. Additionally, vision transformers excel in multiclass tissue classification of CRC histology images, achieving 93.3% and 95% accuracy for Vision Transformer and Compact Convolutional Transformer, respectively [7]. Evaluating CRC detection methodology based on vision transformers on the CRC-5000 dataset, the study demonstrates superior diagnosis performance compared to recent works, including those incorporating vision transformers [8]. Another comparison between vision transformers and CNNs for histopathology image classification highlights vision transformers' superior performance in three out of four tissue types, including colon cancer [9]. An overview of transformer use in medical image analysis, particularly histopathology, discusses the advantages over traditional CNNs and reviews recent studies utilizing transformers across various medical image analysis tasks [10].

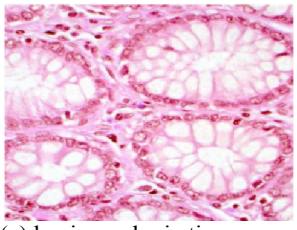
II. MATERIALS AND METHODS

2.1 Dataset

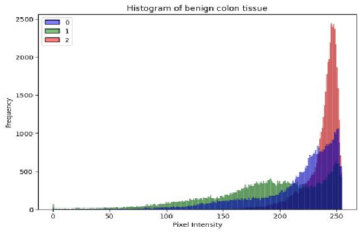
The utilized dataset, LC25000 Lung And Colon Histopathological Image Dataset, is a publicly available collection comprising 25,000 histopathology images. The dataset encompasses five distinct classes, with three classes dedicated to lung cancer and two classes specifically associated with colon cancer. For our study, we focused on two classes within the colon category, namely colon_aca (colon adenocarcinomas) and colon_n (benign colonic tissues), each containing 5,000 images. The images are of dimensions 768 x 768 pixels. The colon_aca class contains images representing cancerous colon tissue, while the colon_n class comprises images depicting benign colonic tissues.

2.2 Colon cancer tissue and non colon cancer tissue

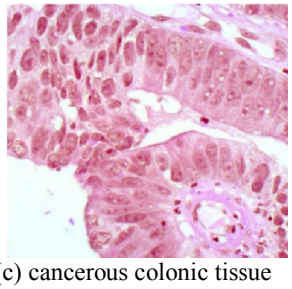
Colon cancer cells display irregular shapes and sizes, accompanied by distinctive nuclear abnormalities like enlarged nuclei, heightened nuclear staining (hyperchromasia), and noticeable nucleoli prominence. In contrast, benign colon tissue exhibits cells characterized by well-defined and organized glandular structures. These cells maintain a consistent morphology with regular nuclear features, and the tissue preserves its normal architecture without any disruptions or invasive tendencies.



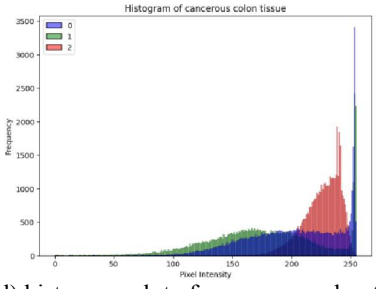
(a) benign colonic tissue



(b) histogram plot of benign colon tissue



(c) cancerous colonic tissue



(d) histogram plot of cancerous colon tissue

Fig. 1. Histopathology images of colon tissue and its corresponding histogram plot

2.3 Image Pre-processing

The image data of size 10,000 images is augumented by normalizing ,Resizing, RandomFlipping, rotatiing Zooming the images hence generating a dataset of size 50,000 images.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

Normalization involves changing the range of pixel value from 0 to 2555 to 0 to1 range.

$$I_{norm} = \frac{I_{original} - I_{min}}{I_{max} - I_{min}} (new_{max} - new_{min}) + new_{min} \quad (1)$$

I_{norm} is the pixel value after normalization, $I_{original}$ is the original pixel value, I_{max} and I_{min} are the maximum and minimum pixel values of the image ,new_{min} and new_{max} are the required range of pixel values after normalization set between 0 to 1.

Resizing an image in the context of image processing usually involves changing the dimensions of the image.

$$new_{width} = \sqrt{(\frac{originalwidth}{originalheight})}\ x\ target\_area (2)$$

$$new_{height} = targetarea/new_{width} \quad (3)$$

2.4 Vision Transformers

In real-time computer vision applications, Vision Transformers (ViTs) have proven to be highly effective for tasks such as object detection, image segmentation, image classification, and action recognition. This is attributed to their unique ability to capture both global and local features within an image. At the core of the Vision Transformer is the self-attention mechanism, a fundamental principle that enables the model to discern and evaluate interdependencies and connections within input sequences.

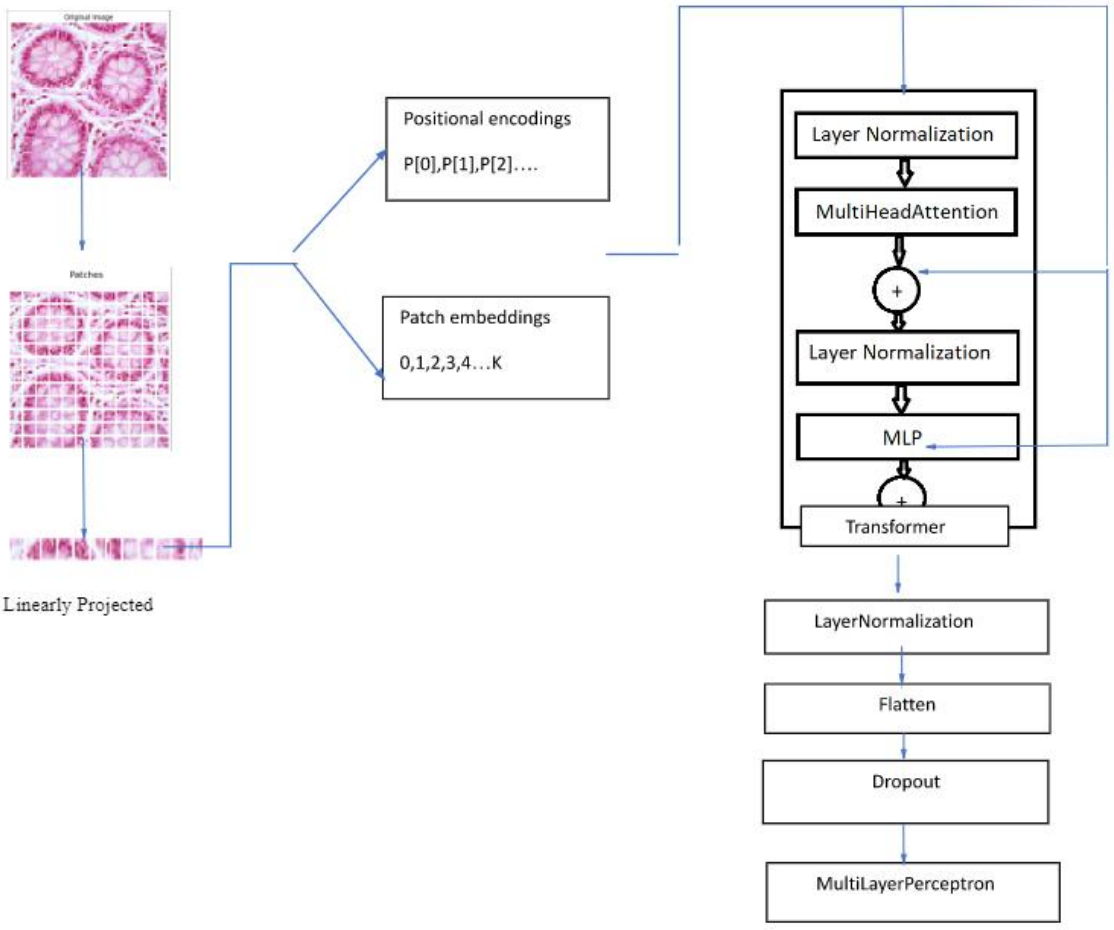


Fig.2. Vision Transformer based classifier block diagram

In our dataset, the input image, sized (768, 768, 3), undergoes a transformative process. It is first divided into 144 patches, and these patches are then encoded using the Patch Encoder layer. This layer performs linear projection on the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

patch features, effecting dimensionality reduction, and subsequently incorporates positional information using embeddings. This step is crucial in the Vision Transformer architecture as it readies the input for subsequent transformer layers, which excel in capturing global dependencies and local dependancies within the image.

The patch-encoded output is then fed into the transformer. Here, the self-attention mechanism plays a pivotal role, allowing the model to capture long-range dependencies within the input sequences. This transformed information is further processed by a feedforward neural network, facilitating the capture of complex, non-linear relationships within the patches.

The ultimate output from the transformer and feedforward network is directed to an MLP head, serving as the final stage for binary classification. Specifically, this head distinguishes between cancerous colon tissue and benign colonic tissues.

Patches are linearly projected to (P,K,V), P is query, k is Key and V is value.

$$Attention(P,K,V) = softmax\left(\frac{P.K^T}{\sqrt{d_k}}\right).V \quad (4)$$

2.5 GradCAM

Gradient-weighted Class Activation Mapping (Grad-CAM) leverages gradients associated with a target concept, whether in a classification network or a sequence of words in a captioning network, flowing into the last dense layer. This process generates a coarse localization map that highlights significant regions in the image crucial for predicting the given concept. The process involves passing the input through the model, capturing both the layer output and loss. Subsequently, we calculate the gradient of the desired model layer's output concerning the model loss. This gradient is then processed to isolate sections relevant to the prediction, which are subsequently resized, and rescaled to allow overlaying the heat-map onto the original image. In the initial step, gradients are computed within the final layer of the Vision Transformer (ViT), with a specific focus on comprehending the impact of these gradients on class predictions.

Compute the gradient of output y_c with respect to the feature map activations J^k of the dense layer, which is $\frac{\partial y_c}{\partial J^k}$

Perform global average pooling of gradients, over width dimension i and height dimension j,

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y_c}{\partial J_{ij}^k} \quad (5)$$

Perform a weighted combination of the feature map activation J^k for the weights α_k^c,

$$O_{Grad-CAM}^c = ReLU(\sum_K \alpha_k^c J^k) \quad (6)$$

III. MODEL TRAINING

The features extracted from the Vision Transformer are input into a Multilayer Perceptron (MLP) for classification. In this process, 80% of the data is allocated to the training set, 10% for testing, and an additional 10% for validation. The output from the transformer is then passed through Layer Normalization, which stabilizes and normalizes layer activations, facilitating faster convergence in training. The addition of a small constant (epsilon) to the variance prevents division by zero. Subsequently, a Flatten layer reshapes the tensor into a one-dimensional vector, commonly employed in transitioning from convolutional or recurrent layers to fully connected layers. Following this, a Dropout layer randomly nullifies a fraction (here, 50% or 0.5) of input units during training, serving as a regularization technique to prevent overfitting. Finally, the processed data is fed into the Multilayer Perceptron. The MLP's constructor initializes the model with layer pairs, each comprising a dense layer with Gated Linear Unit (GELU) activation and a dropout layer with a specified rate, introducing non-linearity and preventing overfitting. The forward pass (call method) efficiently navigates the input tensor through these layers, allowing for a dynamic and customizable architecture. The method iterates through the layer pairs, sequentially applying each layer to the input tensor. The input to MLP is of size (None, 9216) and the output from MLP is of size (None, 1024) The model is trained for 60 epochs with batch size of 50, using the ADAM optimizer. In our study to assess the effectiveness of vision transformers in colon cancer detection, we employed a 2D Convolutional Neural Network (CNN) model. The model begins with an input layer designed for images of size (768, 768). This is followed by two convolutional layers, each employing a (3, 3) filter, aimed at extracting intricate features from the input images. Subsequently, max-pooling layers with a (2, 2) pool size are applied to downsample the spatial dimensions and retain essential features. Afterwards, a flatten layer is

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

incorporated to transform the 2D feature maps into a vector, facilitating the transition to the fully connected layers. Two dense layers follow, the first with 64 neurons and ReLU activation, enabling the model to learn complex representations. The second dense layer serves a similar purpose, promoting further abstraction of features. The final dense layer utilizes a sigmoid activation function, providing a binary classification output for colon cancer detection. The 2D CNN model is trained for 60 epochs with batch size 0f 50 using the ADAM optimizer.

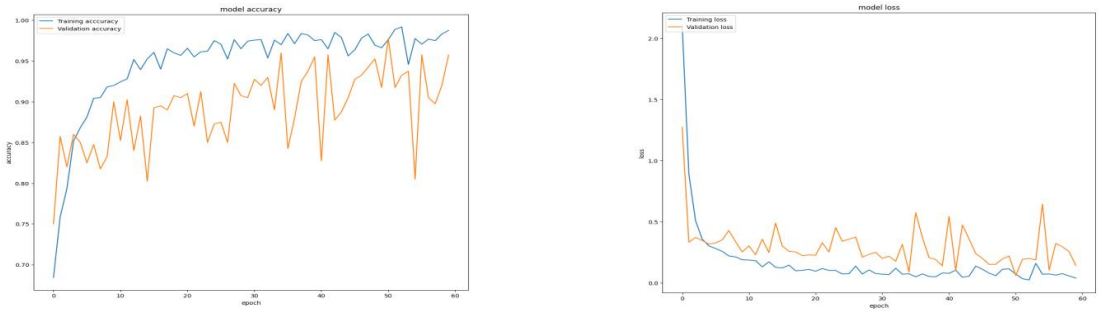


Fig.3. Loss Curve and Accuracy Curve of training and validation set.

IV. RESULTS AND DISCUSSION

The evaluation of the models in this study involved a comprehensive assessment using various performance metrics to gauge their classification capabilities. The metrics employed for this evaluation include Precision (P), Recall (R), and F1-score (F1) for each individual class. These metrics provide a detailed perspective on the models' adeptness in classifying instances for specific categories.

4.1 Accuracy

This metric reflects the proportion of instances that were predicted correctly in relation to the total number of instances within the dataset.(refer to Equation 7)

$$A = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

Where TP stands for True Positive , TN stands for True Negative, FP stands for False Positive, FN stands for False Negative.

4.2 Precision

This metric quantifies the proportion of true positive predictions among all positive predictions made by the model. Mathematically, it's represented as in Eqn. 8.

$$P = \frac{TP}{TP+FP} \quad (8)$$

Where TP stands for True Positives and FP stands for False Positives. Precision serves as an indicator of the model's accuracy by assessing its ability to minimize false positives.

4.3 Recall

Commonly known as Sensitivity or the True Positive Rate, Recall measures the proportion of true positive predictions in comparison to all actual positive instances (refer to Equation 9).

$$R = \frac{TP}{TP+FN} \quad (9)$$

Where TP stands for True Positives and FN stands for False Negatives. Recall provides insight into the model's ability to identify all relevant instances.

4.4 F1-Score

Functioning as the harmonic mean of precision and recall, the F1-score acts as a comprehensive metric that synthesizes the model's performance with respect to both false positives and false negatives (refer to Equation 10).

$$F1 = \frac{2\,X\,P\,X\,R}{P+R} \quad (10)$$

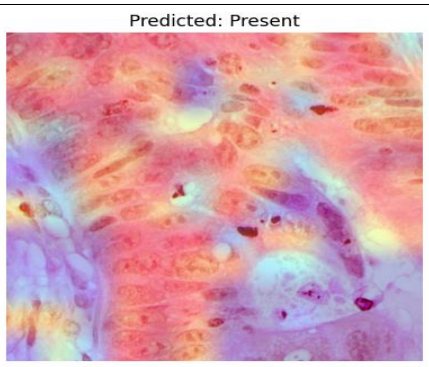
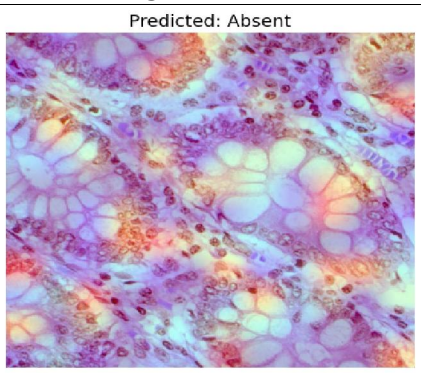
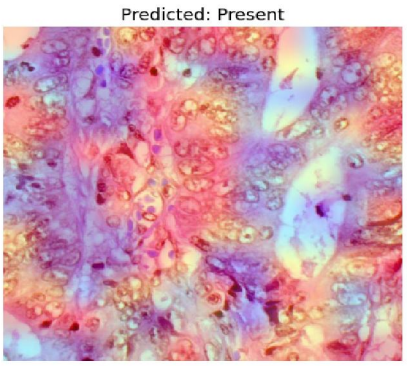
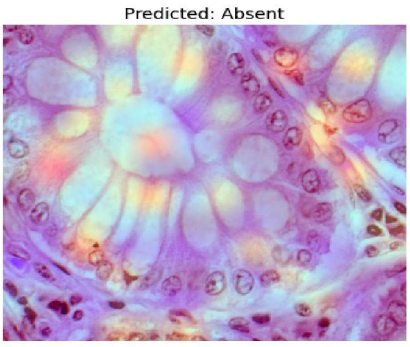
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

TABLE 1: Evaluation Metrics for Vision Transformer Classifier and 2D CNN Model

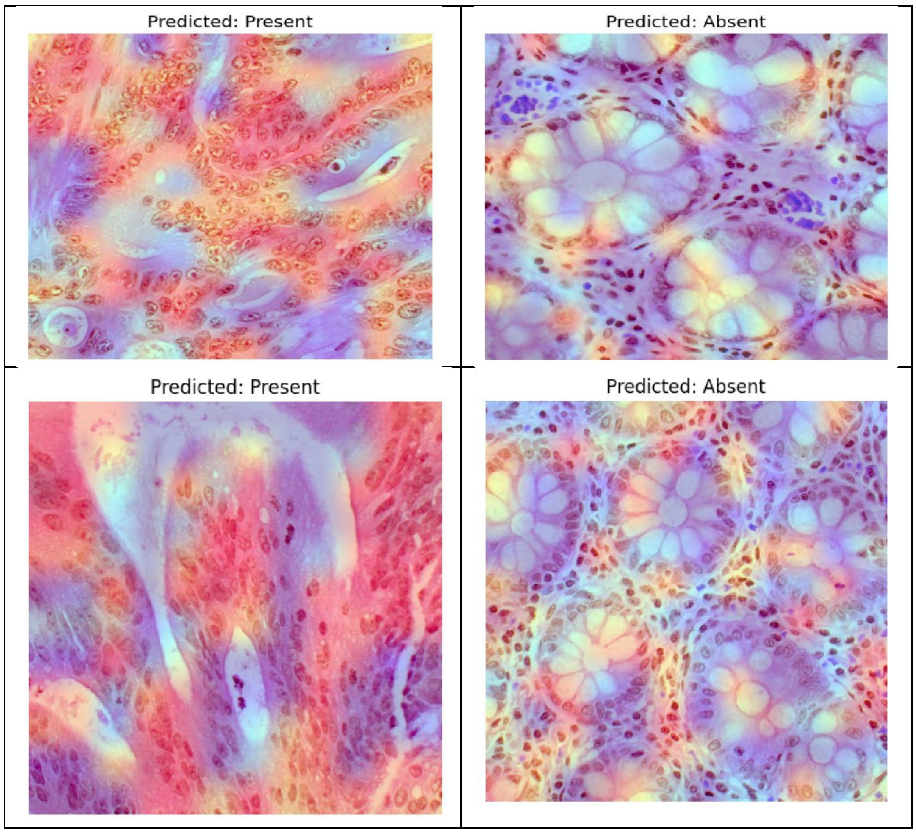
Performance metrics	Vision Transformers	2D CNN
Accuracy	96%	80.48%
Precision	0.99	0.87
Recall	0.92	0.79
F1 score	0.96	0.83

The substantial increase in accuracy from 80.48% with CNN to 96% with Vision Transformers (ViT) highlights the remarkable capacity of ViT in capturing intricate patterns and features within histopathology images. This improvement strongly suggests that ViT excels in learning complex spatial relationships and representations embedded in the data. Leveraging attention mechanisms, ViT models demonstrate a superior ability to focus on pertinent regions of the input image, allowing them to discern subtle details crucial for accurate classification. Furthermore, ViT's intrinsic design facilitates the effective capture of global context information, a critical aspect in histopathology image analysis where the arrangement and relationships among cells and structures can provide vital insights into pathological conditions. This enhanced understanding of the global context likely contributes significantly to the observed superior performance of Vision Transformers in comparison to traditional CNN architectures

TABLE II: Grad-CAM Result Trained Using Vision Transformers

Cancerous Colon Tissue	Benign Colon Tissue
	
	

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024



In the GradCAM results, adenocarcinomas, characterized by dense cancerous cells, are prominently displayed in red. In contrast, normal colon cells are indicated by a distinctive yellow color.

V. CONCLUSION

This investigation utilized Vision Transformers (ViT) in conjunction with a MultiLayer Perceptron (MLP) for the classification of colon tissue images. The methodology involved image segmentation into patches, with the integration of position encodings and patch embeddings into the transformer model. The resulting output from the transformer model was then fed into the MLP for the final classification. Subsequently, the trained Vision Transformer model was employed with GradCAM to highlight cancer cells and non-cancerous cells within colon tissue images. This application of GradCAM aids in quantifying the presence of cancer cells in an image, thereby providing valuable insights into the stage of cancer.

In the comparative analysis between ViT and 2D CNN models, both were trained, and notably, the Vision Transformer outperformed the 2D CNN. This superior performance can be attributed to ViT's self-attention mechanism, leading to an impressive accuracy of 96%. Consequently, the study concludes that Vision Transformers demonstrate enhanced efficacy in the detection of colon cancer cells in histopathology images, showcasing their potential for improved diagnostic accuracy and clinical utility in cancer detection.

VI. ACKNOWLEDGMENT

The author express their gratitude to Borkowski AA, Bui MM, Thomas LB, Wilson CP, DeLand LA, Mastorides SM. Lung for sharing the dataset used in this study.

FUNDING STATEMENT

The author(s) did not receive any special funding for this study.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

AVAILABILITY OF DATA AND MATERIALS

Thedataset is available at the link https://academictorrents.com/details/7a638ed187a6180fd6e464b3666a6ea0499af4af, and it is accessible to the public.

REFERENCES

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., &Houlsby, N. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv.org.

[2] Paul, S., & Chen, P.-Y. 2021. Vision Transformers are robust learners. arXiv.org.

[3] Gao, X., Qian, Y., & Gao, A. 2021.Covid-VIT: Classification of covid-19 from CT chest images based on Vision Transformer models. arXiv.org.

[4] Cai H;FengX;YinR;ZhaoY;GuoL;FanX;Liao J; (2023, February). Mist: Multiple instance learning network based on Swin Transformer for whole slide image classification of colorectal adenomas. The Journal of pathology., 259(2):125-135.

[5] Uparkar, O. et al. 2023. Vision Transformer Outperforms Deep Convolutional Neural Network-based Model in Classifying X-ray Images, ScienceDirect.com., 218: 2338-2349.

[6] Xu, H., Xu, Q., Cong, F., Kang, J., Han, C., Liu, Z., Madabhushi, A., & Lu, C. 2023. Vision Transformers for Computational Histopathology: IEEE Journals & Magazine: IEEE Xplore.

[7] XieJuanying, Liu Ran, Luttrell Joseph, Zhang Chaoyang. 2019. Deep Learning Based Analysis of Histopathological Images of Breast Cancer, Frontiers in Genetics.

[8] Ayyad, S.M.; Shehata, M.; Shalaby, A.; Abou El-Ghar, M.; Ghazal, M.; El-Melegy, M.;Abdel-Hamid, N.B.; Labib, L.M.; Ali, H.A.; El-Baz, A. 2021. Role of AI and Histopathological Images in Detecting Prostate Cancer: A Survey. MDPI., 21(8): 2586.

[9] Deininger, Luca &Stimpel, Bernhard &Yuce, Anil & Abbasi-Sureshjani, Samaneh&Schönenberger, Simon & Ocampo, Paolo &Korski, Konstanty&Gaire, Fabien. (2022). A comparative study between vision transformers and CNNs in digital pathology.

[10] Zhang, Y., Wang, J., Gorriz, J. M., & Wang, S. 2023. Deep Learning and Vision Transformer for medical image analysis. MDPI, 9(7): 147.