Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
16/1 (2023), 47-57. DOI: http://dx.doi.org/10.21609/jiki.v16i1.1100

Face Spoofing Detection using Inception-v3 on RGB Modal and Depth Modal

Yuni Arti[1,2], Aniati Murni Arymurthy[1]

[1]Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
[2]National Research and Innovation Agency (BRIN), Tangerang Selatan, Indonesia

E-mail: yuni.arti@ui.ac.id, aniati@cs.ui.ac.id

Abstract

Face spoofing can provide inaccurate face verification results in the face recognition system. Deep learning has been widely used to solve face spoofing problems. In face spoofing detection, it is unnecessary to use the entire network layer to represent the difference between real and spoof features. This study detects face spoofing by cutting the Inception-v3 network and utilizing RGB modal, depth, and fusion approaches. The results showed that face spoofing detection has a good performance on the RGB and fusion models. Both models have better performance than the depth model because RGB modal can represent the difference between real and spoof features, and RGB modal dominate the fusion model. The RGB model has accuracy, precision, recall, F1-score, and AUC values obtained respectively 98.78%, 99.22%, 99.31.2%, 99.27%, and 0.9997 while the fusion model is 98.5%, 99.31%, 98.88%. 99.09%, and 0.9995, respectively. Our proposed method with cutting the Inception-v3 network to mixed6 successfully outperforms the previous study with accuracy up to 100% using the MSU MFSD benchmark dataset.

Keywords: *face spoofing, real, spoof, Inception-v3, depth*

1. Introduction

iOTENTIK is a Government Certification Authority (Penyelenggara Sertifikat Elektronik, PSrE) that issue digital certificates to Government Employees (Aparatur Sipil Negara, ASN). One of the uses of digital certificates is for digital signatures. iOTENTIK has a certificate issuance application. The application has been able to verify registration remotely using facial biometric to handle ASN who spread throughout Indonesia. However, some problems were found when verifying a face, e.g., a photo position, the structural components (beard, glasses, and mask), the dark photos, and fraud (face spoofing). These problems made the system to be unable to verify correctly, with the percentage of face similarity being less accurate.

Each problem in face verification has different handling. The main problem in face verification and related to security issues is face spoofing. Some cases of face spoofing are found in certificate issuance application, such as photos are not taken directly, but using printed photos. The application has not been able to distinguish between real or spoof for the photo, so the percentage of face similarity for this problem is actually high.

Generally, spoof attacks are divided into three attacks, i.e., photo attacks, video attacks, and 3D attacks. Photo attacks use photos of genuine user faces. These photos are captured, printed and displayed on device screen. The video attacks replay the genuine user video during the authentication process. 3D attacks use a genuine user face mask that has similar shape and characteristics with the real face. Most of the face spoofing studies use photo and video attacks because those attacks often occured in the field.

Several issues related to face spoofing are encountered, i.e., generalization of unseen data. Most of the models are trained and tested with the same data spoof attack, so the ability to generalize unseen data is lacking. Another issue is how the system can distinguish spoof images by utilizing the information contained in the image (spatial, temporal, noise, and depth) and transfer learning that can improve the model performance in face spoofing detection. The deep learning approach has been widely used to solve those issues. Generally, that study uses single modal (RGB), such as [1], [2], [3], and [4]. Some studies use depth information, such as [5], [6], and [7]. The studies proved the effectiveness of depth information in face spoofing detection.

One of the studies that use a multi-modal dataset (RGB, depth and Infrared/IR) is Zhang et al. [8]. The study provided good accuracy and generalization by utilizing multi-modal. The True Positive Rate (TPR) and Average Classification Error Rate (ACER) values that are generated using multi-modal RGB, depth, and IR are 96.7% and 2.4%, respectively, while from multi-modal RGB and depth 86.1% and 5.0%. This result is better than using only a single modal.

Nagpal & Dubey [4] studied face spoofing detection using several state-of-the-art CNN architectures, such as Inception-v3, ResNet50 and ResNet152. Then the performance of each model is compared. Inception-v3 gives good result in face spoofing detection when transfer learning is carried out at the fully connected layer. This study uses the entire Inception-v3 layer without cutting the network and produces training, validation, and testing accuracy of 94.63%, 96.47%, and 96.13%, respectively.

Inception-v3 is the pre-trained model that is quite popular and one of the state-of-the-art models in the image classification task. Szegezy et al. [9] proposed some Inception architectures, one of which is Inception-v3. Inception-v3 is a combination of improvements from Inception-v2 with some advantages, such as factorization into smaller convolutions and asymmetric convolutions to reduce the number of parameters without reducing network efficiency, auxiliary classifier as a regularizer, and efficient grid size reduction so that it can save computation.

The previous research [4] did not include information about how far the CNN architecture can work well on face spoofing detection problems. Seeing the success of Inception-v3 from Nagpal & Dubey [4] in handling face spoofing detection and some of the advantages of Inception-v3, this study focuses on transfer learning by cutting the Inception-v3 network to get the best feature representation that can distinguish spoof and real images. Inspired by the use of depth information in face spoofing detection [5], [6], [7], and [8] as well as the use of multi-modal, and fusion approach [8], this study uses RGB modal and depth modal, with fusion approach to generalize the data well in face spoofing detection. The fusion approach used is feature fusion, which combines RGB features and depth features. Face spoofing is defined as a binary classification problem in this study, determining whether a face image is real or spoof.

The main contributions of this study are:

a. Detects face spoofing using the RGB modal, depth modal, and fusion approach to obtain good generalizations. The depth modal used is taken from RGB modal using the dense Depth model.
b. Transfer learning using pre-trained model Inception-v3 by cutting the network until mixed5, mixed6, and mixed7 layer to get the right features that can classify face images well.
c. Transfer learning on the RGB modal, the depth modal, and fusion of both modals using the right features to get the best model for face spoofing detection.

2. Related Works

There have many studies on face spoofing using deep learning. Generally, these studies use a single modal (RGB modal) and different image information. Liu *et al.* [1] proposed a method Zero Shot Face Anti-Spoofing (ZSFA) with the concept of Deep Tree Network (DTN) to learn the homogeneous features in the early tree node until different features are obtained on each tree node. The model is trained and tested using 13 types of spoof attacks to recognize various kinds of spoof attacks. With DTN, the model can detect unknown attacks well.

Yang et al. [2] proposed a method that considers global temporal and local spatial information. The proposed method is Spatio-Temporal Anti-Spoof Network (STASN), which combines the LSTM and CNN methods. The method can distinguish spoof faces from a variety of clues, such as borders, moire patterns, reflection artifacts, etc.

Jourabloo et al. [3] studied face spoofing detection using noise information on spoof faces. The proposed architecture uses a modified CNN architecture. The architecture that was built succeeded in visualizing the spoof noise contained in the spoof images. In the spoof image, the noise is visible, while in the live image there is no noise.

Nagpal & Dubey [4] studied face spoofing detection using RGB information with Inception-v3, ResNet-50, and ResNet-152. That study proposed face spoofing detection using transfer learning at the last fully connected layer with a lower learning rate.

Liu et al. [5] proposed a novel two-stream CNN-based approach for the face spoofing detection. The first CNN stream is used to extract the face region into patches and the second stream is used to estimate face depth from a full face image. Both results are combined to obtain the spoof score. A face image or video clip is classified as a spoof if the spoof score is above a pre-defined threshold. The use of extracted patches combined with depth information provides a promising approach, which can distinguish the spoof from live faces.

Liu et al. [6] proposed a method that considers depth information with pixel-wise supervision and Remote Photoplethysmography (rPPG) signal with sequence-wise supervision. The face depth and rPPG signal are combined to distinguish live and spoof faces. Using auxiliary supervision that combines image depth information and rPPG signal provides good performance.

Wang et al. [7] proposed a method to estimate depth estimation from multiple RGB frames and a depth-supervised architecture that can encode spatiotemporal information for face spoofing detection. The proposed method can get spoof patterns accurately and efficiently under depth-supervision.

The CASIA SURF dataset is a multi-modal dataset for face spoofing detection. One of the studies that use a multi-modal dataset is Zhang et al. [8]. Zhang et al. proposed ResNet-18 as the backbone and Squeeze and Excitation as the fusion module to select the more informative features in each modal and combined them into multi-modal features.

The latest face spoofing research uses one-class learning [10], wherein handling various attacks, the model only uses live face images. The image is trained with two generators to produce latent features representing various real face properties in the embedding space and generate spoofing cues. Furthermore, Feature Correlation Network (FCN) determines whether the inputted latent features represent live characteristics or not. This method completes face spoofing detection of various attacks using only one-class learning.

3. Methodology

3.1 Dataset

This study uses a dataset from Kaggle, i.e., ID R&D facial anti-spoofing challenge dataset from Biometric Technology Provider ID R&D, USA (https://www.kaggle.com/boksman/spoof-raw). The dataset consists of real and spoof faces, which spoof faces are taken from printed photos or videos. Image size varies with the smallest 76x76 pixels and the largest 1387x1387 pixels.

Table 1. Dataset.

Class	Training	Validation	Testing	Total
Real	1282	76	238	1596
Spoof	6160	390	1158	7708
Total	7442	466	1396	9304

The dataset is divided into three, i.e., 80% training data, 5% validation data, and 15% testing data, as shown in Table 1. Example of real faces and spoof faces are shown in Fig. 1. The dataset used only provides RGB images, as shown in Fig. 1 (top). The spoof face image is an image is not taken directly, but from a printed photo or video replay so that the image is blurred or not sharp.

A depth illustration of the RGB images is shown in Fig. 1 (bottom). The depth images are obtained from the conversion of RGB images to depth using the dense depth model [11]. The depth images appear to have the same color but there are differences in dark and light color. Spoof faces will have a flat depth so that all pixels will have the same depth to the image plane. All parts of the face are dark, as shown in Fig. 1 (right bottom). Real faces will have different depth for each part of the face. As shown in Fig. 1 (left bottom), the eyes and mouth can be distinguished by lighter color.

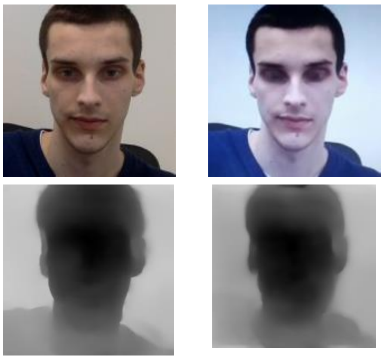

Fig. 1. Face images: real (left) and spoof (right); RGB (top) and depth (bottom).

3.2 Method

Face anti-spoofing provides good accuracy and generalization in multi-modal datasets (RGB, depth, IR) [8]. Spatial information, such as depth, is known to distinguish whether a face is real or spoof. The spoof face will have a flat depth so that all pixels will have the same depth as the camera. Unlike the spoof face, the real face will have different depths, e.g., the nose is closer to the camera than the cheek when facing the camera [6]. Therefore, this study uses RGB and depth modal, then uses a feature fusion approach to obtain a good generalization.

Face spoofing detection is conducted on RGB and depth modal with the same stages and methods. The process flow of RGB and depth model is shown in Fig. 2. Modal is the term that shows the information of the image, and it is used as input of the process, while the model is the term that shows the output of the process. RGB modal shows image that contain Red, Green, and Blue color information, while depth modal shows image that contain pixel depth information from RGB image. The depth of the real face will have a different depth, while the depth of the spoof face will have the same depth.

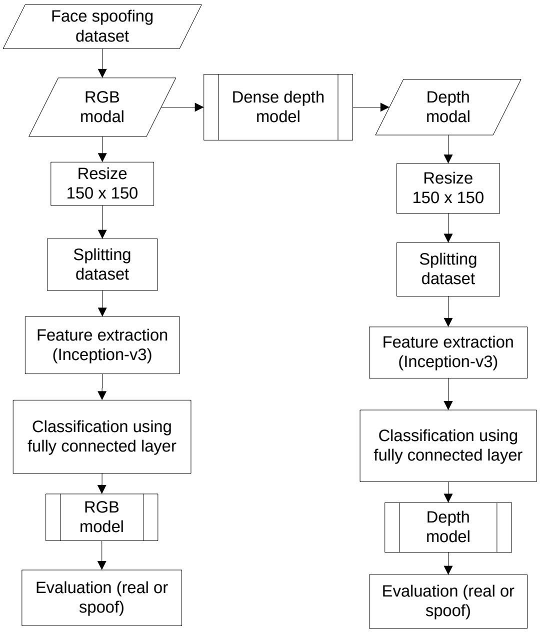
Face spoofing dataset
RGB modal
Dense depth model
Depth modal
Resize 150 x 150
Resize 150 x 150
Splitting dataset
Splitting dataset
Feature extraction (Inception-v3)
Feature extraction (Inception-v3)
Classification using fully connected layer
Classification using fully connected layer
RGB model
Depth model
Evaluation (real or spoof)
Evaluation (real or spoof)

Fig. 2. The flowchart of RGB model and depth model.

Depth modal is obtained from the RGB modal using a dense depth model [11]. The dense depth model was trained using the Mannequin Challenge dataset. The model can predict accurate and dense depth, which shows high level of details and sharp depth discontinuities (changes in pixel intensity). The predicted depth can describe an accurate sequence of depths for objects contained in the image. Therefore, this study uses the depth estimation from the study of Li et al. [11]. The resulting depth can help distinguish the pixel depth of the real and spoof face.

Dense depth model process [11] is shown as Fig. 3. The input consists of an RGB, a binary mask of human and initial depth of environment (i.e., non-human regions). The output is a dense depth map that includes the people and the environment.

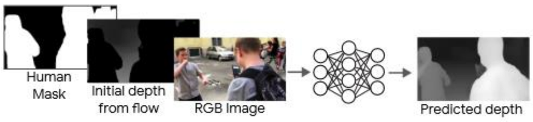

Fig. 3. Dense depth model process [11].

The architecture used of dense depth model is a variant of the hourglass network of [12], as shown in Fig. 4. The architecture consists of a set of convolutions (a variant of the Inception module) and downsampling, followed by a set of convolutions and upsampling blocks. There are skip connections that add back features from high resolutions. The upsampling layer used is bilinear upsampling layer.

As seen in Fig. 2, the image in each modal is resized to 150x150 pixels so the computation is not heavy. Furthermore, the dataset is divided into training, validation, and testing data, as listed in Table 1. After that, feature extraction is conducted for the three datasets using Inception-v3 [9]. Then classification process uses a fully connected layer. The model is trained using training data, then the model is evaluated during training with validation data. Finally, testing data is used to evaluate the model, whether the model can predict correctly.

The model obtained for each modal is evaluated using accuracy, precision, recall, and F1-score (confusion matrix) to see how far the model can perform real or spoof face detection in the three models. Furthermore, the evaluation is also conducted using a precision-recall curve. This curve are commonly used to evaluate the performance of models with imbalanced datasets and provide more informative information [13].

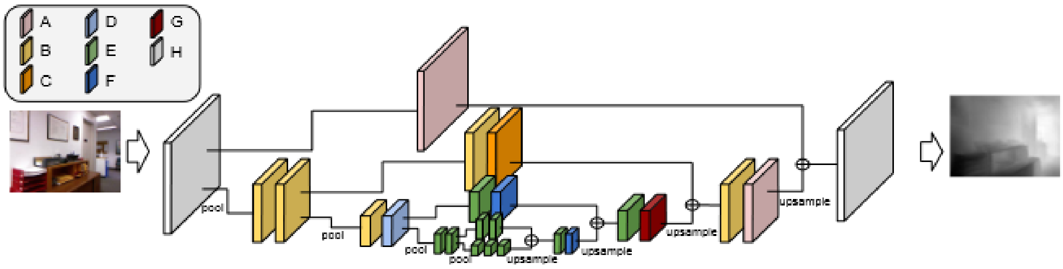

Fig. 4. The architecture of dense depth model [12].

The proposed method of this study is shown in Fig. 5. The proposed method uses a feature fusion approach. RGB features and depth features are generated from feature extraction using Inception-v3. Then, feature fusion is conducted by concatenating RGB features and depth features using concatenate layer. After the feature fusion process, the process is continued with classification using a fully connected layer. The model is trained using training data, then the model is evaluated during training with validation data. The last stage, the evaluation of the model.

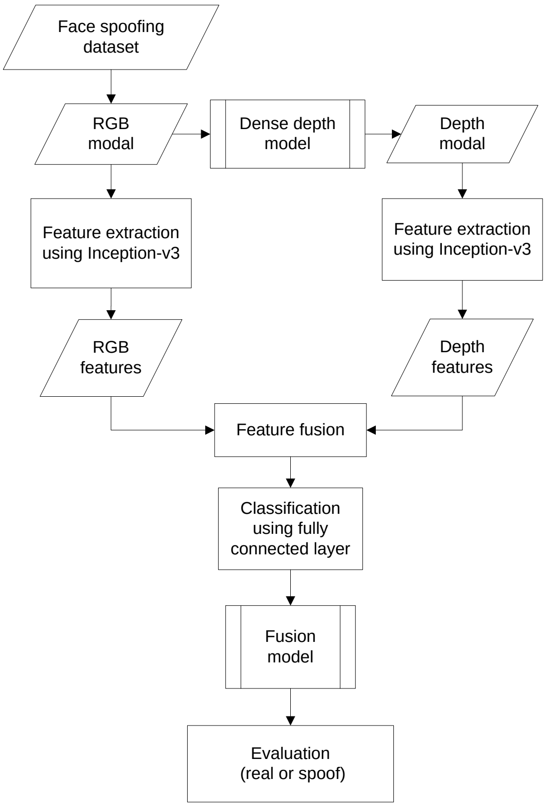
Face spoofing dataset
RGB modal
Dense depth model
Depth modal
Feature extraction using Inception-v3
Feature extraction using Inception-v3
RGB features
Depth features
Feature fusion
Classification using fully connected layer
Fusion model
Evaluation (real or spoof)

Fig. 5. The proposed method using fusion model.

The network architecture used in Fig. 4 dan Fig. 5 is shown in Fig. 6. Fig. 6 shows the feature extraction process using Inception-v3 and the classification process using a fully connected layer. Based on the study from Mednikov et al. [14], cutting the Inception-v3 network to mixed_6a then adding the network and retraining with a new dataset will get better performance. Therefore, this study is conducted by cutting the Inception-v3 network, then adding the network and training with the face spoofing dataset. The Inception-v3 architecture used is the Inception A, reduction A, and Inception B modules, as shown in Fig. 6. These modules are used in the feature extraction process for each modal.

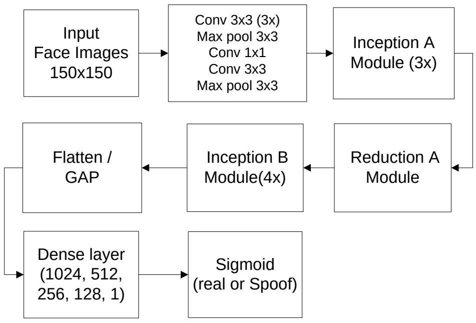

Fig. 6. The proposed method using fusion model.

The Inception A module consists of three blocks, i.e., mixed0, mixed1, and mixed2 layer. The Inception A module is shown in Fig. 7. The reduction module A includes mixed3 block layer, as shown in Fig. 8. The Inception B module consists of four (4) blocks, i.e., mixed4, mixed5, mixed6, and mixed7 layer, as shown in Fig. 9.

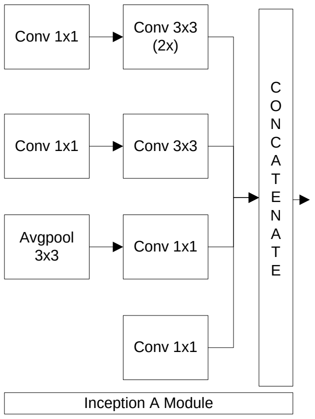

Fig. 7. The Inception A module.

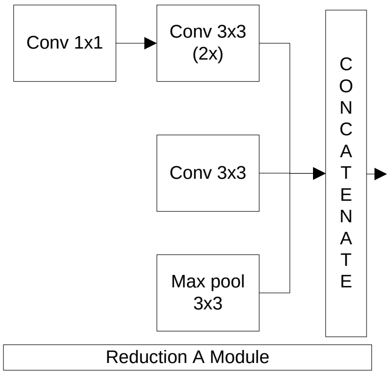

Fig. 8. The Reduction A module

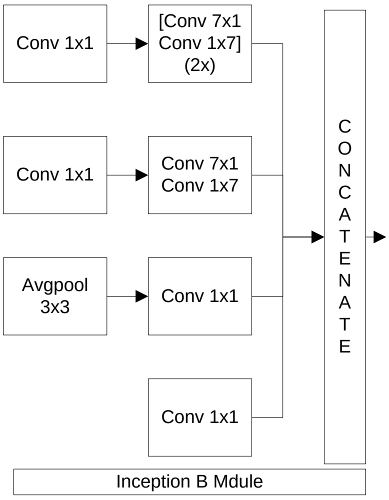

Fig. 9. The Inception B module.

The mixed layer shows the factorization process on the Inception-v3 architecture. The mixed layer is the block layer contained in each Inception-v3 module. For example, mixed0 block layer in Inception A module. This block consists of several convolutions, such as 1x1 convolution followed by two 3x3 convolution, 1x1 convolution followed by 3x3 convolution, average pooling followed by 1x1 convolution and 1x1 convolution. The last process is to concatenate all the outputs of the convolution. Factorization in the mixed0 layer is replacing 5x5 convolution with two 3x3 convolution to reduce the number of parameters.

The classification layer, as shown in Fig.6 consists of flatten layer or GAP layer (adjusted by experiment), fully connected layer (dense layer) with the number of units, 1024, 512, 256, 128, and 1. The classification used is binary classification with a Sigmoid activation function to determine whether a face is real or spoof.

3.3 Experimenal Scenario

The experiment scenario of this study is divided into two (2), as follows:

a. Transfer learning to get the best Inception-v3 layer output (Inception B module), representing the difference between real and spoof features. This transfer learning is conducted three times using mixed5 layer, mixed6 layer, and mixed7 layer output.
b. Transfer learning to get the best model for face spoofing detection using the best Inception layer output, representing the difference between real and spoof features. Transfer learning is conducted on the RGB modal, the depth modal and the fusion of both the modals.

4. Results and Analysis

4.1 Experimental Results

Face spoofing detection experiments on RGB modal, depth modal, and fusion (RGB and depth) are conducted using hyperparamater tuning as follows:

a. Optimizer: RMSProp and Adam, with a learning rate of 0.0001 and a batch size of 32; Loss metric: Binary Crossentropy and Evaluation metric: Binary Accuracy;
b. Transfer learning is trained with ImageNet weights and modified on the classification layer. The output of the base model uses flatten and Global Average Pooling (GAP). The model is modified by adding several dense layers (RELU), which the last dense layer with the number of nodes 1 is for binary classification (Sigmoid). Furthermore, to reduce overfitting, regularization is conducted by adding five Batch Normalization, and four Dropouts (0.5).

The first experiment is conducted to get the best Inception B module output layer, representing the difference between real and spoof. In this experiment, transfer learning is conducted three times using mixed5 layer, mixed6 layer, and mixed7 layer output on the RGB modal. The parameter used are RMSProp optimizer, learning rate 0.0001, batch size 32, and Flatten on the classification layer.

Based on Fig. 10, feature maps in the mixed5 and mixed6 layer are almost the same. The model can predict the real and spoof classes well so that both have almost same accuracy, which the accuracy of the mixed6 layer is bit higher than the mixed5 layer, as shown in Table 2. Otherwise the mixed7 layer has the lowest accuracy, possibly because there is lost information when feature extraction, so that is difficult to distinguish between real and spoof classes. Therefore, this study uses mixed6 layer output for the next experiment.

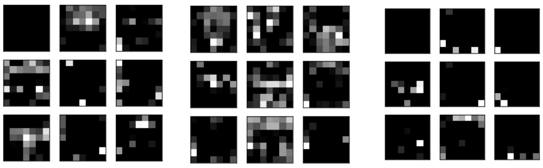

Fig. 10. Feature map of RGB modal. Mixed5 layer (left), mixed6 layer (middle), and mixed7 layer (right).

Table 2. Transfer learning Inception-v3.

Model	Layer	Epoch	Train acc	Valid acc	Test acc
RGB	Mixed5	100	0.9996	0.9721	0.9842
	Mixed6	100	1	0.9721	0.9857
	Mixed7	100	0.9999	0.9721	0.9764
Depth	Mixed5	10	0.9816	0.9206	0.9226
	Mixed6	10	0.973	0.9249	0.9255
	Mixed7	10	0.9793	0.9249	0.909

The second experiment is conducted to get the best model for face spoofing detection, using mixed6 layer output. In this experiment, transfer learning is conducted on the RGB modal, the depth modal, and fusion both of modals. Based on the experimental results in the RGB modal, the highest accuracy was obtained when using Adam optimizer, GAP, and epoch 100. As shown in Table 3, the training, validation, and testing accuracy are 100%, 97.64%, and 98.78%, respectively. The model learns well from the training data and has a good generalization from the validation data. The testing accuracy of 98.78% shows the model can predict testing data close to the actual value, i.e., from 1396 face images, 1379 are predicted correctly.

Table 3. The results of the RGB model.

RGB	Optimizer	Layer	Train acc	Valid acc	Test acc
Mixed6 layer	RMSProp	GAP	0.9991	0.9742	0.9814
		Flatten	1	0.9721	0.9857
	Adam	GAP	1	0.9764	0.9878
		Flatten	0.9996	0.9721	0.9857

The learning curve of the RGB model shows promising results (good fit), and there is no overfitting. Based on the curve in Fig. 11, training and validation accuracy increase steadily, and the gap between the two is relatively small. Likewise, training and validation loss decrease steadily, and the gap is also relatively small.

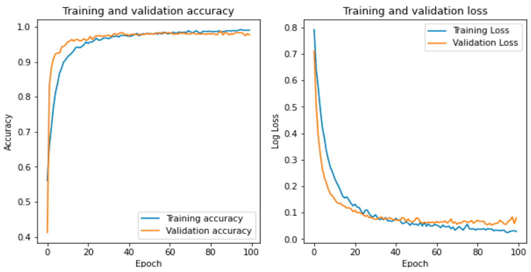

Fig. 11. The learning curve of the RGB model.

The best prediction results of the RGB model are shown in Table 4. The RGB model can detect and classify face images correctly. Only 17 face images are classified incorrectly, i.e., nine real images are predicted to be spoof class and eight spoof images are predicted to be a real class. The incorrect prediction results of the RGB model are shown in Fig. 12. The model cannot distinguish real and spoof images properly in dark lighting and blurry condition because those images are predicted incorrectly.

Table 4. The confusion matrix of the RGB model.

		Predicted	
		Real	Spoof
Actual	Real	229	9
	Spoof	8	1150

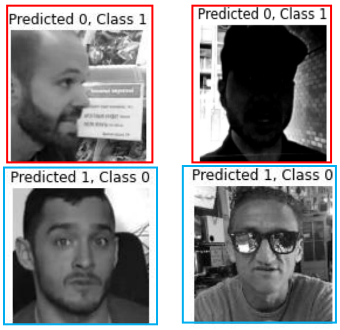

Fig. 12. Incorrect prediction of the RGB model. Real face images (blue box) and spoof face images (red box) are predicted incorrectly.

Based on the experimental results in the depth modal, the highest accuracy was obtained when using Adam optimizer, Flatten, and epoch 10. As shown in Table 5, the training, validation, and testing accuracy are 98.05%, 92.49%, and 92.84%, respectively. The testing accuracy of 92.84% shows that from 1396 face images (real and spoof), 1296 are predicted correctly. However, the accuracy of the Depth model is lower than the RGB model.

Table 5. The results of the depth model.

Depth	Optimizer	Layer	Train acc	Valid acc	Test acc
Mixed6 layer	RMSProp	GAP	0.8722	0.8648	0.8532
		Flatten	0.973	0.9249	0.9255
	Adam	GAP	0.8788	0.8755	0.8689
		Flatten	0.9805	0.9249	0.9284

The learning curve of the Depth model shows promising results (good fit) and there is no overfitting. Based on the curve in Fig. 13, training and validation accuracy increase steadily, and the gap between the two is relatively large at the beginning. However, after epoch 8 the gap between the two is getting smaller. Likewise, training and validation loss decrease steadily and after epoch 8 the gap is getting smaller.

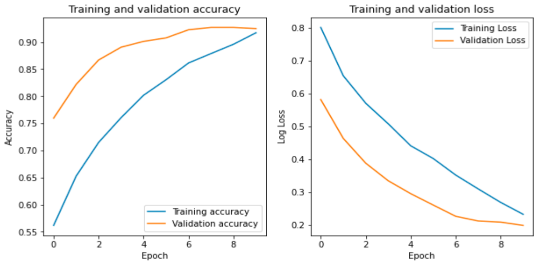

Fig. 13. The learning curve of the depth model.

The best prediction results of the Depth model are shown in Table 6. The Depth model can detect and classify face images correctly. However, the number of mispredicted images is more than the RGB model, which is 100 face images. Among them are 74 real images predicted to be spoof class and 26 spoof images predicted to be real class. The model cannot distinguish real and spoof images properly with less clear depth contrast or blends with the image background so that it is mispredicted, as shown Fig. 14. The condition of the depth images is caused by the original image, both real and spoof, has less/dark lighting and blurry condition.

Table 6. The confusion matrix of the depth model.

		Predicted	
		Real	Spoof
Actual	Real	164	74
	Spoof	26	1132

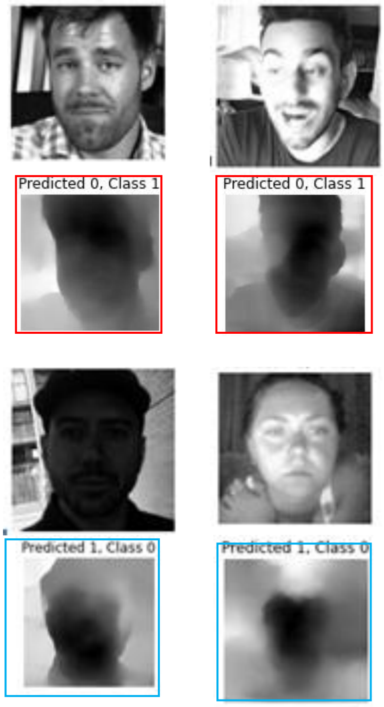

Fig. 14. Incorrect prediction of the depth model. Real face images (blue box) and spoof face images (red box) are predicted incorrectly.

Based on the experimental results with the fusion approach, the highest accuracy was obtained when using Adam optimizer, GAP, and epoch 100. As shown in Table 7, the training, validation, and testing accuracy are 100%, 98.28%, and 98.50%, respectively. The testing accuracy 98.50% shows from 1396 face images (real and spoof), 1375 are predicted correctly. The Fusion model results are almost same as the RGB model, possibly because the RGB features are more dominant than the depth features.

Table 7. The results of the fusion model.

Fusion	Optimizer	Layer	Train acc	Valid acc	Test acc
Mixed6 layer	RMSProp	GAP	0.9996	0.9742	0.9792
		Flatten	0.9999	0.9721	0.9771
	Adam	GAP	1	0.9828	0.985
		Flatten	0.9991	0.9721	0.9778

The learning curve of the Fusion model when using Adam optimizer and GAP shows promising results (good fit), and there is no overfitting. Based on the curve in Fig. 15, training accuracy and validation accuracy increase steadily, and the gap between the two is relatively small at epoch 100. Likewise, training and validation losses decrease steadily, and the gap is also relatively small at epoch 100.

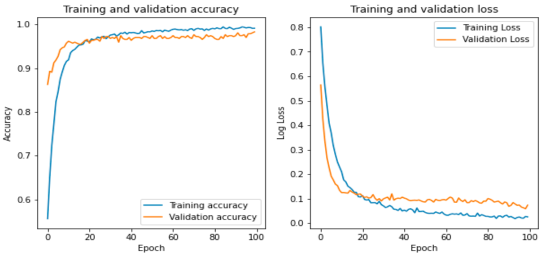

Fig. 15. The learning curve of the fusion model.

The best prediction results of the Fusion model are shown in Table 8. The RGB model can detect and classify face images correctly. Only 21 face images are classified incorrectly, i.e., eight real images are predicted to be spoof class and 13 spoof images are predicted to be a real class. The incorrect prediction results of the Fusion model are shown in Fig. 16. As in the RGB model, the Fusion model cannot distinguish real and spoof images properly in dark lighting and blurry conditions because those images are mispredicted.

Table 8. The confusion matrix of the fusion model.

		Predicted	
		Real	**Spoof**
Actual	**Real**	230	8
	Spoof	13	1145

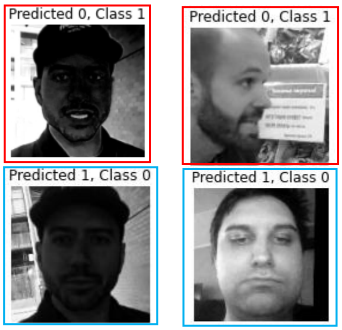

Fig. 16. Incorrect prediction of the fusion model. Real face images (blue box) and spoof face images (red box) are predicted incorrectly.

4.2 Evaluation

Based on Tabel 9, the RGB model and Fusion model results are almost same, but the RGB model has accuracy, precision, recall, and F1-score higher than the Fusion model. The RGB modal has sufficient color intensity to obtain certain features that can distinguish real and spoof face [15]. Therefore, RGB modal can represent real and spoof features well.

The RGB model provides accuracy of 98.78%, and be able to predict testing data close to the actual value, i.e., from 1396 face images (real and spoof), 1379 are predicted correctly. The prediction results can be seen in Table 4. The precision value is 99.22%, from 1159 prediction results, 1150 face images are predicted correctly as spoof class. The recall value is 99.31%, from 1158 spoof images, 1150 face images are predicted correctly as spoof class. The F1-score of the RGB model is 99.27%, almost the same as the Fusion model of 99.09%. In contrast to the Depth model, which only results in an F1-score of 95.77%.

Table 9. The performance evaluation of the RGB, depth and fusion model.

Model	**Acc**	***Precision***	***Recall***	***F1-score***
RGB	0.9878	0.9922	0.9931	0.9927
Depth	0.9284	0.9386	0.9775	0.9577
Fusion	0.985	0.9931	0.9888	0.9909

The precision-recall curve of the RGB, Depth and Fusion model is shown in Fig. 17. Baseline (blue dotted line) shows model without skill, model that cannot distinguish between classes and will predict a random class or a constant class. The best performance model is represented by a curve that is above Baseline and above the other curves.

Based on Fig.17, the RGB and Fusion model are the models that have the best performance compared to the Depth model. The precision and recall value of the the two models is high along the threshold. The RGB and Fusion curve are above Baseline line and above the Depth curve.

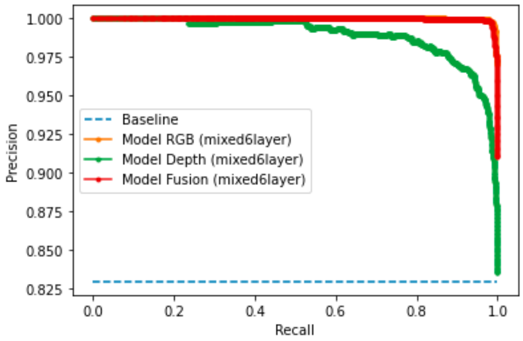

Fig. 17. The precision-recall curve.

The AUC of the RGB model and the Fusion model is also high, i.e., 0.9997 for the RGB model and 0.9995 for the Fusion model. The AUC of the Depth model is the lowest, which is 0.9893. The AUC value of the model is close to 1, indicating that the model can distinguish real and spoof class correctly.

This study also uses McNemar's test to compare models. Statistical tests such as McNemar’s test can be used to compare models [16]. The statistical test in this study was carried out only on the RGB model and the Fusion model with the aim of seeing whether the use of Depth modal has an effect on the Fusion model. The hypothesis used, namely:

H0: Addition of modal depth has no effect (same error proportion).
H1: Addition of modal depth has an effect (different error proportion).

The calculation of McNemar’s statistic test uses the predictive data shown in the contingency table Tabel 10. This table shows the amount of data that is predicted correctly or incorrectly by the RGB model and the Fusion model based on real/spoof class.

Table 10. The contingency tabel.

		Fusion Model	
		0 (Real)	**1 (Spoof)**
RGB Model	**0 (Real)**	A = 230	B = 7
	1 (Spoof)	C = 13	D = 1146

The McNemar’s test value is calculated by equation (1):

$$\chi^2 = \frac{(|C-B|-1)^2}{C+B} \quad (1)$$

where, C is the number of images that are predicted to be spoof by the RGB model and the number of images that are predicted to be real by the Fusion model; B is the number of images that are predicted to be real by the RGB model and the number of images that are predicted to be spoof by the Fusion model.

The significance level (α) used is 0.05. The resulting χ^2 value is 1.250 and the resulting p-value is 0.264. Based on the results of McNemar's test with a significance level of 5%, the p-value > α. Therefore, it can be stated that there is no significant difference with the addition of Depth modal in face spoofing detection, because the proportion of errors in both models has the same proportion.

4.3 Evaluation with MSU MFSD Dataset

The proposed method of this study is face spoofing detection in RGB modal, depth modal, and fusion approach using transfer learning by cutting the Inception-v3 network to the mixed6 layer. Nagpal & Dubey [4] conducted a study on face spoofing detection only on RGB modal using the same architecture, but uses the entire Inception-v3 layer.

Therefore, we evaluate and compare performance with the study of Nagpal & Dubey [4]. The dataset used is the MSU Mobile Face Spoofing Database (MFSD) benchmark dataset [16]. Our proposed method uses following parameters: learning rate 0.0001, batch size 32, Adam optimizer, and GAP. The metric used in this experimental evaluation is accuracy. The experimental results show that our proposed method is superior to research [4], with training, validation and testing accuracy up to 100%, as shown in Table 11.

Table 11. The evaluation and comparison (MSU MFSD dataset).

Model	Layer	Train Acc	Valid Acc	Test Acc
Nagpal & Dubey (2019) [4]	All layers	0.9463	0.9647	0.9613
Proposed Method	Mixed6 layer	1	1	1

5. Conclusion

Based on the results, face spoofing detection by cutting the Inception-v3 network until the mixed6 layer has good performance. The models can learn different features between real and spoof classes without using all layers. Face spoofing detection has a good performance on the RGB and fusion models. Both models have better performance than the depth model because RGB modal can represent real and spoof features well, and RGB modal dominate the fusion model. In contrast to the depth modal generated from the dense Depth model, it may not represent the difference between real and spoof features. This condition is also likely to cause the performance of the Fusion model to be lower than the RGB model. In evaluating and comparing methods, our proposed method successfully outperforms the previous study with accuracy up to 100%.

The RGB, Depth, and Fusion model cannot predict correctly for face images with dark lighting and blur conditions. Images with these conditions are predicted incorrectly, some are real classes and some are spoof classes. RGB image in dark lighting conditions will have almost the same pixel intensity, approaching 255, so the image tends to be dark in color. The image with the blur condition has almost the same pixel intensity, but is close to 0 so that the image is light in color. These condition makes the depth image has a less clear depth contrast. Therefore, RGB image with these conditions, both real and spoof, are predicted to be incorrectly. The model has difficulty distinguishing images with these conditions.

For future study, image preprocessing is needed, e.g., histogram equalization for dark lighting images and sharpening filter for blurry images. Image preprocessing with histogram equalization gives a good performance on deep learning in study of Yue & Lu [17].

Based on McNemar’s test, the addition of Depth modal in face spoofing detection has no effect, so the use of the depth and fusion modal in the model cannot increase the accuracy of face spoofing detection. This is probably because the depth information produced is less accurate, making it difficult to distinguish between real and spoof facial features. Therefore, it is necessary to carry out further exploration of accurate and precise depth estimation in distinguishing the real face depth feature and spoof face, so that when combined with the RGB modal it can increase the accuracy of face spoofing detection. Therefore, the use of depth information can be considered in the detection of face spoofing.

Acknowledgement

This study is fully supported by Beasiswa Saintek-Pusat Pendidikan dan Pelatihan Kementerian Riset dan Teknologi/Badan Riset dan Inovasi Nasional (Pusdiklat Kemenristek/BRIN). We gratefully acknowledge Beasiswa Saintek Pusdiklat Kemenristek/BRIN.

References

[1] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, “Deep Tree Learning for Zero-Shot Face Anti-Spoofing,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 4675–4684. doi: 10.1109/CVPR.2019.00481.

[2] X. Yang *et al.*, “Face Anti-Spoofing: Model Matters, so Does Data,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 3502–3511. doi: 10.1109/CVPR.2019.00362.

[3] A. Jourabloo, Y. Liu, and X. Liu, “Face De-Spoofing: Anti-Spoofing via Noise Modeling.” arXiv, Jul. 26, 2018. Accessed: Jul. 15, 2022. [Online]. Available: http://arxiv.org/abs/1807.09968

[4] C. Nagpal and S. R. Dubey, “A Performance Evaluation of Convolutional Neural Networks for Face Anti Spoofing,” in *2019 International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, Jul. 2019, pp. 1–8. doi: 10.1109/IJCNN.2019.8852422.

[5] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, “Face anti-spoofing using patch and depth-based CNNs,” in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, Denver, CO, Oct. 2017, pp. 319–328. doi: 10.1109/BTAS.2017.8272713.

[6] Y. Liu, A. Jourabloo, and X. Liu, “Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, Jun. 2018, pp. 389–398. doi: 10.1109/CVPR.2018.00048.

[7] Z. Wang *et al.*, “Exploiting temporal and depth information for multi-frame face anti-spoofing.” arXiv, Mar. 05, 2019. Accessed: Jul. 19, 2022. [Online]. Available: http://arxiv.org/abs/1811.05118

[8] S. Zhang *et al.*, “A Dataset and Benchmark for Large-Scale Multi-Modal Face Anti-Spoofing,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 919–928. doi: 10.1109/CVPR.2019.00101.

[9] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2818–2826. doi: 10.1109/CVPR.2016.308.

[10] S. Lim, Y. Gwak, W. Kim, J.-H. Roh, and S. Cho, “One-Class Learning Method Based on Live Correlation Loss for Face Anti-Spoofing,” *IEEE Access*, vol. 8, pp. 201635–201648, 2020, doi: 10.1109/ACCESS.2020.3035747.

[11] Z. Li *et al.*, “Learning the Depths of Moving People by Watching Frozen People.” arXiv, Apr. 24, 2019. Accessed: Jul. 15, 2022. [Online]. Available: http://arxiv.org/abs/1904.11111

[12] W. Chen, Z. Fu, D. Yang, and J. Deng, “Single-Image Depth Perception in the Wild.” arXiv, Jan. 06, 2017. Accessed: Jul. 20, 2022. [Online]. Available: http://arxiv.org/abs/1604.03901

[13] T. Saito and M. Rehmsmeier, “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets,” *PLOS ONE*, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: 10.1371/journal.pone.0118432.

[14] Y. Mednikov, S. Nehemia, B. Zheng, O. Benzaquen, and D. Lederman, “Transfer Representation Learning using Inception-V3 for the Detection of Masses in Mammography,” in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Honolulu, HI, Jul. 2018, pp. 2587–2590. doi: 10.1109/EMBC.2018.8512750.

[15] L. Souza, M. Pamplona, L. Oliveira, and J. Papa, “How far did we get in face spoofing detection?,” *Eng. Appl. Artif. Intell.*, vol. 72, pp. 368–381, Jun. 2018, doi: 10.1016/j.engappai.2018.04.013.

[16] Di Wen, Hu Han, and A. K. Jain, “Face Spoof Detection With Image Distortion Analysis,” *IEEE Trans. Inf. Forensics Secur.*, vol. 10, no. 4, pp. 746–761, Apr. 2015, doi: 10.1109/TIFS.2015.2400395.

[17] G. Yue and L. Lu, “Face Recognition Based on Histogram Equalization and Convolution Neural Network,” in *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Hangzhou, Aug. 2018, pp. 336–339. doi: 10.1109/IHMSC.2018.00084.