

# Comparison of Transfer Learning Using VGG16, MobileNetV2, and ResNet50 for Pornography Image Detection

Christopher Ade SihWardana<sup>1</sup>, Sani Muhamad Isa<sup>2</sup>

<sup>1</sup>Department of Master Computer Science, Bina Nusantara University, Jakarta, Indonesia

<sup>2</sup>Department of Computer Science, Bina Nusantara University, Jakarta, Indonesia

## Article Info

### Article history:

Received Jun 20, 2025

Revised Jul 28, 2025

Accepted Oct 25, 2025

### Keywords:

Pornography Image Detection

CNN

Skin Color Segmentation

Color Space

Transfer Learning

## ABSTRACT

The rapid growth of digital technology is vital for the Indonesian Scout to reach and interact with its members. The National Indonesian Scout (*Kwarnas*) uses the “*Ayo Pramuka*” social media application to support this. However, such platforms risk exposing users, especially teenagers, to harmful content like pornography. This research applies Computer Vision and Transfer Learning Convolutional Neural Networks (CNNs) to detect pornographic images automatically. The objective is to identify the CNN model (VGG16, MobileNet V2, ResNet 50) with the highest detection accuracy and determine the impact of color space preprocessing. The method includes two stages first, image preprocessing by converting RGB images to HSV and YCbCr second, feature extraction using pre-trained CNNs with freezing and fine-tuning. A dataset of 4060 images was used for training and testing. Without preprocessing, VGG16 achieved the best accuracy of 99.01%. When RGB images were converted to HSV, ResNet 50 produced the highest accuracy of 99.51%. The findings show that combining color space transformation and Transfer Learning CNN significantly improves pornographic content detection in the “*Ayo Pramuka*” Application, enhancing safe digital engagement for Indonesian Scouts.

This is an open access article under the [CC BY-SA](#) license.



## Corresponding Author:

Christopher Ade SihWardana

Department of Master Computer Science

Bina Nusantara University

Jl. Raya Kb. Jeruk No.27, RT.1/RW.9, Kemanggisian, Kec. Palmerah, Kota Jakarta Barat, Jakarta 11530

Email: christopher.sihwardana@binus.ac.id

## 1. INTRODUCTION

*Gerakan Pramuka (Kwarnas)* is Indonesia's largest non-formal education organization, focusing on character development, leadership, and social awareness for the younger generation [1], [2]. With technological advancements, *Gerakan Pramuka (Kwarnas)* has adapted through digital innovation, including the “*Ayo Pramuka*” application, which has attracted more than four million users [2]. This social media platform enhances member engagement and facilitates collaboration. However, the adoption of digital technology also presents risks, particularly exposure to inappropriate content such as pornography, which threatens the moral values upheld by *Gerakan Pramuka (Kwarnas)* [3], [4]. Online media is a primary source of pornographic content accessed by adolescents, posing significant risks to physical, mental, and moral well-being [4]. Studies highlight that virtual social organizations often struggle to maintain strong social bonds; however, members engaged in social media-based platforms may develop a heightened sense of responsibility toward the organization's mission [5]. This highlights the need for a detection system that can effectively prevent the spread of inappropriate materials.

Advanced technological solutions have been explored to mitigate this risk. Machine learning, particularly Convolutional Neural Networks (CNN), has proven effective in detecting adult content [6], [7], [12]. CNN analyzes image data by progressively extracting features from

basic shapes to complex structures, enabling it to differentiate pornographic from non-explicit content with high accuracy while reducing false positives.

Despite advancements in CNN-based pornography detection, many studies have underutilized preprocessing techniques, which are crucial for improving model accuracy. Preprocessing methods such as skin color detection, edge detection, and color space transformation significantly enhance feature extraction [8], [9], [11], [26]. However, some studies rely solely on raw RGB image input.

A hybrid CNN model combined with a Restricted Gaussian-Bernoulli Boltzmann Machine (RBM) was introduced in [6], achieving 97.80% accuracy without extensive preprocessing. However, the study reported challenges in skin color detection, with an accuracy limitation of 89.34% due to variations in human skin tones. This limitation underscores the need for advanced preprocessing techniques such as color space transformation to improve detection reliability. Similarly, in [7], a Deep Convolutional Neural Network (DCNN) with transfer learning through fine-tuning was used to detect pornographic images, achieving 78% accuracy. However, the study did not incorporate advanced preprocessing techniques, leaving the model vulnerable to background noise and reducing detection performance. A skin color detection approach using the YCbCr color space combined with the eigenporn method was proposed in [8]. While this method successfully identified skin areas and color, it had significant limitations, particularly in handling large datasets. The eigenporn method requires analyzing each image individually and determining the nearest distance, leading to inefficiencies in classification when processing large-scale data. Consequently, the model experienced increased computational time and reduced scalability.

Preprocessing techniques such as color space transformation have significantly enhanced feature extraction, particularly in distinguishing skin tones and texture patterns, as discussed in [9], [28], [29]. However, in [9], resistance to lighting variations, reliance on manually set threshold values, and underutilization of spatial information and adaptive capabilities were reported as major drawbacks.

Therefore, this study proposes the integration of advanced color space transformations with transfer learning using pre-trained CNN models, including VGG16, MobileNetV2, and ResNet50. Both the Freezing Layer and Fine-Tuning methods are applied to improve the accuracy and efficiency of the detection system. The objectives of this research are expected to strengthen content moderation in the “*Ayo Pramuka*” application, thereby protecting the younger generation from exposure to inappropriate content and preserving the integrity of *Gerakan Pramuka (Kwarnas)*.

## 2. RESEARCH METHOD

In this study, several stages were carried out. The first stage is dataset collection, the second is preprocessing, the third is data division, the fourth is feature extraction, the fifth is dataset experimentation, the sixth is model building, the seventh is model training, the eighth is testing, and the ninth is detection results. The experimental flow presented in Figure 1 provides a more detailed explanation.

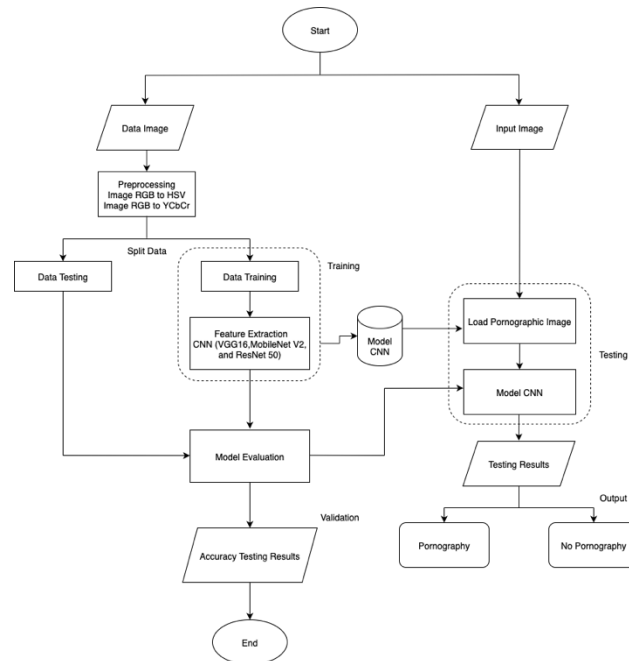


Figure 1. Research Stages

Based on Figure 1, it can be explained that this research experiment was carried out through the following stages:

- Data preprocessing.** At this stage, the image preprocessing process is carried out to prepare the data before using it in model training. The images used are converted from RGB (Red, Green, Blue) to HSV (Hue, Saturation, Value) format and vice versa. This conversion aims to highlight color characteristics that can help improve detection accuracy. The data used in the study is divided into two parts: training data to train the model and testing data to test the model's performance after training.
- Feature extraction.** At this stage, feature extraction uses several Convolutional Neural Network (CNN) architectures, namely VGG16, MobileNet V2, and ResNet50 [12], [22], [23]. This process aims to extract features and then analyze their values for the classification process stage. Training data split at a 90:10 ratio using a machine learning model will be prepared before being processed in the CNN feature extraction model. This allows the model to understand image patterns that are the leading indicators in distinguishing pornographic and non-pornographic images, model training, and evaluation. After feature extraction, the data is trained using the CNN model. This model is designed to recognize complex image patterns through several convolutional and pooling layers. After the training process, the model is evaluated using test data. This evaluation aims to measure the performance of the model in classifying images.
- Pornography detection.** The image to be tested is entered into the system at this stage. The trained CNN model processes the image and provides a classification result indicating whether it falls into the pornography or non-pornography category.
- Performance measurement.** Model performance is measured using evaluation metrics such as accuracy, precision, recall, and F1 score. These metrics evaluate how well the model can classify correctly and avoid prediction errors.
- Best result.** The best results are obtained from testing model performance based on various CNN architectures and training parameters. The model with the best performance is determined from the highest accuracy value and other evaluation metrics.

## 2.1 Adult Content Dataset

The initial step in this study was data collection, utilizing secondary data obtained from a publicly available repository, namely the Adult Content Dataset, which can be accessed at [https://figshare.com/articles/dataset/Adult\\_content\\_dataset/13456484](https://figshare.com/articles/dataset/Adult_content_dataset/13456484). This dataset consists of images labeled as either pornographic or non-pornographic based on their visual characteristics.

Following the re-sorting process, the final dataset consisted of 4060 images, with an equal distribution of 2030 pornographic and 2030 non-pornographic samples. Pornographic images typically featured very revealing or vulgar clothing that exposed most of the body, whereas non-pornographic images exhibited the opposite characteristics as modest clothing or visuals that did not explicitly highlight the human body.

After classification, a pre-processing step was applied by resizing all images. Oversized images can slow down the computation process, while undersized images may degrade detection accuracy. Therefore, all images were resized to  $128 \times 128$  pixels to maintain consistency, preserve sufficient detail, and enhance computational efficiency. An example of the resized images is shown in Figure 2.

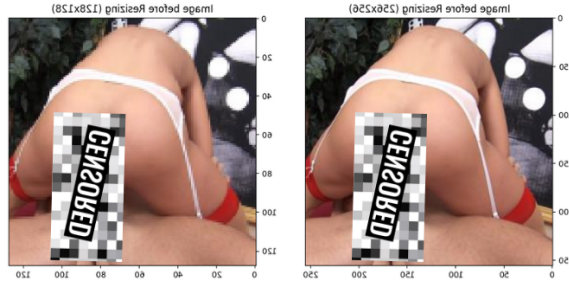


Figure 2. Image Resizing Example

## 2.2 Preprocessing

As shown in Figure 1, the preprocessing stage of the pornography image dataset involves converting RGB images into HSV and YCbCr color spaces. This transformation is part of a strategy to improve the performance of Transfer Learning models, namely VGG16, MobileNetV2, and ResNet50, for pornography image classification. HSV and YCbCr were chosen due to their proven effectiveness in skin color detection and visual segmentation, especially under varying lighting conditions. HSV separates chromatic content from luminance, making it more robust in detecting skin-colored areas [28], while YCbCr isolates luminance from chrominance components, which helps reduce background interference [29]. By applying these color space conversions, training time is reduced, and detection accuracy is improved, especially in models that benefit from a clearer representation of skin color features. The conversion process from RGB to HSV and YCbCr follows a standard formula presented as follows:

$$r = \frac{R}{(R+G+B)}, \quad g = \frac{G}{(R+G+B)}, \quad b = \frac{B}{(R+G+B)} \quad (1)$$

$$H = \begin{cases} 0, & \text{if } S = 0 \\ \frac{60 * (g-b)}{S * V}, & \text{if } V = r \\ 60 * \left[ 2 + \frac{b-r}{S * V} \right], & \text{if } V = g \\ 60 * \left[ 4 + \frac{r-g}{S * V} \right], & \text{if } V = b \end{cases} \quad (2)$$

$$S = \begin{cases} 0, & V = 0 \\ 1 - \frac{\min(r,g,b)}{V}, & V > 0 \end{cases} \quad (3)$$

$$V = \max(r, g, b) \quad (4)$$

As an example of changing an RGB image to HSV, it can be seen in Figure 2.

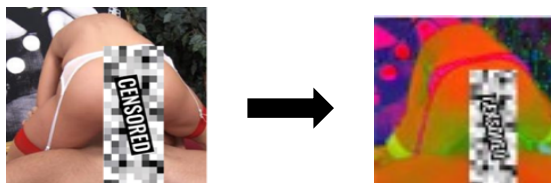


Figure 3. Preprocessing RGB to HSV

The following image is a preprocessing of RGB to YCbCr. Figure 4 shows the results of changing the color space.

### 3. RESULTS AND DISCUSSION



Figure 4. Preprocessing RGB to YCbCr

The formula below is a preprocessing of an RGB image into a YCbCr image:

$$Y = 0.299R + 0.587G + 0.114B. \quad (5)$$

$$Cb = 128 - 0.168736R - 0.331264G + 0.5B \quad (6)$$

$$Cr = 128 + 0.5R - 0.418688G - 0.081312 \quad (7)$$

Once the RGB images are transformed into HSV and YCbCr formats, they are converted into arrays and saved according to their respective classes, which are categorized as pornography and non-pornography. The data will then be processed using CNN models, including VGG16, MobileNet V2, and ResNet50, and classified through the Normalized Softmax Transform.

#### 3.1. Proposed Method

The proposed method uses VGG16, MobileNet V2, and ResNet50-based Convolutional Neural Network (CNN) architectures to classify images into two classes. A data-driven approach is used to understand the semantic characteristics through a deep CNN. This section discusses network construction, training procedures, and model testing details.

The model receives input as a colored image of size (128x128x3). The network structure consists of five main blocks, each containing multiple Conv2D layers with ReLU activation and ending with a MaxPooling2D layer (pool size 2x2, stride 2x2). The first and second blocks consist of two convolutional layers with 64 and 128 filters, respectively. The third to fifth blocks have three convolutional layers with 256 and 512 filters, respectively. 'Same' padding is used on all convolutional layers to preserve the spatial dimension. After the convolution process, the output is converted into a vector using a flattened layer, after which our model continues with three Dense layers with the first layer of 512 units, the second layer with 128 units, and the last layer with two units using softmax activation to predict two classes (pornography and non-pornography). ReLU activation is applied to all dense layers to prevent the propagation of negative values. The model is compiled using the Adam optimizer with a learning rate of 0.0001 to speed up convergence. Early Stopping stops training if the validation accuracy does not improve significantly, while Model Checkpoint stores the best model weights based on the validation accuracy. Image Data Generator is used for data augmentation and generates data batches in real-time during training.

Table 1. Training Parameters

|             |                                     |
|-------------|-------------------------------------|
| Target size | img_width = 128<br>img_height = 128 |
| Classes     | Pornography<br>Non Pornography      |
| Class       | Binary                              |
| Batch size  | 24                                  |

We use images with a resolution of  $128 \times 128$  pixels ( $\text{img\_width} = 128$ ,  $\text{img\_height} = 128$ ) as model input. Using a binary classification approach, we classify the images into two classes: pornography and non-pornography. During the model training process, the data is processed in batch sizes of 24 epochs per iteration to ensure efficiency in processing and stability of weight updates.

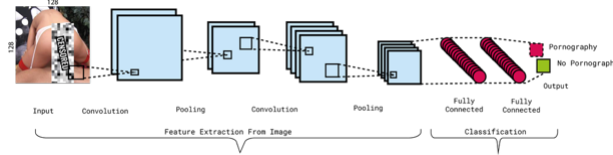


Figure 5. Structure of the proposed method

Table 2. Trained parameters

| Layer Type                    | Output Shape         | Param     |
|-------------------------------|----------------------|-----------|
| input_layer<br>(InputLayer)   | (None, 128, 128, 3)  | 0         |
| block1_conv1<br>(Conv2D)      | (None, 128, 128, 64) | 1,792     |
| block1_conv2<br>(Conv2D)      | (None, 128, 128, 64) | 36,928    |
| block1_pool<br>(MaxPooling2D) | (None, 64, 64, 64)   | 0         |
| block2_conv1<br>(Conv2D)      | (None, 64, 64, 128)  | 73,856    |
| block2_conv2<br>(Conv2D)      | (None, 64, 64, 128)  | 147,584   |
| block2_pool<br>(MaxPooling2D) | (None, 32, 32, 128)  | 0         |
| block3_conv1<br>(Conv2D)      | (None, 32, 32, 256)  | 295,168   |
| block3_conv2<br>(Conv2D)      | (None, 32, 32, 256)  | 590,080   |
| block3_conv3<br>(Conv2D)      | (None, 32, 32, 256)  | 590,080   |
| block3_pool<br>(MaxPooling2D) | (None, 16, 16, 256)  | 0         |
| block4_conv1<br>(Conv2D)      | (None, 16, 16, 512)  | 1,180,160 |
| block4_conv2<br>(Conv2D)      | (None, 16, 16, 512)  | 2,359,808 |
| block4_conv3<br>(Conv2D)      | (None, 16, 16, 512)  | 2,359,808 |
| block4_pool<br>(MaxPooling2D) | (None, 8, 8, 512)    | 0         |
| block5_conv1<br>(Conv2D)      | (None, 8, 8, 512)    | 2,359,808 |
| block5_conv2<br>(Conv2D)      | (None, 8, 8, 512)    | 2,359,808 |
| block5_conv3<br>(Conv2D)      | (None, 8, 8, 512)    | 2,359,808 |
| block5_pool<br>(MaxPooling2D) | (None, 4, 4, 512)    | 0         |
| flatten (Flatten)             | (None, 8192)         | 0         |
| dense (Dense)                 | (None, 512)          | 4,194,816 |

Steps per epoch refers to the batch size used to supply instructional data to the model in the fit generator and validation steps processes. The test data used are listed in Table 2.

In the process of training the model with the fine-tuning technique, freezing is performed on most of the initial layers of the network, while only the final layer is retrained. To train this final layer, the Adam optimizer is used with a learning rate of 0.0001.

### 3.2 Evaluation Metrics

The next step after testing to evaluate the classification technique's effectiveness is to use a confusion matrix, a formula that shows a model or algorithm's performance. Various evaluation metrics, including accuracy, precision, recall, and F1Score, are often used in classification to assess model performance. The following is the confusion matrix formula.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

$$F1\ Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (11)$$

#### 4. DISCUSSION

The dataset contains 4060 explicit images for the training and testing process. It is partitioned into 90% training data, comprising 3654 images, and 10% testing data, with 406 images. This dataset consists of images labeled as either pornographic or non-pornographic based on their visual characteristics. The original labeling was conducted manually by the dataset creators, without additional information regarding validation procedures or expert involvement.

To ensure alignment with the Indonesian legal context, a conservative re-sorting process was conducted by referring to explicit indicators outlined in Law of the Republic of Indonesia No. 44 of 2008 on Pornography, such as frontal nudity, explicit sexual activity, or similar visual representations [15].

The next step is the preprocessing stage, after the dataset needed for this research is collected. Preprocessing is done by performing an augmentation process using ImageDataGenerator, which is applied to the dataset. The testing strategy consists of three approaches:

1. Input images are utilized in their original RGB format without preprocessing, employing transfer learning models such as VGG16, MobileNet V2, and ResNet50 for analysis.
2. The input images are preprocessed by converting them from RGB to HSV. Then, transfer learning models VGG16, MobileNet V2, and ResNet50 are used for evaluation.
3. The third strategy processes the input image from the RGB color space to YCbCr.

Using VGG16, MobileNet V2, and ResNet50 models, each model will be tested using a freezing layer in the initial layer to reduce the risk of overfitting and speed up the training process and fine-tuning methods. Figure 6 visualizes the training data augmentation results.

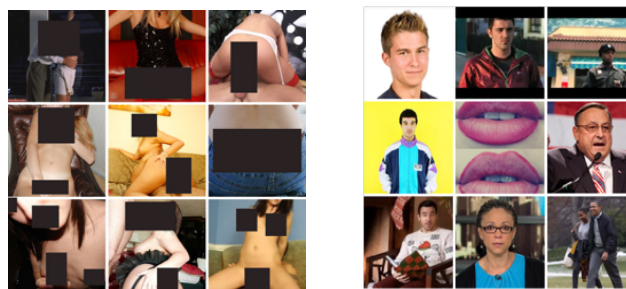


Figure 6. Sample of the augmented training images

Table 3. Performance Evaluation of the VGG16 Model

| Methods     | Color Space | Accuracy | Precision | Recall | F1 Score |
|-------------|-------------|----------|-----------|--------|----------|
| Freezing    | RGB         | 98.28%   | 99.07%    | 97.71% | 98.38%   |
| Fine-Tuning | RGB         | 98.77%   | 99.45%    | 97.84% | 98.64%   |
| Freezing    | HSV         | 95.57%   | 98.04%    | 93.46% | 95.69%   |
| Fine-Tuning | HSV         | -        | -         | -      | -        |
| Freezing    | YCbCr       | 96.14%   | 98.51%    | 93.87% | 96.14%   |
| Fine-Tuning | YCbCr       | 97.04%   | 96.60%    | 97.75% | 97.07%   |

The best model in Table 3 is VGG16, which achieves 98.77% accuracy without preprocessing on the input image, utilizing the RGB color space. The VGG16 model with RGB input demonstrates excellent stability in classifying pornography images, as shown in the performance graph. In contrast, experiments with the color transformation from RGB to HSV (Hue, Saturation, Value) and RGB to YCbCr were carried out to evaluate the potential for accuracy improvement. The results show that the transformation from RGB to HSV only achieves an accuracy of 95.57%, while the transformation to YCbCr yields 97.04% using the fine-tuning method and 96.14% using the freezing layer method. However, the VGG16 + HSV configuration with fine-tuning consistently failed to converge, producing unstable accuracy and loss, and therefore could not generate valid classification results. This indicates that VGG16 is optimized for RGB input, and transformations into other color spaces may reduce important information relevant to pornography image detection. Furthermore, preprocessing adds an extra step without significantly improving performance, making direct RGB input more accurate and computationally efficient.

Table 4. Performance Evaluation of the MobileNet V2 Model

| Methods     | Color Space | Accuracy | Precision | Recall | F1 Score |
|-------------|-------------|----------|-----------|--------|----------|
| Freezing    | RGB         | 99.01%   | 98.53%    | 99.50% | 99.01%   |
| Fine-Tuning | RGB         | 99.01%   | 99.51%    | 98.54% | 99.02%   |
| Freezing    | HSV         | 90.39%   | 89.62%    | 91.79% | 90.69%   |
| Fine-Tuning | HSV         | 95.81%   | 96.98%    | 94.61% | 95.78%   |
| Freezing    | YCbCr       | 98.52%   | 97.89%    | 98.94% | 98.41%   |
| Fine-Tuning | YCbCr       | 98.03%   | 99.05%    | 97.20% | 98.11%   |

The best model in Table 4 is MobileNet V2 with the fine-tuning method, achieving 99.01% accuracy without preprocessing on the input image, where the input image is without preprocessing using the RGB color space. As for the addition of preprocessing, namely RGB to HSV in the MobileNet V2 model, it does not affect the increase in accuracy, even the accuracy obtained is not better than RGB and YCbCr, namely 90.39% for the Freezing model, 95.81% for the Fine-Tuning model in HSV color space while in YCbCr color space the accuracy for Freezing is 97.73%, and 98.13% for Fine-Tuning, MobileNet V2 still shows competitive accuracy, especially with the fine-tuning method, which improves precision and recalls significantly.

The MobileNet V2 model is particularly notable for its ability to maintain high accuracy despite using a lighter and more efficient model, allowing for fast and accurate detection. The fine-tuning method consistently yielded better results than freezing, indicating that the model can adapt well to pornography detection tasks, resulting in optimal performance in a wide range of color spaces. The stability evaluation metric F1 score balances precision and recall across the tested color spaces. The regularization applied during fine-tuning helps prevent the model from overfitting and ensures that the model remains well generalized.

Table 5. Performance Evaluation of the ResNet 50 Model

| Methods     | Color Space | Accuracy | Precision | Recall | F1 Score |
|-------------|-------------|----------|-----------|--------|----------|
| Freezing    | RGB         | 86.45%   | 98.17%    | 80.83% | 88.66%   |
| Fine-Tuning | RGB         | 85.96%   | 78.77%    | 93.30% | 85.42%   |
| Freezing    | HSV         | 99.51%   | 100%      | 99.02% | 99.51%   |
| Fine-Tuning | HSV         | 99.01%   | 99.52%    | 98.11% | 98.81%   |
| Freezing    | YCbCr       | 98.52%   | 88.18%    | 80.19% | 95.95%   |
| Fine-Tuning | YCbCr       | -        | -         | -      | -        |

The best model in Table 5 is ResNet50 with the freezing method, achieving 99.51% accuracy using HSV preprocessing. This configuration shows that when RGB images are transformed into HSV and convolutional blocks 5 to 1 are frozen, the model extracts more discriminative features relevant to pornographic content. The freezing strategy helps reduce overfitting by limiting parameter updates, while HSV offers a more intuitive representation of skin tones and body patterns, making detection more robust against lighting variations. In contrast, experiments with RGB input achieved only 89.70% accuracy with freezing and 85.96% with fine-tuning, which improved to 99.01% after applying HSV preprocessing. Conversely, the RGB to YCbCr transformation did not significantly improve performance, and in the fine-tuning configuration, it failed to converge. This failure can be attributed to a representational mismatch between YCbCr preprocessing and the pretrained filters of ResNet50, which were optimized for RGB images from ImageNet. The separation of luminance and chrominance in YCbCr created a distribution gap that the model's frozen or partially fine-tuned layers could not adapt to, resulting in classification failure.

Table 6. The result of the experiments of each model

| Model            | Methods         | Data Image | Accuracy      | Val_Loss      | F1 Score      |
|------------------|-----------------|------------|---------------|---------------|---------------|
| VGG16            | Freezing        | RGB        | 98.28%        | 0.0398        | 98.38%        |
| VGG16            | Fine-Tuning     | RGB        | 98.77%        | 0.0620        | 98.64%        |
| VGG16            | Freezing        | HSV        | 95.57%        | 0.1874        | 95.69%        |
| VGG16            | Fine-Tuning     | HSV        | -             | -             | -             |
| VGG16            | Freezing        | YCbCr      | 96.14%        | 0.1035        | 96.14%        |
| VGG16            | Fine-Tuning     | YCbCr      | 97.04%        | 0.1474        | 97.07%        |
| MobileNet V2     | Freezing        | RGB        | 98.62%        | 0.0533        | 99.01%        |
| MobileNet V2     | Fine-Tuning     | RGB        | 97.64%        | 0.1589        | 99.02%        |
| MobileNet V2     | Freezing        | HSV        | 90.39%        | 0.0712        | 90.69%        |
| MobileNet V2     | Fine-Tuning     | HSV        | 95.81%        | 0.9581        | 95.78%        |
| MobileNet V2     | Freezing        | YCbCr      | 97.73%        | 0.0593        | 98.41%        |
| MobileNet V2     | Fine-Tuning     | YCbCr      | 98.13%        | 0.1443        | 98.11%        |
| ResNet 50        | Freezing        | RGB        | 86.45%        | 0.2891        | 88.66%        |
| ResNet 50        | Fine-Tuning     | RGB        | 85.96%        | 0.8122        | 85.42%        |
| <b>ResNet 50</b> | <b>Freezing</b> | <b>HSV</b> | <b>99.51%</b> | <b>0.0131</b> | <b>99.51%</b> |
| ResNet 50        | Fine-Tuning     | HSV        | 99.01%        | 0.0450        | 99.01%        |
| ResNet 50        | Freezing        | YCbCr      | 88.18%        | 0.3372        | 87.37%        |
| ResNet 50        | Fine-Tuning     | YCbCr      | -             | -             | -             |

The analysis of RGB to HSV or YCbCr preprocessing reveals significant variations in model performance. The most notable improvement occurs with ResNet50 in the freezing phase, where HSV preprocessing increases the F1 Score from 88.66% (RGB) to 99.51% (HSV), representing a substantial 10.85% performance boost. This indicates that HSV preprocessing enhances the model's ability to detect key visual patterns in pornographic images.

However, in lighter models such as VGG16 and MobileNetV2, HSV preprocessing results in performance drops ranging from approximately 2.7% to 8.32% when compared to RGB inputs. A performance change of even 2–3% in automated filtering systems can be critical, especially in minimizing false positives and false negatives. In the context of content moderation systems like “Ayo Pramuka,” this improvement translates into fewer misclassifications, helping maintain a safer digital environment for young users.

Feeding images in HSV or YCbCr formats may lead to a feature mismatch, particularly in the initial layers, thereby reducing sensitivity to important visual cues, especially if fine-tuning is limited or shallow.

In VGG16 and MobileNetV2, the performance with YCbCr preprocessing also decreases slightly (0.6% to 2%) relative to RGB. Notably, ResNet50 suffers a performance drop

from 88.66% to 87.37% when using YCbCr preprocessing, suggesting this color transformation may not be suitable for deep models in this context.

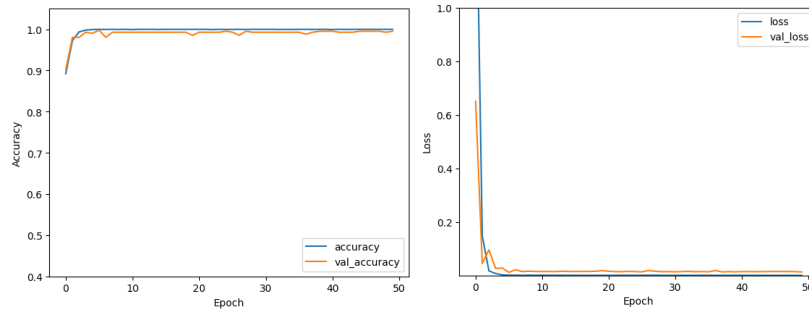


Figure 7. Accuracy and loss curves of the ResNet50 + HSV model during training and validation

Based on the performance illustrated in Figure 7, the combination of the ResNet50 model with HSV color space transformation demonstrated the most superior results during conventional training and validation. This is evidenced by stable and converging accuracy and loss curves, indicating that the model consistently recognizes visual patterns throughout the learning process.

The evaluation results are obtained using the confusion matrix, which contains the model training results in testing, obtained with TP = 203, FP = 0, FN = 2, and TN = 201, as shown in Figure 8.

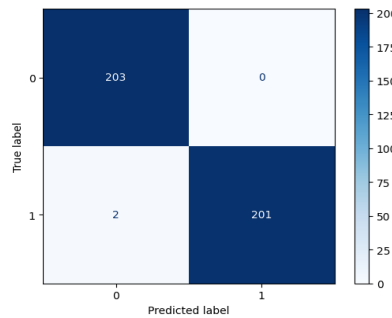


Figure 8. Confusion Matrix of the Best Model

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{203 + 201}{203 + 201 + 0 + 2} = 0.9951 = 99.51\% \quad (8)$$

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{203}{203 + 0} = 0.10 \\ &= 100\% \end{aligned} \quad (9)$$

$$\begin{aligned} \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{203}{203 + 2} = 0.9902 \\ &= 99.02\% \end{aligned} \quad (10)$$

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot 0.1000 \cdot 0.9902}{0.1000 + 0.9902} = 0.9951 = 99.51\% \quad (11)$$

Efficiency-wise, HSV preprocessing consistently reduces training and testing times across all models. For example, in the ResNet50 Freezing Layer, the training time decreases from 1618.2 ms (RGB) to 1513.8 ms (HSV), an improvement of approximately 6.4%. Similarly, MobileNetV2 (Freezing phase) shows a 16.7% reduction in processing time when using HSV. Meanwhile, YCbCr preprocessing shows only marginal improvements in efficiency without significant gains in detection accuracy.

#### 4.1 5-Fold Cross-Validation Evaluation

To assess the model's generalization capability and to reduce dependency on a single data split, experiments were conducted using 5-fold cross-validation. Each fold represents a different test data subset. This technique partitions the dataset into five balanced subsets, in which training and testing are alternated across each fold.

Table 7. presents the results of the cross-validation using the ResNet50 model combined with HSV preprocessing. The model consistently achieved high performance, with accuracy scores ranging from 96.21% to 97.76% and loss values ranging from 0.0747 to 0.0991. The mean accuracy reached 96.92% and the mean loss was 0.0865, indicating stable performance across various data partitions.

Table 7. 5-Fold Cross Validation Results (ResNet50 + HSV)

| Fold        | Training Accuracy % | Validation Accuracy % | Validation Loss |
|-------------|---------------------|-----------------------|-----------------|
| 1           | 98.25%              | 96.20%                | 0.0991          |
| 2           | 98.60%              | 97.77%                | 0.0747          |
| 3           | 97.94%              | 96.46%                | 0.0891          |
| 4           | 98.10%              | 96.86%                | 0.0826          |
| 5           | 98.45%              | 97.34%                | 0.0811          |
| <b>Mean</b> | <b>98.27%</b>       | <b>96.92%</b>         | <b>0.0865</b>   |

Results in Table 7. demonstrate that the ResNet50 + HSV model maintained consistently high performance across all folds. Training accuracy ranged narrowly between 97.94% and 98.60%, showing model stability during learning. Meanwhile, validation accuracy remained in the range of 96.20% to 97.77%, and validation loss ranged between 0.0747 and 0.0991. The small standard deviation across these folds suggests that the model did not rely on specific data partitions but instead generalized well across varied data subsets. These findings confirm that the use of HSV preprocessing combined with ResNet50, and the freezing method is robust and resilient even under different training-validation splits.

The visual representation of this performance consistency is shown in Figure 9, which illustrates the accuracy and loss trends for each fold of the cross-validation.

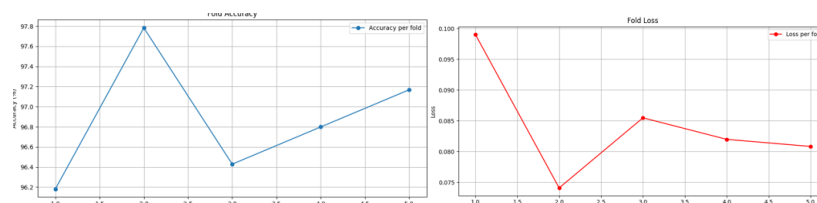


Figure 9. Accuracy and Loss of Each Fold in 5-Fold Cross Validation (ResNet50 + HSV)

The graph above shows that the trained model performed well during training. Training accuracy reaches 100% at the 6th epoch and stays until the end of the training, while validation accuracy stabilizes in the range of 99.26% to 99.51% after the 4th epoch. Training loss started to decrease significantly, approaching zero. In contrast, validation loss reached a minimum value of 0.0261 at the 4th epoch and remained stable from 0.01 to 0.02 until the end of training. This stability shows the strong generalization ability of the model, with no indication of overfitting despite the training accuracy reaching 100%. The model also showed rapid convergence, with optimal performance achieved within the first few epochs, especially at the 4th epoch.

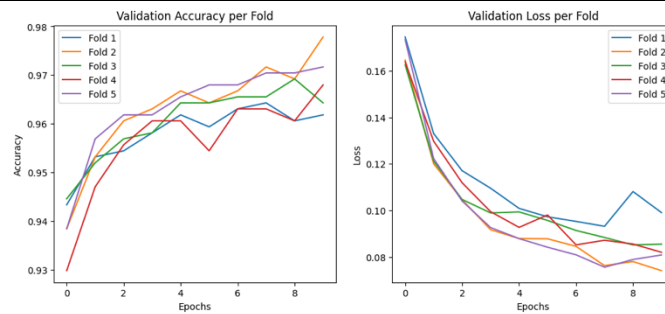


Figure 10. Validation Accuracy and Loss per Epoch for Each Fold in 5-Fold Cross Validation (ResNet50 + HSV)

As shown in Figure 10, the model's validation accuracy increased steadily across epochs in all folds, stabilizing above 96% by the final epoch. Similarly, validation loss decreased consistently, with most folds converging below 0.09. These trends further confirm the model's consistent learning behavior and ability to generalize well across different subsets of the data.

The following presents the training and testing time evaluation results of three deep learning models, VGG16, MobileNet V2, and ResNet 50. Each model was tested using two training methods, Freezing and Fine-Tuning, and different types of image data representations with three color spaces, RGB, HSV, and YCbCr. This table analyzes the time efficiency required by each model based on the training method and the type of data used. The results obtained are expected to significantly contribute to selecting the optimal model for the Let's Scout application in detecting pornographic content, especially in computational time efficiency. Available in Table 8.

Table 8. Result of experiment Time of each model

| Model               | Methods         | Data Image | Training time (ms) | Testing time (ms) | Total time (ms) |
|---------------------|-----------------|------------|--------------------|-------------------|-----------------|
| VGG16               | Freezing        | RGB        | 2740.5             | 304.5             | 3045            |
| VGG16               | Fine-Tuning     | RGB        | 2975.4             | 330.6             | 3306            |
| VGG16               | Freezing        | HSV        | 2662.2             | 295.8             | 2958            |
| VGG16               | Fine-Tuning     | HSV        | -                  | -                 | -               |
| VGG16               | Freezing        | YCbCr      | 2818.8             | 313.2             | 3132            |
| VGG16               | Fine-Tuning     | YCbCr      | 2844.9             | 316.1             | 3161            |
| MobileNet V2        | Freezing        | RGB        | 783                | 87                | 870             |
| MobileNet V2        | Fine-Tuning     | RGB        | 809.1              | 89.9              | 899             |
| <b>MobileNet V2</b> | <b>Freezing</b> | <b>HSV</b> | <b>652.5</b>       | <b>72.5</b>       | <b>725</b>      |
| MobileNet V2        | Fine-Tuning     | HSV        | 704.7              | 78.3              | 783             |
| MobileNet V2        | Freezing        | YCbCr      | 783                | 87                | 870             |
| MobileNet V2        | Fine-Tuning     | YCbCr      | 809.1              | 89.9              | 899             |
| ResNet 50           | Freezing        | RGB        | 1618.2             | 179.8             | 1798            |
| ResNet 50           | Fine-Tuning     | RGB        | 2009.7             | 223.3             | 2233            |
| ResNet 50           | Freezing        | HSV        | 1513.8             | 168.2             | 1682            |
| ResNet 50           | Fine-Tuning     | HSV        | 1905.3             | 211.7             | 2117            |
| ResNet 50           | Freezing        | YCbCr      | 1598.4             | 177.6             | 1776            |
| ResNet 50           | Fine-Tuning     | YCbCr      | -                  | -                 | -               |

Table 8 shows that MobileNet V2 is the most efficient model using the Freezing method, with the lowest total time of 725ms with HSV data. At the same time, VGG16 has the slowest time, with a total time of 9088ms using HSV color space, the Freezing method, which indicates that this model requires more computation, and ResNet 50 is in between, with the

lowest total time of 1682ms with HSV color space using the Freezing method. HSV provides a more significant processing time improvement than RGB, especially in more complex models such as ResNet50 and MobileNet V2. HSV data provides faster training and testing times, especially on MobileNet V2. Therefore, MobileNet V2 is the best choice for applications that require time efficiency, especially for pornography detection with HSV input data.

## 5. CONCLUSION

This study investigated the influence of RGB, HSV, and YCbCr color space transformations on transfer learning models (VGG16, MobileNetV2, and ResNet50) for pornographic image classification. The experimental results demonstrated that HSV preprocessing substantially improved performance, with the best configuration achieved by ResNet50 using the freezing method, reaching an accuracy of 99.51%. MobileNetV2 also delivered competitive results in resource-constrained environments, while HSV transformation generally enhanced robustness under varying lighting conditions.

Further evaluation using 5-Fold Cross Validation confirmed the generalization capability of the ResNet50 + HSV configuration, yielding a mean accuracy of 96.92% and an average loss of 0.0865. These findings indicate stable performance across different data partitions, with no significant evidence of overfitting despite the limited dataset size.

However, several limitations should be acknowledged. The combinations of VGG16 + HSV and ResNet50 + YCbCr consistently failed to produce valid classifications, often resulting in false positives. This limitation is likely caused by a mismatch between non-RGB inputs and pretrained convolutional filters optimized for RGB-based ImageNet features. When fine-tuning does not extend to the early layers, the model struggles to adapt to altered input distributions, leading to ineffective feature extraction. In addition, the dataset used in this study was relatively small and lacked demographic metadata such as ethnicity, skin tone, or cultural background, which may introduce algorithmic bias and reduce the generalizability of the findings. While the dataset was filtered according to Indonesia's legal definition of pornography to ensure normative relevance, the absence of visual diversity remains an open limitation.

In conclusion, this study highlights that integrating color space transformations with CNN-based transfer learning can enhance the accuracy and efficiency of pornography detection systems. Implementing this approach in the “*Ayo Pramuka*” application has the potential to strengthen automated content moderation, thereby contributing to a safer digital environment for youth and supporting the educational mission of *Gerakan Pramuka* (Kwarnas).

## REFERENCES

- [1] Kwarnas. (2023a). *Dasa Karya Gerakan Pramuka*. <https://pramuka.or.id/dasa-karya/>
- [2] Kwarnas. (2023b). *Visi, Misi, dan Tujuan Gerakan Pramuka*. <https://pramuka.or.id/visi-misi-dan-tujuan/>
- [3] Nasrullah, R. (2016). *Media Sosial: Perspektif Komunikasi, Budaya, dan Sosioteknologi*. Simbiosis Rekatama Media.
- [4] Yutifa, U., Santoso, H. B., & Hasibuan, Z. A. (2015). Analisis Penggunaan Media Sosial dalam Penyebaran
- [5] Ahmad, S. Z., Ahmad, N., & Bakar, A. R. A. (2019). Reflections of Entrepreneurs of Small and Medium-Sized Enterprises Concerning the Adoption of Social Media and Its Impact on Performance Outcomes: Evidence from the UAE. *Telematics and Informatics*, 36, 1–17. <https://doi.org/10.1016/j.tele.2018.11.007>
- [6] Karamizadeh, S., Chaeikar, S. S., & Jolfaei, A. (2023). Adult Content Image Recognition by Boltzmann Machine Limited and Deep Learning. *Evolutionary Intelligence*, 16, 1185–1195. <https://doi.org/10.1007/s12065-022-00729-8>
- [7] Setiono, K. (2021). Deteksi Citra Pornografi Memanfaatkan Deep Convolutional Neural Network. *Journal of Intelligent Systems and Computation*, 3(1), 1–6.
- [8] Liantoni, F., & Wahyudi, Y. R. (2022). Deteksi Citra Pornografi Menggunakan Skin Color Detection dan Eigenporn Berdasarkan Ruang Warna YCbCr. *Jurnal MNEMONIC*, 5(2), 145–150.

- [9] Rahman, A., Tarigan, T., & Sari, R. M. (2022). HSV Color Model and Logical Filter for Human Skin Detection. *AIP Conference Proceedings*, 2457, 040003. <https://doi.org/10.1063/5.0098747>
- [10] Jannah, A., & Pratiwi, N. (2025). Deteksi Pornografi pada Citra Karakter Animasi dengan HVS dan YCBCR Menggunakan Naive Bayes. *Jurnal Ilmu Penelitian Dan Pembelajaran Informatika*, 10(1), 337–344.
- [11] Kolkur, S., Kalbande, D., Shimpi, P., Bapat, C., & Jatakia, J. (2017). Human Skin Detection Using RGB, HSV, and YCbCr Color Models. *ArXiv Preprint*. <https://arxiv.org/abs/1708.02694>
- [12] Albahli, S. (2023). Transfer Learning for Pornographic Image Detection Using CNN Models. *Computer Systems Science & Engineering*, 43(2), 701–717.
- [13] Aldahoul, N., Omar, M. K., Mohamed, A., Hussain, F., & Rahman, A. (2021a). An Evaluation of Traditional and CNN-Based Feature Descriptors for Cartoon Pornography Detection. *IEEE Access*, 9, 1–16. <https://doi.org/10.1109/ACCESS.2021.3064920>
- [14] Aldahoul, N., Omar, M. K., Mohamed, A., Hussain, F., & Rahman, A. (2021b). Transfer Detection of YOLO to Focus CNN's Attention on Nude Regions for Adult Content Detection. *Symmetry*, 13(1), 1–26. <https://doi.org/10.3390/sym13010020>
- [15] Chen, J., Wu, Z., Zhou, Z., Li, Z., & Liu, X. (2020). A Pornographic Images Recognition Model based on Deep One-Class Classification With Visual Attention Mechanism. *IEEE Access*, 8, 1–13. <https://doi.org/10.1109/ACCESS.2020.2977463>
- [16] Hor, S. L., Phung, S. L., & Bouzerdoum, A. (2022). Deep Active Learning for Pornography Recognition Using ResNet. *International Journal of Technology*, 13(6), 1261–1270.
- [17] Mazinani, M. R., Hoseini, S. M., & Ahmadi, K. D. (2020). Time-Sensitive Adaptive Model for Adult Image Classification. *Computing and Informatics*, 39(1), 1282–1310.
- [18] Moreira, D. C., & Fechine, J. M. (2018). A Machine Learning-based Forensic Discriminator of Pornographic and Bikini Images. *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 1–8. <https://doi.org/10.1109/IJCNN.2018.8489335>
- [19] Nashrullah, F., Wibowo, S. A., & Budiman, G. (2020). Investigasi Parameter Epoch pada Arsitektur ResNet-50 untuk Klasifikasi Pornografi. *Journal of Computer, Electronic, and Telecommunication*, 1(1), 1–8.
- [20] Nikam, A. V. (2018). Locating and Detecting Nipple for Pornographic Image Identification. *Journal of Computer Sciences and Engineering*, 6(7), 1343–1347.
- [21] Phan, D. D., Nguyen, T. A., Le, Q. M., Tran, B. M., & Nguyen, V. V. (2021). A Novel Pornographic Visual Content Classifier based on Sensitive Object Detection. *International Journal of Advanced Computer Science and Applications*, 12(5), 787–797. <https://doi.org/10.14569/IJACSA.2021.0120595>
- [22] Sharma, R. (2020). Using Transfer Learning to Classify Pornographic Images. *Proceedings of the Conference on Data Science and Machine Learning Applications (CDMA)*, 43–46. <https://doi.org/10.1109/CDMA47397.2020.00013>
- [23] Surinta, O., & Khamket, T. (2019). Recognizing Pornographic Images using Deep Convolutional Neural Networks. *Proceedings of the International Conference on Digital Arts, Media, and Technology*, 150–154. <https://doi.org/10.1109/ICDAMT.2019.8713477>
- [24] Patterson, J., & Gibson, A. (2017). *Deep Learning: A Practitioner's Approach*. O'Reilly Media.
- [25] Lin, X., Qin, F., Peng, Y., & Shao, Y. (2021). Fine-Grained Pornographic Image Recognition with Multiple Feature Fusion Transfer Learning. *International Journal of Machine Learning and Cybernetics*, 12(1), 73–86. <https://doi.org/10.1007/s13042-020-01223-5>

- [26] Naji, S., Jalab, H. A., & Kareem, S. A. (2020). A Survey on Skin Detection in Colored Images. *Artificial Intelligence Review*, 52(2), 1041–1087. <https://doi.org/10.1007/s10462-018-9644-y>
- [27] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 25, 1097–1105.
- [28] Ravanbakhsh, E., Rezaei, M., Namjoo, E., & Choobdar, P. (2022). Comparison between EM and FCM algorithms in skin tone extraction. *ArXiv Preprint*. <https://arxiv.org/abs/2211.09979>
- [29] Arya, S. J., Asish, A., Shine, B. S. F., Sreelakshmi, J. L., & Varghese, E. (2021). Skin Detection Using YCbCr Colour Space for UAV-Based Disaster Management. In *Innovations in Computer Science and Engineering* (pp. 633–641). Springer. [https://doi.org/10.1007/978-981-33-4543-0\\_67](https://doi.org/10.1007/978-981-33-4543-0_67)