

A convolutional neural network framework for classifying inappropriate online video contents

Tanatorn Tanantong, Patcharajak Yongwattana

Thammasat Research Unit in Data Innovation and Artificial Intelligence, Department of Computer Science,
Faculty of Science and Technology, Thammasat University, Khlong Luang, Thailand

Article Info

Article history:

Received Jan 31, 2022

Revised Jul 22, 2022

Accepted Aug 20, 2022

Keywords:

Convolution neural network

Deep learning

Pre-trained model

Transfer learning

Video content classification

ABSTRACT

In the digital world, the Internet and online media especially video media are convenient and easy to access. It leads to problems of inappropriate content media consumption among children and youths. However, measures or methods to control the inappropriate content for children and young people are still a challenge for management. In this research, an automated model was developed and presented to classify the content on online video media using a deep learning technique namely convolution neural networks (CNN). For data collection and preparation, the researchers collected video clips from movies and television (TV) series from websites that distribute the clips online. It consists of different types of content: i) sexually inappropriate content; ii) violently inappropriate content; and iii) general content. The collected video clip data was then extracted into frames and then used for developing the automatically-content-classifying model with algorithm CNN, analyzing and comparing the result of CNN model performance. For enhancing the model performance, a transfer learning approach and different regularization techniques were adopted in order to find the most suitable method to create high-performance modeling to classify content in video clips, movies and TV series published online.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Tanatorn Tanantong

Thammasat Research Unit in Data Innovation and Artificial Intelligence,

Department of Computer Science, Faculty of Science and Technology, Thammasat University

99 Phaholyothin Road, Khlong Luang, Pathum Thani, 12121, Thailand

Email: tanatorn@sci.tu.ac.th

1. INTRODUCTION

Since the number of internet users is increasing every year, and video viewing has become increasingly popular in the online world such as educational video media, social video media, news video media, economic video media and entertainment video media and so on, a large amount of video media has been distributed online. This includes video material that contains various forms of inappropriate content such as sexually explicit and violent content [1]. Distribution of these videos may lead to imitation of inappropriate behaviors among viewers. Therefore, the development of effective methods to control inappropriate content on online video media is necessary especially for children and young people who need special consideration for the suitability online video materials [2]. At present, deep learning (DL) is a machine learning (ML) method which is highly effective in applying image classification [3]–[6]. One of the most popular and effective DL algorithms for complex data classification is the convolution neural network (CNN) [7]–[11]. Ali and Senan [12] has collected the research data related to the application of the deep learning CNN algorithm in classifying the video clips inappropriate for young adults. Karpathy *et al.* [13] studied and applied the CNN along with considering the location and time factor data (spatio-temporal information) to be used in the development of models for large-scale video clip

classification on YouTube. The efficacy of the model was tested with the YouTube sports-1M dataset containing 1 million video clips, 487 categories. In the study [14], the CNN model built with an AlexNet architecture [15], a GoogLeNet architecture [16], and a combination of AlexNet and GoogLeNet architectures were compared for their efficiency. Its objective was to develop and create a model for classifying sexually explicit image frames. The method was tested on the normalized difference phenology index (NDPI) dataset [17] containing videos in which the content consists of sexually explicit clips and generic content clips. Research by Khan *et al.* [18] has developed an automated system for analyzing and removing inappropriate content on video clips. In research [19], a CNN model using a MobileNet architecture has been developed to identify violently inappropriate image frames in video clips. The model has been built and tested on various datasets: i) the crowd violence dataset [20]; ii) the hockey fight dataset [21]; and iii) the violent scene detection dataset (VSD) [22]. For data preparation in CNN modeling, the research presents key-frame selection method by considering selecting from complete data and able to see objects in images clearly. In research Lin *et al.* [23] presented a CNN approach using multiple models to classify sexually explicit images to enhance extraction and image cognitive abilities. Regarding the experiment with the proposed method with a standard dataset published online for research not safe for work (NSFW) dataset [24], Zhou *et al.* [25] developed the application of CNN technology in conjunction with analysis of the object movement within the video clip by considering the still images and frame animation in video (optical flow and acceleration field vector) in order to detect violently inappropriate content. Perez *et al.* [26] proposed a CNN model using still image data and motion data of objects within a video clip (optical flow vector data and mpeg motion vector) to detect sexually explicit videos. In addition to the application of CNN models in two-dimensional (2D) for the video classification, in the above-mentioned researches, Accattoli *et al.* [27] presented the application of three-dimensional (3D) convolutional neural networks to increase efficiency to classify images of violently inappropriate video, such as fighting scenes by experimenting with methods with three public datasets containing violent content; the hockey fight dataset, the crowd violence dataset, and the movie violence dataset [20].

From the research review, it has been found that the application and development of CNN models are popular and effective for categorizing video clips from various sources. However, studies and researches on improving the efficiency of CNN models are an interesting topic and essential when applying the developed models to real-world situations where there might be potential limitations, which can reduce the model's performance. The example of researches developing and applying various techniques to improve the efficiency of the CNN model, is described by [28]–[31]. He *et al.* [28] conducted on the optimization of the learning model by means of preparing and improving the image data before data processing with various techniques: i) normalization, the technique for improving the pixel data of the image to the specified scaling range by improving the pixel data of the red, green, blue (RGB) mode image with the color gamut from 0-255 to become the data with the value between 0-1; ii) data augmentation (DA) technique which consists of random rotation, color jitter, and random crop; and iii) principal component analysis (PCA). Zhang *et al.* [29] examined the results of experimental model performance improvement techniques (regularization) with the CNN model using L2 regularization (L2) weight decay, DA, and drop-out. The results showed that the technique regularization aforementioned can help improve and optimize the accuracy outcomes for learning models by 4-5%. Kasche and Nordström [30] studied the techniques to improve model performance (regularization) for solving the problems of neural networks overfitting. The experiments were conducted to compare the efficiency of various techniques, including an early stopping (ES) technique, a L1 regularization technique, a L2 technique, and a dropout technique, which randomly closes nodes in the neural network. The results showed that the dropout model optimization technique with maximized the highest accuracy outcomes for the learning model. In research [31], a method to improve the efficiency of image classification with convolutional neural networks was presented using the technique of adding images (DA) consisting of random adjustment of color jitter, rotation, reduction or enlargement (zoom), flipping the image. The results were found that it can increase the image recognition efficiency by 3% higher.

To improve the performance of traditional CNN models, transfer learning is one of the effective approach [32] that can solve a problem for building CNN models when there is a limited amount of training data. In transfer learning, the knowledge of existing trained models is transferred to build new classification models. There are several research studies that investigated on classification performance when using transfer learning, e.g., sports video classification using pre-trained neural network [33], classification of lung cancer based on CNN and transfer learning with GoogLeNet [34], fruit recognition using pre-trained models [35], and leaf disease classification based on a pre-trained model [36]. In research Ramesh and Mahesh [33], reported the comparison performance results between a CNN model and a pre-trained CNN model for classifying sports categories from video collected from YouTube. Two examples of pre-trained model, i.e., Alexnet and Googlenet, were employed to build the pre-trained CNN model. The experimental results demonstrated that the pre-trained CNN model performed higher accuracy than the CNN model, with the accuracy of 92.68% and 88.75%, respectively. AL-Huseiny and Sajit [34] developed a CNN classification model using transfer learning with GoogLeNet for detecting lung cancer. The model was trained and validated by using computed tomography (CT) scan images. For obtained classification results, the CNN model using transfer learning,

GoogLeNet, yielded the higher performance with accuracy of 94.38%, compared to the original CNN model evaluated with the same dataset which scored 89.88%. In order to find a suitable transfer learning architecture for the CNN-based classification in real-world application, Tan and Le [37] presented a systematically study on the CNN model development under a limitation of computational resources. Experiments on scaling pre-trained CNN models and number of parameters (e.g., network depth, width and resolution) were evaluated with ImageNet datasets for optimizing the model performance. Based on the evaluated results, a new scaling method, namely, EfficientNets-B7, yield the highest accuracy of 84.3%. The results showed that the proposed EfficientNet model can be effectively applied to develop pre-trained CNN classification with optimized resource utilization. Duong *et al.* [35] applied a transfer learning approach for fruit classification. The pre-trained models, EfficientNet and MixNet, were adopted for building the CNN model and evaluating with a fruit dataset (48,905 images). Based on experimental results, EfficientNet-B1 outperformed the other configurations with precision, recall, and F1 score of 95%, 95% and 95%, respectively. In addition, the MixNet-Small method yield the best result for MixNet configurations with precision, recall, and F1 score of 94%, 94% and 93%, respectively. The evaluated results illustrated that adoption of EfficientNet and MixNet with CNN models can effectively be improved the performance of the fruit classification. Atila *et al.* [36] developed pre-trained CNN models to classify leaf diseases. EfficientNet and state-of-the-art pre-trained CNN models, e.g., AlexNet, residual network (ResNet) 50, visual geometry group (VGG) 16 and Inception V3., were employed for building the classification models of plant leaf diseases. The experimental results obtained demonstrated that EfficientNet-B4 and EfficientNet-B5 models gave the top performance compared to other models with accuracy of 99.91% and 99.97%, respectively. As reported from these research studies, the transfer learning method can effectively be applied to enhance the performance of the CNN model for classifying various types and resolutions of images.

Based on the existing studies above, this research purposes to study and apply CNN for reviewing and categorizing the content of online video clips by referring to the criteria for rating the appropriateness of Thai television programs according to the announcement of national broadcasting and telecommunications commission (NBTC) [38]. The process to develop a model for reviewing and classifying content in video clips is: i) collection of data from YouTube and various websites that publish video clips in movies and TV series. Then, the video clip data is converted into still image frame data; ii) generating a solution for categorizing the content of a video clip based on the extracted still image frame data divided into sexually explicit content, violently inappropriate content and general content; iii) modeling by using still image frame data obtained from the data preparation process to create a deep learning model with CNN algorithm; iv) performance improvement of the CNN model derived from the previous step by using techniques to improve the efficiency of the model; and v) measuring the performance of the standard CNN model and its comparison with the enhanced CNN models with other techniques. However, for improving the performance of the obtained CNN classification model, a transfer learning method and model optimization techniques were also applied in this study.

This research paper is organized as: section 2 describes the proposed framework and methodologies. Section 3 describes the improvement of the model performance. Section 4 presents the model evaluation and experimental results. Section 5 provides conclusions.

2. METHOD

A framework for developing an automated model for video content classification is presented in this research. It started with collecting online video clips and converting them into image frame datasets. The obtained image frames were annotated with three video content categories, i.e., sexually explicit content, violently inappropriate content and general content, and were extracted features for developing a CNN model. The CNN model performance was improved using model optimization techniques (DA, L2 and ES). Finally, all image frames were automatically classified into three categories (sexually explicit content, violently inappropriate content and general content) based on the developed CNN model. The process of developing an automatic video content classification model can be summarized as shown in Figure 1.

2.1. Data collection and preparation

Data collection and preparation is an important starting step for the development of a machine learning model because the efficiency of the model depends on the quality and suitability of the data used in modeling processing. The first step of this research starts with collecting video clips of movies and TV series from YouTube and video clip websites. In the next step, the researcher analyzes and considers the content of the video clip according to the criteria for rating the appropriateness of Thai television programs according to the announcement of the broadcasting committee, The NBTC [38]. The content of the video clips analyzed in accordance with the above criteria was then used for the development of the CNN model and the model's efficiency was tested. To develop a researcher model, 42 minutes of sexually explicit content, 42 minutes of

violently inappropriate content, and 53 minutes of general content were used. For model performance testing, 10 minutes of sexually explicit content, 10 minutes of violently inappropriate videos, and 10 minutes of general content were employed.

The next step is to prepare the data where the researchers took the video content, processed and extracted the data into image frames. The extracted image frames represented the data in a 2second video clip, and the image frames were resized to the same standard size of 224×224 pixels. Then the obtained data were divided into 3 groups: i) sexually inappropriate content; ii) violently inappropriate content; and iii) general content. To categorize the image frame data into different groups, the frames in which the elements can be seen clearly were chosen. Then, they were grouped according to the criteria for rating the appropriateness of Thai television programs. The contents of each group are illustrated in Figures 2 to 4, respectively.

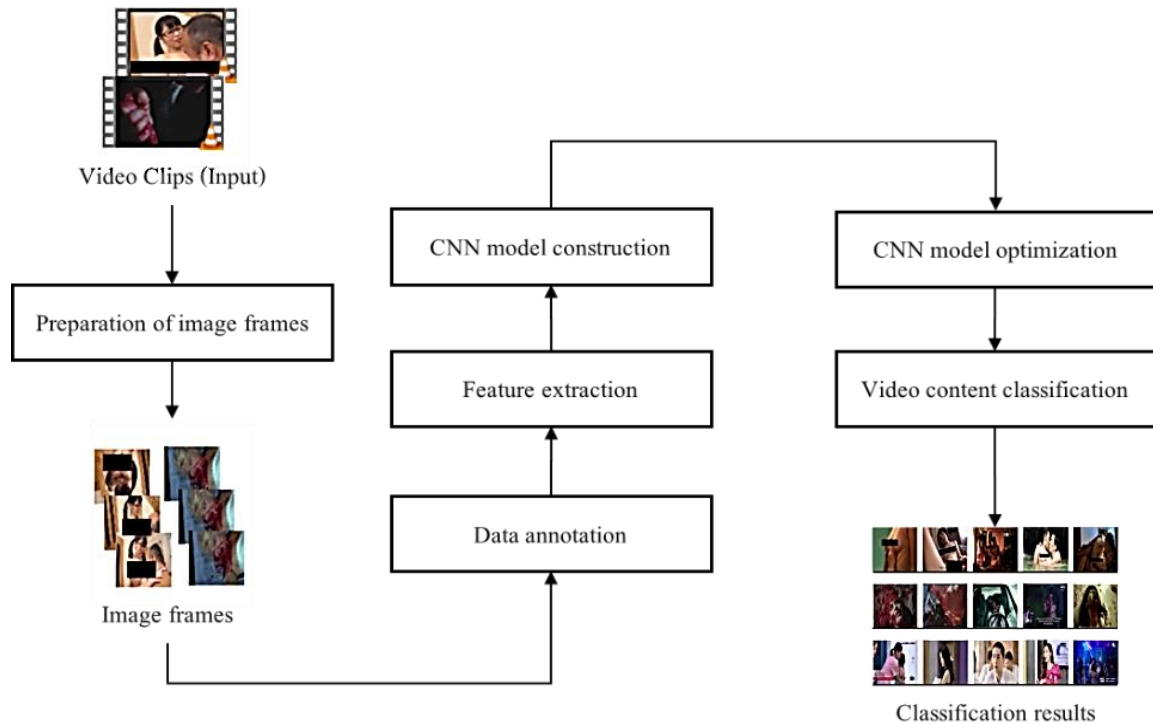


Figure 1. A framework for video content classification



Figure 2. An example of sexually explicit image frame data published on a website



Figure 3. An example of violently inappropriate image frame data published on a website



Figure 4. An example of generic image frame data published on a website

2.2. CNN model

In this research, the researcher has chosen CNN to create a model for classifying the content of a video clip. CNN is a deep learning approach based on simulations of human vision that can classify objects by considering features of the objects. The components of CNN consist of: i) convolution layer, the layer that serves as feature extraction of various elements in the image, such as the nature of the lines, color and color contrast; and ii) fully-connected (FC) layer that serves to learn and remember various features of the object in the image and consider categorizing the images according to the data used to create the model [39]. Regarding the model development step, the researcher selected the CNN modeling architecture presented in the research study [40] which consists of six layers of convolutional layer (Conv) and three layers of FC layer. The model receives the input as color images RGB sized 224×224 pixels as shown in Figure 5.

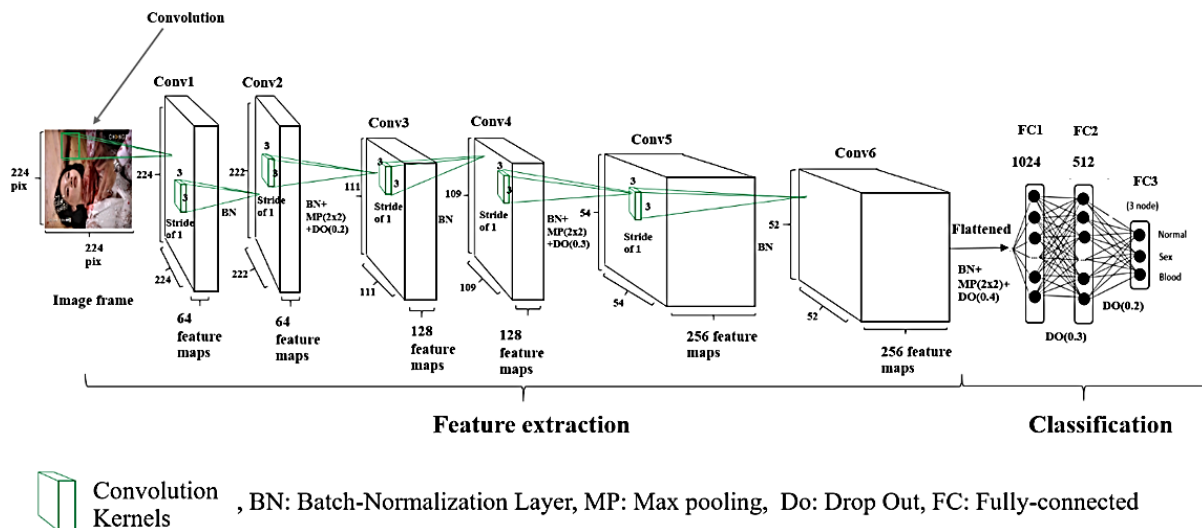


Figure 5. A CNN architecture

For the model development process, it started from Conv1 where the edges of the image frames (padding) were improved and expanded by 2 pixels per side, resulting in a new frame size of 226×226 pixels. Kernels sized (3×3 Pixel) were used as filters. The results of the feature extraction in the Conv1 were 64 sets of the feature map of 224×224 pixel images. The result was then be adjusted to be in an appropriate format with batch normalization technique, which standardizes the data using mean and standard deviation. Afterward, the data were passed to the next layer, Conv2.

In the Conv2 layer, the data from layer 1 (Conv1) were extracted from the attributes with kernels attribute filter (3×3 Pixel). The result is a total of 64 attribute maps sized 222×222 pixel. Characteristics were then adjusted into a proper form with a batch normalization technique. And the attribute data were then adjusted to be less complex with a max pooling technique, which is a down sampling method. Filtering the kernels pooling data of 2×2 pixel size, attribute data were reduced to 111×111 pixel and the dropout technique was used to make the neural network less complicated by randomly closing the node in the neural network with a dropout rate of 0.2. The results were then sent to the next layer, Conv3.

In the Conv3 step, feature extraction was applied as in the Conv1, but 128 kernel attribute filters (size equal to 3×3 pixels) were defined. The result of the extraction was 128 sets of feature data sized 111×111 Pixel. The datasets were adjusted into a proper format with Batch Normalization technique and the data were then forwarded to the next layer, Conv4. For the Conv4, the feature extraction process was the same as the Conv2

layer, and the attribute filtering process was the same as the Conv3 layer. The result of this Conv4 layer was the images sized 109×109 Pixel. Then the dataset was adjusted using batch normalization technique, max pooling technique, and randomly closed nodes in the neural network with a dropout rate of 0.3. This resulted in 128 attribute sets of 54×54 Pixel size. The data were then forwarded to the next layer, Conv5.

For the Conv5 layer, the feature extraction procedure was the same as the Conv1 layer, but 256 attribute filters were defined (sized equal to 3×3 pixels), which obtains a total of 256 attribute datasets of 54×54-pixel size to be then adjusted into the proper form with the batch normalization technique and the data were forwarded to Conv6. In the Conv6 layer, the same feature extraction step was the same as the Conv2 layer, but with 256 kernal attribute filters (sized equal to 3×3 pixels) and data manipulation using batch normalization, max pooling, and node shutdown techniques in a neural network with a dropout rate of 0.4. A total of 256 sets of 26×26-pixel attribute data were obtained.

After going through those 6 layers of convolution successfully, the next step was to take the resulting attribute data to convert into a one-dimensional (1D) vector format using a flattened technique and the 1D vector data were passed into the data classification process, which is a layer of the FC layer artificial neural network, with a total of 3 layers. The FC layer 1 (FC1) defined the total number of processing units as 1,024 nodes, and the Dropout technique was used with a value of 0.3. Next in the FC layer 2 (FC2) layer, the number of processing units was set as 512 nodes, and the dropout rate was 0.2. In the final layer, FC layer 3 (FC3), there were a total of 3 nodes in this layer. Each node was classified by classes of content: i) sexually inappropriate content; ii) violently inappropriate content; and iii) general content. The FC3 layer is called the output layer, which is the layer of CNN in which content classification results are restored by the defined classes.

3. CNN MODEL OPTIMIZATION

CNN model optimization is the optimization of the model before it is put into practice. It helps improve and prevent problems that occur during the model's learning process. In this study, the researcher has chosen various regularizations to improve the performance of the CNN model: i) DA technique [41] is a technique to create additional images by using the existing image or sample data to modify the image such as distortion, clipping and rotation. These techniques help solve the problem of insufficient images or data to teach the model. The more data the model has, the more opportunities to predict more accurate results; ii) L2 technique [42] is the process of adjusting the weight of the neural network. This technique simplifies the CNN model structure, reducing the likelihood of model overfitting; iii) the ES technique [43], [44] is a method for detecting overfitting problems during the learning of the CNN model taking into account the tolerance of the test data in each round of the model's learning. If the performance outcome trends are not improving, the CNN model then terminates the learning process and create a model as a result. In this study, the researchers selected the Python programming language to develop a CNN model using the core model development libraries, i.e., Keras and Tensorflow. The model was developed on the computer system using central processing unit (CPU): advanced micro devices (AMD) Ryzen5 5600X 6-core processor; graphics processing unit (GPU): Nvidia geforce rtx3080 and the main memory or random access memory (RAM) of 16 GB was used. To build a model for classifying the architecture-based video content is presented in Figure 5. The first step was to create a model using CNN learning techniques, and the next step was to improve CNN model performance with regularization techniques which consists of DA, L2 and ES as presented in Table 1.

Table 1. CNN modeling for classifying online video content with optimization techniques

Experiment setting	CNN	DA	L2	ES
3.1	✓	-	-	-
3.2	✓	✓	-	-
3.3	✓	-	✓	-
3.4	✓	-	-	✓
3.5	✓	✓	✓	✓

3.1. Constructing the model using CNN

To create a CNN model for categorizing the content of video clips, the first step started when the researchers extracted image frame data consisting of 3 classes of content: i) 1,268 image frames for sexually inappropriate content; ii) 1,245 image frames for vio-lently inappropriate content; and iii) 1,608 image frames of general content. A total of 4,121 image frames are used. The image frame data were then divided into a training dataset and a validation dataset by specifying the proportion of the data as 80% and 20% respectively. In this step, the initial values used in CNN modeling were defined as follows: 1. The number of image data for each model training cycle (batch size) was set as 32 images. 2. The number of modeling cycles (epoch) was

set as 120 cycles, and the algorithm for adjusting the weight and bias values in the model was set with the Adam optimization algorithm. 3. The model's learning rate (learning rate) was set as 0.01.

3.2. Improving the performance of CNN models with DA techniques

In this experiment, the performance of the CNN model was improved by using DA techniques, and new image data were added into the model's learning process to help solve the problems of insufficient amount of data. The images added to the model's learning process were the original images modified with randomly rotating the image back from left to right, and randomly skewing the images. The addition of new image data allows the model to learn in more diverse forms and gradually improves its ability to recognize images of various types.

3.3. Improving the performance of CNN models with L2

The constructed CNN model (in section 3.1) was used to optimize the model with the L2 technique, which was applied in both Conv and FC layer. As reported by Zhang *et al.* [29] the purpose of adapting the L2 is to simplify the CNN model and reduce the likelihood of model overfitting. In order to build the optimizing CNN model, in the experiment, the L2 factor was assigned to 0.01.

3.4. Improving CNN model performance with ES techniques

In this experiment, the ES technique was applied to improve the performance of the CNN model. The ES technique is another technique that can help reduce the likelihood of the overfitting problems when using CNN model. This technique was used in the model learning process. The loss function value was considered in each cycle of the model's learning curve (patience variable is set 10). If it is found that the trend of learning efficiency of the model in each cycle does not improve, the algorithm will then stop the learning of the CNN model in the cycle. To detect such a trend, this ES technique does not only prevent overfitting, but also reduces the processing time for building models.

3.5. Improving the performance of CNN models with DA, L2 and ES

The CNN model developed in section 3.1 was used to improve efficiency by using all 3 of the regularization techniques; i) the DA technique that solves the problem of insufficient amount of data for modeling; ii) the L2 technique that simplifies CNN models and reduces the likelihood of model overfitting; and iii) the ES technique that prevents overfitting and reduces model development time.

4. MODEL EVALUATION AND EXPERIMENTAL RESULTS

This section presents an evaluation of the performance and results of a video content classification model, which have been explained in section 3 consisting of: i) the CNN model development experiment; ii) the CNN model development experiment with DA technique (CNN model+DA); iii) the CNN model development experiment with L2 technique (CNN Model+L2); iv) the experimental CNN model development with ES (CNN Model+ES); and v) the CNN model experimental development with DA, L2 and ES (CNN Model+DA+L2+ES). The efficiency of the models was evaluated in various trials with the training dataset, the validation dataset and the test dataset as shown in Figure 6.

Regarding the measuring tool for CNN model efficiency, the accuracy, precision, recall, and harmonic mean (F1 score) were used. These values were obtained by processing and calculating the data in the confusion matrix, which consists of true positive (TP) and true negative (TN) data, which mean the amount of data could correctly predict that the image frames were considered inappropriate and appropriate respectively. False positive (FP) and false negative (FN) data mean the number of image frames were incorrectly predicted to be inappropriate and appropriate respectively. Various data were taken and calculated as percentage as in the following formula.

$$Accuracy (\%) = \frac{(TP+TN)}{(TP+TN+FP+FN)} \times 100 \quad (1)$$

$$Precision (\%) = \frac{TP}{(TP+FP)} \times 100 \quad (2)$$

$$Recall (\%) = \frac{TP}{(TP+FN)} \times 100 \quad (3)$$

$$F1 \text{ score } (\%) = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \times 100 \quad (4)$$

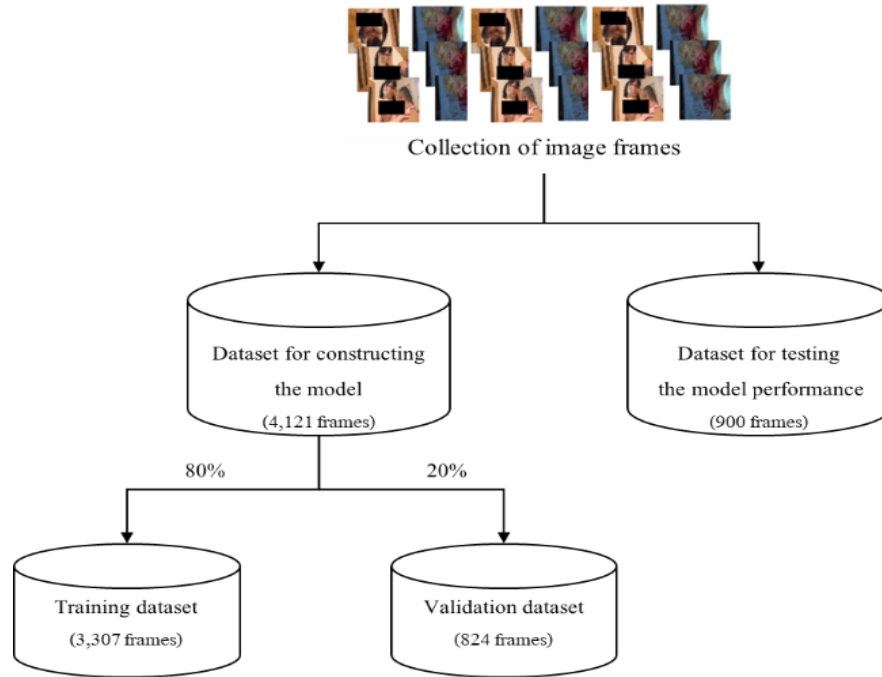


Figure 6. Data segmentation for constructing the CNN classification model and evaluating the performance of the classification model

4.1. CNN model performance results on training dataset and validation dataset

To test the performance of the CNN model in combination with various optimization techniques for the training dataset and the Validation dataset in each of the 5 experiments (CNN model, CNN model+DA, CNN model+L2, CNN model+ES and CNN model+DA+L2+ES), the model's performance was tested with 120 cycles of learning and processing (epoch) cycles of the model. The results of validity of each cycle for the training dataset and the validation dataset are revealed in Figures 7 and 8, respectively.



Figure 7. Results in each epoch of the CNN model evaluated with the training dataset

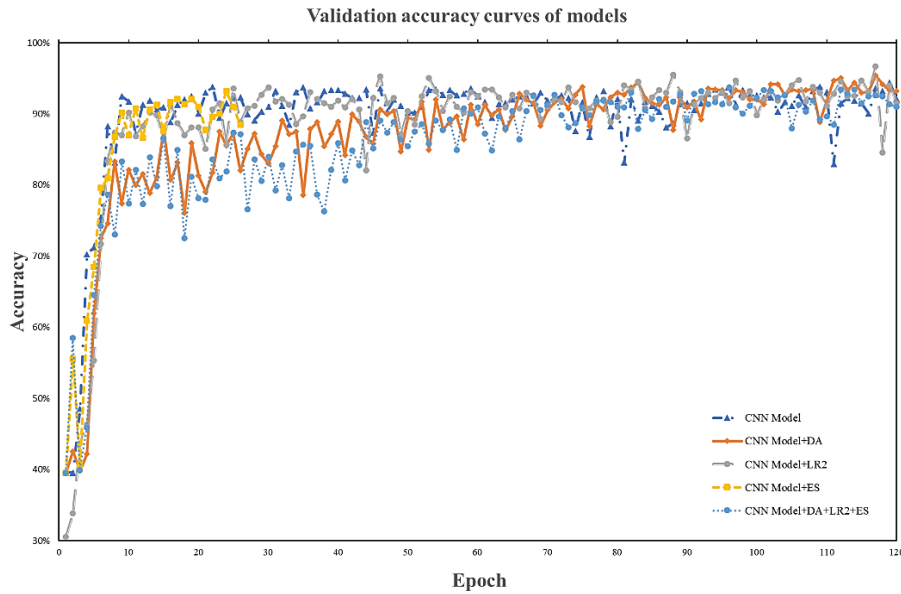


Figure 8. Results in each epoch of the CNN model evaluated with the validation dataset

Table 2 presents the mean accuracy results from the five model performance tests. It was found that the experiment in section 3.1 (CNN Model) gave the highest accuracy in the experiments. The accuracy values were 97.4% and 89.6% when tested with the training dataset and the validation dataset, respectively. In section 3.3 (CNN Model+L2) yielded the highest accuracy compared to other performance improvement trials. The accuracy values for the Training dataset and the Validation dataset were 95.7% and 89.1%, respectively. The ES experiment described in section 3.4 (CNN Model+ES) yielded the second highest validity results, being 93.2% and 83.1% for the training dataset and the validation dataset, respectively. However, the experiment according to section 3.4, the number of learning cycles was less than others. The other trial was 34 cycles, as the ES technique stops the learning process of the CNN model to reduce the likelihood of overfitting.

Table 2. Experimental results in CNN modeling for content classification of video clips

Experiments	Epoch	Avg. training acc. (%)	S.D. training acc.	Max training acc. (%)	Min training acc. (%)	Avg. validation acc. (%)	S.D. validation acc.	Max validation acc. (%)	Min validation acc. (%)	Time spent on creating models (seconds)
CNN	120	97.4	0.05	99.8	66.0	89.6	0.08	94.2	39.7	2,505
CNN+DA	120	88.2	0.07	96.1	59.2	86.3	0.10	94.9	39.7	2,479
CNN+L2	120	95.7	0.04	99.1	65.0	89.1	0.99	95.4	28.5	2,690
CNN+ES	34	93.2	0.07	98.7	66.1	83.1	0.14	93.6	38.3	724
CNN+DA+L2+ES	120	87.4	0.06	94.3	59.8	85.3	0.09	93.0	39.7	2,726

4.2. CNN model performance results on test dataset

In this stage, it is to test the performance of the CNN model with the datasets that are not used as model learning data in order to simulate the actual situation in which the model can be applied to other video clips on the websites. In the simulating test, the researcher used the data divided into test datasets to experiment with the five developed models (CNN model, CNN model+DA, CNN model+L2, CNN model+ES, and CNN model+DA+L2+ES). The test was conducted on video content data for all three categories, totaling 900 image frames. The data extracted from 300 frames consist of 300 frames of violent inappropriate video clips, 300 frames of sexually explicit video clips and 300 frames of generic content.

Table 3 details the test performance results of the various experiments with the test dataset. The data in this table was found that the results of the first experiment (3.1 CNN model) gave accuracy, precision, recall, and F1 score were 66%, 71%, 66% and 62%, respectively. For the use of optimization techniques in combination with CNN models, the fifth trial (3.5 CNN model+DA+L2+ES) gave the best performance compared to the other performance improvement experiments. It gave accuracy, precision, recall, and F1-score as 81%, 83%, 81% and 80% respectively.

Nevertheless, the optimization of the CNN model using the DA technique in the second experiment (3.2 CNN model+DA) yielded a second accuracy result. And because this technique helps improve and increase the amount of image data used to teach the model to be diverse, this allows the CNN model+DA to classify data from the never-before-seen test set better than the CNN model from other experiments that have not applied this technique. For the third experiment (3.3 CNN model+L2), the results showed that the L2 technique improves the predictive performance of the models better than those that do not use model optimization techniques (experiment 3.1) with an Accuracy greater than 3%, and in experiment 3.4 (CNN model+ES) using the ES technique it resulted in accuracy similar to that of the non-improved CNN model. However, in experiment 3.4, it took 3 times shorter time to learn and generate the CNN models than other experiments, being approximately 724 seconds, as shown in Table 3.

Table 3. CNN model performance results with test datasets

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
3.1 CNN	66	71	66	62
3.2 CNN+DA	75	80	76	75
3.3 CNN+L2	69	74	69	66
3.4 CNN+ES	65	66	65	61
3.5 CNN+DA+L2+ES	81	83	81	80

4.3. Improving CNN Model performance using transfer learning

According to a literature survey [32] a transfer learning method can currently be adopted to improve the performance of traditional machine learning methods by transferring information from a related domain. There are many applications that successfully applied transfer learning to enhance the model performance, e.g., sports video classification using pre-trained neural network [33], classification of lung cancer using pre-trained convolutional neural networks [34], automated fruit recognition using pre-trained models [35], and plant leaf disease classification using a pre-trained model [36]. As suggested in [37], for improving the CNN model with the transfer learning technique, a pre-trained model, namely EfficientNet-B0, which is suitable with the dataset used in this study (image size of 224x224 pixel), was used. To construct the pre-trained CNN model, CNN layers in EfficientNet-B0 were frozen and a last FC layer (an output layer) in EfficientNet-B0 was fixed to 3 classes, i.e., a sexually inappropriate class, a violently inappropriate class, and a general class. In order to improve the pre-trained CNN model performance, optimization techniques, DA, L2 and ES, were applied in experiments as detailed in Table 4.

Table 5 demonstrates the experimental results of the pre-trained CNN models with test datasets. The first experiment (5.1 pre-trained CNN) shows accuracy, precision, recall, and F1-score as 84%, 87%, 84%, and 84% respectively. To enhance the performance of the pre-trained CNN models, optimization techniques (DA, L2, and ES), were employed in experiments 5.2 to 5.5. The obtained experimental results illustrate that the pre-trained CNN model combining with DA, L2, and ES optimization techniques yields the highest performance compared to other experiments. It provided accuracy, precision, recall, and F1-score were 87%, 89%, 87% and 87%, respectively. In addition, the optimization of the pre-trained CNN model using the DA technique in the experiment 5.2 (pre-trained CNN model+DA) yielded a second accuracy result (86%). For experiments 5.3 and 5.4, both experiments yielded the same accuracy value of 85% and achieved higher accuracy values than the experiment 5.1, which were not employed optimization techniques.

Table 4. Pre-trained CNN modeling for classifying online video content with optimization techniques

Experiment setting	Pre-trained CNN	DA	L2	ES
4.1	✓	-	-	-
4.2	✓	✓	-	-
4.3	✓	-	✓	-
4.4	✓	-	-	✓
4.5	✓	✓	✓	✓

Table 5. Pre-trained CNN model performance results with test datasets

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
5.1 Pre-trained CNN	84	87	84	84
5.2 Pre-trained CNN+DA	86	89	86	86
5.3 Pre-trained CNN+L2	85	88	85	85
5.4 Pre-trained CNN+ES	85	87	85	85
5.5 Pre-trained CNN+DA+L2+ES	87	89	87	87

5. CONCLUSION

This research presents a framework for developing a deep learning approach using CNN algorithms for categorizing the content of online video clips by referring to the criteria for categorizing content of Thai television programs enforced by the office of the Thai NBTC. The researcher collected video clips of movies and TV series published online on the websites. Then the data were prepared by bringing the video clip to create image frames and using the resulting image frames to create a data label divided into the following classes of content: i) sexually inappropriate content; ii) violently inappropriate content; and iii) general content. Meanwhile, the researchers used the datasets in the process of creating a CNN model and a pre-trained CNN model for video content classification, and improved the performance of the CNN model and the pre-train CNN model using various optimization techniques, namely DA, L2, and ES. The model's performance is evaluated with the training dataset, the validation dataset and the test dataset. The result of the training dataset showed that CNN modeling experiments (experiment 3.1) and CNN models using the L2 (experiment 3.3) shared similar efficiency but showed higher efficiency than other experiments. The mean results for accuracy were 97.4% and 95.7% respectively. For the experiments with the validation dataset, the results tended to be in the same direction as the training dataset as for experiments 3.1 and 3.3, the Accuracy values were 89.6% and 89.1% respectively. The test datasets, which were not used in the modeling process like the training datasets and the validation datasets, it was found that experiment 3.5, the CNN model created with the application of DA, L2, and ES, had the best results when compared with other experiments. The results for accuracy, precision, recall and F1 score were 81%, 83%, 81%, and 80% respectively. However, in order to enhance the performance of the CNN model classification, a transfer learning method and model optimization techniques were also adopted in this study. The pre-trained models, namely EfficientNet-B0, were employed with the obtained CNN models and then were evaluated with the test dataset. The experimental results of the pre-trained CNN models with test datasets demonstrate that the pre-trained CNN model using DA, L2, and ES yield the highest accuracy, precision, recall and F1 score of 87%, 89%, 87% and 87%, respectively. These obtained results show that the CNN model with transfer learning (pre-trained model) can effectively help to improve the classification performance. From the methods and experimental results presented above, it was found that the framework for developing a model for content classification of online video clips generated by the CNN algorithm's deep learning approach yielded a high level of accuracy. Therefore, the proposed framework can be used to develop the automatic analysis and classification system in other content classes in the realtime online video clips, especially the classification of content that may be inappropriate for children and youth.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support provided by Faculty of Science and Technology, Thammasat University, Contract No. 2/2564. This work was also supported by Thammasat Research Unit in Data Innovation and Artificial Intelligence.

REFERENCES




- [1] M. B. Short, L. Black, A. H. Smith, C. T. Wetterneck, and D. E. Wells, "A review of internet pornography use research: Methodology and content from the past 10 years," *Cyberpsychology, Behavior, and Social Networking*, vol. 15, no. 1, pp. 13–23, Jan. 2012, doi: 10.1089/cyber.2010.0477.
- [2] K. Krajangsaeng, K. Chanasith, and T. Chantuk, "Causal factors affecting violent behavior in adolescents," *Journal of Humanities and Social Sciences Thonburi University*, vol. 12, no. 27, pp. 97–110, 2018.
- [3] A. Musha, A. Al Mamun, A. Tahabilder, M. J. Hossen, B. Hossen, and S. Ranjbari, "A deep learning approach for COVID-19 and pneumonia detection from chest X-ray images," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 4, pp. 3655–3664, Aug. 2022, doi: 10.11591/ijece.v12i4.pp3655-3664.
- [4] I. Salehin *et al.*, "Analysis of student sentiment during video class with multi-layer deep learning approach," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 4, pp. 3981–3993, Aug. 2022, doi: 10.11591/ijece.v12i4.pp3981-3993.
- [5] P. Vasavi, A. Punitha, and T. V. N. Rao, "Crop leaf disease detection and classification using machine learning and deep learning algorithms by visual symptoms: A review," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 2, pp. 2079–2086, Apr. 2022, doi: 10.11591/ijece.v12i2.pp2079-2086.
- [6] T. A. Sadoon and M. H. Ali, "Deep learning model for glioma, meningioma and pituitary classification," *International Journal of Advances in Applied Sciences*, vol. 10, no. 1, pp. 88–98, Mar. 2021, doi: 10.11591/ijaas.v10.i1.pp88-98.
- [7] A. A. M. Al-Saffar, H. Tao, and M. A. Talab, "Review of deep convolution neural network in image classification," in *2017 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, Oct. 2017, pp. 26–31. doi: 10.1109/ICRAMET.2017.8253139.
- [8] O. Dahmane, M. Khelifi, M. Beladgham, and I. Kadri, "Pneumonia detection based on transfer learning and a combination of VGG19 and a CNN built from scratch," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 3, pp. 1469–1480, Dec. 2021, doi: 10.11591/ijeecs.v24.i3.pp1469-1480.
- [9] N. Nafi'iyah and A. Yuniarti, "A convolutional neural network for skin cancer classification," *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 11, no. 1, pp. 76–84, Apr. 2022, doi: 10.11591/ijict.v11i1.pp76-84.

- [10] M. M. Ben Ismail, "Insult detection using a partitioned CNN-LSTM model," *Computer Science and Information Technologies*, vol. 1, no. 2, pp. 84–92, Jul. 2020, doi: 10.11591/csit.v1i2.p84-92.
- [11] A. Srinivasulu and A. Pushpa, "Disease prediction in big data healthcare using extended convolutional neural network techniques," *International Journal of Advances in Applied Sciences*, vol. 9, no. 2, pp. 85–92, Jun. 2020, doi: 10.11591/ijaas.v9.i2.pp85-92.
- [12] A. Ali and N. Senan, "A review on violence video classification using convolutional neural networks," in *Advances in Intelligent Systems and Computing*, 2017, pp. 130–140. doi: 10.1007/978-3-319-51281-5_14.
- [13] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1725–1732. doi: 10.1109/CVPR.2014.223.
- [14] M. Moustafa, "Applying deep learning to classify pornographic images and videos," Nov. 2015, [Online]. Available: <http://arxiv.org/abs/1511.08899>
- [15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, vol. 25, pp. 1097–1105. [Online]. Available: <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>
- [16] C. Szegedy et al., "A large annotated corpus for learning natural language inference," in *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1–9. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html
- [17] NPDI, "Pornography database," 2013. <https://sites.google.com/site/pornography-database/> (accessed Sep. 01, 2021).
- [18] S. U. Khan, I. U. Haq, S. Rho, S. W. Baik, and M. Y. Lee, "Cover the violence: A novel deep-learning-based approach towards violence-detection in movies," *Applied Sciences*, vol. 9, no. 22, pp. 4963–4974, Nov. 2019, doi: 10.3390/app9224963.
- [19] A. G. Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," Apr. 2017, [Online]. Available: <http://arxiv.org/abs/1704.04861>
- [20] E. B. Nievas, O. D. Suarez, G. B. Garcia, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011, pp. 332–339. doi: 10.1007/978-3-642-23678-5_39.
- [21] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Jun. 2012, pp. 1–6. doi: 10.1109/CVPRW.2012.6239348.
- [22] C.-H. Demarty, C. Penet, M. Soleymani, and G. Gravier, "VSD, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation," *Multimedia Tools and Applications*, vol. 74, no. 17, pp. 7379–7404, Sep. 2015, doi: 10.1007/s11042-014-1984-4.
- [23] X. Lin, F. Qin, Y. Peng, and Y. Shao, "Fine-grained pornographic image recognition with multiple feature fusion transfer learning," *International Journal of Machine Learning and Cybernetics*, vol. 12, no. 1, pp. 73–86, Jan. 2021, doi: 10.1007/s13042-020-01157-9.
- [24] A. Kim, "NSFW dataset," *github*. https://github.com/alex000kim/nsfw_data_scraper (accessed Oct. 01, 2021).
- [25] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violent interaction detection in video based on deep learning," *Journal of Physics: Conference Series*, vol. 844, no. 1, pp. 12044–12052, Jun. 2017, doi: 10.1088/1742-6596/844/1/012044.
- [26] M. Perez et al., "Video pornography detection through deep learning techniques and motion information," *Neurocomputing*, vol. 230, pp. 279–293, Mar. 2017, doi: 10.1016/j.neucom.2016.12.017.
- [27] S. Accattoli, P. Sernani, N. Falcionelli, D. N. Mekuria, and A. F. Dragoni, "Violence detection in videos by combining 3D convolutional neural networks and support vector machines," *Applied Artificial Intelligence*, vol. 34, no. 4, pp. 329–344, Mar. 2020, doi: 10.1080/08839514.2020.1723876.
- [28] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 558–567. doi: 10.1109/CVPR.2019.00065.
- [29] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, Mar. 2021, doi: 10.1145/3446776.
- [30] B. J. Kasche and F. Nordström, "Regularization methods in neural networks," M.S. thesis, Dept. of Statist., Uppsala Univ., Uppsala, Sweden, 2020. [Online]. Available: <http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-403486>
- [31] P. Cheewaparakobkit, "Improving the performance of an image classification with convolutional neural network model by using image augmentations technique," *TNI Journal of Engineering and Technology*, vol. 7, no. 1, pp. 59–64, 2019, [Online]. Available: <https://ph01.tci-thaijo.org/index.php/TNIJournal/article/view/182318>
- [32] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, pp. 1–40, Dec. 2016, doi: 10.1186/s40537-016-0043-6.
- [33] M. Ramesh and K. Mahesh, "A performance analysis of pre-trained neural network and design of CNN for sports video classification," in *2020 International Conference on Communication and Signal Processing (ICCSP)*, Jul. 2020, pp. 0213–0216. doi: 10.1109/ICCSP48568.2020.9182113.
- [34] M. S. AL-Huseiny and A. S. Sajit, "Transfer learning with GoogLeNet for detection of lung cancer," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, pp. 1078–1086, May 2021, doi: 10.11591/ijeecs.v22.i2.pp1078-1086.
- [35] L. T. Duong, P. T. Nguyen, C. Di Sipio, and D. Di Ruscio, "Automated fruit recognition using EfficientNet and MixNet," *Computers and Electronics in Agriculture*, vol. 171, pp. 1–10, Apr. 2020, doi: 10.1016/j.compag.2020.105326.
- [36] Ü. Atıla, M. Uçar, K. Akyol, and E. Uçar, "Plant leaf disease classification using EfficientNet deep learning model," *Ecological Informatics*, vol. 61, pp. 1–20, Mar. 2021, doi: 10.1016/j.ecoinf.2020.101182.
- [37] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *36th International Conference on Machine Learning, ICML 2019*, May 2019, pp. 6105–6114. [Online]. Available: <http://proceedings.mlr.press/v97/tan19a.html>
- [38] National Broadcasting and Telecommunications Commission Thailand, "Rating the appropriateness of Thai television programs." shorturl.at/efzAN (accessed Sep. 01, 2021).
- [39] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)*, Aug. 2017, pp. 1–6. doi: 10.1109/ICEngTechnol.2017.8308186.
- [40] M. Bhubé, "Classifying fashion with a keras CNN." shorturl.at/aQR15 (accessed Jan. 20, 2022).
- [41] P. Murugan and S. Durairaj, "Regularization and optimization strategies in deep convolutional neural network," Dec. 2017, [Online]. Available: <http://arxiv.org/abs/1712.04711>
- [42] A. Krogh and J. Hertz, "A simple weight decay can improve generalization," in *Advances in Neural Information Processing Systems*, 1992, vol. 4, pp. 950–957. [Online]. Available: https://proceedings.neurips.cc/paper/1991/file/8eefcfd5990e441_f0fb6f3fad709e21-Paper.pdf




- [43] L. Prechelt, “Early stopping-but when?,” in *Neural Networks: Tricks of the trade (Lecture Notes in Computer Science)*, 2012, pp. 53–67. doi: 10.1007/3-540-49430-8_3.
- [44] B. Jason, “Use early stopping to halt the training of neural networks at the right time,” 2018. <https://machinelearningmastery.com/how-to-stop-training-deep-neural-networks-at-the-right-time-using-early-stopping/> (accessed Jan. 01, 2022).

BIOGRAPHIES OF AUTHORS



Tanatorn Tanantong    received a bachelor's degree and master's degree in Computer Engineering from Suranaree University of Technology, Thailand in 2005 and 2008, respectively. In 2010, he received a scholarship from Thailand Research Fund (TRF) under The Royal Golden Jubilee Ph.D. Program and got his Ph.D. in Computer Science from Sirindhorn International Institute of Technology (SIIT), Thammasat University, Thailand in 2015. He is currently an Assistant Professor at Department of Computer Science (CS), Faculty of Science and Technology, Thammasat University, Thailand. He is also the head of Thammasat Research Unit in Data Innovation and Artificial Intelligence. In 2016, he was a visiting professor at School of Computing, Informatics, and Decision Systems Engineering, Arizona State University in USA. In 2017, he was also a Post-Doctoral researcher at Japan Advanced Institute of Science and Technology in Japan. In recent years, he has received the Ph.D. Dissertation Award (Good Level) from The National Research Council of Thailand in 2019. His interesting research includes Artificial Intelligence, Data Mining, Medical & Health Informatics, and Hospital Information Systems. He can be contacted at email: tanatorn@sci.tu.ac.th.



Patcharajak Yongwattana    received a bachelor's degree in Computer Science from Faculty of Science and Technology, Thammasat university, Thailand in 2021. He was a research assistant under supervision of Assistant Professor Dr. Tanatorn Tanantong at Thammasat Research Unit in Data Innovation and Artificial Intelligence from July 2020 to September 2021. He passed the final round of the National Software Contest (NSC) Thailand in 2021. He is currently the Data Scientist at True Corporation, Thailand. His research interests are in Deep Learning and Image Processing. He can be contacted at email: Patcharajak.yong@outlook.co.th.