International Journal of Autonomous and Adaptive Communications Systems

ISSN online: 1754-8640 - ISSN print: 1754-8632
https://www.inderscience.com/ijaacs

Exposing deepfakes in online communication: detection based on ensemble strategy

Jie Xu, Guoqiang Wang, Tianxiong Zhou

DOI: 10.1504/IJAACS.2022.10049685

Article History:
Received: 17 November 2021
Accepted: 16 December 2021
Published online: 10 January 2024

Exposing deepfakes in online communication: detection based on ensemble strategy

Jie Xu*, Guoqiang Wang and Tianxiong Zhou

Changzhou Meteorological Bureau,
Changzhou, 213125, Jiangsu, China
Email: 80829917@qq.com
Email: 158118831@qq.com
Email: 602896860@qq.com
*Corresponding author

Abstract: In recent years, deepfake techniques have appeared in people's lives. As a product of deep learning, it can generate realistic face-swapping videos. Due to its high fidelity, deepfake is often used to produce porn videos and guide the public opinion, so as to pose a great threat to social stability. Previous studies have been able to improve detection accuracy. This paper aims to improve the detection ability of existing schemes by using the ensemble learning scheme from the perspective of model learning. Specifically, our scheme includes feature extraction, feature selection, feature classification, and a combination strategy. The experimental results on several datasets demonstrate that our scheme can effectively improve the detection ability of the model.

Keywords: deepfake detection; ensemble strategy; online communication; video forensics; deep learning.

Reference to this paper should be made as follows: Xu, J., Wang, G. and Zhou, T. (2024) 'Exposing deepfakes in online communication: detection based on ensemble strategy', *Int. J. Autonomous and Adaptive Communications Systems*, Vol. 17, No. 1, pp.24–38.

Biographical notes: Jie Xu is the Deputy Director of Meteorological Exploration Center of Changzhou Meteorological Bureau. He obtained a Bachelor's degree from the Northwest University of Technology. His research interests include comprehensive meteorological exploration.

Guoqiang Wang is the Director of Meteorological Exploration Center of Changzhou Meteorological Bureau. He obtained a Bachelor's degree from the Northwest University of Technology. His research interests include comprehensive meteorological exploration.

Tianxiong Zhou is the Master of the Meteorological Exploration Center Observation Station, Changzhou Meteorological Bureau. He obtained a Bachelor's degree from the Northwest University of Technology. His research interests include comprehensive meteorological exploration.

1 Introduction

In recent years, the development of the deepfake algorithm has led to the emergence of a large number of face-swapping videos. The deepfake algorithm, which is based on deep learning techniques, is able to replace the face of one individual in a video with a synthesised face of another individual. The core of deepfake video generation is to generate a realistic face that resembles the original face and has the same expression as the target face. The implementation of the algorithm usually relies on Auto-encoders, which is trained using a large number of face images to map the expression of the target face to the source face. The generated faces are then replaced on the target video to increase the realism of generated video through post-processing operations.

It is undeniable that deepfake can be used for positive purposes, such as movies about Chechnya that replace faces of actors in order to protect the identity of victims. However, a large number of deepfake videos circulating on the internet are still intended for negative purposes, with pornography accounting for the vast majority of them. In fact, the earliest deepfake video was a pornographic video of a famous actress, uploaded by a Reddit user named *deepfakes* in December 2017, which means that the deepfake algorithm could not escape being used for negative purposes since it was proposed. Except for generating privacy-invasive videos, deepfake algorithms are also applied to guide public opinion and even influence political politics. In 2019, a manipulated video of Nancy Pelosi was circulating on social media networks, showing Pelosi speaking in a muddled manner and looking very confused. The video was quickly reposted by President Trump and watched more than 1.4 million times, which was actually attempting to mislead voters at the time. Likewise, deepfake videos were also used in India's parliamentary elections to attack rivals by the Bharatiya Janata Party. In the coup of the Gabonese Republic, deepfake videos were also involved, which affected the political situation to some degree. Actually, even for someone without the relevant expertise, it's quite simple to generate deepfake videos, especially after the advent of video generation software such as Faceswap, Deepfacelab, etc. The real threat of Deepfake is that it's not in the hands of a few, but owned by the majority. Preventing threats brought by deepfake videos has become a significant issue to be addressed for people all around the world.

The past two years have witnessed the advance of the deepfake detection algorithm. Researchers have proposed many effective detection algorithms for the defects in the video, which have been well performed in various datasets. More and more scientific research institutions and commercial companies are also joining the queue for deepfake detection research. Recently, AWS, Facebook, and Microsoft have joined together to build the Deepfake Detection Challenge, attempting to inspire researchers around the world to build innovative new technologies to help detect deepfake videos. However, most existing works still rely on supervised learning, which leads to limited generalisation capability. In Deepfake Detection Challenge, the best detection accuracy was only 65.18% against the black-box dataset, although the top-performing model on the public dataset achieved 82.56% accuracy. It can be seen that the existing detection algorithms still have shortcomings in detection performance, so improving the detection ability is one of the important problems that need to be solved.

In this paper, we introduce an ensemble strategy in deepfake detection tasks to improve the performance of detection algorithms. Specifically, several feature extraction modules are employed to extract CNN features, which are then fed into ensemble classifiers. The main contributions of the paper are as follows.

1 We introduce ensemble strategies to improve the performance of existing detection neural networks, achieving superior results on various datasets. Experimental results also demonstrate the superiority of this strategy.

2 Through several comparative experiments, we confirm that ensemble strategies can further improve the detection accuracy of algorithms even for features extracted using a weak neural network.

In following, we will first introduce the related literature in Section 2. Then the background knowledge is described in Section 3. After that, we introduce the proposed scheme in Section 4. The experimental results are presented in Section 5, and the conclusion is finally summarised in Section 6.

2 Related work

In this section, we describe the recent work related to deepfake detection.

Data security has been a hot topic in recent years (Yang et al., 2018, 2020). Due to the characteristics of deepfake, i.e., difficulty of identity, fast propagation speed, and powerful destructive ability, it will lead to a risk of damage to personal privacy data, social stability, and even national security.

The majority of deepfake detection algorithms rely on deep learning. Based on deepfake dataset training, the subsequent feature extraction and classifier construction can be implemented. In recent years, deepfake detection in the visual aspect has attracted a lot of attention, which can be divided into image-based deepfake detection and video-based deepfake detection.

2.1 Image-based deepfake detection

According to different judging principles, image-based deepfake detection can be divided into the following four categories. The first category is the deepfake detection approach based on traditional image forensics. The second is the deepfake detection approach based on the customised modification of CNN architecture. The third is to train the classifier to detect deepfake images by analysing and extracting the different features of real and fake images. The fourth is the deepfake detection approach based on fingerprint features.

Deepfake detection approach based on traditional image forensics: Deepfake detection approaches based on traditional image forensics mostly rely on specific tampering and utilise frequency domain features and statistical features of images to distinguish. Fridrich and Kodovsky (2012) extracted the steganographic features of digital images by assembling the noise component model, and then the noise features were widely used in the field of image forensics. Lukáš et al. (2006) proposed to utilise camera equipment fingerprint light response nonuniformity of digital images to detect deepfake images.

However, these image-based forensics technologies pay more attention to local abnormal features, and can easily avoid the detection of deepfake detection methods based on traditional image forensics.

Deepfake detection approach based on the customised modification of CNN architecture: With the development of artificial intelligence, image forensics gradually integrate with deep learning. (Nataraj et al., 2019) achieved the detection of GAN-based generated fake images

by extracting cooccurrence matrices on RGB channel in pixel domain and constructing a pixel-level image detection model based on CNN. (Liu and Pun, 2018) proposed a new deep fusion network to locate the tampered area of the image by tracking the boundary. (Mo et al., 2018) modified the architecture of CNN (such as the number of layers and activation functions) and realised the detection of deepfake images in the form of supervised learning. The disadvantage is that it is vulnerable to attacks from adversarial examples.

Deepfake detection approach based on comparison of real and fake image features: Zhang et al. utilised Sped up robust features (SURF) (Bay et al., 2006) and Bag of words (BoW) to extract image features (Zhang et al., 2017). Yang et al. (2019) trained the classifier by extracting the different features of facial marker points between the real images and the fake images. However, with the development of GAN and deepfake generation model, the differences between real and fake images would gradually narrow and even disappear.

Hsu et al. (2020) proposed a novel Common fake feature network (CFFN), which can be divided into two stages. In the first stage, a large amount of real-fake image pairs are generated based on GAN, so that CFFN learns to identify features of real and fake images. The second stage is a CNN with the cross-level pseudo-feature providing capability. The fake images could be recognised based on the identified features extracted in the first stage.

Deepfake detection approach based on fingerprint feature: Zhang *et al.* proposed a classifier based on spectrum input, namely AutoGAN (Zhang et al., 2019b). It can realise the accurate detection of deepfake images generated by prevalent GAN models such as CycleGAN. However, deepfake generation methods would avoid the detection of deepfake detection model by selecting GAN without fingerprint features. Moreover, with the rapid development of GAN, the fingerprint features extracted by the above detection methods are not universal.

2.2 Video-based deepfake detection

Video-based deepfake tends to tamper specific areas of the face frame by frame. Visual artefacts and visual noise would appear in frames, and the continuity of the space-time state of the person would be inconsistent between frames. Therefore, the majority of static feature-based deepfake detection methods cannot be directly applied to video-based deepfake detection. The existing video-based deepfake detection methods can be divided into two categories. The first is based on inter-frame difference, and the second is based on intra-frame visual artefacts.

Deepfake detection approach based on inter-frame difference: Deepfake detection models generally use the static facial dataset for model training, which cannot accurately forge physiological information such as blinking, breathing, and heartbeat. Moreover, since deepfake video is generated frame by frame, continuous frames of it show differences in temporal and spatial distribution. Therefore, video-based deepfake detection methods can be implemented based on the rationality of physiological information.

In the existing deepfake generation methods, the reconstruction of the target video is mostly realised on the basis of frame-by-frame operations. And the time consistency between video frames cannot be effectively enhanced in the synthesis stage. Güera and Delp (2018) proposed a time-sensing pipeline method to detect deepfake videos based on CNN and long short-term memory (LSTM). However, this method is weak in robustness and vulnerable to attack from adversarial examples. On the basis of the above method, Sabir et al. (2019) extracted the features of the frame sequences by ResNet50 and DenseNet, which realised the detection of deepfake video by utilising the temporal difference between frame groups.

In addition, the optical flow method can be combined with CNN to realise video-based deepfake detection. The optical flow method utilises the change of pixels of video frames in time domain to find the corresponding relationship between the two adjacent frames, and calculates the optical flow vector of adjacent frames. Amerini et al. (2019) applied Lucas-Kanade (LK) and PWC-Net to estimate the faked video frames as optical flow. The differences in direction, size, and quantity between the optical flow vector formed frame by frame around the face in the real video and that of the deepfake video can be captured by CNN.

Deepfake detection approach based on intra-frame visual artefacts: Generally speaking, these methods extract discriminative features through exploring intra-frame visual artefacts. And these differences can be distinguished by deep learning models and other classification algorithms.

Afchar et al. (2018) proposed a neural network, namely MesoNet, which combined Meso-4 with MesoInception-4. The architecture is lightweight while it maintains high performance. They also proved that the features of eyes and mouth play a vital role in deepfake video detection. Zhang et al. (2019a) applied lightweight flow modules to replace the global average pooling layer and full connection layer of CNN, so that greatly reduces the computational cost when extracting the noise features of frames.

Deepfake videos generally need to accurately match the target face to the original videos based on face affine deformation technologies, such as scale, rotate and cut, so that the resolution of the synthesised video may be inconsistent between the facial region and the surrounding environment. Nguyen et al. (2019) combined capsule network with video-based deepfake detection methods. The Capsule network uses the vector to measure the probability of visual artefact features in deepfake videos. Compared with CNN, which represents neurons and weights with the scalar, the capsule network effectively reduces the detection error caused by the twist angle and direction differences of the video characters. Moreover, this model uses less training data to maximise the retention of valuable information. and is superior to CNN in feature extraction and noise resistance.

Sabour et al. (2017) also proved that the capsule network can accurately describe the hierarchical relationship between object components. On the basis of the mutation of pixel relations in some regions of deepfake videos, (Koopman et al., 2018) proposed a detection method based on photo response non-uniformity (PRNU). Moreover, according to the differences in intra-frame artefacts or inherent features between real and fake videos, Matern et al. (2019) designed a video-based deepfake detection method based on visual features such as eyes, teeth, and facial contouring. In this method, it can accurately detect fake videos only if certain prerequisites are met (such as eyes being open).

3 Preliminary knowledge

In machine learning, the goal of the supervised learning algorithm is to construct a stable model with good performance in all aspects. However, the actual situation tends to be not ideal. That is, sometimes we can only obtain multiple preference models, which are called weak supervised model. Ensemble learning is to combine these multiple weak supervised models to obtain a better and more comprehensive supervised model. The potential idea of ensemble learning is that even if a weak classifier makes a wrong prediction, other weak classifiers can correct the error.

Ensemble learning has good strategies on datasets of all sizes (Polikar, 2012). A large dataset can be divided into multiple small datasets, and then multiple learned models can be combined. Small datasets can be sampled by the Bootstrap method, so as to obtain multiple datasets. Then multiple learned models can also be combined.

Two branches of ensemble learning are Bagging and Boosting.

Bagging: It is the abbreviation of bootstrap aggregating (Bühlmann, 2012). Among them, bootstrap is a sampling method with playback, and the goal is to obtain the distribution of statistics and confidence intervals. The specific steps are as follows.

1 Extract a certain number of samples from the original samples by resampling method (sampling with playback).

2 Calculate the desired statistics T according to the extracted samples.

3 Repeat the above steps N times (generally greater than 1000) to obtain N statistics.

4 On the basis of these N statistics, confidence intervals for statistics can be calculated.

In the Bagging method, the Bootstrap method is utilised to obtain N datasets from the overall datasets by sampling with playback, and each model can be learned on each dataset. The final prediction results are obtained by the output of N models.

Taking random forest as an example, a forest is established in a random way. The forest is composed of a great number of decision trees, and each decision tree of the random forest is not relevant. The Bootstrap method needs to be used when each decision tree is learned. In random forest, there are two random sampling processes: sampling rows (number of data) and columns (characteristics of data) of input data.

For row sampling, take a playback way for N times, the sampled N data may be repeated. Therefore, each tree cannot cover all samples in the training process, which may not lead to overfitting relatively. Then, for column sampling, select m from M features ($m \ll M$), and the decision tree can be learned.

When predicting, each tree in the random forest predicts the input and votes. And the input sample belongs to the category which has the most votes. In other words, each classifier (each tree) is relatively weak, and the combined classifiers (votes) are strong oppositely.

Boosting It is a machine learning algorithm that can be used to reduce the deviation in supervised learning, which is also to learn a series of weak classifiers and combine them into a strong classifier (Wu et al., 2021). The representative algorithm of Boosting is AdaBoost (Adaptive Boosting) (Schapire, 2013). In the process of initialisation, each training case is assigned an equal weight of $\frac{1}{N}$, and then AdaBoost is used to train t rounds with the training set. After each training, a large weight is assigned to the training case which fails to train. That is to say, the learning algorithm is allowed to focus on the more difficult training

cases in the subsequent learning, so as to obtain a prediction function sequence $h_1, \ldots, h_m$. Among them, h_i also has a certain weight, and the weight of the prediction function with great prediction effect is larger, and vice versa. The final prediction function H adopts the weighted voting method for classification problems, and the weighted average method is used to discriminate the new cases for regression problems.

Boosting is an idea that says 'change by mistake', and Gradient Boosting is a method for function (model) optimisation under this idea. Firstly, the function is decomposed into addable forms. Then, m iterations are performed to reduce the loss in the gradient direction, and an excellent model is finally obtained. It is worth mentioning that the reduced part of the model in the gradient direction can be considered as a 'small' or 'weak' model. Finally, these 'weak' models can be combined into a better model by weighting.

Although both Bagging and Boosting adopt a sampling-learning-combining approach, there are some differences in details. For instance, each training set in Bagging is not related to each other, that is, each base classifier is not related to each other. However, the training set in Boosting is adjusted on the results of the previous round, which makes it impossible to compute in parallel. Moreover, the prediction function is uniform in Bagging, but weighted in Boosting.

When learning, there is commonly a trade-off between bias and variance. Therefore, in order to achieve the optical effect, the model must take into account bias and variance, that is, to adopt strategies to make the two more balanced.

From the algorithm itself, in order to ensure the stability of the model, Bagging focuses on the voting combination of multiple base models. Therefore, each base model is relatively complex to reduce the deviation. However, the strategy adopted in Boosting is to reduce the deviation of the previous round in each learning. On the basis of ensuring the deviation, each base classifier is simplified to make the variance smaller.

4 The proposed method

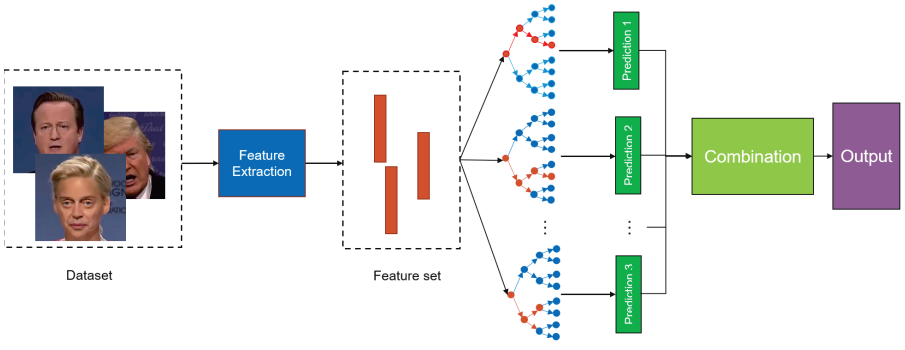
The main purpose of this research is to evaluate the effectiveness of ensemble strategies in deepfake detection tasks.

4.1 Overview

Ensemble learning has shown its advantages in various fields. In this paper, as shown in Figure 1, we design the corresponding scheme for deepfake detection, including feature extraction, feature selection, feature classification, and prediction fusion.

The feature extraction module adopts the feature extractor which is used in the best scheme of DFDC competition (Dolhansky et al., 2020). The excellent performance in the competition is the premise for it to extract high-quality forgery features. The feature selection module randomly selects the extracted features to construct the corresponding dataset for subsequent classification. In the feature classification module, we adopt a random forest classifier, which contains multiple decision trees to further improve the performance of ensemble learning. The final prediction fusion module adopts different fusion strategies to calculate multiple prediction results, so as to obtain the best fusion method.

Figure 1 The overview of the proposed method (see online version for colours)



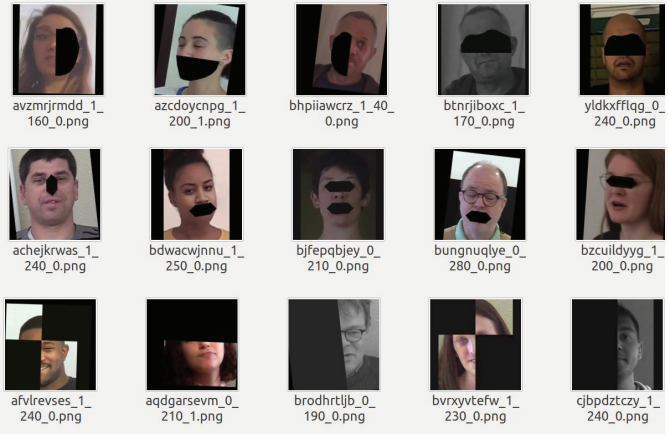
4.2 Feature extraction module

Feature extraction is a greatly significant step in the whole scheme, and the quality of features has an important impact on the subsequent steps.

To this end, we adopt the more excellent feature extractor in the existing scheme to extract features. Scheme 1 and 2 are as follows.

Scheme 1 (Method of Selim Seferbekov): A variety of data expansion methods are adopted to increase the detection ability of the algorithm for unknown data. As shown in Figure 2, multiple masking methods are used to reduce the influence of facial features on deepfake detection, so that the classifier pays more attention to the information of the deepfake area. The scheme integrates multiple EfficientnetB3 to obtain the best results in DFDC competition (ELmoufidi and Amoun, 2021).

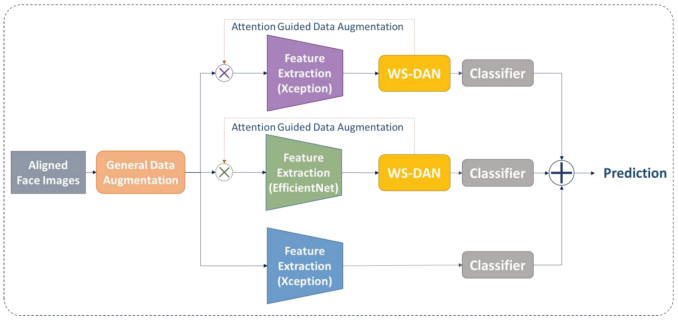
Figure 2 Method of Selim Seferbekov (see online version for colours)



Scheme 2 (Method of \WM/): A WS-DAN model (Yang et al., 2021) and an Xception model (Chollet, 2017) with different feature extractors are integrated to generate the face prediction of each frame, and then the aligned faces are the input. The prediction of each video is

calculated by the average of the prediction of all frames. The scheme introduces WS-DAN as the part of the network architecture to enhance the capability of data augmentation, which greatly helps to improve the performance of the model. As shown in Figure 3, for each training image, WS-DAN generates the attention map to represent the discriminant part of the object through weakly supervised learning. Then, WS-DAN augments the images guided by these attention maps, including attention clipping and attention decline.

Figure 3 Method of \WM/ (see online version for colours)



WS-DAN improves classification accuracy in two ways.

1 In the first stage, since more features are extracted from the discriminant part, the visual quality of the image is better.

2 In the second stage, the attention map provides the exact location of the object, which ensures that the model can observe the object more closely and further improve the performance.

4.3 Feature selection module

In this paper, we adopt the feature selection method in random forests. There are two objectives of feature selection. The first objective is to find the characteristic variables that are highly correlated with the strain. The second is to select a small number of characteristic variables which can fully predict the results of the variables.

The steps of general feature selection include:

1 Preliminary estimation and ranking

   a The characteristic variables in random forests are sorted in descending order of VI (Variable Importance).

   b Determine the deletion ratio, and remove the indicators that are not important in the corresponding ratio from the current feature variables, so as to obtain a new feature set.

   c Create a new random forest with a new feature set and calculate VI (Variable Importance) of each feature in the feature set and sort it.

   d Repeat the above steps until m features are left.

2 According to each feature set obtained in step 1 and the random forest established by them, the corresponding out-of-bag error rate (OOB err (Ciss, 2015)) is calculated, and the feature set with the lowest out-of-bag error rate is used as the final feature set.

4.4 Feature classification module

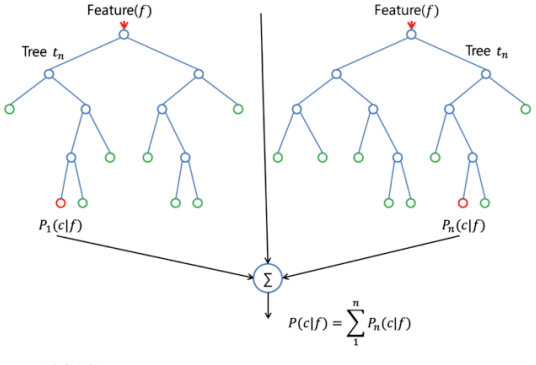
Random forest is an important ensemble learning method based on Bagging, which has high accuracy (Paul et al., 2018). At the same time, due to the introduction of randomness, the model is not easy to over-fit, and has a great ability to resist noise. It determines the final classification results according to the voting selection of the decision tree. For decision tree, if a set can be divided into multiple categories, the information of each category can be defined as follows:

$$I(X = x_i) = -\log_2 p(x_i) \tag{1}$$

Among them, $I(x)$ represents the information of random variables, and $p(x_i)$ represents the probability of occurrence. Entropy is used to measure the uncertainty of variables. When the entropy is larger, the uncertainty of variables belonging to a certain category is greater, and vice versa. Information gain is a measure used for feature selection in the random forest algorithm. The larger the information gain is, the better the selection of a feature is. Information, entropy, and information gain are the basis of decision tree and the basis for decision tree to determine the order of feature selection.

The structure of a random forest is shown in Figure 4. Random forest is an improvement based on Bagging, and its corresponding strategy is to select n samples from the sample set by Bootstrap sampling. Randomly select k attributes from all attributes, and then utilise information gain and Gini index method to find the best segmentation attributes to establish CART decision tree, where k controls the introduction of randomness. Repeat the above process to establish m classifiers, and these trees can be utilised to form a random forest, and then the prediction results are obtained by averaging.

Figure 4 The structure of a random forest (see online version for colours)



Source: Cotta (2018)

4.5 Prediction fusion module

In the final prediction fusion module, the prediction of different strategies is fused to obtain the final prediction results. The adopted strategies are as follows.

Simple average method: Weighted average method requires that the weight is nonnegative and the sum of weights is 1. Generally speaking, the weighted average method is adopted when the performance of the individual learning models varies greatly, and the simple average method is adopted when the performance is similar. For instance, the performance can be judged by the accuracy of the test set.

Voting method: Voting is the main method used for classification (Randhawa et al., 2018). In absolute majority voting method, if a mark gets more than half the vote, it is predicted as the mark. Otherwise refuse to predict. This method is effective in learning tasks with high-reliability requirements.

In the relative majority voting method, the prediction result is the marker with the most votes. If multiple markers get the highest votes at the same time, one is randomly selected.

Learning method: This method is typically combined with another learner. Stacking is a typical representative (Džeroski and Ženko, 2004). The individual learner is called primary learner, and the learner used for combination is called the secondary learner or meta-learner.

The basic idea of the learning method is as follows. First, the primary learner is trained with the initial dataset, and then the output of the primary learner is taken as the feature. Generally speaking, the primary learner is used to train the unused samples to generate the samples of the secondary learner by the leave-one-out method or cross-validation method. And the output features and the corresponding initial training set are combined into a new dataset for training the secondary learner.

5 Experiments

In this section, we first introduce our experimental setup. Then, the performance of ensemble learning algorithm is verified by experiments on multiple datasets.

5.1 Experimental setup

5.1.1 Dataset

We employed two datasets for effectiveness validation, including FaceForensic++ and DFDC.

FaceForensic++: It is a dataset composed of 1000 original video sequences, which are generated by four face operation methods: Deepfake, Face2Face, FaceSwap, and NeuralTexture. The original data comes from 977 YouTube videos, all of which contain front faces without occlusion, which makes the tampering method easy to generate realistic forgery videos. FaceForensic++ contains four sub-datasets, each of which consists of 1000 forgery videos and 1000 real videos. In our experiments, the training set contains 720 forgery videos and corresponding real videos, the validation set contains 140 forgery videos and corresponding real videos, and the test set contains 140 forgery videos and corresponding real videos.

DFDC++: It consists of face-swapping videos which are generated by a variety of methods, and these methods cover some of the most popular face-swapping technologies when creating the dataset. The source data of the dataset involves 3,426 objects, with an average of 14.4 videos per object. There are 48,190 videos in the dataset, and the average length of each video is 68.8 seconds. The training set includes 119,154 video clips, involving 486 different objects. Among these, 100,000 videos contain Deepfake content, meaning that 83.9% of the videos in the dataset are synthetic. The validation set is a public test set used to calculate the ranking position in DFDC competition. It contains 4,000 ten-second videos, half of which (2,000 videos) contain Deepfake content. The test set contains 10,000 ten-second videos. Like the public test set, half of them are Deepfake videos. However, the difference between the two is that in the private test set, half of the videos come from the network, and the other half comes from the source data.

5.1.2 Details

In the experiment, we use face clipping algorithms including Retinaface and MTCNN to extract faces in the video. All face images are unified to the size of 224*224.

In the training process, the stochastic gradient descent method is adopted, and the batch size is set to be 32. The learning rate starts from 0.0001, and then decreases by 0.9 after each epoch. The training process will be terminated after 20 epochs. We evaluate the classification ability of the model by calculating Area Under Curve.

5.2 Deepfake detection based on single model

We first test the detection effect of different backbones under a single model, so as to apply the ensemble learning algorithm to improve the detection effect according to the ability of a single model. As shown in Table 1, with the increase of model depth and complexity, the effect of the model is further improved.

Table 1 AUC performance based on single model

Dataset	*VGG16*	*VGG19*	*Resnet18*	*Resnet34*	*Resnet50*	*Resnet101*
Raw_DF	98.7	99.02	98.1	98.6	99.36	99.39
Raw_FS	98.51	98.6	95.95	97.62	98.66	98.86
Raw_F2F	98.26	98.25	97.34	96.97	97.9	97.96
Raw_NT	93	92.66	91.39	92.91	94.88	95.44
C23_DF	97.03	96.74	96.28	96.85	97.86	97.97
C23_FS	97.26	97.62	94.92	96.47	98.31	98.28
C23_F2F	95.99	96.55	94.11	94.42	96.68	96.56
C23_NT	87.34	87.67	84.19	86.83	90.8	91.86
C40_DF	92.48	92.84	89.65	91.25	94.78	95.77
C40_FS	90.54	90.95	82.58	87.1	94.29	95.4
C40_F2F	86.56	86.56	78.06	83.35	87.94	89.03
C40_NT	73.06	73.39	71.03	72.16	78.92	77.99

5.3 Deepfake detection based on ensemble learning

We first verify the effectiveness of the combination strategy on the Deepfake dataset. Specifically, we initially apply squeezenet as a feature extraction model, and adopt different

combination strategies for experiments. It can be seen from Table 2 that the strategy based on feature fusion is significantly higher than other combination strategies.

Table 2 comparison between different combination strategies

Combination strategy	*Description*	*Loss*	*Acc*
Average	–	14.94	94.75
Voting	–	14.94	94.79
Meta	Sigmoid layer	30.64	94.91
Meta	Feature	11.92	95.97

Furthermore, we utilised the feature fusion strategy to verify the effectiveness on multiple datasets. Experimental results are shown in Table 3, and we found that the detection ability of the model was further improved.

Table 3 Detection performance based on feature fusion

Quality	*Manipulation method*	*Accuracy*	*Single_best_acc*	*Improve*
c23	FaceSwap	91.09	89.28	1.81
c23	NeuralTextures	80.13	77.68	2.45
c40	Deepfakes	89.29	85.97	3.32
c40	Face2Face	78.32	75.09	3.23
c40	FaceSwap	82.11	79.08	3.03
c40	NeuralTextures	69.26	67.43	1.82

We also tested on DFDC dataset. As shown in Table 4, experimental results can prove that by adopting a variety of feature extractors, we obtain effective features, and the introduction of ensemble learning further improves the detection performance of the algorithm.

Table 4 Detection performance based on superior feature extractors

Face extractor	*Feature extractor*	*Loss*	*Accuracy*	*Ensemble accuracy*
RetinaFace	WSDAN-Xception	20.90	93.82	94.53
MTCNN	EfficientNet-b7	18.38	92.78	95.26

6 Conclusion

Although the development of deep learning brings convenience to people, the existence of Deepfake technology undoubtedly poses a serious threat to personal privacy and social stability. In this paper, based on the existing research, we propose the ensemble learning-based detection scheme to improve the algorithm performance. The high-quality forgery features are obtained through the existing feature extraction model, so as to confirm the detection effectiveness in the subsequent steps. The final test results are further obtained through feature selection, feature classification and combination strategy. The effectiveness of the scheme is verified on multiple datasets. It can be proved that ensemble learning is an effective way to improve the detection performance of existing algorithms.

References

Afchar, D., Nozick, V., Yamagishi, J. and Echizen, I. (2018) 'Mesonet: a compact facial video forgery detection network', *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, Hong Kong, pp.1–7.

Amerini, I., Galteri, L., Caldelli, R. and Del Bimbo, A. (2019) 'Deepfake video detection through optical flow based cnn', *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, Seoul, Korea (South), pp.1205–1207.

Bay, H., Tuytelaars, T. and Van Gool, L. (2006) 'Surf: Speeded up robust features', *European Conference on Computer Vision*, Springer, Graz, Austria, pp.404–417.

Bühlmann, P. (2012) 'Bagging, boosting and ensemble methods', *Handbook of computational statistics*, Springer, Humboldt University, Berlin, pp.985–1022.

Chollet, F. (2017) 'Xception: Deep learning with depthwise separable convolutions', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp.1251–1258.

Ciss, S. (2015) *Generalization Error and Out-of-bag Bounds in Random (Uniform) Forests*, ffhal01110524v2.

Cotta, P.V.P. (2018) *Random Forest Structure*, https://github.com/paulovpcotta/pokemon-udacity-final

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M. and Ferrer, C.C. (2020) *The Deepfake Detection Challenge (DFDC) Dataset*, arXiv preprint arXiv:2006.07397.

Džeroski, S. and Ženko, B. (2004) 'Is combining classifiers with stacking better than selecting the best one?', *Machine learning*, Vol. 54, No. 3, pp.255–273.

ELmoufidi, A. and Amoun, H. (2021) *EfficientNetB3 Architecture for Diabetic Retinopathy Assessment using Fundus Images*.

Fridrich, J. and Kodovsky, J. (2012) 'Rich models for steganalysis of digital images', *IEEE Transactions on Information Forensics and Security*, Vol. 7, No. 3, pp.868–882.

Güera, D. and Delp, E.J. (2018) 'Deepfake video detection using recurrent neural networks', *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, IEEE, Auckland, New Zealand, pp.1–6.

Hsu, C.C., Zhuang, Y.X. and Lee, C.Y. (2020) 'Deep fake image detection based on pairwise learning', *Applied Sciences*, Vol. 10, No. 1, p.370.

Koopman, M., Rodriguez, A.M. and Geradts, Z. (2018) 'Detection of deepfake video manipulation', *The 20th Irish Machine Vision and Image Processing Conference (IMVIP)*, Belfast, Northern Ireland, pp.133–136.

Liu, B. and Pun, C.M. (2018) 'Deep fusion network for splicing forgery localization', *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, Munich, Germany, pp.237–251.

Lukáš, J., Fridrich, J. and Goljan, M. (2006) 'Detecting digital image forgeries using sensor pattern noise', *Security, Steganography, and Watermarking of Multimedia Contents VIII*, Vol. 6072, International Society for Optics and Photonics, p.60720Y.

Matern, F., Riess, C. and Stamminger, M. (2019) 'Exploiting visual artifacts to expose deepfakes and face manipulations', *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, IEEE, Waikoloa, HI, USA, pp.83–92.

Mo, H., Chen, B. and Luo, W. (2018) 'Fake faces identification via convolutional neural network', *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security*, Innsbruck Austria, pp.43–47.

Nataraj, L., Mohammed, T.M., Manjunath, B., Chandrasekaran, S., Flenner, A., Bappy, J.H. and Roy-Chowdhury, A.K. (2019) 'Detecting gan generated fake images using co-occurrence matrices', *Electronic Imaging*, Vol. 2019, No. 5, pp.532–1.

Nguyen, H.H., Yamagishi, J. and Echizen, I. (2019) 'Capsule-forensics: Using capsule networks to detect forged images and videos', *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Brighton, UK, pp.2307–2311.

Paul, A., Mukherjee, D.P., Das, P., Gangopadhyay, A., Chintha, A.R. and Kundu, S. (2018) 'Improved random forest for classification', *IEEE Transactions on Image Processing*, Vol. 27, No. 8, pp.4012–4024.

Polikar, R. (2012) 'Ensemble learning', *Ensemble Machine Learning*, Springer, Boston, MA, pp.1–34.

Randhawa, K., Loo, C.K., Seera, M., Lim, C.P. and Nandi, A.K. (2018) 'Credit card fraud detection using adaboost and majority voting', *IEEE Access*, Vol. 6, 14277–14284.

Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I. and Natarajan, P. (2019) 'Recurrent convolutional strategies for face manipulation detection in videos', *Interfaces (GUI)*, Vol. 3, No. 1, pp.80–87.

Sabour, S., Frosst, N. and Hinton, G.E. (2017) *Dynamic Routing Between Capsules*, arXiv preprint arXiv:1710.09829.

Schapire, R.E. (2013) 'Explaining adaboost', *Empirical Inference*, Springer, Berlin, Heidelberg, pp.37–52.

Wu, Y., Liu, L., Xie, Z., Chow, K.H. and Wei, W. (2021) 'Boosting ensemble accuracy by revisiting ensemble diversity metrics', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.16469–16477.

Yang, X., Li, Y. and Lyu, S. (2019) 'Exposing deep fakes using inconsistent head poses', *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Brighton, UK, pp.8261–8265.

Yang, Y., Liu, X., Deng, R.H. and Li, Y. (2020) 'Lightweight sharable and traceable secure mobile health system', *IEEE Transactions on Dependable and Secure Computing*, Vol. 17, No. 1, pp.78–91.

Yang, Y., Zheng, X., Liu, X., Zhong, S. and Chang, V. (2018) ' 'Cross-domain dynamic anonymous authenticated group key management with symptom-matching for e-health social system', *Future Generation Computer Systems*, Vol. 84, pp.160–176.

Yang, Z., Wang, Z., Luo, L., Gan, H. and Zhang, T. (2021) 'SWS-DAN: Subtler WS-DAN for fine-grained image classification', *Journal of Visual Communication and Image Representation*, Vol. 79, p.103245.

Zhang, P., Zou, F., Wu, Z., Dai, N., Mark, S., Fu, M., Zhao, J. and Li, K. (2019a) 'Feathernets: convolutional neural networks as light as feather for face anti-spoofing', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Long Beach, CA, USA, pp.1574–1583.

Zhang, X., Karaman, S. and Chang, S.F. (2019b) 'Detecting and simulating artifacts in gan fake images', *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, Delft, Netherlands, pp.1–6.

Zhang, Y., Zheng, L. and Thing, V.L. (2017) 'Automated face swapping and its detection', *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*, IEEE, Singapore, pp.15–19.