Interpretable-through-prototypes deepfake detection for diffusion models

Agil Aghasanli*
Lancaster University
Lancaster, United Kingdom
a.aghasanli@lancaster.ac.uk

Dmitry Kangin*
Lancaster University
Lancaster, United Kingdom
d.kangin1@lancaster.ac.uk

Plamen Angelov
Lancaster University
Lancaster, United Kingdom
p.angelov@lancaster.ac.uk

Abstract

The process of recognizing and distinguishing between real content and content generated by deep learning algorithms, often referred to as deepfakes, is known as deepfake detection. In order to counter the rising threat of deepfakes and maintain the integrity of digital media, research is now being done to create more reliable and precise detection techniques. Deep learning models, such as Stable Diffusion, have been able to generate more detailed and less blurry images in recent years. In this paper, we develop a deepfake detection technique to distinguish original and fake images generated by various Diffusion Models. The developed methodology for deepfake detection takes advantage of features from fine-tuned Vision Transformers (ViTs), combined with existing classifiers such as Support Vector Machines (SVM). We demonstrate the proposed methodology's ability of interpretability-through-prototypes by analysing support vectors of the SVMs. Additionally, due to the novelty of the topic, there is a lack of open datasets for deepfake detection. Therefore, to evaluate the methodology, we have also created custom datasets based on various generative techniques of Diffusion Models on open datasets (ImageNet, FFHQ, Oxford-IIIT Pet). The code is available at `https://github.com/lira-centre/DeepfakeDetection`.

1. Introduction

A variety of social and security challenges have emerged in the current digital environment as a result of the widespread use of deepfakes, which convincingly modify media produced using cutting-edge deep learning algorithms. This endangers verifiability and validity of digital media. The availability of existing software platforms, which can be easily accessible by end-users, opens the door to dangerous applications. The study [37] shows that spreading false information through artificially generated content, such as text and visuals, can deceive millions of users and reduce trust in social media platforms; therefore, it may have detrimental effects on society. Additionally, various records [27] show a dramatic increase in deepfake-based theft and fraud over recent years. In certain similar settings, such as immersive virtual reality, it has been demonstrated to interfere with human psychological mechanisms, such as implanting false memories in individuals who have been exposed to fake content continually [16, 33].

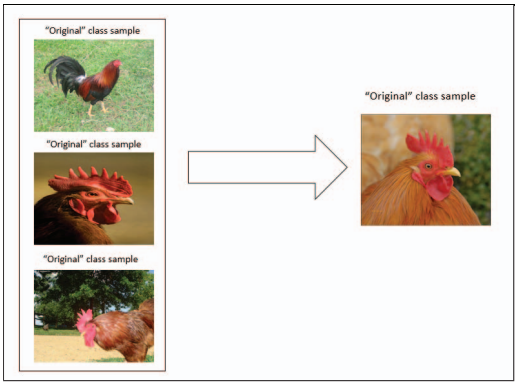


Figure 1: Example of SVM explainability: the image on the right side represents the sample query image, while the left side shows images corresponding to the closest (top to bottom) three support vectors in feature space. For more details, see subsection 4.3

In order to analyze and spot visual anomalies and inconsistencies in media, deep learning has emerged as an effective tool for deepfake detection. Deep learning models have the potential to mitigate the risks deepfakes pose. Another crucial step in addressing the problem of DeepFakes is creation of datasets using deepfake generating techniques. These datasets offer representative samples of changed media to academics and industry specialists, enabling creation and assessment of reliable detection systems. Some big technology companies, such as Meta [12], have partnered with academics to create datasets to over-

*Dmitry Kangin and Agil Aghasanli are both first authors of this paper.

come the issues of deepfakes. By training deep learning techniques on those datasets, we can effectively recognize and distinguish between fake and original content.

In this paper, we propose a methodology based on Vision Transformers to distinguish deepfake images, which are generated by diffusion models, from the original ones. We have also created custom datasets to train our models by utilizing various Latent Diffusion and Stable Diffusion model-based fake generation techniques on open datasets (ImageNet [9], Oxford-IIIT Pet [28], FFHQ [21]). In addition, we also provide explainability to understand the decision-making process through eXplainable Deep Neural Network (xDNN) [1] and Support Vector Machine (SVM) classifiers. xDNN is a prototype-based interpretable architecture that showed great performance on various tasks, such as semantic segmentation of satellite images for Earth Observation [42] and detection on CT scans for Covid-19 identification [35]. This paper aims to contribute to this field by focusing on three key aspects:

- We have developed a deepfake detection model based on fine-tuned and non-fine-tuned (pre-trained on ImageNet-1K) Vision Transformer features and various classifiers to perform comparative analysis.
- We have created datasets for the deepfake detection task based on images generated using different fake generation techniques with diffusion models, such as class-conditional and unconditional fake image generation with Latent Diffusion models and a text-guided image-to-image generation using Stable Diffusion models.
- We provide interpretability to understand how and why particular predictions are made. We use SVM and xDNN classifiers to figure out models' behavior by analyzing the closest support vectors and prototypes for each classifier, respectively.

2. Related Work

Deepfake Generation. With the development of deep learning models and algorithms, deepfake generating techniques have advanced quickly. By training discriminator and generator networks in an adversarial way, GANs [15] have demonstrated an outstanding ability in synthesising very realistic and convincing deepfakes. During training, the discriminator network tries to separate fake material from original content while the generator network generates fake content. Several GAN-based techniques [3, 26, 14, 41] have been applied to generate high-quality DeepFake images, specifically in the domain of DeepFake Face generation. Moreover, FaceForensics++ [32] is an open dataset that contains images generated by GAN, and it has been used as a benchmark dataset for DeepFake Detection Challenge.

Utilising Variational Autoencoders (VAE) [22] is yet another noteworthy method for deepfake generation. By mapping random noise to the learnt latent space, a Variational Autoencoder (VAE) learns a compressed representation of the data distribution and enables controlled generation of fake samples. Various modifications of VAE [34, 38, 40] have been developed in recent years and have shown great performance in creating DeepFake images on open datasets, such as CIFAR-10 [24], MNIST [10], FFHQ and ImageNet. Additionally, Celeb-DF [25] and DeeperForensics-1.0 [20] have been extensively used for benchmarking of deepfake detection models, where images are generated with Autoencoder models in those datasets.

The generation of realistic fake samples with sharper features and less blurriness is made possible by diffusion models [18], which involve repeatedly updating an initial noise distribution to match the desired data distribution. Diffusion models are able to generate more highly detailed images than GANs [11] and VAEs [19], which make the deepfake detection task hypothetically more complex. However, there exist very few open datasets [4] for benchmarking of deepfake Detection on Diffusion Model generated images.

Deepfake Detection. To classify deepfake images, some methods use Convolutional Neural Networks (CNN) [23, 5]. The work [6] utilizes optical flow fields in order to exploit inter-frame correlations. Target-specific region extraction layer for the CNN architecture [36] has also been used to feed only the most important information to the model. The proposed model [2] is based on feature extraction from the trained CNN model and XGBoost for classification. The composite method [29], which consists of state-of-the-art Deep Learning models, is also used on the DeepForensics++ dataset.

In addition, Vision Transformers are also applied to the deepfake detection problem in recent years. The use of EfficientNet as a feature extractor for the Vision Transformers model [7], and combining CNN features with patch embeddings [17] have been analyzed to distinguish fake and original contents.

Not only there is a lack of open datasets for deepfake detection with images generated by diffusion models, but also very few works [8, 30] have examined the detection of diffusion models' fake images.

3. Methodology

3.1. Dataset Creation

In this section, we describe the creation process for four different datasets for deepfake detection based on open datasets (Oxford-IIIT Pet, ImageNet, FFHQ) with the help of several fake generation techniques of Latent Diffusion

and Stable Diffusion Models (class-conditional, unconditional, image-to-image generation).

Oxford-IIIT Pet Deepfake dataset. A popular dataset for fine-grained image classification tasks is the Oxford-IIIT Pet Dataset. It is made up of images of cats and dogs from 37 different breeds. We have created a new dataset for deepfake detection based on Oxford-IIIT Pet in this work. To create the dataset, we used the Stable Diffusion Model's text-guided image-to-image fake generation technique [31]. This approach is based on Denoising Diffusion Implicit Models (DDIM), and requires a text prompt and an input image for conditioning to generate fake images. We took all images from the Oxford-IIIT Pet dataset (approximately 7400 images) and combined them with the same number of fake images, which are generated by the Stable Diffusion Model's image-to-image generation approach for each image in the dataset with resolution 448x320 while setting class names as a text prompt; so, the new Oxford-IIIT Pet Deepfake dataset consists of roughly 14800 images as a whole. We set the DDIM steps to 30, the downsampling factor to 8, and the strength of noising to 0.75 (1.0 refers to the total destruction of information in the initial image).

Figure 2: Original samples from Oxford-IIIT Pet dataset and corresponding fake generated images by Stable Diffusion's image-to-image technique

ImageNet-1K Subset Deepfake dataset with class-conditional approach. The ImageNet-1K dataset, sometimes referred to as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset, is a frequently used benchmark dataset in the field of computer vision. It consists of a set of labeled images representing 1,000 distinct classes with approximately 1300 images for each class for the training. We have selected only a small subset of the ImageNet-1K dataset, which consists of 10 classes, for this work; then, we created a deepfake detection dataset based on that. We have used Latent Diffusion Model's (LDM) class-conditional image synthesis approach with the pre-trained LDM on ImageNet-1K. It uses the images from the sample (specific class images) to generate high-quality fake images. We have generated 1000 images for each class in the selected subset of ImageNet-1K by class conditioning with resolution 256×256 and combined them with the original images in the same subset; hence, the new deepfake detection dataset based on the subset of ImageNet-1K consists of around 23000 images.

Figure 3: Examples of generated images on ImageNet with class-conditional LDM

ImageNet-1K subset deepfake dataset with image-to-image approach. To create this dataset, we have used the same subset of the ImageNet-1K dataset, but the image-to-image generation technique of the Stable Diffusion Model is utilized at this time. Similarly, the same number of images are generated in the original subset of ImageNet-1K by the image-to-image approach for each image with class names as a text prompt. Overall, this deepfake detection dataset is composed of around 26000 images. To generate corresponding fake images, we again set the parameter values to 30, 8, 0.75 for DDIM steps, downsampling factor and strength of noising, respectively.

FFHQ Deepfake dataset. FFHQ, or "Flickr-Faces-HQ," is a dataset that is often used in the fields of computer vision and machine learning. It is made up of high-quality, high-resolution face photos that were gathered from the photo-sharing website Flickr. FFHQ dataset for deepfake detection was built from 54000 original images from FFHQ and merged with 50000 fake generated human faces based on unconditional image synthesizing of Latent Diffusion Model, which pre-trained on FFHQ. Unconditional image synthesizing is generating new images without conditioning and any specific input where just only trained dataset features are concerned during the generation of fake images. To generate fake images, batch size was set to 10, number of steps ddim sampling was 50 and η for ddim sampling was 1.

Figure 4: Sample generated human faces by unconditional LDM

3.2. Deepfake Detection Model

The architecture used in this work is composed of two main parts: the feature extractor and the classifier. Vision Transformer [13] (ViT) takes advantage of the transformer architecture to recognize pictures as sequences of patches rather than as a grid of pixels. Instead of convolutional processing, it divides images into fixed-sized patches and flattens them; then, it feeds all this sequential data to transformer layers. The ViT model enables the attention mechanism to attend to both local patch-level information as well as the overall image-level context by incorporating the class token in the input sequence. It gives the model the ability to discover representations that include both fine-grained local information and high-level semantic information necessary for tasks like picture categorization. We used the specific Vision Transformer model, named Vit-L-32, and trained several classifiers (SVM, KNN, Naive Bayes, xDNN) based on extracted 1024 dimensional class tokens of the ViT model. We also compared the classifiers' performance on pre-trained features of ViT (trained on ImageNet-1K) or fine-tuned features of the model trained on custom datasets in this work as a binary classification problem (original or fake), where it will be described more specifically in the section 4 of this paper.

The architecture is shown in Figure 5 in more detail.

4. Results

4.1. Testing on single datasets

We fine-tune a ViT-L-32 architecture, pre-trained on ImageNet-21K, on our custom datasets. We present the problem as the one of binary classification (original or fake), where MLP is used in the classification part as in original ViT models. For every dataset, the model has been trained over five training epochs, with the batch size of 32 for the FFHQ Deepfake dataset and 16 for others. While taking into account that Deepfake FFHQ is the largest dataset among them, we set batch size relatively greater value to accelerate the training process. We provided a training loss over the training steps in Figure 8.

We have evaluated several classifiers (KNN, SVM, Naive Bayes) using both pre-trained and fine-tuned ViT-L-32 architecture as a feature extractor by setting 10% of the whole data for testing. We can clearly see that SVM with pre-trained and MLP Head with fine-tuned features performed best among the classifiers for three datasets (Oxford-IIIT Pet Deepfake, with image-to-image ImageNet Deepfake, class-conditional ImageNet Deepfake) used in this study from Table 1. On the other hand, only SVM performed very well on pre-trained features, while we can say that all classifiers showed competitive results with fine-tuned results. 2D tSNE visualization of image-to-image ImageNet Deepfake subset training and testing samples with fine-tuned and pre-trained features are provided in Appendix A.

Classifier	Accuracy	Features
MLP Head	99.5%	fine-tuned
SVM	96.9%	pre-trained
SVM	93.0%	fine-tuned
KNN	87.2%	fine-tuned
Naive Bayes	85.9%	fine-tuned
KNN	77.2%	pre-trained
Naive Bayes	72.7%	pre-trained

(a) Results on Oxford-IIIT Pet Deepfake dataset

Classifier	Accuracy	Features
SVM	98.0%	pre-trained
MLP Head	96.7%	fine-tuned
SVM	93.4%	fine-tuned
KNN	90.2%	fine-tuned
Naive Bayes	89.5%	fine-tuned
Naive Bayes	79.9%	pre-trained
KNN	79.0%	pre-trained

(b) Results on image-to-image ImageNet Deepfake subset

Classifier	Accuracy	Features
MLP Head	99.7%	fine-tuned
SVM	96.0%	pre-trained
SVM	94.7%	fine-tuned
KNN	93.4%	fine-tuned
Naive Bayes	90.5%	fine-tuned
Naive Bayes	69.2%	pre-trained
KNN	65.8%	pre-trained

(c) Results on class-conditional ImageNet Deepfake subset

Table 1: Testing results on corresponding datasets across classifiers. All classifiers are prototype-based except MLP Head.

We have also tested and compared five classifiers (MLP Head, SVM, KNN, Naive Bayes, xDNN) on the FFHQ Deepfake dataset with fine-tuned features. It can be said that MLP Head classifier performed the best again among all classifiers, and xDNN also showed satisfactory results. The megaClouds layer generated by the xDNN can be visualized by tSNE [39] plots or Voronoi Tessellations to understand further how the model found prototypes and how separable they are. tSNE and Voronoi Tessellation plots, representing the xDNN MegaClouds layer on FFHQ Deepfake, are described in Figure 6 and Figure 7, respectively. As demonstrated in those figures, xDNN found 9 prototypes for the Original class and 18 prototypes for the Fake class. Considering that xDNN is a prototype-based and interpretable model, we also provided explainability for that classifier by extracting rules on the FFHQ Deepfake dataset, which will

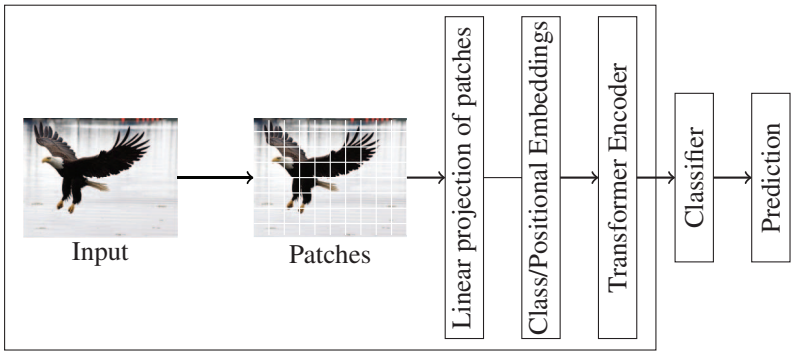


Figure 5: Deepfake detection model architecture

be described in the subsection 4.3 of this paper.

Classifier	Accuracy	Features
MLP Head	99.4%	fine-tuned
SVM	97.2%	fine-tuned
KNN	94.4%	fine-tuned
xDNN	91.1%	fine-tuned
Naive Bayes	85.2%	fine-tuned

Table 2: Results on FFHQ Deepfake dataset

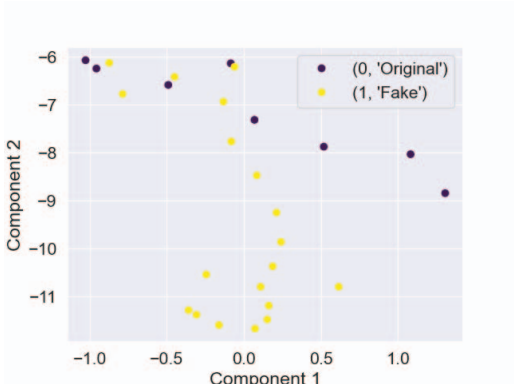


Figure 6: TSNE plot of xDNN MegaClouds layer on Deepfake FFHQ dataset

Model parameters The model implementation uses Python 3.10, torch 2.0.1 (with cuda support) and sklearn 1.2.0. SVM implementation uses RBF kernel; the rest of the parameters are default for sklearn implementation. For Naive Bayes, sklearn implementation of Gaussian Naive Bayes, which can be imported as GaussianNB from sklearn.naive_bayes, is used with default parameters. We also used sklearn implementation of KNN, while the best K value is selected based on the error rate across various K values between 1 and 40. The MLP architecture represents final two layers, finetuned to provide a binary classification output.

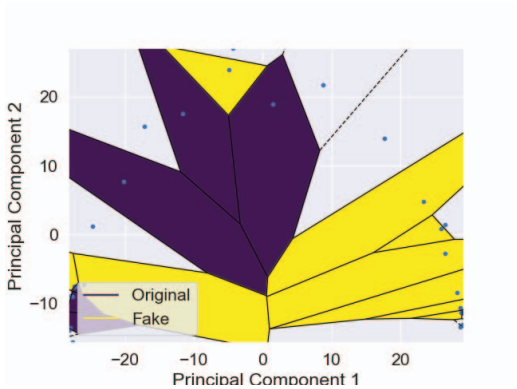
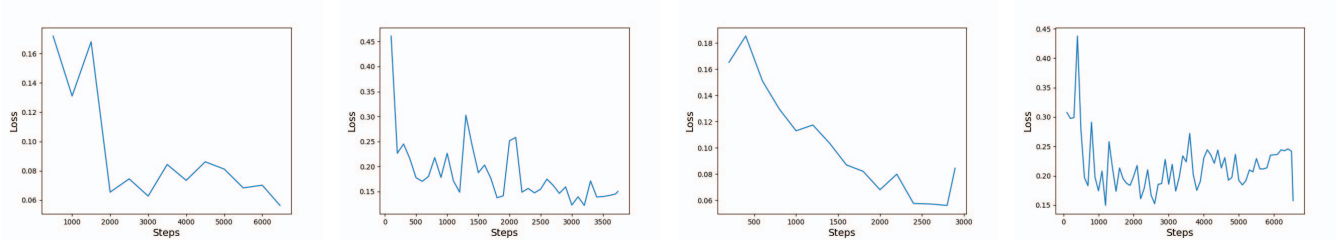


Figure 7: Voronoi Tessellation of xDNN MegaClouds layer on Deepfake FFHQ dataset

4.2. Testing with cross-dataset approach

We have tested classifiers on cross-dataset domains. Cross-dataset testing in machine learning refers to the evaluation of a trained model on one or more datasets that differ from the datasets used for training. By examining a model's capacity to produce precise predictions on unseen data from many sources or domains, it is possible to evaluate a model's generalization performance. When a model is applied to data that is different from the data it was trained on, this method aids in identification of any biases, overfitting, or performance restrictions.

The classifiers have been with both pre-trained and fine-tuned features, trained on the image-to-image ImageNet Deepfake subset and testing on the Oxford-IIIT Pet Deepfake dataset and vice versa. Both datasets have been cre-



(a) Deepfake FFHQ dataset fine-tuning (b) Deepfake Oxford-IIIT Pet dataset fine-tuning (c) Class-conditional ImageNet Deepfake dataset fine-tuning (d) Image-to-image ImageNet Deepfake dataset fine-tuning

Figure 8: Fine-tuning loss over the training steps

ated by image-to-image generation technique of the Stable Diffusion Model. Table 3 represents classification accuracies for the classifiers in all domains. It indicates that all classifiers can better generalize the data with fine-tuned features, which can be interpreted that they are more specific to distinguishing fake and original images. 2D tSNE plot of testing samples of Oxford-IIIT Pet dataset with fine-tuned features on image-to-image ImageNet Deepfake subset is provided in Appendix A. According to that plot, we can see that there are a large number of scattered points between the two classes compared to other tSNE plots, which can explain why KNN showed 50.1% accuracy while considering that it predicts a new data point based on its nearest neighbors.

Classifier	Accuracy	Features
MLP Head	81.3%	fine-tuned
SVM	80.9%	fine-tuned
KNN	79.4%	fine-tuned
Naive Bayes	75.7%	fine-tuned
SVM	74.8%	pre-trained
Naive Bayes	72.4%	pre-trained
KNN	50.0%	pre-trained

(a) Classification results on image-to-image ImageNet Deepfake subset by training on Oxford-IIIT Pet Deepfake dataset

Classifier	Accuracy	Features
MLP Head	84.0%	fine-tuned
SVM	80.5%	fine-tuned
Naive Bayes	76.4%	fine-tuned
KNN	75.0%	fine-tuned
SVM	64.7%	pre-trained
Naive Bayes	55.5%	pre-trained
KNN	50.1%	pre-trained

(b) Classification results on Oxford-IIIT Pet Deepfake dataset by training on image-to-image ImageNet Deepfake subset

Table 3: Results of cross-dataset approach

4.3. Interpretability

The interpretability for classifiers such as SVM or xDNN is provided through similarity through prototypes.

SVM. To interpret model's behaviour in particular situations, we have selected the three closest support vectors to the specific testing samples in the feature space. To calculate the distance between the sample and support vectors, we have used the Euclidean distance in n-dimensional feature space (in a case, $n = 1024$-dimensional) as represented in equation 1, and selected corresponding support vector samples from the training dataset to visually interpret the model behaviour through prototypes:

$$d\left(p,q\right) = \sqrt{\sum_{i=1}^{n}\left(q_i - p_i\right)^2} \tag{1}$$

xDNN. On selected use cases, we demonstrate interpretability with a prototype-based classifier xDNN. We calculated the similarity score (Euclidean distance in feature space) between an input image and all identified prototypes, so, we were able to extract the rules for each specific sample to explain the model's behavior as described:

$$\begin{aligned} R_{\mathrm{c}}:\ & IF\ (I \sim \hat{I}_1)\ OR\ (I \sim \hat{I}_2)\ OR\ ... \\ & OR\ (I \sim \hat{I}_{\mathrm{p}})\ THEN\ (class\ c) \end{aligned} \tag{2}$$

where I stands for the input image, P is the number of identified prototypes, and c is the class. The $\sim$ sign simply denotes the similarity.

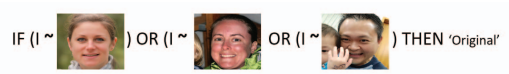


Figure 9: Use case example of xDNN Interpretability on 'Original' class of FFHQ Deepfake dataset with top 3 closest prototypes on feature space

5. Conclusion

In this paper, we developed a deepfake detection model based on Vision Transformer features to differentiate fake and original images, which are generated diffusion models. We created four different datasets as deepfake counterparts to open datasets (ImageNet, FFHQ, Oxford-IIIT Pet) with Stable Diffusion and Latent Diffusion models. We also presented interpretability techniques based on SVM and xDNN results to understand the reason why particular predictions are made by models. Various classifiers (SVM, KNN, Naive Bayes, xDNN) are compared with fine-tuned and non-fine-tuned (pre-trained on ImageNet-1K) features of Vision Transformer architecture (ViT-L-32) based on each custom deepfake detection dataset used in this work.

We demonstrate that foundation models such as ViT can be successful in telling apart real and fake images, including in the interpretable-through-prototypes learning scenarios. We showed that classifiers performed particularly well on single datasets on fine-tuned features, while SVM also showed great results on features of Vision Transformer pre-trained on ImageNet-1K. Moreover, the eXplainable Deep Neural Network (xDNN) model showed satisfactory results on the FFHQ Deepfake dataset, which is specifically a large dataset compared to others, with fine-tuned features. It can also be shown that the results of all classifiers with fine-tuned features outperform the results with pre-trained weights in the cross-dataset domain. In future work, we will focus on creating larger and more representative deepfake datasets to perform comparative analysis in the cross-dataset domain while considering that classifiers demonstrated better generalization ability with features of the fine-tuned model on relatively small deepfake detection datasets (Oxford-IIIT Pet Deepfake, image-to-image ImageNet subset Deepfake).

Acknowledgements

This work is supported by ELSA – European Lighthouse on Secure and Safe AI funded by the European Union under grant agreement No. 101070617. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

[1] Plamen Angelov and Eduardo Soares. Towards explainable deep neural networks (xdnn). *Neural Networks*, 130:185–194, 2020.

[2] Nency Bansal, Turki Aljrees, Dhirendra Prasad Yadav, Kamred Udham Singh, Ankit Kumar, Gyanendra Kumar Verma, and Teekam Singh. Real-time advanced computational intelligence for deep fake video detection. *Applied Sciences*, 13(5), 2023.

[3] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: Fine-grained image generation through asymmetric training. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2764–2773, 2017.

[4] Jordan J. Bird and Ahmad Lotfi. Cifake: Image classification and explainable identification of ai-generated synthetic images. *ArXiv*, abs/2303.14126, 2023.

[5] L. Bondi, Edoardo Daniele Cannas, Paolo Bestagini, and Stefano Tubaro. Training strategies and data augmentations in cnn-based deepfake video detection. *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2020.

[6] Roberto Caldelli, Leonardo Galteri, Irene Amerini, and Alberto Del Bimbo. Optical flow based cnn for detection of unlearnt deepfake manipulations. *Pattern Recognition Letters*, 146:31–37, 2021.

[7] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and F. Falchi. Combining efficientnet and vision transformers for video deepfake detection. In *International Conference on Image Analysis and Processing*, 2021.

[8] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. *ArXiv*, abs/2211.00680, 2022.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[10] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[11] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *ArXiv*, abs/2105.05233, 2021.

[12] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. The deepfake detection challenge dataset. *ArXiv*, abs/2006.07397, 2020.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[14] Zheng Fang, Zhen Liu, Tingting Liu, Chih-Chieh Hung, Jiangjian Xiao, and Guangjin Feng. Facial expression gan for voice-driven face generation. *Vis. Comput.*, 38(3):1151–1164, mar 2022.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[16] Jeffrey T. Hancock and Jeremy N. Bailenson. The social impact of deepfakes. *Cyberpsychology, Behavior, and Social Networking*, 24(3):149–152, 2021. PMID: 33760669.

[17] Young-Jin Heo, Woon-Ha Yeo, and Byung-Gyu Kim. Deepfake detection algorithm based on improved vision transformer. *Applied Intelligence*, 53(7):7512–7527, jul 2022.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[19] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47:1–47:33, 2021.

[20] Liming Jiang, Wayne Wu, Ren Li, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2886–2895, 2020.

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2018.

[22] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[23] Aditi Kohli and Abhinav Gupta. Detecting deepfake, faceswap and face2face facial forgeries using frequency cnn. *Multimedia Tools Appl.*, 80(12):18461–18478, may 2021.

[24] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[25] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3204–3213, 2019.

[26] Mingcong Liu, Qiang Li, Zekui Qin, Guoxin Zhang, Pengfei Wan, and Wen Zheng. Blendgan: Implicitly gan blending for arbitrary stylized face generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29710–29722. Curran Associates, Inc., 2021.

[27] Mekhail Mustak, Joni Salminen, Matti Mäntymäki, Arafat Rahman, and Yogesh K. Dwivedi. Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research*, 154:113368, 2023.

[28] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 2012.

[29] Md. Shohel Rana and Andrew H. Sung. Deepfakestack: A deep ensemble-based learning technique for deepfake detection. In *2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, pages 70–75, 2020.

[30] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. *ArXiv*, abs/2210.14571, 2022.

[31] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021.

[32] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, 2019.

[33] Kathryn Y Segovia and Jeremy N Bailenson. Virtually true: Children's acquisition of false memories in virtual reality. *Media Psychology*, 12(4):371–393, 2009.

[34] Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang, and Tarek Abdelzaher. ControlVAE: Controllable variational autoencoder. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8655–8664. PMLR, 13–18 Jul 2020.

[35] Eduardo Soares, Plamen Angelov, Sarah Biaso, Marcelo Cury, and Daniel Abe. A large multiclass dataset of CT scans for COVID-19 identification. *Evolving Systems*, June 2023.

[36] Van-Nhan Tran, Suk-Hwan Lee, Hoanh-Su Le, and Ki-Ryong Kwon. High performance deepfake video detection on cnn-based with attention target-specific regions and manual distillation extraction. *Applied Sciences*, 11(16), 2021.

[37] Cristian Vaccari and Andrew Chadwick. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1):2056305120903408, 2020.

[38] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19667–19679. Curran Associates, Inc., 2020.

[39] Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

[40] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, 2015.

[41] Xian Zhang, Xin Wang, Canghong Shi, Zhe Yan, Xiaojie Li, Bin Kong, Siwei Lyu, Bin Zhu, Jiancheng Lv, Youbing Yin, Qi Song, Xi Wu, and Imran Mumtaz. De-gan: Domain embedded gan for high quality face image inpainting. *Pattern Recogn.*, 124(C), apr 2022.

[42] Ziyang Zhang, Plamen Angelov, Eduardo Soares, Nicolas Longepe, and Pierre Philippe Mathieu. An interpretable deep semantic segmentation method for earth observation. In *2022 IEEE 11th International Conference on Intelligent Systems (IS)*, pages 1–8, 2022.