Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2022, 2022:81
https://doi.org/10.28919/cmbn/7611
ISSN: 2052-2541

HISTOPATHOLOGY OF LUNG CANCER CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORK WITH GAMMA CORRECTION

WAHYUDI SETIAWAN*, MUHAMMAD MUSHLIH SUHADI, HUSNI, YOGA DWITYA PRAMUDITA

Department of Informatics, University of Trunojoyo Madura, Bangkalan, East Java 69162, Indonesia

Abstract: Lung cancer is a degenerative disease that causes the growth of abnormal tissue to become autonomous and malignant. There are two types of lung cancer: adenocarcinoma and squamous cell carcinoma. Both types of cancer can be examined with varied techniques, one of them is histopathological examination. This examination is performed by expert analysts who can distinguish normal tissue from others. Manual inspection requires a long observation time and energy, therefore a computer-based classification system is created. In this study, Convolutional Neural Network (CNN) with gamma correction was implemented. Gamma correction is a process to adjust the image light, while CNN is for feature extraction and classification. The CNN built consists of one gamma correction layer, three convolution layers, three max-pooling layers, and two fully connected layers. The data consists of 3,000 images from the public dataset Lc25000. It has a normal and two lung cancer classes i.e adenocarcinoma and squamous cell carcinoma. In this study, we used gamma values of 0.8, 1.0, and 1.2. The testing was carried out using five-fold cross-validation. It obtained the highest accuracy of 87.16% with a gamma value of 1.2.

Keywords: convolutional neural network; gamma correction; histopathology; lung cancer; images classification.

2010 AMS Subject Classification: 92C50.

*Corresponding author
E-mail address: wsetiawan@trunojoyo.ac.id
Received July 15, 2022

1. INTRODUCTION

The lungs are part of the respiratory system that plays a role in carrying oxygen throughout the body. Abnormal cell growth in the lung can occur causing cancer or pulmonary carcinoma. This is caused by the poor environment and an unhealthy lifestyle [1]. Symptoms of lung cancer only appear when the tumor is large enough or has spread to other parts. The earlier cancer is diagnosed the higher the treatment success rate [2]. However, if it is indicated to spread to other organs, the chance of recovery is lower.

Based on data from GLOBOCAN (Global Burden Of Cancer), lung cancer is the most deaths cancer in Indonesia, with up to 30,843 incidents in 2020 [3]. It increased 80% compared to 2018 [4], [5]. Histopathological examination is required to diagnose lung cancer [5]. It depends on the level of expert experience and the decision can vary. To assist this inspection process, a computer-based system was created.

Computer-based lung cancer screening has been used by researchers using certain methods to get better performance. In 2019, Pan et al. [6] classified the histopathology of lung cancer using CNN. The CNN architecture used consists of seven layers. The result has an accuracy of 86%.

There are related works Rahman et al. [7] have investigated the effect of image enhancement using gamma correction. The results show a better than state of the art techniques. Furthermore, Shin et al. [8] have compared two methods of classifying Colorectal Cancer into two classes. The research compares the Histogram of Oriented Gradient (HOG) as feature extraction, Support Vector Machine (SVM), and CNN as a classifier. CNN consists of three convolution layers, three max-pooling, and two fully connected layers. CNN obtained a better accuracy up to 92%.

Then, Sarmiento and Fondon [9] classified the severity of breast cancer histopathology. The research uses gamma correction as a preprocessing, Segmentation Fractal Texture Analysis (SFTA) as feature extraction, and Support Vector Machine (SVM) as a classifier of the four classes of breast cancer. The accuracy obtained was 79.2%.

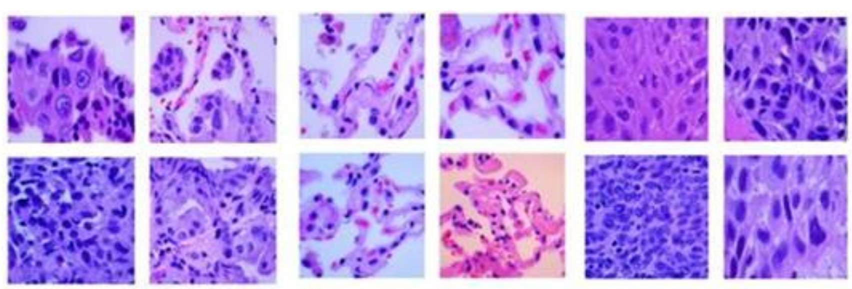
Bhattacharjee et al. [10] have classified two classes of prostate cancer histopathology. The research applies Gamma Correction for preprocessing. The CNN and Ensemble Machine Learning were compared to produce the best accuracy. The results show CNN obtained an accuracy of 94% and Ensemble Machine Learning of 92%. Based on previous research, in this study, the CNN with gamma correction was implemented for classifying the histopathology of lung cancer images.

2. PRELIMINARIES

A. Lung Cancer

Lung cancer is a disease caused by abnormal cell growth in the lungs that can spread to nearby tissues and other organs of the body. There are two types of lung cancer: adenocarcinoma and squamous cell carcinoma [11]. The diagnosis of lung cancer can be through histopathological examination.

Histopathology examines the condition of body organs based on microscopic tissue observations to determine whether they are indicated by disease or not. Generally, this examination is assisted by bronchoscopy. Histopathology is one of the gold standards in medical examinations [12]. An example of lung cancer histopathology images is shown in Figure 1.



(a) (b) (c)

Figure 1. Sample of Dataset (a) Adenocarcinoma (b) Normal (c) Squamous cell carcinoma from Lc25000 dataset

Adenocarcinoma is a type of lung cancer that undergoes glandular formation. It is characterized by tissue swelling. The cancer is characterized by long-lasting cough symptoms even after treatment. This cancer progresses slowly and is often diagnosed before spreading to other organs [13].

Squamous cell carcinoma (SCC) is another type of lung cancer that undergoes a process of keratinization and is characterized by the presence of polygonal-shaped cells [9]. This condition does not cause any symptoms in its early stages. This causes cancer to be diagnosed only after it has spread to other body parts. Therefore, early detection is an important point in increasing success during treatment. If the delayed diagnosis, the patient's five-year survival chance is less than 20% [10].

B. Gamma Correction

Gamma correction is a method to adjust the image light. It is necessary to manually calibrate the gamma value, The gamma correction formula can be seen in equation 1 [14].

$$g = u^{\gamma} \qquad (1)$$

g= Gamma correction; u= input image; γ= Gamma value

The input image in equation 1 is a positive number that has been normalized [9]. The normalization is carried out on a scale of zero to one. If it is less than one, the image will be brighter, vice-versa [9], [10], [14].

C. Convolutional Neural Network

A CNN is one of deep learning that can be used to classify an image [15]. CNN is widely implemented in image processing because it can well recognize the representation of features. The illustration of simple CNN can be seen in Figure 2.

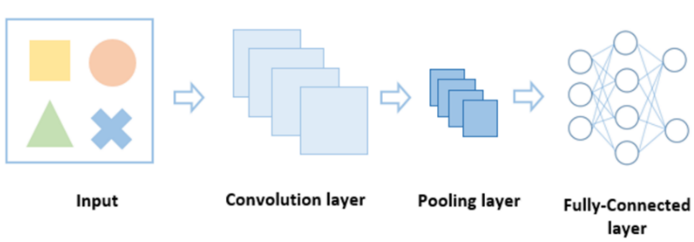


Figure 2. *Convolutional Neural Network*

CNN is divided into two main parts whose role is to perform feature extraction and classification. The feature extraction layer is composed of convolution and pooling, while the classification layer consists of a Fully-Connected Layer.

C.1 Convolutional Layer

A convolutional layer is the first layer that receives an image. The convolution layer utilizes a series of kernels to extract features from the image. The result of this extraction is commonly referred to as a feature map which is used as input for the next layer [16]. The feature map formula is shown in equation 2.

$$FM[i]_{j,k} = \sum_m \sum_n N_{[j-m,k-n]} \; F_{(m,n)} + bF \tag{2}$$

$FM[i]$: i-th feature map; F : convolutional filter; N : image

bF : bias on filter; j, k; pixels of image; m, n : pixels on the convolution filter

C.2 Pooling layer

Pooling works by sub-sampling the image. Max-pooling is the most widely applied pooling option today, where the output is the maximum value taken from the nearest grid [16].

C.3 Flatten and Fully Connected Layer

Before entering the Fully-Connected Layer (FCL), there is a Flatten process. At this stage, the feature map is converted into a one-dimensional vector so that it can be processed at the FCL [17]. FCL has a similar concept to conventional/artificial neural networks. FCL has a series of nodes that are connected from a layer to the nearest part [18]. For FCL architecture, see Figure 3.

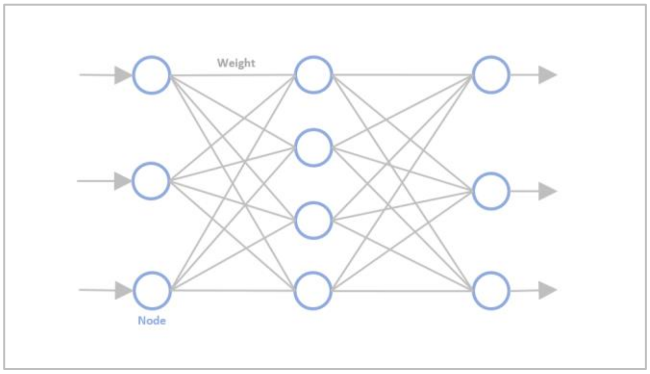


Figure 3. *Fully-Connected Layer*

FCL has an important role in determining which features are the most correlated in a certain class. So, they can be classified linearly. There is a multiplication of the input node with the weight continuously until it reaches the output. The calculation of this layer is formulated in equation 3 [16].

$$z_k = \sum_{m=1}^{n} X_m * w_{m,j} + b_{0,k} \tag{3}$$

z_k = output for hidden layer node at index k; X_m = value at node X at index m;
$w_{m,j}$ = weight X_m; $b_{0,k}$ = bias z_k

C.4 ReLU

ReLU (Rectified Linear Unit) is an activation function that can be regarded as a normalization process that compares the maximum value between the input and zero. The output of ReLU has a value of zero if the input value is negative, if the input value is positive then the output will remain [19]. The ReLU can be formulated in equation 4.

$$f(x) = \max\ (0, x) \tag{4}$$

C.5 Softmax

Softmax is an activation function that is used to determine the probability of a correlated class. This probability is obtained by normalizing the value in the previous process and interpreted with *a* value range between zero to one [20]. Suppose the probability of y_1, is obtained by calculating the exponential power of y_1 divided by the sum of each exponential to the power of the element *y*. The softmax formula is shown in equation 5.

$$S(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{m} e^{y_j}} \text{ for } i \tag{5}$$

$y = [y_1, \ldots, y_m]^T$ is vector m element. e is exponential 2,71828. $\sum_{j=1}^{m} S\ (y_i) = 1$ for *softmax*.

C.6 Categorical Cross Entropy

Categorical Cross Entropy is a multi-class loss function that is useful for calculating error values at the training phase [20]. In calculating the loss, the value of the prediction class (*p*) will be compared with the actual class value (*s*). Then, it will be updated using Gradient Descent to obtain a smaller error value. An example of a loss can be seen in Figure 4.

$$p\begin{bmatrix}0.7185\\0.0717\\0.2098\end{bmatrix}, s\begin{bmatrix}1\\0\\0\end{bmatrix} \rightarrow \boxed{-\sum_i s_i \log p_i} = 0.1436$$

Figure 4. Loss function example

C.7 Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is an optimization used to update the weight and bias parameters in the Neural Network. This optimization process is aimed at finding parameters that can minimize the loss value. The SGD optimization formula can be considered in Equation 6 [21].

$$\emptyset_n = \emptyset_0 - a(\Delta\emptyset_0) \tag{6}$$

$\emptyset_0$ = weight; $\emptyset_n$ = new weight; a = learning rate; $\Delta\emptyset_0$= weight gradient

C.8 *k*-fold cross-validation

A phase of training and testing data using *k*-fold cross-validation. Suppose the fold value is equal to five, then the data will be divided into five equal parts. One part for testing data, the other four parts for training data, and this combination keep changing. This *k*-fold process is used to get the best combination of data during training.

D. Research Method

This section discusses the steps for creating a lung cancer histopathology image classification. The data is divided into two parts, training and testing data. We used five-fold cross-validation. Furthermore, the data was processed with gamma correction and CNN. The result of the process is a model used for data testing. The output shows one of the lung cancer classes. Figure 5 is the proposed architecture system.

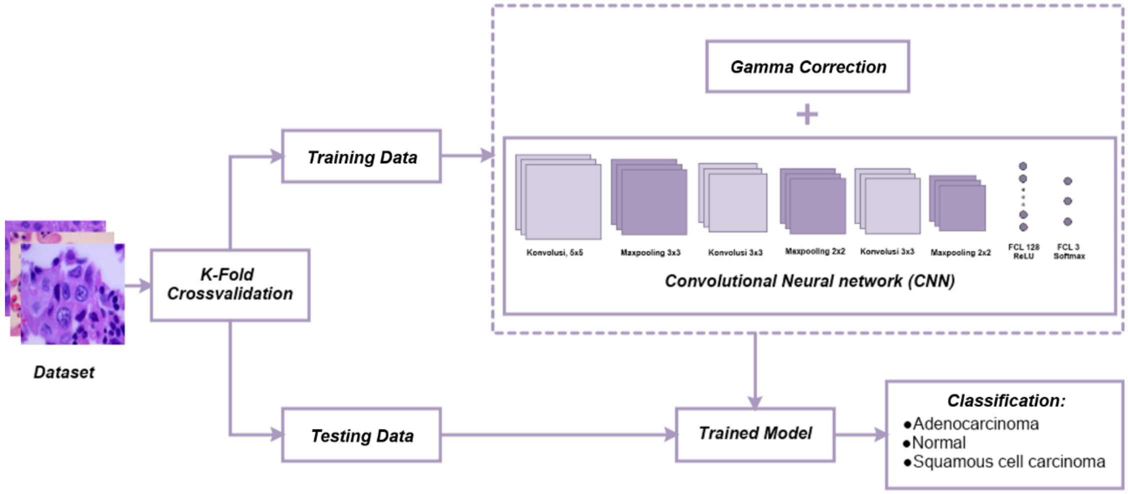


Figure 5. Proposed architecture system

The architecture system consists of the following parts:

1. Training data is the part of the dataset that trains the algorithm for building the model.
2. Testing data generated by the model when the training phase is complete.
3. Gamma correction is a preprocessing that is used to adjust the image light level.
4. CNN is an algorithm for feature extraction and classification
5. *k* fold cross validation. Steps that are useful for obtaining the optimal combination of training and testing data with *k*=5.
6. Trained Model. The trained model contains parameter information obtained after the training phase is complete, a trained model can already predict the class.
7. Classification is the output predicted by the model. There are three classes: adenocarcinoma, normal, and squamous cell carcinoma.

E. CNN Architecture

CNN architecture commonly has several convolutional layers, max-pooling, and activation functions. In this study, gamma correction was added before the first convolutional layer. The proposed CNN architecture is shown in Figure 6.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, 128, 128, 3)]     0
_________________________________________________________________
koreksi_gamma_1 (koreksi_gam (None, 128, 128, 3)       0
_________________________________________________________________
conv2d (Conv2D)              (None, 128, 128, 32)      2432
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 42, 42, 32)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 42, 42, 32)        9248
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 21, 21, 32)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 21, 21, 32)        9248
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 10, 10, 32)        0
_________________________________________________________________
flatten (Flatten)            (None, 3200)              0
_________________________________________________________________
dense (Dense)                (None, 256)               819456
_________________________________________________________________
dense_1 (Dense)              (None, 3)                 771
=================================================================
Total params: 841,155
Trainable params: 820,227
Non-trainable params: 20,928
```

Figure 6. Proposed CNN architecture

The architecture consists of:

1. The input layer has a size of 128×128×3 with RGB channels.
2. Gamma Correction Layer which adjusts the lighting of the image.
3. Convolution layer with 32 kernels measuring 5×5, zero padding, activation ReLU which produces 128×128×32 output.
4. Max-pooling with a size of 3x3 produces an output of 42×42×32.
5. Convolution layer with 32 kernels measuring 3×3, zero padding, activation ReLU which produces a 42×42×32 output.
6. Max-pooling with a size of 2×2 produces an output of 21×21×32.
7. Convolution layer with 32 kernels measuring 3×3, zero padding, activation ReLU which produces 21×21×32 output.
8. Max-pooling with a size of 2×2 produces an output of 10×10×32.

9. Flatten.
10. Fully Connected Layer, 256 nodes with ReLU activation.
11. Fully Connected Layer, 3 nodes with Softmax activation.

CNN architecture has a gamma correction layer, three convolution layers, three pooling layers, and two fully-connected layers. The loss function used is categorical cross-entropy. It handles multiclass classification with the specified optimization Stochastic Gradient Descent (SGD). The training data is taken from the five-fold cross-validation with a ratio of 8/10. The model that has been built is trained with training data to be able to recognize the parameters that match the characteristics of the data. The training process on the model uses 50 epochs, and the optimization uses SGD with learning rates of 0.01.

Furthermore, the testing phase. testing data is taken from the five-fold cross-validation that has been done previously with a ratio of 2/10. The trained model was used for testing data. results. The class prediction is obtained, then compared with the actual class to evaluate the performance of the model.

F. Material

The dataset used is LC25000. The data are histopathology of lung cancer images in JPEG format. It is from Academic Torrents (https://academictorrents.com/details/7a638ed187a6180fd6e464b3666a6ea0499af4af). The dataset released in 2019 was obtained from James A. Haley Veterans Hospital, Tampa, Florida, USA. The lung cancer dataset is divided into 3 classes: adenocarcinoma, normal, and squamous cell carcinoma. Experts have limited the classification process based on visual observations.

Adenocarcinoma shows that the lungs undergo a process of glandular formation which is characterized by swelling of the tissue. Normal indicates that the lung tissue is healthy. Squamous Cell Carcinoma (SCC) shows that the lungs undergo a process of keratinization which is characterized by the presence of polygonal-shaped cells.

The dataset, which consists of 3,000 images, is divided into five folds, where each fold in turn will be used as training data and testing data. The data sharing ratio is 80% for training data and 20% for testing data. There is a scenario where the gamma correction is used. It affects the intensity of image light. The scenario aims to find out which model has the best performance for classifying lung cancer.

3. Main Results

A. The Experiment Environment

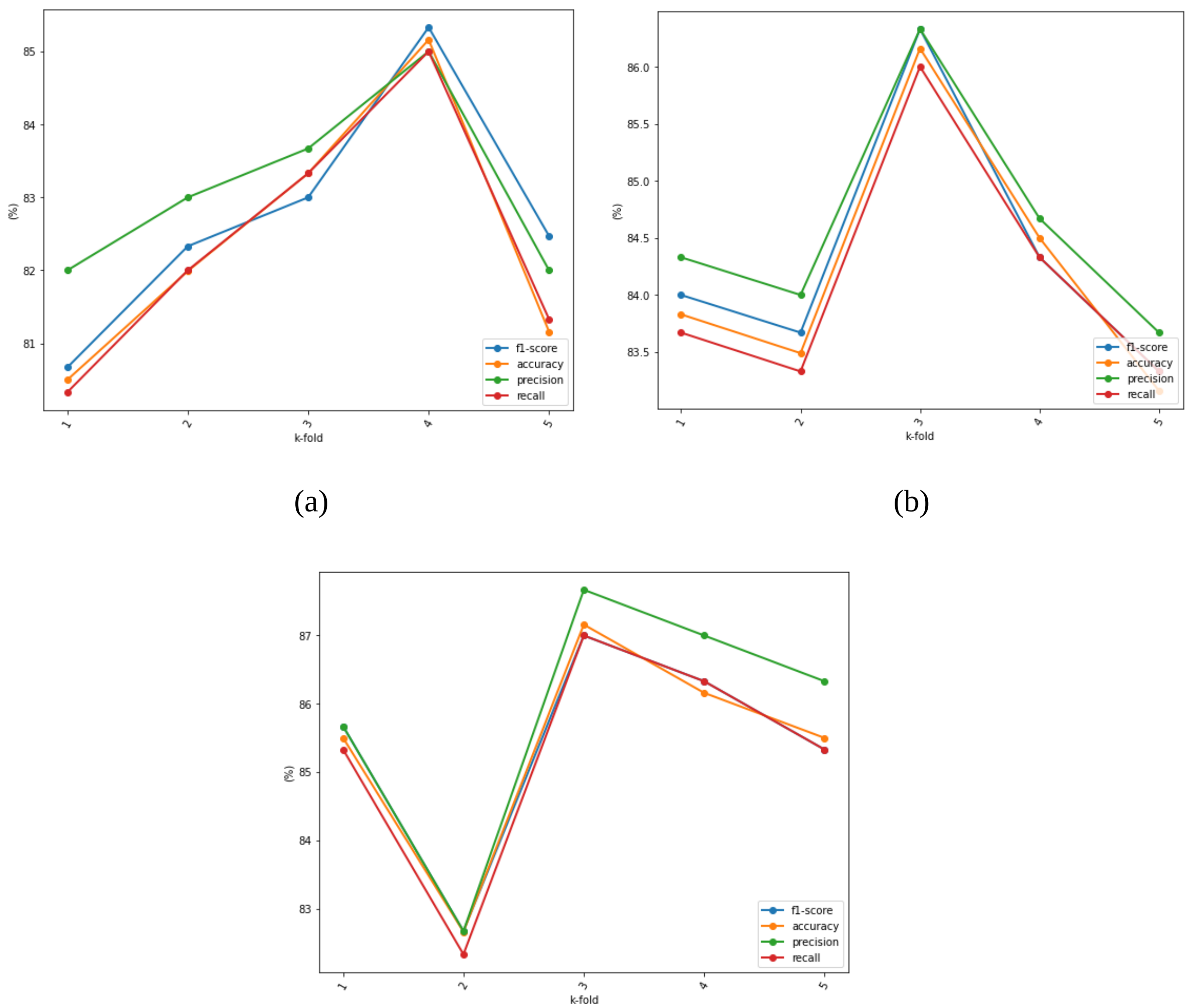
This section describes the device specifications used when testing the system, both device specifications, platforms, and python libraries. The hardware used is Intel i7-9750H, GPU GTX 1650 4GB, RAM 16 GB. Google colab platform specifications are accessed for free. This google collab platform is used for the training and testing process. Google colab is run online and has a daily usage limit for computing. The python library used is shown in Table 1.

Table 1. Python library

Phyton library	Description
os	access from system files and folder directories
NumPy	storing and processing array data
OpenCV	reading and managing images
matplotlib	displaying images
scikit-learn	share data and display confusion matrix
TensorFlow	model building
Keras	adding layers to the model
seaborn	plotting from array
Tkinter	build desktop view

B. The Experiment Results

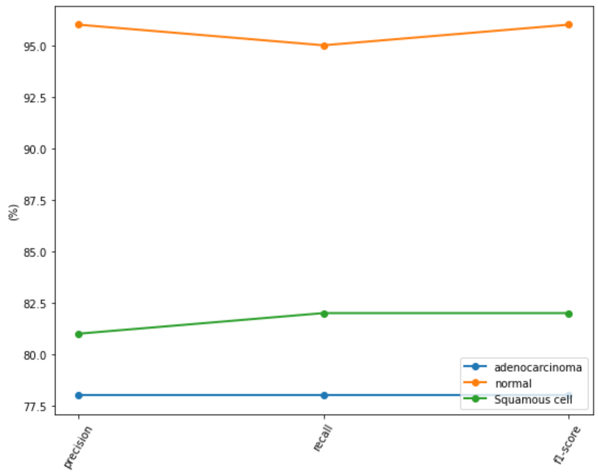
The experiment was about the effect of three values of gamma correction. It used the same configuration with epochs 50 and a learning rate of 0.01. Epoch is a network training cycle that uses all parameter data while learning rate is a parameter that affects the learning speed. The gamma values are 0.8, 1.0, and 1.2. It used five-fold cross-validation. The testing results of the three scenarios are shown in Figures 7 and 8.



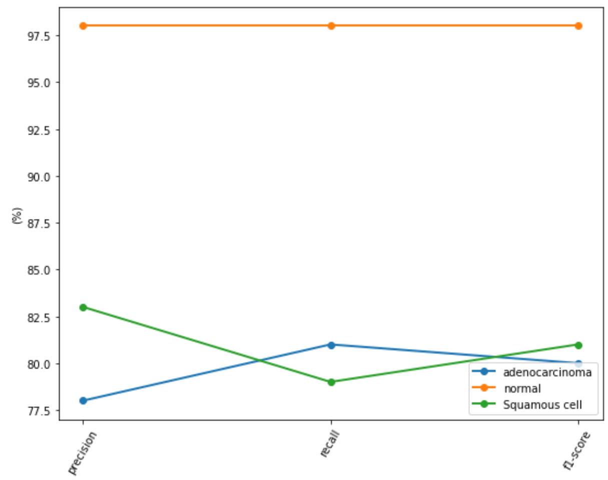
(a) (b)

(c)

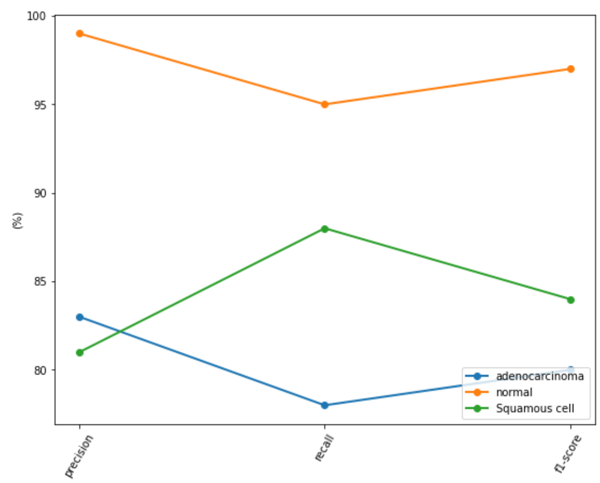
Figure 7. Performance measures (f1-score, accuracy, precision and recall) using gamma value (a) 0.8 (b) 1.0 and (c) 1.2



(a)



(b)



(c)

Figure 8. Best performance (precision, recall, f1-score) of each breast cancer class using gamma value (a) 0.8 (b) 1.0 and (c) 1.2

Figure 7 shows the testing results of three scenarios using different gamma values. The results show from the first scenario, that the best performance is in the fourth fold. While in the second and third scenarios, it is in the third fold. The performance measures consist of f1-score, accuracy, precision, and recall. In the first scenario, the fourth fold delivered the performance measures are 85.33%, 85.16%, 85%, and 85%, respectively. Meanwhile, for the second scenario, the best performance measures are 86.33%, 86.16%, 86.33%, and 86%. For the third scenario, it produced 87%, 87.16%, 87.67%, and 87% for performance measures. From the testing scenario with three gamma values, the best results are found in the third scenario. It can be concluded that the higher the gamma value, the better the performance obtained. Furthermore, Figure 8 shows the best performance of each breast cancer class based on each scenario.

Figure 8 shows the best precision, recall, and f1-score from each class (adenocarcinoma, normal and squamous cell carcinoma) based on three gamma values. The testing results show that the normal class can perform better recognition than the other two classes. Meanwhile, the lowest recognition was in the adenocarcinoma class.

Based on the testing scenario that has been carried out, the model obtains better results when the gamma correction process is carried out with a gamma value configuration of 1.2 with an accuracy of 87.16%, F1-score 87%, Precision 86.67%, and Recall 87%. The image processed using a gamma value of 1.2 looks clearer than before. While the gamma of 0.8 obtains the lowest accuracy when compared to other scenarios. This is because the increase in light intensity in the dataset results in a faded image and makes it more difficult for the system to recognize the image. In addition, the adenocarcinoma and squamous cell carcinoma classes have differences that are difficult to identify with the naked eye so the system obtains lower accuracy.

4. CONCLUSION

The testing has been carried out, classification using gamma correction and CNN on the classification of lung cancer based on histopathological images with 3 classes, with 5-fold Cross Validation. The data consists of 2,400 training data and 600 test data. It can be seen that the accuracy obtained is better if the image is preprocessed using gamma correction (gamma value = 1.2). Accuracy up to 87.16%. Histopathological images appear clearer if the gamma correction process is carried out. In addition, the Convolutional Neural Network has also been successfully used to identify features from the image for the classification process.

From the research that has been done, there are several suggestions which are presented as follows:

1. In the next research, devices with higher specifications can be used to handle the computational process.

2. A more complex CNN architecture can be used to get better performance, such as Alexnet, VGG, Resnet, or Inception.
3. Subsequent research can add the amount of data processed, it can support better performance.

ACKNOWLEDGEMENT

This research is funded by DRTPM Kemdikbudristek DIKTI for Penelitian Dasar Kompetitif Nasional No. 2220/UN46.4.1/PT.01.03/2022.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] K. Inamura, Lung cancer: understanding its molecular pathology and the 2015 WHO classification, Front. Oncol. 7 (2017), 193. https://doi.org/10.3389/fonc.2017.00193.

[2] N. Aliyah, E. Pranggono, B. Andriyoko, Kanker Paru: Sebuah Kajian Singkat, Indones. J. Chest Emerg. Med. 4 (2016), 28–32.

[3] The Global Cancer Observatory, Cancer Incident in Indonesia, 2020. [Online]. Available: https://gco.iarc.fr/today/data/factsheets/populations/360-indonesia-fact-sheets.pdf.

[4] J. Ferlay, M. Colombet, I. Soerjomataram, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods, Int. J. Cancer. 144 (2018), 1941–1953. https://doi.org/10.1002/ijc.31937.

[5] S. Anttila, P. Boffetta, eds., Occupational cancers, Springer, 2014. https://doi.org/10.1007/978-3-030-30766-0.

[6] H. Pan, Z. Xu, J. Huang, An effective approach for robust lung cancer cell detection, in: G. Wu, P. Coupé, Y. Zhan, B. Munsell, D. Rueckert (Eds.), Patch-Based Techniques in Medical Imaging, Springer International Publishing, Cham, 2015: pp. 87–94. https://doi.org/10.1007/978-3-319-28194-0_11.

[7] S. Rahman, M.M. Rahman, M. Abdullah-Al-Wadud, et al. An adaptive gamma correction for image enhancement, J. Image Video Proc. 2016 (2016), 35. https://doi.org/10.1186/s13640-016-0138-1.

[8] Y. Shin, I. Balasingham, Comparison of hand-craft feature based SVM and CNN based deep learning framework

for automatic polyp classification, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, Seogwipo, 2017: pp. 3277–3280. https://doi.org/10.1109/EMBC.2017.8037556.

[9] A. Sarmiento, I. Fondón, Automatic breast cancer grading of histological images based on colour and texture descriptors, in: A. Campilho, F. Karray, B. ter Haar Romeny (Eds.), Image Analysis and Recognition, Springer International Publishing, Cham, 2018: pp. 887–894. https://doi.org/10.1007/978-3-319-93000-8_101.

[10] S. Bhattacharjee, C.H. Kim, D. Prakash, et al. An efficient lightweight CNN and ensemble machine learning classification of prostate tissue using multilevel feature analysis, Appl. Sci. 10 (2020), 8013. https://doi.org/10.3390/app10228013.

[11] T. Huang, J. Li, C. Zhang, et al. Distinguishing lung adenocarcinoma from lung squamous cell carcinoma by two hypomethylated and three hypermethylated genes: a meta-analysis, PLoS ONE. 11 (2016), e0149088. https://doi.org/10.1371/journal.pone.0149088.

[12] C.H. Marquette, M.C. Copin, F. Wallet, et al. Diagnostic tests for pneumonia in ventilated patients: prospective evaluation of diagnostic accuracy using histology as a diagnostic gold standard, Amer. J. Respir. Critic. Care Med. 151 (1995), 1878–1888. https://doi.org/10.1164/ajrccm.151.6.7767535.

[13] B. S. Hernowo, Karsinoma Paru. Bandung Integrated Respiratory Care, 2012. http://pustaka.unpad.ac.id/wp-content/uploads/2017/01/KarsinomaParu.pdf.

[14] A. Hendrawan, Analisa Peningkatan Kualitas Citra Bawah Air Berbasis Koreksi Gamma Untuk Pencocokan Gambar Pada Algoritma SIFT, J. Transform. 12 (2014), 27-34. https://doi.org/10.26623/transformatika.v12i1.89.

[15] S. Albawi, T. A. M. Mohammed, S. Alzawi, Layers of a convolutional neural network, in ICET2017, 2017, pp. 1–6. https://wiki.tum.de/display/lfdv/Layers+of+a+Convolutional+Neural+Network.

[16] A. Hibatullah and I. Maliki, Penerapan Metode Convolutional Neural Network Pada Pengenalan Pola Citra Sandi Rumput, 2019. http://elibrary.unikom.ac.id/id/eprint/1529.

[17] J. Jin, A. Dundar, E. Culurciello, Flattened convolutional neural networks for feedforward acceleration, in: 3rd Int. Conf. Learn. Represent. ICLR 2015, 2015. https://doi.org/10.48550/arXiv.1412.5474.

[18] S. Mangal, A. Chaurasia, A. Khajanchi, Convolution Neural Networks for diagnosing colon and lung cancer histopathological images, (2020). http://arxiv.org/abs/2009.03878.

[19] A.F. Agarap, Deep learning using rectified linear units (ReLU), (2019). http://arxiv.org/abs/1803.08375.

[20] W. Liu, Y. Wen, Z. Yu, et al. Large-margin softmax loss for convolutional neural networks, (2017). http://arxiv.org/abs/1612.02295 (accessed August 11, 2022).

[21] P. Achlioptas Stanford, Stochastic gradient descent in theory and practice, Stanford, pp. 1–16, 2019.