IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 40, No. 6, December, 2023, pp. 2539-2552
Journal homepage: http://iieta.org/journals/ts

Leveraging Deep Learning for Identification of Illicit Images in Digital Forensic Investigations

Mustafa Eriş*, Mustafa Kaya

Department of Digital Forensics Engineering, College of Technology, Firat University, Elazig 23100, Turkey

Corresponding Author Email: meris@firat.edu.tr

https://doi.org/10.18280/ts.400617

Received: 24 April 2023
Revised: 21 June 2023
Accepted: 12 September 2023
Available online: 30 December 2023

Keywords:
deep learning, pornography detection, child sexual abuse detection, digital forensics, convolutional neural networks, obscene image detection

ABSTRACT

Background: The distribution of illicit material such as obscene images and child sexual abuse (CSA) content constitutes a serious crime in numerous jurisdictions worldwide. The identification and detection of such content is a crucial component of forensic investigations, and is integral to the apprehension and prosecution of offenders. Traditional methods for identifying such material are predominantly manual, requiring expert review, and are consequently time-consuming and susceptible to human error. Deep learning, with its proficiency in discerning complex patterns and features in large-scale data, offers a promising alternative for the automated and accurate detection of obscene and CSA content. This research advances a deep learning model for detecting obscene images and CSA in digital forensic evidence. Methods: A dataset of 3000 obscene and 3000 non-obscene images was compiled, with the obscene images sourced from a social sharing platform, Reddit. Images were classified as obscene if they portrayed sexual organs or activities for the primary purpose of eliciting sexual arousal. A convolutional neural network (CNN) based deep learning model was then developed to detect the obscene content. The efficacy of the proposed model was compared with existing methodologies using the NPDI benchmark dataset. Pre-trained CNNs were used to extract feature vectors from the images which were subsequently used in conjunction with a neural network classifier to categorise the images. To detect CSA, the UTKFace dataset was employed to identify minors in images, using a lightweight CNN model with skip connections to recognise juvenile faces. Results: The proposed model for obscene content detection demonstrated a robust performance, achieving an accuracy of 99.8% and 99.4% respectively in the training and testing segments of the NPDI dataset. This model has potential applicability to both image and video files. Meanwhile, the model for identifying minors achieved an accuracy of 99.6% and 99.0% in the training and testing segments of the UTKFace dataset respectively. Conclusion: The findings of this study underscore the efficacy and high performance of the deep learning models proposed for the detection of obscene images and CSA content. These results highlight the potential for these models to be employed in digital forensic investigations for automated content detection, which would significantly advance efforts in combatting the distribution of illicit content.

1. INTRODUCTION

1.1 Background

In the contemporary digital age, facilitated by the proliferation of the Internet and technological advancements, access to diverse data types has become increasingly seamless. This ease of access undeniably enhances various facets of life, including education, science, and healthcare, but it simultaneously augments the dissemination of harmful content.

Numerous online content-sharing platforms are monitored by Law Enforcement Agencies (LEAs) and automatic content recognition systems to curb the propagation of inappropriate content on the Internet [1]. Upon detecting such content, countermeasures range from blocking the offending platforms or websites to removing the harmful content. Content intended for specific age groups also necessitates stringent control systems, particularly to safeguard children from potential online hazards.

Globally, content deemed inappropriate for public consumption predominantly involves the sharing of pornographic material. Despite numerous platforms implementing restrictions or outright bans on the distribution of such content, the illicit dissemination of this material persists [2].

In the current landscape, access to sites and platforms dedicated to sharing pornographic content varies greatly, with outright prohibition in some countries, and age-verified accessibility in others. Furthermore, child abuse and child pornography (collectively referred to as CSA) are universally recognized as criminal activities [3]. The detection and blocking of such criminal content, particularly after it has been disseminated, becomes critically important.

The increasing accessibility and demand for such content heightens the probability of encountering this data in crime-related evidence. Additionally, pornographic content disseminated on the Internet often finds its way into digital media. The rapid detection of such content not only aids in blocking data published on the Internet but also proves instrumental in identifying objectionable content during forensic digital evidence investigations, thereby expediting crime resolution. The detection of pornographic content on confiscated digital media can significantly accelerate the comprehensive investigation of the crime [4].

1.2 Literature review

1.2.1 Early studies based on traditional techniques for pornography detection

Numerous methodologies for detecting pornographic content in video and image data have been proposed in academic literature. A common objective in many of these previous studies has been the detection of human skin, given its prevalence in most pornographic material. Techniques such as color-based identification or manually applied features to localized areas were employed to discern human skin [5-7].

However, these methods possess a significant limitation: all data in which skin color is dominant can be erroneously classified as pornographic content. For instance, skin color is prominently observed in certain contexts such as sports competitions, beach scenes, and fashion shows, yet these scenarios are not pornographic in nature. Consequently, the use of skin color detection as an isolated technique can result in a high rate of false positives, reducing its effectiveness as a standalone tool. Moreover, its inability to function effectively with grayscale data allows for easy circumvention of the skin color detection method.

To achieve more accurate results, subsequent studies used classifiers with multiple manually extracted regional features. Some studies have employed the Bag-of-Visual-Words (BoVW) method in conjunction with features like SIFT, SURF, and ORB, using classical machine learning classifiers such as Support Vector Machines (SVMs) [8-12].

1.2.2 Deep learning-based studies for pornography detection

While traditional feature extraction methods have achieved partial success, they have not yet reached the desired level in terms of reducing false positives and augmenting overall accuracy. The promising results garnered from applying deep learning techniques to image data have incentivized researchers to utilize these methods for detecting pornographic content. Early studies in this domain recommended leveraging successful ImageNet pre-trained models as feature extractors. Some of these studies fused the decisions from multiple pre-trained models to detect pornography in images [13]. In contrast, other studies introduced custom CNN architectures for this purpose, as exemplified in the study [14]. However, these custom architectures often require additional data for training, making it challenging to compare these studies with the broader literature.

Numerous studies have explored diverse approaches to enhance the detection of inappropriate content in images and videos. Some researchers have proposed integrating motion data from videos with image data for training. For instance, Perez et al. [15] combined GoogleNet image features with motion features extracted using optical flow and MPEG motion vectors, achieving an impressive accuracy of 97.9% on the NPDI dataset. Wehrmann et al. [16] used GoogleNet and ResNet CNN models for image data feature extraction and employed the LSTM model to capture temporal connections in video data, achieving a peak accuracy of 95.6%.

Various methods for detecting and safeguarding against pornography have been explored across different domains. Yousaf and Nawaz [17] introduced a model for automatically filtering inappropriate content in YouTube videos. Their model utilized the EfficientNet-B7 CNN model for extracting features from video frames and incorporated a BiLSTM model with 128 hidden units to capture temporal information, achieving a commendable accuracy of 95.66%. Gautam and Vishwakarma [18] proposed a Frame Sequence ConvNet that employed the ResNet-18 backbone to analyze sequential features of NPDI videos. By extracting features from sequential frames using the ResNet-18 model, their approach reached a peak accuracy of 98.3% on the NPDI dataset. Furthermore, Yang and Xu [19] developed a generative adversarial network (GAN) model for automatically censoring inappropriate content in educational platform streams. They trained their model specifically on indoor images from the NPDI dataset for binary classification, achieving an average accuracy of 98% in detecting inappropriate content within videos.

Conversely, some studies have reframed the detection of pornographic images as an object detection problem. For example, Mallmann et al. [20] aimed to censor specific body areas using object detection models like the Faster R-CNN and SSD with different backbones. They achieved a maximum mean average precision (mAP) of 63.5% using the Faster R-CNN model with the Inceptionv2 backbone. Similarly, Hor et al. [21] aimed to detect inappropriate body parts in videos using object detection models, with the EfficientDet model achieving the highest accuracy, averaging 75% on a dataset derived from the NPDI dataset.

1.2.3 Studies about age estimation

Historically, studies on age determination have relied on handcrafted features for facial and body analysis [22]. For instance, Hajizadeh and Ebrahimnezhad [23] achieved an 87% accuracy in classifying images into four age groups by utilizing probabilistic neural networks and HOG features extracted from various facial regions. Eidinger et al. [24] deployed LBP and FPLBP methods in tandem with dropout-SVM to determine age and sex from facial images, obtaining accuracies of 66.6% and 45.1% on the Gallegher and Adience datasets, respectively, for 7 and 8 age groups. Sai et al. [25] employed LBP, Gabor, and biological features, coupled with ELM classification, achieving a 70% accuracy for age group classification.

Recent studies have underscored the superiority of deep learning models, particularly Convolutional Neural Networks (CNNs), over traditional approaches. One of the pioneering studies by Wang et al. [26] introduced a two-layer CNN model, which outperformed handcrafted features in age estimation tasks. Chen et al. [27] proposed a ranking-CNN model, which enhanced age estimation results by amalgamating binary CNN classifiers. Anda et al. [28] developed a VGG-16-based CNN model, demonstrating successful results in identifying borderline age groups. Castrillón-Santana et al. [29] improved the performance of non-adult face classification by integrating local face descriptors with deep CNN model features.

In the context of Child Sexual Abuse (CSA) detection, age determination using facial features has gained considerable attention, largely due to the availability of labeled data. Sae-Bae et al. [30] proposed a CSA detection system that combined skin tone detection and age estimation from facial images, achieving accuracies of 83% and 96.5% in detecting explicit-like images and child faces, respectively. Yiallourou et al. [31] considered various factors such as age, gender, lighting, and the number of people in the pictures to classify CSA material, utilizing LBP features for age estimation. Macedo et al. [32] integrated age estimation and pornography detection, achieving an accuracy of 79.84% by identifying faces with the MTCNN face detector and classifying child faces using a fine-tuned VGG-16 CNN model. Similarly, Gangwar et al. [33] proposed an attention-based CNN model for age classification and pornography detection, achieving accuracies of 92.7% and 97.1% on CSA and NPDI dataset classifications, respectively.

In our work, we propose a system with fewer parameters than other models and also test the model with a novel dataset collected from the internet. Given that transfer learning is known to enhance accuracy [34] when data is limited, we employed models trained with the ImageNet dataset, which contains considerably more data than pornography benchmark datasets, in our proposed model.

1.3 Problem statement

The surge in data directly influences the volume of data potentially associated with crime [35]. Digital forensics encompasses the processes of preventing crimes before they occur, responding as they are committed, and revealing and examining evidence after a crime has been perpetrated. The impact of today's data growth is palpable in this field [36]. Particularly after crimes are committed, the demand for experts needed to scrutinize the evidence is escalating daily. At this juncture, there is a clear need for software tools that provide automatic detection and analysis methodologies for data expected to be investigated in digital environments [37].

Various commercial and open-source software tools are employed in forensic evidence analysis [38]. These tools ease the examination process and present the data in the digital environment in a comprehensive and understandable manner. The tools include features such as data carving, visualization, indexing, data type grouping, performing operating system-specific analyses, and reporting. However, conveying the content analysis of this data to the examiner often constitutes the most exhausting and time-consuming portion of the investigation. To expedite this step, efforts are being made to develop software capable of performing automatic content analysis, particularly for image data [39-41].

Nonetheless, evidence examiners using automatic content analysis tools in digital evidence investigations have reported that these tools typically function by filtering the hash values of known data. This is a limiting factor for the evidence review process, indicating a need for tools that evaluate content without dependency on hash values [42]. Additionally, a significant proportion of participants in this study (61.7%) reported experiencing sluggish performance with the tools used, while a considerable percentage (23.4%) voiced concerns about the accuracy of the applied software. These challenges underscore the immediate need for improved tools and software solutions that enhance efficiency and accuracy.

In conclusion, the exponential increase in data directly affects the amount of data linked to criminal activities, thereby necessitating advancements in digital forensics. The critical need for automated detection and analysis methods is evident, especially considering the rising number of experts required to scrutinize evidence post-crime. While many commercial and open-source software tools facilitate the examination of digital evidence, content analysis remains a daunting and protracted task. This underscores the pressing need for tools that can appraise content without solely relying on hash values, and that also address concerns related to performance and accuracy. The findings of this study highlight the urgency of developing superior tools and software solutions that can boost efficiency and accuracy in evidence examination, thereby tackling the challenges faced by practitioners in the field.

1.4 Motivation and contributions

Motivations:

This study is motivated by the identified limitations and challenges that forensic practitioners encounter in the realm of digital forensics. The burgeoning volume of data, dependencies on hash values, and concerns surrounding performance and accuracy have collectively underscored the need for enhanced tools and software solutions. By addressing these motivations, this study seeks to advance the development of automated content analysis methods to augment the efficiency, reliability, and comprehensiveness of digital evidence examination. The ultimate aim is to equip practitioners with advanced tools that can expedite the investigative process and facilitate more effective analysis of digital data in forensic investigations.

In this study, we propose the use of the OCDet model in digital evidence investigations to automatically present pertinent data to the examiner. In addition, we introduce the IMPD model to detect underage individuals in explicit content, should it exist. Our goal is to swiftly and automatically detect CSA material using the proposed dual-model system. With this proposed system, the digital evidence examination process will be expedited, and the demand for manpower will be reduced. While the developed model is intended for forensic evidence examination, its near real-time operation speed allows for its application in areas such as broadcast content censorship and online media analysis.

Contributions:

This work presents the following contributions:

- We propose a robust deep learning-based model to assist in forensic evidence investigations.
- We present a new dataset, collected from the internet, for obscene image detection.
- Our proposed OCDet model, to our knowledge, achieves state-of-the-art performance on the NPDI benchmark dataset.
- We propose a lightweight, high-performance IMPD CNN model with skip connections to detect underage faces in images.
- We propose a CSA detection system by integrating the aforementioned models.
- Both proposed models boast an accuracy rate of over 99%.

2. MATERIAL

A deep learning-based model was proposed using two different datasets to detect obscene data. One of the datasets used is the NPDI dataset [11], which is widely used in the literature. In addition, a second data set was created by making use of the pictures obtained from the internet links and published publicly. These two datasets contain a variety of backgrounds, people from different ethnicities, and difficult and easy examples. The UTKFace [43] dataset is also used for immature face detection in this work. The UTKFace dataset consists of images that include faces of people from different ethnicity, age, and gender.

2.1 Pornography-800 dataset

The Pornography-800 dataset (NPDI Dataset) contains video data in two different categories (Porn-nonPorn). This dataset includes 400 pornographic and 400 non-pornographic videos. The non-pornographic videos are divided into 2 levels and these levels are named easy and hard levels. Easy-level non-pornographic videos contain landscapes without people, everyday pictures without human skin density, etc. The hard-level non-pornographic videos show wrestling competitions, people on the beach, etc. The length of the videos in the dataset range from 3 minutes to 30 minutes. These long videos in the dataset were divided into shorter shots, and each shot's key frame was saved as an image. Each video is divided into an average of 20 shots. The resulting dataset has 16727 frames (images) extracted from the videos. In this way, a dataset was created to detect pornographic content on both video and pictures. The NPDI dataset was used to have large data and include compelling data in this study. Sample images from the NPDI dataset are shown in Figure 1. As shown in Figure 1, the dataset contains different images with various backgrounds and difficulties.

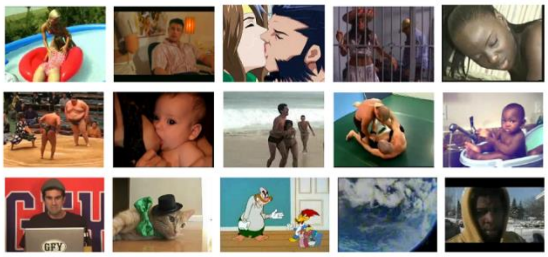


Figure 1. NPDI dataset sample images

Additionally, it should be noted that the dataset used in this study consisted of videos with varying quality and content, which led to the inclusion of mislabeled images or images of very low quality. It is important to highlight that the image data was extracted from the frames within these videos. Therefore, the training set was inspected manually, and 1711 images were moved to non-pornographic class. Samples from removed images are shown in Figure 2.

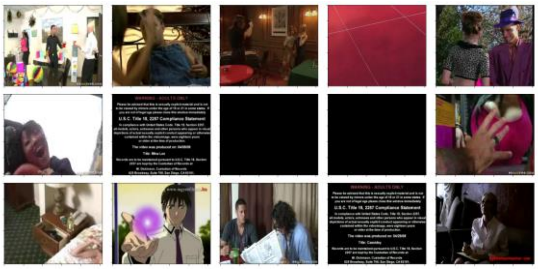


Figure 2. Removed images from the NPDI dataset

As can be seen in Figure 2, as these images are extracted from videos, some images may not include any humans or are just not pornographic. Therefore, these images were used as negative examples in the training phase.

2.2 Internet NSFW dataset

A second dataset was used to test the generalization ability of the proposed model, which achieved successful results in the experiments performed on the NPDI dataset. The second dataset was created using the image links published on the GitHub platform by Bazarov [44]. In the published project, there are links to 1,589,331 pictures in total, divided into 159 different categories. This amount corresponds to approximately 500 GB of data. However, the images in these links are not verified (incorrectly labeled data may be found), corrupted data may be found, and duplicate data may be found. This study created a new data group with 3000 images randomly selected from these links to be used in the obscene class. Since the source of the created data set is the data shared openly on the Internet by people, it is possible that the images are deleted over time, that they are shared on the same platform more than once, or that they have very different content. For this reason, the data must be passed through a preprocessing step. After downloading images, all data was controlled and validated by a participant and authors manually. In the validation step, all images were checked for corrupted or irrelevant content. Images were labeled as obscene if they primarily intend to elicit sexual arousal in individuals through depictions of sexual organs or activities. Nearly all images in obscene class have a person with explicit body parts which are accepted as obscene in studies [20, 21]. It is important to emphasize that within this dataset, data explicitly depicting organs considered obscene in the aforementioned studies are categorized as "obscene," while the labeling does not encompass any obscenity inferred from poses or diverse facial expressions. Furthermore, it is noteworthy that the dataset predominantly comprises indoor images, with outdoor images representing a minority. While this distribution aligns with the data commonly encountered in both online sources and forensic investigations, it raises a concern that should not be overlooked. In addition to collected obscene images, for the creation of non-obscene images, 3000 randomly selected images from the Google Open Images dataset, 90% of the images obtained from there have “human” labels which means these images contain people in them. This approach has significantly enhanced the utility of the dataset in distinguishing between individuals with normal conditions and those exhibiting explicit behaviors, thereby improving its efficacy for accurate classification. Sample images from the newly created dataset are shown in Figure 3.

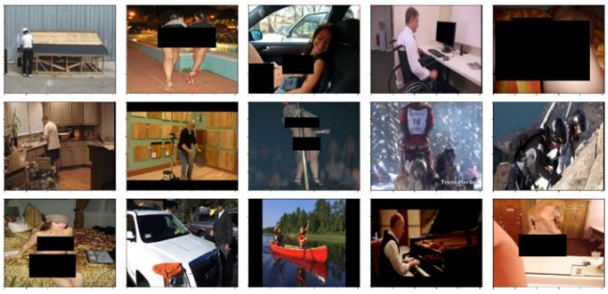


Figure 3. Sample images from the newly created dataset

After creating and preprocessing datasets, image counts of each dataset are shown in Figure 4. As shown in Figure 4, the NPDI dataset contains approximately 6000 explicit and 10000 non-explicit images in two groups classified as easy and difficult. In the NFSW data set, there are approximately 3000 data belonging to the explicit and non-explicit classes. Due to the imbalance of the classes in the NPDI data set, the data from the non-explicit class, where the hard and easy data are mixed, were randomly selected to be equal to the number of data in the training set. In this way, it was ensured that the data set was balanced.

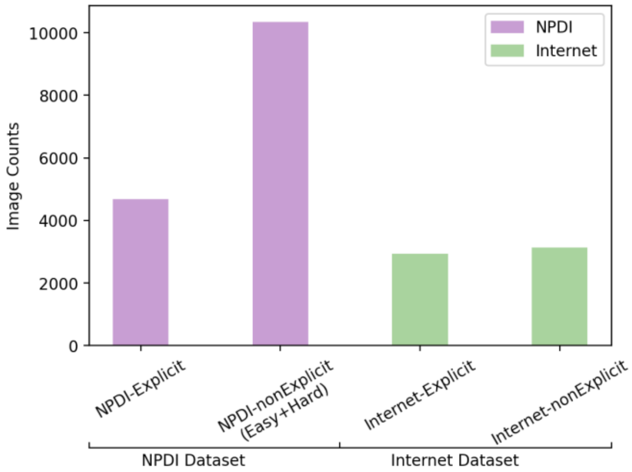


Figure 4. Image counts of final datasets

2.3 UTKFace dataset

UTKFace dataset is a large-scale face dataset with a long age range from 0 to 116 years old. This dataset contains over 20,000 face images. These images are labeled according to the age, gender, and ethnicity of the people. In addition, these images cover large variations in pose, facial expression, illumination, occlusion, and resolution. Sample images of the UTKFace dataset are given in Figure 5.

Figure 5. Sample images from the UTKFace dataset

In the UTKFace dataset, all images are labeled with their age and other information. Since we only categorize people into two classes (mature and immature), we labeled images of people under 18 as immature. All other images are used in mature classes. The age distribution of all images is given in Figure 6. In our experiments, we used cropped face images from this dataset to make the model more robust.

We annotated all images with two labels. After the annotation process, we obtained 4229 images for the immature class and 19475 images for the mature class. Although there was an imbalance between the obtained classes, there was no overfitting problem in the experiments, since there was sufficient data for the minority immature class.

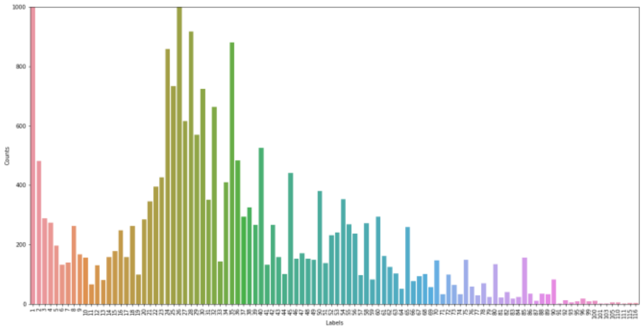


Figure 6. Age distribution of UTKFace dataset

3. PROPOSED METHOD

The proposed method consists of two parts: the obscene detection part and the age detection part. The model has four main steps namely preprocessing, feature extraction, classification and decision fusion.

Image preprocessing: Image preprocessing pipeline consists of two primary elements: image resizing and image normalization. Image resizing is performed to standardize the dimensions of all input images. By resizing images to a consistent resolution or aspect ratio, we eliminate variations in size and simplify the subsequent computational processes. Following image resizing, image normalization is applied to adjust the pixel values of the images to a common scale. This process involves techniques such as mean subtraction and min-max scaling. Normalization helps to mitigate the impact of varying pixel value ranges across images, promoting fair comparisons and improving the convergence of subsequent analysis algorithms. Additionally, within our proposed model, a crucial preprocessing step involves the detection and cropping of all faces present in the images. This essential process is seamlessly integrated into our preprocessing pipeline. The formulas of preprocessing steps are given in Eqs. (1)-(5).

Feature extraction: In our proposed model, a significant stage of the pipeline is dedicated to feature extraction. This step involves the utilization of two distinct Convolutional Neural Network (CNN) models designed for specific tasks: obscenity detection and immature face detection. One of these models comprises a pretrained CNN, while the other is newly created. These CNN models are employed to extract deep features from preprocessed images.

The pretrained CNN model, which has been trained on a large-scale dataset, possesses the capability to learn high-level representations of general image features. By utilizing this pretrained model, we leverage the knowledge gained from extensive prior training to extract discriminative features related to obscenity detection.

For the task of immature face detection, a separate CNN model is specifically designed and trained from scratch. This model is tailored to capture facial features associated with immaturity, enabling effective identification of immature faces in the images. The formulas for extracting features from image input with CNN models are given in Eq. (6).

Classification: In the classification step of our proposed model, we employed fully connected layers to classify the deep features extracted from the images. These fully connected layers serve as the final component of our model's architecture, responsible for mapping the extracted features to the corresponding class labels.

The deep features obtained from the feature extraction step encode rich and discriminative information about the input images. However, they are in a high-dimensional space and not directly interpretable for classification purposes. Therefore, we utilize fully connected layers, also known as dense layers, to transform these deep features into meaningful predictions. Formulas used for transforming deep features into predictions are given in Eqs. (7)-(9).

Decision fusion: Finally, the results obtained from the two parts were combined in the decision fusion step, and the CSA detection classification was performed. The formula for fusing decision of the classification results are given in Eq. (10).

The flow diagram of the proposed method is given in Figure 7.

The proposed model exhibits a reduced number of trainable parameters in comparison to other relevant studies documented in the literature. For instance, a comparable model in [33] was reported to possess 9M parameters for tasks involving pornography and age group detection. In contrast, our model consists of merely 4M parameters, achieving an approximate reduction of 50% in terms of trainable parameters. Moreover, our model demonstrates superior performance on the NPDI dataset for pornography detection. Conversely, in contrast to the approach in [15, 16, 19] that involves the utilization of continuous data types such as video, our training process employs a reduced amount of data. In conclusion, our model outperforms alternative approaches while simultaneously requiring a smaller dataset and less computational power.

In the proposed model, we used CNN models to classify obscenity and immaturity. The model gets input data and applies the face detection algorithm to the input image to get faces. After face detection, face images and original input images are fed to CNN architectures as shown in Figure 7. For the obscene detection task, we used a pre-trained MobileNetV3-large CNN model. This model was first used only as a feature extractor for training newly added fully connected layers by freezing the weights in the network. After a short training, we unfree all layers except the batch normalization layer of the MobileNetV3-large model and trained the network secondly in an end-to-end fashion. The second training was performed with a much lower learning rate not to distort already trained layer weights. The architecture of the MobileNet-based model OCDet model is given in Figure 8.

Although the MobileNetV3-large model looks small in Figure 8, it is a large model with 69 convolution layers, excluding the classification layer. At the same time, we also processed face images with other skip connection CNN that we created for immature detection. The architecture of the IMPD model is given in Figure 9. As seen from here, our network has one skip connection and 5 convolution layers followed by two dense layers with dropout. This model is a lightweight model according to the OCDet model.

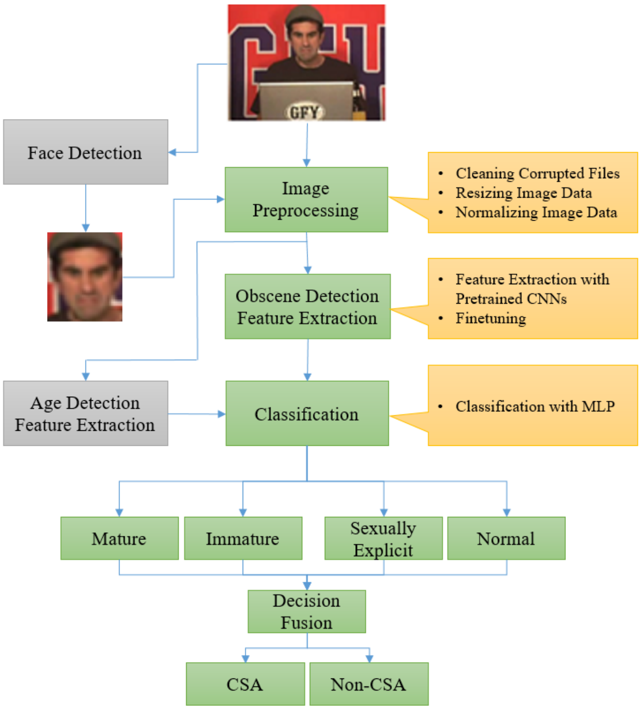


Figure 7. Flow Diagram of the proposed method

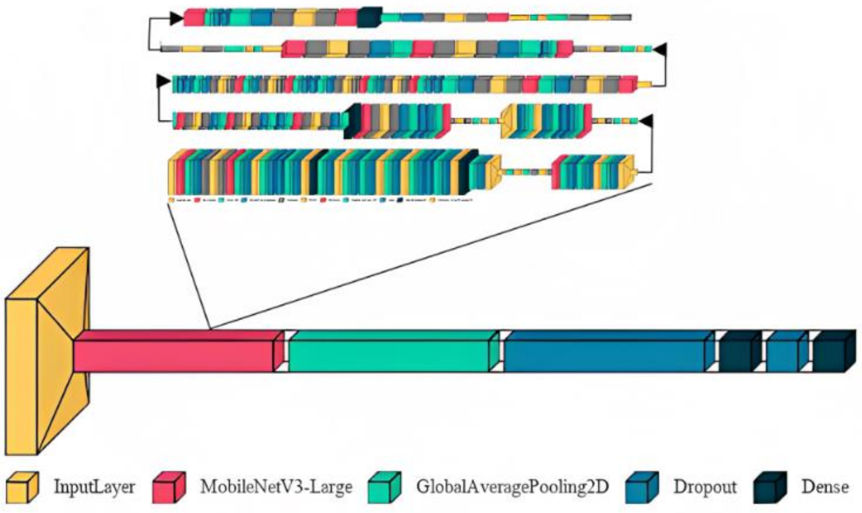


Figure 8. The architecture of the fine-tuned OCDet CNN model

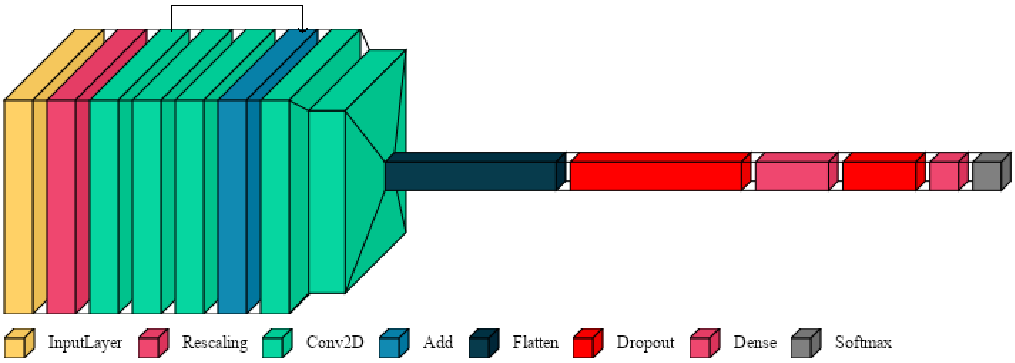


Figure 9. The architecture of the IMPD model

We trained the IMPD model from scratch in an end-to-end fashion. As a result, both models get successful results in the relevant tasks, ensuring that the result obtained in the decision fusion stage is robust. Parameter numbers, kernel sizes, output sizes, and other details of the models are given in Table 3 and Table 5.

The process steps of the proposed model are as follows.

Step 0: Resize and rescale the input images.

Since the datasets contain various-sized images, we resized images of NPDI and the manually collected dataset into 224 × 244. On the other hand, the UTKFace dataset cropped face images were already resized into 200 × 200 images, and we used these images. The formula notation of the resizing process is given in Eq. (1).

$$X^{(i)} = R(X^{(i)}, [h, w]) \tag{1}$$

where, $X^{(i)}$ represents the i.th image in the dataset, and R is the resizing function that resizes input images to h × w sized images. In this way, it is ensured that all images are 224 × 244 in size. After resizing, images are rescaled using the method in Eq. (2).

$$X^{(i)} = X^{(i)}./255 \tag{2}$$

where, $X^{(i)}$ is the i.th input image. All pixel values are rescaled into an interval of [0,1] by dividing every pixel value of images by 255. This way, we prevent our models from the exploding gradient problem. In step 0, Eq. (1) is applied to all datasets except the UTKFace dataset, while Eq. (2) is applied to all datasets.

Step 1: Detect faces in images and preprocess face images.

Face detection and preprocessing are performed by Eqs. (3), (4), and (5). In Eq. (3) faces in images are cropped and face images are resized and rescaled using Eqs. (4) and (5).

$$\begin{gathered} F^{(i)} = DetectFace\left(X^{(j)}\right) \\ i \in \{1,2,\ldots,m\}, j \in \{1,2,\ldots,n\} \end{gathered} \tag{3}$$

$$F^{(i)} = R(F^{(i)}, [h, w]) \tag{4}$$

$$F^{(i)} = F^{(i)}./255 \tag{5}$$

where, $F^{(i)}$ denotes the i.th face image detected in the j.th image. Here m is the number of faces detected from images in the dataset. And n is the number of images in the dataset, R is the resizing function. DetectFace is the face detection function that returns face images cropped from the original images. After then this, as shown in Eq. (4), cropped face images are resized into h x w since our IMPD model expects images in 200 × 200 × 3 shape. Finally, in Eq. (5), resized face images are also rescaled.

Step 2: Feed original images and face images to CNN models.

In Eq. (6), the calculation of output in a CNN layer is given.

$$X_{i,j}^{l} = \sum_{s1=0}^{s1-1} \sum_{s2=0}^{s2-1} w_{s1,s2} X_{(i+s1)(j+s2)}^{l-1} \tag{6}$$

where, X^{l-1} is the input image, *i*, and *j* are the pixel indexes of images, *l* is the layer index, *s1*, and *s2* are the kernel sizes of the convolution layers and X^{l} is the output of the l. th CNN layer. In this situation, *X0* is the original image data. Here every new output is calculated by multiplying inputs with corresponding w weights of the convolution layer. And the input of convolution layers is the output of the layer before the l.th layer. The network architectures of the CNN models are given in Figure 8 and Figure 9.

Step 3: Classify images using MLP classifiers.

In Eq. (7), the calculation of output in a fully connected layer is given.

$$z^{(i)} = a(I * W^{(i)} + B^{(i)}) \tag{7}$$

where, $z^{(i)}$ is the i.th fully connected layer output in MLP, *I* is the input of the fully connected layer, $W^{(i)}$ is the weights of the fully connected layer, $B^{(i)}$ is the bias values of the fully connected layer, and a is the activation function. In our experiments, we used the ReLU activation function in CNN and MLP layers except for the last layer of the MLP, wherein we used the softmax activation function.

Step 4: Get softmax probabilities of predicted classes and final predictions.

In softmax activation step, outputs of the model are calculated using Eq. (8) and then the final decision is calculated as given in Eq. (9).

$$sfmx(z) = \begin{cases} o^{(i)} = \dfrac{e^{z^{(i)}}}{\sum_{j=1}^{k} e^{z^{(j)}}} \\ i \in \{1,2,3,\ldots k\} \end{cases} \tag{8}$$

$$p = \text{Max}(o) \tag{9}$$

where, *sfmx* is the softmax activation function and *z* is the input vector to the softmax activation function. Here *e* is the exponentiation constant, $o^{(i)}$ is the i.th element of the output probabilities, $z^{(i)}$ is the *i*.th element of the input vector, and *k* is the number of classes. Finally, *Max* is a function to find the index of the maximum value in vector *o* and obtain class prediction *p*. This step is applied to both CNN model outputs, and two classification results are obtained.

Step 5: Combine prediction results to decide CSA existence as given in Eq. (10).

$$csa(p1, p2) = \begin{cases} 1, p1 = p2 = 1 \\ 0, p2 = 0 \end{cases} \tag{10}$$

where, *csa* is the function for deciding CSA existence. *p1* and *p2* are predictions of IMPD and OCDet. These prediction values get the value 1 if the content contains immature or obscenity; otherwise, they take the value 0. When two predictions are 1, the image contains both immature and obscenity which means it probably contains CSA.

4. RESULTS

The proposed CNN-based IMPD and OCDet models were programmed using Python programming language on a personal computer with 32 GB memory, an Intel Xeon E5 processor, and an 8 GB graphic processor unit with Cuda. We used classification evaluation metrics in our experiments to observe our proposed models. The mathematical formulas of the performance metrics used in experiments are given below [45] in Eqs. (11)-(14).

$$Precision = \frac{TP}{TP+FP} \tag{11}$$

$$Recall = \frac{TP}{TP+FN} \tag{12}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{13}$$

$$F1 - Score = 2 \times \frac{Precision\ x\ Recall}{Precision+Recall} \tag{14}$$

where, *TP, TN, FP*, and *FN* are the counts of true positives, true negatives, false positives, and false negatives, respectively. These values are obtained from the confusion matrices of the proposed models. The resulting confusion matrices of the OCDet and IMPD models are given in Figure 10.

As can be seen from Figure 10, the classification accuracies of the proposed models are 99.35% and 99.03% for obscenity and immature detection. The training curves of the two models are given in Figure 11.

According to Figure 11, models are not over-fitted and achieve high accuracies on both training and test data. To understand the performance of the models in more detail, we extracted TP, TN, FP, and FN values from confusion matrices. With these values, the performances of the proposed models are calculated using Eqs (11)-(14) according to various evaluation metrics. Obtained results are given in Table 1.

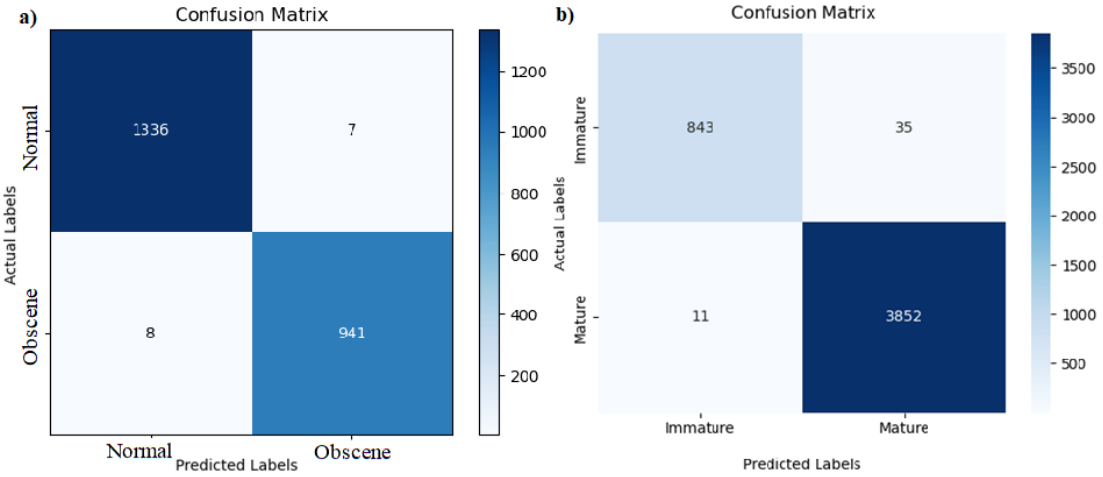
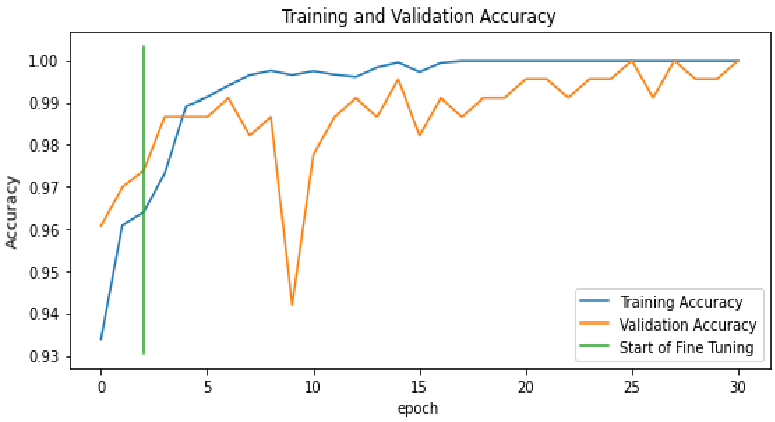
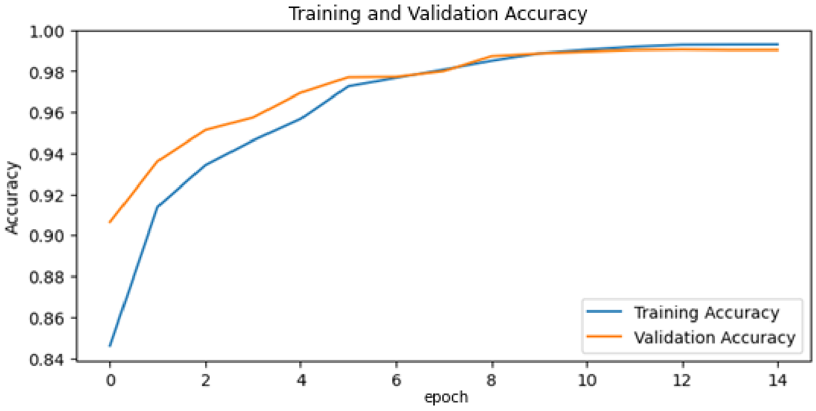


Figure 10. a) Confusion matrix of the OCDet model b) Confusion matrix of the IMPD model



(a)



(b)

Figure 11. (a) Training curves of the OCDet model (b) Training curves of the IMPD model

Table 1. Evaluation results of the proposed models

	OCDet Model				IMPD Model			
	TP	TN	FP	FN	TP	TN	FP	FN
	941	1336	7	8	843	3852	11	35
Precision	0.9926				0.9871			
Recall	0.9916				0.9601			
Accuracy	0.9935				0.9903			
F1-Score	0.9921				0.9734			

As shown in Table 1, both models achieve over 99% success on the NPDI and UTKFace datasets. The evaluation results guarantee the accuracy of the proposed CSA detection model by combining the obtained results from the two models. Unfortunately, it has not been possible to test it on real data since it is illegal to access such data.

5. DISCUSSION

In this study, we proposed two CNN models for detecting obscenity and immaturity in image and video data to detect CSA finally. We proposed two models for the specific tasks; IMPD for immature person detection and OCDet for obscene content detection. The OCDet model achieved an accuracy of 99.8% which is state-of-the-art accuracy in classifying the NPDI dataset. And it has fewer trainable parameters than other successful studies in the literature. Similarly, a lightweight CNN model with residual blocks is proposed to classify immature faces in images. This model also demonstrated exceptional performance on the UTKFace dataset, achieving an impressive classification accuracy of 99.0%.

For the OCDet model, we first used a pre-trained model on the ImageNet dataset as a feature extractor, and then we fine-tuned this model to obtain the best results. For the selection of this model, we compared different popular CNN models' feature extraction performances by adding three fully connected layers after the CNN models listed in Table 2 and only trained newly added layers with features obtained from the CNN models. An overview of the model created with pre-trained models is given in Figure 12.

In Figure 12, *m, n,* and *k* are dimensions of the output activation maps of the CNN models -could have different values for different models- and *p* is the probability of the predicted class. After the global average pooling layer, the activation map is resized by calculating the average along the first two axes. n denotes neurons in fully connected layers. We added two fully connected layers with 512 units after the pre-trained model. In the training phase, we deactivated half of the connections with the dropout layer to prevent the model from overfitting. Obtained results of various CNN models are given in Table 2.

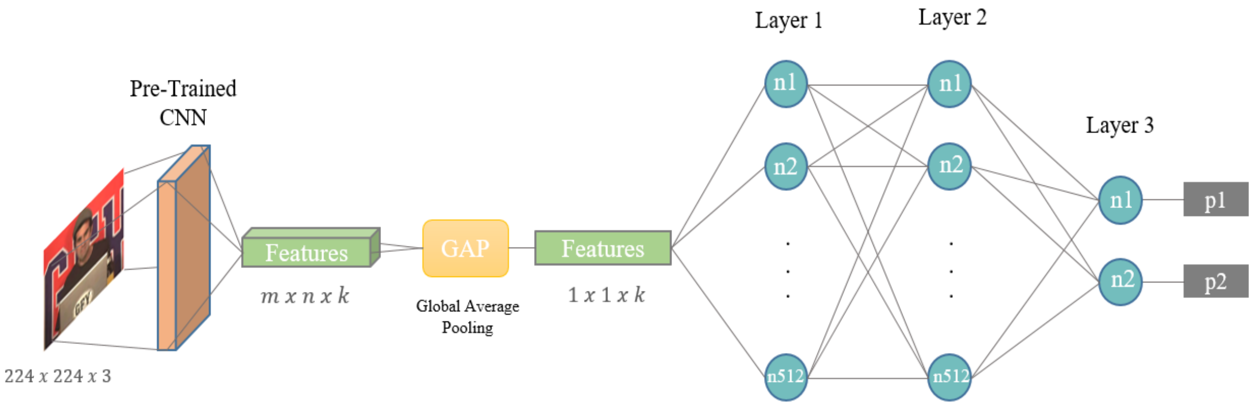


Figure 12. The architecture used to compare pre-trained CNNs

Table 2. Obtained results on the NPDI dataset without fine-tuning using pre-trained CNNs

	Model	Train Accuracy	Test Accuracy	Average Inference Time (sn)
1	Xception	0.9844	0.9664	0.44
2	VGG19	0.9611	0.9642	0.18
3	VGG16	0.9474	0.9581	**0.09**
4	InceptionV3	0.9785	0.9585	0.68
5	InceptionResNetV2	0.9776	0.9685	1.07
6	EfficientNetB4	0.9887	0.9673	1.04
7	EfficientNetB0	0.9914	0.9764	0.57
8	ResNet50	0.9790	0.9734	0.44
9	ResNet101	0.9881	0.9773	0.55
10	MobileNetV3Large	0.9891	**0.9786**	0.46
11	DenseNet121	0.9805	0.9666	0.85
12	EfficientNetB7	**0.9948**	0.9742	0.41
13	MobileNetV2	0.9124	0.9219	1.33

Table 2 shows the best performance on test data belonging to the MobileNetV3-large model. Besides, the second-best performance on training data also belongs to it. Some models' test results are higher than training because dropout layers are used in the created model. During the training phase, the dropout layer prevents some activations from being active and ensures that the whole model is trained in a balanced way, but all model activations are used during testing. This sometimes causes test results to be higher. According to the average inference times (calculated using 10 random images), the best models are VGG models, which can process 10 images in a second. On the other hand, the MobileNetV3-large model is coming after the VGG models with Xception, Resnet50, and EfficinetNetB7 models and they process one image in 0.4 seconds.

The MobileNetV3-large model was chosen as the base model because the EfficientNet models are closer to being over-fitted and the MobileNetV3large model got the best results in the test dataset. We did not change the structure given in Figure 10 because it has already obtained good results. Instead, we aimed to achieve higher performances by fine-tuning the model with newly added fully connected layers. In this process, we first froze pre-trained weights of the MobileNetV3-large model and only changed the weights of newly added fully connected layers by using transfer learning. After a short training, we unfreeze all layers of the MobileNetV3-large model and trained all layers using the parameters given in Table 3.

Table 3. Training details of the proposed model

	Transfer Learning	Fine-Tuning
Epochs	3	30
Optimizer	Adam [46]	Adam
Learning Rate	0.001	0.0001
Trainable Parameters	164,226	4,341,858
Total Parameters	164,226	4,390,658

We achieved very good results with these parameters only using static images on the NPDI dataset. After transfer learning, we decreased the learning rate parameter to prevent pre-trained model weights from changing too much, as stated by Li et al. [47]. Because shallower layers of the pre-trained model are already very good at detecting low-level features like edges, changing these weights too much would destroy the pre-learned feature extraction methods in the model.

The proposed OCDet model achieved an average accuracy of 99.8% and 99.4%, respectively on training and test data of the NPDI dataset. While preparing the data for training, we first selected the same number of random data as the training set as the test data among the hard and easy test data. Then we divided the data into two groups 80% training data and 20% validation data. We trained the model for 30 epochs as the model started to over-fit after this point. In addition, the trained model was evaluated using an internet dataset, which was previously introduced in Section 2. The internet dataset comprises 6000 images, with an equal distribution of 3000 obscene and 3000 normal images. Notably, all images in this dataset feature humans in various poses, while the obscene-labeled images predominantly depict individuals in indoor settings, simulating real-life scenarios. When we evaluated the trained model on the NPDI dataset with our newly created dataset, the model achieved an accuracy of 98% on this evaluation task. This result shows the generalizing power of the model. We also trained the model with the second dataset by dividing the dataset into 80% training and 20% test data. This time the model achieved 100% and 99.2% accuracy on training and test data, respectively. The results show that the proposed model is suitable for robustly detecting obscenity in images and videos.

Table 4. Comparison of the proposed method with the literature

Work	Dataset	Data	Method	Overall Accuracy
[13]	NPDI	Keyframes	Transfer Learning	94.0%
[11]	NPDI	Keyframes	BoVW	86.5%
[15]	NPDI	Videos	Transfer Learning+ Motion Features	97.9%
[16]	NPDI	Sequential frames	Transfer Learning+LSTM	95.6%
[21]	NPDI	KeyFrames	EfficientDet	75.0%
[19]	NPDI	Videos	Generative Adversarial Network	98.0%
[18]	NPDI	Sequential frames	Transfer Learning+CNN	98.3%
[14]	NPDI+extra data	Keyframes	CNN	99%
Ours	NPDI	Keyframes	Transfer Learning + CNN	**99.35%**

Upon comparison between the proposed model in this study and existing models in the literature, it is evident that our model achieves superior results. In a previous study [13], the authors employed AlexNet and GoogleNet as feature extractors, focusing solely on training the last layer added on top of these pre-trained models. In contrast, our work involves training two separate models for the same task, thereby increasing the complexity of the approach. In a previous study [11], a BovW method is used but this model is overperformed by current deep learning-based models. In a previous study [15], authors used a separate feature extraction method to enhance the performance of the model but this method requires more data and relies on manual feature extraction methods. Similarly in the studies [16, 18, 19], authors used sequential images as input to their model. these models also require sequential frames of videos while our model can be trained with only single frames. Also, our model outperforms these models thanks to its effective feature extraction ability. In a recent study [14], the authors incorporated additional data during the training of their model, thereby introducing challenges in comparing their results with the existing literature. In contrast, our model effectively addresses these issues by utilizing single frames for training and employing a single model dedicated to a specific task, facilitating fair and meaningful comparisons with the literature. Furthermore, our model exhibits a reduced number of trainable parameters in comparison to these models, resulting in lower time complexity.

We compared the results obtained with the benchmark dataset with the literature and summarized the comparison in Table 4. As can be seen, our model over-performs the other studies on the NPDI dataset. Besides, our model works with frames of the videos, so the proposed model is faster than the models that require a sequence of images. In addition, we verified the success of the model in the collected private dataset, ensuring that the model performance is high on different data.

We proposed the IMPD model to classify face images into two classes. For detecting faces, we used the pre-trained model proposed by Karakuş et al. [48] as the model has a high accuracy rate on a big dataset and can detect faces in different light, angle, and alignment conditions. The parameters of the proposed classification model are given in Table 5.

Table 5. Parameters of the IMPD model

Layer Name	Description	Output Shape	Trainable Parameters
Input	Model input layer to get images with desired input size	200×200×3	0
Conv2D_1	Num. of filters=32, kernel size=7, stride=1	200×200×32	4736
Conv2D_2	Num. of filters=64, kernel size=3, stride=1	200×200×64	18496
Conv2D_3	Num. of filters=32, kernel size=3, stride=1	200×200×32	18464
Add	Layer for adding outputs of Conv2D_1 and Conv2D_3	200×200×32	0
Conv2D_4	Num. of filters=128, kernel size=3, stride=3	66×66×128	36992
Conv2D_5	Num. of filters=256, kernel size=3, stride=2	32×32×256	295168
Flatten	Flattening layer	1×1×262144	0
Dense_1	Num. of units=512, activation=relu	512	134218240
Dense_2	Num. of units=2, activation=softmax	2	1026

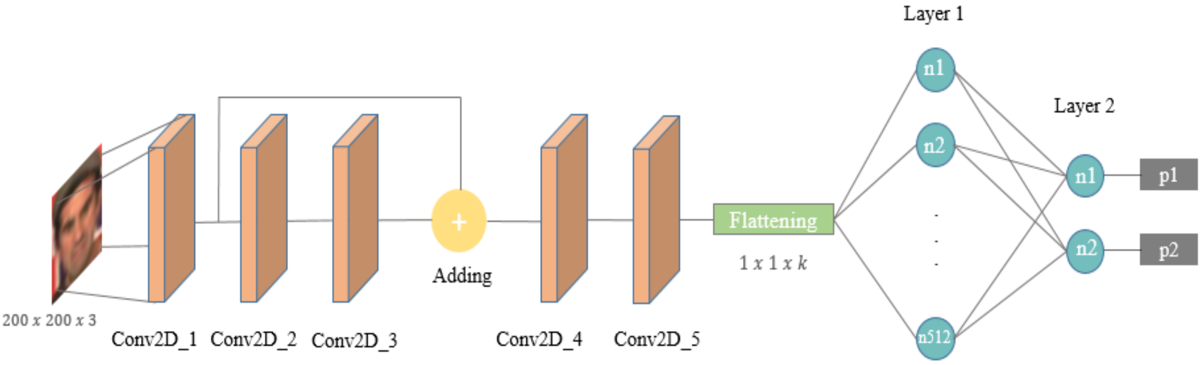


Figure 13. The architecture of the IMPD model

We trained the proposed IMPD model from scratch in an end-to-end fashion. In the model, we first increased the input images' number of channels in the first two layers while keeping activations map sizes unchanged using the same padding. Then we decreased the number of channels and added first and third convolution filter activations to each other to combine low-level and high-level features. Thus, the model becomes more resistant to overfitting, and representations are learned better [49]. We then decreased the dimensions of the activation maps by using convolution layers with valid padding while increasing the number of channels. The results show that the model is achieving good results on the UTKFace dataset. The proposed model achieved a maximum accuracy of 99.16% on the UTKFace dataset. In our experiments, we divided the dataset into training and validation sets whose ratios are 80% and 20%. The architecture of the proposed method is given in Figure 13.

Training results show that there is no sign of overfitting and the model puts a robust performance on the UTKFace dataset. The proposed model achieved an accuracy of 99.7% and 99.0% on training and test data, respectively. Besides, this model's inference time is 5 times lower than the OCDet model. The model can process an image in 0.1 seconds. Moreover, the model can be run simultaneously with the other model without delay.

The results show that both proposed models have achieved successful results on the relevant data. Considering the success of the models, it can be deduced that the proposed system works with a 98% success rate in the worst case in the decision combination stage.

In future studies, this model can be developed to have only one common CNN model for feature extraction from both face images and obscene or normal images. This way, the disadvantages such as running two models simultaneously, arising from the use of two different models can be avoided. Besides the OCDet model can be used for video stream content filtering or content moderation applications thanks to its high speed of image processing. Similarly, the IMPD model can be used in content filtering systems, online platforms, educational platforms, and legal proceedings to identify and restrict access to age-inappropriate or explicit content. Its application promotes safer online environments, ensures age-appropriate content delivery, and aids in evidence analysis for legal cases.

6. CONCLUSION

This study aimed to accelerate forensic evidence investigations through the proposal of obscenity and CSA detection model. The model consisted of two sub-models, namely the OCDet model and the IMPD model, both created using CNN architectures for obscenity detection and immature person detection tasks. The proposed models achieved an accuracy of 99.35% and 99.03% on related tasks, respectively. The OCDet model achieved state-of-the-art results on the NPDI dataset, while the IMPD model exhibited high performance in finding immature faces. Moreover, the proposed methods exhibited exceptional performance by achieving superior results with a reduced number of trainable parameters, effectively halving the parameter count compared to a model that demonstrated performance comparable to our own. This achievement highlights the efficiency and effectiveness of the proposed methods in optimizing model complexity while maintaining high-performance levels. The proposed models demonstrate significant performance improvements over existing studies, surpassing them by a notable margin of 1% to 15% in terms of the accuracy evaluation metric. These results highlight the superior efficacy and effectiveness of the proposed models compared to their counterparts in the literature. This remarkable outcome attests to the effectiveness of the proposed models in surpassing the performance of previously proposed studies and models. The proposed models demonstrated the potential to automate and improve the accuracy of evidence examination processes by automatically filtering.

The significance of the proposed models lies in their potential impact on forensic evidence investigations, where rapid and accurate evidence acquisition is crucial. By minimizing the reliance on human experts and reducing human-induced errors, the proposed models offer a valuable tool for law enforcement agencies combating the distribution of illegal content, especially in the case of CSA detection—a crime of global concern. Considering that commercial software often fails to meet the expectations of forensic experts, especially in terms of speed and accuracy, the proposed methods put forth rapid and robust solutions tailored to meet the specific requirements of experts in the field. By addressing the limitations of commercial alternatives, the proposed methods provide effective and reliable tools that align with the distinct needs of forensic experts.

However, further research is needed to address potential challenges and improve the proposed models. Future directions may include exploring more advanced CNN architectures to use only one model for both obscenity and immature person detection tasks and incorporating additional data sources to even enhance the generalization power of the models.

In summary, this study has contributed to the field of forensic evidence investigations by proposing robust models for obscenity and CSA detection. The potential impact of these models lies in their ability to automate the detection process, improve accuracy, and aid law enforcement agencies in combating illegal content distribution. With further research and advancements in the proposed models, limitations can be overcome and their practical utility can be enhanced.

REFERENCES

[1] Medzini, R. (2022). Enhanced self-regulation: The case of Facebook's content governance. New Media & Society, 24(10): 2227-2251. http://doi.org/10.1177/1461444821989352

[2] Hines, A. (2021). The rise in the distribution of child pornography on the web: 3 essential tools to help law enforcement (Doctoral dissertation, Utica College).

[3] Salian, D., Khatun, S. (2020). Legal framework on child pornography: A perspective. In Digital Forensic Science. IntechOpen. http://doi.org/10.5772/intechopen.92716

[4] Spalazzi, L., Paolanti, M., Frontoni, E. (2021). An offline parallel architecture for forensic multimedia classification. Multimedia Tools and Applications, 81: 22715-22730. http://doi.org/10.1007/s11042-021-10819-x

[5] Basilio, J.A.M., Torres, G.A., Perez, G.S., Medina, L.K. T., Meana, H.M.P. (2011). Explicit image detection using YCbCr space color model as skin detection. Applications of Mathematics and Computer Engineering, 123-128.

[6] Jones, M.J., Rehg, J.M. (2002). Statistical color models with application to skin detection. International Journal of Computer Vision, 46: 81-96. https://doi.org/10.1023/A:1013200319198

[7] Lumini, A., Nanni, L. (2020). Fair comparison of skin detection approaches on publicly available datasets. Expert Systems with Applications, 160: 113677. https://doi.org/10.1016/j.eswa.2020.113677

[8] Deselaers, T., Pimenidis, L., Ney, H. (2008). Bag-of-visual-words models for adult image classification and filtering. In 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, pp. 1-4. http://doi.org/10.1109/ICPR.2008.4761366

[9] Geng, Z., Zhuo, L., Zhang, J., Li, X. (2015). A comparative study of local feature extraction algorithms for web pornographic image recognition. In 2015 IEEE International Conference on Progress in Informatics and Computing (PIC), Nanjing, China, pp. 87-92. http://doi.org/10.1109/PIC.2015.7489815

[10] Lopes, A.P.B., de Avila, S.E., Peixoto, A.N., Oliveira, R.S., Coelho, M.D.M., Araújo, A.D.A. (2009). Nude detection in video using bag-of-visual-features. In 2009 XXII Brazilian Symposium on Computer Graphics and Image Processing, Rio de Janeiro, Brazil, pp. 224-231. http://doi.org/10.1109/SIBGRAPI.2009.32

[11] Avila, S., Thome, N., Cord, M., Valle, E., AraúJo, A. D. A. (2013). Pooling in image representation: The visual codeword point of view. Computer Vision and Image Understanding, 117(5): 453-465. https://doi.org/10.1016/j.cviu.2012.09.007

[12] Moreira, D., Avila, S., Perez, M., Moraes, D., Testoni, V., Valle, E., Rocha, A. (2016). Pornography classification: The hidden clues in video space–time. Forensic Science International, 268: 46-61. https://doi.org/10.1016/j.forsciint.2016.09.010

[13] Moustafa, M. (2015). Applying deep learning to classify pornographic images and videos. arXiv preprint arXiv:1511.08899. https://doi.org/10.48550/arXiv.1511.08899

[14] Karamizadeh, S., Shojae Chaeikar, S., Jolfaei, A. (2023). Adult content image recognition by Boltzmann machine limited and deep learning. Evolutionary Intelligence, 16(4): 1185-1194. https://doi.org/10.1007/s12065-022-00729-8

[15] Perez, M., Avila, S., Moreira, D., Moraes, D., Testoni, V., Valle, E., Rocha, A. (2017). Video pornography detection through deep learning techniques and motion information. Neurocomputing, 230: 279-293. https://doi.org/10.1016/j.neucom.2016.12.017

[16] Wehrmann, J., Simões, G.S., Barros, R.C., Cavalcante, V.F. (2018). Adult content detection in videos with convolutional and recurrent neural networks. Neurocomputing, 272: 432-438. https://doi.org/10.1016/j.neucom.2017.07.012

[17] Yousaf, K., Nawaz, T. (2022). A deep learning-based approach for inappropriate content detection and classification of YouTube videos. IEEE Access, 10: 16283-16298. http://doi.org/10.1109/ACCESS.2022.3147519

[18] Gautam, N., Vishwakarma, D.K. (2022). Obscenity detection in videos through a sequential convnet pipeline classifier. IEEE Transactions on Cognitive and Developmental Systems, 15(1): 310-318. http://doi.org/10.1109/TCDS.2022.3158613

[19] Yang, Y., Xu, S. (2022). Tackling explicit material from online video conferencing software for education using deep attention neural architectures. Computational Intelligence and Neuroscience, Article ID 6334802. https://doi.org/10.1155/2022/6334802

[20] Mallmann, J., Santin, A.O., Viegas, E.K., dos Santos, R.R., Geremias, J. (2020). PPCensor: Architecture for real-time pornography detection in video streaming. Future Generation Computer Systems, 112: 945-955. http://doi.org/10.1016/j.future.2020.06.017

[21] Hor, S.L., Karim, H.A., Abdullah, M.H.L., AlDahoul, N., Mansor, S., Fauzi, M.F.A., Wazir, A.S.B. (2021). An evaluation of state-of-the-art object detectors for pornography detection. In 2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Kuala Terengganu, Malaysia, pp. 191-196. http://doi.org/10.1109/ICSIPA52582.2021.9576796

[22] Rhodes, M.G. (2009). Age estimation of faces: A review. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 23(1): 1-12. https://doi.org/10.1002/acp.1442

[23] Hajizadeh, M.A., Ebrahimnezhad, H. (2011). Classification of age groups from facial image using histograms of oriented gradients. In 2011 7th Iranian Conference on Machine Vision and Image Processing, Tehran, Iran, pp. 1-5. http://doi.org/10.1109/IranianMVIP.2011.6121582

[24] Eidinger, E., Enbar, R., Hassner, T. (2014). Age and gender estimation of unfiltered faces. IEEE Transactions on Information Forensics and Security, 9(12): 2170-2179. https://doi.org/10.1109/Tifs.2014.2359646

[25] Sai, P.K., Wang, J.G., Teoh, E.K. (2015). Facial age range estimation with extreme learning machines. Neurocomputing, 149: 364-372. https://doi.org/10.1016/j.neucom.2014.03.074

[26] Wang, X., Guo, R., Kambhamettu, C. (2015). Deeply-learned feature for age estimation. In 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, pp. 534-541. http://doi.org/10.1109/WACV.2015.77

[27] Chen, S., Zhang, C., Dong, M., Le, J., Rao, M. (2017). Using ranking-CNN for age estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 5183-5192. http://doi.org/10.1109/CVPR.2017.86

[28] Anda, F., Lillis, D., Kanta, A., Becker, B.A., Bou-Harb, E., Le-Khac, N.A., Scanlon, M. (2019). Improving the accuracy of automated facial age estimation to aid CSEM investigations. Digital Investigation, 28: S142. http://doi.org/10.1016/j.diin.2019.01.024

[29] Castrillón-Santana, M., Lorenzo-Navarro, J., Travieso-González, C.M., Freire-Obregón, D., Alonso-Hernandez, J.B. (2018). Evaluation of local descriptors and CNNs for non-adult detection in visual content. Pattern Recognition Letters, 113: 10-18. http://doi.org/10.1016/j.patrec.2017.03.016

[30] Sae-Bae, N., Sun, X., Sencar, H.T., Memon, N.D. (2014). Towards automatic detection of child pornography. In 2014 IEEE International Conference on Image Processing (ICIP), pp. 5332-5336. http://doi.org/10.1109/ICIP.2014.7026079

[31] Yiallourou, E., Demetriou, R., Lanitis, A. (2017). On the detection of images containing child-pornographic material. In 2017 24th International Conference on Telecommunications (ICT), Limassol, Cyprus, pp. 1-5. http:/doi.org/10.1109/ICT.2017.7998260

[32] Macedo, J., Costa, F., dos Santos, J.A. (2018). A benchmark methodology for child pornography detection. In 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Parana, Brazil, pp. 455-462. http://doi.org/10.1109/SIBGRAPI.2018.00065

[33] Gangwar, A., González-Castro, V., Alegre, E., Fidalgo, E. (2021). AttM-CNN: Attention and metric learning based CNN for pornography, age and Child Sexual Abuse (CSA) detection in images. Neurocomputing, 445: 81-104. https://doi.org/10.1016/j.neucom.2021.02.056

[34] Raghu, M., Zhang, C., Kleinberg, J., Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. Advances in Neural Information Processing Systems, 32. https://doi.org/10.48550/arXiv.1902.07208

[35] Montasari, R., Hill, R. (2019). Next-generation digital forensics: Challenges and future paradigms. In 2019 IEEE 12th International Conference on Global Security, Safety and Sustainability (ICGS3), London, UK, pp. 205-212. http://doi.org/10.1109/ICGS3.2019.8688020

[36] Paul Joseph, D., Norman, J. (2019). An analysis of digital forensics in cyber security. In First International Conference on Artificial Intelligence and Cognitive Computing: AICC 2018, pp. 701-708. http://doi.org/10.1007/978-981-13-1580-0_67

[37] Sikos, L.F. (2021). AI in digital forensics: Ontology engineering for cybercrime investigations. Wiley Interdisciplinary Reviews: Forensic Science, 3(3): e1394. https://doi.org/10.1002/wfs2.1394

[38] Barik, K., Abirami, A., Konar, K., Das, S. (2022). Research perspective on digital forensic tools and investigation process. Illumination of Artificial Intelligence in Cybersecurity and Forensics, 71-95. https://doi.org/10.1007/978-3-030-93453-8_4

[39] Forensics, O.E. (2022). Encase Forensic Features. Available from: https://www.opentext.com/products/encase-forensic#features.

[40] Forensics, O. (2022). Oxygen Forensics Detective. Available from: https://www.oxygen-forensic.com/en/products/oxygen-forensic-detective.

[41] Yosifon, O. (2022). How AI and Machine Learning Are Impacting Digital Investigations. Available from: https://cellebrite.com/en/how-ai-and-machine-learning-is-impacting-digital-investigations/.

[42] Sanchez, L., Grajeda, C., Baggili, I., Hall, C. (2019). A practitioner survey exploring the value of forensic tools,

AI, filtering, & safer presentation for investigating child sexual abuse material (CSAM). Digital Investigation, 29: S124-S142. https://doi.org/10.1016/j.diin.2019.04.005

[43] Zhang, Z., Song, Y., Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 5810-5818. http://doi.org/10.1109/CVPR.2017.463

[44] Bazarov, E. (2023). NSFW data source URLs. Available from: https://github.com/EBazarov/nsfw_data_source_urls.

[45] Majid, S., Alenezi, F., Masood, S., Ahmad, M., Gündüz, E.S., Polat, K. (2022). Attention based CNN model for fire detection and localization in real-world images. Expert Systems with Applications, 189: 116114. http://doi.org/10.1016/j.eswa.2021.116114

[46] Kingma, D.P., Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. https://doi.org/10.48550/arXiv.1412.6980

[47] Li, H., Chaudhari, P., Yang, H., Lam, M., Ravichandran, A., Bhotika, R., Soatto, S. (2020). Rethinking the hyperparameters for fine-tuning. arXiv preprint arXiv:2002.11770. https://doi.org/10.48550/arXiv.2002.11770

[48] Karakuş, S., Kaya, M., Tuncer, S.A., Bahşi, M.T., Açikoğlu, M. (2022). A deep learning based fast face detection and recognition algorithm for forensic analysis. In 2022 10th International Symposium on Digital Forensics and Security (ISDFS), Istanbul, Turkey, pp. 1-6. http://doi.org/10.1109/ISDFS55398.2022.9800785

[49] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778. https://doi.org/10.48550/arXiv.1512.03385

NOMENCLATURE

X	Input image for the model
h	Height of the image file
w	Width of the image file
m	Number of faces in an image
n	Number of images in dataset
F	Cropped face image
R	Resizing function
s1	Height of the convolution kernel
s2	Width of the convolution kernel
z	Output of fully connected layer
a	Activation function of fully connected layer
I	Input of fully connected layer
W	Weights of fully connected layer
B	Biases of fully connected layer
sfmx	Softmax activation function
o	Probability of a class in classification
e	Exponentiation constant
p	Final decision of the classification model
k	Number of classes in classification task
Max	Maximum function
p1	Final decision of the first classification model
p2	Final decision of the second classification model
csa	Decision fusion function

Superscripts

i	Index of specific variables
l	Layer index of convolution layers