qrAs per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology
Volume 10, Issue 02, pages 508-519.
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol10issue02.html
DOI: https://doi.org/10.46501/IJMTST1002070

Unmasking Deep Fakes Using Neural Network

Tejaswi.K[1] | PushpaLatha.M[2] | Raziya Sulthana.SK[3]

[1]Department of Information Technology, KKR&KSR Institute of Technology and Sciences , Vinjanamdu, Guntur, Andhra Pradesh, India.
tejaswisuresh09@gmail.com
[2]Department of Computer Science and Engineering, Chalapathi Institute of Engineering and Technology, Lam, Guntur, Andhra Pradesh, India.
[3]Department of Computer Science and Engineering, Chalapathi Institute of Engineering and Technology, Lam, Guntur, Andhra Pradesh, India.

To Cite this Article
Tejaswi.K, PushpaLatha.M, Raziya Sulthana.SK.A, Unmasking Deep Fakes Using Neural Network, International Journal for Modern Trends in Science and Technology, 2024, 10(02), pages. 58-519.https://doi.org/10.46501/IJMTST1002070

Article Info
Received: 28 January 2024; Accepted: 19 February 2024; Published: 25 February 2024.

ABSTRACT

In many real-world applications, deep generative models have recently produced impressive results, producing diverse and high-resolution samples from large and complicated data sets. As a result of this advancement, there is an immediate need for automated methods to identify these artificial intelligence (AI)-generated fake images because fake digital contents have multiplied, causing increased anxiety and mistrust in image content. When examined more closely, several face editing algorithms do show artifacts in some areas that are frequently invisible to the unaided eye, even if they appear to produce genuine human faces. Here, we describe a straightforward method for identifying these so-called DeepFakes, or phony face photos. Our approach starts with a traditional frequency domain analysis and moves on to a simple classifier. Our approach obtained good accuracies in fully unsupervised scenarios and demonstrated extremely strong results using only a few annotated training samples, in contrast to earlier systems that require enormous volumes of labeled data to be fed into the system. We created a new benchmark, Faces-HQ, for the evaluation of high resolution face photos by combining multiple public data sets, including both actual and synthetic faces. Using only 20 annotated samples for training, our method achieves a 100% classification accuracy given such high-resolution images. In a follow-up study, our approach obtains 100% supervised accuracy and 96% unsupervised accuracy while evaluating the medium-resolution photos of the CelebA data set.

Keywords: GAN images, DeepFake, Image forensic, Forgery detection

1. INTRODUCTION

A massive amount of new digital object contents has emerged in recent years due to the growing complexity of smart phones and the expansion of social networks. With the wide spread usage of digital photographs, there has been an increase in methods for manipulating the contents of photos. These methods were tedious, time-consuming, and needed a high level of computer vision subject expertise, making them unattainable for the majority of users until recently. But those restrictions have rapidly disappeared, thanks to the latest developments in machine learning and the availability of vast mounts of training data. This has led to adramatic reduction in the time required to fabricate and manipulate digital content, enabling even novice users to alter content as they see fit.

Specifically,deep generative models have been widely applied recently to generate realistic-looking fakeimages. Deep neural networks, the foundation of these models ,may roughly represent the actual data distribution of a particular training set. As a result, onecan add variants and sample from the learnt distribution. Generative Adversarial Networks (GAN)[11] and Variational Autoencoders (VAE)[16] are two of the most popular and effective methods. Recently, state-of-the-art results have been pushed, with GAN techniquesin particular boosting the quality and resolution of the images produced [4,14, 15]. Deepgenerative models are thereby creating a new avenue for AI-based fakepicture synthesis, which will hasten the spread of high-quality manipulated image content. Even though image forgery detection has advanced significantly, it is still a challenging task because most existing systems rely on deeplearning approaches that need a lot of labeled training data. In this study, we especially address the topic of false face detection among these manufactured image components. We provide a new machine learning based method to identify the nature of these images. Our methodology is based on a traditional frequency analysis of the pictures, which shows distinct behaviors at high frequencies.

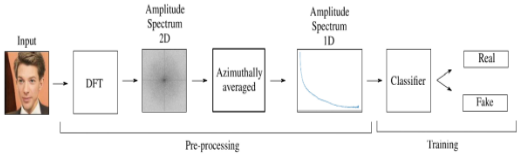


Fig. 2: Overview of the processing pipeline of our approach. It contains two main blocks, a feature extraction block using DFT and a training block, where a classifier uses the new transformed features to determine whether the face is real or not. Notice that input images are transformed to grey-scale before DFT.

Our approach uses a basic supervised or unsupervised classifier after frequency domain analysis to identify such artifacts. You'll see that our proposed pipeline doesn't require or include large amounts of data, which is a highly useful feature for circumstances when data is scarce. For our experimental evaluation, we also offer a new data set, Faces-HQ, which we used to supplement the CelebA and FaceForensics++datasets[25].

Generally speaking, our contributions are summarized as follows: We describe a novel artificial face identification classification pipeline based on frequency domain analysis. We provide a high-quality image collection called Faces-HQ, which consists of both real and fictional faces, gathered from various public sources. Extensive trials on high and medium-resolution photos of the Faces- HQ and CelebA data sets showed 100% accuracy, demonstrating how we successfully learn to detect forgeries. Furthermore, 91% accuracy was attained in the evaluation of the FaceForensics++ data set using low-resolution movies.

RELATED WORK

we concentrate on identifying manipulated images produced by GAN-based techniques. Conventional image forensics techniques, such as local noise estimation [23], pattern analysis [10], illumination modification [9], and steganalysis feature classification [7], can be categorized based on the picture features they aim to analyze. But since the discovery of deep learning, the field of computer vision has drastically shifted toward neural network methods. A couple of recent works that use convolutional neural networks (CNN) as its foundation are [8], [28]. These CNN-based methods likewise seek to implicitly capture the previously described picture properties.

An important development in generative models was the introduction of an adversarial framework (GAN) by Goodfellow et al. in 2014. Specifically, there has been a notable advancement in picture generation, which has resulted in notable advancements in artificial faces, among other areas [5]. As a result, throughout the past several months, new picture and video alteration techniques known as DeepFake have surfaced and become well-known online. The pursuit of identifying GAN-generated images or movies has garnered significant attention in the field of digital image forensics.

The absence of blinking [18] is one issue that arises when videos are produced artificially. The lack of training photos, particularly ones in which the person has their eyes closed, is the cause of this. However, this detection can be evaded by include closed-eye photos in the training set. Finding unusual head positions is another extended technique [26] for identifying manipulated digital content. Conversely, the studies [17], [22] examine the color-space characteristics of real and GAN-generated images, then utilize the difference to categorize the images.Some methods [2], [21], [27] use CNNs to discern GAN's output from actual photos instead than relying on overt deficiencies or failures. Similarly, [12] applies Recurrent Neural Networks (RNNs) upon CNNs to add temporal domain information and [13] presents a deep forgery discriminator with a contrastive loss function. In particular, the generator may be modified to learn a countermeasure for any differentiable forensic by integrating them into the discriminator of the GAN.

2. LITERATURE REVIEW

[1] Advances in AI, machine learning, and deep learning have resulted in the creation of methods and resources for altering multimedia in recent years. Although the main uses of these technologies have been in the fields of entertainment and education, unscrupulous people have also taken advantage of them to produce Deepfakes, which are high-quality fake images, audio files, and movies. Various ways have been detailed in the literature to solve this issue, and in this work, a systematic literature review is carried out to synthesize and analyse 112 pertinent papers from 2018 to 2020. The review classifies the methods into four categories: deep learning, statistical, blockchain, and conventional machine learning. It concludes that deep learning-based methods are more effective than other ways in identifying Deepfakes.

[2] According to the text, deep learning is an effective method that is applied in many different domains, including computer vision and natural language processing. Deep learning technology is also used in deepfakes, which are altered photos and videos that are identical to real ones. In order to help researchers comprehend and compare the most recent methods and datasets in this field, the paper offers a thorough analysis of deepfakes production and detection techniques employing deep learning algorithms.

Because of their advanced manipulation tactics, deepfake videos—which use deep learning technology to replace a person's face, emotion, or speech with someone else—pose a severe threat. This study offers a unique neural network-based method that uses key video frame extraction to detect fake films with high accuracy and low computing requirements. Recognizing deepfake videos on social media is important to limit their misleading impact. Even with little training data, the suggested model—which combines a convolutional neural network (CNN) and a classifier network—achieves remarkable performance in identifying extremely compressed deepfake films.

[4] The problem of video counterfeiting has grown significantly in importance, especially on social media platforms, with the development of deepfake videos. In order to detect false movies, this research presents a neural network-based method that combines a classifier network and a convolutional neural network (CNN). After comparing several CNN architectures, the study determines that XceptionNet is the best model. The suggested classifier is then combined with this model for classification. The system can identify compressed movies in social media and uses the FaceForensics++ dataset.

[5] The paper provides a thorough analysis of deepfake detection through the use of deep learning techniques, emphasizing the growing threat that deepfake technology poses and the requirement for efficient detection techniques. The abstract emphasizes the significance of creating tools to distinguish between reality and false information, as well as the effects of deep learning across a range of fields and the possible concerns related to deepfake technology. In order to provide a better understanding of deepfake generation, identification, latest developments, weaknesses of existing security methods, and areas requiring further investigation, the study categorizes deepfake detection methods based on their applications, including video detection, image detection, audio detection, and hybrid multimedia detection. According to the findings, the most widely used deep learning technique in papers for video deepfake detection and accuracy parameter enhancement is Conventional Neural Networks (CNN) methodology.

[6] Deep learning has made incredible progress, resulting in the creation of highly realistic AI- generated videos

called deepfakes. Deepfakes use generative models to manipulate facial features and create altered identities or expressions that look incredibly real. These synthetic media creations can be used to deceive or harm individuals and pose a threat to our legal, political, and social systems. To tackle this issue, researchers are actively working on detecting deepfake content to protect privacy and combat the spread of manipulated media. This article provides a comprehensive study on the methods used to create deepfake images and videos for face and expression replacement. It also discusses publicly available datasets that can be used to benchmark and evaluate deepfake detection systems. The study explores various detection approaches and highlights the challenges involved in identifying deepfake face and expression swaps. Additionally, it outlines future research directions to further enhance deepfake detection methods. The goal is to develop robust and effective solutions that can safeguard the authenticity and trustworthiness of visual media.

[7] This research paper explores the creation and detection of audio deepfakes. The first section provides an overview of deepfakes in general. The second section focuses on the specific methods used for audio deepfakes and compares them. The results discuss various techniques for detecting audio deepfakes, including analysing statistical properties, examining media consistency, and utilizing machine learning and deep learning algorithms. Some of the methods used for detection include Support Vector Machines (SVMs), Decision Trees (DTs), Convolutional Neural Networks (CNNs), Siamese CNNs, Deep Neural Networks (DNNs), and a combination of CNNs and Recurrent Neural Networks (RNNs). The accuracy of these methods varied, with SVM achieving the highest accuracy of 99% and DT achieving the

lowest at 73.33%. The Equal Error Rate (EER) and t-DCF were also reported in some studies, with different methods performing best in different scenarios.

[8] The rise of deepfakes has indeed made the authentication of digital media a critical need in our society. With the advancements in Generative Adversarial Networks (GANs), it has become increasingly challenging to identify synthetic media. Deepfakes, which are synthetic videos that manipulate faces and voices, pose a significant threat to trust and privacy in digital content. They can be misused for political gain, defamation, and tarnishing the reputation of public figures. People often struggle to distinguish between authentic and manipulated images and videos, highlighting the importance of automated systems that can accurately classify the validity of digital content. While many deepfake detection methods focus on spatial information in single frames, there are promising approaches that also consider temporal inconsistencies in manipulated videos. In our research, we propose a hybrid deep learning approach that combines spatial, spectral, and temporal content to differentiate real and fake videos. By leveraging the Discrete Cosine transform, we can capture spectral features of individual frames, improving deepfake detection. Our multimodal network explores new features and achieved a 61.95% accuracy on the Facebook Deepfake Detection Challenge (DFDC) dataset. It's exciting to see advancements in this field to combat the challenges posed by deepfakes.

[9] With the advancement of deep learning and computer vision technologies, ever- expanding methods have been made possible for anyone to produce fake yet remarkably lifelike images and films. These technologies are referred to as deepfakemethodology.Withdeepfake, face change in images and videos may be done with a high degree of realism, inventiveness. Deepfake recordings have been frequently shared online, with most of them aimed against politicians or well-known individuals. However, other approaches have been described in the literature to address the problems raised by deepfake. In this study, we conduct a review by examining and contrasting two main areas of research: (1) significant advancements in deepfake models; and (2) commonly utilized deepfake tools. Additionally, we have created two distinct taxonomies for deepfake tools and models. The underlying algorithms, datasets that these models and tools have used, and accuracy of these models and tools are also compared. Numerous difficulties and unresolved problems have also been noted.

[10] For a variety of uses, from social networking to border security, the capacity to authenticate a person's face in photos and videos might be crucial. Changing one's looks to resemble a target identity is a direct biometric attack tactic against the security of facialrecognition systems. The ability to identify such

attacks as distinct identities from their target is necessary for defense against them. On the other hand, this might be seen as a digital media fabrication from a forensics standpoint. Being able to identify items that are uncommon in authentic media is necessary for identifying such frauds. In this paper, two scenarios where faces in digital media can be classified as real or phony are examined from the viewpoints of the attacker and the defense . Firstly, we will investigate the defender's function by looking at how authentic videos may be separated from deepfakes. Videos in which one person's face has been replaced with another's are the most prevalent type of deepfakes; these are frequently called "face-swaps". Second, by looking at a problem that is becoming more and more important to border security, we will investigate the attacker's involvement. The technique of combining two or more people's faces into one image is known as "face morphing."

[11] Recently, deep generative models have shown impressive results for numerous real- world applications, producing diverse and high-resolution samples from intricate data sets.The proliferation of fraudulent digital materials as a result of this advancement has raised concerns and stoked public mistrust of image content, making the need for automated methods to identify these artificial intelligence (AI)-generated fake images imperative.Even though many face-editing algorithms appear to create realistic human faces, closer inspection reveals that they actually include artifacts in several areas that are frequently invisible to the untrained sight. In this study, we describe a straightforward method for identifying so-called DeepFakes, or phony facial photographs.Our approach is predicated on a simple classifier that comes after a traditional frequency domain study. Our technique demonstrated very strong results using only a few annotated training samples, and even produced good accuracies in entirely unsupervised circumstances, in contrast to earlier systems that require enormous amounts of labeled data to be fed into the system. We created a new benchmarkcalled Faces-HQ by combining many public data sets of genuine and synthetic faces for the evaluation of high- resolution face photos.

[12]Our method produced excellent results with only a few annotated training examples, in contrast to other systems that require enormous volumes of labeled data to be fed in. It even managed to obtain good accuracies in Over the past few decades, there has been a rapid advancement in artificial intelligence, machine learning, and deep learning, leading to the development of new techniques and tools for manipulating multimedia.A face-swapping method called "deepfakes" enables anyone to swap faces in a video with remarkably lifelike effects. Despite its usefulness, this tactic can have a significant negative impact on society if used maliciously, as in the case of disseminating false information or indulging incyberstalking.Asaresult,identifyingdeepfakesbecomes crucial.

[13]With just a few annotated training samples, our method produced extremely strong results and even achieved good accuracies in comparison to earlier systems that require enormous volumes of labeled data to be fed into them. With the rapid advancement of artificial intelligence, machine learning.In 2020, 78% of Canadian firms had at least one successful hack, according to the 2020 Cyber Threat Defense Report [1]. Such attacks can have a range of negative effects, from privacy violations to significant financial losses for people, businesses, and nations. Experts estimate that by 2025, the annual worldwide loss resulting from cybercrime will amount to 10.5 trillion US dollars [2]. It is more important than ever to anticipate and prevent cyberattacks in light of these concerning data. Artificial Intelligence (AI) and Machine Learning (ML)-based solutions are becoming more and more necessary for our essential infrastructure to provide timely services at scale [3]. There are significant worries concerning the security and safety of machine learning (ML) systems due to our growing reliance on them. Serious ethical concerns were raised, particularly with the advent of potent machine learning algorithms that can be used to create phony visual, textual, or auditory content that has a high potential to trick people.

[14]High-quality face photos can be produced by today's image generating technologies, and it might be challenging for humans to verify the authenticity of these photographs. The goal of this work is to leverage the benefits of deep learning technologies to enhance the detection of face swapping forgeries, or deepfake detection. In order to address the issue of subpar detection performance on cross-data sets, this work uses spatial enhancement technologies to create a unified and improved data collection from several sources. The novel

deepfake detection architecture, which consists of 20 network layers, is suggested as the deepfake detection model by utilizing the benefits of Inception and ResNet networks. Hyper parameter variables are improved in order to enhance the suggested model even further. The experiment's outcome demonstrates that, in terms of accuracy, loss value, AUC, number of parameters, and FLOPs, the suggested network outperformed popular techniques including ResNeXt50, ResNet101, XceptionNet, and VGG19. All things considered, the techniques presented in this work can aid in broadening the data set, improving the detection of deepfake contents, and successfully optimizing network models.

[15]A deepfake refers to content that is created by artificial intelligence and appears authentic to humans. It is primarily generated using artificial neural networks, a branch of machine learning, and commonly involves manipulating and generating human imagery. While deepfakes have creative applications like realistic video dubbing and historical figure reanimation, they are also associated with unethical and malicious uses, such as spreading misinformation and impersonating individuals. To understand where the threats are moving and how to mitigate them, we need a clear view of the technology's, challenges, limitations, capabilities, and trajectory. Unfortunately, to the best of our knowledge, there are no other works which present the techniques, advancements, and challenges, in a technical and encompassing way

[16]In this paper,the authors talk about the developments in Deep fake technology, which modifies multimedia material realistically and has advantages as well as disadvantages. It focuses on classifying creation methods, evaluating data sets, and employing deep neural networks to detect Deep fakes. The conclusion emphasizes the necessity for continuous progress by highlighting the present shortcomings in detecting techniques. The report notes that corporations are making an attempt to address these issues, but it emphasizes how critical it is to improve data integrity and put in place extra security measures. It also predicts an increase in AI-driven Deep fake propaganda in the future, highlighting the necessity of ongoing technological developments to stay ahead in this changing environment. The review's overall goal is to improve and streamline comprehension of Deep fake detection in facial picture and video applications.

[17] In this paper, the authors investigate how adversarial attacks can affect deep neural networks, specifically with regard to the creation of misleading images that could lead classifiers astray. It draws attention to the difficulties in developing reliable and efficient techniques for producing these adversarial cases. Inspired by adversarial examples, the abstract presents two new generating models: CGAN-F and CGAN-Adv. Through the use of conditional generative adversarial networks and a novel training approach, these models seek to directly generate adaptive attack instances. The paper concludes by highlighting the effectiveness of the suggested techniques in producing assault images, demonstrating enhanced resilience and decreased production expenses in contrast to conventional approaches such as the Fast Gradient Sign Method. All things considered, the research advances our knowledge of how to create strong adversarial instances to target deep neural network classifiers.

[18] In this paper, the authors looks at how multimedia content manipulation is changing and emphasizes how quickly realistic false photos and movies are being created. It highlights how this development is dual in nature, offering serious security risks in addition to interesting possibilities in the creative arts, particularly with the emergence of deepfakes. The abstract emphasizes how critical it is to develop automated methods for identifying fraudulent multimedia information, particularly in order to stop potential criminal activity and public opinion manipulation. The conclusion considers how artificial intelligence has shaped multimedia forensics during the last fifteen years. It highlights the continuous arms race between the development of forensic tools and realistic fakes, and it makes a case for more study to maintain information integrity in the face of emerging challenges.

[19] In this paper, the authors address the problem of deepfake videos with cutting-edge artificial intelligence algorithms. The abstract presents an innovative technique that uses optical flow fields to analyse temporal variations in order to differentiate between authentic and fraudulent videos. Early testing on the Face Forensics++ dataset demonstrate that our method performs better than existing algorithms that concentrate on individual frames.

The conclusion highlights the creative application of temporal cues and makes recommendations for further study to evaluate the method's consistency over a range of datasets and investigate possible synergies with well-established frame-based approaches

[20]In this paper, the authors with the recent developments in media generation, focuses on the identification of faked images and videos. A suggested technique detects several kinds of spoofs, like replay attacks and computer-generated films, by using capsule networks, which are commonly used for computer vision. Tests indicate that the technique works well for detecting a variety of falsified content. The use of random noise during training is introduced in the study, and it works well. Future work in the research community aims to address mixed attacks and anomalies, enhance robustness, and evaluate resilience against adversarial attacks

3. METHODS

A. Frequency Domain Analysis

Frequency domain analysis is of utmost importance in signal processing theory and applications. In particular in the computer vision domain, the repetitive nature or the frequency

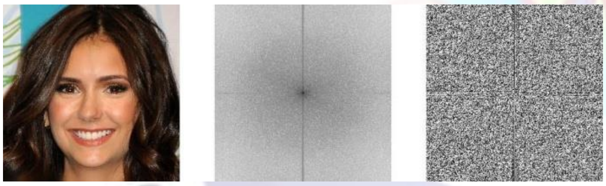


Fig. 3: Example of a DFT applied to a sample. (Left) Inputimage[1].(Center)PowerSpectrum.(Right)PhaseSpectrum.

characteristics of images can be analyzed on a space defined by Fourier transform. Such transformation consists in a spectral decomposition of the input data indicating how the signal's energy is distributed over a range of frequencies. Methods based on frequency domain analysis have shown wide applications in image processing, such as image analysis, image filtering, image reconstruction and image compression.

Discrete Fourier Transform: The Discrete Fourier Transform (DFT) is a mathematical technique to decompose a dis- crete signal into sinusoidal components of various frequencies ranging from 0 (i.e., constant frequency, corresponding to the image mean value) up to the maximum representable fre- quency, given the spatial resolution. It is the discrete analogonofthecontinuousFourierTransformforsignalssampledonequidistantpoints.For2-dimensionaldataofsize$M×N$,itcanbecomputedas

$$X_{k,A} = \sum_{n=0}^{N 1}\sum_{0}^{M 1} x_{n,m} \cdot e^{-\frac{i2\pi kn}{N}} \cdot e^{-\frac{i2\pi Am}{M}}. \qquad (1)$$

The frequency-domain representation of a signal (*Xk*) carries information about the signal's amplitude and phase at each frequency. Fig. 3 depicts the complex output information (power and phase). Notice that the amplitude spectrum is the square root power spectrum

1) *Azimuthal Average:* After applying a Fourier Transform to a sample image, the information is represented in a new domain but within the same dimensionality. Therefore, given that we work with images, the output still contains 2D information. We apply azimuthal averaging to compute a robust 1D representation of the FFT power spectrum. It can be seen as a compression, gathering and averaging similar

2) A.ClassifierAlgorithms

Classificationisthetasktolearnageneralmappingfromtheattributespacetodescreteclasses using specificexamples

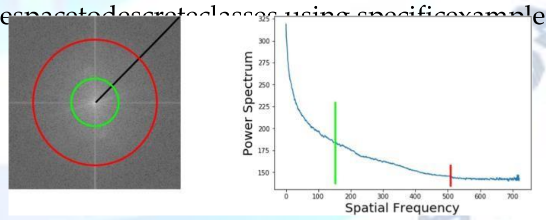


Fig.4: Example of anazimuthal average. (Left) Power Spectrum 2D. (Right) Power Spectrum 1D.

Each frequency component is the radial average from the 2Dspectrum of instances, each represented by a vector of attribute values and their acordinglable.

1)Logistic Regression: One of the technically simplest(linear) classification algorithms is the Logistic Regression (LR). It is a simple statistical model that employs a logistic function (seeFormula (4)) to model a binary dependent variable. The output from the hypothesis h is the estimated probability. This is used to infer how confident predicted value can be given an inputx.Logistic regression is formulated as

$$h_{\mathbf{w}}(\mathbf{x})=\frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}} \qquad (2)$$

The underlying algorithm of maximum likelihood estimation

determines the regression coefficient w for the binary dependent variable. The algorithms tops when the convergence criterion is metor the maximum number of iterations is reached.

2)Support Vector Machines:Support Vector Machines (SVMs)[3],[6] are among the most widely used learning algorithms for (non-linear)data classification.The target of theSVM formulation is to produce a model (based on the trainingdata) which will identify an optimal separating hyperplane,maximizing the margin between different classes.Given a training set of instance-label pairs (xi, yi),i= 1, ..., l where xi∈Rnandy∈{1,−1}l, Training of SVMs is implemented by the solution of the following optimization problem.

$$\min_{w,b,\xi} \frac{1}{2}||\mathbf{w}||^2+C\sum_{i=1}^{l}\xi_i \qquad (3)$$
$$\text{s.t.} \quad y_i(\mathbf{w}^T\varphi(x_i)+b)\geq 1-\xi_i,\xi_i\geq 0,$$

where w and b are the parameters of our classifier, ξ is the slack variable and C >0 the penalty parameter of the error term. Here training vectors xi are mapped into a higher dimensional space by the function φ. The training objective of SVMs is to find alinear separating hyperplane with the maximal margin in this higher dimensional space.

3)K-Means Clustering:While supervised classification algorithms like SVM and LR rely on labeled training example to learn a classification, we also want to test the detection performance inthe absence of any labeled data.Clustering is an unsupervised machine learning technique which finds similarities in the data points and group similar data points together.The key assumption is that nearby points in the features ace exhibit similar qualities and they can be clustered together.Clustering can be done using different techniques likeK-means clustering.

$$J=\sum_{k=1}^{K}\sum_{i=1}^{m}||x_i-\mu_k||^2 \qquad (4)$$

where K and m are the numberof clusters and samples respectively.

A common approach to heuristically approximate solutions is to iteratively identify nearby features based on the distancescalculated from initial centroids µ. Then, these features re assigned to the closest cluster and the centroids are re-estimated. Since the amount of clusters is determined by theuser, it canbe easily employed in classification where we divide data into Kclusters with Kequal to or greater than the number of classes.

4. PROPOSED METHOD

In order to verify our approach,we also evaluate on the CelebA data set [19], which contains medium-resolution images, and on the Face Forensics++ data set [25],which contain slow-re solution video sequences.

A. Faces-HQ

1) Dataset: to the best of our knowledge, currently no public dataset is providing high resolution images with annotated fake and real faces. Therefore, we have created ourown data set from established sources, called Faces-HQ2. Inorder to have a sufficient variety of faces, we have chosen todownload and label theimages available from the CelebA-HQ dataset [14], Flickr-Faces-HQ dataset [15],100K Faces project [1] .In total, we have collected 40K high quality images,half of them real and the other half fake faces.TableI contains a summary.

	#ofsamples	category	label
CelebA-HQdataset[14]	10000	Fake	0
Flickr-Faces-HQdataset[15]	10000	Fake	0
100KFacesproject[1]	10000	Real	1
www.thispersondoesnotexist.com	10000	Real	1

TABLEI:*Faces-HQ*datasetstructure.

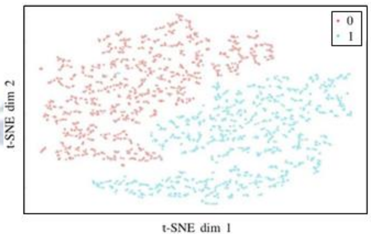


Fig.5: T-SNE visualization of 1D Power Spectrum on a random subset from Faces-HQ data set. We used a perplexity of 4 and 4000 iterations to produce the plot

On the onehand, atpre-processing time,we take the whole dataset and we transform every sample from the spatial domain to the1D frequency do-main, reducing1024x1024x3 high quality color images to722 features(1DPowerSpectrum).This method is formed by a Discrete Fourier Transform followed by an azimuthally average. The transformation can be substantially optimized by employing the Fast Fourier Transform. Notice that after applying the transformation, we use only the power spectrum since it already contains enough information for the classifier.A first visualization (see Fig. 5) using t-Distributed Stochastic Neighbor Embedding [20](t-SNE) reveals a clear clustering off a kean real samples in this feature space.

On the other hand, once the pre-processing step is finished, we start training the classifierengine. First of all, we divide the transformed data into training and testing sets, with20% for the testing stage and use the remaining 80% as thetraining set. Then, we train a classifier with the training data and finally evaluate the accuracy on the testing set. Our goal is to distinguish, real and fake faces, thus we need to use a binary classifier.

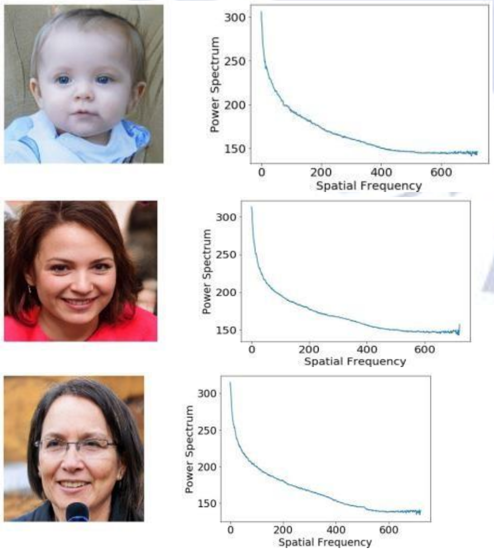


Fig a

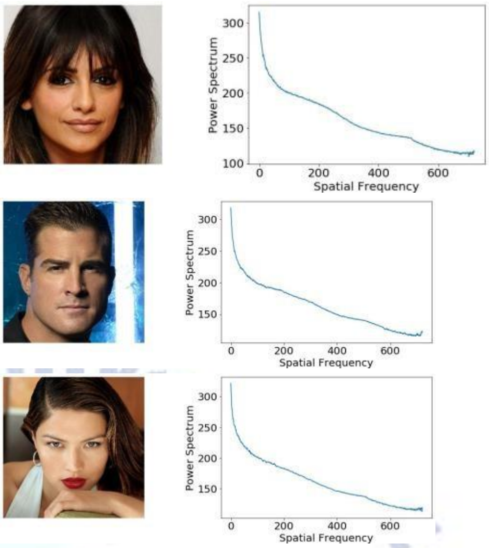


Fig b celebA HQ-dataset

Fig. 6: Samples from the different data sets gathered on Faces-HQ data set. It is possible to observe on the 1D Power Spectrum some similitudes between images belonging to the same class and differences otherwise

For instance, real faces (b) and (d) do not have flat regions at high frequencies, whereas fake (a) and (b) have them.

samples	0% (train) - 20% (test)		
	SVM	Logistic Reg.	K-Means
4000	100%	100%	82%
1000	100%	100%	82%
100	100%	100%	81%
20	100%	100%	75%

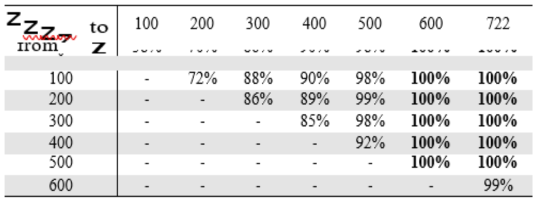
TABLE II: Test accuracy using SVM, logistic regression and k-means classifier under different data settings.

3)Method 1D Power Spectrum: looking at Fig. 6, one can observe that there is a certain repetitive behavior or pattern onthe 1D Power Spectrum on those images that belong to thesame class. Just by checking individual samples, it is possibleto conclude that real and fake images behave in notice able different spectra at high frequencies, and therefore they canbeeasily classified.Driven by this phenomenon,we have evaluated a significant subset of images (4000 in total, 1000 of each sub-data set) and we have computed basic statistics to try to find a more general representation that help to simplify the problem. Fig. 1 plots the mean and the standard deviation of each sub-data set and corroborates the observable and distinguishable trend that real and fake images have. Motivated by this observations we have carried out a set of tests to determine the extent to which our approach successfully detects deepfakes and how much data is needed to train the model. In our experiments, we have implemented one classifier based on support vector machines (SVMs) with a radial basis function kernel and one based on logistic regression. We have run an initial experiment using 80% of the data for training and 20% for testing. We have utilized this configuration for different amount of samples (4000, 1000, 100, 20) equally distributed (see Table II).

zzzz to from z	100	200	300	400	500	600	722
0	58%	69%	85%	89%	98%	**100%**	**100%**
100	-	72%	86%	89%	98%	**100%**	**100%**
200	-	-	85%	87%	99%	**100%**	**100%**
300	-	-	-	84%	98%	**100%**	**100%**
400	-	-	-	-	93%	**100%**	**100%**
500	-	-	-	-	-	**100%**	**100%**
600	-	-	-	-	-	-	100%

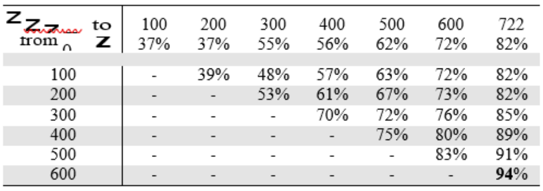
TABLE III: Test accuracy using SVM classifier.

After testing the effectiveness and efficiency of our trans- formed features, we have conducted a another round of experiments to determine the impact of different frequency components. Given the 722 features from 1D Power Spectrum, we have analyzed the relevance of different frequencies by grouping them into 28 sub-sections. Table III,Table IV and Table V show the accuracy results on SVM, logistic regression and K-means respectively. The rows indicate where the chunk of frequencies starts, and the column where it ends. For example, there is a chunk with 0.86 accuracy that contains frequencies from 100 to 300.



zzz to from z	100	200	300	400	500	600	722
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
100	-	72%	88%	90%	98%	**100%**	**100%**
200	-	-	86%	89%	99%	**100%**	**100%**
300	-	-	-	85%	98%	**100%**	**100%**
400	-	-	-	-	92%	**100%**	**100%**
500	-	-	-	-	-	**100%**	**100%**
600	-	-	-	-	-	-	99%

TABLE IV: Test accuracy using logistic regression classifier.



zzz to from 0 z	100 37%	200 37%	300 55%	400 56%	500 62%	600 72%	722 82%
100	-	39%	48%	57%	63%	72%	82%
200	-	-	53%	61%	67%	73%	82%
300	-	-	-	70%	72%	76%	85%
400	-	-	-	-	75%	80%	89%
500	-	-	-	-	-	83%	91%
600	-	-	-	-	-	-	**94%**

TABLE V: Test accuracy using k-means classifier.

A. CelebAData set: CelebFaces Attributes (CelebA) data set [19] consists of 202,599 celebrity face images with 40 variations in facial attributes. The dimensions of the face images are 178x218x3, which can be considered to be a medium resolution in our context.

Training setting:

In order to train our forgery detection classifier we need both real and fake images. We used the real images from the *CelebA* data set. On the same set, we then train a DCGAN [24] to generate realistic but fake images. We split the data set into 162,770 images for training and 39,829 for testing, and we crop and resize the initial 178x218x3 size images to 128x128x3. Once the model is trained, we can conduct the classification experiments on medium-resolution scale**.**

7. RESULTS & DISCUSSION

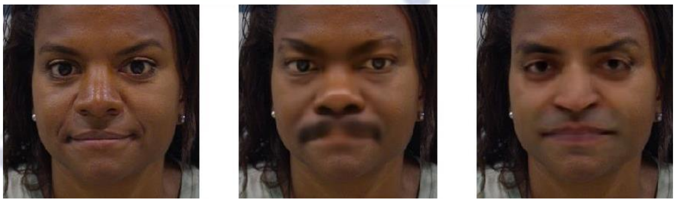
Results: We follow the same procedure as in the previous experiments. Table VI) shows perfect classification accuracy in the supervised, and also very good results in unsupervised clustering.

	80% (train) - 20% (test)		
samples	**SVM**	**Logistic Reg.**	**K-Means**
000	**00%**	**100%**	**96%**

TABLE VI: Test accuracy using SVM, logistic regression and k-means classifier.

FaceForensics++

Data set: FaceForensics++ [25] is a forensics data set consisting of video sequences that have been modified with different automated face manipulation methods. Additionally,



(a) Example of one real face (left) and two deepfake faces, fake 1 (center) and fake 2 (right). Notice that the modifications only affect the inner face.

it is hosting DeepFakeDetection Data set. In particular, this data set contains 363 original sequences from 28 paid actors in 16 different scenes as well as over 3000 manipulated videos using DeepFakes and their corresponding binary masks. All videos contain a trackable, mostly frontal face without occlusions which enables automated tampering methods to generate realistic forgeries.

Method 1D Power Spectrum: As in the previous experiments, Fig. 7 shows that deepfake images have a noticeably different frequency characteristic. Despite of having a similar behaviour along the spatial frequency, there is a clear offset between the real the fakes that allows the images to be classified.

# samples	80% (train) - 20% (test)	
	VM	Logistic Reg.
2000	5%	8%
1000	2%	6%
200	7%	3%
20	6%	6%

TABLE VII: Test accuracy using SVM classifier and logistic regression classifier under different data settings.

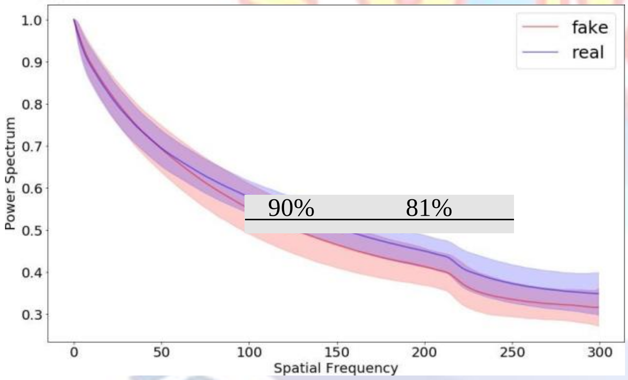
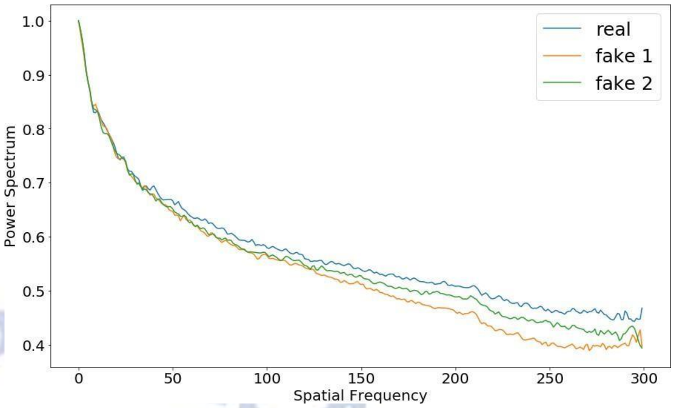


Fig. 8: 1D Power Spectrum statistics from DeepFakeDetectiondata set.

Table VII contains the classification accuracy for the supervised algorithms. These results confirm the robustness of frequency components as classification features. Nevertheless, in this case, we have observed a slightly different behaviour with respect to *Faces-HQ* accuracy results (see Table II). The problem is become harder for low-resolution inputs. Hence, the accuracy starts to decrease when the number of samples is smaller than 1000, specially, for the logistic regression.



The dependency on samples and the non-perfect classifica- tion accuracy can be understood by looking at Fig. 8. We can see how the standard deviations from the real and the deepfake statistics overlap with each other, meaning that some samples will be misclassified. As a result, it is not recommendable to reduce the number of features, since now the classifiers are much more sensitive to the number of features.

Finally, we compute the average classification rate per video, applying a simple majority vote over the single frame classifications. Table VIII shows the accuracy test results, which are relatively higher than the previous ones based on a frame by frame evaluation.

SVM	Logistic Reg.
90%	81%

TABLE VIII: Test accuracy per video using SVM classifier and logistic regression classifier

8. CONCLUSIONS

In this paper, we described and evaluated the efficacy of a new method to expose AI-generated fake faces images. Our approach is based on a high-frequency component analysis. We performed extensive experiments to demonstrate the ro- bustness of our pipeline independently of the source image. We show that our method is able to detect high- and medium-resolution deepfake images on two data sets with data from various GANs with 100% accuracy. Low-resolution content is harder to identify since the available frequency spectrum is much smaller. Nevertheless, we are able to identify low- resolution fakes in a popular benchmark with 91% accuracy.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] 100,000 faces generated. https://generated.photos/.

[2] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pages 1–7. IEEE, 2018.

[3] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory, pages 144–152. ACM, 1992.

[4] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096, 2018.

[5] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel,A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228, 2018.

[6] C. Cortes and V. Vapnik. Support-vector networks. Machine learning, 20(3):273–297, 1995.

[7] D. Cozzolino, D. Gragnaniello, and L. Verdoliva. Image forgery localization through the fusion of camera-based, feature-based and pixel- based techniques. In 2014 IEEE International Conference on Image Processing (ICIP), pages 5302–5306. IEEE, 2014.

[8] D. Cozzolino and L. Verdoliva. Noiseprint: a cnn-based camera model fingerprint. IEEE Transactions on Information Forensics and Security, 2019.

[9] T. J. De Carvalho, C. Riess, E. Angelopoulou, H. Pedrini, and A. de Rezende Rocha. Exposing digital image forgeries by illumination color classification. IEEE Transactions on Information Forensics and Security, 8(7):1182–1194, 2013.

[10] P. Ferrara, T. Bianchi, A. De Rosa, and A. Piva. Image forgery localization via fine-grained analysis of cfa artifacts. IEEE Transactions on Information Forensics and Security, 7(5):1566–1577, 2012.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley,S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014.

[12] D. Gu¨era and E. J. Delp. Deepfake video detection using recurrent neural networks. In 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 1–6. IEEE, 2018.

[13] C.-C. Hsu, C.-Y. Lee, and Y.-X. Zhuang. Learning to detect fake face images in the wild. In 2018 International Symposium on Computer, Consumer and Control (IS3C), pages 388–391. IEEE, 2018.

[14] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017.

[15] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4401–4410, 2019.

[16] D. P. Kingma and M. Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

[17] H. Li, B. Li, S. Tan, and J. Huang. Detection of deep network generated images using disparities in color components. arXiv preprint arXiv:1808.07276, 2018.

[18] Y. Li, M.-C. Chang, and S. Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. arXiv preprint arXiv:1806.02877, 2018.

[19] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision, pages 3730–3738, 2015

[20] L.v.d Maaten and G. Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605, 2008.

[21] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva. Detection of gan-generated fake images over social networks. In 2018 IEEE Con- ference on Multimedia Information Processing and Retrieval (MIPR),

[22] S.Mc Closkey and M. Albright. Detecting gan-generated imagery using color cues. arXiv preprint arXiv:1812.08247, 2018.

[23] X. Pan, X. Zhang, and S. Lyu. Exposing image splicing with inconsistent local noise variances. In 2012 IEEE International Conference on Computational Photography (ICCP), pages 1–10. IEEE, 2012.

[24] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015.

[25] A.Ro¨ssler,D.Cozzolino,L.Verdoliva,C.Riess, J.Thies, and M.Nießner.FaceForensics++: Learning to detect manipulated facial images. In International Conference on Computer Vision (ICCV), 2019

[26] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 8261–8265. IEEE, 2019.