[1]**Subrata Sinha**
[2]**Saurav Mali**
[3]**Utpala Borgohain**
[4]**Ujjal Saikia**
[5]**Gunadeep Chetia**
[6]**Smriti Priya Medhi**
[7]**Debashree Borthakur**
[8]**Jayanta Aich**
[9]**Gunjan Mukherjee**
[10]**Dolan Ghosh**
[11]**Rajeev Sarmah**

A Machine Learning Approach for Detection and Classification of Colon Cancer using Convolutional Neural Network Architecture

Journal of Electrical Systems

Abstract: **-** In both developing and developed countries, cancer has become one of the most fearsome health ailments in the 21st century. Among the different variants of cancer, colon cancer is the 3rd most prevalent cancer globally and the 2nd deadliest form of cancer. Early diagnosis is very critical for the successful treatment of colon cancer. In recent times, machine-learning approaches have made significant strides in IoT-assisted healthcare systems. Computer-aided diagnostic systems driven by deep learning algorithms can detect colon cancer with great accuracy, assisting the medical profession in quick diagnosis and developing quick remedies against it. In our approach, we have developed a deep learning model based on convolutional neural network architecture to accurately classify and detect colon cancer. The developed model has been trained with 10,000 histopathological images of the colon, divided into two classes: colon adenocarcinoma and colon benign tissues, each containing 5000 images. We have also employed various performance metrics in our research to monitor the performance of our machine-learning model. The proposed model has been trained for 30 epochs with a batch size of 32 and achieved an overall accuracy of 98.79%.

Keywords: Colon Cancer, Internet of Things (IoT), Artificial Intelligence(AI), Machine Learning(ML), Convolutional Neural Network (CNN).

I. INTRODUCTION

The World Health Organization reports that colon cancer is the second most common cause of cancer-related deaths worldwide and the third most common disease to be diagnosed. More than 1.9 million new cases of colorectal cancer were detected in 2020, and was responsible for more than 930,000 deaths globally, accounting to 9.4% of all cancer-related deaths. [1-3]. By 2040 the burden of colorectal cancer will increase to 3.2 million new cases per year and 1.6 million deaths per year. The yearly incidence rates (AARs) colon cancer in men in India is 4.4 per 100,000 and in women, the AAR for colon cancer is 3.9 per 100,000. In India, colon cancer is ranked ninth among women and eighth among men [4]. Globally, the incidence of colorectal cancer has shown a rising trend, and it is currently 19.5 per 100,000 people, with 15.2 cases in India. Men's colon cancer incidence has increased by 6.5%, while women's incidence has increased by 10.4% [5]. In India, colorectal cancer ranks sixth among cancer types that cause a loss of years of life adjusted for disability [6].

As per the American Cancer Society (ACS), adenocarcinoma is the primary cause of colon cancer, representing about 96% of all colon cancer cases [20]. Present-day cancer screening technique requires a great deal of labor and time. To recognize colon histopathological images, pathologists need to gain a thorough understanding of the field by analyzing labelled histological images. This leads to an extensive waste of resources and becomes quite a labor-intensive task. Consequently, higher diagnostic precision and speed are required [9]. Histopathological examination of specimens (such as glass slides stained with hematoxylin and eosin, or H&E) is typically performed under light microscopy in routine clinical pathology diagnosis. Consequently, in order to process and interpret histopathological images automatically, Medical Image Analysis (MIA) is necessary. Colon cancer can be classified using an MIA system, which can also be used to accurately and objectively analyze the disease's various grades [21].

[11]*Corresponding author: Prof. Rajeev Sarmah, Professor, Department of Biotechnology, Assam down town University, Assam
[1-2,8-10]Department of Computational Sciences, Brainware University, Kolkata-700125
[3-4]Centre for Computer Science and Applications, Dibrugarh University, Dibrugarh-786004, Assam
[5]Administrative Branch, Dibrugarh University, Dibrugarh-786004, Assam, India
[6-7]Assam Don Bosco University, Airport Road, Azara, Guwahati--781017, Assam, India

Whole slide images (WSIs) are digital representations of glass slides that are acquired by specialized scanning equipment [17]. With the introduction of whole slide imaging, machine learning, deep learning, and medical image analysis techniques have been applied to assist pathologists in evaluating WSIs and making cancer diagnoses. Deep convolutional neural networks (CNNs) in particular have demonstrated state-of-the-art performance in numerous computer vision applications [18–19]. Colon cancer has a good prognosis if detected early. Deep learning models based on CNN architecture will help health professionals in early detection of colon cancer, and can formulate a quick treatment plan against the ailment.

Our present work focuses on developing a machine learning model, based on CNN architecture for accurate detection and classification of colon cancer. In due course of our research, we have implemented various optimization techniques to gain a deeper understanding of how different optimization techniques affect the image classification tasks.

II. RELATED WORKS

Recent breakthroughs in intelligent medicine have led to a significant focus on artificial intelligence in recent years [7]. The development of deep learning techniques has been quite useful in computer aided medical diagnosis. Since the past decade, numerous studies have been conducted in developing optimal deep learning models capable of accurately detecting colon cancer. In this section, we shall be discussing some of the deep learning models that have been developed by different researchers for detection of colon cancer.

In the year 2020, Xu et al., [8] developed a deep learning model based on CNN architecture for the detection and classification of colon cancer. The model was trained with a total of 307 histopathological images of the colon and the size of the images was 768x 768 pixels which was reduced to a resolution of about 384x 384 pixels to accommodate memory. Among the 307 images, 85 were of normal tissues, and 222 contained various proportions of colorectal cancer. The dataset was divided into training and testing sets, which were randomly selected. The developed model obtained an optimum accuracy of about 99.9% for normal slides and 94.8% for cancer slides. In the following year, Hasan et al., [9] developed a DCNN model for the detection and classification of colon cancer. Transfer learning was implemented in the proposed model. The model was trained on LC25000 datasets containing about 10000 histopathological images of the colon, divided into 2 classes- adenocarcinoma and benign tumour tissue of the colon. The image resolution of 128x128 pixels was taken, and the model was trained for 50 epochs. The model achieved an optimal accuracy of 99.80%. Further, Tasnim et al., [10] developed multiple deep learning models for colon cancer classification. In their work, they also compared the performances of the developed models with other pre-trained models such as MobileNetV2. The models were trained on 25000 colon histopathological images. The dataset was divided into 2 classes of images each containing 12500 images of cancerous and non-cancerous images. The models were trained for 50 epochs. One of the models used max-pooling as a pooling layer and consisted of 3 convolutional layers and 2 max-pooling layers, and the other model used average pooling as a pooling layer, consisting of 3 convolutional layers and 2 average pooling layers. Both the model achieved quite good accuracy rates of 97.49% and 95.48% respectively.

In 2022, Talukdar et al., [11] developed a hybrid ensemble feature extraction model for the detection of lung and colon cancer. The model was trained on the LC25000 colon dataset consisting of about 2800 colon histopathological images, which was later augmented to about 10000 images, divided into 2 classes: colon adenocarcinoma and benign colon tissues. The empirical findings on the LC25000 suggested that the developed model achieved an accuracy of 96.61%. for colon cancer identification. Furthermore, Sakr et al., [12] proposed multiple deep-learning models based on CNN architecture to detect colon cancer. The proposed models were trained on the LC25000 Lung and colon histopathological images dataset. The colon dataset was divided into 2 classes -benign and adenocarcinoma, each class containing 5000 images. The resolution of the images was 768x768 pixels which were reduced to 180x180 pixels to lessen the training time of the models, and the dataset was divided into training sets (80%) to train the models, 10% of the images were used for testing and the rest for validating the models. Several layers were varied in the models that were developed. Out of the 6 models developed, the 6th model which consisted of 12 layers, 8 convolution layers and 2 max pool layers achieved the best accuracy rate of 99.50%. Also, Ho et al., [13] proposed a deep learning algorithm comprising a Faster Region-Based Convolutional Neural Network (Faster-RCNN) architecture for instance segmentation with a ResNet-101 feature extraction. The model was trained on a dataset containing 66,191 histopathological images which were

obtained from H&E-stained colonic specimen slides. The specimens belonged to Singapore General Hospital's pathology archives and The Cancer Genome Atlas (TCGA). The model achieved an AUC of 91.7%.

In 2023, Sakthipriya et al., [14] developed a customed CNN model for classifying colon cancer. The developed model was based on other pre-trained models such as VGG16, ResNet50 and InceptionV3. It achieved an accuracy of 93.6%. In the same year, Karthikeyan et al., [15] proposed a deep learning model based on CNN architecture and LSTM algorithm. The model was trained on a dataset of 334 images divided into 3 tumor classes. Data augmentation was performed in the proposed method, to increase the size of the dataset. The proposed model was trained for only 8 epochs and obtained an accuracy of 91%.

In the above cited literature, most of the models that have been discussed have either been developed on pre-existing deep learning models such as VGG19, VGG16, ResNet50, etc, or have been trained with substantially low number of images. In our work we have developed a novel model for colon cancer detection using colon histopathological images. Substantiality number of layers has been added to the proposed model for optimal prediction of colon cancer.

III. MATERIALS AND METHODOLOGY

A. Dataset

In our approach, we have trained our machine learning model with a dataset containing 10000 histopathological images of the colon. The dataset was divided into 2 classes: colon adenocarcinoma and normal benign tissue of the colon, each containing 5000 images respectively. We have downloaded the dataset from Kaggle, an online data science repository from google. The original dataset (LC25000) consisted of 25000 histopathological images of 2 variants of cancer: lung cancer and colon cancer. In our work, we have only considered 2 classes of colon histopathological images [16].

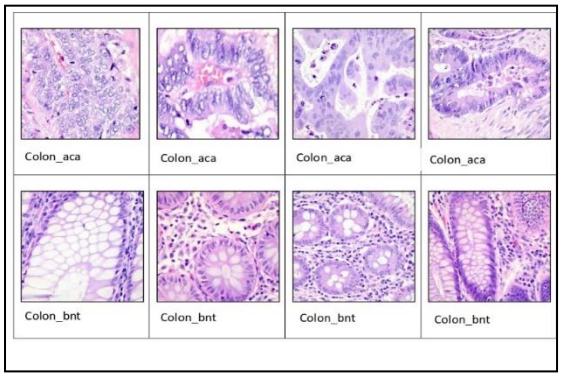

Figure 1: Class wise sample images from the dataset

B. Proposed CNN architecture

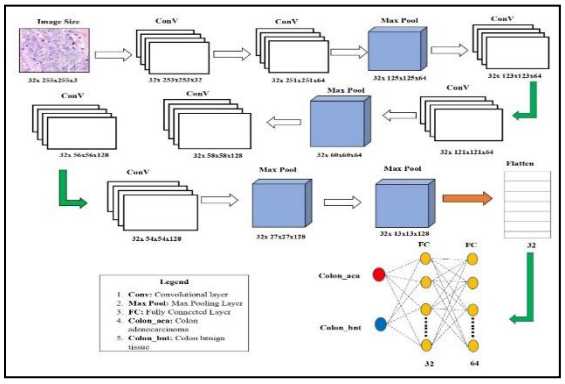

Figure 2: CNN architecture of the proposed model

The CNN model used for this study had 7 convolutional layers, 4 maxpool layers, 2 fully connected layer, and 1 output layer. The layer design followed a schema of: [{(2 Conv x 1 MaxPool)* 2}+ {(3 Conv x 2 MaxPool)* 1}+ 1Dropout + 2 FC + 1 Output]. During training, 80% of the images were used for training, 10% for testing, and 10% for validation. The kernel size used was 3x3, which performs convolution on the input image through

element-by-element matrix multiplication. To reduce the feature map's size, pooling layers were added after the convolutional layers to accommodate computational memory. Following numerous convolutional and pooling layers, 2 fully connected layers were added (To reduce the amount of human supervision required) to connect neurons between two different layers using weights, biases, and neurons. The flattened vector of the input image is then processed through those fully connected layers, where standard mathematical operations occur to initiate the classification process. Finally, the SoftMax activation function was applied to the output layer to determine which model inputs should fire in the forward direction and which ones should not.

C. *Proposed Colon Cancer Classification Model*

The dataset containing 2 classes of histopathological images of the colon were embedded into our proposed CNN model. The proposed CNN model was trained, tested and validated on a dataset of histopathological images. 80% of the dataset was divided into training set, 10 % for testing the model and the remaining 10% for validation set, to validate the performance of the model, during training phase. We have performed data augmentation by means of random flipping and rotation, to increase the size and diversity of the dataset, which shall help the model to generalize better to new and unseen data. The batch size was set to 32. Adam optimizer was used to lessen the loss, occurring during training of the model and the learning rate was set to 0.0001. The model was trained for a period of 30 epochs.

D. Experimental Setup

Our proposed model has been trained with a total of 10000 histopathological images of colon divided into 2 classes. Then the dataset was divided into 3 categories, training set consisted of 80% of the images, used to train the model. The validation and testing dataset consisted of 10% each of the overall dataset. The size of the images was reduced to 150x 150 pixels to accommodate computational memory.

The Google Collab setup (GPUNvidia V100, GPU Memory: 16 GB) along with 56 GB of RAM was used to carry out our experiments, Python 3.9 was required for our experiment, and the CNN model was built with Tensorflow v2.12 as the back end and Keras Framework v2.11.0 as a high-level API.

IV. RESULTS AND DISCUSSIONS

A. *Classification Curve of the Proposed Model*

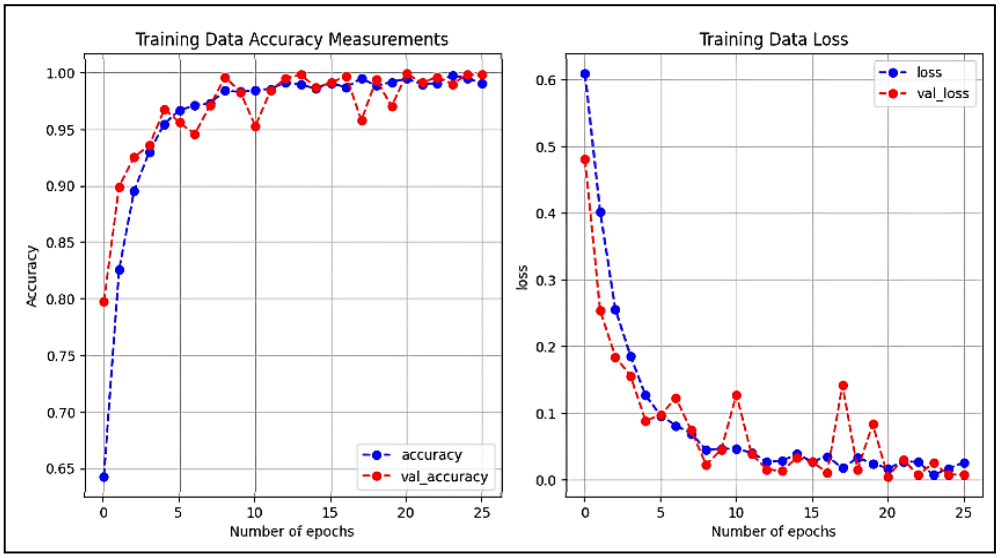

Figure 3. *(a)Training and Validation accuracy graph (b) Training and validation loss graph of the proposed model*

Figure 3 depicts the training performance of our proposed model. The iteration is continued for 30 epochs, and accuracy along with loss performance was monitored throughout the training phase of the proposed model. With increase in number of epochs, improvement in training accuracy along with validation accuracy was observed throughout the training phase. The training accuracy along with validation accuracy increased simultaneously with a slight variation. The training loss also decreased gradually, with increase in the number of epochs, depicting that the model has been trained quite well.

B. Confusion Matrix of the Proposed Model

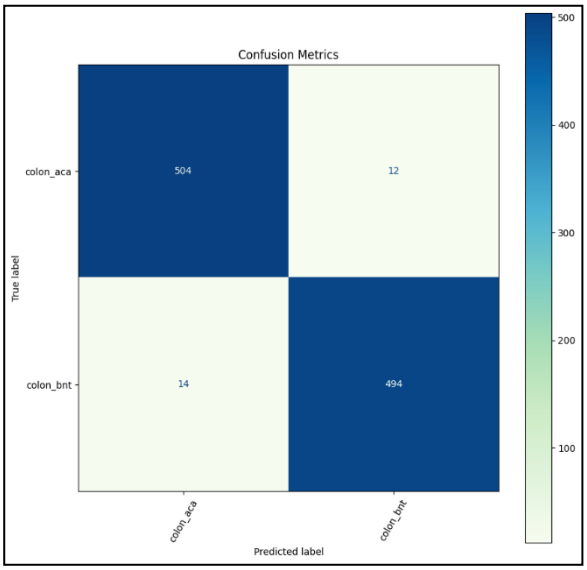

Fig 4: Confusion Matrix of the Proposed Model

Fig. 4 depicts the total instances of true positives, true negatives, false positives, and false negatives that our model has correctly classified on a dataset containing two classes of colon images, one of the is malignant. From the figure, we can see that our model has correctly predicted 504 instances of adenocarcinoma (colon_aca) images as true positives. In the other class, a total of 494 instances of colon benign normal tissues were correctly predicted as colon benign normal tissues, and 14 instances of benign normal tissue images were incorrectly predicted as adenocarcinoma. We have calculated the different assessment metrics from the total instances of TP, TN, FP, and FN presented in the confusion matrix and is discussed below.

Table 1: Proposed Model performance on colon cancer dataset with Adam optimizer (with learning rate 0.0001).

Class	**Assessment Metrics**						
	Accuracy (%)	**Precision (%)**	**Sensitivity (%)**	**Specificity (%)**	**F1-S (%)**	**NPV (%)**	**MCC (%)**
Colon_aca	97.46	97.30	97.67	97.24	97.49	97.63	94.92
Colon_bnt	97.46	97.63	97.24	97.67	97.44	97.30	94.92

In table 1, calculation of various performance metrics such as accuracy, precision, sensitivity, specificity, NPV and MPV have been presented to evaluate the class-wise performance of the proposed model. The model was trained for 30 epochs with a batch size of 32. Adam optimizer was used to lessen the loss, occurring during training of the model and a learning rate of 0.0001 was set. From the table we observe that our proposed model was able to achieve quite optimal performance in all of the assessment metrics for two classes of colon images.

C. Confidence Report of the proposed model

The confidence report of our colon cancer prediction model provides important insights into the accuracy and reliability of the model's predictions. 6 randomly colon histopathological images were picked to develop the confidence report. In the 1st instance, we can see that our model has correctly classify the class containing benign normal tissue images as benign normal tissues with a confidence of 100%. Further, actual class of colon adenocarcinoma was also correctly predicted as adenocarcinoma with a confidence of 100%. The same result was also observed in case of the other 4 presented images in the confidence report.

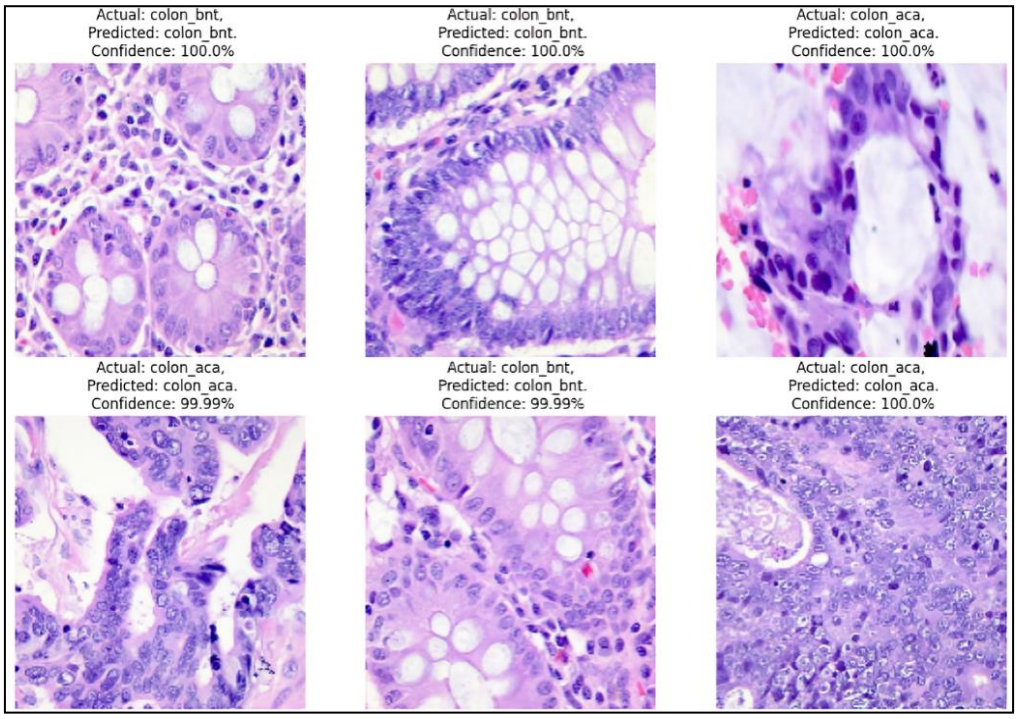

Figure 5: Confidence Report of the Proposed Model

V. CONSLUSION

Colon cancer causes a significant amount of distress to both the physical and mental health of the patient suffering from it. It is the 2nd most leading cause of cancer death occurring in the world. Early screening of this cancer is only the effective way to prevent, and early detection so that necessary remedies can be adopted against the disease. With the application of deep learning in medical imaging, screening has become more efficient and has been assisting physicians in detection of the cancer at an early stage, thus improving the overall prognosis of the disease. In our work, the proposed model has achieved a remarkable accuracy of 98.79% and was quite successful in classifying the two classes of histopathological colon images, with one class being malignant adenocarcinoma. The prediction accuracy in our proposed work is observed to be better than most of the above-cited models.

In the future, we shall be performing various hyper-parameter optimization techniques in our approaches to overall improve the performance of the machine learning models. Further, we shall also try to train the models with a more diverse dataset containing multiple classes of colon cancer images, to make our model more robust so that, it can classify different types of colon cancer. And also try to implement the deep learning models into an AI-based mobile environment, to make it accessible to general public.

ACKNOWLEDGMENT

S.S thanks university administration for giving all possible support to conduct the research work.

REFERENCES

[1] World Health Organization, "Global Cancer Observatory colon fact sheets," 2020. [Online]. Available: https://gco.iarc.fr/today/online-analysis-pie.

[2] World Health Organization, "Cancer," July 14, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer.

[3] J. Ferlay et al., "Global Cancer Observatory: Cancer Today," 2020. [Online]. Available: https://gco.iarc.fr/today.

[4] National Cancer Registry Programme, "Three-year report of the population-based cancer registries- 2009-2011," Indian Council of Medical Research (ICMR), Bangalore, India, 2013.

[5] B. Joseph, S. Ghosh, V. Dhakad, and S. Desai, "Trends of colorectal cancer in central India: An institutional review," Indian Journal of Applied Research, pp. 63-64, 2020. [Online]. Available: https://doi.org/10.36106/ijar/1712280.

[6] J. Shaik, "Economic burden of cancer and their variations along with incident trends, challenges, and opportunities in India," Authorea, vol. 1, no. 8, 2020.

[7] X. Gu et al., "Brain tumor MR image classification using convolutional dictionary learning with local constraint," Frontiers in Neuroscience, vol. 15, p. 679847, 2021.

[8] L. Xu et al., "Colorectal Cancer Detection Based on Deep Learning," Journal of Pathology Informatics, vol. 11, p. 28, 2020. [Online]. Available: https://doi.org/10.4103/jpi.jpi_68_19.

[9] M. I. Hasan, M. S. Ali, M. H. Rahman, and M. K. Islam, "Automated Detection and Characterization of Colon Cancer with Deep Convolutional Neural Networks," Journal of Healthcare Engineering, vol. 2022, Article ID 5269913, 12 pages, 2022. [Online]. Available: https://doi.org/10.1155/2022/5269913.

[10] Z. Tasnim et al., "Deep Learning PredictionModel for Colon Cancer Patient using CNN- based Classification," International Journal of Advanced Computer Science and Applications, vol. 12, no. 8, pp. 687-696, 2021.

[11] R. R. Talukdar et al., "Hybrid ensemble feature extraction model for efficient colon cancer identification," Journal of Ambient Intelligence and Humanized Computing, vol. 13, no. 6, pp. 5783-5794, 2022.

[12] A. S. Sakr et al., "An Efficient Deep Learning Approach for Colon Cancer Detection," Applied Sciences, vol. 12, p. 8450, 2022. [Online]. Available: https://doi.org/10.3390/app12178450.

[13] C. Ho et al., "A promising deep learning-assistive algorithm for histopathological screening of colorectal cancer," Scientific Reports, vol. 12, no. 1, p. 2222, 2022.

[14] N. Sakthipriya et al., "Deep Learning-Based Colon Cancer Classification Using Pre-Trained Custom Convolutional Neural Network with Histopathological Images," International Journal of Creative Research Thoughts (IJCRT), vol. 11, no. 4, pp. c397-c401, 2023.

[15] A. Karthikeyan, S. Jothilakshmi, and S. Suthir, "Colorectal cancer detection based on convolutional neural networks (CNN) and ranking algorithm," Measurement: Sensors, p. 100976, 2023.

[16] A. A. Borkowski et al., "Lung and Colon Cancer Histopathological Image Dataset (LC25000)," arXiv:1912.12142v1 [eess.IV], 2019. [Online]. Available: https://www.kaggle.com/datasets/biplobdey/lung-and-colon-cancer.

[17] S. Mukhopadhyay et al., "Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: A multicenter blinded randomized noninferiority study of 1992 cases (pivotal study)," American Journal of Surgical Pathology, vol. 42, p. 39, 2018.

[18] K. He et al., "Delving deep into rectifers: Surpassing human-level performance on imagenet classifcation," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.

[19] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7640, pp. 115, 2017.

[20] B. Levin et al., "Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: A joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology," Gastroenterology, vol. 134, pp. 1570–1595, 2008. [Online]. Available: https://doi.org/10.1053/j.gastro.2008.02.002.

[21] M. Dal Molin et al., "Clinicopathological correlates of activating GNAS mutations in intraductal papillary mucinous neoplasm (IPMN) of the pancreas," Annals of Surgical Oncology, vol. 20, pp. 3802–3808, 2013. [Online]. Available: https://doi.org/10.1245/s10434-013-3096-1.