

Diagnosis of lung and colon cancer based on clinical pathology images using convolutional neural network and CLAHE framework

Sugondo Hadiyoso*, Suci Aulia, Indrarini Dyah Irawati

School of Applied Science, Telkom University, Bandung, West Java, 40257, Indonesia

ABSTRACT


Cancer is a non-contagious disease that is the leading cause of death globally. The most common types of cancer with high mortality are lung and colon cancer. One of the efforts to reduce cases of death is early diagnosis followed by medical therapy. Tissue sampling and clinical pathological examination are the gold standard in cancer diagnosis. However, in some cases, pathological examination of tissue to the cell level requires high accuracy, depending on the contrast of the pathological image, and the experience of the clinician. Therefore, we need an image processing approach combined with artificial intelligence for automatic classification. In this study, a method is proposed for automatic classification of lung and colon cancer based on a deep learning approach. The object of the image that is classified is the histopathological image of normal tissue, benign cancer, and malignant cancer. Convolutional neural network (CNN) with VGG16 architecture and Contrast Limited Adaptive Histogram Equalization (CLAHE) were employed for demonstration of classification on 25000 histopathological images. The simulation results show that the proposed method is able to classify with a maximum accuracy of 98.96%. The system performance using CLAHE shows a higher detection accuracy than without using CLAHE and is consistent for all epoch scenarios. The comparative study shows that the proposed method outperforms some previous studies. With this proposed method, it is hoped that it can help clinicians in diagnosing cancer automatically, with low cost, high accuracy, and fast processing on large datasets.

Keywords: Cancer, Classification, Deep learning, Histopathological, CNN.

OPEN ACCESS 

Received: February 7, 2022
Revised: December 5, 2022
Accepted: December 16, 2022

Corresponding Author:
Sugondo Hadiyoso
sugondo@telkomuniversity.ac.id

 **Copyright:** The Author(s). This is an open access article distributed under the terms of the [Creative Commons Attribution License \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted distribution provided the original author and source are cited.

Publisher:
[Chaoyang University of Technology](https://www.telkomuniversity.ac.id/)
ISSN: 1727-2394 (Print)
ISSN: 1727-7841 (Online)

1. INTRODUCTION

Lung and colon cancer has been identified being one of the leading causes of death worldwide. It is one of the most dangerous tumors that can harm a person's health (Chang and Abu-Amara, 2021). It has the highest mortality rate of all tumors, and it is also the leading cause of cancer death in both men and women (Siegel et al., 2017).

In recent years, deep learning algorithms for detecting and identifying lung and colon cancer using histopathological slide, X-rays, MRIs, Endoscopic, and CT scans images have become a main research topic (Ali and Ali, 2021). Since their reemergence, Convolutional Neural Networks (CNN) were used in almost all deep learning algorithms for medical image classification (Sharma et al., 2020). Convolutional neural networks are divided into two types, each with a fundamentally different working concept. The two types of convolutional neural networks are traditional CNN and separable CNN (Elnakib et al., 2020).

Several research groups have made progress in the field of lung cancer detection using computer vision based on histopathological images in recent years. Saif et al. (2020) used VGG16 as CNN model to diagnose 1500 of lung cancer images, the result achieve 98.55% of accuracy. Ayad et al. (2020) proposed a CT-based lung segmentation method that used the weighted Softmax function. Using the LIDC- IDRI CT lung images

database, the system obtained a maximum segmentation accuracy of 98.90%. Another research by Masud. et al. (2021). was obtained Digital Image Processing (DIP) and Deep Learning (DL) to classify five types of lung and colon cancer from histopathological slide images with the highest 96.33% of accuracy. Hatuwal and Thapa (2020) looked at benign tissue, squamous cell carcinoma, and adenocarcinoma to categorize lung cancer forms. The training and testing accuracy of the CNN model were 96.11 and 97.2 percent, respectively. Since CNN has showed promise in classifying lung and colon cancer (Saleh et al., 2021), the performance of this algorithm is being assessed in this study. On the other approach, Contrast Limited Adaptive Histogram Equalization (CLAHE) is gaining traction in the medical imagery categorization field due to its simple models for image enhancement (Sepasian et al., 2008; Ma et al., 2018; Stimper et al., 2019; Buddha et al., 2020).

As a result, in this study, we combine VGG16 as the CNN model and CLAHE to classify the histopathology slide images of lung and colon cancer. In this proposed research, image preprocessing to balance the colors and maintain the overall detail for optimal feature learning is an important part of this research.

2. MATERIALS AND METHODS

2.1 Histopathological Image Dataset

This study used histopathological image dataset of lung and colon cancer collected from the LC25000 dataset by Borkowski et al. (2019). Color image with the .jpeg format totaling 25,000 data consisting of five categories, such as: Colon Adenocarcinoma, Benign Colonic Tissue, Lung Adenocarcinoma, Benign Lung Tissue, and Lung Squamous Cell Carcinoma. The class name, class ID and number of samples of each type are shown in Table 1. Sample histopathological images from each class used in this study are presented in Fig. 1.

2.2 Convolutional Neural Network (CNN)

Convolutional layer, pooling layers, and full connections are the basic stages in CNN architecture (Kalaivani et al., 2020; Ali and Ali, 2021), as depicted in Fig. 2.

Convolutional layers are made up of multiple nodes that extract data from the input images. To fulfill the basic purpose of features extraction on input images, these layers use a huge number of kernels or filters. Many feature maps are constructed inside the step-1 convolutional process employing detector features to obtain the first convolutional layer using Equation (1).

$$(f * g)(t) \cong \int_{-\infty}^{\infty} f(\tau)g(t - \tau)d\tau \quad (1)$$

Here f is input image, g is kernel or filter matrix. This convolutional are in time domain (t) and the shifting are

based on τ . Pooling layers are commonly used after convolutional layers. The main goal of these layers is to reduce the input data's dimension, in spatial domain are width and height before transferring into the next layers. The computing efficiency of CNN models is aided by these layers. Fully-connected layers are entirely connected to one end of the CNN network's preceding layers (Mamdouh et al., 2021). These elements can help learn output probabilities, which are used to assess the model's accuracy.

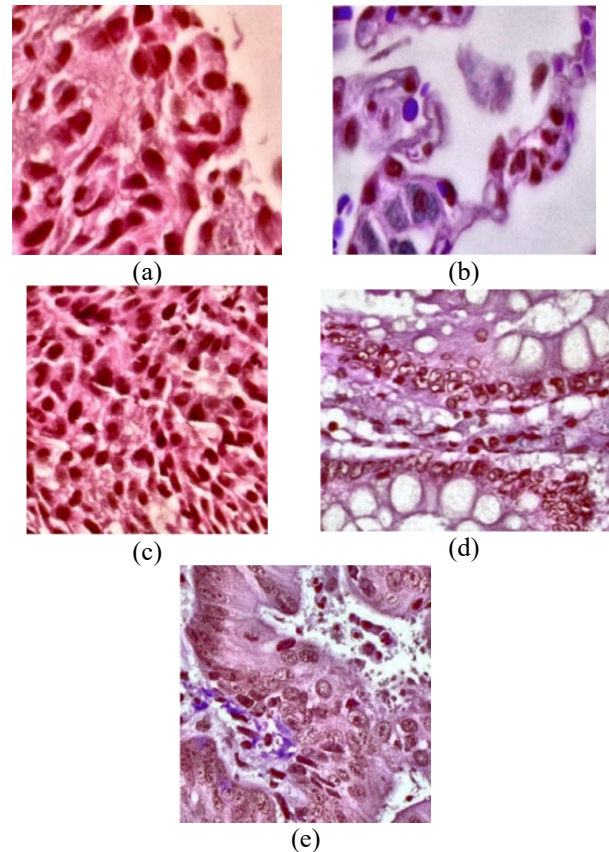


Fig. 1. Histopathological image (a) lung_ACA (b) lung_N (c) lung_SCC (d) colon_N (e) colon_ACA

2.3 Architecture VGG16

The CNN architecture employed in this study is Visual Geometry Group 16, also known as VGG16. Fig. 3 presents the VGG16 architecture. Layer Cov1 uses a fixed size RGB image of 224×224 . This input goes through several convolution layers and is filtered with a very small dimension of 3×3 . One of the configurations also uses a 1×1 convolution filter, which is a linear transformation followed by a non-linear transformation. Max pooling consists of 5 layers which are used for spatial polling with a 2×2 pixel window and lowering the sample resolution by 2 factors. The three Fully Connected (Fc) layers follow a convoluted layer stack with 4096 channels in the first 2 Fc layers and 1000 channels for the last Fc. This layer connects the entire network.

Table 1. LC25000 dataset description

| The type of cancer | Class name | Class ID | Number of samples |
|------------------------------|------------|----------|-------------------|
| Colon adenocarcinoma | Colon_ACA | 0 | 5000 |
| Benign colonic tissue | Colon_N | 1 | 5000 |
| Lung adenocarcinoma | Lung_ACA | 2 | 5000 |
| Benign lung tissue | Lung_N | 3 | 5000 |
| Lung squamous cell carcinoma | Lung SCC | 4 | 5000 |

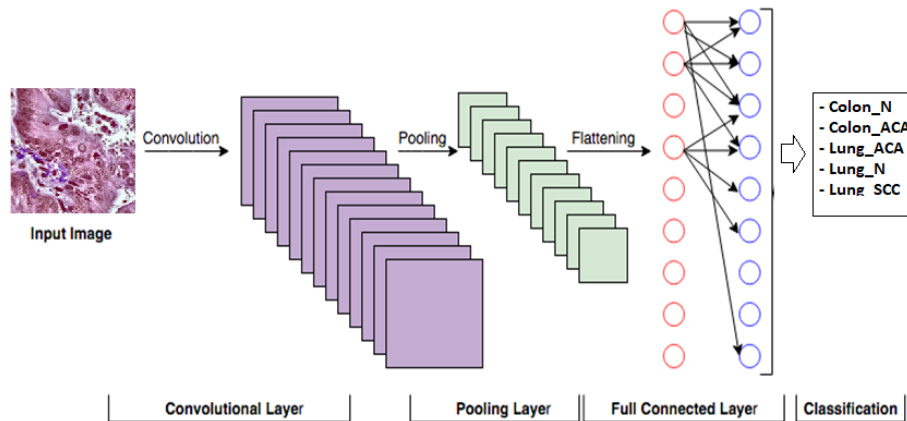


Fig. 2. Basic architecture of CNN

2.4 Contrast Limited Adaptive Histogram Equalization (CLAHE)

CLAHE is a more advanced variant of the Adaptive Histogram Equalization algorithm (AHE). This is utilized in image contrast improvisation, equalizes the lightness and prevents the images from noise (Maheshan et al., 2018). The OpenCV CLAHE API is utilized to improve the image in this approach. First, the image is converted to grayscale, and then the grayscale image is supplied into the OpenCV CLAHE API. The API delivers a new image that is an upgraded version of the original. The processes involved in CLAHE are represented in Table 2 (Pizer et al., 1990).

2.5 Proposed Framework

The general proposed framework for lung and colon cancer classification is presented in Fig. 4. There are two main phases, namely training, and testing. The input data is a labeled RGB image consisting of five classes. The ratio of the number of images for training and testing is 80:20. All input images are preprocessed using CLAHE for contrast enhancement. The following process is feature extraction using the convolution operation and pooling and classification. At the training stage, variations in the number of epochs are applied to find the optimum parameters so that the prediction model produces the highest accuracy. The model with the best parameters is stored and used in the testing phase.

The basic CNN architecture used in this study is VGG16 with a transfer learning system on ImageNet. Then the model that is generated using a transfer learning system is called the pre-trained model. The pre-trained model is divided into two main components, namely the base model

and the custom head network. After the Pre-trained model is generated, the next step is to create a custom head model which will later be added to the main model. Meanwhile, the optimization system used in this study is the Adam Optimizer with parameters including: learning rate 1e-5, decay step 755, decay rate 0.9, and for the loss function using categorical cross-entropy. Furthermore, at the training stage using a batch size of 32 and a checkpoint model that aims to store the model with the best accuracy in each epochs scenario. A summary of the CNN-architecture parameters which were employed in this study is presented in Table 3. The performance test of the proposed method was carried out on images with CLAHE and without CLAHE. The test uses 10 epochs, 20 epochs, 30 epochs, 40 epochs and 50 epochs to find the best performance.

Table 3. CNN parameter which is used in this study

| Variables | Value |
|--------------------------------|--------------------------|
| Architecture | VGG16 |
| Image size | 224 × 224 |
| Epochs | 10, 20, 30, 40 and 50 |
| Batch size | 32 |
| Filters | 64 |
| 2D Convolutional layers (size) | 3 (3 × 3) |
| Convolutional layer activation | Relu |
| Dense layer activation | Softmax |
| Compiler optimizer | Adam |
| Compiler loss | Categorical crossentropy |
| Learning rate | 1e-5 |
| Decay steps | 755 |
| Decay rate | 0.9 |

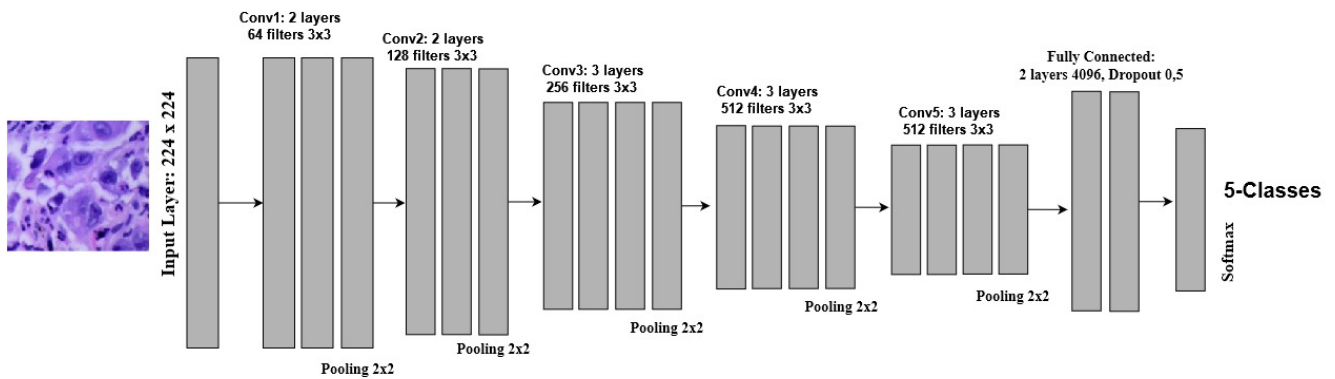


Fig. 3. VGG16 architecture (Li and Guo, 2018)

Table 2. The steps of CLAHE algorithm

| | |
|------------|---|
| Data input | : Image (img) |
| Step 1 | : Read the image as img |
| Step 2 | : Plot the histogram of img |
| Step 3 | : Using OpenCV to apply the CLAHE API function |
| Step 4 | : Save the new image from step 3 as “clahe_img.jpg” |
| Result | “clahe_img.jpg” is the enhanced image |

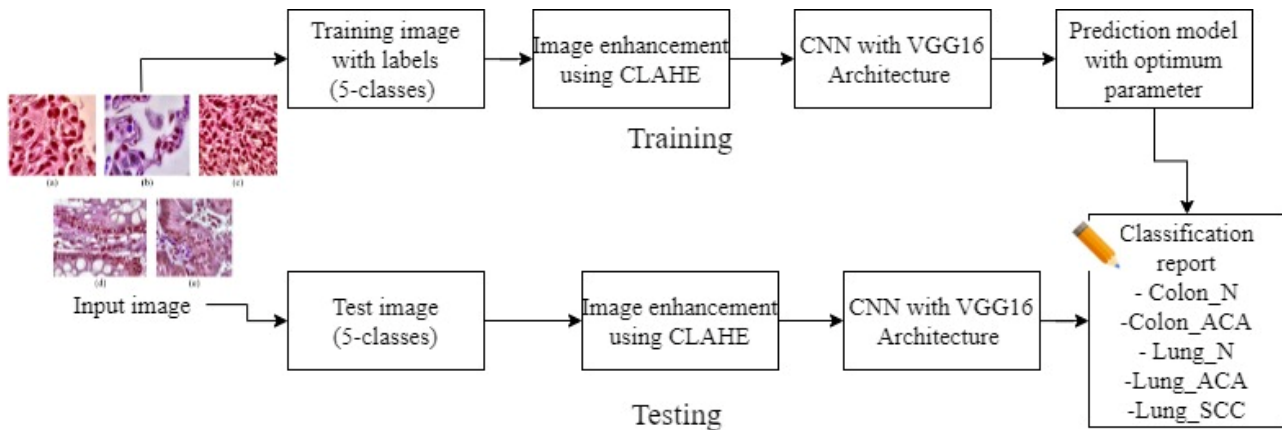


Fig. 4. Proposed classification framework

Testing the proposed method using the Google Colab Pro application with high tensor processing unit and High-RAM hardware. A total of 25000 images were processed consisting of 5,000 test images and 20,000 training images representing all histopathological images of cancer type. The test image consists of 1021 colon_ACA, 1000 colon_N, 985 lung_ACA, 989 lung_N, and 1005 lung_SCC.

3. RESULTS AND DISCUSSION

Fig. 5 shows the raw images and CLAHE images for each type of cancer. Visually, the CLAHE image seen more contrasted with the clear boundaries than the raw image. So, it is hoped that using CLAHE will get better system performance in terms of higher classification accuracy. Fig. 6 shows the mean feature map of each type of histopathological cancer image from one of the

convolutional stages. From the generated feature map, it can be seen that there are differences in characteristics between classes. Direct observation showed that the histopathological images of lung and colon have different characteristics.

Table 4 is a summary of the accuracy of the classification simulation results on 5000 test images using the proposed model. Table 4 shows the highest accuracy of 98.96% with specificity and sensitivity of 99.74% and 98.96%, respectively. The number of epochs affects the accuracy which is linear with increasing accuracy both using CLAHE and without CLAHE. From this test, it is also known that the CLAHE technique is able to generate higher accuracy than without CLAHE for all scenarios with the number of epochs as shown in Fig. 7. Simple analysis that CLAHE increases the contrast of the image so that better characterization results will be obtained.

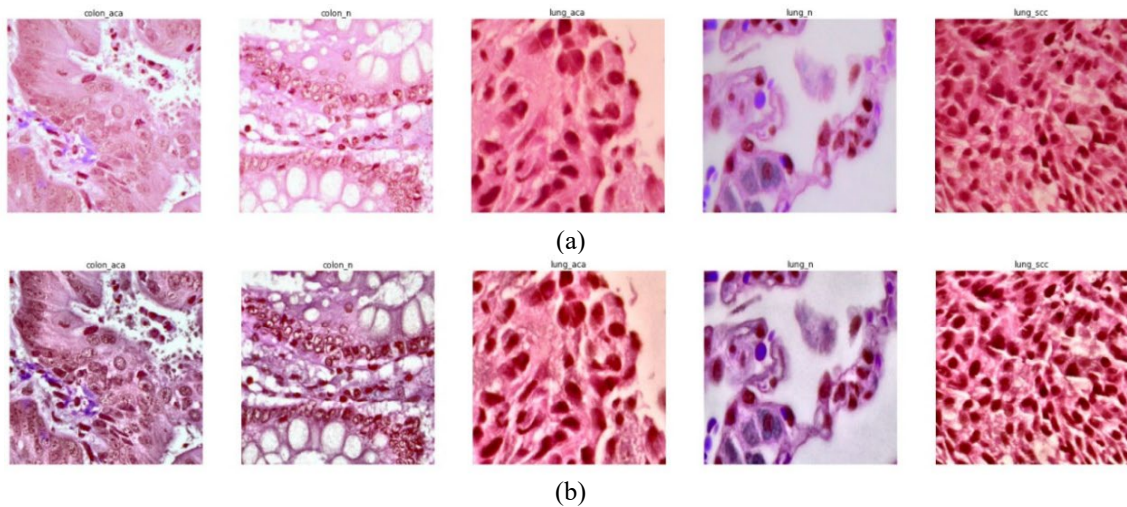


Fig. 5. Histopathological image (a) raw (b) result of CLAHE

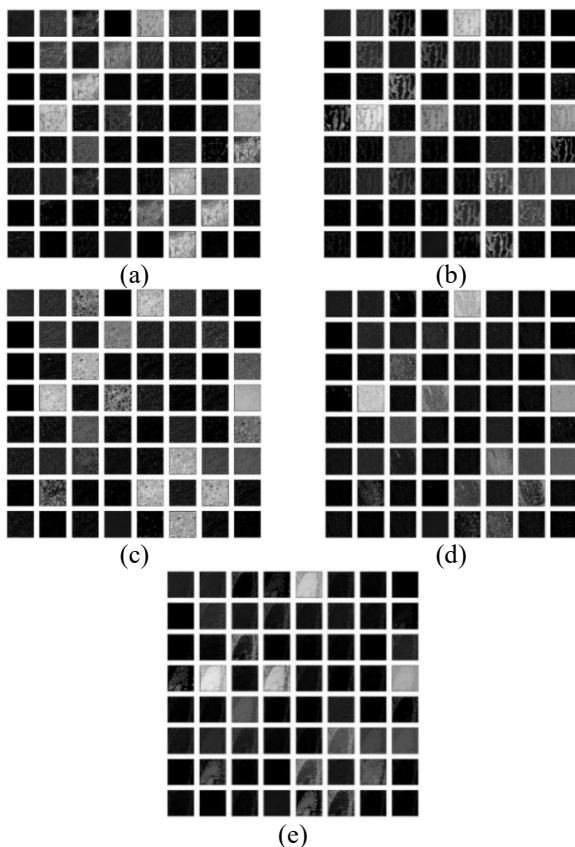


Fig. 6. Feature map of each class (a) lung_ACA (b) lung_N (c) lung_SCC (d) colon_ACA (e) colon_N

Based on the confusion matrix presented in Table 5, it is known that the detection accuracy for normal cases reaches 100% for normal lung pathology. Meanwhile, the detection accuracy of lung ACA generates the highest detection error compared to other types of pathology. Classification errors in Lung ACA are generally misclassified to Lung SCC because the histopathological images have a high similarity as shown in Fig. 1(a) and Fig. 1(c).

Table 4. Accuracy for each scenario

| Epoch | Accuracy (%) | |
|-------|--------------|--------------|
| | Non-CLAHE | CLAHE |
| 10 | 98.04 | 98.5 |
| 20 | 98.82 | 98.82 |
| 30 | 98.78 | 98.86 |
| 40 | 98.86 | 98.86 |
| 50 | 98.86 | 98.96 |

Based on the confusion matrix, the proposed method is also capable of classifying histopathological types of lung and colon images with high accuracy. The proposed method generates a small error, whether the lung image is detected as a colon image or vice versa.

The performance of the proposed method is then compared with several previous studies which used the same dataset or other cancer histopathological datasets. Table 6 shows a comparison of the performance of this study with several previous similar studies over a two-year period. From Table 3 it can be seen that the classification accuracy of our proposed method outperforms almost all other studies except the study by Nishio et al. (2021) and Abbas et al. (2020). In study by Nishio et al. (2021) and Abbas et al. (2020), the highest accuracy achieved was slightly higher than our study, but in the case of three classes of lung histopathological image classification. Actually, the comparison cannot be done directly, even in several studies that use the same dataset, they have different test scenarios, for example the case of two-class or three-class classification. However, the histopathological image classification method proposed in this study is capable of generating high accuracy in application to large datasets. In the end, it is hoped that the proposed method can be used by clinicians in assisting clinical diagnosis.

4. CONCLUSION

The classification method based on CNN with VGG16

architecture and CLAHE can classify histopathological cancer images of 5 classes consisting of lung and colon cancer. The CNN architecture consists of a convolutional layer, a max-pooling layer, and a fully connected layer, and a classification layer to classify the five classes. The experimental results show that the proposed system outperforms the previous research. The maximum accuracy reaches 98.96%, with 98.96% Precision, 0.26% False Positive Rate and 98.96% F1 score. In addition, the results of early diagnosis are shown based on the sensitivity and specificity parameters obtained at 98.96% and 99.74%,

respectively. The test scenario using CLAHE shows a higher accuracy performance than without using CLAHE. Based on the simulation, the proposed method produces high accuracy and reliability so that it is expected to be used to support clinical diagnosis and early detection of cancer based on clinical pathology images. In further research, we focused on preprocessing input, testing with grayscale images, and exploring the use of other CNN architectures, such as: resnet, alexnet, or VGG19.

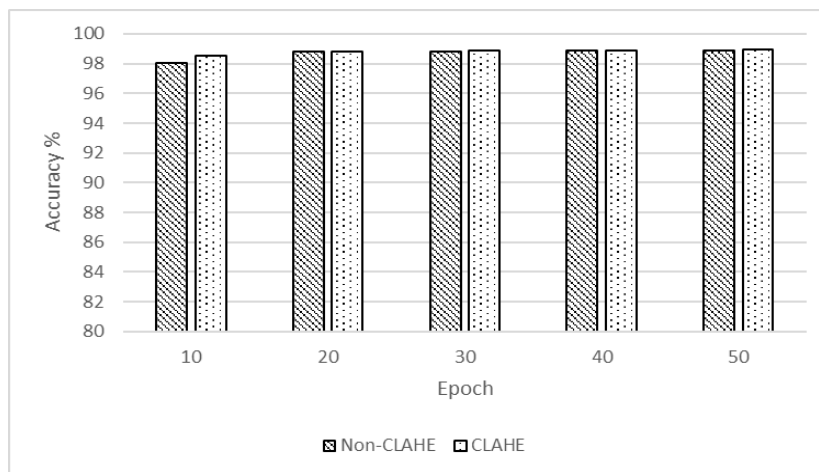


Fig. 7. Performance comparison of CLAHE

Table 5. Confusion matrix (using CLAHE and 50 epoch)

| | | Predicted class | | | | | Acc. (%) |
|--------------|-----------|-----------------|---------|----------|--------|----------|----------|
| | | colon ACA | colon N | lung ACA | lung N | lung SCC | |
| Ground truth | colon_ACA | 1011 | 0 | 6 | 0 | 4 | 99.02 |
| | colon_N | 3 | 997 | 0 | 0 | 0 | 99.7 |
| | lung_ACA | 6 | 0 | 963 | 1 | 15 | 97.7 |
| | lung_N | 0 | 0 | 0 | 989 | 0 | 100 |
| | lung_SCC | 2 | 1 | 14 | 0 | 988 | 98.3 |
| Average | | | | | | 98.96 | |

Table 6. Comparison with previous studies

| Author | Image type/dataset | Image/number of classes | Method | ACC (%) |
|-----------------------------|---|----------------------------------|--|----------------|
| Mangal et al., 2020 | Histopathological / LC25000 | Lung and colon / 3 and 2 classes | CNN-RMSprop | 97.89 |
| Nishio et al., 2021 | Histopathological / private dataset and LC25000 | Lung / 5 and 3 classes | Homology-based image processing + Machine learning | 78.3 and 99.33 |
| Abbas et al., 2020 | Histopathological / LC25000 | Lung / 3 classes | CNN-ResNet | 99.8 |
| Masud et al., 2021 | Histopathological / LC25000 | Lung and Colon / 5 classes | CNN | 96.33 |
| Hatuwal and Thapa, 2020 | Histopathological / LC25000 | Lung / 3 classes | CNN | 97.2 |
| Saleh et al., 2021 | CT-Images | Lung / 4 classes | CNN-SVM | 97.91 |
| Khalid Bukhari et al., 2020 | Histopathological / LC25000 | Colon | CNN-ResNet | 93.91 |
| Irawati et al., 2021 | Histopathological / LC25000 | Lung / 3 classes | Compressive sensing - KNN | 88 |
| This study | Histopathological / LC25000 | Lung and Colon / 5 classes | CNN-CLAHE-VGG16 | 98.96 |

ACKNOWLEDGMENT

This research would like to thank the support of School Applied Science, Telkom University

REFERENCES

- Abbas, M.A., Bukhari, S.U.K., Syed, A., Shah, S.S. 2020. The histopathological diagnosis of adenocarcinoma squamous cells carcinoma of lungs by artificial intelligence: A comparative study of convolutional neural networks, *MedRxiv*, 1–13.
- Ali, M., Ali, R. 2021. Multi-input dual-stream capsule network for improved lung and colon cancer classification. *Diagnostics*, 11, 1–18.
- Ayad, H., Ghindawi, I.W., Kadhm, M.S. 2020. Lung segmentation using proposed deep learning architecture. *International Journal of Online and Biomedical Engineering*, 16, 141–147.
- Borkowski, A.A., Bui, M.M., Thomas, L.B., Wilson, C.P., DeLand, L.A., Mastorides, S.M. 2019. Lung and colon cancer histopathological image dataset (LC25000). *ArXiv:1912.12142v1*, 1–2.
- Chang, Y., Abu-Amara, F. 2021. An efficient hybrid classifier for cancer detection. *International Journal of Online and Biomedical Engineering*, 17, 76–97.
- Elnakib, A., Amer, H.M., Abou-Chadi, F.E.Z. 2020. Early lung cancer detection using deep learning optimization. *International Journal of Online and Biomedical Engineering*, 16, 82–94.
- Hatuwal, B.K., Thapa, H.C. 2020. Lung cancer detection using convolutional neural network on histopathological images. *International Journal of Computer Trends and Technology*, 68, 21–24.
- Irawati, I.D., Hadiyoso, S., Fahmi, A. 2021. Compressive sensing in lung cancer images for telemedicine application. *ACM International Conference Proceeding Series*, 55–61.
- Kalaivani, N., Manimaran, N., Sophia, S., D. Devi, D. 2020. Deep learning based lung cancer detection and classification. *IOP Conference Series: Materials Science and Engineering*, 994, 1–5.
- Khalid Bukhari, S.U., Syed, A., Arsalan Bokhari, S.K., Hussain, S.S., Armaghan, S.U., Hussain Shah, S.S. 2020. The histological diagnosis of colonic adenocarcinoma by applying partial self supervised learning. *MedRxiv*, 2020, 1–11.
- Li, Y.T., Guo, J.I. 2018. A VGG-16 based Faster RCNN Model for PCB error inspection in industrial AOI applications. *IEEE International Conference on Consumer Electronics-Taiwan*, 1–2.
- Ma, J., Fan, X., Yang, S.X., Zhang, X., Zhu, X. 2018. Contrast limited adaptive histogram equalization-based fusion in YIQ and HSI color spaces for underwater image enhancement. *International Journal of Pattern Recognition and Artificial Intelligence*, 32, 1–26.
- Maheshan, M.S., Harish, B.S., Nagadarshan, N. 2018. On the use of image enhancement technique towards robust sclera segmentation. *Procedia Computer Science*, 143, 466–473.
- Mamdouh, R., El-Khamisy, N., Amer, K., Riad, A., El-Bakry, H.M. 2021. A new model for image segmentation based on deep learning. *International Journal of Online and Biomedical Engineering*, 17, 28–47.
- Mangal, S., Chaurasia, A., Khajanchi, A. 2020. Convolution neural networks for diagnosing colon and lung cancer histopathological images. *Computer Science, Engineering ArXiv*, 2020, 1–10.
- Masud, M., Sikder, N., Nahid, A. Al, Bairagi, A.K., Alzain, M.A. 2021. A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework. *Sensors (Switzerland)*, 21, 1–21.
- Buddha, H.K., Meka, J.S., Choppala, P. 2020. OCR image enhancement & implementation by using CLAHE algorithm. *Mukt Shabd Journal*, 9, 3595–3599.
- Nishio, M., Nishio, M., Jimbo, N., Nakane, K. 2021. Homology-based image processing for automatic classification of histopathological images of lung tissue. *Cancers*, 13, 1–12.
- Pizer, S.M., Johnston, R.E., Ericksen, J.P., Yankaskas, B.C., Muller, K.E. 1990. Contrast-limited adaptive histogram equalization: Speed and effectiveness. *Proceedings of the First Conference on Visualization in Biomedical Computing*, 1990, 337–345.
- Saif, A., Qasim, Y.R.H., Al-Sameai, H.A.M.H., Farhan Ali, O.A., Hassan, A.A.M. 2020. Multi paths technique on convolutional neural network for lung cancer detection based on histopathological images. *International Journal of Advanced Networking and Applications*, 12, 4549–4554.
- Saleh, A.Y., Chin, C.K., Penshie, V., Al-Absi, H.R.H. 2021. Lung cancer medical images classification using hybrid cnn-svm. *International Journal of Advances in Intelligent Informatics*, 7, 151–162.
- Sepasian, M., Balachandran, W., Mares, C. 2008. Image Enhancement for fingerprint minutiae-based algorithms using CLAHE, Standard deviation analysis and sliding neighborhood. *Lecture Notes in Engineering and Computer Science*, 2173, 1199–1203.
- Sharma, P., Bora, K., Kasugai, K., Balabantaray, B.K. 2020. Two stage classification with CNN for colorectal cancer detection. *Oncologie*, 22, 129–145.
- Siegel, R.L., Miller, K.D., Jemal, A. 2017. Cancer statistics, 2017. *CA Cancer Journal for Clinicians*, 67, 7–30.
- Stimper, V., Bauer, S., Ernstorfer, R., Schölkopf, B., Xian, R.P. 2019. Multidimensional contrast limited adaptive histogram equalization. *IEEE Access*, 7, 165437–165447.