Multiclass Colorectal Cancer Histology Images Classification Using Vision Transformers

Magdy Abd-elghany Zeid
Computer Science Department
Obour High Institute for Management and Informatics
Cairo, Egypt
magdy_zeid83@oi.edu.eg

Khaled El-Bahnasy
Computer Science Department
Obour High Institute for Management and Informatics
Cairo, Egypt
khaled.bahnasy@oi.edu.eg

S.E.Abo-youssef
Mathematics and Computer Science Department
Faculty of Science
Al-Azhar University
Cairo, Egypt
abuyousf@hotmail.com

2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS) | 978-1-6654-4076-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICIS52592.2021.9694125

***Abstract*— Colorectal cancer (CRC) is the third most diagnosed cancer form globally and the second leading cause of cancer-related death after lung cancer. A precise histological categorization of CRC tissue is critical for diagnosis and patient management decisions. However, the variety of tissue patterns in CRC histological images makes the classification a challenging problem. This study applies Vision Transformers, a new class of deep-learning models in computer vision, to perform a multiclass tissue classification of a publicly available CRC histology images dataset. The data set consists of 5000 images with eight categories of tissue. We trained two variants of Transformers, namely Vision Transformer and Compact Convolutional Transformer, and achieved 93.3% and 95% accuracy, respectively. Our results outperform the original paper (87.4%) on the same dataset. Furthermore, our study highlights the opportunities of using Transformers in the histopathological image domain.**

Keywords—Biomedical image processing, Colorectal cancer detection, Deep Learning, Pathology, Vision Transformers

I. INTRODUCTION

Colorectal cancer (CRC) is a kind of malignancy that begins in the rectum or colon. CRC is the third most frequent type of tumor in men in Egypt, after urinary bladder and lymphohematopoietic tumors, and the fifth most common type in females, after bladder tumors[1]. GLOBOCAN 2020 data[2] presented that the third highest diagnosed cancer form in the world is CRC (10%) after breast (11.7%) and lung (11.4%) cancers. CRC is the second prominent reason of cancer-related death (9.4%) following lung cancer, as shown in fig. 1.

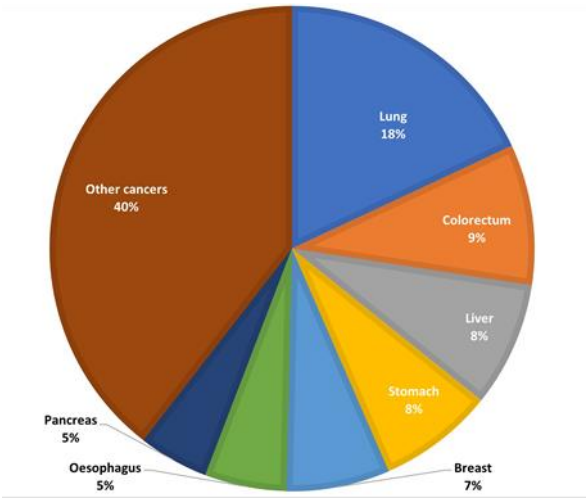

Fig. 1. Estimated number of death cases worldwide in 2020

Cancer is a category of disorders defined by the uncontrolled development of aberrant cells with the latent ability to invade other tissues[3]. CRC is often initiated by polyps which is a proliferation of a noncancerous mucosal epithelial cells. The polyps can develop slowly taking many years(10–20) into cancer. Polyp or adenoma is the most frequent kind and developed from granular cells that generate mucus for the large intestine's lining[4]. Only around 10% of all adenomas develop into aggressive carcinoma; however, the risk increases with the size of the polyp. Adenocarcinoma is invasive cancer that develops from such polyps and accounts for 96% of all CRCs[5].

Human solid tumors are complicated masses composed of a variety of different tissue types. Besides clonal tumor cells, they include necrotic regions, immune cell penetration, tumor stroma, and islets of residual benign tissue. Histopathological assessment of Hematoxylin and Eosin tissue-stained specimens can differentiate these various tissue types. In CRC, tumor structure evolves over time and is associated with patient prognosis [6]. As a result, assessing the CRC tissue component is a critical histopathological task.

Although manual examination of histopathological slides is still crucial in experimental practice, automatic image processing enables the quantitative and fast study of malignant tissues[7]. Early diagnosis is critical for determining the most appropriate care plan and increasing the patient's survival probability. As a result, automated techniques are required to save time and reduce the possibility of human mistakes. Recently, much work has been devoted to diagnosing and forecasting several types of cancer using artificial intelligence[8], [9].

The revival of artificial neural networks via deep learning [10] has increased the accuracy of numerous pattern recognition applications. Histological images are generally composed of a complicated set of patterns A traditional machine learning classifier requires extensive feature analysis which is a function to expert interpretation and prior knowledge. This process, called feature engineering, is frequently tedious and time-exhausting. The need for feature engineering can be eliminated by Deep learning that automatically and immediately learn representative features from unprocessed input instances such as images of malignant tissues received from patient for investigative reasons [11]. In the biomedical domain, promising outcomes in image-based diagnoses have been made in the field of diagnostic pathology[12]. Within digital pathology, supervised deep learning has demonstrated promising results for quantifying and classifying digitized tissue samples, especially for tasks previously deemed too difficult to achieve using standard image analysis methods [11].

Kather et al. (2016)[7] demonstrated that all published methods for CRC histological image classification suffer from

two main limitations: first, they take into account only two groups of tissues (tumor and stroma), rendering them inappropriate for additional heterogeneous chunks of the tumor; and secondly, each study used its private images dataset, preventing quantitative judgement of classifiers performance. While freely accessible benchmarking datasets be present for image categorization issues for example MNIST handwriting recognition and CIFAR 100 image datasets, histopathological tissue classification does not have such data. Kather et al. collected, assessed, and freely distributed a complete image dataset of all-important tissue categories seen in colorectal cancer specimens. They compared various up-to-date texture characteristics and algorithms against the dataset to discover the best one suitable for a classification of tissue multiclass challenge. Transformers are a new class of deep-learning models that have gained some traction in computer vision. Transformers depend on a simple yet effective attention process, which concentrates on certain input pieces to produce more efficient outputs. Now, they are regarded as cutting-edge models for sequential data, particularly natural language processing techniques for instance machine translations [13]. Inspired by the success of Transformers in natural language processing, recent studies attempt to apply Transformers to images directly [14], [15].

This study aims to perform a multiclass tissue classification of colorectal cancer using Vision Transformers. We will use Kather colorectal cancer histology dataset " textures Collection of colorectal cancer histology"[16], which is freely accessible online, to compare our results of multiclass tissue classification with the original paper[7] and to highlight the opportunities of using Transformers in the histopathological image domain. The remainder of the article is structured as follows: The second section, we will cover related work. In section three we will explain the methods. While section four describes our results and discussions. Lastly, Section five concludes the article.

II. Related work

Previous studies indicate the possibility for an AI-based technique that includes machine learning into pathologist diagnosis systems for cancer in order to accelerate the diagnosis process, reduce error and improve accuracy. The adoption of deep learning-based systems for digital image recognition and detection is a significant advancement in digital pathology[17].

In the context of CRC histology images, Linder et al. (2012) [18]and Bianconi et al. (2015) [19] apply texture-based techniques to categorize CRC tissue types. They extract texture features and then input them into a classification algorithm to predict the tissue type. However, these Methods achieve accuracies between 97 – 99.8%, but they are inappropriate for multiclass tissue classification because they address just two tissue types (tumor and stroma).

In 2016, Kather et al.[7] published about 5,000 histological colored images of human for different cancer tissue types in a dataset. They evaluated the performance of classification for a broad variety of texture characteristics and algorithms using this dataset. Consequently, they are establishing a new benchmark for tissues multiclass separation with accuracy of 87.4% for eight classes. They make their histology image dataset freely available to the public through the Creative Commons license in addition to urging more academics to utilize the dataset as a standard in their study.

In 2019, Rizalputri et al.[20] utilized several multi-classification techniques to the Kather colorectal histology dataset; K-Nearest Neighbor (KNN), Convolutional Neural Networks (CNN), Random Forest, and Logistic Regression. purpose of their study is to compare each technique's performance and identify the optimal algorithm; the best method is CNN, which achieves an accuracy of 82.2 %.

In 2020, Yazdi M and Erfankhah H[21] proposed extracting four local features, including regional structural knowledge, regional geometric knowledge, regional energetic knowledge, and regional patterns, to describe the histology images' enormous textural diversity in a feature space. The Riesz transform and the monogenic local binary patterns are employed to extract these characteristics. They evaluated their approach against two multiclass histology image datasets, namely the Kather and Kimiapath24 datasets. They obtained a classification and retrieval performance of more than 90% in two datasets.

In 2021, Ohata et al[22]. proposed the automated identification of eight tissues observed in colorectal cancer histological assessment using the Kather Colorectal Histology dataset. They employ Transfer Learning based on CNN architectures. They alter the structure of CNNs in order to extract characteristics from images and feed them into many machine learning algorithms. DenseNet169 with SVM (RBF) obtained the highest outcomes with 92.08 % and 92.12 % for accuracy and F1-Score respectively.

III. Methods

This section will describe the dataset, explain data pre-processing, and illustrate the proposed model's design and implementation.

A. Dataset description

This study we will use Kather et al. colorectal cancer histology dataset, which is freely accessible online [17]. By focusing on histology tiles from colorectal cancer patients, the data set works as a more intriguing MNIST or CIFAR10 challenge for biologists. The data set contains eight distinct tissue classes ('TUMOR', 'STROMA', 'COMPLEX', 'LYMPHO', 'DEBRIS', 'MUCOSA', 'ADIPOSE', and 'EMPTY'). A sample of Kather colorectal cancer histology dataset eight classes is shown in fig. 2.

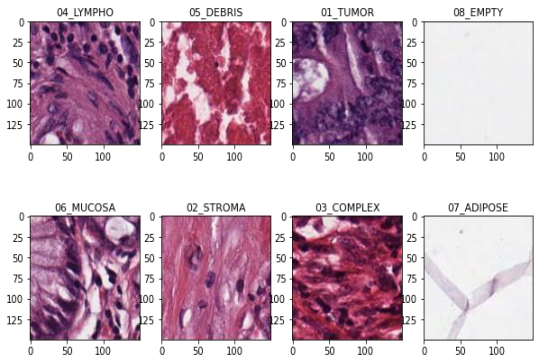

Fig. 2. The eight classes of Kather colorectal cancer histology dataset.

The dataset contains textures extracted from histology images of colorectal human cancer. It consists of two archives: the first is "Kathertexture2016imagetiles5000.zip," a zipped package including 5000 histology images with a resolution of 150 * 150 px (74 * 74 m). Every image corresponds to one of eight distinct tissue types. The dataset contains an equal

number of images from each class (balanced dataset). The second, Kathertexture2016largerimages10.zip": a zipped package containing ten bigger histology images, each measuring 5000 by 5000 px in size. These images show a combination of multiple tissue types.

All images are RGB, with a pixel size of 0.495 m, and were scanned using an Aperio ScanScope at a amplification of 20x. Histopathological specimens are completely anonymized images of human colorectal adenocarcinomas (primary tumors) preserved in formalin and embedded in paraffin. The dataset contains textures extracted from histology images of human colorectal cancer. Each image corresponds to one of eight distinct tissue types [16].

B. Data Pre-processing

Tuning of deep learning model hyperparameters is an iterative process, so it is required a separate testing set (validation set) used during training. Furthermore, another unseen testing set is required for the final evaluation of the model. Therefore, we split the dataset into three parts (training, validation, and testing datasets). We split the data into 80% of the dataset for training and 20% dataset for testing. Then we divide the testing set into two groups, 30% for validation during the training and hyperparameter tuning processes to ensure no overfitting occurs, and 70% for evaluating model accuracy once the training and hyperparameter tuning processes are complete. All images are normalized by standardizing data. In this manner, the optimizer will converge more quickly. Fig. 3 shows image samples before and after standardization.

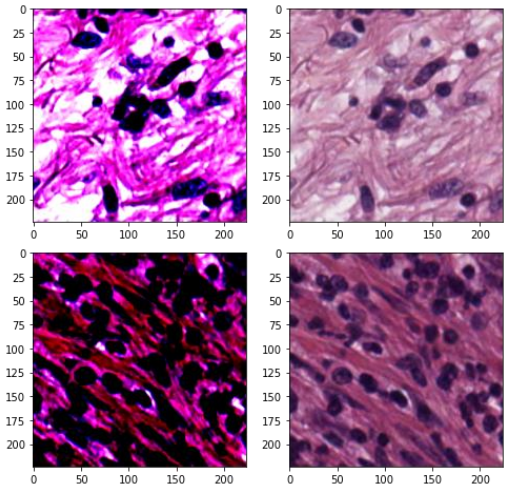

Fig. 3. image samples before (left) and after standardization (right).

We utilized a variety of augmentation approaches to improve our runtime data, avoiding the model from overfitting and allowing it to generalize and achieve excellent performance.

Data augmentation is a type of regularization performed at the dataset level to mitigate overfitting and improve generalization performance by supplementing the training dataset without modifying the proposed classification approach[23]. Medical image datasets are typically tiny and challenging to collect. On the other hand, Convolutional Neural Networks have been proven to perform effectively with data augmentation in the categorization of liver injuries, brain scan analysis, skin condition classification, and a variety of other medical imaging issues[24].

Although the dataset contains 5000 images, the deep learning algorithm is starving for data, and its accuracy improved with more data [25]. Data augmentation is viewed as a means of compensating for the shortage of training data. The term "data augmentation" refers to the process of expanding a dataset by performing different modifications on existing data, such as scaling, flipping, and lightning, in order to produce new images while maintaining the target class's identity[26].

Deep learning has been shown to perform effectively with larger datasets. Data augmentation increases the total number of images (by introducing new modified images) available at runtime, allowing the model to train more effectively. We used various data augmentation procedures, involving Random Translation, zooming, rotation modifications, and random horizontal flipping.

C. Vision Transformer

As demonstrated in Dosovitskiy et al. [14] paper, Vision Transformer (ViT) architecture is primarily based on the vanilla Transformer [27], which has gained much popularity in latest years for its advanced efficiency in machine translations and various NLP applications [28].

The Transformer is an encoder-decoder design that enables a simultaneous processing of chronological information without using a recurrent neural network. Transformer models' effectiveness is primarily due to the self-attention technique, which is hypothesized to obtain long-range connections between sequence parts.

ViT is presented as a means of extending the conventional Transformer's application to image categorization. The primary objective is to generalize them to non-textual modalities with no incorporation of any data-specific design. ViT uses the Transformer's encoder module to carry out the classification by modeling a sequence of image pieces to a semantic tag. Contrasting traditional CNN designs, which generally utilize filters with a limited receptive field, ViT's attention mechanism enables it to attend across several parts of the image and interpret the information throughout the whole image. Fig. 4 demonstrates the model's entire end-to-end architecture. It generally consists of:

- The embedding layer,
- The encoder.
- The final classifier head.

The initial action is to split the training set image into non-overlapping patches. The Transformer treats each patch as a distinct token. Therefore, an image of [*c, h, w*] dimension gives n sequence patches of dimension [*c, p, p*] (where c is the amount of channels, *h* represents the height, *w* denotes the width, and *p* is the patch size), where *n* is the number of patches is calculated by dividing the *h*w* over *p* square.

Typically, a patch size of 16 or 32 is selected since a lower patch size resulting in a more extended sequence and vice versa. However, we select the patch size of 15 because our image dimension is (150,150).

1) Embeding Layer

The patch sequence is linearly projected onto a 1D vector using a trainable linear projection (embedding matrix E) before passing to the encoder. The embedded patches after that integrated with a learnable embedding classification token x_{class} which is necessary for the classification job. The

positional embeddings E_{pos} is attached to the patch representations to retain positional information. E_{pos} dimension is ((n+1), D) where D is the vector size. The resultant sequence of embedded patches is represented with the token z_0 as in (1) where $E \in R^{(p^2c)}$, $E_{pos} \in R^{(n+1)\times D}$

$$z_0 = \left[x_{class};\ x_p^1 E;\ x_p^2 E;\ \ldots;\ x_p^n E\right] + E_{pos} \quad (1)$$

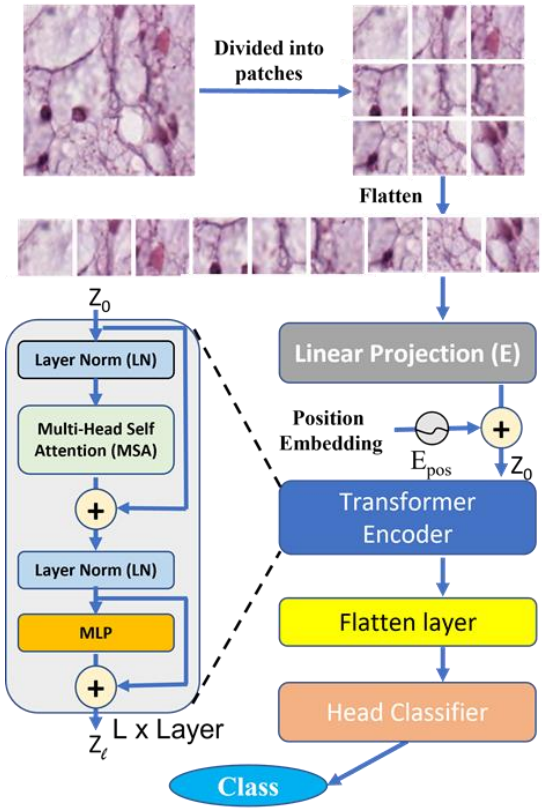

Fig. 4. Vision Transformer (ViT) model overview.

2) Vision Transformer Encoder

The encoder is made from L identical layers. Every single layer consists of a multiheaded self-attention (MSA) block as shown in (2) and a multilayer perceptron (MLP) block as in (3). a normalization layer (LN) is added before each block, and skip connections are applied after each block. The MLP is composed of two layers with a non-linear activation function GELU.

$$z_l' = MSA\left(LN(z_{l-1})\right) + z_{l-1}, \quad l = 1 \ldots L \quad (2)$$

$$z_l = MLP\left(LN(z_l')\right) + z_l', \quad l = 1 \ldots L \quad (3)$$

$$y = LN(z_L^0) \quad (4)$$

The first component in the sequence is taken at the final layer of the encoder and fed to the head classifier to predict the class label, as shown in (4). Unlike the original paper method described above, which prefixes the sequence of encoded patches with a learnable embedding classification Token to act as the image representation, we use Flatten layer that reshaped the final Transformer block's outputs and used as the classifier head's image representation input.

D. Compact Convolutional Transformer

ViT architectures usually demand a larger dataset and a more extended pre-training period [14]. This is mainly because, in contrast to CNNs, ViTs lack well-informed inductive biases. This raises the question: why can we not combine the advantages of transformers and convolution in a single network design? Hassani et al. [29] propose a method for doing this. They introduced the Compact Convolutional Transformer (CCT) architecture, which employs a convolutional-based patching approach that retains local information and, unlike the original ViT, can encode relationships between patches. CCTs are convolutions with short strides that enable efficient tokenization while maintaining local spatial relations. Additionally, they provide a sequential pooling approach called SeqPool, which pools the transformer encoder's sequence-based information. SeqPool obviates the requirement for an additional Classification Token. Fig. 5 demonstrates the model's entire end-to-end architecture. The authors of the original study employ Auto Augment to enhance regularization. Nevertheless, we will employ standard geometric augmentation such as random cropping and flipping to avoid the auto-augmentation computational cost.

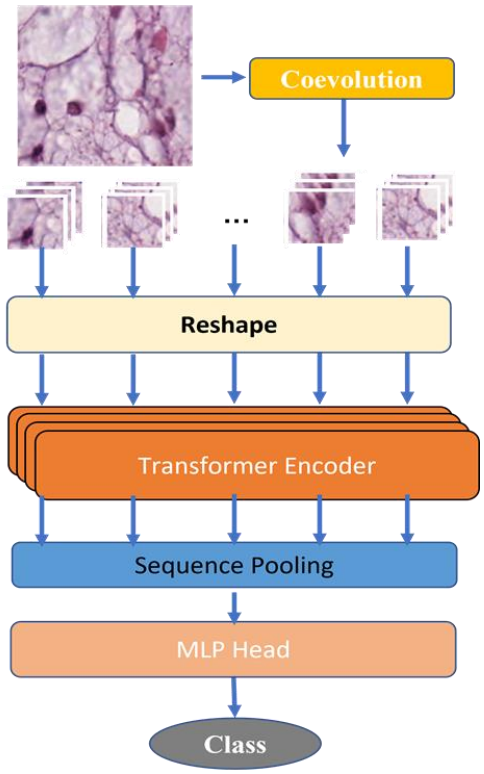

Fig. 5. Compact Convolutional Transformer (CCT) model overview.

E. Models Implemntation and Training

We use google Colab with GPU to train the models. Python 3.7.12 is used to implement the models, with Tensor Flow 2.6 serving as a backend to the Keras 2.6 library. The tensorflow-addons library is required to be installed on the Colab kernel. We employ the Adam optimization technique for both ViT and CCT models and initialize the learning rate with 0.001. Table I lists the hyperparameters that were applied during the training phase. The training process takes around 60 minutes for the ViT model and around 98 minutes for the CCT model.

TABLE I. TRAINING HYPERPARAMETERS

Hyperparameter	ViT Model	CCT model
Batch size	64	32
Number of epochs	100	100
Optimizer	Adam	Adam
Learning rate (Initial)	0.001	0.001
Weight decay	0.0001	0.0001
Patch size	15	Not applicable
Projection dimension	128	128
Transformer layers	8	4

IV. RESULTS AND DISCUSSION

In this section, we will represent and explain the result of our work. We will use the accuracy, precision, recall, and f1-score to assess our model. The accuracy and loss curves for both the training and validation datasets during the ViT and CCT models training are shown in Fig. 6 and 7.

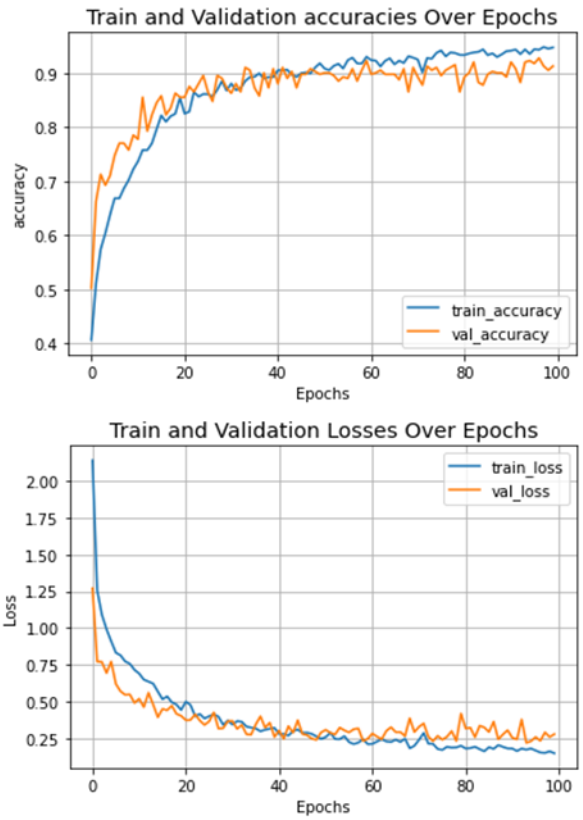

Fig. 6. Accuracy and Loss curves for validation and training sets during ViT model training.

The accuracy of classification models is a widely used criterion. It compares the total number of accurate predictions against the total number of guesses.

Recall (Sensitivity) measures the amount of properly categorized positive cases to the entire quantity of positive cases in the dataset, as defined in (5), where TP and FP denote the count of True Positive and False Positive instances, respectively, where TN and FN denote for the amount of instances of True Negative and False Negative, respectively.

$$recall = \frac{TP}{TP\ +\ FN} \tag{5}$$

Precision expresses the ratio of properly categorized positive cases to the overall amount of predicted positive instances, as shown in (6).

$$Precision = \frac{TP}{TP\ +\ FP} \tag{6}$$

F1-score measures the precision and recall harmonic mean [30] as seen in (7).

$$F1\_score = \frac{2\,TP}{2TP + FP + FN} \tag{7}$$

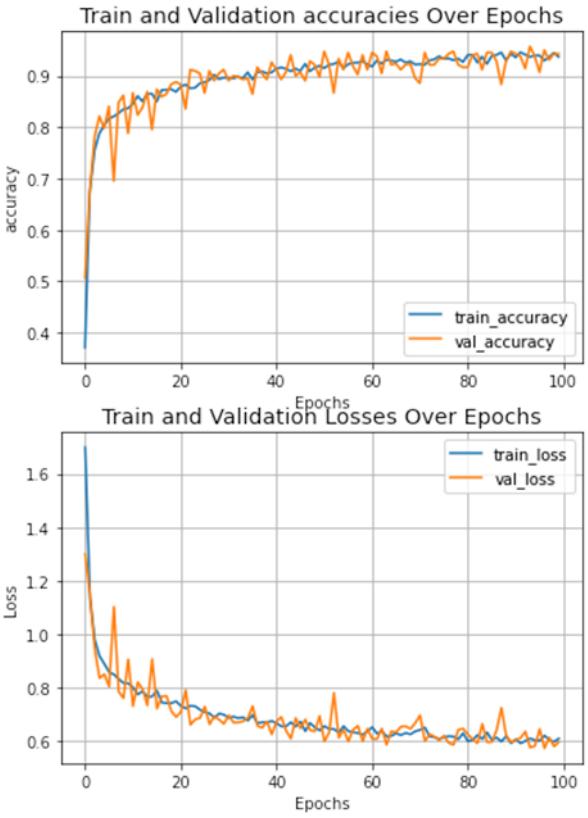

Fig. 7. Accuracy and Loss curves for validation and training sets during CCT model training.

The CCT model achieves the best metrics performance, It gives an overall 94.75% accuracy, and its f1-score is 94.74% on the testing set. In comparison, the Vit model produces an accuracy of 93.3% while gives a f1-score of 93.3%. That was expected because CCT architecture employs a convolutional-based patching approach that retains local information and encodes relationships between patches. In contrast, the ViT model lacks well-informed inductive biases. Table II summaries the overall accuracy, F1-score, recall, and precision obtained by both ViT and CCT models on the testing set. However, Tables III and IV report each class evaluation metrics on the testing set (recall, precision, and f1-score) for ViT and CCT models.

Also, As shown in fig. 6 and 7, the Vit model loss curve begins to overfit from epoch 50. In contrast, the CCT model does not suffer from overfitting until the end of the training. Therefore, we can increase the iteration number.

TABLE II. PERFORMANCE MEASURES OF MODELS ON TESTING SET.

Model	Accuracy	Recall	Precision	F1-score
ViT Model	93.3	93.33	93.44	93.3
CCT Model	94,75	94.75	94.8	94.74

TABLE III. VIT MODEL RESULT SUMMARY FOR EACH CLASS.

class	precision	recall	f1-score
TUMOR	0.99	0.98	0.98
STROMA	0.84	0.87	0.86
COMPLEX	0.95	0.88	0.91
LYMPH	0.99	1.00	1.00
'DEBRIS	0.90	0.98	0.94
MUCOSA	0.95	0.98	0.96
ADIPOSE'	0.87	0.88	0.88
BACKGROUND	0.97	0.90	0.93
weighted average	0.93	0.93	0.93

TABLE IV. CCT MODEL RESULT SUMMARY FOR EACH CLASS.

class	precision	recall	f1-score
TUMOR	0.97	0.99	0.98
STROMA	0.87	0.90	0.88
COMPLEX	0.94	0.92	0.93
LYMPH	0.99	0.99	0.99
'DEBRIS	0.93	0.97	0.95
MUCOSA	0.97	0.98	0.98
ADIPOSE'	0.96	0.87	0.91
BACKGROUND	0.96	0.96	0.96
weighted avg	0.95	0.95	0.95

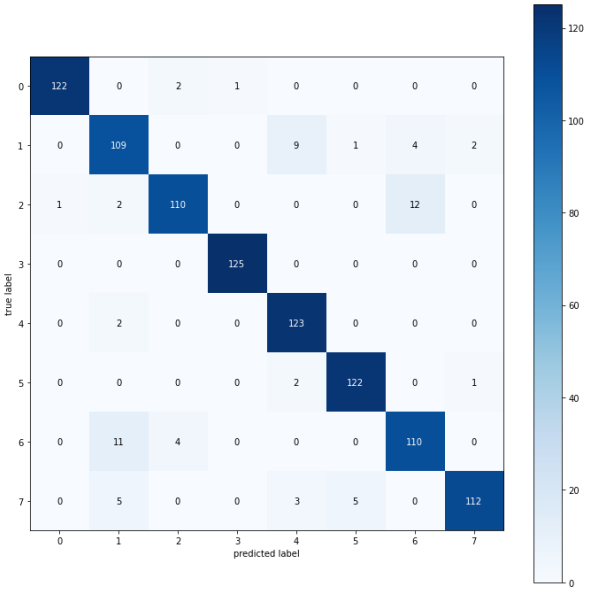

Fig. 8. The confusion matrix for Vit model

We generated the confusion matrix for each class to determine the error distribution across the various classes. Fig. 8 and 9 illustrate confusion matrices for our two models. From the analysis of confusion matrices, The Vit model is most sensitive to the Lymph class. While most errors come from misclassifying Complex tissue class with Adipose tissue class, then from misclassifying Adipose tissue class with Stroma tissue class, and lastly from misclassifying Stroma tissue class with Debris tissue class.

On the other hand, The CCT model is most sensitive to the Tumor tissue and Lymph tissue classes. While most errors come from misclassifying Adipose tissue class with Stroma tissue class, then from misclassifying Adipose tissue class with Complex tissue class.

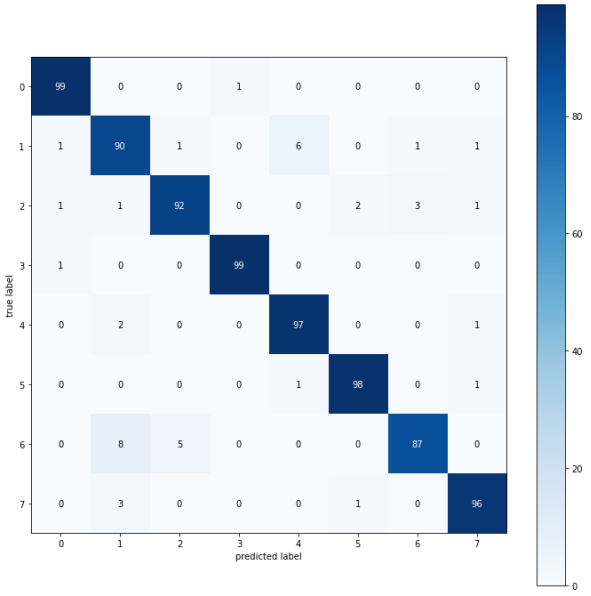

Fig. 9. The confusion matrix for CCT model

Table V compares our two models' accuracy results with previous works on the same dataset. Our CCT and ViT model contributes to CRC multiclass tissue classification research by introducing significantly better performance than the original paper and other work done on this dataset. Moreover, to our knowledge, this study is the first attempt to apply Vision Transformers in the field of histopathology images. In our opinion, it seems that there are promising opportunities from using Vision Transformers in the field of histopathology images recognition.

TABLE V. COMPARISON BETWEEN OUR MODELS AND PREVIOUS LITERATURES.

Author	Year	accuracy
Kather et al[8] original paper	2016	87.4 %
Rizalputri et al[20]	2019	82.2 %
Yazdi et al. [21]	2020	93.4 %
Ohata et al. [22]	2021	92.08 %
Our Vit Model		93.3 %
Our CCT Model		95 %

V. CONCLUSION

In this paper, we have been applied Vision Transformers, a new methodology of deep-learning models in computer vision, to perform a multiclass tissue classification of a publicly available CRC histology images dataset. We used Kather et al. colorectal cancer histology dataset [16], which is freely accessible through the following link (https://doi.org/10.5281/zenodo.53169). We trained two variants of Transformers. First, the Vision Transformer model achieved 93.3% accuracy and 93.3% f1-sore. The second, the Compact Convolutional Transformer model that achieved 94.75% accuracy and 94.75% f1-sore. Our results outperform the original paper (87.4%) on the same dataset and other approaches for multiclass tissue classification of CRC histology images. This study showed promising opportunities for using Vision Transformers in the field of histopathology image recognition domain. We plan to expand this experiment to perform experimentation with other datasets and different types of cancer for future work.

REFERENCES

[1] H. T Selim, Y. E Hossein, E. E Hassan, and M. D Mohammed, "Awareness about Risk Factors of Colorectal Cancer among Employees at Minia University," *Minia Scientific Nursing Journal*, vol. 9, no. 1, pp. 40–49, 2021.

[2] IARC - WHO, "Colorectum fact sheet - Globocan 2020." 2020. https://gco.iarc.fr/today/data/factsheets/cancers/10_8_9-Colorectum-fact-sheet.pdf

[3] V. Gyanani, J. C. Haley, and R. Goswami, "Challenges of Current Anticancer Treatment Approaches with Focus on Liposomal Drug Delivery Systems," *Pharmaceuticals*, vol. 14, no. 9, p. 835, 2021.

[4] P. Rawla, T. Sunkara, and A. Barsouk, "Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors," *Przeglad gastroenterologiczny*, vol. 14, no. 2, p. 89, 2019.

[5] S. L. Stewart, J. M. Wike, I. Kato, D. R. Lewis, and F. Michaud, "A population-based study of colorectal cancer histology in the United States, 1998–2001," *Cancer*, vol. 107, no. S5, 2006.

[6] M. Egeblad, E. S. Nakasone, and Z. Werb, "Tumors as organs: complex tissues that interface with the entire organism," *Developmental cell*, vol. 18, no. 6, pp. 884–901, 2010.

[7] J. N. Kather *et al.*, "Multi-class texture analysis in colorectal cancer histology," *Scientific reports*, vol. 6, no. 1, 2016.

[8] H. Wang *et al.*, "Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features," *Journal of Medical Imaging*, vol. 1, 2014.

[9] D. Bychkov *et al.*, "Deep learning based tissue analysis predicts outcome in colorectal cancer," *Scientific reports*, vol. 8, no. 1, pp. 1–11, 2018.

[10] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[11] D. Bychkov *et al.*, "Deep learning based tissue analysis predicts outcome in colorectal cancer," *Scientific reports*, vol. 8, 2018.

[12] G. Litjens *et al.*, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific reports*, vol. 6, no. 1, pp. 1–11, 2016.

[13] Q. Wang *et al.*, "Learning deep transformer models for machine translation," *arXiv preprint arXiv:1906.01787*, 2019.

[14] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[15] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li, "HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, 2019.

[16] J. N. Kather *et al.*, "Collection of textures in colorectal cancer histology," *doi. org, May*, 2016.

[17] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *Journal of pathology informatics*, vol. 7, 2016.

[18] N. Linder *et al.*, "Identification of tumor epithelium and stroma in tissue microarrays using texture analysis," *Diagnostic pathology*, vol. 7, no. 1, pp. 1–11, 2012.

[19] F. Bianconi, A. Álvarez-Larrán, and A. Fernández, "Discrimination between tumour epithelium and stroma via perception-based features," *Neurocomputing*, vol. 154, 2015.

[20] L. N. Rizalputri, T. Pranata, N. S. Tanjung, H. M. Auliya, S. Harimurti, and I. Anshori, "Colorectal histology CSV multi-classification accuracy comparison using various machine learning models," in *2019 International Conference on Electrical Engineering and Informatics (ICEEI)*, 2019.

[21] M. Yazdi and H. Erfankhah, "Multiclass histology image retrieval, classification using Riesz transform and local binary pattern features," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 8, no. 6, pp. 595–607, 2020.

[22] E. F. Ohata, J. V. S. das Chagas, G. M. Bezerra, M. M. Hassan, V. H. C. de Albuquerque, and P. P. Reboucas Filho, "A novel transfer learning approach for the classification of histological images of colorectal cancer," *The Journal of Supercomputing*, pp. 1–26, 2021.

[23] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1, no. 2. MIT press Cambridge, 2016.

[24] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[25] H. Chen *et al.*, "Assessing impacts of data volume and data set balance in using deep learning approach to human activity recognition," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017, pp. 1160–1165.

[26] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *2018 international interdisciplinary PhD workshop*, 2018.

[27] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[29] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the Big Data Paradigm with Compact Transformers," *CoRR*, vol. abs/2104.05704, 2021, [Online]. Available: https://arxiv.org/abs/2104.05704

[30] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, 2020.