p-ISSN 2301–4156 | e-ISSN 2460–5719

# Deep Transfer Learning to Improve Classification Accuracy in Dermoscopic Images of Skin Cancer

**Qorry Aina Fitroh[1], Shofwatul 'Uyun[2]**

[1,2] Program Studi Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Sunan Kalijaga, Yogyakarta 55281 INDONESIA (tel.: 0274-512474; fax: 0274-586117; email: [1]qorryafitroh@gmail.com, [2]shofwatul.uyun@uin-suka.ac.id)

[Received: 27 December 2022, Revised: 2 March 2023]
Corresponding Author: Qorry Aina Fitroh

**ABSTRACT** — Benign and malignant cancers are the most common skin cancer types. It is essential to know skin cancer symptoms with an early diagnosis to provide an appropriate treatment and reduce the mortality rate. Dermoscopic image is one of the diagnostic media that many researchers have developed. It provides more optimal results in computational-based diagnosis than visual detection. Deep learning and transfer learning are two models that have been used successfully in computational-based analysis, although optimization is still needed. In this study, transfer learning was used to separate dermoscopic images of skin cancer into two categories: benign and malignant. This study used 2,000 images to increase previous research's accuracy conducted on the Kaggle public dataset containing 3,297 images. Two pretrained models, namely VGG-16 and residual network (ResNet)-50, were compared and used as feature extractors. Fine-tuning was conducted by adding a flatten layer, two dense layers with the ReLU activation function, and one dense layer with the Softmax activation function to classify images into two categories. Hyperparameter tuning on the optimizer, batch size, learning rate, and epoch were performed to get each model's best performance parameter combination. Before hyperparameter tuning, the model was tested by resizing the input image using different sizes. The results of model testing on the VGG-16 model gave the best performance at an image size of 128 × 128 pixels with a combination of Adam parameters as an optimizer, batch size of 64, learning rate of 0.001, and epoch of 10 with an accuracy value of 91% and loss of 0.2712. The ResNet-50 model yielded better accuracy of 94% and a loss of 0.2198 using the optimizer parameter RMSprop, batch size of 64, learning rate of 0.0001, and epoch of 100. The results indicate that the proposed method provides good accuracies and can assist dermatologist in the early detection of skin cancer.

**KEYWORDS** — Transfer Learning, Deep Learning, CNN, VGG-16, ResNet-50, Skin Cancer, Classification.

## I. INTRODUCTION

Cancer is one of the causes of death. According to WHO data in 2020, nearly 10 million deaths were caused by this disease [1]. Various types of cancer can spread throughout the human body and skin cancer is one of the fastest-growing cancers which can cause death [2]. This type of cancer appears on the outermost layer of skin, namely the epidermis, which is visible to the naked eyes [3]. Malignant and benign cancers are the deadliest types of skin cancer [4]. Malignant cancer can spread and develop in the patient's body by infiltrating other tissues and organs and uncontrollably growing and spreading. In comparison, benign cancer can grow but does not spread. However, it is crucial to understand the common signs and symptoms of benign cancer turning into malignant. Early diagnosis at the onset of symptoms can help sufferers have a greater chance of survival [5]. The initial stage of skin cancer diagnosis is done through a visual examination, followed by a biopsy and histopathological examination [6]. Another diagnostic technique currently being developed is based on computational analysis of dermoscopic images for automatic classification of skin lesions that can assist doctors in detecting disease and making diagnostic decisions, as well as enabling fast and inexpensive access [7], [8].

Dermoscopic image is the most popular skin imaging technique. This technique has proven more accurate than a direct visual examination by medical personnel since it allows visualization of various features that are unable to be directly seen by the eyes, enabling the structure beneath the surface of the skin lesion can be analyzed and different types of lesions can be distinguished using better visualization [9], [10]. The use of dermoscopic images has improved skin cancer diagnosis compared to direct visual examination. Research classified skin cancer images using two different dermoscopic image datasets, namely HAM10000 with 10,000 images and BCN20000 with 19,424 images [3]. The research achieved an accuracy of 94.92%. The same dataset (HAM10000) was used and yielded an accuracy of 91.43% [2]. Dermoscopic images of skin cancer with a total of 4,000 images were used in a classification using convolutional neural network (CNN) with VGG-16 and performing hyperparameter tuning [11]. Using this method, an accuracy of 99.7% was obtained. The 6,594 public datasets from Kaggle were trained using the VGG-16 model to generate an accuracy of 93.18% [4]. Research by comparing the number of two different datasets, namely 10,000 and 3,297 images for classifying skin cancer, has shown that using a dataset with a larger number of images does not significantly improve classification accuracy [6].

Science and technological advancement supporting the detection of skin cancer based on image analysis has shown a continuous enhancement [12]. In certain skin cancer classifications, skin cancer image analysis using classical machine learning algorithms can yield good results. However, this algorithm is ineffective for more complex diagnostic cases in clinical practices as the process of feature extraction and feature selection requires a long time [13]. In addition, errors and data loss at the processing stage can significantly affect the classification quality [7], [14]. Various challenges from classical machine learning can be overcome using the deep learning method since this method has demonstrated significant performances that can analyze data from large datasets more

quickly and accurately [8], [13], [15]. CNN is a deep learning method that provides significant results with good accuracy in image recognition [16], [17]. It was utilized to detect melanoma dermoscopic images by obtaining an average sensitivity and specificity of 86.6% and 71.3%, respectively [18]. The classification of melanoma dermoscopic images using CNN yielded an average sensitivity value of 89.4% and an average specificity of 68.2% [19].

Research to detect breast cancer with CNN was implemented through three scenarios: training CNN from scratch (full training), using pretrained CNN for feature extraction, and fine-tuning the pretrained CNN model [20]. The best accuracy was obtained from the second and third methods, with the fine-tuning model's accuracy being only 0.008 higher than that of the feature-extraction model. This study has demonstrated that the application of transfer learning on CNN is a promising solution in a breast cancer detection. Transfer learning is proven to increase accuracy and accelerate the training process [15]. Transfer learning was used to classify human facial images with the Inception-v3 pretrained model to obtain an average accuracy of 93% [21]. In addition, image classification into two classes was carried out using deep CNN with the pretrained VGG-16 model and fine-tuning and image augmentation to obtain an accuracy of 95.40% [22]. Meanwhile, pretrained CNN models, namely VGG-16, VGG-19, and residual network (ResNet)-50, were compared with two scenarios (training from scratch and transfer learning) [23]. The best accuracy was 92.60% with the transfer learning scenario using VGG-16, and fine-tuning was performed. Transfer learning with various pretrained models was compared in classifying white blood cells [24] and skin cancer [25]. Both demonstrated the best performance when ResNet-50 was used as a pretrained model.

This study optimizes the classification of skin cancer into two categories (benign and malignant) through transfer learning and fine-tuning the optimizer, batch size, learning rate, and epoch. The pretrained models used in this study were VGG-16 and ResNet-50. The dataset was acquired from the Kaggle database as dermoscopic images of skin cancer.

This research optimizes the classification of skin cancer into two categories (benign and malignant) through the use of transfer learning and fine-tuning the optimizer, batch size, learning rate, and epoch. The pretrained models used in this study were VGG-16 and ResNet-50. The dataset was acquired from the Kaggle database in the form of dermoscopic images of skin cancer.

## II. METHODS

This study employed two pretained CNN models, namely VGG-16 and ResNet-50. Both models were implemented in the same dataset, which consisted of 2,000 skin cancer images. As for the technique, deep transfer learning was applied to classify dermoscopic images of skin cancer into two categories, namely benign and malignant. The image processing flow is shown in Figure 1. Beginning with image input, the process continued with image preprocessing through data augmentation. Then, feature extraction was performed on the two pretrained models, which were then modified by fine-tuning the fully connected layers. To determine the best hyperparameters in both models, training was carried out on the training data, followed by testing the hyperparameters against the testing data to obtain the best

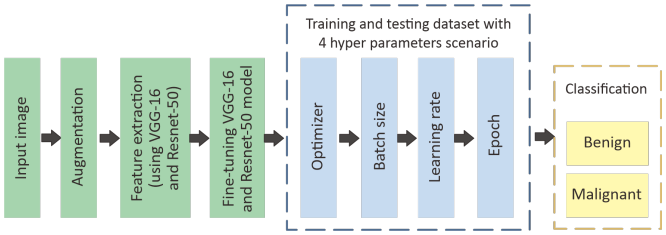


**Figure 1.** Image processing flow.

TABLE I
DATASET DESCRIPTION

| | Training | Validation | Testing | Total |
|---|---|---|---|---|
| Benign | 640 | 160 | 200 | 1,000 |
| Malignant | 640 | 160 | 200 | 1,000 |
| Total | 1,280 | 320 | 400 | 2,000 |

accuracy. The hyperparameter test was conducted using four scenarios with the first scenario being the optimizer. After getting the best type of optimizer for each model, then it was applied to the second scenario, namely batch size. The best batch size value was used for the subsequent test scenario, namely the learning rate. The last scenario was the epoch value test.

### A. DATASET

This research used a public dataset derived from the Kaggle database. The dataset was RGB images comprising 1,800 benign images and 1,497 malignant images with an image size of 224 × 224 × 3 pixels. The total number of images in the dataset was 3,297, but only 2,000 images were used in this study. Table I describes the dataset used in this study. The dataset was divided into training data and test data with a ratio of 80:20. Furthermore, 20% of the training data was used for validation. In contrast, the remaining data were used for model training.

### B. DATA PREPROCESSING

The preprocessing stage included image resizing and image augmentation. Images were resized and included in preprocessing scenarios to determine the optimal performance. In image resizing, the image size shifted from the original 224 × 224 pixels to 128 × 128 pixels and 64 × 64 pixels. Image augmentation aims to enlarge training data that is done artificially. The augmentations performed included rotation, zoom, horizontal flip, and translation.

### C. FEATURE EXTRACTION

The transfer learning employed a pretrained CNN model as a feature extractor. Furthermore, the extracted features were used for training on a new layer [26]. The pretrained model was trained on a large dataset, namely ImageNet, consisting of various image categories so that the model learned the image representation well, and then this feature could be used for knowledge transfer and as a feature extractor on different datasets [22]. When using a pretrained model as a feature extractor, the lower layer is frozen, and the upper layer is customized. This study used two different pretrained models to classify skin cancer images, namely VGG-16 and ResNet-50.

#### 1) VGG-16

Visual geometry group (VGG) is an architecture proposed by Simonyan and Zisserman at the University of Oxford and won the ILSVRC 2014 competition. It was trained using the ImageNet ILSVRC dataset of 1.3 million data with 1,000

p-ISSN 2301–4156 | e-ISSN 2460–5719

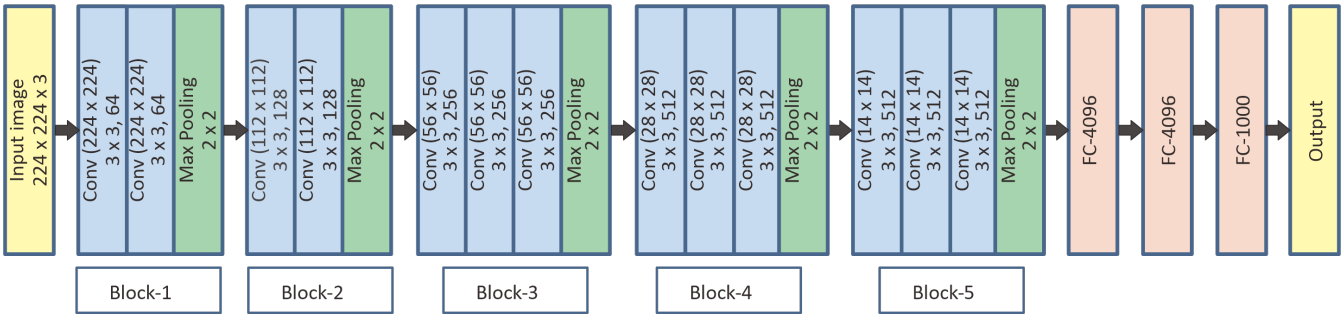


**Figure 2.** VGG-16 architecture.

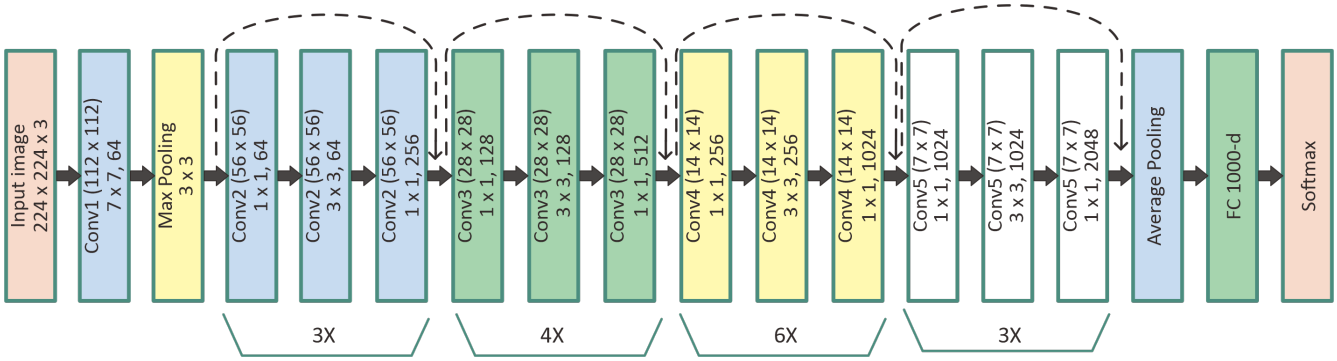


**Figure 3.** ResNet-50 architecture.

classes [27]. This architecture is recognized as one of the most advanced architecture models ever made, so the pretrained VGG-16 model is an excellent option for transfer learning [4], [28]. Figure 2 depicts the VGG-16 architecture, which consists of five blocks. The first and second blocks each contain two convolution layers and max pooling with filters of 64 and 128, respectively. The third to fifth blocks contain one max pooling layer and three convolution layers with filters of 256, 512, and 512, respectively. In this study, VGG-16 was modified by adding one flatten layer, two dense layers with the ReLU activation function, and one dense layer with the Softmax activation function to improve model accuracy.

2) RESNET-50

Microsoft Research Asia (MRA) in 2015 proposed a deep neural network model with a very deep network called ResNet, whose aim is to ensure that the upper layer's performance is as good as that of the layer below it without any vanishing gradients and optimization problems [26]. ResNet won the 2015 ILSVRC competition with the architecture on ResNet-50, which is a ResNet consisting of 50 layers. As the name implies, this network uses residual learning. In simple terms, residuals are a process of reducing features that have been learned from the input layer [24]. The ResNet-50 architecture, as depicted in Figure 3, consists of a series of convolution blocks with average pooling. There are five convolution layer blocks. After being inputted, the image passed through the convolution layer with 64 filters and a kernel size of 7 × 7 (layer conv1) followed by max pooling. Additionally, in conv2, layers were grouped in pairs according to how ResNet operated. This process was repeated until the fifth convolution layer, after which the average pooling was performed. Softmax was used in the last layer for classification.

### *D. FINE-TUNING SCENARIO*

Fine-tuning of the VGG-16 and ResNet-50 models is shown in Figure 4. The upper layers of the models, both VGG-16 and ResNet-50, which consisted of the fully connected layer and

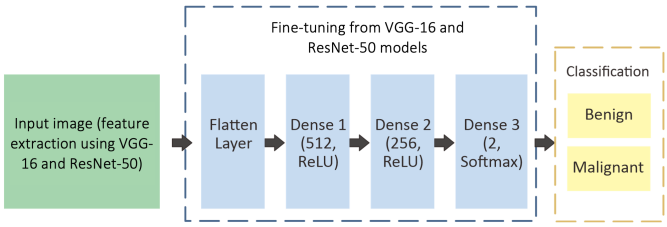


**Figure 4.** Fine-tuning of VGG-16 and ResNet-50 models.

the classification layer, were eliminated and replaced with two dense layers with activation functions ReLU, each layer consisting of 512 and 256 neurons. The third dense layer consisted of two neurons with the Softmax activation function. The images were then classified into two different cancer categories: benign and malignant. The feature extraction results in the pretrained model were then used as input for fine-tuning. This scenario applied to the VGG-16 and ResNet-50 models. In addition, hyperparameter tuning (which included the optimizer scenario, batch size, learning rate, and epoch), was performed to achieve the best performance.

The Softmax activation function is frequently used for multiclass classification. This function appears in almost all output layers of deep neural network architectures [29]. The Softmax activation function itself is a generalization of Sigmoid. The fundamental difference between Sigmoid and Softmax is that Sigmoid can only be used for binary classification, whereas Softmax can also be used for multivariate classification. Using the region-based CNN (R-CNN) method, the Softmax and Sigmoid activation functions were used to do the binary classification of face detection on comic book characters [30]. The classification with the Sigmoid activation function showed slightly better performance than Softmax, although the results were not significantly different. Meanwhile, Softmax and Sigmoid activation functions applied to the CNN model for binary classification yielded scores of 97.3% and 70%, respectively [31].

TABLE II
PARAMETER OF THE PROPOSED VGG-16 MODEL

| Layer | Shape of Output | Parameters |
|---|---|---|
| VGG-16 | 4, 4, 512 | 14,714,688 |
| Flatten | 8,192 | 0 |
| Dense_1 | 512 | 4,194,816 |
| ReLU_1 | 512 | 0 |
| Dense_2 | 256 | 131,328 |
| ReLU_2 | 256 | 0 |
| Dense_3 | 2 | 514 |
| Total parameters | | 19,041,346 |
| Trainable parameters | | 4,326,658 |
| Nontrainable parameters | | 14,714,688 |

TABLE III
PARAMETER OF THE PROPOSED RESNET-50 MODEL

| Layer | Shape of Output | Parameters |
|---|---|---|
| ResNet-50 | 7, 7, 2048 | 23,587,712 |
| Flatten | 100,352 | 0 |
| Dense_1 | 512 | 51,380,736 |
| ReLU_1 | 512 | 0 |
| Dense_2 | 256 | 131,328 |
| ReLU_2 | 256 | 0 |
| Dense_3 | 2 | 514 |
| Total parameters | | 75,100,290 |
| Trainable parameters | | 51,512,578 |
| Nontrainable parameters | | 23,587,712 |

Research by comparing the activation functions of Softmax and Sigmoid to do the binary classification and Softmax for multiclass using five types of neural networks, namely separable CNN (SCNN), recurrent neural network (RNN), long short-term memory (LSTM), gated recurrent unit (GRU), and fully connected network (FCN) by applying three scenarios, revealed that Softmax achieved the highest classification accuracy in conducting multiclass classification [32]. In contrast, for the binary classification, SCNN obtained the best accuracy in both activation functions. In addition, for the five types of neural networks applied, the Softmax activation function obtained greater accuracy than the Sigmoid. Therefore, using the Softmax activation function for binary classification, which is considered a multiclass classification with two classes, can enhance the accuracy of the five types of neural networks. For that reason, this study proposed applying the Softmax activation function at the output layer to classify cancer images into two categories: benign and malignant.

### E. OPERATION EVALUATION

Using the confusion matrix, the level of success was determined by measuring the performance of the experiments that were carried out. The results of the classification process were represented through four terms in the confusion matrix, namely true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The performance evaluation was calculated based on the TP, TN, FP, and FN values, from which the accuracy, precision, sensitivity, and f1-score of the model being tested could be obtained. The performance evaluation accuracy measures the closeness of the predicted values to the actual values. The accuracy value calculation was conducted using (1). The amount of data that are classified as TP compared to the total data in the positive classification category is called the precision and is calculated using (2). Sensitivity or recall compares the number of positive data classified as TP to the total positive data and is shown in (3). The f1-score value is a combination of precision and sensitivity, which can be obtained using (4).

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

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

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

$$F1-Score = \frac{2\times TP}{2\times(TP+FP+FN)}. \quad (4)$$

## III. RESULT AND DISCUSSION

The experimental process involving fine-tuning was carried out by dividing the dataset as described in Table I. The images used for training had the largest proportion, with 640 images in each class, to ensure that the model learned the representation pattern properly. To test the best hyperparameters, the training was carried out on the training data. Furthermore, the hyperparameters that were applied to the training data were used in testing with test data.

Before performing hyperparameter tuning, an image resizing test was carried out to obtain the best image size of the VGG-16 and ResNet-50 models to achieve optimal accuracy. Testing was conducted using predetermined hyperparameters, namely the Adam optimizer, batch size of 64, learning rate of 0.001, and epoch of 10. The image sizes tested were 64 × 64, 128 × 128, and 224 × 224 pixels. Based on the results of the image resizing test, the best accuracy for the VGG-16 model was at an image size of 128 × 128 pixels with an accuracy level of 0.9100. On the other hand, the best size of the ResNet-50 was at an image size of 224 × 224 pixels with an accuracy value of 0.9050. Comparatively, the VGG-16 model with an image size of 224 × 224 pixels yielded a lower accuracy value of 0.8850, while the ResNet-50 model was tested with an image size of 128 × 128 pixels and yielded an accuracy value of 0.8975. For an image size of 64 × 64 pixels, both VGG-16 and ResNet-50 produced the lowest accuracy among the other image sizes tested: 0.8875 for VGG-16 and 0.8875 for ResNet-50. Based on the results of image size testing, it can be said that the accuracy of VGG-16 is better than the ResNet-50 with various predetermined factors.

Table II shows the parameters of the proposed VGG-16 model after fine-tuning. The output image size after the VGG-16 model with an input image size of 128 × 128 pixels was 4 × 4, and the resulting number of parameters was 14,714,468. After that, a flatten layer with an output size of 8.192 was added. Lastly, Dense_1, Dense_2, and Dense_3 was added and yielded 4,194,816; 131,328; and 514 parameters. The parameters generated by ResNet-50 with an input image size of 224 × 224 pixels were significantly greater than those generated by VGG-16. The parameters of the proposed ResNet-50 model are displayed in Table III. The size of the image after feature extraction was 7 × 7 with 23,587,712 parameters. Using the same additional layers as the VGG-16 model, parameters produced by ResNet-50 for each dense layer were 51,380,736; 131,328; and 514.

After testing the image size and performing fine-tuning, hyperparameter testing with four scenarios was conducted, namely optimizers (Adam, SGD, and RMSprop), batch sizes (32, 64, and 128), learning rates (0.01, 0.001, and 0.0001), and

JURNAL NASIONAL TEKNIK ELEKTRO DAN TEKNOLOGI INFORMASI
p-ISSN 2301–4156 | e-ISSN 2460–5719

TABLE IV
RESULTS OF SCENARIO TESTING ON THE VGG-16 DAN RESNET-50 MODELS

| Model | Scenario 1: Optimizer (batch size of 64, learning rate of 0.001, epoch of 10) | | | Scenario 2: Batch Size (optimizer: Adam–VGG-16; RMSProp–ResNet-50, learning rate of 0.001, epoch of 10) | | |
|---|---|---|---|---|---|---|
| | **Optimizer** | **Test Accuracy** | **Test Loss** | **Batch Size** | **Test Accuracy** | **Test Loss** |
| VGG-16 | Adam | 0.9100 | 0.2712 | 32 | 0.8975 | 0.3029 |
| | SGD | 0.8875 | 0.2933 | 64 | 0.9100 | 0.2712 |
| | RMSprop | 0.8875 | 0.2448 | 128 | 0.8900 | 0.2931 |
| ResNet-50 | Adam | 0.9050 | 0.2548 | 32 | 0.8825 | 0.2785 |
| | SGD | 0.8925 | 0.2629 | 64 | 0.9075 | 0.2420 |
| | RMSprop | 0.9075 | 0.2420 | 128 | 0.8725 | 0.3031 |
| | Scenario 3: Learning Rate (optimizer: Adam–VGG-16; RMSProp–ResNet-50, batch size of 64, epoch of 10) | | | Scenario 4: Epoch (optimizer: Adam–VGG-16; RMSProp–ResNet-50, batch size of 64, learning rate: 0.001–VGG-16; 0.0001–ResNet-50) | | |
| | **Learning rate** | **Test Accuracy** | **Test Loss** | **Epochs** | **Test Accuracy** | **Test Loss** |
| VGG-16 | 0.01 | 0.8975 | 0.2811 | 10 | 0.9100 | 0.2712 |
| | 0.001 | 0.9100 | 0.2712 | 50 | 0.9025 | 0.2261 |
| | 0.0001 | 0.8700 | 0.2982 | 100 | 0.8975 | 0.2641 |
| ResNet-50 | 0.01 | 0.8925 | 0.2494 | 10 | 0.9150 | 0.2342 |
| | 0.001 | 0.9075 | 0.2420 | 50 | 0.9075 | 0.2255 |
| | 0.0001 | 0.9150 | 0.2342 | 100 | 0.9400 | 0.2198 |

epochs (10, 50, and 100). The first scenario involved an optimizer with other predetermined hyperparameters: the batch size of 64, the learning rate of 0.001, and the epoch of 10. Table IV summarizes the results of testing the VGG-16 and ResNet-50 models in the four scenarios. The first scenario on the VGG-16 Adam optimizer model had better performance than SGD and RMSprop on the VGG-16 model. It differs from the ResNet-50 model, which obtained the best performance, which was equal to 0.9075, when using the RMSprop optimizer.

The results of testing the second scenario, namely the batch size with a different optimizer, revealed that the performance of the two models did not improve from the first scenario. It indicates that the best accuracy in both models is from the same batch size as determined at the beginning of the test, namely 64. The following scenario was learning rate testing, with the learning rate values tested being 0.01, 0.001, and 0.0001. This scenario resulted in different learning rate values for the two models. The ResNet-50 model obtained better accuracy than the VGG-16 model, with a learning rate of 0.0001 and an accuracy value of 0.9150. In contrast, the accuracy of the VGG-16 model remained unchanged from the previous scenario, with a learning rate of 0.001 and an accuracy of 0.91100.

After testing the optimizer, batch size, and learning rate, tests were carried out with the last scenario, namely the epoch test. The fourth scenario tested the epoch values of 10, 50, and 100. As shown in Table IV, the results derived from the two models are different. Based on the epoch test, it can be concluded that the effect of the epoch on the VGG-16 model is that as the epoch value increases, the model's accuracy decreases, with the best accuracy being achieved at epoch 10. It differs from the ResNet-50, which had the best performance at epoch 100, with an accuracy value of 0.9400 and loss of 0.2198. The accuracy of the ResNet-50 model obtained was more optimal than previous research, which applied the same image size, batch size, and epoch to obtain an accuracy of 0.87 [33].

The hyperparameters with the best performance for each model were obtained after running the tests with the previously specified scenarios. For the VGG-16 model, the best

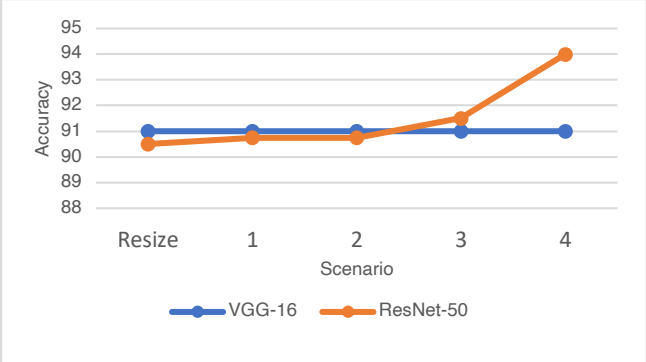


**Figure 5.** Graph of the best accuracy in each scenario.

performance was provided by an image of 128 × 128 pixels with the Adam optimizer, the batch size of 64, the learning rate of 0.001, and epoch of 10. The results of testing accuracy and loss using training data and validation with the best hyperparameters on VGG- 16 revealed that the difference in the accuracy of the training and validation data was insignificant. Meanwhile, the loss graphs can be interpreted that the losses from the training and validation data not overfitting because, at the end of the epoch, the loss graphs hardly make a significant difference. For the ResNet-50 model, the best performance was an image of 224 × 224 pixels. The hyperparameters with the best combination for the ResNet-50 model were the RMSprop optimizer, the batch of size 64, the learning rate of 0.0001, and epoch of 100. The test results on training and validation data on the validation accuracy graph for each epoch were relatively stable. The loss graph, on the other hand, is considered a good fit because it does not show a significant difference between the training data and the validation.

After the two models were trained using training and validation data, tests were carried out on the new dataset, namely test data, which comprised 400 images. Testing the model using the test dataset yielded values used to measure the model performance. Using the VGG-16 model, the number of benign images correctly classified as the benign category was 177. Meanwhile, there were 187 malignant images that were

TABLE V
COMPARISON OF PROPOSED STATE-OF-THE-ART MODELS

| Techniques/Models | Dataset | Accuracy (%) |
|---|---|---|
| SVM and KNN [34] | Kaggle | 70.61 |
| Random Forest [35] | Kaggle | 84.20 |
| CNN [33] | Kaggle | 87.87 |
| Transfer Learning VGG-16 [6] | Kaggle dan HAM10000 | 88.00 |
| Transfer Learning VGG-16 [12] | Kaggle | 89.09 |
| Transfer Learning VGG-16 | Kaggle | 91.00 |
| Transfer Learning ResNet-50 | Kaggle | 94.00 |

correctly classified as a malignant category. The number of benign images that the model incorrectly predicted as malignant was 23. Conversely, the number of malignant images that the model incorrectly predicted as benign category was 13. Not much different from the VGG-16 model, the ResNet-50 model correctly classified 185 benign images and 191 malignant images. Then, there were 15 benign images that were incorrectly identified as malignant category and 9 malignant images that were incorrectly identified as benign category. Thus, most test images can be classified according to their categories.

Performance was also evaluated based on the precision, sensitivity, and f1-score. Using (2), (3), and (4), the average value of precision, sensitivity, and f1-score for VGG-16 was 0.91. It demonstrates that the proposed VGG-16 model can provide good performance in classifying dermoscopic images of skin cancer. Meanwhile, it is evident that the ResNet-50 model could classify images better than the VGG-16 model. Performance measurement of the proposed ResNet-50 model showed that the average value of precision, sensitivity, and f1-score generated by the ResNet-50 model was 0.94.

The best results for each scenario that has been tested are modeled in a graph in Figure 5. After implementing the resize scenario, the following scenarios are implemented: scenario 1, namely optimizer; scenario 2, namely batch size; scenario 3, namely learning rate; and scenario 4, namely epoch. As can be seen in Figure 5, the accuracy of the VGG-16 model does not increase in any scenario. It suggests that the resize scenario with an image size of 128 × 128 pixels and the hyperparameter tuning that was determined at the beginning of the tests yields the best results. It is different from the ResNet-50 model, which almost always experienced an increase in accuracy in every given scenario, except for scenario 1 (optimizer) and scenario 2 (batch size), where the optimizer and batch size both had the same best accuracy value. Therefore, the last scenario in the ResNet-50 model was the scenario that yielded optimal accuracy. Meanwhile, for the VGG-16 model, the best accuracy was obtained since the resize scenario.

A comparison of the performance results of the proposed model with other models is presented in Table V. Based on the accuracy value, the proposed model provided better accuracy than other models using the same dataset, namely skin cancer dermoscopic images from Kaggle classifying images into two categories (benign and malignant categories). Using the same model, namely VGG-16, this study demonstrated better performance based on the accuracy results, which was 91%, compared to previous research with 89.09% [12]. The same datasets, namely Kaggle and the HAM10000 datasets, that were tested using pretrained VGG-16 and ResNet-50 models yielded a not-significantly-different accuracy, namely 88% for the VGG-16 model and 87% for the ResNet-50 model [6]. Despite the fact that these accuracy results were from two different datasets, the accuracies yielded were alike. It means that both the proposed VGG-16 and ResNet-50 models provide better performance. In addition, applying the ResNet-50 model and the same fine-tuning scenario, except for Dense_3, in which the Sigmoid activation function was used, generated an accuracy of 87.87% at epoch 150 [33]. The ResNet-50 model proposed in this study applied a smaller epoch, namely 100, and attained the best performance, namely 94%, among other models listed in Table V.

## IV. CONCLUSION

This study employed deep transfer learning techniques with pretrained deep CNN VGG-16 and ResNet-50. The training, validation, and test data used were derived from the Kaggle dataset. The model was tested using various hyperparameters, namely the optimizer consisting of Adam, SGD, and RMSprop. In addition, the model was also tested with batch size parameters of 32, 64, and 128; learning rates of 0.01, 0.001, and 0.0001; and epochs of 10, 50, and 100. The proposed model generated the best performance, with 91% accuracy for the VGG-16 model and 94% for the ResNet-50 model. There are still many gaps in improving accuracy in classifying skin cancer images. Using different pretrained models, fine-tuning scenarios, and other hyperparameter tuning does not rule out the possibility of increasing the model's accuracy in classifying skin cancer images.

## CONFLICT OF INTEREST

The authors declare no conflict of interest is found in paper entitled "Deep Transfer Learning to Improve Classification Accuracy in Dermoscopic Images of Skin Cancer."

## ACKNOWLEDGMENT

Thanks to the Directorate General of Islamic Education, the Ministry of Religious Affairs of the Republic of Indonesia, and all parties who have supported the continuation of this research.

## REFERENCES

[1] WHO (2022) "Cancer," [Online], https://www.who.int/news-room/fact-sheets/detail/cancer, access date: 7-Jun-2022.

[2] M.S. Ali *et al.*, "An Enhanced Technique of Skin Cancer Classification Using Deep Convolutional Neural Network with Transfer Learning Models," *Mach. Learn. Appl.*, Vol. 5, pp. 1–8, Sep. 2021, doi: 10.1016/j.mlwa.2021.100036.

[3] M.A. Kassem, K.M. Hosny, dan M.M. Fouad, "Skin Lesions Classification into Eight Classes for ISIC 2019 Using Deep Convolutional Neural Network and Transfer Learning," *IEEE Access*, Vol. 8, pp. 114822–114832, Jun. 2020, doi: 10.1109/ACCESS.2020.3003890.

[4] M.R. Hasan *et al.*, "Comparative Analysis of Skin Cancer (Benign vs. Malignant) Detection Using Convolutional Neural Networks," *J. Healthc. Eng.*, Vol. 2021, pp. 1–17, 2021, doi: 10.1155/2021/5895156.

[5] P.P. Naik, "Cutaneous Malignant Melanoma: A Review of Early Diagnosis and Management," *World J. Oncol.*, Vol. 12, No. 1, pp. 7–19, Feb. 2021, doi: 10.14740/wjon1349.

[6] S. Bechelli dan J. Delhommelle, "Machine Learning and Deep Learning Algorithms for Skin Cancer Classification from Dermoscopic Images," *Bioeng.*, Vol. 9, No. 3, pp. 1–18, Mar. 2022, doi: 10.3390/bioengineering9030097.

[7] T.J. Brinker *et al.*, "Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review," *J. Med. Internet Res.*, Vol. 20, No. 10, pp. 1–8, Oct. 2018, doi: 10.2196/11936.

[8] P.M.M. Pereira *et al.*, "Skin Lesion Classification Enhancement Using Border-Line Features – The Melanoma vs Nevus Problem," *Biomed.*

*Signal Process. Control*, Vol. 57, pp. 1-8, Mar. 2020, doi: 10.1016/j.bspc.2019.101765.

[9] G. Argenziano *et al.*, "Dermoscopy Improves Accuracy of Primary Care Physicians to Triage Lesions Suggestive of Skin Cancer," *J. Clin. Oncol.*, Vol. 24, No. 12, pp. 1877–1882, Apr. 2006, doi: 10.1200/JCO.2005.05.0864.

[10] S.M. Jaisakthi, P. Mirunalini, dan C. Aravindan, "Automated Skin Lesion Segmentation Of Dermoscopic Images Using Grabcut And K-Means," *IET Comput. Vis.*, Vol. 12, No. 8, pp. 1088–1095, Dec. 2018, doi: 10.1049/iet-cvi.2018.5289.

[11] R. Agustina, R. Magdalena, dan N.K.C. Pratiwi, "Klasifikasi Kanker Kulit Menggunakan Metode Convolutional Neural Network dengan Arsitektur VGG-16," *ELKOMIKA J. Tek. Energi Elektr. Tek. Telekomun. Tek. Elektron.*, Vol. 10, No. 2, pp. 446–457, Apr. 2022, doi: 10.26760/elkomika.v10i2.446.

[12] V. Anand *et al.*, "An Enhanced Transfer Learning Based Classification for Diagnosis of Skin Cancer," *Diagnostics*, Vol. 12, No. 7, Jul. 2022, doi: 10.3390/diagnostics12071628.

[13] Y. Wu *et al.*, "Skin Cancer Classification With Deep Learning: A Systematic Review," *Front. Oncol.*, Vol. 12, pp. 1–20, Jul. 2022, doi: 10.3389/fonc.2022.893972.

[14] M.A. Kassem, K.M. Hosny, R. Damaševičius, dan M.M. Eltoukhy, "Machine learning and Deep Learning Methods for Skin Lesion Classification and Diagnosis: A Systematic Review," *Diagnostics*, Vol. 11, No. 8, pp. 1–29, Jul. 2021, doi: 10.3390/diagnostics11081390.

[15] A. Saber *et al.*, "A Novel Deep-Learning Model for Automatic Detection and Classification of Breast Cancer Using the Transfer-Learning Technique," *IEEE Access*, Vol. 9, pp. 71194–71209, May 2021, doi: 10.1109/ACCESS.2021.3079204.

[16] R. Refianti, A.B. Mutiara, dan R.P. Priyandini, "Classification of Melanoma Skin Cancer Using Convolutional Neural Network," *Int. J. Adv. Comput. Sci., Appl.*, Vol. 10, No. 3, pp. 409–417, Mar. 2019, doi: 10.14569/IJACSA.2019.0100353.

[17] S. Saidah, I.P.Y.N. Suparta, and E. Suhartono, "Modifikasi Convolutional Neural Network Arsitektur GoogLeNet dengan Dull Razor Filtering untuk Klasifikasi Kanker Kulit," *J. Nas. Tek. Elekt., Teknol. Inf.*, Vol. 11, No. 2, pp. 148–153, May 2022, doi: 10.22146/jnteti.v11i2.2739.

[18] H.A. Haenssle *et al.*, "Man Against Machine: Diagnostic Performance of a Deep Learning Convolutional Neural Network for Dermoscopic Melanoma Recognition in Comparison to 58 Dermatologists," *Ann. Oncol.*, Vol. 29, No. 8, pp. 1836–1842, Aug. 2018, doi: 10.1093/annonc/mdy166.

[19] T.J. Brinker *et al.*, "A Convolutional Neural Network Trained with Dermoscopic Images Performed on Par With 145 Dermatologists in a Clinical Melanoma Image Classification Task," *Eur. J. Cancer*, Vol. 111, pp. 148–154, Apr. 2019, doi: 10.1016/j.ejca.2019.02.005.

[20] S. Guan and M. Loew, "Breast Cancer Detection Using Transfer Learning In Convolutional Neural Networks," *2017 Appl. Imag. Pattern Recognit. Workshop (AIPR)*, 2017, pp. 1–8, doi: 10.1109/AIPR.2017.8457948.

[21] M. Hussain, J.J. Bird, dan D.R. Faria, "A Study on CNN Transfer Learning for Image Classification," in *Advances in Computational Intelligence Systems*, A. Lotfi *et al.*, Eds., Cham, Switzerland: Springer Cham, 2019, pp. 191–202, doi: 10.1007/978-3-319-97982-3_16.

[22] S. Tammina, "Transfer Learning Using VGG-16 with Deep Convolutional Neural Network for Classifying Images," *Int. J. Sci., Res. Publ.*, Vol. 9, No. 10, pp. 143–150, Oct. 2019, doi: 10.29322/ijsrp.9.10.2019.p9420.

[23] Shallu dan R. Mehra, "Breast Cancer Histology Images Classification: Training from Scratch or Transfer Learning?" *ICT Express*, Vol. 4, No. 4, pp. 247–254, Dec. 2018, doi: 10.1016/j.icte.2018.10.007.

[24] D.M. Wonohadidjojo, "Perbandingan Convolutional Neural Network pada Transfer Learning Method untuk Mengklasifikasikan Sel Darah Putih," *Ultimics J. Tek. Inform.*, Vol. 13, No. 1, pp. 51–57, Jun. 2021, doi: 10.31937/ti.v13i1.2040.

[25] Z. Aliyah, A. Arifianto, and F. Sthevanie, "Classifying Skin Cancer in Digital Images Using Convolutional Neural Network with Augmentation," *Indonesia J. Comput. (Indo-JC)*, Vol. 5, No. 2, pp. 55–66, Sep. 2020, doi: 10.21108/indojc.2020.5.2.455.

[26] A. Abubakar, M. Ajuji, and I.U. Yahya, "Comparison of Deep Transfer Learning Techniques in Human Skin Burns Discrimination," *Appl. Syst. Innov.*, Vol. 3, No. 2, pp. 1–15, Apr. 2020, doi: 10.3390/asi3020020.

[27] A.V. Ikechukwu, S. Murali, R. Deepu, and R. C. Shivamurthy, "ResNet-50 vs VGG-19 vs Training from Scratch: A Comparative Analysis of the Segmentation and Classification of Pneumonia from Chest X-Ray Images," *Glob. Transit. Proc.*, Vol. 2, No. 2, pp. 375–381, Nov. 2021, doi: 10.1016/j.gltp.2021.08.027.

[28] R.A. Khan, A. Crenn, A. Meyer, and S. Bouakaz, "A Novel Database of Children's Spontaneous Facial Expressions (LIRIS-CSE)," *Image, Vis. Comput.*, Vol. 83–84, pp. 61–69, Mar.–Apr. 2019, doi: 10.1016/j.imavis.2019.02.004.

[29] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation Functions: Comparison of Trends in Practice and Research for Deep Learning," 2018, *arXiv*:1811.03378.

[30] X. Qin *et al.*, "A Faster R-CNN Based Method for Comic Characters Face Detection," *2017 14th IAPR Int. Conf. Doc. Anal., Recogn. (ICDAR)*, 2017, pp. 1074–1080, doi: 10.1109/ICDAR.2017.178.

[31] M. Korpusik and J. Glass, "Convolutional Neural Networks and Multitask Strategies for Semantic Mapping of Natural Language Input to a Structured Database," *2018 IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*, 2018, pp. 6174–6178, doi: 10.1109/ICASSP.2018.8461769.

[32] E. Solovyeva and A. Abdullah, "Binary and Multiclass Text Classification by Means of Separable Convolutional Neural Network," *Inventions*, Vol. 6, No. 4, Oct. 2021, doi: 10.3390/inventions6040070.

[33] K. Manasa and D.G.V. Murthy, "Skin Cancer Detection Using VGG-16," *Eur. J. Mol., Clin. Med.*, Vol. 8, No. 1, pp. 1419–1426, Jan. 2021.

[34] M. Faruk and N. Nafi'iyah, "Klasifikasi Kanker Kulit Berdasarkan Fitur Tekstur, Fitur Warna Citra Menggunakan SVM dan KNN," *Telematika*, Vol. 13, No. 2, pp. 100–109, Aug. 2020, doi: 10.35671/telematika.v13i2.987.

[35] N. Khasanah *et al.*, "Skin Cancer Classification Using Random Forest Algorithm," *J. Ilm. Sisfotenika*, Vol. 11, No. 2, pp. 137–147, Jul. 2021, doi: 10.30700/jst.v11i2.1122.