A Lightweight Reliably Quantified Deepfake Detection Approach

Tianyi Wang
The University of Hong Kong, Department of Computer Science

Kam Pui Chow
The University of Hong Kong, Department of Computer Science

A LIGHTWEIGHT RELIABLY QUANTIFIED DEEPFAKE DETECTION APPROACH

Tianyi Wang and Kam Pui Chow*

The University of Hong Kong
Department of Computer Science
Hong Kong, China
{tywang,chow}@cs.hku.hk

ABSTRACT

Deepfake has brought huge threats to society such that everyone can become a potential victim. Current Deepfake detection approaches have unsatisfactory performance in either accuracy or efficiency. Meanwhile, most models are only evaluated on different benchmark test datasets with different accuracies, which could not imitate the real-life Deepfake unknown population. As Deepfake cases have already been raised and brought challenges at the court, it is disappointed that no existing work has studied the model reliability and attempted to make the detection model act as the evidence at the court. We propose a lightweight Deepfake detection deep learning approach using the convolutional neural network backbone and the efficient convolutional attention mechanism, outperforming the state-of-the-art baseline models on each benchmark test dataset. Furthermore, a real-life Deepfake content is usually unknown about the corresponding source dataset or manipulation technique. We conduct a model reliability study using statistical random sampling from the available benchmark datasets to imitate the real-life Deepfake cases. A sufficient number of trials for model evaluation with random sampling derives the 95% and 90% confidence intervals, informing the reliable accuracy information of the proposed model. As a result, the reliably quantified detection model derives satisfactory accuracy and error rate to be applicable at the court for civil cases and provides an informative scheme to analyze future satisfactory approaches for criminal cases at the court.

Keywords: Deepfake Detection, Confidence Interval, Civil Case, Deep Learning

1. INTRODUCTION

The fraud of Deepfake images and videos has brought threats to human lives and caused challenges for prosecutions at the court (Shao, 2019; Harwell, 2021). In 2017, the Reddit user 'deepfakes' (deepfakes, 2019) announced to be able to generate high-quality celebrity pornography, which was the first appearance of the so-called Deepfake technique. Deepfake refers to a machine learning based face synthesis technique that swaps the face of another person onto the target one. Representative victims include famous singer Ariana Grande and actress Emma Watson (Kelion, 2018). As the circulating fake videos on the internet become hyper-realistic, Deepfake has been nominated as the most serious artificial intelligence crime threat in 2020. With various packaged applications and source code (Kemelmacher-Shlizerman, 2016; L. Li, Bao, Yang, Chen, & Wen, 2019; Natsume, Yatagawa, & Morishima, 2018) publicly available on the internet, anyone can become a potential victim of Deepfake (Melville, 2019; Kietzmann, Lee, McCarthy, & Kietzmann, 2020; Tolosana, Vera-

*Corresponding author.

Rodriguez, Fierrez, Morales, & Ortega-Garcia, 2020).

Existing Deepfake detection approaches mainly rely on deep learning models trained on the public benchmark datasets. Early solutions (Afchar, Nozick, Yamagishi, & Echizen, 2018; Nguyen, Yamagishi, & Echizen, 2019; Zhang, Zuo, & Zhang, 2018) frequently adopt convolutional neural networks (CNNs) solely as the model backbone but perform poorly on unseen datasets and manipulation techniques. To enhance the cross-dataset performance on unseen data, recent methods (Wodajo & Atnafu, 2021; Zhao et al., 2021; Luo, Zhang, Yan, & Liu, 2021) attempt various strategies by introducing more model parameters and adopting the well-performed attention mechanism from the transformer architecture (Vaswani et al., 2017). However, regardless of the largely improved but still unsatisfied cross-dataset performance, models with heavy parameters suffer from time-consuming problems. Moreover, existing work only evaluates the model performance on each benchmark test dataset and the accuracy varies depending on the dataset adopted in the experiments. On the contrary, people usually have no clue on a real-life Deepfake video about the source dataset or manipulation that the video is based on. To our knowledge, no existing approach has come up with a reliable detection model that applies to any arbitrary candidate Deepfake image, regardless of the source dataset and the manipulation technique, at a satisfactory accuracy rate with certain levels of confidence, which can perform as the evidence for prosecutions at the court.

In this paper, we address the time-consuming issue of the existing detection models with heavy parameters by introducing a lightweight approach using the CNN backbone network and the efficient convolutional attention mechanism. In specific, a candidate Deepfake facial image is passed to the CNN backbone and extracted the determinant facial features. Channel attention and spatial attention are sequentially performed upon the extracted image features for further refining. The detection result is determined based on the refined features following the attention mechanisms. The proposed model outperforms the existing state-of-the-art Deepfake detection models with high efficiency in the training process. Furthermore, to imitate the real-life Deepfake threat cases, we conduct scientific random sampling from the available benchmark test datasets and evaluate the model quantitatively with a sufficient number of trials and different sample sizes accordingly. We study the model reliability by investigating the 90% and 95% confidence intervals for the detection accuracy on any arbitrary candidate image after the intervals settle and converge with a sufficient number of trails. As a result, the model reliability study proves that our approach is satisfied to be applied to civil cases with respect to the balance of probability standard, and the scheme is informative for utilizing future detection models with promising performance in crime cases.

2. RELATED WORK

The original Deepfake technique refers to the deep learning face-swap technique utilizing autoencoders (Kingma & Welling, 2014) with a shared encoder and two unique decoders. For a pair of source and target identities, the shared encoder learns common facial features from the two identities, and the unique decoders each takes charge of generating faces with one of the identities. Later approaches have focused on boosting the autoencoder synthesized faces with further tuning techniques such as smoothing and blurring to eliminate obvious synthesis traces. Recently, the utilization of generative adversarial network (GAN) (Goodfellow et al., 2014) has achieved significant success in Deepfake generation. Specifically, the GAN architecture contains a generator to synthesize the fake faces and a discriminator to detect fake from the generated ones. The battle between generator and discriminator gradually enhances the quality and authenticity of the GAN generated fake faces.

Detection work has been proposed since the first occurrence of Deepfake contents. Early approaches (Afchar et al., 2018; Nguyen et al., 2019; Zhang et al., 2018) mainly exploit the traditional CNN architecture for feature learn-

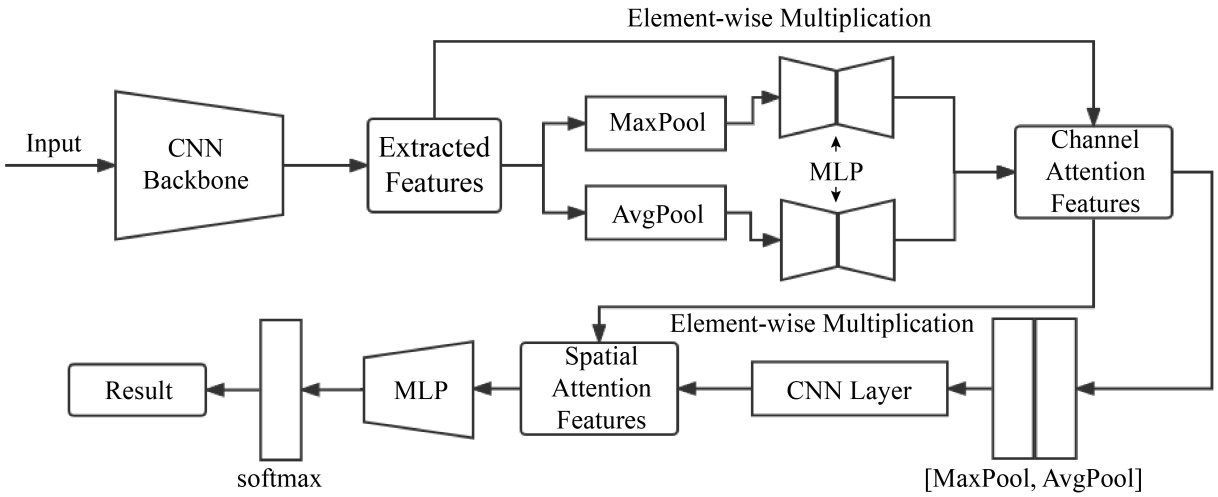

Figure 1: Framework of the proposed model.

ing and perform the classification of real and fake. Later, researchers frequently adopted the well-trained powerful CNN backbones for performance improvements (Rossler et al., 2019). The most recent detection work (Wodajo & Atnafu, 2021; Zhao et al., 2021; Luo et al., 2021) has started to combine the idea of CNN backbones with transformer attention techniques to learn better global information. As the matter of fact, these approaches either fail to achieve promising detection performance or suffer the time-consuming problem due to the heavy parameters.

3. METHODOLOGY

In this section, we present our proposed lightweight Deepfake detection model and the statistical scheme for model reliability study. The model framework is as shown in Figure 1. We first introduce the CNN backbone adopted for facial feature extraction, and then we explain the idea of the efficient convolutional attention mechanism. After that, we illustrate the scheme for the model reliability study.

3.1 CNN Backbone for Facial Feature Extraction

Convolutional neural network (CNN) has been proved its capability on images in visual tasks, especially for feature extraction. The input image passed through each convolutional layer is processed by a shifting filter window convolved across the width and height of the input volume. While shifting within the input image, each local receptive field is convolved and contributes as one element in the extracted feature map. In other words, a convolutional layer performs local feature extraction by convolving on each receptive field located by the shifting filter window.

In our detection approach, we perform facial feature extraction using a stack of convolutional layers that mainly follows the VGG (Simonyan & Zisserman, 2015) architecture with 16 convolutional layers. In particular, the convolutional layers are grouped into five blocks, where the first two blocks each contains two convolutional layers, while the other three each contains four. A max-pooling layer, a batch normalization (Ioffe & Szegedy, 2015) operation, and a ReLU (Xu, Wang, Chen, & Li, 2015) activation are applied in between every two blocks. For a colored input image with dimension $3\times224\times224$, the convolutional stack extracts a feature map with dimension $512\times7\times7$, where the first dimension refers to the number of channels and the rest two refer to height and width. The extracted facial features are then fed to the convolutional attention mechanisms to study the correlations among local features.

3.2 Convolutional Attention Mechanism

Although CNN is able to refine the useful features from the input image, only feature elements within the same receptive field are considered the inter-relationships locally. The idea of attention mechanism raised in the transformer architecture studies relationship between features that are far from each other and has achieved huge success in the natural language processing (NLP) domain. The transformer architecture is recently widely adopted to deal with images by studying the correlations between any two feature elements within the feature map. Although achieves relatively good performance, the transformer architecture suffers the time-consuming issue due to heavy parameters.

In this study, to maintain a lightweight efficient Deepfake detection model, we propose the convolutional attention mechanisms to study the local facial features extracted by the convolutional stack. In particular, we adopt the CBAM (Woo, Park, Lee, & Kweon, 2018) design and conduct channel attention and spatial attention sequentially on the extracted feature map. The channel attention takes charge of analyzing the inter-channel relationship of features. For the locally extracted features, max-pooling and average-pooling are each operated in one stream to narrow the features to have width and height both 1 in each channel and passed through a multi-layer perceptron (MLP). Results from the two streams are element-wisely summed up and performed sigmoid activation function for the channel attention score map. The channel attention score map M_c is computed by

$$M_c = \sigma(\text{MLP}(\text{AvgPool}(X))+\text{MLP}(\text{MaxPool}(X))), \tag{1}$$

where X represents the extracted local facial features and σ denotes the sigmoid function. The channel attention score is element-wisely multiplied back to the local facial features by

$$F_c = M_c \odot X, \tag{2}$$

where $\odot$ denotes the element-wise multiplication and F_c denotes the multiplication result. The spatial attention is followed to be performed on F_c to investigate the inter-spatial relationships. In specific, a max-pooling and an average-pooling are each conducted on F_c and concatenated together. The concatenated pooling results are fed to a convolutional layer and computed a spatial attention score map using the sigmoid activation function. The spatial attention score map M_s is computed by

$$M_s = \sigma(\text{Conv}(\text{Concat}(\text{MaxPool}(F_c), \text{AvgPool}(F_c)))), \tag{3}$$

and the result is element-wisely multiplied back to F_c by

$$F_s = M_s \odot F_c, \tag{4}$$

where F_s is the ultimate multiplication result that contains local facial feature information along with inter-channel and inter-spatial feature relationships.

It is worth noting that the convolutional attention mechanism does not modify the dimension of the locally extracted facial features. Thereafter, we flatten the feature map and append an MLP to gradually decrease the dimension to match the binary classification for real and fake and apply the softmax function to derive the final prediction on real and fake.

3.3 Model Reliability Study

The existing work has evaluated the models for both within- and cross-dataset experiments, deriving performance statistics for each benchmark test dataset. However, in real-life cases, we have no clue of the source dataset or manipulation technique of the candidate Deepfake content. Therefore, none of the existing work is able to make a claim about the model performance against an unknown Deepfake content in real-life. We thus conduct a model reliability study to make such a claim about our model.

To imitate the real-life unknown Deepfake data corpus, we combine the available benchmark test datasets as the population. Then, we sample a list of image data to evaluate the Deepfake detection accuracy and repeat the process

for a sufficient number of trials. We keep a balanced ratio for real and fake samples to derive a convincing result. The mean value $\bar{x}$ and sample standard deviation s can be computed based on the evaluated accuracies. After that, we derive the confidence interval CI by

$$\text{CI} = \bar{x} \pm t_c \frac{s}{\sqrt{n}}, \tag{5}$$

where n denotes the sample size. The parameter t_c for confidence level c can be computed by

$$t_c = I_{n-1} \frac{1-c}{2}, \tag{6}$$

where I_{n-1} is the student's inverse cumulative distribution function with n-1 degrees of freedom. Since bias occurs when the sample size is insufficient to represent the population distribution, we repeat the process of confidence interval computation for multiple attempts with different sample sizes until the confidence interval converges and settles. The model reliability study scheme can be summarized as Algorithm 1.

Algorithm 1: Model Reliability Study

```
Let real_data be the list of the pre-processed real face images
Let fake_data be the list of the pre-processed fake face images
Let n be the number of sampling trials for every sample size
Let s_lst be the list of int representing different sample sizes
Let model be the well-trained detection model
for i ← 0 to len(s_lst) − 1 do
    acc_lst = []
    s = s_lst[i]
    for j ← 0 to range(n) − 1 do
        shuffle real_data and fake_data
        samples = first s/2 images in real_data + first s/2 images in fake_data
        accuracy ← model(samples)
        acc_lst.append(accuracy)
    end
    compute mean and std of acc_lst
    compute the 90% and 95% confidence intervals
end
```

4. EXPERIMENT

In this section, we present the experimental results and discuss the findings. We first introduce the dataset adopted in this study. Then, we briefly demonstrate the implementation details of the model and the reliability study. Thereafter, we illustrate the results and discuss the findings accordingly.

4.1 Datasets

Following the convention in the existing work, we adopted the FaceForensics++ (FF++) (Rossler et al., 2019) dataset as the training dataset of our proposed model. FF++ contains 1,000 real videos acquired from YouTube and 4,000 fake videos synthesized using four different face manipulation techniques (FaceSwap (FS), Deepfakes (DF), Face2Face (F2F) (Thies, Zollhofer, Stamminger, Theobalt, & Niessner, 2016), and NeuralTextures (NT) (Thies, Zollhöfer, & Nießner, 2019)), where 1,000 fake videos are generated by each technique. The dataset has provided an official split list with the ratio 720:140:140 for training, validation, and test datasets with different video compression levels, namely, raw, HQ (c23), and LQ (c40). To match up with the real-life video qualities, we adopted the HQ dataset and followed the official split and trained our model accordingly.

After the model is trained and tested on FF++, we also wanted to know the model performance against unseen datasets with unseen manipulations. Besides FF++, we considered other existing benchmark Deepfake datasets with different manipulation techniques for the cross-dataset evaluation, namely, Deepfake Detection Challenge (DFDC) (Dolhansky et al., 2020), Celeb-DF (Y. Li, Yang, Sun, Qi, & Lyu, 2020), and DeeperForensics-1.0 (DF-1.0) (Jiang, Li, Wu, Qian, & Loy, 2020). DFDC includes videos manipulated by eight different techniques, Celeb-DF is a high-quality dataset generated using improved FaceSwap technique, and DF-1.0

contains videos synthesized by DF-VAE and has seven levels of perturbation and distortion intentionally added. Every dataset contains a relatively similar amount of real videos as the fake ones for experimental purposes.

For each cross-dataset evaluation, we acquired the officially provided test datasets in the experiment. In the model reliability study, the population is composed of all test datasets. For all datasets, image frames are randomly extracted from each candidate video.

4.2 Implementation Details

The DLIB (King, 2021) library is used to crop the face area from each image frame, and all face images are uniform to the size of 224×224. The model is restricted by the loss function

$$L = \Sigma_{i=1}^{2} t_i \log(p_i), \tag{7}$$

where t_i is the ground truth value and p_i is the softmax prediction for class i upon the final output from the last fully connected layer. In the reliability study, we set the sample size to 10 and gradually increased the value until the confidence interval is settled. Two values for the number of trials with each sample size are selected, i.e., 500 and 3,000. Parameter t_c and confidence interval CI are computed using SciPy (Reddy, 2022) in python.

4.3 Results and Discussions

The proposed Deepfake detection model is trained on FF++ and first evaluated on the FF++ test dataset. Then, to verify the model transferability on unseen Deepfake manipulation techniques, we performed cross-dataset evaluation by testing the model ability on several other benchmark test datasets. We further adopted the existing state-of-the-art Deepfake detection models that have source code available and trained the models on the same dataset following the same experimental settings as ours. The purpose of training the existing models is to set up a comparative test and have an idea of how good the proposed model is. We tested each well-trained comparative model on the same group of unseen datasets and recorded their performance. Following the convention of Deepfake detection, we utilized accuracy and area under the receiver operating characteristic (ROC) curve (AUC) score as our evaluation metrics, where the accuracy is the proportion level of the number of correct classification decisions by the detection model and the AUC score describes the probability that a random positive example is positioned to the right of a random negative example, i.e., how good the positive examples and negative examples are classified apart. The accuracy is usually more reliable on a balanced dataset while the AUC score is more applicable to imbalanced datasets.

We present the experimental results in Table 1. It can be observed that our proposed Deepfake detection model has outperformed the existing state-of-the-art approaches for both within-dataset and cross-dataset evaluations. For the experiment on FF++, our model achieves the highest AUC value of 97.36% against other models, which means the real and fake faces are nearly perfectly classified apart. As for the cross-dataset evaluation on the unseen benchmark datasets, all models suffer a significant performance drop compared to that on FF++. It is reasonable to obtain such results because it is usually more difficult for models to detect Deepfake especially when the candidate images are synthesized using other manipulation techniques different from the ones seen in the training dataset. Among all the comparative models, the latest MAT (Zhao et al., 2021) and Two-Stream (Luo et al., 2021) have achieved relatively better performance than the older ones (Afchar et al., 2018; Nguyen et al., 2019; Zhang et al., 2018) using traditional CNN architectures. The classic Xception (Rossler et al., 2019) model has also shown promising performance over the older ones due to its robust XceptionNet (Chollet, 2017) backbone. The transformer-based CViT (Wodajo & Atnafu, 2021) approach achieves good performance on some datasets but is unstable on the rest ones.

In order to make the detection models straightforward to be used by law reinforcement in actual criminal prosecution and defence cases, we studied the model reliability following the workflow in Algorithm 1. As Table 2 shown,

Method	Test Dataset							
	FF++		DFDC		Celeb-DF		DF-1.0	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
MesoNet	61.03	58.13	50.02	50.16	36.73	50.01	50.05	50.21
Capsule	76.40	83.44	51.30	56.16	61.96	59.93	59.29	61.46
FFD	82.29	82.48	59.44	59.47	46.19	55.86	53.69	53.81
CViT	83.05	91.08	60.76	67.43	53.26	63.60	54.97	58.52
MAT	87.50	94.85	63.16	69.56	44.78	57.20	56.90	61.72
Two-Stream	88.17	94.93	59.93	64.80	52.95	60.90	55.83	62.54
Xception	90.08	96.51	58.77	66.95	54.24	65.86	54.76	67.03
Our Approach	**91.67**	**97.36**	**64.78**	**70.21**	**63.35**	**66.04**	**69.84**	**79.12**

Table 1: Frame-level comparative tests accuracy (%) and AUC scores (%) on the testing datasets after trained on FF++.

the mean value and the confidence intervals have gradually settled as the sample size achieves 30,000 with 500 trials for each sample size. The values are rounded to four decimal places (i.e., to 0.01%). To ensure that the number of trials for each sample size is large enough, we conducted another reliability study with 3,000 trials. As Table 3 shown, the experiment with 3,000 trials for each sample size derives similar confidence intervals as the sample size reaches 30,000. The standard deviation values for the two experiments both decrease as expected. For a sufficient number of trials with large sample size, the sampled dataset is a reasonable imitation of the unknown Deepfake distribution. Therefore, based on the outcome in Table 3, we are 95% confident to claim that our Deepfake detection model has an accuracy between 74.25% and 74.41% on the classification of real and fake upon an arbitrary candidate image. In other words, the error rate of the model is between 25.59% and 25.75% accordingly. Meanwhile, we are 90% confident to claim that our Deepfake detection model has an accuracy between 74.26% and 74.39% on the classification of real and fake upon an arbitrary candidate image. And the corresponding error rate is between 25.61% and 25.74%.

We now present an illustrative example in which the above results are useful to be applied. In March 2021, a Bucks County mom was accused of creating Deepfake videos of the underage girls on her daughter's cheerleader team and threatening them to quit the team (Chinchilla, 2021). The videos exhibit the girls that were naked, drinking alcohol, or vaping, and are accused to be fake. Two months later in May, the prosecutors admitted that they could not prove the fake-video claims (Harwell, 2021). We applied our well-trained model to the videos that have brought challenges to the prosecutors. Due to sensitive content, we only acquired the vaping image frames from the news clip (Edition, 2021). As a result, 75 out of 77 image frames are classified as fake by our model. Some representative frames that are classified as fake are displayed in Figure 2. Based on the model reliability study result, we are 95% confident to claim that our model has an accuracy between 74.25% and 74.41% to correctly classify the cheerleader vaping clip as fake, and the error rate for the classification is between 25.59% and 25.75%.

Although successfully outperforms the state-of-the-art baseline models on every benchmark Deepfake dataset, the classification accuracy with 95% confidence interval as demonstrated in the model reliability study is not satisfied to act as evidence for criminal cases at the court. However, as the defendant and the victim are usually two individuals, the civil case is more applicable to the Deepfake prosecution. According to the words of Lord Nicholas in Re H (Minors) [1996] AC 563, based on the balance of probability standard, a court is satisfied an event occurred if the court considers that, on the evidence, the oc-

Sample Size	10	100	1,000	5,000	10,000	20,000	30,000
95% Interval	65.23–84.69	71.27–77.80	73.25–75.35	73.87–74.76	74.04–74.63	74.14–74.50	74.20–74.44
90% Interval	66.80–83.12	71.80–77.27	73.42–75.18	73.94–74.68	74.09–74.58	74.17–74.47	74.22–74.42
Mean	74.96	74.53	74.30	74.31	74.34	74.32	74.32
Std	13.11	4.40	1.42	0.60	0.40	0.24	0.17

Table 2: 95% and 90% confidence interval of the model accuracy (%) with 500 trials for different sample sizes.

Sample Size	10	100	1,000	5,000	10,000	20,000	30,000
95% interval	68.10–80.91	72.29–76.34	73.69–74.97	74.07–74.62	74.14–74.51	74.21–74.44	74.25–74.41
90% interval	69.13–79.88	72.62–76.01	73.79–74.86	74.11–74.57	74.17–74.48	74.23–74.42	74.26–74.39
mean	74.50	74.32	74.33	74.34	74.32	74.32	74.33
std	13.86	4.37	1.38	0.59	0.40	0.25	0.17

Table 3: 95% and 90% confidence interval of the model accuracy (%) with 3,000 trials for different sample sizes.

currence of the event was more likely than not, in other words, larger than the 50% likelihood. Thus, our Deepfake detection model is reliable with an accuracy between 74.25% and 74.41% on any candidate fake image with the 95% confidence interval in civil cases according to the balance of probability standard.

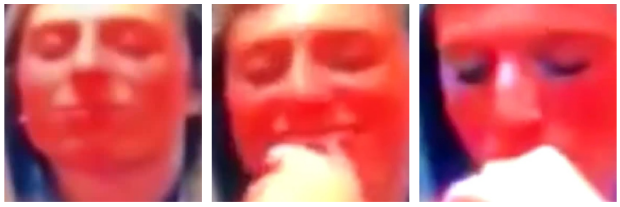

Figure 2: Representative frames from the cheerleader video that are classified as fake by our model.

5. CONCLUSION

In this study, we present a lightweight Deepfake detection model using the CNN backbone and convolutional attention mechanism, outperforming the state-of-the-art detection models with promising performance for both within-dataset and cross-dataset evaluations. Besides, we propose the reliability study to ensure the model is reliably quantified using statistical techniques with random sampling, which has never been considered in any existing Deepfake detection work. The derived model accuracy level with a 95% confidence interval has satisfied the civil case balance of probability standard and is proved to be reliable for Deepfake related civil cases at the court. Although not sufficient to be adopted as the evidence for criminal cases, our model reliability study has provided a clear scheme for future satisfactory detection models on Deepfake.

REFERENCES

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018, Dec). Mesonet: a compact facial video forgery detection network. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. doi: 10.1109/wifs.2018.8630761

Chinchilla, R. (2021). *Mom made deepfake nudes of daughter's cheer teammates to harass them: Police.* `shorturl.at/bfFKZ`. (Accessed: 2022-03-05)

Chollet, F. (2017). *Xception: Deep learning with depthwise separable convolutions.*

deepfakes. (2019). *Faceswap.* `https://github.com/deepfakes/faceswap`. (Accessed: 2022-03-05)

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C.

(2020). *The deepfake detection challenge (dfdc) dataset.*

Edition, I. (2021). *Cheerleader's mom allegedly made deepfakes of teammates.* shorturl.at/mtGHS. (Accessed: 2022-03-05)

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 27). Curran Associates, Inc.

Harwell, D. (2021). *Remember the 'deepfake cheerleader mom'? prosecutors now admit they can't prove fake-video claims.* shorturl.at/hwBFQ. (Accessed: 2022-03-05)

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd international conference on international conference on machine learning - volume 37* (p. 448–456). JMLR.org.

Jiang, L., Li, R., Wu, W., Qian, C., & Loy, C. C. (2020). DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In *Cvpr* (pp. 2889–2898).

Kelion, L. (2018). *Deepfake porn videos deleted from internet by Gfycat.* shorturl.at/hqxV5. (Accessed: 2022-03-05)

Kemelmacher-Shlizerman, I. (2016, July). Transfiguring portraits. *ACM Trans. Graph.*, *35*(4). Retrieved from https://doi.org/10.1145/2897824.2925871 doi: 10.1145/2897824.2925871

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons*, *63*(2), 135-146. (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING) doi: https://doi.org/10.1016/j.bushor.2019.11.006

King, D. (2021). *dlib 19.22.1.* https://pypi.org/project/dlib/. (Accessed: 2022-03-05)

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In Y. Bengio & Y. LeCun (Eds.), *2nd international conference on learning representations, ICLR 2014, banff, ab, canada, april 14-16, 2014, conference track proceedings.*

Li, L., Bao, J., Yang, H., Chen, D., & Wen, F. (2019). Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*.

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Ieee conference on computer vision and patten recognition (cvpr)* (pp. 3207–3216).

Luo, Y., Zhang, Y., Yan, J., & Liu, W. (2021, June). Generalizing face forgery detection with high-frequency features. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 16317-16326).

Melville, K. (2019). *The insidious rise of deepfake porn videos — and one woman who won't be silenced.* shorturl.at/euwT5. (Accessed: 2021-08-29)

Natsume, R., Yatagawa, T., & Morishima, S. (2018). Rsgan: Face swapping and editing using face and hair representation in latent spaces. In *Acm siggraph 2018 posters.* New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3230744.3230818 doi: 10.1145/3230744.3230818

Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). *Use of a capsule network to detect fake images and videos.*

Reddy, T. (2022). *Scipy.* https://scipy.org/. (Accessed: 2022-03-05)

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019, October). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the ieee/cvf international conference on computer vision (iccv)* (pp.

1–11).

Shao, G. (2019). *What 'deepfakes' are and how they may be dangerous.* shorturl.at/bnJZ7. (Accessed: 2022-03-05)

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Y. Bengio & Y. LeCun (Eds.), *3rd international conference on learning representations, ICLR 2015, san diego, ca, usa, may 7-9, 2015, conference track proceedings.* Retrieved from http://arxiv.org/abs/1409.1556

Thies, J., Zollhöfer, M., & Nießner, M. (2019, July). Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.*, *38*(4). doi: 10.1145/3306346.3323035

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Niessner, M. (2016, June). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)* (pp. 2387–2395).

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, *64*, 131-148. doi: https://doi.org/10.1016/j.inffus.2020.06.014

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Wodajo, D., & Atnafu, S. (2021). *Deepfake video detection using convolutional vision transformer* (Vol. abs/2102.11126). Retrieved from https://arxiv.org/abs/2102.11126

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018, September). Cbam: Convolutional block attention module. In *Proceedings of the european conference on computer vision (eccv).*

Xu, B., Wang, N., Chen, T., & Li, M. (2015). *Empirical evaluation of rectified activations in convolutional network.*

Zhang, K., Zuo, W., & Zhang, L. (2018). Ffdnet: Toward a fast and flexible solution for CNN based image denoising. *IEEE Transactions on Image Processing*.

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021, June). Multi-attentional deepfake detection. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 2185-2194).