IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 20, 2022, accepted February 8, 2022, date of publication February 10, 2022, date of current version March 10, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3150924

Malignancy Detection in Lung and Colon Histopathology Images Using Transfer Learning With Class Selective Image Processing

SHAHID MEHMOOD[1], TAHER M. GHAZAL[2,3], (Member, IEEE), MUHAMMAD ADNAN KHAN[1,4], MUHAMMAD ZUBAIR[5], (Senior Member, IEEE), MUHAMMAD TAHIR NASEEM[1,6], TAUQEER FAIZ[3], AND MUNIR AHMAD[7], (Member, IEEE)

[1]Faculty of Computing, Riphah School of Computing and Innovation, Riphah International University, Lahore Campus, Lahore 54000, Pakistan
[2]Center for Cyber Security, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor 43600, Malaysia
[3]School of Information Technology, Skyline University College, University City of Sharjah, Al Sharjah, United Arab Emirates
[4]Pattern Recognition and Machine Learning Laboratory, Department of Software, Gachon University, Seongnam-si 13557, South Korea
[5]Faculty of Computing, Riphah International University, Islamabad 45000, Pakistan
[6]Human Ecology Research Center, Yeungnam University, Gyeongsan-si 712-749, Republic of Korea
[7]School of Computer Science, National College of Business Administration and Economics, Lahore 54000, Pakistan

Corresponding authors: Munir Ahmad (munir@ncbae.edu.pk) and Muhammad Adnan Khan (adnan@gachon.ac.kr)

ABSTRACT Cancer accounts for a huge mortality rate due to its aggressiveness, colossal potential of metastasis, and heterogeneity (causing resistance against chemotherapy). Lung and colon cancers are among the most prevalent types of cancer around the globe that can occur in both males and females. Early and accurate diagnosis of these cancers can substantially improve the quality of treatment as well as the survival rate of cancer patients. We propose a highly accurate and computationally efficient model for the swift and accurate diagnosis of lung and colon cancers as an alternative to current cancer detection methods. In this study, a large dataset of lung and colon histopathology images was employed for training and the validation process. The dataset is comprised of 25000 histopathology images of lung and colon tissues equally divided into 5 classes. A pretrained neural network (AlexNet) was tuned by modifying the four of its layers before training it on the dataset. Initial classification results were promising for all classes of images except for one class with an overall accuracy of 89%. To improve the overall accuracy and keep the model computationally efficient, instead of implementing image enhancement techniques on the entire dataset, the quality of images of the underperforming class was improved by applying a contrast enhancement technique which is fairly simple and efficient. The implementation of the proposed methodology has not only improved the overall accuracy from 89% to 98.4% but has also proved computationally efficient.

INDEX TERMS Colon cancer, convolutional neural networks, histopathology, image processing, lung cancer, transfer learning.

I. INTRODUCTION

Worldwide, cancer ranks second in terms of causes of death. Well over 19 million new cases of cancer and 9.95 million deaths were reported worldwide in 2020 [1]. The human body is comprised of trillions of cells, these cells grow and multiply to form new cells as per body requirements through a process called cell division. It is normal for cells to die and be replaced by new ones when they reach a certain age or degree of damage. If this process breaks down, damaged cells start to grow and multiply resulting in the formation of tumors. These tumors could be malignant (cancerous) or benign [2]. Cancer could affect any human body organ; however, the foremost common affected areas are the colon, lungs, liver, breasts, rectum, brain, prostate, stomach, and skin. The most common cancers to cause death equally in males and females are lung and colon cancers. In 2020, 4.14 million new instances of lung and colorectal cancer with 2.7 million deaths were recorded globally [3].

Behavioral characteristics such as a high BMI, use of cigarettes and alcohol may contribute to the development of cancer. Radiation and ultraviolet rays are physical

The associate editor coordinating the review of this manuscript and approving it for publication was Abdullah Iliyasu.

carcinogens and some biological and hereditary ones [3]. Cancer symptoms include discomfort, exhaustion, nausea, chronic cough, breathlessness, loss of weight, muscle aches, bleeding, and bruises [4]. However, none of these symptoms is specific to cancer, nor are they present in every patient. Consequently, detecting cancer is difficult without a comprehensive diagnostic test for instance a biopsy, PET scan, CT scan, MRI scan, and ultrasound. Across many cases, patients develop no or few signs of the disease, although, by the time symptoms emerge, it is often too late.

Some people inherit the gene responsible for cancer from their parents. People with an inherited risk of developing cancer should have regular checkups. These screening systems are highly expensive, and yet many people are unable to afford them. About 70 percent of cancer-related deaths are reported from low- and middle-income countries suggest the World Health Organization. [3]. These countries must build a large number of laboratories and diagnostic facilities with the required equipment, as well as train medical workers to execute diagnostic processes. Additionally, people who live below the poverty line must be able to afford the charges of these tests.

This problem may be solved by a field completely different from medicine and healthcare. Pathology utilizes Machine Learning (ML) in a wide variety of ways, such as the detection of diseases and the development of intelligent systems that, based on a patient's symptoms, prescribe conventional medicines [5]. Machine learning approaches have been used to classify and predict several kinds of biomedical data. Deep learning (DL) techniques enable machines to analyze high-dimensional data such as pictures, multidimensional anatomical representations, and video. As a subset of ML, DL deals with algorithms based on the anatomy and function of the brain [6].

In clinical practice, accurately categorizing histopathology pictures is critical for obtaining a trustworthy diagnosis of illnesses. This type of activity could be automated using ML algorithms, specifically DL, to replace the exhausting and expensive physical effort of human specialists while also meeting the demands for high accuracy, large data sets, and other factors. Transfer learning (TL) is used in visual categorization to solve cross-domain learning issues by transferring useful knowledge from the given dataset to the task domain [5]. Through TL, knowledge from different yet related sources is transferred to target learners for improved performance on target domains. This can decrease the need for an enormous amount of data to construct target learners. TL is becoming an increasingly prevalent and auspicious area within ML because of its broad range of application prospects [7]. Different types of TL approaches include feature representation transfer, which includes cross-domain and cross-view knowledge exchange, and classifier-based information exchange, which includes Support Vector Machine (SVM)-based, TrAdaboost, and generative models [8].

TL is often performed with DL models trained for the ImageNet competition, an image classification competition with a large and challenging dataset. Pre-trained AlexNet architecture is employed in this study. To utilize the elaborately learned features by AlexNet in a new image classification task, a few layers of the model are modified while the weights in the transferred layers are kept frozen.

The rest of this article is laid out as follows: A brief outline of previous research similar to ours is given in section 2. In section 3 employed dataset and its contents are described. Section 4 is dedicated to the elaboration of the proposed research model while the experimental environment and discussion on its results are presented in section 5. Finally, the article is concluded with a summary of the outcomes.

II. RELATED WORK

Sirinukunwattana *et al.* [10] proposed a spatially constrained neural network for detecting the nucleus in histopathology images of colon cancer. For the classification of cell nuclei, a novel adjacent group predictor was used. The maximum accuracy attained was 97.1%. Although their model obtained good results, its computational efficiency did not measure up since it took an average of 50 minutes to run a single slide. Shein *et al.* [11] developed an ML architecture for the classification of lung nodule malignancy. A pooling method has been used to crop different areas of interest from feature maps and max-pooling was used multiple times to extract features from CT scans of lung nodules. This approach is distinct in that it did not employ any segmentation or feature extraction methods on the CT scan images used. Based solely on their machine learning model for lung nodule identification, they obtained an accuracy of 87.14 percent. Three deep structured algorithms have been used to automatically extract features from CT images of a lung nodule in [11], including stacked denoising autoencoder (SDAE), convolutional neural network (CNN), and deep belief network (DBN). The highest accuracy achieved was 89% by applying CNN. Selvanambi *et al.* [12] performed lung cancer detection by employing recurrent neural network (RNN) with damped least-squares (DLS) method and glowworm swarm optimization algorithm and achieved 98% accuracy.

Filho *et al.* [13] utilized image segmentation for the preprocessing of Lung nodule CT images. Authors applied index basic taxic weights along with standardized taxic weights for recognition of malignant and benign patterns and CNN for classification and attained an accuracy of 92.6%. Yuan *et al.* [14] presented a CNN-based model for the automatic detection of polyps in colonoscopy videos using AlexNet architecture. Before the application of CNN for classification several preprocessing operations namely, edge detection, morphology, and intensity adjustments were performed. Their model acquired an accuracy of 91.4%. The FP per frame result suggests that this method may still be improved. It may fail to locate a polyp if its border is the same color as the backdrop. In an informative frame, a large

number of non-polyp candidates may also be created. This will increase execution time while also increasing the chance of decreased accuracy. For the classification of benign and malignant lung nodules in CT images in [15], RestNet50 and SVM with Radial Basis Function (RBF) kernel were used. The employed methodology obtained an accuracy of 92.8%. Masood *et al.* [16] proposed the DFCNet model, which is based on a deep CNN, for classifying pulmonary nodules in CT images into four stages of lung cancer and reached 84.5% accuracy. In the case of malignant tumors, the study is flawed in that it employs several datasets with different scan settings, which results in false-positive results. Using the same scan parameters on every dataset will maximize classification results. For the detection of polyps in colonoscopy videos, Mo *et al.* [17] used a faster region-based convolutional neural network (Faster R-CNN) on four distinct datasets and acquired 98.5% average accuracy. Urban *et al.* [18] developed and implemented deep CNNs for the detection of polyps in expert-labeled pictures extracted from over 2000 colonoscopies. Their model accomplished an accuracy of 96.4%. To reduce the network size, binarized weights were applied with convolutional neural networks by the model proposed in [19] for the classification of colonoscopy frames with an accuracy of 90.28%. In [20] to reduce dimensionality and complexity, Wolf heuristic features using minimum repetition were selected. A joint learning comprehensive neural network optimized for AdaBoost was used to identify the normal and anomalous lung structures with high accuracy of 98.42%.

Suresh and Mohan [21] proposed an eight-layer CNN architecture for categorizing CT images of lung lesions into one of three groups. Segmentation was applied on images to extract nodule regions of interest in collaboration with experts. In addition, generative adversarial networks were used to augment the dataset. The classification accuracy of their proposed model was 93.9%. For the detection of pulmonary nodules, Masud *et al.* [22] proposed a light deep learning approach relying on CNN architecture with only four convolutional layers. Each convolutional layer is comprised of two successive convolutional blocks, a connecting convolutional block, non-linear activation functions after each block, and a pooling block. The authors found their proposed model suitable for real-time CT image analysis because it has fewer flops and parameters than state-of-the-art CNN architectures. The proposed model acquired 97.9% accuracy. In [23] lung cancer CT scans were preprocessed by utilizing an approach that preserves the brightness of images at multiple levels and eliminates noise. An improved neural network was employed for the segmentation of the affected regions and extraction of features. An ensemble classifier was used to perform feature classification. Classification accuracy achieved by the model was 96.2%. Bukhari *et al.* [24] applied three pre-trained CNNs (ResNet-50, ResNet-34, and ResNet-18) to evaluate the histopathology images of colonic adenocarcinoma from two different databases and acquired 96.4% accuracy. Mangal *et al.* [25] analyzed digital pathology images of adenocarcinoma and squamous cell carcinoma of the colon and lung and trained a shallow neural network for its classification, and achieved an accuracy of 97.8%. B.K. Hatuwal and H.C. Thapa employed a convolutional neural network to classify Lung cancer histopathology images and achieved an accuracy of 96.11% and 97.2% for training and validation respectively [26].

A. LIMITATIONS OF THE RELATED WORK

The previous studies have the following limitations:

- Inefficient algorithms, resulting in higher computational cost and time utilization [10], [11], [14].
- Multiple datasets with varying scan settings, resulting in false-positive results [16].
- Worked on a smaller number of classes or a less diverse dataset [19], [20], [24]–[26].
- Reported low accuracy [11], [13], [14], [19], [21]

The major contributions of this paper are outlined below:

- Most earlier cancer detection studies focused on a single form of cancer, however in this study, we used our model to identify lung and colon cancer at the same time.
- Although few studies implemented image processing techniques to improve the quality of images before classification, however, in these studies, image processing was applied to the whole dataset resulting in higher time utilization and computing cost. We proposed Class Selective Image Processing (CSIP) which is a novel strategy. Instead of preprocessing all images in a data set, images of a selected class are improved. This strategy not only improves the classification accuracy of the model but also saves time and computing costs.
- Improved learning of the pre-trained deep learning model by employing CSIP on histopathological slide images which enhanced their contrast using histogram equalization.
- Presented state-of-the-art results for automated lung and colon cancer detection using histopathological slide images.
- The experiment was conducted on a comprehensive and balanced dataset comprised of 25000 histopathology images divided equally into five classes of lungs and colon. The multiple classes and a large number of images included in the proposed method made the model more accurate and reliable.

III. EMPLOYED CANCER DATASET

In this study, we used A. Borkowski and his colleagues' new dataset LC25000, published in 2020 [27]. There are 25000 images of lung and colon tissues in this collection, which are classified into five categories. Three types of lung tissue images are adenocarcinoma, squamous cell carcinoma, and benign. While colon images are of two categories adenocarcinoma and benign tissues.

The LC25000 dataset was primarily comprised of 1250 images of pathology slides of lung and colon tissues. The dataset was augmented by flipping and rotating the images in a variety of conditions as a result, the dataset was expanded to 25,000 images divided into five categories, with 5000 images per category. Images were resized to 768 × 768

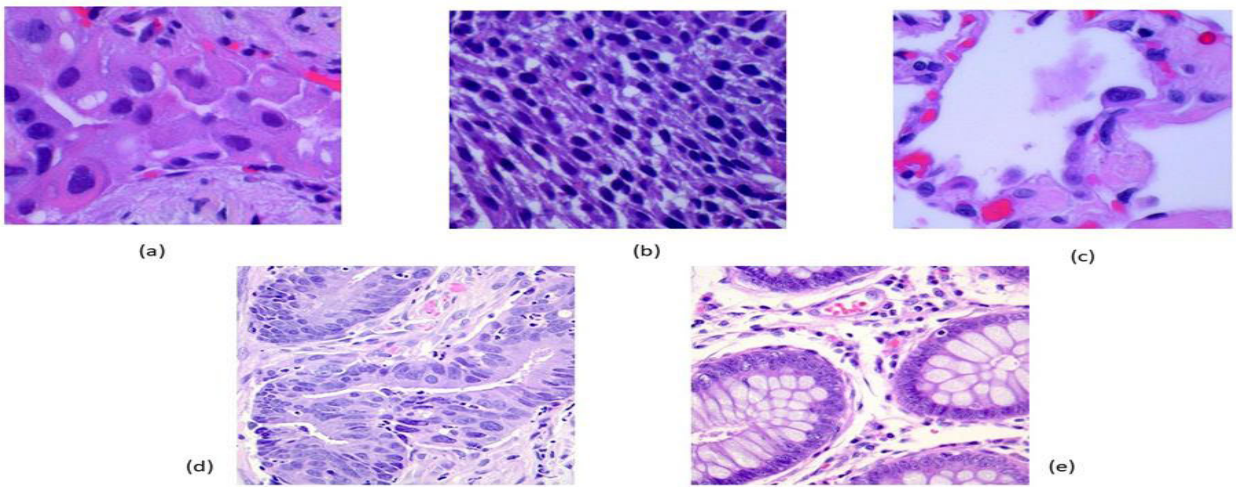

FIGURE 1. Histopathological images from dataset. (a) Lung Adenocarcinoma. (b) Lung benig. (c) Lung squamous cell. (d) Colon adenocarcinoma. (e) Colon benign.

TABLE 1. Description of the employed dataset.

Image Type	Class ID	Class Title	Total Images
Colon Adenocarcinoma	0	Colon_aca	5000
Colon Benign	1	Colon_n	5000
Lung Adenocarcinoma	2	Lung_aca	5000
Lung Benign	3	Lung_n	5000
Lung Squamous Cell Carcinoma	4	Lung_scc	5000

TABLE 2. Training options and parameters.

Training Options	Parameters
Image Size	227x227x3
No of Epochs	60
Initial Learning Rate	0.00001
Momentum	0.9
Solver	SGDM

before applying the augmentation. To ensure privacy and free use, these images are validated and are compliant with Health Insurance Portability and Accountability Act (HIPAA). Sample images from the dataset are presented in figure1. Table1 exhibits the assigned names and IDs to each class of images in the dataset.

IV. MATERIALS AND METHODS

TL is a machine learning technique that entails adapting a model created for one task to another. DL encompasses using networks that have been pre-trained on a large dataset to build a new network architecture that can then be used on a new dataset after fine-tuning. Given the time and resources required to train deep CNNs, the TL method massively reduces the training effort. It is quite often faster and easier than building and training a network from the ground up. TL is typically performed with DL models trained for a large image classification task, ImageNet competition [7].

A pretrained model AlexNet is used in this study for the classification of cancer images. AlexNet developed in 2012 by Alex Khrizveski is an award-winning convolutional neural network well known for its high accuracy on challenging datasets. This network is comprised of 65,000 neurons and has more than 60 million parameters. ImageNet database is a huge database of images with over 14 million images of over 20,000 classes. AlexNet has been trained on the ImageNet database using over a million images of more than one thousand classes. The architecture of the network is comprised of five convolutional layers along with three fully connected layers [28].

The first layer and the last three layers of the model are modified to tune the model according to the requirement of this research. To match the size limitations of the model, images were resized to $227 \times 227 \times 3$. The data set was divided in such a way that eighty percent of the whole images were reserved for the training and twenty percent for the validation. Training options and other parameters employed for training are shown in Table 2.

The proposed methodology is exhibited in figure 2. As a first step, the histopathology images are acquired and converted to $227 \times 227 \times 3$ size as per the model's specification. Providing the model with resized images for training and validation is the next step. The preliminary results are evaluated in terms of accuracy, precision, F1-score, recall, specificity, and misclassification rate. In case of unsatisfactory results, CSIP strategy is applied. The suggested CSIP technique identifies the underperforming class and uses Histogram Equalization (HE) to improve the picture quality of the identified class. The model is executed again with improved images. The proposed approach focuses on improving the images of the underperforming classes only instead of processing the entire dataset, as this reduces processing time and effort.

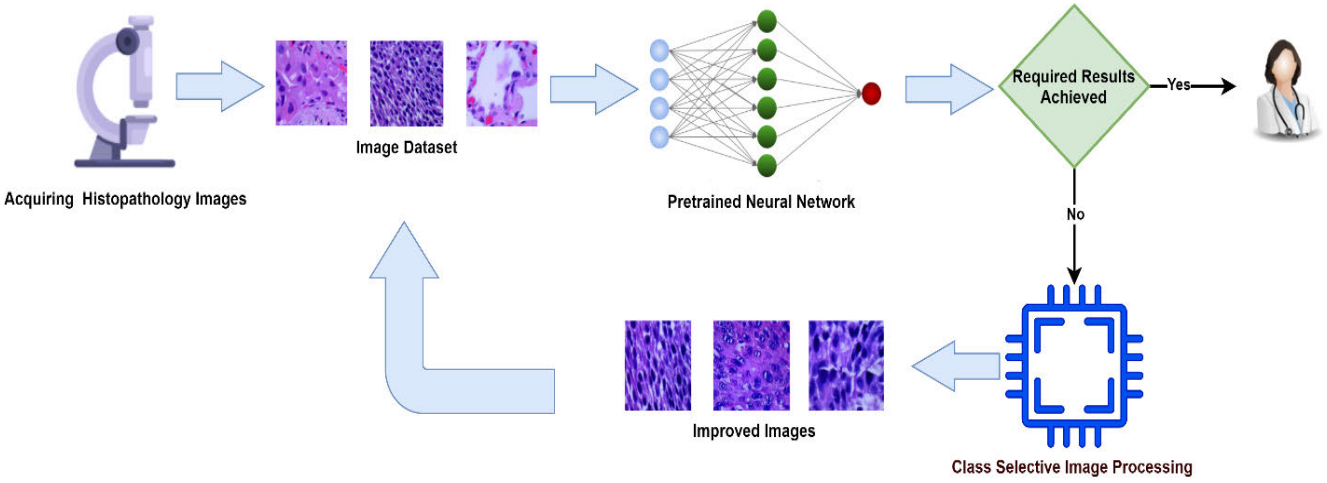

FIGURE 2. Proposed model for transfer learning and class selective image processing.

Image contrast enhancement is one of the most effective techniques for image quality improvement. The histogram equalization process is applied to the selected images as it is an efficient and computationally fast technique for contrast enhancement [29]. Histogram equalization redistributes the intensity values to accomplish contrast enhancement [30]. All images of the underperforming class were replaced with histogram equalized images and then the dataset was reused for training and validation.

V. RESULTS AND DISCUSSIONS

When compared to existing studies on cancer detection in the lungs and colon, our proposed methodology produced promising results. There are 25000 images in our dataset equally divided into 5 classes. We alienated the dataset in such a way that 80% of images in each class were used for the training of the AlexNet model and 20% of images for testing. In the performance analysis, we considered precision, sensitivity, specificity, f-1 score, recall, and misclassification rate in addition to accuracy. Listed below are the performance metrics utilized in this study.

$$Specificity = \frac{TN}{TN + FP} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$Accuracy = \frac{TP + TN}{TN + TP + FN + FP} \tag{5}$$

$$Misclassification\ Rate = 1 - (\frac{TP + TN}{TN + TP + FN + FP}) \tag{6}$$

The training progress and confusion matrix in Figure 3 and Figure 4, depicts the model's performance before the application of CSIP. The confusion matrix presents the overall accuracy, it also depicts the precision, and recall values of each class. Overall accuracy achieved by the model is 89.9%, precision of colon_aca, colon_n, lung_aca, lung_n, lung_scc, are 98.8%, 99.7%, 67.4%, 99.9%, and 96% respectively. Recall values of colon_aca, colon_n, lung_aca, lung_n, and lung_scc are 99.1%, 99.9%,97.1%,99.8%, and 53.3% respectively. Table 3 and figure 5 present the model's accuracy, precision, F-1 score, recall, specificity, and misclassification rate achieved for each class of the dataset. The highest accuracy is obtained by lung_n class which is 99.94% along with the lowest misclassification rate of 0.0006. The colon_n class attained an accuracy of 99.92% with 0.0008 misclassification rate. The accuracy achieved by colon_aca is 99.58% and the misclassification rate is 0.0042. When compared to the aforementioned classes, the lung_scc and lung_aca classes performed poorly, with an accuracy of 90.22 percent and 89.84 percent, respectively, and misclassification rates of 0.978 and 0.1016. In the case of F1-score lung_n, colon_n, and colon_aca classes obtained 99.85%,99.80%, and 98.80% respectively while lung_aca and lung_scc scored 79.56% and 68.55%.

To improve the overall classification accuracy of the model we implemented the contrast enhancement technique (HE) on the images of underperforming classes only so that the time and computational cost could be saved. HE technique was applied on all 5000 images of the lung_scc class. Essentially, the HE technique straightens the histogram by uniformly spreading all grey levels. As a result, the enhanced image's mean brightness changes dramatically. The following steps make up the HE's fundamental procedure. Suppose the input image $S(i, j)$ contains a sum of Y pixels in the dynamic range $\in [Z_0, Z_{G-1}]$, where G denotes the number of distinct grey levels. Following expression calculates the normalized histogram (NH) or the probability density function (PDF) $F(Z_k)$ of the image's intensity level Z_k:

$$F\ (Z_k) = \frac{n_k}{Y} \quad for\ 0 \leq k \leq G - 1 \tag{7}$$

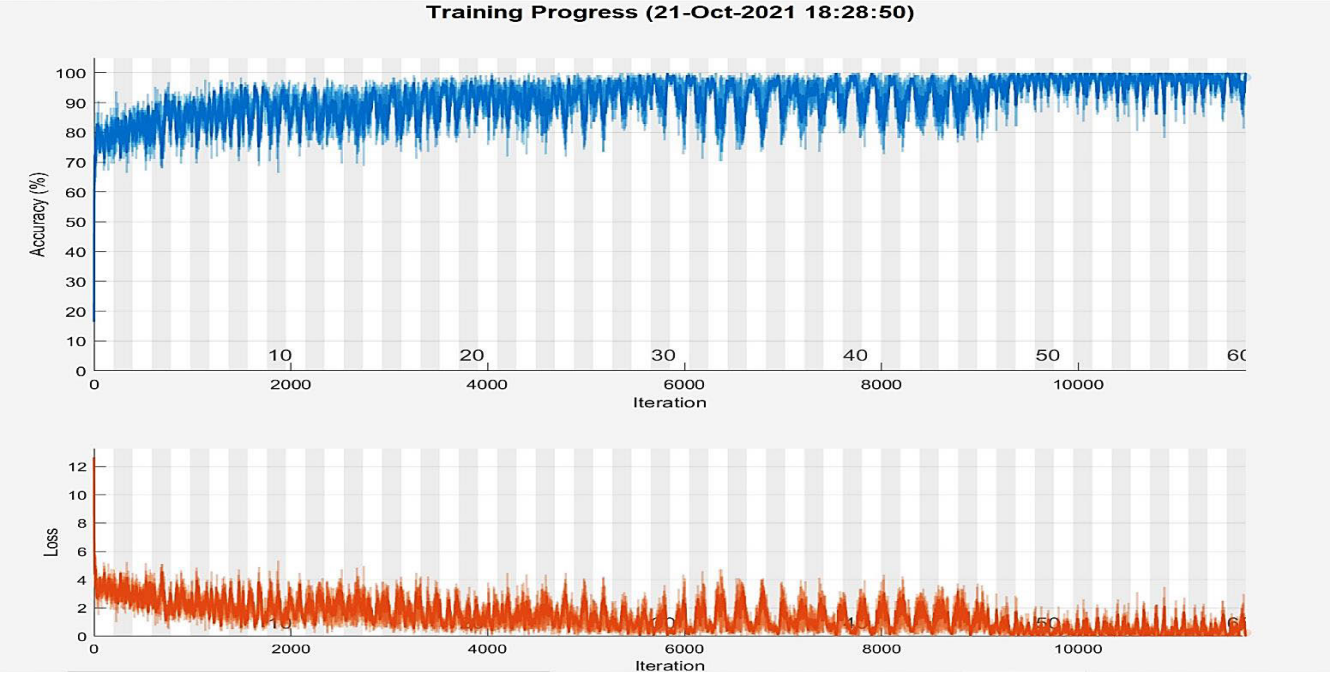

FIGURE 3. Training progress of AlexNet before application of CSIP.

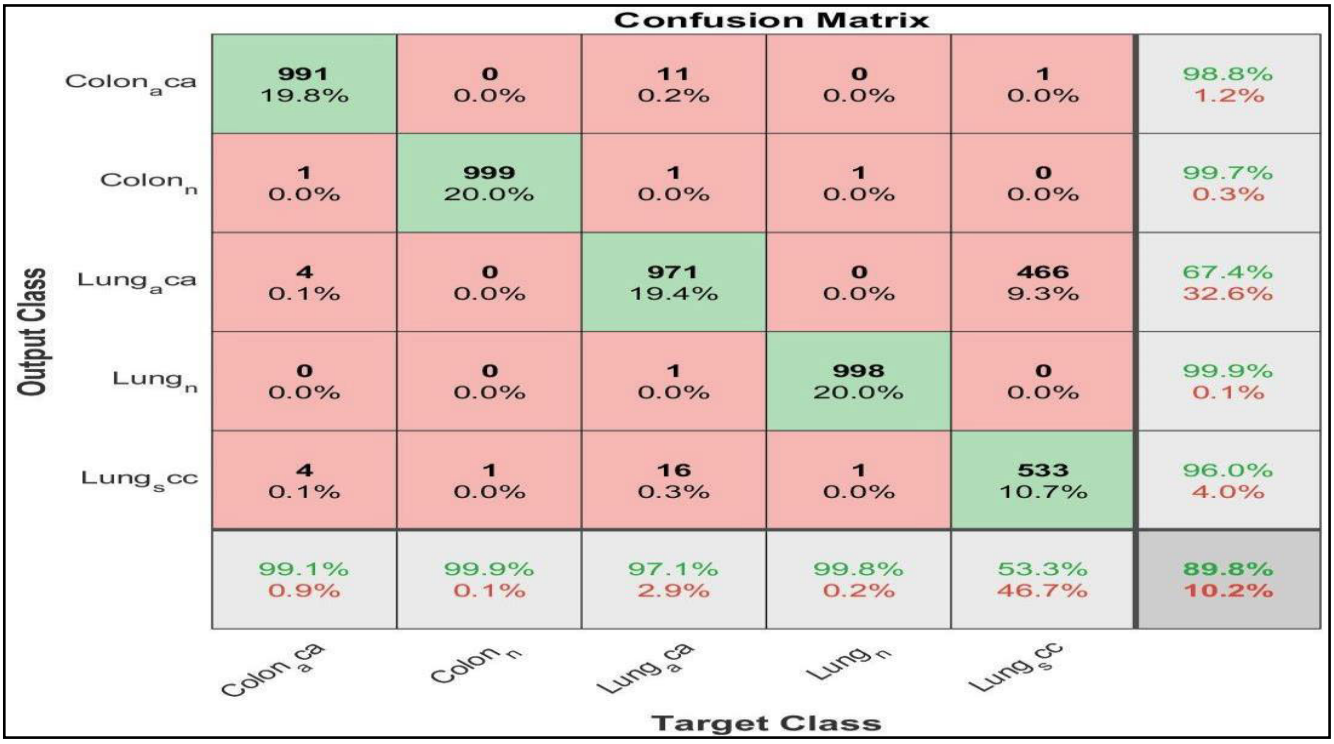

FIGURE 4. Confusion matrix of AlexNet before application of CSIP.

where n_k denotes the sum of pixels and Z_k represents the intensity level. The histogram of image S is the plot of Z_k vs. n_k while the cumulative density function (CDF) is given by [31]:

$$CDF(Z_k) = \sum_{i=0}^{k} F(Z_i) \tag{8}$$

The relevant image is projected into the whole dynamic range $\in [Z_0, Z_{G-1}]$, by HE using the CDF, which is denoted by the following relation [32], [33]:

$$C(Z) = Z_0 + (Z_{G-1}, Z_0)(\sum_{i=0}^{k} F\,(Z_i)\,(Z)) \tag{9}$$

TABLE 3. Performance of the model before CSIP.

Type of Tissue	Accuracy	Precision	F-1 Score	Recall	Specificity	Misclassification Rate	Overall Accuracy
Colon_aca	99.58%	98.80%	98.95%	99.10%	99.70%	0.0042	89.9%
Colon_n	99.92%	99.70%	99.80%	99.90%	99.93%	0.0008	
Lung_aca	89.84%	67.38%	79.56%	96.14%	88.25%	0.1016	
Lung_n	99.94%	99.90%	99.85%	99.80%	99.98%	0.0006	
Lung_scc	90.22%	96.04%	68.55%	53.30%	99.45%	0.0978	

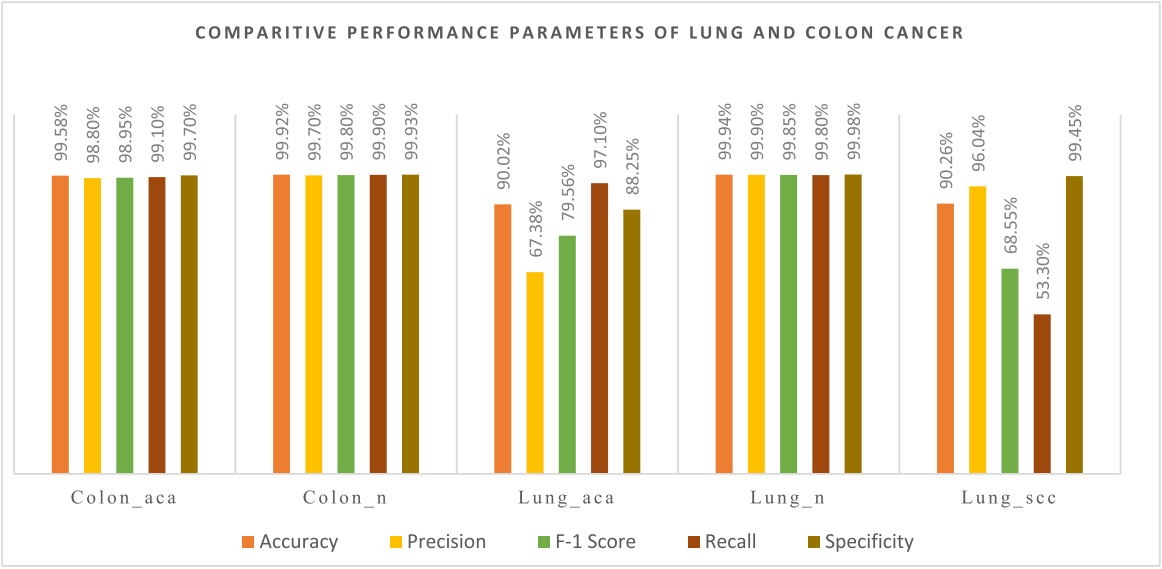

FIGURE 5. Performance of the model on each class.

If the image has been processed using the classic, HE method, the statistical expectation E(.) of the resulting picture S is determined by:

$$E(S) = S_m = \frac{1}{2}(Z_{G-1}, Z_0) \tag{10}$$

where S_m denotes the average intensity of the resulting picture. Consequently, the image's intensity is substantially shifted to the grey level's mean. To illustrate the results of the applied image processing technique Figure 6 shows the original image and histogram equalized image along with their respective histograms.

It is evident in the confusion matrix that AlexNet performed very well on all classes of cancer images except the lung_scc class and attained an overall accuracy of 89.8%. Colon_aca class acquired 99.58%, 98.80%, and 98.95% accuracy, precision, and F-1 score respectively. The accuracy, precision, and F-1 scores achieved by Colon_n class are 99.92%, 99.70%, and 99.80% respectively. Lung_aca class obtained 90.02% accuracy, 67.38% precision, and 79.56% F-1 score. Accuracy, precision and F-1 score in the case of Lung_n reached 99.94%, 99.90%, and 99.85% respectively. In the case of Lung_scc class accuracy is 90.22%, precision is 96.04% while the F-1 score is 68.55% only. Therefore, images in lung_scc need to be processed for quality improvement. The HE technique has been applied to all 5000 images in the lung_scc class. Figure 6 presents sample images before and after the application of HE. The difference of contrast is visible in the sample images. After optimizing the picture quality of the selected class, the dataset is updated and the model is run for additional 60 epochs. The adoption of CSIP resulted in a considerable increase in terms of accuracy and other parameters in the experimental findings. Figure 7 depicts the training progress and Figure 8 presents the confusion matrix after CSIP. Accuracy of Lung_scc class reached 100% after implementation of the proposed CSIP.

Figure 9 and Table 4 show the comparison of accuracy and other parameters among all five classes of images in the dataset. To determine how well the proposed model performs on previously undiscovered data, it has been rigorously tested on test data. As part of ensuring an unbiased evaluation, the test data was also pre-processed and structured the same way. The model has excellent performance across all categories but stands out in identifying lung squamous cell cancer with 100% prediction accuracy, precision, recall, f1-score, and specificity. Similarly, the model has a 99.94%, 99.64%, 99.12%, and 98.98% accuracy for lung benign, lung adenocarcinomas, colon benign, and colon adenocarcinoma

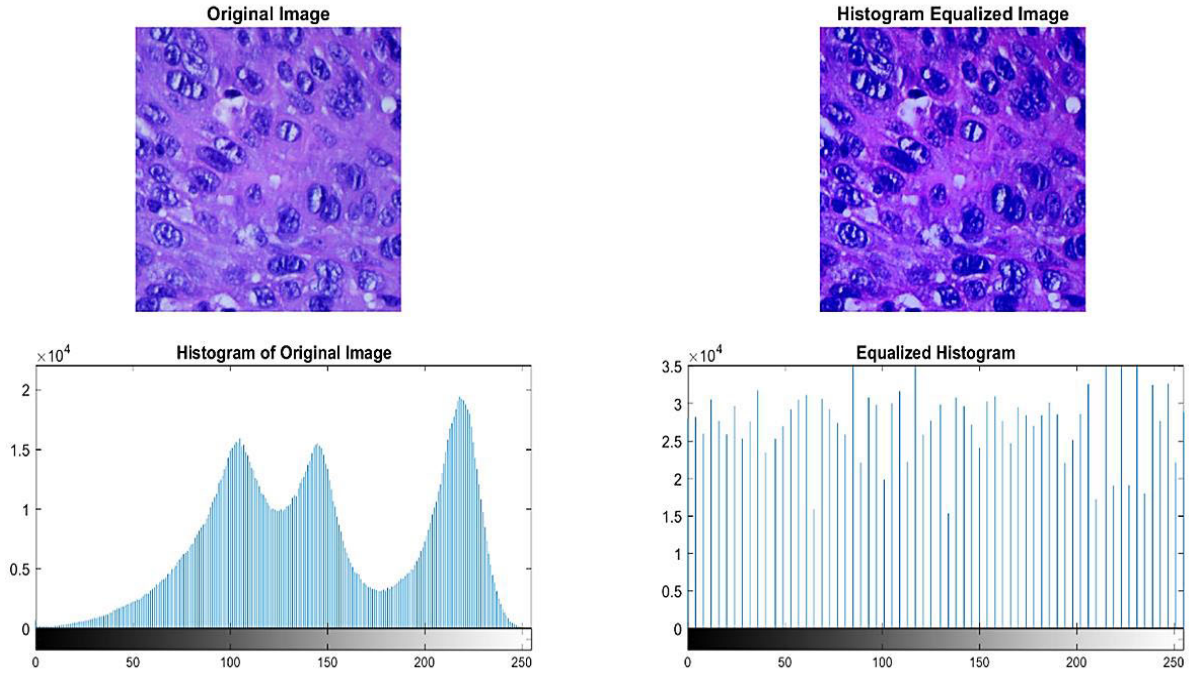

FIGURE 6. A sample image and its histogram before and after histogram equalization.

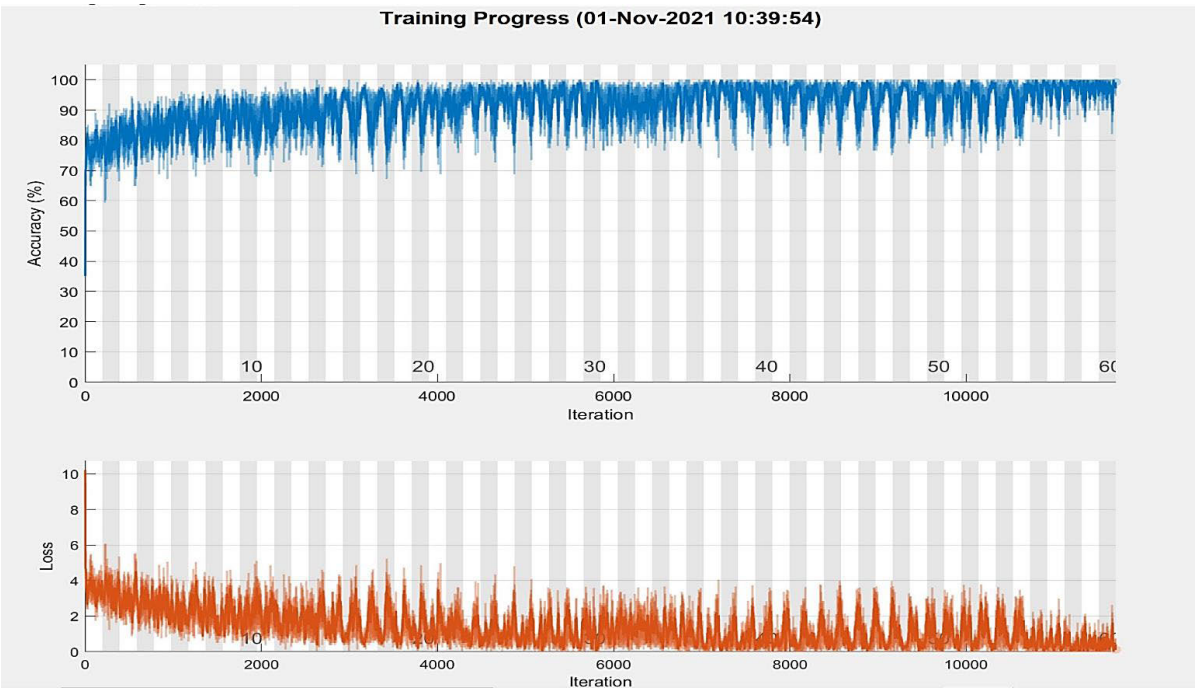

FIGURE 7. Training progress of the model after class selective image processing.

respectively. All of the classes demonstrated high sensitivity and specificity as well.

The key rationale for the enhanced performance is the quality of the features obtained as a consequence of the used CSIP approach, which enabled the model to function better. Table 5 compares the model's performance in terms of classification accuracy and misclassification before and after the application of CSIP. We compare the obtained results to the outcomes of cutting-edge techniques to substantiate the effectiveness of the suggested model. However, because we employed a unique dataset that is substantially distinct from the ones used in the mentioned publications, these results are largely incomparable. They have nevertheless been put in comparison because of the same objective.

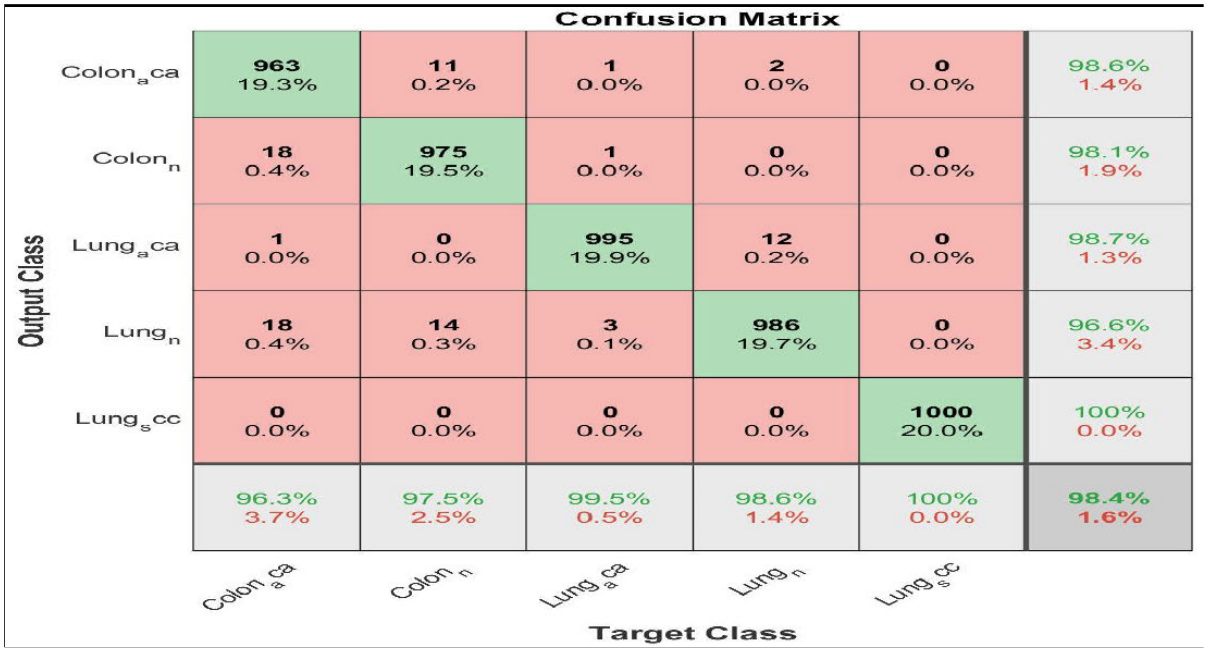

FIGURE 8. Confusion matrix of the model after class selective image processing.

TABLE 4. Post image processing results.

Type of Tissue	Accuracy	Precision	F-1 Score	Recall	Specificity	Misclassification Rate	Overall Accuracy
Colon_aca	98.98%	98.57%	97.42%	96.30%	99.65%	0.0102	**98.4%**
Colon_n	99.12%	98.09%	97.79%	97.50%	99.53%	0.0088	
Lung_aca	99.64%	98.71%	99.10%	99.50%	99.68%	0.0036	
Lung_n	99.02%	96.57%	97.58%	98.60%	99.13%	0.0098	
Lung_scc	100.00%	100.00%	100.00%	100.00%	100.00%	0	

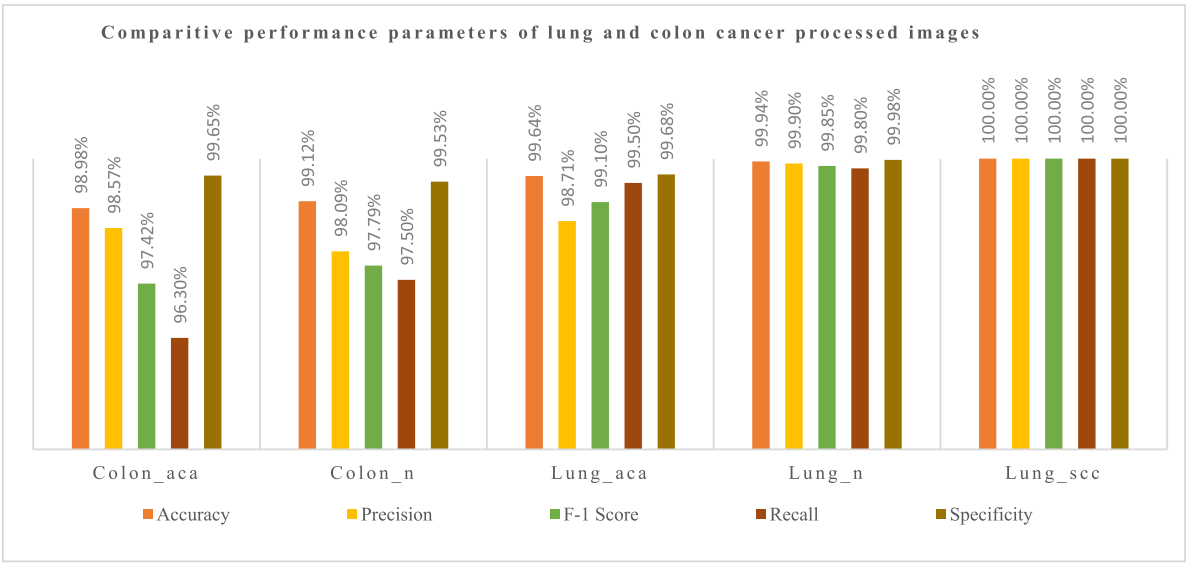

FIGURE 9. Performance of the model on each class after CSIP application.

The proposed model outperforms state-of-the-art methods in terms of classification accuracy when lung and colon anomalies are classified concurrently, as shown in Table 6.

Our approach outperforms most known methods for cancer detection, as shown in Table 6; the only exception is the research described in [20]. The dataset utilized in [20]

TABLE 5. Overall performance of proposed CSIP-TL model.

	Training		Validation	
	Accuracy	Misclassification Rate	Accuracy	Misclassification Rate
Before CSIP	93.8%	0.985875	89.9%	0.040554313
Post CSIP	99.85%	0.003640	98.4%	0.00464

TABLE 6. Comparison of lung and colon cancer classification results of other methods and datasets.

Article Title	Year	Type of Cancer	Type of Images	Model/Classifier	Accuracy (%)
D.C. Filho et.al, [13]	2018	Lung Cancer	CT Images	Convolutional Neural Network	92.63
D. Nobrega et.al, [15]	2020	Lung Cancer	CT Images	RestNet50 + SVM RBF	93.19
A. Masood et.al, [16]	2018	Lung Cancer	CT Images	DFCNet	89.5
G. Urban et.al, [18]	2018	Colon Cancer	Colonoscopy Frames	Convolutional Neural Network	96.4
M. Akbari et.al, [19]	2018	Colon Cancer	Colonoscopy Frames	Convolutional Neural Network	90.28
P. M. Shakeel et.al, [20]	2019	Lung Cancer	CT Images	DITNN	98.42
S. Suresh et.al, [21]	2020	Lung Cancer	CT Images	Convolutional Neural Network	93.9
P. M. Shakeel et. al, [23]	2020	Lung Cancer	CT Images	EM	96.2
S.U.K. Bukhari et.al, [24]	2020	Colon Cancer	Histopathology Images	RESNET-50	93.91
S. Mangal et.al, [25]	2020	Lung Cancer	Histopathology Images	Convolutional Neural Network	97.8 9
S. Mangal et.al, [25]	2020	Colon Cancer	Histopathology Images	Convolutional Neural Network	96.61
B.K. Hatuwal et.al, [26]	2020	Lung Cancer	Histopathology Images	Convolutional Neural Network	97.2
W. Shen et.al, [34]	2017	Lung Cancer	CT Images	MC-CNN	87.14
Z. Yuan et.al, [35]	2017	Colon Cancer	Colonoscopy Frames	AlexNet	91.47
T. Babu et.al, [36]	2018	Colon Cancer	Histopathology Images	RF	85.3
M. Masud et, al, [38]	2020	Lung Cancer	CT scan	Convolutional Neural Network	97.9
M. Masud et.al, [39]	2021	Lung & Colon Cancer	Histopathology Images	Convolutional Neural Network	96.33
Proposed	**2021**	**Lung & Colon Cancer**	**Histopathology Images**	**CNN with CSIP**	**98.4**

only included CT scan pictures of the lungs, whereas our dataset includes both lung and colon histopathological images, making a direct comparison impossible. Studies cited in [24]–[26], [39] worked on the LC25000 dataset. A study in reference [24] only looked at the colon samples and demonstrated lower accuracy and a lower F-measure than the proposed method. Even though relatively better accuracy was reported by [25], [26], they classified either lung images or colon images. The obtained findings are not precisely comparable since we considered all classes of the dataset images for classification. In the case of reference [39], multiclass classification is performed on the LC25000 dataset but reported accuracy, precision, and F-1 score are lower than the proposed model. As a result of the discussions in this section, it may be inferred that the suggested approach is successful in accurately identifying lung and colon cancer tissue.

VI. CONCLUSION

Lung and colon cancers are among the leading causes of fatality worldwide. Early and accurate diagnosis of these cancers can significantly improve therapeutic outcomes and survival rates. The goal of this study was to detect lung and colon cancer accurately and efficiently. Transfer learning is employed for this detection on a dataset of 25000 histopathology images of lung and colon tissues. Our deployed DL network's overall accuracy was initially 89. %, however, after applying the proposed CSIP approach,

accuracy was significantly enhanced and reached 98.8%. The proposed methodology has not only outperformed state-of-the-art methods for lung and colon cancer detection in terms of accuracy but has also reduced the time and computational cost. We believe that our proposed methodology can also be implicated in the effective diagnosis of other diseases.

REFERENCES

[1] World Fact Sheet. *International Agency for Research on Cancer*. Accessed: Oct. 17, 2021. [Online]. Available: https://gco.iarc.fr/today/data/factsheets/populations/900-world-fact-sheets.pdf

[2] (May 5, 2021). *What is Cancer*. National Cancer Institute. Accessed: Oct. 17, 2021. [Online]. Available: https://www.cancer.gov/about-cancer/understanding/what-is-cancer

[3] (Sep. 21, 2021). *Cancer*. World Health Organization. Accessed: Oct. 18, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[4] *Cancer Symptoms and Causes*. Mayo Clinic. Accessed: Oct. 18, 2021. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/cancer/symptoms-causes/syc-20370588

[5] S. Das, S. Biswas, A. Paul, and A. Dey, "AI Doctor: An intelligent approach for medical diagnosis," in *Industry Interactive Innovations in Science, Engineering and Technology*. Singapore: Springer, 2018, pp. 173–183.

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

[7] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021, doi: 10.1109/JPROC.2020.3004555.

[8] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 5, pp. 1019–1034, May 2015.

[9] J. Fan, J. Lee, and Y. Lee, "A transfer learning architecture based on a support vector machine for histopathology image classification," *Appl. Sci.*, vol. 11, pp. 1–16, 2021.

[10] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1196–1206, May 2016.

[11] W. Sun, B. Zheng, and W. Qian, "Automatic feature learning using multichannel ROI based on deep structured algorithms for computerized lung cancer diagnosis," *Comput. Biol. Med.*, vol. 89, pp. 530–539, Oct. 2017.

[12] R. Selvanambi, J. Natarajan, M. Karuppiah, S. K. H. Islam, M. M. Hassan, and G. Fortino, "Lung cancer prediction using higher-order recurrent neural network based on glowworm swarm optimization," *Neural Comput. Appl.*, vol. 32, no. 9, pp. 4373–4386, 2018.

[13] A. O. de Carvalho Filho, A. C. Silva, A. C. de Paiva, R. A. Nunes, and M. Gattass, "Classification of patterns of benignity and malignancy based on CT using topology-based phylogenetic diversity index and convolutional neural network," *Pattern Recognit.*, vol. 81, pp. 200–212, Sep. 2018.

[14] Z. Yuan, M. IzadyYazdanabadi, D. Mokkapati, R. Panvalkar, J. Y. Shin, N. Tajbakhsh, S. Gurudu, and J. Liang, "Automatic polyp detection in colonoscopy videos," *Med. Imag.*, vol. 10133, pp. 718–727, Feb. 2017.

[15] R. V. M. da Nóbrega, P. P. Rebouças Filho, M. B. Rodrigues, S. P. P. da Silva, C. M. J. M. D. Júnior, and V. H. C. de Albuquerque, "Lung nodule malignancy classification in chest computed tomography images using transfer learning and convolutional neural networks," *Neural Comput. Appl.*, vol. 32, no. 15, pp. 11065–11082, Aug. 2020.

[16] A. Masood, B. Sheng, P. Li, X. Hou, X. Wei, J. Qin, and D. Feng, "Computer-assisted decision support system in pulmonary cancer detection and stage classification on CT images," *J. Biomed. Inf.*, vol. 79, pp. 117–128, Mar. 2018.

[17] X. Mo, K. Tao, Q. Wang, and G. Wang, "An efficient approach for polyps detection in endoscopic videos based on faster R-CNN," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3929–3934.

[18] G. Urban, P. Tripathi, T. Alkayali, M. Mittal, F. Jalali, W. Karnes, and P. Baldi, "Deep learning localizes and identifies polyps in real-time With 96% accuracy in screening colonoscopy," *Gastroenterology*, vol. 155, pp. 1069–1078, Oct. 2018.

[19] M. Akbari, M. Mohrekesh, S. Rafiei, S. M. R. Soroushmehr, N. Karimi, S. Samavi, and K. Najarian, "Classification of informative frames in colonoscopy videos using convolutional neural networks with binarized weights," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Honolulu, HI, USA, Jul. 2018, pp. 65–68.

[20] P. M. Shakeel, A. Tolba, Z. Al-Makhadmeh, and M. M. Jaber, "Automatic detection of lung cancer from biomedical data set using discrete AdaBoost optimized ensemble learning generalized neural networks," *Neural Comput. Appl.*, vol. 32, no. 3, pp. 777–790, Feb. 2020.

[21] S. Suresh and S. Mohan, "ROI-based feature learning for efficient true positive prediction using convolutional neural network for lung cancer diagnosis," *Neural Comput. Appl.*, vol. 32, no. 20, pp. 15989–16009, Oct. 2020.

[22] M. Masud, G. Muhammad, M. S. Hossain, H. Alhumyani, S. S. Alshamrani, O. Cheikhrouhou, and S. Ibrahim, "Light deep model for pulmonary nodule detection from CT scan images for mobile devices," *Wireless Commun. Mobile Comput.*, vol. 2020, pp. 1–8, Jul. 2020.

[23] P. M. Shakeel, M. A. Burhanuddin, and M. I. Desa, "Automatic lung cancer detection from CT image using improved deep neural network and ensemble classifier," *Neural Comput. Appl.*, vol. 32, pp. 1–14, Apr. 2020.

[24] S. U. K. Bukhari, A. Syed, S. K. A. Bokhari, S. S. Hussain, S. U. Armaghan, and S. S. H. Shah, "The histological diagnosis of colonic adenocarcinoma by applying partial self supervised learning," *MedRxiv*, vol. 2020, pp. 1–11, Jan. 2020.

[25] S. Mangal, A. Chaurasia, and A. Khajanchi, "Convolution neural networks for diagnosing colon and lung cancer histopathological images," 2020, *arXiv:2009.03878*.

[26] B. K. Hatuwal and H. C. Thapa, "Lung cancer detection using convolutional neural network on histopathological images," *Int. J. Comput. Trends Technol.*, vol. 68, no. 10, pp. 21–24, Oct. 2020.

[27] A. A. Borkowski, M. M. Bui, L. Brannon Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides, "Lung and colon cancer histopathological image dataset (LC25000)," 2019, *arXiv:1912.12142*.

[28] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, 2020. 10.1007/s10462-020-09825-6.

[29] B. S. Rao, "Dynamic histogram equalization for contrast enhancement for digital images," *Appl. Soft Comput.*, vol. 89, Apr. 2020, Art. no. 106114.

[30] P. Shanmugavadivu and K. Balasubramanian, "Particle swarm optimized multi-objective histogram equalization for image enhancement," *Opt. Laser Technol.*, vol. 57, pp. 243–251, Apr. 2014.

[31] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. New York, NY, USA: Prentice-Hall, 2002.

[32] S. D. Chen and A. R. Ramli, "Preserving brightness in histogram equalization-based contrast enhancement techniques," *Digit Signal Process*, vol. 14, no. 5, pp. 413–428, 2004.

[33] K. G. Dhal, A. Das, S. Ray, J. Gálvez, and S. Das, "Histogram equalization variants as optimization problems: A review," *Arch. Comput. Methods Eng.*, vol. 28, no. 3, pp. 1471–1496, May 2021, doi: 10.1007/s11831-020-09425-1.

[34] W. Shen, M. Zhou, F. Yang, D. Yu, D. Dong, C. Yang, Y. Zang, and J. Tian, "Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification," *Pattern Recognit.*, vol. 61, pp. 663–673, Jan. 2017.

[35] Z. Yuan, M. I. Yazdanabadi, D. Mokkapati, R. Panvalkar, J. Y. Shin, N. Tajbakhsh, S. Gurudu, and J. Liang, "Automatic polyp detection in colonoscopy videos," *Proc. SPIE*, vol. 10133, Feb. 2017, Art. no. 101332K.

[36] T. Babu, D. Gupta, T. Singh, and S. Hameed, "Colon cancer prediction on different magnified colon biopsy images," in *Proc. 10th Int. Conf. Adv. Comput. (ICoAC)*, Chennai, India, Dec. 2018, pp. 277–280.

[37] P. M. Shakeel, M. A. Burhanuddin, and M. I. Desa, "Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks," *Measurement*, vol. 145, pp. 702–712, Oct. 2019, doi: 10.1016/j.measurement.2019.05.027.

[38] M. Masud, G. Muhammad, M. S. Hossain, H. Alhumyani, S. S. Alshamrani, O. Cheikhrouhou, and S. Ibrahim, "Light deep model for pulmonary nodule detection from CT scan images for mobile devices," *Wireless Commun. Mobile Comput.*, vol. 2020, pp. 1–8, Jul. 2020.

[39] M. Masud, N. Sikder, A.-A. Nahid, A. K. Bairagi, and M. A. AlZain, "A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework," *Sensors*, vol. 21, no. 3, p. 748, Jan. 2021, doi: 10.3390/s21030748.

SHAHID MEHMOOD received the master's degree in computer science from Riphah International University, Lahore, Pakistan. He is currently pursuing the Ph.D. degree in computing with Riphah International University, Islamabad, Pakistan. He joined the Riphah School of Computing and Innovation, Riphah International University, Lahore, as a Lecturer, in 2019. His research interests include machine learning, deep learning, image processing, and big data analytics.

TAHER M. GHAZAL (Member, IEEE) received the B.Sc. degree in software engineering from Al Ain University, in 2011, the M.Sc. degree in information technology management from The British University in Dubai, associated with The University of Manchester and The University of Edinburgh, in 2013, and the Ph.D. degree in IT/software engineering from Damascus University, in 2019. He is currently pursuing the Ph.D. degree in information science and technology with Universiti Kebangsaan Malaysia. He has more than ten years of extensive and diverse experience as an Instructor, a Tutor, a Researcher, a Teacher, a IT Support/Specialist Engineer, and a Business/Systems Analyst. He served in Engineering, Computer Science, ICT, Head of STEM, and Innovation Departments. He was also involved in quality assurance, accreditation, and data analysis, in several governmental and private educational institutions under KHDA, Ministry of Education, and Ministry of Higher Education and Scientific Research, United Arab Emirates. His research interests include the IoT, IT, artificial intelligence, information systems, software engineering, web developing, building info. modeling, quality of education, management, big data, quality of software, and project management. He is actively involved in community services in the projects and research field.

MUHAMMAD ADNAN KHAN received the B.S. and M.Phil. degrees from the International Islamic University, Islamabad, Pakistan, by obtaining a scholarship award from the Punjab Information and Technology Board, Government of Punjab, Pakistan, and the Ph.D. degree from ISRA University, Islamabad, by obtaining a scholarship award from the Higher Education Commission, Islamabad, in 2016. Before joining Gachon University, he has worked in various academic and industrial roles in Pakistan. He is currently working as an Assistant Professor at the Pattern Recognition and Machine Learning Laboratory, Department of Software, Gachon University, Republic of Korea, and an Associate Professor with the Faculty of Computing, Riphah School of Computing and Innovation, Riphah International University, Lahore Campus, Lahore, Pakistan. He has been teaching graduate and undergraduate students in computer science and engineering for the past 12 years. Currently, he is guiding five Ph.D. scholars and six M.Phil. scholars. He has published more than 200 research articles with Cumulative JCR-IF of over 390 in reputed international journals as well as international conferences. His research interests primarily include machine learning, MUD, image processing and medical diagnosis, and channel estimation in multi-carrier communication systems using soft computing.

MUHAMMAD ZUBAIR (Senior Member, IEEE) received the M.S. degree in information technology from Hamdard University Islamabad and the Ph.D. degree in EE from the International Islamic University, Islamabad, Pakistan. He is currently working as a Professor at the Riphah Faculty of Computing, Riphah International University, Islamabad. His research interests include multimedia analysis and processing with focus on medical, information security, machine learning, and deep learning.

MUHAMMAD TAHIR NASEEM received the B.S. degree in CS from the University of the Punjab, Lahore, Pakistan, the M.S. degree in EE from the International Islamic University, Islamabad, Pakistan, and the Ph.D. degree in EE from Isra University, Hyderabad, Pakistan. He is currently working as an Associate Professor at the Riphah School of Computing and Innovation (RSCI), Riphah International University, Lahore, and a Research Professor at the Human Ecology Research Center, Yeungnam University, Gyeongsan-si, Republic of Korea. His research interests include multimedia analysis and processing with focus on medical, information security, machine learning, and deep learning.

TAUQEER FAIZ received the master's degree from Staffordshire University, U.K. He is a Researcher in the area of information systems. He holds a rich experience of teaching, research, and consultancy. He has taught subjects like business analytics, machine learning, data structures, mobile applications development, 3D games development, and virtual reality. He has also served as a Senior IT Trainer for Microsoft, Android, and Apple at different institutions of Unite Arab Emirates. He is working as an Instructor at Skyline University College, Al Sharjah, where he conducts the laboratory sessions and assist students in their projects. His research interests include business analytics and virtual reality with a focus on gamification, shareholder engagement, sustainability, and e-commerce start-ups. He is an active member of Microsoft, Android, and Unity official platforms.

MUNIR AHMAD (Member, IEEE) received the Master of Computer Science degree from the Virtual University of Pakistan, in 2018. He is currently pursuing the Ph.D. degree in computer science with the School of Computer Science, National College of Business Administration and Economics. He has spent several years in industry. He is currently working as the Executive Director/Head of the IT Department, United International Group, Lahore, Pakistan. He has vast experience in data management and efficient utilization of resources at multinational organizations. He has conducted many research studies on sentiment analysis and utilization of AI for prediction on various healthcare issues. His research interests include data mining, big data, and artificial intelligence.