



## Deep Active Learning for Pornography Recognition Using ResNet

Sui Lyn Hor<sup>1\*</sup>, Nouar Aldahoul<sup>1\*</sup>, Hezerul Abdul Karim<sup>1\*</sup>, Mohd Haris Lye<sup>1</sup>, Sarina Mansor<sup>1</sup>,  
Mohammad Faizal Ahmad Fauzi<sup>1</sup>, Abdulaziz Saleh Ba Wazir<sup>1</sup>

<sup>1</sup>Faculty of Engineering, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia

**Abstract.** The demand for nudity and pornographic content detection is increasing due to the prevalence of media products containing sexually explicit content with Internet being the main source. Recent literature has proved the effectiveness of deep learning techniques for adult image and video detection. However, the requirement for a huge dataset with labeled examples poses a restriction in practical use. Several research has shown that training deep models using an active learning framework could reduce the annotation effort, but this approach has yet to be applied for pornography detection. In this paper, the classification efficiency and annotation requirement of fine-tuned ResNet50V2 model trained using an active learning framework in pornographic image recognition was explored by comparing the method's performance using three sampling strategies (random sampling, least confidence sampling, and entropy sampling). The baseline for comparison was a fully supervised learning method. The video frames of the public NPDI dataset were utilized to run a 5-fold cross-validation. The results of the experiments demonstrated that similar average test accuracy of five folds could be obtained using the deep active learning method, with only 60% of labeled samples in the training dataset compared to 100% annotated samples in fully supervised learning.

**Keywords:** Convolutional neural network; Deep active learning; Nudity detection; Pornography recognition

### 1. Introduction

Advancement in technology has facilitated the accessibility of inappropriate content, particularly nudity and pornographic visual content. As a result, internet users could be exposed to such content either intentionally or unintentionally. Studies have shown that early exposure to adult content could promote negative mental health and increase the intention to engage in sexual activities (Zhang & Jemmott, 2014; Camilleri et al., 2021). In addition, unrealistic sexual beliefs may be developed in young pornography consumers (Owens et al., 2012). This poses a concern to society, especially to parents, which calls for a need for control measures to limit the exposure to pornographic images and videos as humans are visual beings (Hyerle, 2000).

Content censorship is one of the preventive measures applied to visually inappropriate content. Therefore, a prerequisite to content filtering is the detection of targeted content. Past work of the co-authors explored pornography recognition using audio features

---

\*Corresponding author's email: [suilhor@gmail.com](mailto:suilhor@gmail.com), [nouar.aldahoul@live.iium.edu.my](mailto:nouar.aldahoul@live.iium.edu.my), [hezerul@mmu.edu.my](mailto:hezerul@mmu.edu.my),  
Tel: +603-83125499, Fax: +603-8318 3029  
doi: [10.14716/ijtech.v13i6.5842](https://doi.org/10.14716/ijtech.v13i6.5842)

(Banaeeyan et al., 2019). In this work, the focus is placed on the detection of pornographic related graphics, where pornography refers to “any sexually explicit material with the aim of sexual arousal or fantasy” (Short et al., 2012).

Automation of pornography recognition using deep learning techniques has greatly improved the efficiency of censorship by reducing the workforce of censorship editors. In 2015, a pornographic video classifier was designed by taking classifications of fine-tuned convolutional neural networks (CNN) such as AlexNet and GoogLeNet on video keyframes into account (Moustafa, 2015). Usage of a support vector machine (SVM) to classify features detected by CNN models was also applied to recognize obscene video frames (Lyn et al., 2020). Furthermore, (Aldahoul et al., 2021) tested the performance of pornography detection in cartoon videos by combining decisions of several fusion approaches that utilize CNN as a feature extractor and SVM as classifier. A multi-level pornographic image classifier was implemented by using ResNet-50 and Mask R-CNN models such that images with low classification probabilities would be sent to the consequent stage while the rest would carry the first stage decisions (Nguyen et al., 2020). Usage of the You Only Look Once v3 (YOLOv3) object detector to focus on image patches containing humans has improved the classification performance of CNN on images with small-scale pornographic content (Aldahoul et al., 2021).

Currently, the performance of pornography detection in images using fully supervised learning or fully trained networks has reached a bottleneck such that improvement is possible only by sacrificing computational resources, time, and cost in the model training process. Inspired by works of other fields that focused on minimizing cost (Imanullah et al., 2019; Lischer et al., 2021), a different approach came into light. In years, training of deep models using an active learning framework, or deep active learning, has been proposed as one of the possible solutions to this obstacle. This method could reduce the annotation cost and effort by not having all the collected data labeled for training. In other words, not all data collected for the purpose of model training would need to be annotated manually by human experts, which is significant when a huge dataset is of concern, and when compared against usage of conventional CNN, which would require 100% of the data collected to be manually annotated. However, it should be noted that the presence of a manual annotator in the preparation or training process is still compulsory (in this stage) to guide the model training performance. This proposed method has been tested in several application areas, such as medical field (Stanitsas et al., 2017; Shao et al., 2018), human face recognition, and object classification (Ranganathan et al., 2017; Wang et al., 2017). Specifically, comparable accuracy was achieved with faster CNN training in cancerous tissue detection using an active learning framework (Stanitsas et al., 2017) while a pairwise-constraint deep CNN designed to categorize nucleus in pathology images managed to save 40% of expert annotation time compared to usage of conventional CNN (Shao et al., 2018). The inclusion of active learning-related criteria in loss function accelerated deep belief model training and allowed a specific accuracy to be reached with less labeling effort (Ranganathan et al., 2017). In addition, using pseudo-labeling technique as a substitute for human annotator for high-confidence samples succeeded in achieving high accuracy at around 60% labeled samples (Wang et al., 2017). Thus far, results from the experiments indicated that active learning could reduce annotation requirement while allowing comparable performances. Therefore, this would be beneficial when classification standards require modification, which would lead to modification in data labels and model retraining.

As pornographic image recognition is a challenging task that involves various depiction

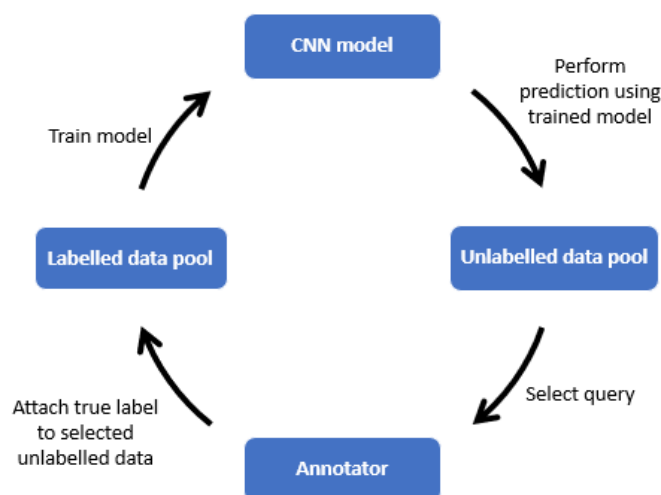
of subject (in terms of object of matter, angle, image type, posture etc.) in which the censoring regulation may change, an update in this regulation would pose a large annotation burden due to the large amount of labeled samples required for model training. This could be mitigated if deep active learning can achieve similar effects of reducing annotation cost while attaining comparable classification accuracy as the method without active learning in this application. However, to the best of the authors' knowledge, deep active learning has not yet been applied to pornography detection. This paper aims to study the performances of deep active learning in pornographic image recognition by training a fine-tuned CNN model, named ResNet50V2 (He et al., 2016), using active learning to perform image classification on video frames of the public NPDI dataset, or Pornography-800 dataset (Avila et al., 2013). Different numbers of samples in the training dataset were annotated compared to the full annotation in the fully supervised learning method. Three common query strategies (random sampling, least confidence sampling and entropy sampling) were employed in this study.

The organizations of the subsequent sections are as follows: Section 2 describes the details of the experiment and performance metrics, Section 3 presents the experimental findings and analysis of the findings, and Section 4 summarises this work and includes possible future works.

## 2. Methods

### 2.1. Overview

Deep active learning was applied by training ResNet50V2 model (He et al., 2016), with transfer learning enabled, using an active learning framework. The overall system architecture of the proposed method is illustrated in Figure 1.



**Figure 1** Deep active learning framework

### 2.2. Dataset

The public NPDI dataset, or Pornography-800 dataset (Avila et al., 2013), was utilized in this work, similar to most other research works involving pornography detection. This dataset contained 16727 keyframes extracted from a total of 800 videos of pornographic and non-pornographic classes that spans almost 80 hours.

The image or frame label assigned depends on the content of the images and not on one of the videos, i.e., the frames of pornographic videos could be non-pornographic if they do not contain any sexually explicit content and vice versa. The distribution of data used in this

work is recorded in Table 1. Despite the imbalanced data distribution, the train-test split used was the one specified by the creators of the NPDI dataset for standardization purposes.

**Table 1** Number of images per class

Image Class	Number of Images
Pornographic	12170
Non-pornographic	4557
Total	16727

### 2.3. Deep Active Learning

Generally, active learning involves an iterative training process in which a subset of unlabeled samples is selected from the unlabeled data pool to be annotated or attached to their true class labels based on the predefined query criterion or criteria. The labeled data pool expands as the model training proceeds, appending newly annotated data to the labeled data pool in the current training cycle. The training ends when the stopping criterion is fulfilled.

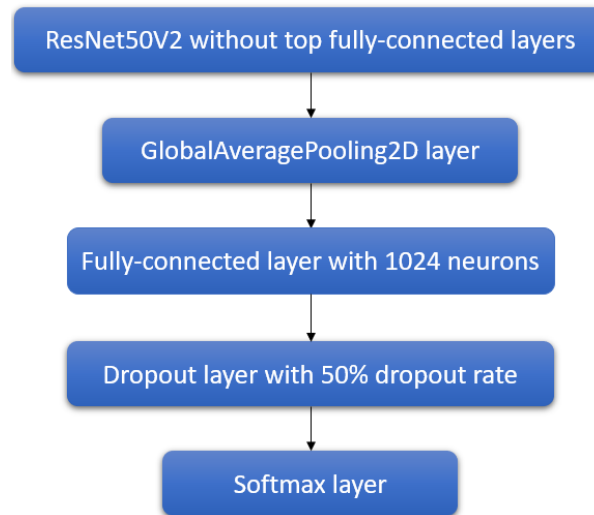
Initially, all data were unlabeled. The initial or first batch of labeled data pool was formed by randomly selecting 20% of unlabeled data from each class, ensuring a balanced initial training dataset. The stopping criterion utilized in the experiment was simply the exhaustion of unlabeled data, i.e., model training ends when all unlabeled data were labeled and used for training. Three sampling strategies were applied in the active learning training framework in order to select the unlabeled data to be used as training sample in the subsequent model training stages by attaching their respective true class labels: (i) random sampling (Settles, 2009), which selected unlabeled samples at random without following any specific pattern (ii) least confidence sampling (Settles, 2009), which picked the data with lowest prediction probability, i.e., the ones the trained model was least confident of their predicted labels, as shown in Equation 1, and (iii) entropy sampling (Shannon, 2001), another uncertainty-based selection strategy that queried samples based on the uncertainty of the trained model by taking all possible label probabilities into account, as shown in Equation 2. A simulated annotation process was performed, i.e., no real-time annotation by humans, such that all true data labels were present but not attached to the corresponding data unless the data were queried.

$$x_{LC} = \operatorname{argmax}_x (1 - P_{\theta}(\hat{y}|x)) \quad (1)$$

$$x_H = \operatorname{argmax}_x (-\sum_i P_{\theta}(y_i|x) \log P_{\theta}(y_i|x)) \quad (2)$$

where  $\hat{y}$  is the class label with the highest posterior probability under the model  $\theta$ , and  $y_i$  ranges over all possible class labels.

Deep active learning involves training of deep learning network using active learning loops. In this case, the deep learning model employed was a CNN model named ResNet50V2 (He et al., 2016). Instead of the fully connected layers at the top of the default network, several different layers were added to the model's top. The deployed model architecture is shown in Figure 2. In addition, the transfer learning technique was applied such that model training was enabled for the higher layers, layers from conv5\_block3\_1 onwards in specifics, of the CNN model pre-trained on the ImageNet classification task (Krizhevsky et al., 2017). This would aid in domain adaptation of the model from source task to the target task.



**Figure 2** Proposed model architecture

Dataset augmentation was applied during model training to expand the training dataset and increase data diversity. The augmentation techniques applied randomly are horizontal flipping, vertical flipping, horizontal shifting, vertical shifting, shearing, rotation, and zooming.

5-fold cross-validation was performed to reduce model biases. The train and test datasets in each fold were defined based on the standard NPDI dataset folds proposed by the dataset authors, which were divided according to videos and not video frames, i.e., all frames of the same video belong to either the training dataset or test dataset. For each fold, the training dataset was divided into two data groups for training and validation purposes, respectively. Specifically, selected randomly, 20% of the training data was used as a validation dataset and the rest for training. Five experiments were run for each fold using different seed values to generalize the model's performance and remove the influence of randomness. The model training hyperparameters are Adam optimizer with a learning rate of 0.001, batch size of 32, the maximum number of epochs per active learning cycle of 5, number of queries per iteration of 500, and early stopping based on validation loss.

As the effect of usage of active learning for model training on reduction of annotation cost was to be studied, testing was performed several times in each experiment conducted to obtain accuracies at different percentages of labeled data out of the whole training dataset.

#### 2.4. Performance Metric

The metric used to evaluate the performance of the proposed method was image classification accuracy, which was calculated using Equation 3.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

where TP denotes the true positive or the number of correctly classified pornographic images; TN refers to the true negative or the number of correctly classified non-pornographic images; FP denotes the false positive or the number of wrongly classified non-pornographic images, and FN refers to the false negative or the number of wrongly classified pornographic images.

#### 2.5. Experimental Setup

The experiment was run on a desktop computer with the specifications listed below. The algorithms were implemented using Python programming language.

- Alienware Aurora R8 device
- Intel(R) Core (TM) i7-8700 Central Processing Unit (CPU) @ 3.20GHz
- Windows 10 Home Single Language
- NVIDIA GeForce GTX 1080 Ti Graphical Processing Unit (GPU)
- 64GB RAM (11.0GB dedicated GPU memory and 31.9GB shared GPU memory)

### 3. Results and Discussion

#### 3.1. Experimental Results

The average test accuracies across five experiments for all five folds obtained at different percentages of labeled data are plotted in Figures 3 to 7, respectively. The total numbers of test data in each fold are indicated in the title of the corresponding graphs. These graphs included the findings obtained using four methods: deep active learning with random sampling (solid red line), deep active learning with least confidence sampling (solid green line), deep active learning with entropy sampling (solid blue line), and fully supervised learning (black dotted line). Due to the different numbers of images in test data per fold, overall average test accuracies of the five folds at several key percentages of labeled data are recorded in Table 2.

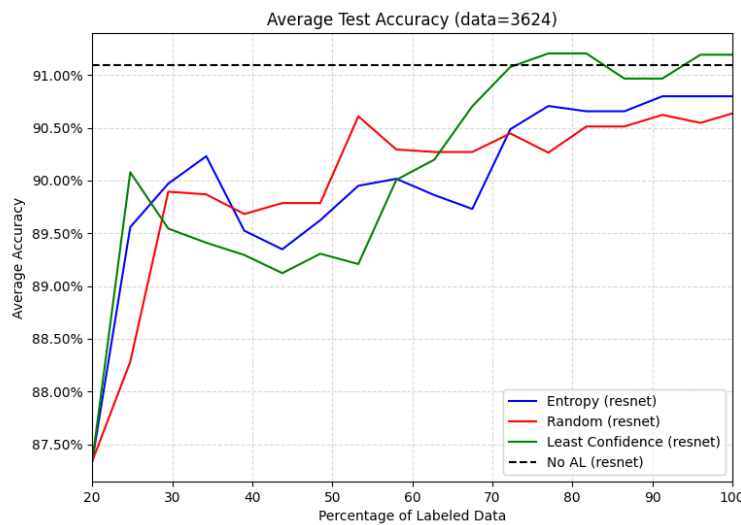


Figure 3 Average test accuracies for fold 0

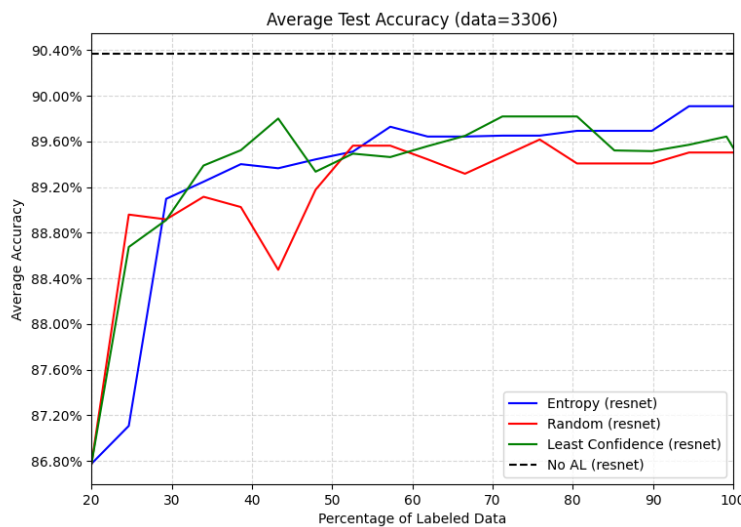


Figure 4 Average test accuracies for fold 1

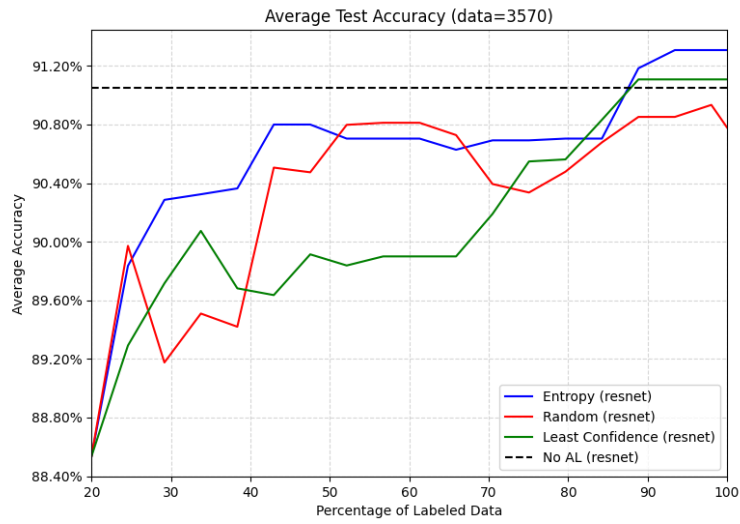


Figure 5 Average test accuracies for fold 2

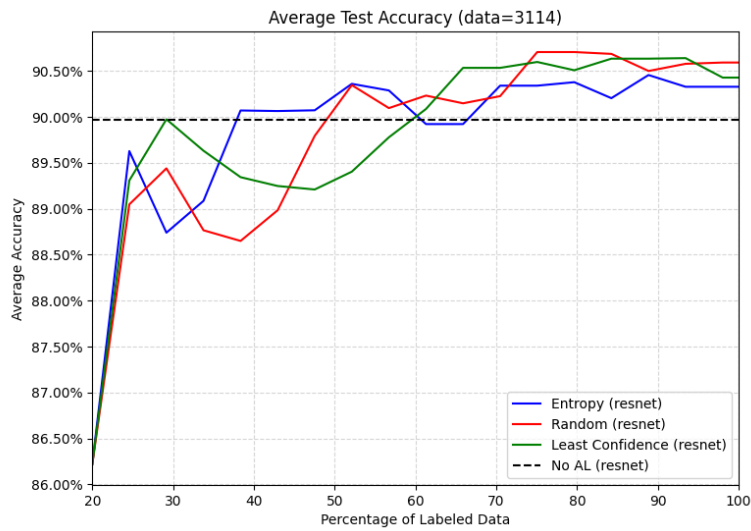


Figure 6 Average test accuracies for fold 3

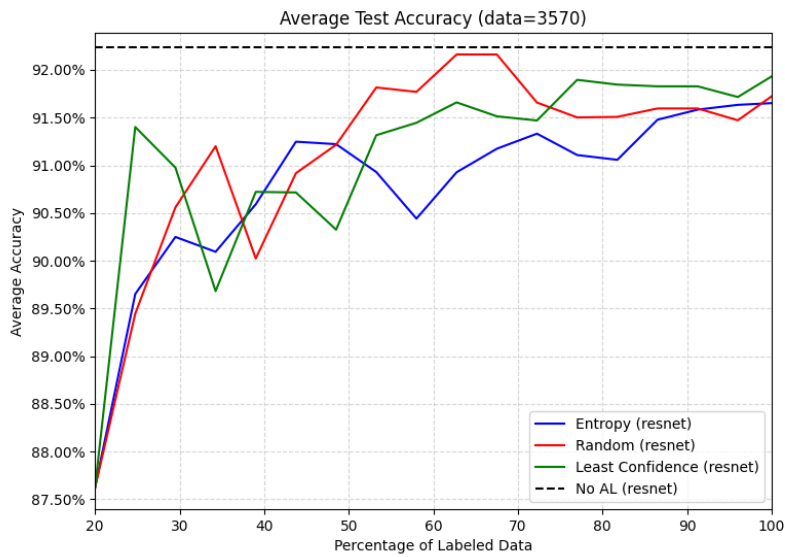


Figure 7 Average test accuracies for fold 4

**Table 2** Overall average test accuracies of five folds for all methods

Percentage of Labeled Data	Average Test Accuracy of Five Folds			
	Deep Active Learning			Fully Supervised Learning (Baseline)
	Random Sampling (Proposed)	Least Confidence Sampling (Proposed)	Entropy Sampling (Proposed)	
20	87.30	87.30	87.30	
40	89.36	89.71	89.99	
60	90.51	90.20	90.14	90.94
80	90.52	90.79	90.50	
100	90.65	90.84	90.80	

### 3.2. Data Analysis and Discussion

From the data in Table 2, the performances of the three querying strategies for the deep active learning method are similar at different percentages of labeled data. The highest average test accuracy of five training folds was obtained using the fully supervised learning method, satisfying the expectation that it serves the role of the highest achiever in terms of performance. However, similar results (difference of less than 1% accuracy) were achieved by the three deep active learning methods utilizing random sampling, least confidence sampling, and entropy sampling strategies with around 60% of training samples annotated. In other words, there is no need to manually label 40% of data to train the CNN model, allowing data labeling costs to be saved. This would be significant when large datasets are involved in practical scenarios.

Interestingly, the top scores that the fully trained CNN model should achieve in folds 0, 2, and 3 (refer to Figures 3, 5, and 6) were defeated by some or all of the deep active learning methods. This could be attributed to the fact that data were sent in batches for training, such that experiments applying the latter methods sent better training batches than the former method, generating better test accuracies.

Looking at deep active learning querying strategies, contrary to the expectation to have the worst performance for the random sampling method (Ranganathan et al., 2017), the results for all three querying strategies tested in this experiment were comparable. This could be proven by the “intertwining” pattern of the average test accuracy plots in Figures 3 to 7. The only obvious exception to this point was the drop and pick up in the performance of the random sampling method between 40% to 70% labeled data in fold 2 (refer to Figure 5). One possible explanation for this situation was that data selected to be annotated and included in the training dataset during this period caused misclassifications, perhaps due to data overfitting, and the performance was only recovered when other distinctive samples were queried. However, this performance drop was neutralized after averaging the results of all training folds. Furthermore, in this case, similar performances were achieved for least confidence sampling and entropy sampling methods; the former querying method would be the better choice to be employed. The reason is that the latter would require more computational resources due to the higher complexity of calculation for sample selection.

Despite the uptrend in performance with the increase in annotated data, the difference was only in a small degree (approximately 3.5% in overall average test accuracy by comparing 20% and 100% labeled data) with the sharpest performance increase observed at the initial stages of training. Note that the CNN model used was simply one with acceptable performance and training resources, and this study aimed to investigate the effect of deep active learning method in reducing annotation cost for pornographic image classification. Therefore, this small performance improvement may be attributed to the superior pre-trained model performance, saturated performance of the selected CNN



model or usage of less effective methods applied for deep active learning. This leads to possible future works exploring different sampling strategies in active learning framework and utilizing different CNN or deep networks.

As the dataset utilized was built using video keyframes, many of the data consisted of repeating characters, objects, and backgrounds. This limited data diversity affected the experimental results regarding model biases, as seen from the differing graph curves in Figures 3 to 7. An effort to reduce such influence was made by performing 5-fold cross-validation taking the average of the folds. Furthermore, it should be highlighted that the poor image quality of the dataset also deteriorated the model performance.

#### 4. Conclusions and Future Work

In summary, the active learning framework applied three query strategies (random sampling, least confidence sampling and entropy sampling). ResNet50V2 CNN was trained for pornography classification to study the effectiveness of deep active learning in annotation cost reduction. Fully supervised learning was performed as a baseline method to measure the model performance against annotation requirements. The popular NPDI dataset consisting of pornographic and non-pornographic video frames was utilized. A 5-fold cross-validation technique was used to minimize the model biases. Experimental results showed that less than 1% difference in average test accuracies of five folds was achieved using the deep active learning methods at 60% of the whole training dataset labeled compared to the fully trained CNN model. This confirmed the effect of deep active learning in reducing annotation costs for the application of pornographic image recognition. For future works, the application of a deep active learning network on different pornography datasets could be studied. Additionally, the utilization of other active learning sampling strategies and other deep learning networks for this application area could be further explored.

#### Acknowledgements

The authors would like to thank the project funders: TM R&D, Malaysia in 2018 under project number MMUE/180029, and IR Fund in 2022, under project number MMUI/220007.

#### References

- Aldahoul, N., Abdul Karim, H., Lye Abdullah, M.H., Ahmad Fauzi, M.F., Ba Wazir, A.S., Mansor, S., See, J., 2021. Transfer Detection of YOLO to Focus CNN's Attention on Nude Regions for Adult Content Detection. *Symmetry*, Volume 13(1), pp. 26
- Aldahoul, N., Karim, H.A., Abdullah, M.H.L., Wazir, A.S.B., Fauzi, M.F.A., Tan, M.J.T., Mansor, S., Lyn, H.S., 2021. An Evaluation of Traditional and CNN-Based Feature Descriptors for Cartoon Pornography Detection. *IEEE Access*, Volume 9, pp. 39910–39925
- 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*, Volume 117(5), pp. 453–465
- Banaeeyan, R., Karim, H.A., Lye, H., Fauzi, M.F.A., Mansor, S., See, J., 2019. Acoustic Pornography Recognition using Fused Pitch and Mel-Frequency Cepstrum Coefficients. *International Journal of Technology*, Volume 10(7), pp. 1335–1343
- Camilleri, C., Perry, J., Sammut, S., 2021. Compulsive Internet Pornography Use and Mental Health: A Cross-Sectional Study in a Sample of University Students in the United States. *Frontiers in Psychology*, Volume 11, pp. 613244

- He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity Mappings in Deep Residual Networks. *In: European Conference on Computer Vision*, pp. 630–645. Springer, Cham
- Hyerle, D., 2000. *A Field Guide to Using Visual Tools*. Association for Supervision and Curriculum Development, Alexandria, Virginia, USA
- Imanullah, M., Yuniarno, E.M., Sooi, A.G., 2019. A Novel Approach in Low-cost Motion Capture System using Color Descriptor and Stereo Webcam. *International Journal of Technology*, Volume 10(5), pp. 942–952
- Krizhevsky, A., Sutskever, I., Hinton, G., 2017. ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, Volume 60(6), pp. 84–90
- Lischer, K., Avila, F., Sahlan, M., Whulanza, Y., 2021. Assessment of Cost-Efficient Thermocycler Prototype for Polymerase Chain Reaction and Loop-Mediated Isothermal Amplification. *International Journal of Technology*, Volume 12(6), pp. 1207–1216
- Lyn, H.S., Mansor, S., Aldahoul, N., Abdul Karim, H., 2020. Convolutional Neural Network-based Transfer Learning and Classification of Visual Contents for Film Censorship. *Journal of Engineering Technology and Applied Physics*, Volume 2(2), pp. 28–35
- Moustafa, M., 2015. Applying Deep Learning to Classify Pornographic Images and Videos. *In: 7<sup>th</sup> Pacific - Rim Symposium on Image and Video Technology (PSIVT 2015)*, Auckland, 26 November, New Zealand
- Nguyen, Q., Tran, H.L., Nguyen, T.T., Phan, D.D., Vu, D.L., 2020. Multi-Level Detector for Pornographic Content Using CNN Models. *In: 2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pp. 1–5
- Owens, E., Behun, R., Manning, J., Reid, R., 2012. The Impact of Internet Pornography on Adolescents: A Review of the Research. *Sexual Addiction & Compulsivity*, Volume 19(12), pp. 99–122
- Ranganathan, H., Venkateswara, H., Chakraborty, S., Panchanathan, S., 2017. Deep Active Learning for Image Classification. *In: 2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3934–3938
- Settles, B., 2009. *Active Learning Literature Survey*. University of Wisconsin-Madison Department of Computer Sciences, Madison, Wisconsin, USA
- Shannon, C.E., 2001. A Mathematical Theory of Communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, Volume 5(1), pp. 3–55
- Shao, W., Sun, L., Zhang, D., 2018. Deep Active Learning for Nucleus Classification in Pathology Images. *In: 2018 IEEE 15<sup>th</sup> International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 199–202
- Short, M., Black, L., Smith, A., Wetterneck, C., Wells, D., 2012. A Review of Internet Pornography Use Research: Methodology and Content from the Past 10 Years. *Cyberpsychology, Behavior, and Social Networking*, Volume 15(1), pp. 13–23
- Stanitsas, P., Cherian, A., Truskinovsky, A., Morellas, V., Papanikolopoulos, N., 2017. Active Convolutional Neural Networks for Cancerous Tissue Recognition. *In: 2017 IEEE International Conference on Image Processing (ICIP)*, pp. 1367–1371
- Wang, K., Zhang, D., Li, Y., Zhang, R., Lin, L., 2017. Cost-Effective Active Learning for Deep Image Classification. *IEEE Transactions on Circuits and Systems for Video Technology*, Volume 27(12), pp. 2591–2600
- Zhang, J., Jemmott, J., 2014. Unintentional Exposure to Online Sexual Content and Sexual Behavior Intentions Among College Students in China. *Asia Pacific Journal of Public Health*, Volume 27(5), pp. 561–571