AIP Conference Proceedings

Skins cancer identification system of HAM10000 skin cancer dataset using convolutional neural network

Cite as: AIP Conference Proceedings **2202**, 020039 (2019); https://doi.org/10.1063/1.5141652
Published Online: 27 December 2019

Ardan Adi Nugroho, Isnandar Slamet and Sugiyanto

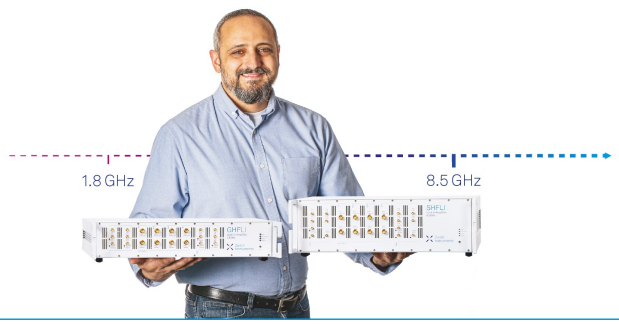
View Online Export Citation

ARTICLES YOU MAY BE INTERESTED IN

Early detection of currency crisis in Indonesia using real interest rate on deposits
AIP Conference Proceedings **2202**, 020041 (2019); https://doi.org/10.1063/1.5141654

Application of stochastic dominance on share return with normal distribution
AIP Conference Proceedings **2202**, 020045 (2019); https://doi.org/10.1063/1.5141658

Modeling the impact of limited treatment resources in the success of typhoid intervention
AIP Conference Proceedings **2202**, 020040 (2019); https://doi.org/10.1063/1.5141653



Trailblazers. New
Meet the Lock-in Amplifiers that measure microwaves.
Zurich Instruments
Find out more

© 2019 Author(s).

Skins Cancer Identification System Of HAM10000 Skin Cancer Dataset Using Convolutional Neural Network

Ardan Adi Nugroho [a)], Isnandar Slamet, Sugiyanto

Department of Statistics, Faculty of Mathematics and Natural Sciences, Sebelas Maret University, Indonesia

[a)]Corresponding author: ardanaran@student.uns.ac.id

Abstract. Cancer disease was one of the significant health problems in the world that cause deaths. Skin cancer is a type of cancer caused by damaged DNA and the body cannot recover the damage. This makes the cells begin to grow and divide uncontrollably. Skin cancer in Indonesia was ranks third after breast and neck cancer. Cases of skin cancer reach 5.9-7.8% of all types of cancer every year. This disease could be identified using the dermatoscopic image. Through those images, doctor decided the skin condition based on the characteristics that exist. Decision making in determining the skin condition depends on the ability and knowledge of the doctor. Therefore, a skin cancer identification system is needed to help make decisions in identifying skin cancer. Convolutional neural network (CNN) was used in this identification system because it can be detecting image and pattern. CNN work through three stages, that is convolutional layer, pooling layer, and fully-connected layer. This identification system is based on the dermoscopy image of HAM10000 skin cancer dataset. Based on this research, the accuracy of training and testing of skins cancer identification system are 80% and 78%.

INTRODUCTION

Skin cancer is the uncontrolled growth of abnormal cells in the skin. There are three most frequent types of skin cancer, basal cell, squamous cell and melanoma [1]. Non-melanoma (basal and squamous cell carcinoma) skin cancer (NMSC) are now the most common types of cancer. NMSCs constitute more than one-third of all cancer in the U.S. with an estimated incidence of over 600.000 cases per year [2]. In the other hand, melanoma is the most dangerous skin cancer, because it can cause death. One of the main problems that make melanoma can cause death is the ability of melanoma, which can spread to other organs such as the heart, liver, lungs, etc. [3]. The main cause of skin cancer is overexposure to UV radiation. This is produced by the sun, but it can also come from other sources, such as solariums (sunbeds). The UV radiation can penetrate the outer skin layer and the deeper layer of the skin, which can damage or kill skin cells, including DNA in the cells. DNA damage can cause a malfunction so that the DNA in the cell mutates. Finally, it can cause uncontrolled cell growth, transformation, and the development of skin cancer [4].

Skin diseases diagnosis is challenging. To diagnose a skin disease, a variety of visual clues may be used such as the individual lesional morphology, the body size distribution, color, scaling, and arrangement of lesions. Especially when each element are analyzed separately, the recognition process will be very complex [5]. There are few clinical methods to diagnose skin disease. To use these methods and achieve a satisfactory diagnostic accuracy, a high level of knowledge is required as the differentiation of skin lesions demands a great deal of experience and knowledge. This makes the results of the diagnosis become subjective and is not always reproducible. Computer with an identification system is more objective and reliable [6].

Convolutional neural network (CNN) was one of artificial intelligence that popular in feature learning and object classification. CNN using high performance of GPUs, it makes CNN capable of training a network on a large-scale dataset so the performance can be better. Many studies show that CNN are surpass human in many computer vision tasks. Therefore, we propose to construct a skins cancer identification system with CNNs [7].

However, the dataset of skin cancer is needed to build CNNs identification system. HAM10000 was used in this system to train the CNN. This dataset contains 10015 dermoscopy images of seven type of skin cancers, namely actinic keratosis, basal cell carcinoma, benign keratosis, dermatofibroma, melanocytic nevi, melanoma and vascular skin

International Conference on Science and Applied Science (ICSAS) 2019
AIP Conf. Proc. 2202, 020039-1–020039-6; https://doi.org/10.1063/1.5141652
Published by AIP Publishing. 978-0-7354-1953-7/$30.00

lesions [8]. Dermoscopy image show more detailed visualization of skin disorder, this can be used to evaluate the color and microscopic structure of the epidermis to papildermis that is not visible to the eye [9].

Based on the problem discussed above. This research design the Skins Cancer Identification System Using Convolutional Neural Network.

METHODS

This paper presents the CNN classifications for targeted disease. The disease is seven types of skin lesion that was existed in HAM10000 dataset, namely actinic keratosis (akiec), basal cell carcinoma (bcc), benign keratosis (bkl), dermatofibroma (df), melanocytic nevi (nv), melanoma (mel) and vascular skin lesions (vasc).

Dataset

This paper using HAM10000 dataset that contains 10015 dermoscopy image of seven skin lesion types. This dataset is deposited at Harvard Dataverse. Table 1 shows a summary of the number of images in the HAM10000 and Table shows the sample dermoscopy pictures used in this system.

TABLE 1. Summary of HAM10000 dataset

Dataset	Total Image	akiec	Bcc	bkl	df	mel	nv	vasc
HAM10000	10015	327	541	1099	155	1113	6705	142

Preprocessing Data

This stage shows the data preparation, this is needed to make the data ready for use as an input. First, the image size is equalized into 90 x 120 pixel. Then normalization is applied, this is done to modify values so that can measure them on the global scale.

The data must be divided into three parts, namely training data, validation data, and test data. This is needed to build and test the system. Sequentially each part of the data amounts to 7212, 2003 and 800 then data augmentation applied to the training data, this used to prevent overfitting and to get the better accuracy.

Building CNN model

Convolutional Layer

Convolutional layer performs a convolutional operation on the output of other function repeatedly. This layer applies the output function as a feature map from the image. Convolution is used to produce the extracted feature from the inputted image. The layer calculates the output of the neurons that connected to the local area in the input, the calculation used is dot product between their weights and the small area connected to the input. Kernel, stride and padding is used as the hyperparameter in this layer. The size of kernel and stride that used for each convolution layer are 3 x 3 and one. Furthermore, the type of padding used are same which means the output dimension is same as the input dimension.

Activation Function

The activation function that used is Rectified Linear Unit (ReLU). The ReLU will eliminate vanishing gradient by applying the activation function element as $f(x) = max(0, x)$ which mean element activation will be done when at the threshold of 0.

Pooling

Pooling is used to reduce the size of the representations and to speed up calculations, as well as to make some of the features it detects a bit more robust. This layer gives CNN the ability to recognize an object. Type of pooling that used is max pooling with the size of pooling layer is 2 x 2 for each layer.

Flatenning

The features map generated from the previous stage is still in the multidimensional form array, so we must to flattening or reshape the feature map. It is needed to convert feature map into a vector so we it can be used as input from the fully-connected layer.

Fully-Connected Layer

The fully-connected layer is a condition that all activation neuron from previous layers is all connected with a neuron in the next layers. This layer is used to transform the data dimension, so classification can be done.

Output

This output is the last layer in this system. This layer shows the type of the disease that used for the classification, namely akiec, bcc, bkl, df, nv, mel, and vasc.

The CNN architecture used in this system shown in Table 2. In the architecture, 90x120x3 images size was used as input. It means the images have size 90 pixels and 120 pixels with three channels, namely red, green, blue (RGB). This system uses four convolution layers with each convolutional layer uses a 3x3 filter size. Then pooling layer uses two layers that uses the max-pooling method.

TABLE 2. CNN architecture

Layer Type	Size	Output Shape
Input	(90,120,3)	
Convolution +ReLU	32(3 x 3) filter	(90,120, 32)
Convolution +ReLU	32(3 x 3) filter	(90, 120, 32)
Max Pooling+Droput	(2 x 2) filter	(45, 60, 32)
Convolution +ReLU	64(3 x 3) filter	(45, 60, 64)
Convolution +ReLU	64(3 x 3) filter	(45, 60, 64)
Max Pooling+Droput	(2 x 2) filter	(22, 30, 64)
Fully-Connected+ReLU+Dropout	256 neuron	7
Softmax	7 way	7

During the training process, changes are made to the parameters of the model to try and minimize the loss function and make predictions as accurately as possible. The optimizer is used to make that changes, it ties together the loss function and model parameters by updating the mode in response to the output of the loss function. In this system Adaptive Moment Estimation (Adam) optimizer is used, Adam optimizer works with momentums of first and second order.

RESULT AND DISCUSSION

Model Evaluation

After going through several processes in the CNN algorithm, the results of training and validation were obtained. This process uses a total of 50 epochs, a learning rate of 0.001. The plot of accuracy and loss over the epochs shown in Figure 1 and Figure 2.

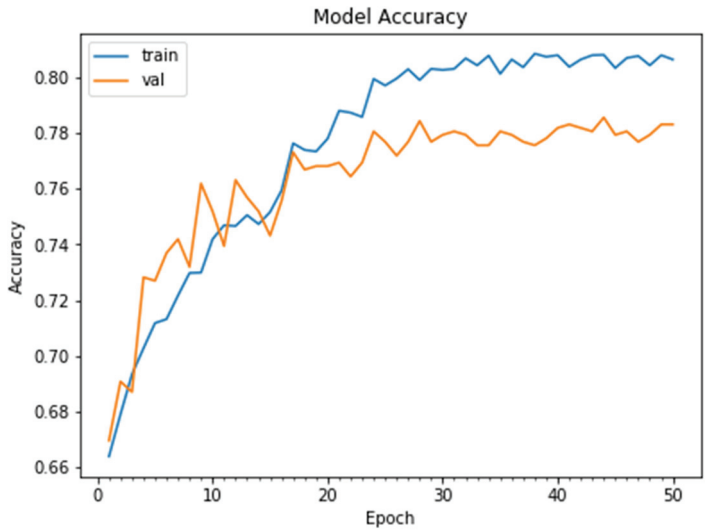


FIGURE 1. Model Accuracy

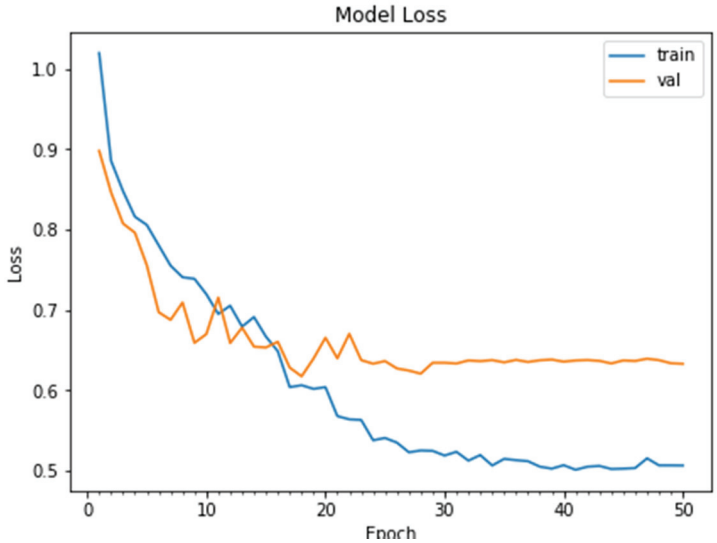


FIGURE 2. Model Loss

Based on Figure 1 and 2 the accuracy of the training set of the model reaches 80% with loss value of 50%. Then the accuracy of the validation set of the model reaches 78% with loss value of 63%. It means the system is expected to perform with 78% accuracy on new data.

Testing Data

This section is discussed about the test of the model, there are confusion matrix and test result that show the number of correct classification and the wrong classification. This classification was used 800 new data from HAM10000 that not used to build the model. The confusion matrix is shown in Table 3 and the result summary is shown in Table 4.

TABLE 3. Confusion matrix

Metrics		Prediction Class						
		Nv (0)	**Mel (1)**	**Bkl (2)**	**Bcc (3)**	**Akiec (4)**	**Vasc(5)**	**Df (6)**
Act Class	Nv (0)	12	10	4	0	2	1	0
	Mel (1)	0	32	4	0	8	0	0
	Bkl (2)	6	4	45	0	23	2	0
	Bcc (3)	1	2	0	1	1	0	0
	Akiec (4)	2	5	23	0	505	10	1
	Vasc (5)	2	1	17	0	41	21	0
	Df (6)	0	3	0	0	0	1	10

TABLE 4. Test result

Disease	Data	True	False
Nv (0)	29	12	17
Mel (1)	44	32	12
Bkl (2)	80	45	35
Bcc (3)	5	1	4
Akiec (4)	546	505	41
Vasc (5)	82	21	61
Df (6)	14	10	4

Based on Table 3 and Table 4 above, the results of the prediction of the model on the testing data of the data show good results. Depend on the result of confusion matrix, the accuracy of the classification can be calculated. So the accuracy produced by the model with 90x120 pixel image input, the learning rate value of 0.001 obtained the accuracy value of 78%. Besides that, Figure 3 shows the fraction that has incorrectly classification. Depend on that Figure, bcc has the highest incorrectly classification and akiec has the lowest classification incorrectly.

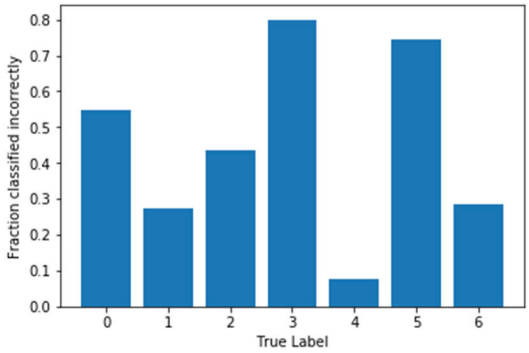


FIGURE 3. Fraction Classified Incorrectly

CONCLUSIONS

The result of CNN model that use input shape with size 90x120 pixel, Adam optimizer, learning rate 0.001 and number of epochs 50 show that the level of training accuracy is 80% and 78% for testing accuracy. Bcc disease is the most difficult to identified by the system and akiec disease is the most likely identity by the system.

REFERENCES

[1] Wagner, A. J., Berliner, N. and Benz, E. J., 2017. Anatomy and Physiology of the Gene, Hematology: Basic Principles and Practice, 63, pp. 3–16.

[2] Jensen, A. O., Lamberg, A. L. and Olesen, A. B., 2010. Epidemiology of non-melanoma skin cancer. Non-Surgical Treatment of Keratinocyte Skin Cancer, 146, pp. 15-24.

[3] Folberg, R., McLean, I. W. and Zimmerman, L. E., 1984. Conjunctival Melanosis and Melanoma. Ophthalmology, 91(6), pp. 673–678.
[4] Qadir, M. I., 2016. Skin cancer: Etiology and management. Pakistan Journal of Pharmaceutical Sciences, 29(3), pp. 999–1003.
[5] Liao, H., Li, Y. and Luo, J. 2017 Skin disease classification versus skin lesion characterization: Achieving robust diagnosis using multi-label deep neural networks. Proceedings - International Conference on Pattern Recognition, pp. 355–360.
[6] Whited, J. D. and Grichnik, J. M., 1998. Does this patient have a mole or a melanoma?. Journal of the American Medical Association, 279(9), pp. 696–701.
[7] Shen, Z. Y. et al., 2019. Deep convolution neural network with scene-centric and object-centric information for object detection. Image and Vision Computing. Elsevier B.V., 85, pp. 14–25.
[8] Tschandl, P., Rosendahl, C. and Kittler, H., 2018. Data descriptor: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data, 5, pp. 1–9.
[9] Kittler, H. et al., 2002. Diagnostic accuracy of dermoscopy. Lancet Oncology, 3(3), pp. 159–165.