An Enhancement Technique to Diagnose Colon and Lung Cancer by using Double CLAHE and Deep Learning

Nora yahia Ibrahim, Amira Samy Talaat
Computers and Systems Department
Electronics Research Institute, Cairo, Egypt

***Abstract*—The most common and deadly cancers are lung and colon cancers. More than a quarter of all cancer cases are caused by them. Early detection of the disease, on the other hand, greatly raises the probability of survival. Image enhancement by Double CLAHE stages and modified neural networks are made to improve classification accuracy and use Deep Learning (DL) algorithms to automate cancer detection. A new Artificial Intelligent classification system is presented in this research to recognize five kinds of colon and lung tissues, three malignant and two benign, with three classes for lung cancer and two classes for colon cancer, based on histological images. The results of the study imply that the suggested system can accurately identify tissues of cancer up to 99.5%. The use of this model will aid medical professionals in the development of an automatic and reliable system for detecting different kinds of colon and lung tumors.**

***Keywords*—Artificial intelligent system; machine learning; cancer detection; image classification; deep learning**

I. Introduction

Cancer is one of the top causes of death worldwide, according to the Organization of World Health. Autonomous growth, genetic instability, and substantial metastatic potential are acquired by cancer cells. Colon and lung are the most affected organs, with the largest number of deaths. Colon cancer is the major cause of 9.2% of cancer mortality worldwide, while lung cancer is the major cause of 18.4% of all cancer mortality [1, 2]. The combined frequency of colon and lung cancer is estimated to be around 17%. Although this is improbable, cancer cell spread across these two organs is highly common in the absence of early diagnosis [3].

Only effective treatment and early detection can now minimise cancer deaths [4]. The faster a patient is diagnosed, the more effective the treatment and the better the patient's chances of survival and healing.

To search for cancer cells and rule out other probable diseases, many tests are performed, including sputum cytology, imaging sets (CT scan, x-ray), and biopsy (tissue sampling). The examination of microscopic histopathology slides by trained pathologists while performing the biopsy is important in determining the diagnosis [5, 6] and identifying tumour forms and subsets [7]. This study uses just histopathology images to diagnose colon and lung cancers automatically.

Health specialists frequently employ histopathological images for analysis, and they are crucial in determining the survival chances of patients. Usually, health specialists had to go through a lengthy process to diagnose cancer by reviewing histopathological images. However, with the technological tools accessible now, this process may be completed with less time and effort [3]. Artificial intelligence systems have recently gained popularity for their capability to analyze data quickly and give conclusions.

II. Literature Review

In biomedical applications, machine learning techniques are used to predict and classify various types of signals and images. Machines can now deal with large-scale data such as anatomical multidimensional videos and images because of deep learning (DL) methods. Deep learning is a machine learning field that builds algorithms to produce an artificial neural network built on the human brain's structure and function [8]. The majority of previous research used DL to categorise lung and colon cancer images simultaneously. Some writers concentrated on colon cancer detection, while others concentrated on lung cancer detection.

A deep learning-based algorithm is used by Masud [9] to classify colon and lung histological images. They used two types of domain modifications to obtain four image categorization feature sets. They joined the properties of the two categories to arrive at the final categorization result. They were 96.33% accurate. By employing a shallow neural network design, Mangal [10] was able to classify colon and lung cancers based on histological images. In classifying lung and colon malignancies, they reached an accuracy of 97% and 96%, respectively.

Hatuwal [11] proposed a deep learning method based on CNN. In the method, they present samples of only lung tissues from the dataset. This approach could only identify two malignant and one benign tissue in the lung, and no information on colon cancer categorization was given. Their suggested lung tissue categorization model attained an accuracy of 97.20%, a recall of 97.33%, and a precision of 97.33%. Sarwinda [12] suggested a classifier of KNN with characteristics retrieved for colon tissues by a DenseNet-121 pretrained network. Their approach mines the information for colon tissues and distinguishes between benign and malignant tissues of the colon. For colon categorization, their

model achieved 98.53% accuracy and 98.63% recall. Their model, however, was unable to collect lung tissues and provided no information about lung classification. According to Kumar [13], DenseNet-121 extracts more significant characteristics than other CNN pre-trained networks. This is because of the use of small links to improve the accuracy and efficiency of the network. Wang [14] built a Python library based on deep learning to detect cancer image categories. In their proposed strategy, they combined the CNN model and the SVM algorithm. The SVM model's overall accuracy was 94%.

Chehade [15] identifies colon and lung cancer subtypes, and the model of XGBoost offers the best classification rate in terms of recall, accuracy, and precision. XGBoost had a 99% accuracy and a 98.8% F1 score.

Hlavcheva [16] employed convolutional neural networks to analyze medical images using deep learning techniques. The dataset was used to compare the accuracy of several CNN designs in classification. The accuracy of 94.6% was achieved using neural network theory and statistical mathematical methodologies.

The study's primary goal is to develop a medical analytical intelligent support system for colon and lung imaging and, using machine learning, develop an automated method for properly classifying the subtypes of lung and colon cancer from histopathological images so we can achieve high levels of accuracy.

The following is a summary of the contributions of this paper:

- We proposed a novel colon and lung Image classification technique by applying the Image enhancement technique combination of DWT (discrete wavelet transform) and Double-CLAHE (Double Contrast Limited Adaptive Histogram Equalization) in the Preprocessing phase of the image.
- CLAHE is applied twice, first for the low frequency decomposed part of the Image DWT component and after the inverse DWT of the reconstructed image, which makes image details more enhanced.
- We proposed a new hybrid combination of an enhanced image from DWT with Double-CLAHE, EfficientNetB7 Deep learning technique, and adding Modified Neural Network method to fully discover the multi-class deep-broad characteristics of the colon and lung Image dataset.
- The proposed method demonstrates outstanding improvement in the performance for the training and testing datasets and gives a very high classification accuracy of 99.5%.

The following is how the article is organized. Section III discusses the datasets on the colon and lungs. Section IV Methodology with implementation details and results evaluation. Section V of Experimental results, comparisons, and conclusion

III. Dataset on Colon and Lung

The proposed technique is tested using the LC25000 dataset [17], a new colon and lung cancer histopathology image dataset that was published in 2020. This collection, which was put up by Andrew A. Borkowski and his colleagues, has 25000 colour images of five lung and colon tissues of different types [18], namely Benign Colonic Tissue, Benign Lung Tissue, Colon Adenocarcinoma, Lung Squamous Cell Carcinoma, and Lung Adenocarcinoma. Table I shows the details of the dataset as well as the allocated class names.

TABLE I. The Details of the LC25000 Database

Cancer Type	***Name of Category***	***Number of Images***
Colonic _Benign _Tissue	Col_Be	5000
Lung _Benign _Tissue	Lun_Be	5000
Colon _Adenocarcinoma	Col_Ad	5000
Lung _ Carcinoma _Squamous _Cell	Lun_Sc	5000
Lung _Adenocarcinoma	Lun_Ad	5000
Total	5	25000

Adenocarcinoma is the most frequent type of colon cancer, accounting for more than 95% of all cases. When a form of polyp (tissue growth) called an adenoma grows in the large intestine, it becomes an adenocarcinoma and progresses to cancer. Lung adenocarcinoma makes up around 40% of all lung tumors, and it affects more females than males. This form of cancer generally starts in glandular cells and spreads to the lungs' alveoli. All tumours that grow in the colon and lungs are not malignant and do not travel to other regions of the body.

These tumours are classified as benign, and they aren't usually fatal. They must, however, be removed surgically and biopsied to determine if malignancy is present. Lastly, lung carcinoma squamous cell is a type of small cell tumour that arises in the airways or bronchi of the lungs. It is the second most frequent kind of lung cancer, accounting for roughly 30% of all cases. Only 500 images of the colon and 750 images of the lung are included in the original LC25000 dataset. They enlarged the dataset to 25,000 images by using augmentation strategies to flip and rotate the original images under various situations (each class has 5000 images).

The original images were 1024 x 768 pixels in size. However, to make them square, they were resized to 768 x 768 pixels before using the augmentation methods. Sample histopathology images from the LC25000 dataset from these five classes are shown in Fig. 1.

IV. Methodology

This section describes the suggested deep learning-based classification method for colon and lung cancer diagnosis. The Convolutional Neural Network is a method for distinguishing cancers from other cells or tissues that has been shown to be effective [19].

The EfficientNet-B7 network was fine-tuned in this paper to classify colon and lung tumours. Histopathology images are

shown in Fig. 2. The proposed method structure consists of three main stages: image pre-processing with enhancement, EfficientNetB7, and the Modified Neural Network (MNN) stages, as shown in Fig. 2.

EfficientNetB7 takes the resized images from the first stage Fig. 3 to train these images for solving the classification problem. EfficientNetB7 is also used to extract the feature maps of colon and lung cancer histopathology images. However, the MNN takes the extracted feature from the second stage as input and a class label as output.

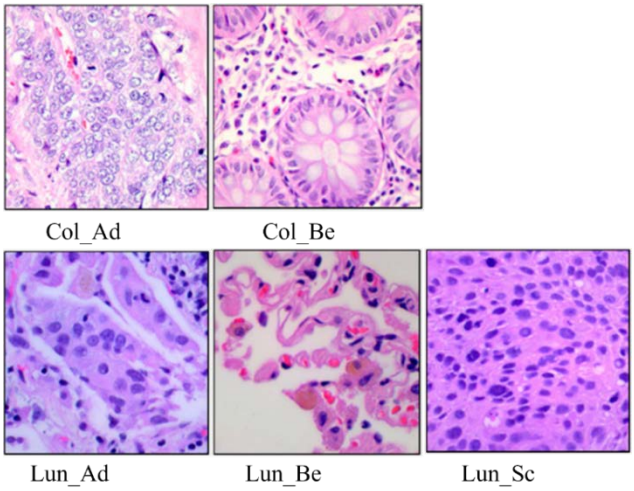

Fig. 1. LC25000 Dataset Sample Images.

The details of the proposed method with MBConv (Mobile Inverted Bottleneck Convolution), the resolution, number of channels, number of levels of each feature map, and the details of the MNN stage are also shown in Fig. 2. In the following subsection, each stage will be described in detail.

A. *Stage of Image Pre-Processing using DWT and Double-CLAHE*

The pre-processing image stage is required before the feature extraction procedure to prepare and clarify the images with labels for training the model.

Blurriness, poor border recognition, artifacts, and overlapping problems in histopathology images were caused by uneven staining of the slide because of human error.

As shown in Fig. 3, The Double-CLAHE approach (Double Contrast Limited Adaptive Histogram Equalization) is intended to eliminate these types of imperfections or uneven staining. The CLAHE method improves image contrast by increasing poor boundary edges in each pixel of an image through restricted amplification [20], as well as improving local contrast in an image. As a result, it's ideal for enhancing the features of histopathological images. This paper proposes a new image enhancing method that combines CLAHE and DWT (Discrete Wavelet Transform). Preprocessing of images in Fig. 2 was done using the DWT and CLAHE approaches in Fig. 3.

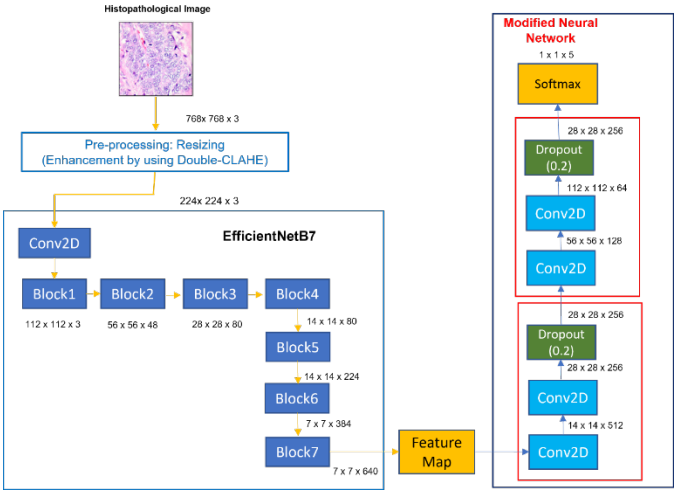

Fig. 2. The Proposed Framework for Image Classification.

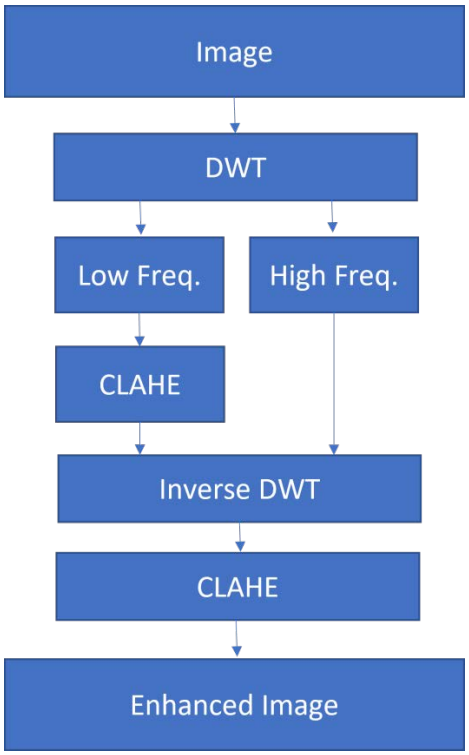

Fig. 3. The Steps of Enhancement by DWT and Double-CLAHE.

The new technique consists of four major steps: The original image is first decomposed into high-frequency and low-frequency components by DWT. The low-frequency values are then boosted by CLAHE while the high-frequency values are left untouched to limit noise amplification. This is because the high-frequency component refers to detailed information and comprises the majority of the original image's noise. Third, reconstruct the image using the inverse DWT of the new coefficients. Finally, after obtaining the reconstructed image, CLAHE is required to enhance the image in order to make the details more abundant.

The colon and lung cancer histopathology images are resized from 768x768 to 224x224 in RGB format to train the suggested model with the dataset.

B. Stage of EfficientNetB7

One of the most powerful CNN structures is EfficientNet. It employs a compound scaling strategy to increase network depth, width, and resolution, resulting in good capacity in a variety of benchmark datasets while using fewer computational resources than other models [21].

EfficientNets come in eight different models, from EfficientNet-B0 to EfficientNet-B7. The simplest model, EfficientNet-B0, is designed automatically by the Neural Architecture Search. Using the compound scaling method, the EfficientNet family is created by scaling up EfficientNetB0. Scaling the network increases model performance by balancing all architecture image resolution, depth, width, and compound coefficients.

Excitation optimization and squeeze in mobile inverted bottleneck convolution (MBConv) [22] is the core of the EfficientNet architecture. Fig. 4 depicts the MBConv concept.

The number of MBConv blocks in the EfficientNet network family varies. The depth, width, resolution, and model size keep increasing as EfficientNetB0 through EfficientNetB7 improve, as well as the accuracy [21]. Efficient-NetB7 exceeds previous CNNs on ImageNet in terms of accuracy, and it is furthermore 6.1x faster and 8.4x smaller than the best available CNN [21]. MBConv is the fundamental building block of the network. The filter size identifies each MBConvX block. It corresponds to X=1 and X=6 which represent the standard ReLU and ReLU6 activation functions, respectively. Fig. 5 depicts the characteristics of the seven blocks' architecture.

Flattening the extracted feature-maps yields a single vector of features once the features are extracted from the dataset images. A Modified Neural Network stage takes this vector as its input.

C. The Modified Neural Network Stage

It is the last stage of the proposed method of the classification process. As clarified in Fig. 2, the Modified Neural Network stage contains four convolution layers, two dropout layers, and a softmax layer. The softmax layer (output layer) of the proposed method is customized with the number of our classes.

The aim of this phase is to add variety to the extracted knowledge and assist it to have a better understanding of the samples, allowing them to be categorized more accurately.

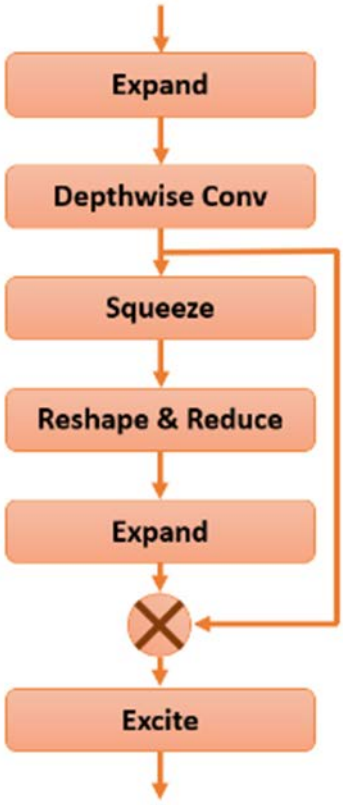

Fig. 4. EfficientNet Basic Building Block (MBConv).

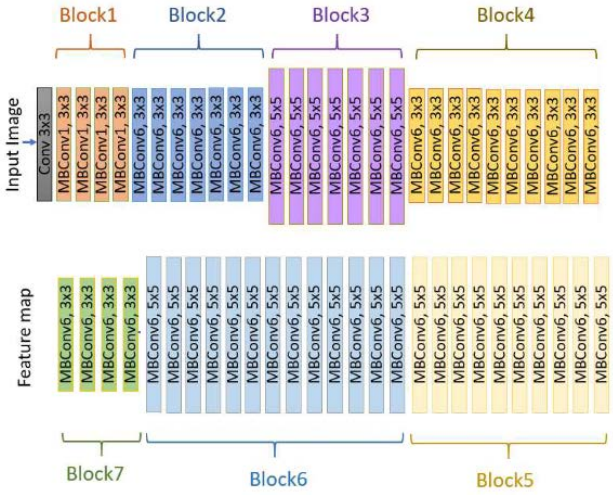

Fig. 5. The Architecture of Block1 to Block 7 of EfficientNetB7.

V. Experimental Results Using

A. The Implemented Details

The obtained outcomes of the implemented experiment are mentioned in this section. The input dataset is split into 85:15, with 85% of the images (randomly chosen) for training and

15% for validating. Because our dataset is balanced (every class contains the same number of images), the system will be less subject to bias while making decisions.

TABLE II. THE IMPLEMENTED DETAILS OF THE PROPOSED CNN MODEL FOR CLASSIFICATION TASK

Variable	***Value***
Image dimensions	224 x 224
Initial channels	3
Dropout	20%
Batch Size	64
Epochs	22
Convolutional layer activation	Relu
Learning rate	0.001
activation of Dense layer	Softmax
Compiler-optimizer	Adam
Compiler-loss	Categorical-cross-entropy

The proposed model was developed using Tensorflow 2.0. As shown in Table II, the system is trained on an image with a size of 768x768 pixels by scaling it to 224x224 pixels, using a batch size of 64 and 22 epochs. For initializing the training, the weights that have been pre-trained by EfficientNetB7 on ImageNet are used, and they are fine-tuned. For training, the ADAM optimizer with a learning rate of 0.001 and a categorical-cross-entropy loss function is utilised. The proposed framework's performance in a classification problem is measured using accuracy, average precision (AP), average recall (AR), and the F1 measure, which will be discussed in the next section.

B. Evaluation of Performance

Machine learning models are evaluated using a variety of criteria. The confusion matrix, as well as associated metric factors like precision, F1-score, accuracy, and recall, are utilized to measure in this paper.

The classifier's accuracy is a measure of its capability to correctly classify instances. It refers to the percentage of valid results or correctly identified samples among all samples.

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

The True Negative, True Positive, False Negative, and False Positive values are represented by TP, TN, FP, and FN, respectively. TP denotes true disease, i.e., the true value is positive, and it is classified positively, indicating that the patient has the disease and that the test is positive.

A false Negative (FN) shows that the patient has the disease while the test is negative, suggesting that the real value is positive but the classification is negative. A False positive (FP) denotes the presence of a disease when none exists, implying that the real value is negative when classed positively. A True Negative (TN) denotes that the patient is healthy and the test is negative, signifying that the true value is negative, and the test is negative.

Precision is denoted as the proportion of correctly identified samples (true positives) to positive samples identified.

$$\text{Precision}=\frac{TP}{TP+FP} \quad (2)$$

Recall is the percent of positive samples of a specific class that are accurately identified. The proportion of real positive samples to total positive samples is used to compute it.

$$\text{Rec}all=\frac{TP}{TP+FN} \quad (3)$$

The F1-score is known as the harmonic average of accuracy and recall.

$$\text{F1-score}=\frac{2\times Precision\times Recall}{Precision+Recall}$$

C. Results

The outcome of the proposed framework for the classification of colon and lung cancer histological images is shown in this section. To determine their performance, the models were evaluated using test data.

Fig. 6 shows the proposed model's each epoch classification accuracy. The experiment has 22 epochs in all. The classification accuracy on the testing subset was 99.36% at the last epoch; however, the greatest results were at epochs 12, 13, and 14, all of which had a 99.47% accuracy. The training accuracy curve increased gradually and almost steadily towards the top, as shown in the figure.

At epoch number 18, the greatest training accuracy was 99.5%, which is quite similar to the accuracy of the previous epoch 99.36%. The curve of testing accuracy is similar to the training accuracy curve, with the outcome improving as the training progresses. At 20 epochs, the curve drops to 96.7%, indicating that the model is able to give a satisfactory classification result even if it is constructed with fewer epochs.

Fig. 7, on the other hand, shows the training and validation loss, which represents the percentage of data loss for each classification attempt.

As seen in Fig. 7, both the training and validation subsets' loss values decreased as the number of epochs grew.

The normalized (ROC) Receiver Operating Characteristic and confusion matrix curves of the testing subset classification at the 18th epoch are shown in Fig. 8. For the test data given labelled categories, the confusion matrix compares the images' true labels against their predicted labels. Only 3% (112 samples) of the testing images (3750 samples) were misclassified, as shown by the normalized confusion matrix.

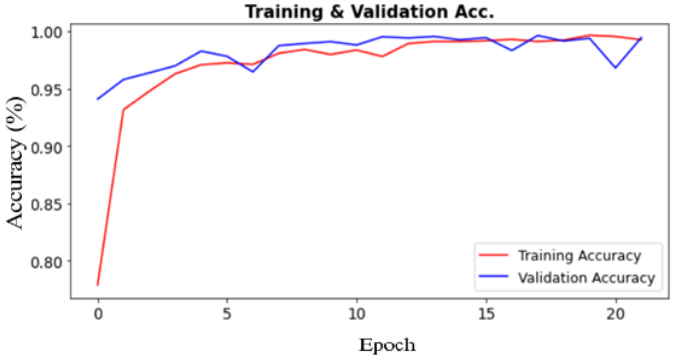

Fig. 6. The Visual Display of the Accuracy Rate of the Proposed Classification Model at Each Epoch.

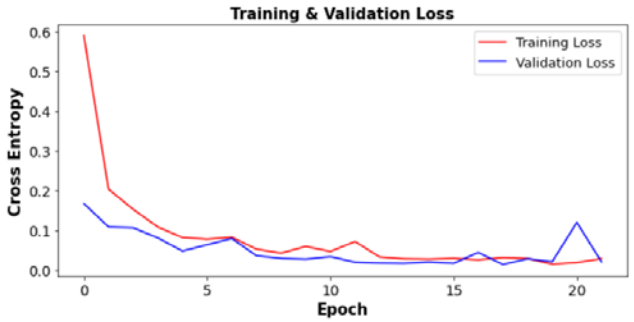

Fig. 7. The Proposed Classification System's Optimal Training and Validation Loss.

The best classification results are in the Col_Be and Lun_Be categories, while the other categories, Lun_Ad, Col_Ad, and Lun_Sc, have the same misclassification rate. These results can also be seen in the ROC curves.

Because the classifier was quite successful at separating the samples, the Lun_Be and Col_Be curves have reached the top-left corner. Overall, the suggested deep learning approach is highly precise in classifying these classes.

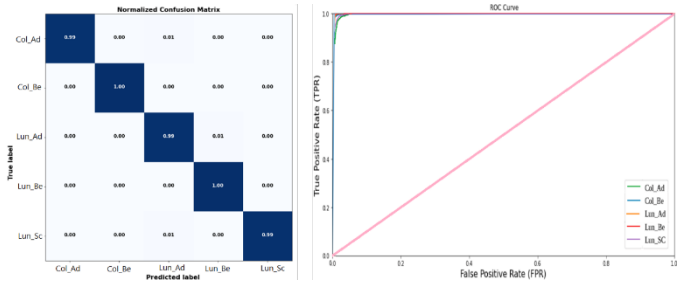

Fig. 8. Representations of Classification Results: (a) Normalized Confusion Matrix (b) Epoch 18th ROC Curve.

Table III displays the recall, F1-score, and precision of the proposed classification model on the test data for five classes of histological images. Table III shows that the average recall, precision, and F1-score for each of the five categories is more than 0.994. In addition, except for Lun_Sc, our classification approach attained the highest precision in all classes.

TABLE III. THE RECALL, PRECISION, AND F1-SCORE FOR HISTOLOGICAL IMAGES OF COLON AND LUNG CANCER

Categories	*Precision*	*Recall*	*F1-score*
Col_Ad	**1.00**	0.995	**1.00**
Col_Be	**1.00**	**1.00**	**1.00**
Lun_Ad	**1.00**	0.98	0.99
Lun_Be	**1.00**	**1.00**	**1.00**
Lun_Sc	0.98	**1.00**	0.99
Average	0.996	0.994	0.996

D. Comparison

We compare our model to current models in the literature, which are given in the introduction, to evaluate the proposed method. Table IV compares the results of the lung and colon cancer subtype classification with other approaches using the same dataset. As shown in Table IV, our system outperforms existing cancer detection technologies in terms of maximal classification accuracy.

TABLE IV. COMPARISON OF THE RESULTS FOR THE SAME DATASET WITH OTHER METHODS

References	*ClassifierModel*	*(%) Accuracy*	*(%) Precision*	*(%) Recall*	*(%) F1-score*
Masud, et al. [9]	CNN	96.33	96.39	96.37	96.38
Mangal [10]	CNN for lung Cancer	97.89	-	-	-
Mangal [10]	CNN for colon Cancer	96.61	-	-	-
Hatuwal [11]	CNN for lung Cancer	97.20	97.33	97.33	0.96
Sarwinda [12]	DenseNet-121-KNN for colon Cancer	98.53	-	98.63	-
Kumar [13]	DenseNet-121-DF	98.60	98.63	98.60	-
Wang [14]	CNN & SVM	94			90
Chehade [15]	XGBoost	99	98.6	99	98.8
Hlavcheva [16]	CNN-D	94.6	-	-	-
Proposed Method	The proposed classifier	**99.5**	**99.6**	**99.4**	**99.6**

E. Conclusion and Future Work

A deep learning technique is presented in this paper to classify images and will help us detect colon and lung cancer more precisely in the future. For this study, we utilised a histopathology image dataset that is freely accessible on Kaggle [17]. The model training accuracy achieved is 99.5% for the colon and lung dataset.

This paper proposed a method for colon and lung cancer classification problems. The method has two main sections: the enhancement of images by DWT and Double-CLAHE stages; and the modified neural network to enhance classification accuracy.

The result shows that: accuracy 99.5%, precision 99.6%, recall 99.4%, and F1-score 99.6%, which proves that the presented method is effective for solving colon and lung cancer classification problems. It also outperforms the previous approaches in terms of performance.

The optimized method that provided improved accuracy must be used in upcoming models. In the future, combining YOLO, 3D-CNN, and a variety of other approaches that are applied to various image datasets will allow us to construct more powerful and effective models in the future.

REFERENCES

[1] Bray, F., et al., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 2018. 68(6): p. 394-424.

[2] Bermúdez, A., et al., Her2-Positive and Microsatellite Instability Status in Gastric Cancer—Clinicopathological Implications. Diagnostics, 2021. 11(6): p. 944.

[3] Toğaçar, M., Disease type detection in lung and colon cancer images using the complement approach of inefficient sets. Computers in Biology and Medicine, 2021. 137: p. 104827.

[4] Sánchez-Peralta, L.F., et al., Deep learning to find colorectal polyps in colonoscopy: A systematic literature review. Artificial intelligence in medicine, 2020. 108: p. 101923.

[5] Travis, W.D., et al., International association for the study of lung cancer/american thoracic society/european respiratory society

international multidisciplinary classification of lung adenocarcinoma. Journal of thoracic oncology, 2011. 6(2): p. 244-285.

[6] Abou Taleb, A.S.T. and A.F. Atiya, A new approach for leukemia identification based on cepstral analysis and wavelet transform. Int J Adv Comput Sci Appl, 2017. 8(7): p. 226-232.

[7] Yu, K., et al., Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun. 2016; 7: 12474. Epub 2016/08/17., https://doi.org/10.1038/ncomms12474 PMID: 27527408.

[8] Schmidhuber, J., Deep learning in neural networks: An overview. Neural networks, 2015. 61: p. 85-117.

[9] Masud, M., et al., A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework. Sensors, 2021. 21(3): p. 748.

[10] Mangal, S., A. Chaurasia, and A. Khajanchi, Convolution Neural Networks for diagnosing colon and lung cancer histopathological images. arXiv preprint arXiv:2009.03878, 2020.

[11] Hatuwal, B.K. and H.C. Thapa, Lung cancer detection using convolutional neural network on histopathological images. Int. J. Comput. Trends Technol, 2020. 68: p. 21-24.

[12] Sarwinda, D., et al. Analysis of Deep Feature Extraction for Colorectal Cancer Detection. in 2020 4th International Conference on Informatics and Computational Sciences (ICICoS). 2020. IEEE.

[13] Kumar, N., et al., An empirical study of handcrafted and dense feature extraction techniques for lung and colon cancer classification from histopathological images. Biomedical Signal Processing and Control, 2022. 75: p. 103596.

[14] Wang, Y., et al., OCTID: a one-class learning-based Python package for tumor image detection. Bioinformatics, 2021. 37(21): p. 3986-3988.

[15] Chehade, A.H., et al., Lung and Colon Cancer Classification Using Medical Imaging: A Feature Engineering Approach. 2022.

[16] Hlavcheva, D., et al. Comparison of CNNs for Lung Biopsy Images Classification. in 2021 IEEE 3rd Ukraine Conference on Electrical and Computer Engineering (UKRCON). 2021. IEEE.

[17] Images, L.a.C.C.H., Kaggle. Available online: https://www.kaggle.com/andrewmvd/lung-and-colon-cancer-histopathological-images, 16 July 2020.

[18] Borkowski, A.A., et al., Lung and colon cancer histopathological image dataset (lc25000). arXiv preprint arXiv:1912.12142, 2019.

[19] Kermany, D.S., et al., Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell, 2018. 172(5): p. 1122-1131. e9.

[20] Zuiderveld, K., Contrast limited adaptive histogram equalization. Graphics gems, 1994: p. 474-485.

[21] Tan, M. and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. in International conference on machine learning. 2019. PMLR.

[22] Howard, A., et al., Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. 2018.