Journal of Information Security and Applications 65 (2022) 103123

Contents lists available at ScienceDirect

Journal of Information Security and Applications

journal homepage: www.elsevier.com/locate/jisa

Image-based malicious Internet content filtering method for child protection

Rasim M. Alguliyev, Fargana J. Abdullayeva [*], Sabira S. Ojagverdiyeva

Institute of Information Technology, Azerbaijan National Academy of Sciences, Baku, Azerbaijan

ARTICLE INFO

Keywords:
Child protection
ChildNet
Convolutional neural network
Content filtering
Not safe for work (NSFW)
Safe for work (SFW)
Deep learning
Harmful content

ABSTRACT

Children and teenagers are among Internet users and they encounter harmful data in the global network. Young users often become the potential victims of pornographic images. Avoiding pornographic images harmful to the child audience is an important research task in the field of detection, computer vision and multimedia. Malicious content can be prevented using various methods. Current paper presents a ChildNet model that filters harmful image content. The pixels of the digital images are used as a data source for recognition of nudity in the images. For each class, a multi-layer deep neural network architecture with five convolution blocks is developed to study the color patterns of undesirable image pixels. The developed neural network consists of 21 layers; the size of the filters is specified as (3 × 3). The filter's size is reduced to increase the accuracy of pixel recognition. The efficiency of the proposed method is tested on real datasets for evaluation purposes and the superior results are obtained from the proposed method in comparison with classical CNN.

1. Introduction

The fast growth of the Internet and the expansion of technology significantly facilitate Internet access through mobile phones, tablets, and other devices. The technological indicators of these devices allow users to access any content regardless of location. Gradually, the number of users increases, along with simultaneous gradual growth of the amount of information. Note that the flood of information contains a wealth of data that may maintain harmful content and may not be suitable for the child audience. Modern children are growing up in the digital world, and they are very active on the Internet. Some information, which contains pornography, torture, cruelty, drugs, alcohol, terrorism, vandalism, and other immoral habits, have recently spread on the web sites and social media propagating undesirable moral and ethical qualities [1].

As noted, children and teenagers suffer differently while using the Internet, however, two most dangerous threats (online sexual desire and access to pornography) attract the attention of experts and related persons. Child pornography is now a problem faced by law enforcement agencies around the world. While numerous laws exist in this regard, it remains a problem and requires more technological solutions, along with social ones.

Children and adolescents are sexually abused over the Internet by criminals, and while being unaware of these dangers, they often become victims. Approximately one in four young users, who regularly use Internet resources, face unwanted sexual images. 25% of adolescents have a risk of being more abused during the year [2].

The Council of Europe's report on "Protecting children against harmful content" classifies the types of malicious information intended for children and includes a clause titled "Sexual relationship in the form of pornographic information and images" [3].

Various technical measures have been taken to protect children from harmful content damaging their health and psychology. The security of children in the network is ensured through the control systems and various filters applied on the Internet traffic, servers and personal computers through the software. The harmful content inappropriate for child psychology, health or age is detected on web pages, and different content prohibitions are applied.

Numerous methods have been proposed for the recognition of pornographic images on the network so far. Several methods are based on content-based image detection [4]. These methods depict the content of the images based on the visual features (as color, texture, relief, etc.). The classification model is based on these features. Pornographic image detection is reduced to the issue of binary classification (non-pornographic or pornographic) [5, 6, 7]. These approaches display the content of the pornographic image as low-level visual features (color, texture,

[*] Corresponding author.
E-mail address: a_farqana@mail.ru (F.J. Abdullayeva).

https://doi.org/10.1016/j.jisa.2022.103123

Available online 2 February 2022

relief, etc.) and then build a classification model by applying machine learning techniques to the vector of these features. Although these types of approaches provide good results, the selection of features here is a complex issue and requires professional staff. To avoid the difficulty in selecting the features, deep learning methods have been applied to computer vision in recent years [8, 9].

Pornographic images often include large naked skin areas. Skin color is the most stable feature of pornographic images. In this regard, the color of skin enables the initial recognition of pornographic images.

In [10], Yahoo offers a classification model for offensive and NSFW (not suitable/safe for work) content inappropriate for children, particularly pornographic images using the convolutional neural network (CNN), and develops its open source code. This code is implemented in the Caffe library of the Python package. An image is fed into the model input, generating the scores ranging from 0 to 1 at the output. The images below the specified threshold based on these scores are classified as undesirable. Since a simple convolutional neural network is used here, the model causes a significant loss during the classification providing low detection accuracy based on the evaluation metrics. Zhou et al. [8] propose the CNN model for the classification of pornographic images. The method consists of two parts: rough detection and slight detection. The rough detection module detects normal images based on smaller skin color regions, whereas the slight detection module detects pornographic images based on larger skin color regions. Colmenares-Guillén et al. [11] offer a filter to remove the visibility of unwanted content on the Internet for children and adolescents. The RSOR algorithm (Recognition, Selection and Operation Regions) is used to detect nudity in the digital image. Nian et al. [12] provide a method based on a deep convolutional neural network to detect pornographic images. Learning algorithm based on two strategies is proposed. The first algorithm is an unstable regulatory strategy that aims to adjust training data at the appropriate time. Then it offers a fast-visual classification method based on the sliding window algorithm for trials.

Recognition accuracy is the best condition for the recognition of pornographic images on the Internet. Available methods can cause significant loss on image recognition, and the recognition accuracy of these methods is very low.

This paper proposes a method with high accuracy classification of undesirable images to prevent harmful image content. A multi-layer neural network architecture with five convolution blocks is built to study the appropriate texture templates of undesirable image pixels for each class. The constructed neural network model contains 21 layers.

The main contributions of the study are:

- Deep ChildNet model with 21 layers consisting of five convolution blocks was proposed, to filtrate the malicious image content on the Internet.
- In the detection of pornographic and non-pornographic images, the feature extraction and selection were not performed.
- The efficiency of the proposed method was evaluated on the Python program package.

The paper is organized as follows. Section 2 introduces the overview of related works. Section 3 describes the proposed ChildNet architecture. Section 4 presents the results of comparative analysis of the proposed method with the existing methods. Section 5 provides experiments for the evaluation of the efficiency of proposed method. Section 6 presents the conclusion.

2. Related work

Recently, the emergence of computer vision, big data, and deep learning algorithms has greatly facilitated the automatic detection of NSFW content images. Huang et al. [13] present the convolution neural network (CNN) ensemble to classify the pornographic and normal images. The ensemble is developed from many CNNs based on the databases containing various pornographic and normal images. Yin et al. [14] propose a hybrid method for identification of a very large part of the naked body in the images for adolescents. The model consists of three parts called color filtering, texture filtering, and dimension filtering. The skin color processing unit performs the function of non-skin color pixels, and a rough texture filtering is performed. Due to the dimension of fractals, skin-like regions are filtered. The model filters non-skin color regions, and then, these regions are treated as a pornographic image according to the threshold values if the skin-color region is larger than the threshold value. Although this is the simplest method, there occur a lot of recognition errors.

Zaidan et al. [15] propose a system that employs two machine learning techniques to detect pornographic images using multi-agent learning, considering color characteristics. The classification system incorporates a Bayesian method that uses grouping histogram technique based on the YCbCr color region. The features of the shape are removed when the skin is detected and then sent to the backpropagation neural network. For classifying the extracted skin region through multi-agent learning, it can determine whether the image is pornographic or not. The main advantage of this research is the identification of pornography and blocking the websites that secretly promote pornography. The hybrid method proposed here is more resistant to change the image size in the classification of pornographic images. Nugroho et al. [16] offer a method that combines two color regions, namely RGB and YCbCr, in the form of skin segment to minimize the errors when specifying a video class in a flawless video category. Yan et al. [17] detect pornographic images based on content analysis. To enhance the representation of visual words, ROI algorithm-based code book has been proposed. Yu and Han [18] evaluate the visioning of the pornographic skin regions to detect and block the pornographic content. A new method for estimating the skin regions in the HSV color environment has been proposed. Sae-Bae et al. [19] propose an automated method for identifying child pornographic images. A child pornographic system was developed, which determines the shades of human skin in the digital images and extracts the features for the detection of naked images and performs age-based classification of human images. Sharma et al. [20] combine the results of text-based recognition, image-based recognition, and text and image-based recognition. The web pages containing secure and pornographic texts and images are detected in different categories. Discovered web pages are classified as naked and secure ones. Using an SVM, a hybrid approach to the detection of the web pages containing malicious content has been proposed. Yin et al. [21] also propose an SVM-based pornographic image detection system. The detection process was performed in two phases. The first phase identifies the features of the skin, face, and shape. The second phase transfers all the features to the SVM classifier to determine whether the image is pornographic or not. In [22], a new decision tree algorithm based on face and body detection have been proposed for the detection of pornographic images. In [23], some additional color-independent features (face detectors) were also considered. After the calculation of features for viewing, SVM was used for classification. Durrell et al. [24] use the distinctive characteristics of vectors and neural networks to define the semantic image content. Pornographic images were identified using Gabor filters, PCA, Correlograms and neural networks. Schettini et al. [25] classifies the pornographic and non-pornographic images on the web pages. CART (Classification And Regression Trees) and SVM were used for the decision-making process and classification. Pornographic images have been detected using only low-level features (color, texture, edge distribution).

3. Proposed ChildNet architecture

ChildNet is a model enable to classify the harmful content on the Internet. The ChildNet model is illustrated in Fig. 1.

The proposed ChildNet model includes five blocks consisting of 21 convolutions layers with the filters of {64, 128, 256, 512, 4608} depth

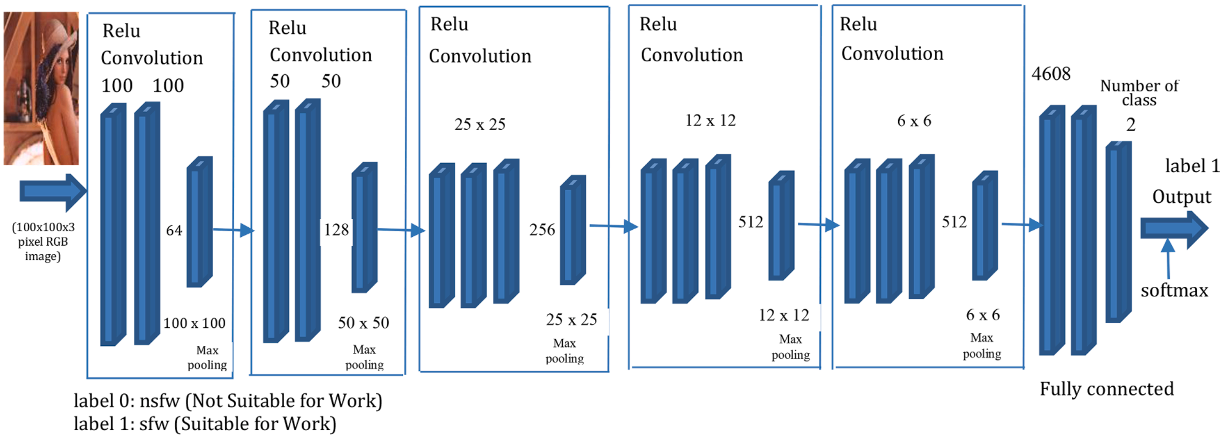
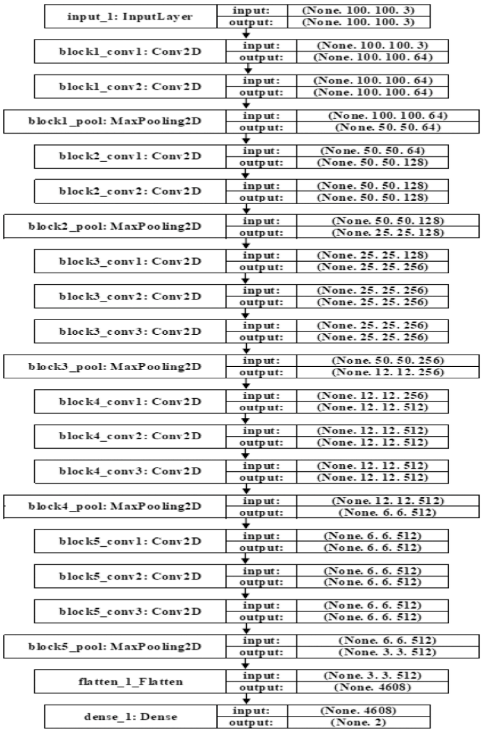


Fig. 1. ChildNet malicious image content filtration model.

and 5 max-pooling layers. The first block of the ChildNet model uses 64 filters, the second block 128 filters, the third block 256 filters, the fourth block 512 filters, and the fifth block 4608 filters. Note that the number of filters in the convolution layers is either remained or doubled in each following layer. The RGB images of (100 × 100) size are fed to the input of the proposed model. The first two fully connected layers of the model include 4608 nodes (neurons). The last fully-connected layer includes *n* number of nodes. Here, *n* is the number of classes, in our case, $n = 2$. Relu activation function is used to obtain nonlinearity in all layers except the last fully connected layer. However, the Softmax regression function is used to perform the classification on the last fully connected layer.

The output parameters of the proposed model are illustrated in Fig. 2. The proposed model contains 21 layers consisting of five convolution blocks, and the filter size is set (3 × 3). The goal is to reduce the size of the filters to increase the accuracy of pixel recognition. Here, the pixel resolution is 100 × 100, and 3 - the RGB (Red, Green, Blue) image format.



input_1: InputLayer	input:	(None. 100. 100. 3)
	output:	(None. 100. 100. 3)
block1_conv1: Conv2D	input:	(None. 100. 100. 3)
	output:	(None. 100. 100. 64)
block1_conv2: Conv2D	input:	(None. 100. 100. 64)
	output:	(None. 100. 100. 64)
block1_pool: MaxPooling2D	input:	(None. 100. 100. 64)
	output:	(None. 50. 50. 64)
block2_conv1: Conv2D	input:	(None. 50. 50. 64)
	output:	(None. 50. 50. 128)
block2_conv2: Conv2D	input:	(None. 50. 50. 128)
	output:	(None. 50. 50. 128)
block2_pool: MaxPooling2D	input:	(None. 50. 50. 128)
	output:	(None. 25. 25. 128)
block3_conv1: Conv2D	input:	(None. 25. 25. 128)
	output:	(None. 25. 25. 256)
block3_conv2: Conv2D	input:	(None. 25. 25. 256)
	output:	(None. 25. 25. 256)
block3_conv3: Conv2D	input:	(None. 25. 25. 256)
	output:	(None. 25. 25. 256)
block3_pool: MaxPooling2D	input:	(None. 50. 50. 256)
	output:	(None. 12. 12. 256)
block4_conv1: Conv2D	input:	(None. 12. 12. 256)
	output:	(None. 12. 12. 512)
block4_conv2: Conv2D	input:	(None. 12. 12. 512)
	output:	(None. 12. 12. 512)
block4_conv3: Conv2D	input:	(None. 12. 12. 512)
	output:	(None. 12. 12. 512)
block4_pool: MaxPooling2D	input:	(None. 12. 12. 512)
	output:	(None. 6. 6. 512)
block5_conv1: Conv2D	input:	(None. 6. 6. 512)
	output:	(None. 6. 6. 512)
block5_conv2: Conv2D	input:	(None. 6. 6. 512)
	output:	(None. 6. 6. 512)
block5_conv3: Conv2D	input:	(None. 6. 6. 512)
	output:	(None. 6. 6. 512)
block5_pool: MaxPooling2D	input:	(None. 6. 6. 512)
	output:	(None. 3. 3. 512)
flatten_1_Flatten	input:	(None. 3. 3. 512)
	output:	(None. 4608)
dense_1: Dense	input:	(None. 4608)
	output:	(None. 2)

Fig. 2. ChildNet architecture.

4. Convolutional neural networks

We can obtain the convolution features through Eq. (1). In this equation, $\mathrm{y}_\mathrm{n}^\mathrm{i}$ – indicates the 1st layer of the n-th feature. When we obtain the features of the 1st level, the convolution kernel is indicated by $\omega^i_{m,\ n}$. b^i_n marks the bias. The typical pattern set related to the 1st layer is indicated by v^i_n.

$$y_n^i = f_i\left(\sum_{m\in v_n^i} y_m^{i-1} \otimes \omega_{m,n}^i + b_n^i\right) \tag{1}$$

Pooling layer. The pooling layer includes equal feature maps as in the previous convolution layer. In this layer, the input of the disjoint zones is divided, and the output is obtained through the specified pooling technique in every zone. Later, the offset and outputs are added to the excitation function [13, 14]. In this layer, the features become more vigorous to resist the deformation.

$$y_n^i = f_i\left(z_n^{i-1} \otimes \omega_n^i + b_n^i\right) \tag{2}$$

Once the window of 1–1 convolution layer features is fixed, we get the value z_n^{i-1} using the pooling algorithm (average-pooling, max-pooling). The weight of the mas is indicated by ω^i_n, whereas the offset is indicated by b^i_n.

Fully connected layer. This layer usually includes the sigmoid neuron or RBF neuron. Overall, fully connected layers are the last layers of the network. The features dimension is reduced. The number of neuron and the types of the input image are identical as in the last layer.

For calculations of the sigmodal neuron of the output layer, we used Eq. (3).

$$y_n^i = f_i\left(\sum_{m=1}^{n_{i-1}} y_m^{i-1} \omega_{m,n}^i + b_n^i\right) \tag{3}$$

n_i indicates the number of neuron in output layer. In 1–1 layer, the typical features are shown by *m*. In the previous layer, which links to *n* neuron in the 1st layer, the weight of the typical features *m* is indicated by y_m.

5. Experiments

The proposed method was tested on NSFW – V1 dataset and NudeNet Classifier dataset v1 [26, 27]. The NSFW – V1 dataset consists of SFW and NSFW classes. In the training dataset, 4000 images were included into SFW class and 4000 images to the NSFW class, whereas in the test dataset, 500 images were included into the SFW class and 500 images to the NSFW class. Dataset elements are described in Table 1.

Table 1
NSFW – V1 dataset elements.

Class	Species	Number of images in train set	Number of images in test set	Total images
0	NSFW	4000	500	4500
1	SFW	4000	500	4500
Total		8000	1000	9000

For both algorithms, the number of iterations is 5, 10, 30, 50, 100, 300, 1500, the batch size is 32, an optimization method is RMSprop, the loss function is categorical cross-entropy and activation function during the model learning is Relu. The proposed ChildNet model was comparatively analysed with CNN and augmented CNN. The loss and accuracy values obtained from each model are presented in Table 2.

As seen in Table 2, the ChildNet model provides more effective results than others. Here, the loss and accuracy values significantly change while the number of iterations of CNN and Augmented CNN models increases in the testing process. The model performs significant losses both in the training and testing process of the neural network when the number of iterations increases. Thus, when the CNN model was tested in 300 iterations, a significant loss has occurred (loss value 0.8736). In addition, for the Augmented CNN model in 300 iterations, the loss value accounts for 8,0590 during training and 8,1945 in the testing process. The proposed ChildNet model performs high accuracy with minimal loss in the training and testing processes of the neural network. Thus, in 300 iterations, the model provides loss value of 0,3681 and accuracy value of 0,8554 during the training process, and loss value of 0,3622 and accuracy value of 0,8792 in the testing process. This significantly overweighs the CNN model with respective loss values of 0.8736 and 0.8530 in training and testing processes.

Generally, when comparing the model prediction values, the ChildNet model performs low loss and high accuracy values in both training and testing processes. As the number of iterations grows, as shown in Fig. 3, the loss curve for the ChildNet model tends to decrease smoothly without any deviations, with the training and testing lines almost overlapping.

Moreover, as illustrated in Fig. 4, providing high accuracy value, the training, and testing curve on accuracy increase and become too close to each other.

When the number of iterations increases, the ChildNet model can achieve higher accuracy value. When testing the model in 1500 iterations, accuracy value accounts for 0,9065, and loss value accounts for 0,2925. The results of the ChildNet model in 1500 iterations are presented in Table 3.

Visual descriptions of the ChildNet results in terms of loss and accuracy function in 1500 iterations are presented in Fig. 5 and Fig. 6, respectively.

The loss and accuracy curves for CNN and augmented CNN models are depicted in Fig. 7, Fig. 8, Fig. 9 and Fig. 10, respectively.

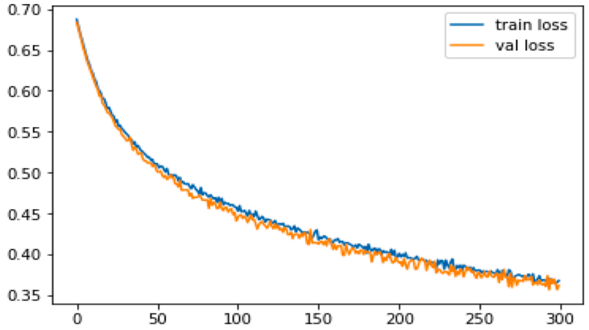


Fig. 3. Loss per 300 epoch in ChildNet model NSFW – V1 dataset.

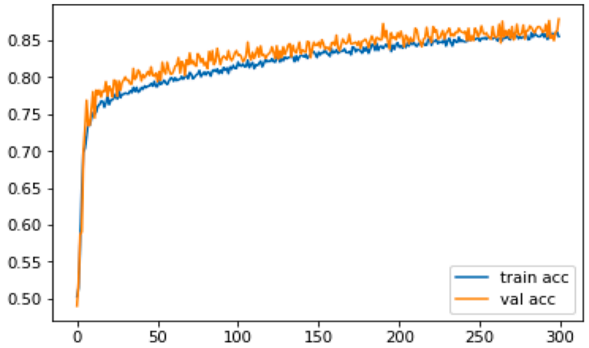


Fig. 4. Accuracy per 300 epoch in ChildNet model on NSFW – V1 dataset.

Table 3
ChildNet model efficiency indicators on NSFW – V1 dataset.

# of epochs			
Method	Class	Metrics	1500
ChildNet	nsfw	Accuracy	0.9065
		Precision	0.8814
		Recall	0.9121
		f1-score	0.9042
	sfw	Accuracy	0.8821
		Precision	0.9023
		Recall	0.8834
		f1-score	0.8922

For comparison of the proposed method with existing ones, the methods' classification results based on various metrics on NSFW dataset are presented in Table 4.

Table 4 shows that the proposed ChildNet model provides better results in each iteration compared to other ones. A characteristic feature

Table 2
Loss and accuracy on CNN, Augmented CNN and ChildNet models on NSFW – V1 dataset.

Number of iterations									
			5	10	30	50	100	300	1500
CNN	Train	Loss	0.2141	0.2019	0.1377	0.1318	0.1400	0.8736	0.9602
		Accuracy	0.9160	0.9258	0.9590	0.9565	0.9668	0.9458	0.9404
	Test	Loss	0.1488	0.2764	0.1097	0.3352	0.3584	0.8530	1.1408
		Accuracy	0.9417	0.8925	0.9624	0.9360	0.9396	0.9460	0.9281
Augmented CNN	Train	Loss	0.2013	0.1583	0.1304	0.1846	0.4640	8.0590	3.7736
		Accuracy	0.9263	0.9451	0.9566	0.9390	0.8980	0.5000	0.7659
	Test	Loss	0.1255	0.1071	0.1382	0.1625	0.1904	8.1945	2.8207
		Accuracy	0.9527	0.9611	0.9716	0.9601	0.9433	0.4916	0.8250
ChildNet	Train	Loss	0.6448	0.6130	0.5496	0.5514	0.4674	0.3681	0.2925
		Accuracy	0.7009	0.7593	0.7586	0.7660	0.8120	0.8554	0.8868
	Test	Loss	0.6396	0.6089	0.5421	0.5483	0.4637	0.3622	0.2917
		Accuracy	0.7300	0.7721	0.7773	0.7805	0.8246	0.8792	0.8833

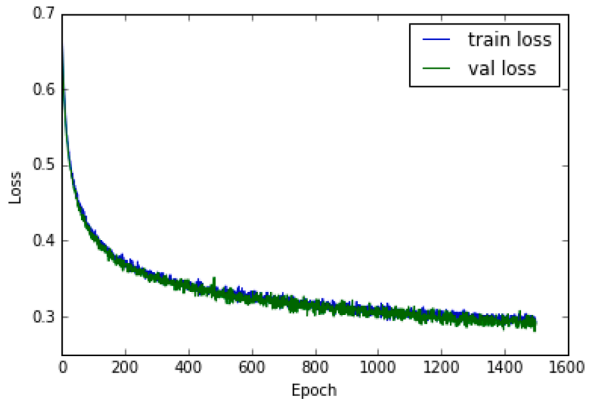


Fig. 5. Loss per 1500 epochs in ChildNet model on NSFW – V1 dataset.

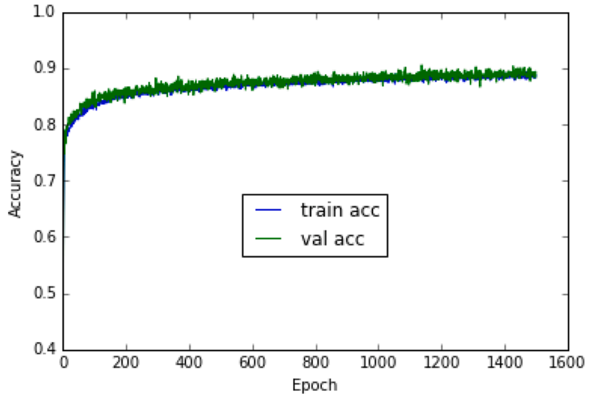


Fig. 6. Accuracy per 1500 epochs in ChildNet model on NSFW – V1 dataset.

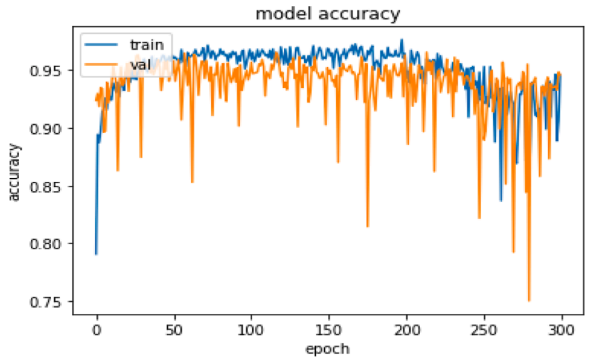


Fig. 7. Loss per 300 epochs in CNN model on NSFW – V1 dataset.

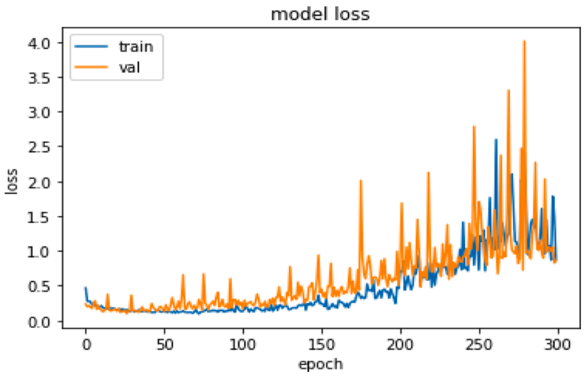


Fig. 8. Accuracy per 300 epochs in CNN model on NSFW – V1 dataset.

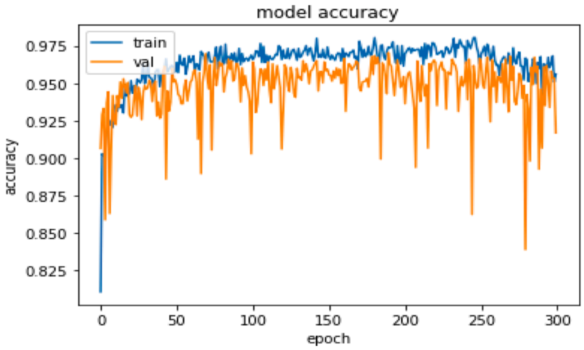


Fig. 9. Loss per 300 epochs in augmented CNN model on NSFW – V1 dataset.

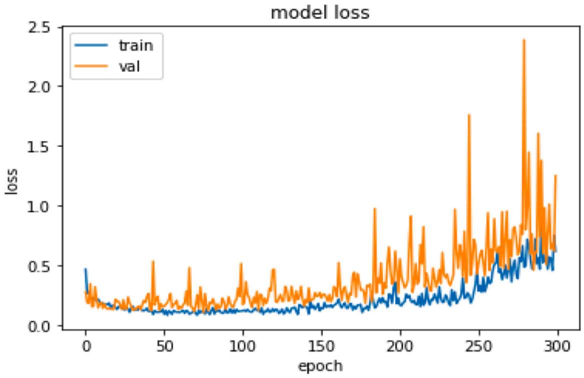


Fig. 10. Accuracy per 300 epochs in augmented CNN model on NSFW – V1 dataset.

of the ChildNet model is the successful classification of large-scale images, and the model attempts to perform more accurate results by increasing the number of iterations, while other methods perform differently. As the number of iterations increases, the performance of the model begins to decrease. According to Table 4, in 300 iterations the ChildNet model recognizes the NSFW class images with 0,8902 accuracies, whereas the simple CNN model is capable of recognizing these class images with 0,5081 accuracies. This value for the augmented CNN model is 0,4491.

To demonstrate the advantage of the proposed method, Fig. 11 presents a visual representation of the NSFW and SFW classification results based on the accuracy metrics of all three methods.

As shown in Fig. 11, ChildNet performs with higher accuracy than other methods.

In total the NSFW-V1 dataset contains 9000 images. There exist 4500 images in the NSFW class and 4500 in SFW class. By launching the ChildNet model on this dataset, the algorithm from 492 images correctly recognized 446 images and added them into the NSFW class, while allowing false recognition it recognized 46 images incorrectly and added them into the SFW class. The algorithm correctly added 434 images from the 492 images of the SFW class into the SFW class, and wrongly added 58 images into the NSFW class. The classification results of the ChildNet algorithm in 1500 iterations on NSFW – V1 dataset are presented in Table 5.

The classification results of the ChildNet algorithm in 300 iterations on NSFW – V1 dataset are presented in Table 6.

As shown in Table 6, the CNN method can recognize the images with very low accuracy. Thus, CNN adds 250 images out of 492 images into the NSFW class and allows 242 image errors, incorrectly adding them into the SFW class, however correctly adding 265 images into the SFW class and incorrectly adding 227 images into the NSFW class.

Table 4
Evaluation of the methods by the various metrics on NSFW – V1 dataset.

# of epochs									
Methods	Class	Metrics	5	10	30	50	100	300	1500
CNN	nsfw	accuracy	0.4939	0.4166	0.5020	0.4268	0.5528	0.5081	0.5142
		precision	0.4801	0.5112	0.5121	0.4813	0.5123	0.5222	0.4812
		recall	0.4923	0.4221	0.5023	0.43031	0.5523	0.5121	0.5111
		f1-score	0.4942	0.4621	0.5123	0.4513	0.5344	0.5223	0.5014
	sfw	accuracy	0.4613	0.5975	0.5142	0.5304	0.4613	0.5386	0.4390
		precision	0.4811	0.5101	0.5122	0.4812	0.5121	0.5203	0.4732
		recall	0.4613	0.6021	0.5142	0.5323	0.4614	0.5422	0.4424
		f1-score	0.4723	0.5503	0.5132	0.5004	0.4823	0.5314	0.4612
Augmented CNN	nsfw	accuracy	0.5020	0.4898	0.5223	0.5441	0.4837	0.4491	0.6565
		precision	0.5232	0.5024	0.5302	0.4823	0.5211	0.4923	0.4933
		recall	0.5022	0.4933	0.5232	0.5024	0.4831	0.4524	0.6623
		f1-score	0.5103	0.4904	0.5312	0.4911	0.5042	0.4744	0.5643
	sfw	accuracy	0.5365	0.5101	0.5325	0.4674	0.5569	0.5264	0.3150
		precision	0.5234	0.5033	0.5342	0.4801	0.5231	0.4921	0.4813
		recall	0.5422	0.5142	0.5314	0.4711	0.5621	0.5323	0.3232
		f1-score	0.5342	0.5102	0.5303	0.4821	0.5422	0.5133	0.3821
ChildNet	nsfw	accuracy	0.6097	0.8109	0.7479	0.8516	0.8597	0.8902	0.9065
		precision	0.8222	0.7603	0.7722	0.7632	0.8022	0.8333	0.8814
		recall	0.6123	0.8123	0.7534	0.8513	0.8612	0.8942	0.9121
		f1-score	0.7024	0.7913	0.7632	0.8104	0.8313	0.8614	0.9042
	sfw	accuracy	0.8658	0.7504	0.7804	0.7378	0.7845	0.8170	0.8821
		precision	0.6922	0.8022	0.7612	0.8334	0.8512	0.8832	0.9023
		recall	0.8712	0.7532	0.7804	0.7433	0.7804	0.8241	0.8834
		f1-score	0.7711	0.7704	0.7722	0.7803	0.8223	0.8503	0.8922

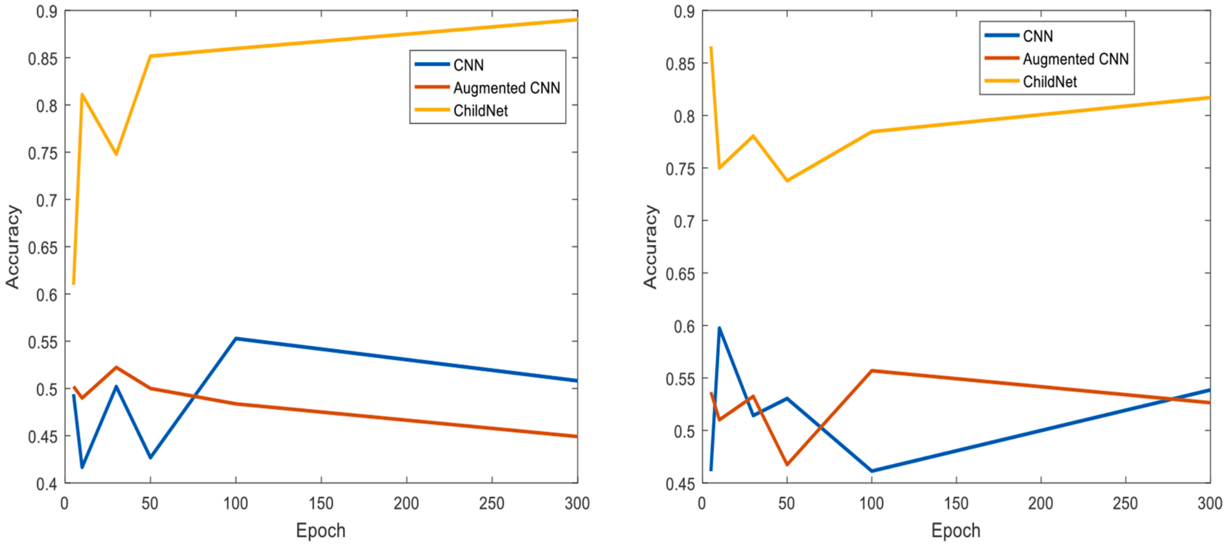


Fig. 11. Comparison of the methods of classification by epochs based on accuracy metric (left: NSFW class, right: SFW class).

Table 5
The number of images classified by ChildNet algorithm in 1500 iterations on NSFW – V1 dataset.

Class	nsfw	sfw	Total
nsfw	446	46	492
sfw	58	434	492

Table 6
The number of images classified by CNN algorithm in 300 iterations NSFW – V1 dataset.

Class	nsfw	sfw	Total
nsfw	250	242	492
sfw	227	265	492

The confusion matrix of the ChildNet method in 1500 iteration on NSFW – V1 dataset is illustrated in Fig. 12.

Fig. 12 shows that in 1500 iterations, the algorithm can classify the NSFW (0) class with 0.91 accuracy performing 0.09 errors, and the SFW (1) class with 0.88 accuracy performing 0.12 errors. This is the expected result. However, in 300 iterations, the CNN algorithm can recognize NSFW (0) class with 0.51 accuracy performing 0.49 errors, and the SFW (1) class with 0.54 accuracy performing 0.46 errors (Fig. 13).

For better demonstration of the results, Fig. 14 visually illustrates the comparison of methods. Notice that in this figure, the values of the accuracy, precision, recall and F1-score metrics clearly demonstrates the difference of the methods.

Accordingly, the proposed network includes a lot of layers and parameters the ChildNet model was also tested on Big Data and we could prove its effectiveness. For this purpose, NudeNet Classifier dataset v1 was used to provide testing process [28]. The dataset consists of safe and nude classes. In the training dataset, 48,672 images were included into

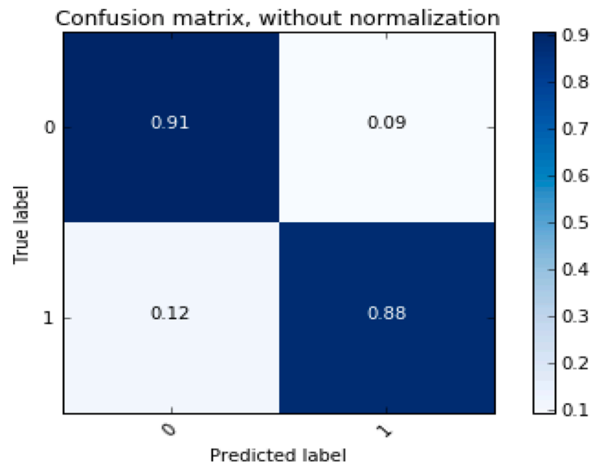


Fig. 12. The confusion matrix of the ChildNet model on NSFW–V1 dataset in 1500 epochs.

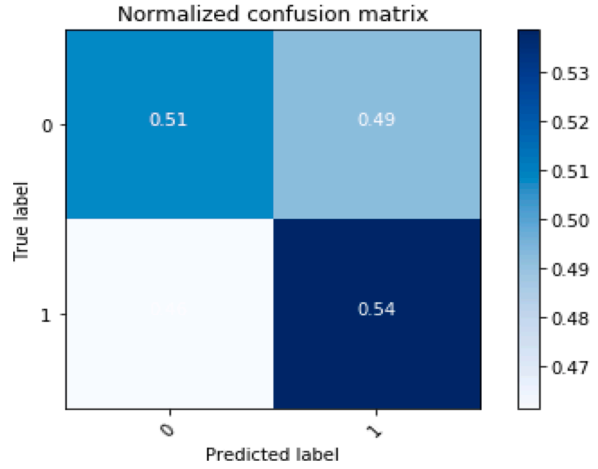


Fig. 13. The confusion matrix of the CNN on NSFW–V1 dataset in 300 epochs.

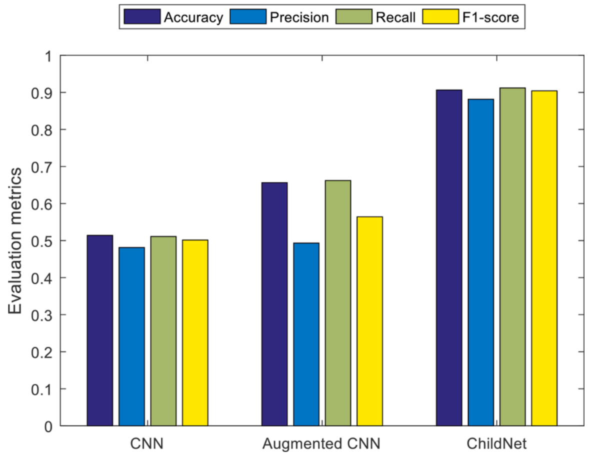


Fig. 14. Performance of the methods in 1500 epochs.

Table 7
NudeNet Classifier dataset v1 elements.

Class	Species	Number of images in train dataset	Number of images in test dataset	Total images
0	nude	48,672	7993	56,665
1	safe	48,672	7993	56,665
Total		97,344	15,986	113,330

safe class and 48,672 images to the nude class, whereas in the test dataset, 7993 images were included into the safe class and 7993 images to the nude class. Dataset elements are described in Table 7.

The experiments on NudeNet Classifier dataset v1 were conducted in 5, 10, 30, 50, 100, 300, 500 iterations, the batch size, an optimization method, the loss function and activation function for both algorithms were taken as same in practice conducted on NSFW – V1 dataset. The results of the experiment are presented in Table 8.

As seen from Table 8, when increasing the number of iterations, the CNN and Augmented CNN algorithms demonstrated anomalous results at different iterations during training and testing phases. In the process of CNN training the loss values of the algorithm in small iterations, for example, at 5, 10, 30, 50 decreased and obtained 0.5856, 0.5697, 0.4916, 0.4665 values, respectively, and in high values of iterations, for example, at 100, 300, 500, the values of this parameter began to increase and obtained 0.5241, 0.8105, 0.7193 values, respectively. When the algorithm allows less loss in low iterations, the accuracy values of the algorithm increased in iterations 5, 10, 30, 50 and obtained 0.7070, 0.7056, 0.7778, 0.7876 values, respectively. In iterations 100, 300, 500, the values of this parameter began to decrease, since the algorithm suffered a lot of losses, and obtained 0.7705, 0.6509, 0.4893 values, respectively. These results demonstrate that the model could not be properly trained the training process. Since the model could not be trained properly in the training phase, the results of the model in the testing phase change with a bigger leap than in the training phase. However, in ideal models, the value of the accuracy parameter in the testing phase should be lower than the value in the training phase. For instance, in 300 iterations, in the training phase of the CNN model, the accuracy parameter obtained 0.6509 value, while in the testing phase it reached 0.7410. This can be considered as a bad result. The same landscape is observed in the Augmented CNN algorithm too.

However, the proposed ChildNet algorithm performed well in each iteration. Thus, when the number of iterations continues to increase, the loss values of the algorithm gradually decrease, and the accuracy values increase. Thus, in the training phase, the loss parameter of the algorithm in iterations 5, 10, 30, 50, 100, 300, 500 gradually decreased and obtained 0.6573, 0.6338, 0.6156, 0.6185, 0.5986, 0.5808, 0.5778 values, respectively, and the accuracy values of the algorithm in the same iterations gradually increased, and obtained 0.6287, 0.6461, 0.6629, 0.6633, 0.6785, 0.6929, 0.6969 values. The shortcomings of the other above mentioned two algorithms are not reflected in the ChildNet model. The model obtained satisfactory results in both training and testing phases. For example, in the 500 iterations, the accuracy parameter of the ChildNet model in the training phase obtained 0. 0.6969 value, while in the testing phase, this value with little difference reached to 0.6885. In ideal models, the value of the accuracy parameter in the training phase should be greater than in the testing phase. This is the expected result.

In conducted experiments, the ChildNet model performed low loss and high accuracy values during both training and testing phases. When the number of iterations grows, as shown in Fig. 15, the accuracy curve for the ChildNet model tends to increase smoothly without any deviations, with the training and testing lines and become too close to each other.

For the comparison of the proposed method with existing ones, the methods' classification results based on various metrics on NudeNet Classifier dataset v1 were presented in Table 9.

As seen from Table 9, the methods tested on the NudeNet Classifier dataset v1 obtained similar results tested on NSFW – V1 dataset. Thus, the ChildNet model recognized the images with high accuracy in all iterations. However, we observe the cases of poor recognition in other methods. Thus, in 300 iterations, the Augmented CNN algorithm did not recognize images of the safe class exactly and obtained 0.0000 values

Table 8
Loss and accuracy on CNN, Augmented CNN and ChildNet models on NudeNet Classifier dataset v1.

Number of iterations									
			5	10	30	50	100	300	500
CNN	Train	Loss	0.5856	0.5697	0.4916	0.4665	0.5241	0.8105	0.7193
		Accuracy	0.7070	0.7056	0.7778	0.7876	0.7705	0.6509	0.4893
	Test	Loss	0.6495	0.5387	0.4550	0.5409	0.5110	0.6182	0.6927
		Accuracy	0.7434	0.6842	0.7127	0.7145	0.7139	0.7410	0.5158
Augmented CNN	Train	Loss	0.5133	0.4842	0.4296	0.5130	1.1951	0.6931	0.6968
		Accuracy	0.7525	0.7728	0.8090	0.7721	0.5891	0.5071	0.5131
	Test	Loss	0.5016	0.4458	0.4245	0.5422	0.6784	0.6933	0.6891
		Accuracy	0.7167	0.7563	0.7958	0.7939	0.5708	0.5000	0.5137
ChildNet	Train	Loss	0.6573	0.6338	0.6156	0.6185	0.5986	0.5808	0.5778
		Accuracy	0.6287	0.6461	0.6629	0.6633	0.6785	0.6929	0.6969
	Test	Loss	0.7286	0.6094	0.7299	0.5834	0.7093	0.5468	0.5805
		Accuracy	0.5998	0.6318	0.6516	0.6571	0.6625	0.6986	0.6885

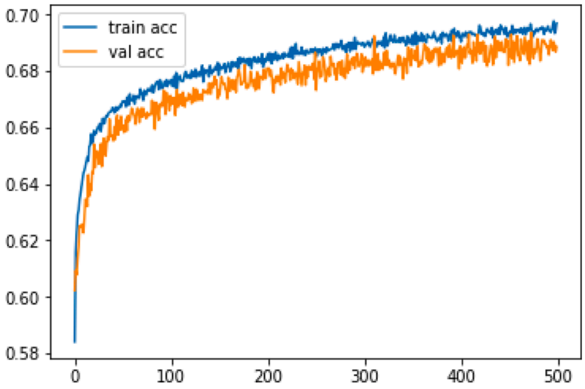


Fig. 15. Accuracy per 500 epochs in ChildNet model on NudeNet Classifier dataset v1.

over every metric. Besides, other algorithms compared with the Childnet model obtained poor results too.

NudeNet Classifier dataset v1 contains 113,330 images. There exist 56,665 images in the nude class and 56,665 in safe class. In testing process the ChildNet model on NudeNet Classifier dataset v1, the algorithm from 7878 images correctly recognized 5831 mages and added them into the nude class, while allowing false recognition it recognized 2047 images incorrectly and added them into the safe class. The algorithm correctly added 5187 images from the 8108 images of the safe class into the safe class, and wrongly added 2921 images into the nude class. The classification results of the ChildNet algorithm in 500 iterations on NudeNet Classifier dataset v1 are provided in Table 10.

As shown in Table 10, the ChildNet method recognized the images well.

Thus, CNN adds 5831 images from 7878 images into the nude class and allows 2047 image errors, incorrectly adding them into the safe class, however correctly adding 5187 images into the safe class and incorrectly adding 2921 images into the nude class. This means that the algorithm can recognize both classes well. The classification results of the CNN algorithm in 500 iterations on NudeNet Classifier dataset v1 are

Table 10
The number of images classified by ChildNet algorithm in 500 iterations on NudeNet Classifier dataset v1.

Class	nude	safe	Total
nude	5831	2047	7878
safe	2921	5187	8108

Table 9
Evaluation of the methods by the various metrics on NudeNet Classifier dataset v1.

# of epochs									
Methods	Class	Metrics	5	10	30	50	100	300	500
CNN	nude	Accuracy	0.4687	0.4083	0.6706	0.5333	0.6021	0.2981	0.0021
		precision	0.4902	0.4843	0.4934	0.4921	0.4911	0.4934	0.5901
		recall	0.4711	0.4123	0.6744	0.5322	0.6002	0.3023	0.0000
		f1-score	0.4823	0.4443	0.5724	0.5134	0.5411	0.3712	0.0000
	safe	Accuracy	0.5324	0.5769	0.3290	0.4628	0.4029	0.6999	0.9985
		precision	0.5133	0.5002	0.5123	0.5133	0.5142	0.5124	0.5111
		recall	0.5342	0.5814	0.3314	0.4623	0.4033	0.7014	1.0024
		f1-score	0.5204	0.5431	0.4003	0.4802	0.4514	0.5923	0.6733
Augmented CNN	nude	Accuracy	0.5765	0.6197	0.5319	0.5082	0.2311	1.0000	0.9986
		precision	0.4932	0.4913	0.4932	0.4913	0.4924	0.4901	0.4911
		recall	0.5814	0.6221	0.5324	0.5121	0.2342	1.0011	1.0000
		f1-score	0.5322	0.5524	0.5141	0.5024	0.3111	0.6603	0.6634
	safe	Accuracy	0.4184	0.3809	0.4697	0.4892	0.7615	0.0000	0.0013
		precision	0.5023	0.5113	0.5123	0.5134	0.5024	0.0000	0.5023
		recall	0.4243	0.3824	0.4734	0.4922	0.7622	0.0000	0.0000
		f1-score	0.4624	0.4412	0.4924	0.5021	0.6112	0.0000	0.0000
ChildNet	nude	Accuracy	0.7632	0.6579	0.7234	0.7005	0.7374	0.7246	0.7401
		precision	0.5723	0.6214	0.6311	0.6343	0.6402	0.6932	0.6722
		recall	0.7615	0.6622	0.7221	0.7001	0.7432	0.7221	0.7413
		f1-score	0.6521	0.6431	0.6732	0.6713	0.6813	0.7121	0.7001
	safe	Accuracy	0.4473	0.6133	0.5833	0.6085	0.5942	0.6809	0.6397
		precision	0.6613	0.6523	0.6814	0.6833	0.7004	0.7233	0.7243
		recall	0.4523	0.6122	0.5831	0.6111	0.5914	0.6832	0.6441
		f1-score	0.5312	0.6331	0.6332	0.6443	0.6442	0.7014	0.6813

presented in Table 11.

As shown in Table 11, the CNN method can recognize the images with very low accuracy. Thus, CNN adds 7867 images from 7878 images into the nude class and allows 11 image errors, incorrectly adding them into the safe class, however correctly adding 11 images into the safe class and incorrectly adding 8097 images into the nude class. This means that the algorithm is almost unable to recognize images that fall into the safe class.

Table 11
The number of images classified by CNN algorithm in 500 iterations on NudeNet Classifier dataset v1.

Class	nude	safe	Total
nude	7867	11	7878
safe	8097	11	8108

The confusion matrix of the ChildNet method in 500 iterations on the NudeNet Classifier dataset v1 is presented in Fig. 16.

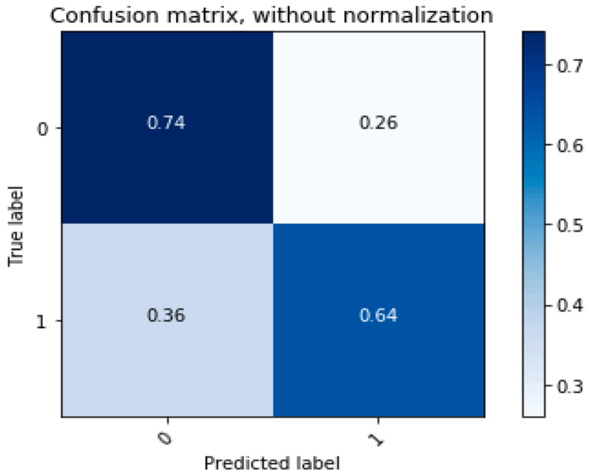


Fig. 16. The confusion matrix of the ChildNet model on NudeNet Classifier dataset v1 in 500 epochs.

Fig. 16 shows that in 500 iterations, the algorithm can classify the nude (0) class with 0.7401 accuracy performing 0.2598 errors, and the safe (1) class with 0.6397 accuracy performing 0.3602 errors. Also, as the number of iterations increases, the value of the ChildNet evaluation metrics approximates into the 1. This is the expected result. However, in 500 iterations, the CNN algorithm couldn't recognize safe class overall.

In the experiments Nvidia GTX geforce 1080 GPU with 11GB of memory was used for training the classification methods. 113,330 images of size 100 × 100 took 9 days of training time to achieve good results from the ChildNet. For 8000 images it took 2 days to train the model.

6. Conclusion

The paper proposed a method for filtering harmful image content on the Internet. The algorithm used a multi-layer neural network architecture consisting of five convolution units to determine whether the digital images are pornographic or not. The high recognition accuracy performed by the method in the testing process using real data proved the model to be the most effective tool in filtering Internet pornography. Future research suggests the development of a new effective method for recognizing the images to be the child pornography.

Authorship statement

All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript. Furthermore, each author certifies that this material or similar material has not been and will not be submitted to or published in any other publication before its appearance in the Journal of Information Security and Applications.

Authorship contributions

Please indicate the specific contributions made by each author (list the authors' initials followed by their surnames, e.g., Y.L. Cheung). The name of each author must appear at least once in each of the three categories below.

Category 1

Conception and design of study: R.M. Alguliyev; acquisition of data: S.S. Ojagverdiyeva; analysis and/or interpretation of data: F.J. Abdullayeva;

Category 2

Category 2 Drafting the manuscript: F.J. Abdullayeva; Revising the manuscript critically for important intellectual content: R.M. Alguliyev;

Category 3

Approval of the version of the manuscript to be published (the names of all authors must be listed): R.M. Alguliyev, F.J. Abdullayeva, S.S. Ojagverdiyeva.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

All persons who have made substantial contributions to the work reported in the manuscript (e.g., technical help, writing and editing assistance, general support), but who do not meet the criteria for authorship, are named in the Acknowledgements and have given us their written permission to be named. If we have not included an Acknowledgements, then that indicates that we have not received substantial contributions from non-authors.

Reference

[1] Alguliyev RM, Ojagverdieva SS. Conceptual model of national intellectual system for children safety in internet environment. Int J Comput Netw Inf Secur 2019;11(3):40–7.
[2] Mitchell KJ, Finkelhor D, Wolak J. The exposure of youth to unwanted sexual material on the internet. Youth Soc 2003;34(3):330–58.
[3] Protecting Children Against Harmful Content. Council of Europe. Strasbourg; 2009. p. 19.
[4] Sui L, Zhang J, Zhuo L, Yang YC. Research on pornographic images recognition method based on visual words in a compressed domain. IET Image Proc 2012;6(1):87–93.
[5] Choi B, Chung B, Ryou J. Adult image detection using Bayesian decision rule weighted by svm probability. In: Proceedings of the 4th IEEE international conference on computer sciences and convergence information technology (ICCIT'09); 2009. p. 659–62.
[6] Zhu H, Zhou S, Wang J, Yin Z. An algorithm of pornographic image detection. In: Proceedings of the 4th IEEE international conference on image and graphics (ICIG); 2007. p. 801–4.
[7] Lopes AP, Avila SE, Peixoto AN, Oliveira RS, Ara′ujo AA. A bag-of-features approach based on hue-sift descriptor for nude detection. In: Proceedings of the 17th European signal processing conference; 2009. p. 1552–6.

[8] Zhou KL, Zhou L, Geng Z, Zhang J, Li XG. Convolutional neural networks based pornographic image classification. In: Proceedings of the 2nd IEEE international conference on multimedia big data (BigMM); 2016. p. 206–9.

[9] Alguliyev RM, Aliguliyev RM, Abdullayeva FJ. Privacy-preserving deep learning algorithm for big personal data analysis. J Ind Inf Integr 2019;15:1–14.

[10] Open Sourcing a Deep Learning Solution for Detecting NSFW Images, https://yahooeng.tumblr.com/post/151148689421/open-sourcing-a-deep-learning-solution-for, https://github.com/yahoo/open_nsfw.

[11] Colmenares-Guillén LE, Velasco FJ. Filter for web pornographic contents based on digital image processing. Int J Combinat Optim Probl Inform 2016;7(2):13–21.

[12] Nian F, Teng L, Wang Y, Mingliang X, Wu J. Pornographic image detection utilizing deep convolutional neural networks. Neurocomputing 2016;210:283–93.

[13] Huang Y, Kong AW. Using a CNN ensemble for detecting pornographic and upskirt images. In: Proceedings of the IEEE 8th international conference on biometrics theory, applications and systems (BTAS); 2016. p. 1–7.

[14] Yin H, Xu X, Ye L. Big skin regions detection for adult image identification. In: Proceedings of the IEEE workshop on digital media and digital content management (DMDCM); 2011. p. 242–7.

[15] Zaidan AA, Karim HA, Ahmad NN, Zaidan BB, Kiah ML. Robust pornography classification solving the image size variation problem based on multi-agent learning. J Circ Syst Comput 2015;24(02):1–37.

[16] Nugroho HA, Hardiyanto D, Adji TB. Negative content filtering for video application. In: Proceedings of the 7th international conference on information technology and electrical engineering (ICITEE); 2015. p. 55–60.

[17] Yan CC, Liu Y, Xie H, Liao Z, Yin J. Extracting salient region for pornographic image detection. J Vis Commun Image Represent 2014;25(5):1130–5.

[18] Yu JJ, Han WS. Skin detection for adult image identification. In: Proceedings of the 16th international conference on advanced communication technology; 2014. p. 645–8.

[19] Sae-Bae N, Sun X, Sencar HT, Memon ND. Towards automatic detection of child pornography. In: Proceedings of the IEEE international conference on image processing (ICIP); 2014. p. 5332–6.

[20] Sharma J, Pathak VK. Automatic pornographic detection in web pages based on images and text data using support vector machine. In: Proceedings of the international conference on soft computing for problem solving (SocProS); 2011. p. 473–83.

[21] Yin H, Huang X, Wei Y. SVM-Based pornographic images detection. Software engineering and knowledge engineering: theory and practice, 115; 2012. p. 751–9.

[22] Shen X, Wei W, Qian Q. The filtering of internet images based on detecting erotogenic-part. In: Proceedings of the IEEE third international conference on natural computation (ICNC); 2007. p. 1–5.

[23] Rowley HA, Jing YS, Baluja S. Large scale image-based adult-content filtering. In: Proceedings of the 1st international conference on computer vision theory and applications; 2006. p. 1–7.

[24] Durrell K, Murray DJ. Pornographic image detection with Gabor filters. In: Proceedings of the applications of artificial neural networks in image processing; 2002. p. 1–9.

[25] Schettini R, Brambilla C, Cusano C, Ciocca G. On the detection of pornographic digital images. In: Proceedings of the visual communications and image processing; 2003. p. 2105–13.

[26] Sajith M. nsfw-v1 dataset, https://github.com/sajithm/nsfw-v1.

[27] NudeNet Classifier dataset v1, https://archive.org/details/NudeNet_classifier_dataset_v1.