International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 01 January 2025
Page No: 98-102
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0013

A Hybrid Approach to Deep Fake Detection Using Error Level Analysis

Mrs. Sushma D. S[1], *Sumanth T.C*[2], *Mehraj*[3], *Likhith.R*[4], *Lohith T. R*[5]

[1]*Assistant Professor, ISE Department, DBIT, Bangalore, Karnataka, India.*

[2,3,4,5]*UG -Information Science and Engineering, DBIT, Bangalore, Karnataka, India.*

Emails: *sushamads99@gmail.com*[1], *sumanthtc15@gmail.com*[2], *mohammadmehraj@gamil.com*[3], *liki.r003@gmail.com*[4], *lohithtr46@gmail.com*[5]

Abstract

The rapid advancement of 'deepfake' video technology— which uses deep learning artificial intelligence algorithms to create fake videos that look real—has given urgency to the question of how policymakers and technology companies should moderate inauthentic content. We conduct an experiment to measure people's alertness to and ability to detect a high-quality deepfake among a set of videos. First, we find that in a natural setting with no content warnings, individuals who are exposed to a deepfake video of neutral content are no more likely to detect anything out of the ordinary (32.9%) compared to a control group who viewed only authentic videos (34.1%). Second, we find that when individuals are given a warning that at least one video in a set of five is a deepfake, only 21.6% of respondents correctly identify the deepfake as the only inauthentic video, while the remainder erroneously select at least one genuine video as a deepfake. The process of recognizing and distinguishing between real content and content generated by deep learning algo rhythm's, often referred to as deepfakes, is known as deepfake detection. In order to counter the rising threat of deepfakes and maintain the integrity of digital media, research is now being done to create more reliable and precise detection techniques. Deep learning models, such as Stable Diffusion, have been able to generate more detailed and less blurry images in recent years. In this paper, we develop a deepfake detection technique to distinguish original and fake images generated by various Diffusion Models. The developed methodology for deepfake detection takes advantage of features from fine-tuned Vision Transformers (ViTs), combined with existing classifiers such as Support Vector Machines (SVM). We demonstrate the proposed methodology's ability of interpretability-through-prototypes by analysing support vectors of the SVMs

Keywords: *Deepfake technology, Deepfake detection, Deep learning algorithms, Vision Transformers (ViTs), Support Vector Machines (SVM)*

1. Introduction

In recent years, deep learning-based generative models, such as Variation Auto encoders (VAEs) and Generative Adversarial Networks (GANs), have made significant advancements in synthesizing realistic visual content, including partial or fully generated images and videos. Enhanced versions of GANs, such as Progressive Growing GANs (PGGAN) [1] and BigGAN [2], have been able to produce photo-realistic images that are almost indistinguishable from authentic content by human observers, especially under time constraints. These generative models are typically employed for image- to-image translation tasks [3], which, when misused, can cause severe ethical and societal challenges. For example, frameworks like CycleGAN can manipulate facial content for malicious purposes, such as creating non-consensual synthetic media in inappropriate contexts [4]. Similarly, GANs can fabricate speech videos, overlaying a synthesized face on existing footage of public figures, which can lead to misinformation and societal disruption. Given the alarming misuse of such technology, there is an urgent need to develop efficient and reliable detection techniques to identify deepfake content, particularly fake facial images. Traditional image forgery detection techniques fall into two categories: active and passive forensics methods. Active methods rely on embedding external signals, such as watermarks, into the original image, which are later retrieved to validate authenticity [6]. However, these techniques fail in the context of GAN-generated images, as there is no original source image available for comparison. On the other hand, passive forensics approaches utilize intrinsic statistical properties to detect tampering or inconsistencies [7,8]. Unfortunately, these methods also struggle with deepfake detection because GAN-generated images are created from random latent vectors rather than being modified versions of pre-existing content. To address these limitations, researchers have explored deep learning-based approaches for fake image detection, framing the problem as a binary classification task (i.e., real vs. fake) [9,10]. For instance, convolutional neural networks (CNNs) have been employed to identify subtle artifacts in deepfake images, and advanced architectures like the Xception network [12] have achieved improved performance in fake face detection tasks [11]. Hybrid ensemble approaches [13] have also been proposed to enhance detection capabilities. However, these studies predominantly focus on partially manipulated images (e.g., face swaps) rather than fully generated synthetic images. Consequently, their models lack the generalization ability to detect entirely fake content produced by various GANs. As multiple advanced GAN models [1–3,14–18] continue to emerge, training deepfake detection systems on all possible GAN outputs becomes computationally expensive and impractical. The dependency on supervised learning strategies [9–11] further limits these methods, as they fail to generalize well to deepfake content generated by GANs that were not included in the training phase. This paper addresses these challenges by introducing a novel Common Fake Feature Network (CFFN) that leverages a pairwise learning approach to improve generalization performance. The proposed architecture enhances feature extraction and learning for fake image detection, particularly for fully synthesized images.

Main Contributions [1-5]

- We propose a novel CFFN-based fake face image detector, which incorporates an enhanced DenseNet backbone with a Siamese network structure to detect fully synthesized fake images effectively.
- The proposed method extracts and analyzes cross-layer features, which significantly improve the performance of fake image detection by capturing finer details of manipulated regions.
- To address generalization issues, we adopt a pairwise learning approach that improves the robustness of the detection system against GANs not included in the training dataset.

The rest of this paper is organized as follows: Section 2 describes the architecture and functioning of the proposed CFFN for fake image detection. Section 3 explains the integration of the pairwise learning approach for enhanced generalization. Experimental results, including performance evaluations on benchmark datasets, are presented in Section 4. Finally, Section 5 concludes the paper and outlines potential directions for future research.

2. Fake Face Image Detection

The detection of fake face images remains one of the most critical challenges in the field of image and video forgery detection. Such synthetic facial images can be maliciously exploited to generate fake identities on social media platforms, leading to illegal activities such as identity theft and unauthorized access to personal information. For example, generative models can synthesize inappropriate or compromising images of public figures, causing significant ethical, social, and legal repercussions. To address these challenges, this section introduces a novel deep learning framework based on a pairwise learning strategy. The proposed method adopts a two-step learning approach, as illustrated in Figure 1, which combines the Common Fake Feature (CFF) extraction with classifier learning. The use of supervised learning in fake face image detection effectively mitigates two key issues: the difficulty of collecting diverse training samples generated by all existing GAN architectures and the requirement to retrain the detection model each time a new GAN-based generator is introduced. The proposed method overcomes these limitations by utilizing pairwise learning, where real and fake image pairs are constructed to derive contrastive loss. This loss function enables the model to learn the discriminative common fake features (CFF) that are shared across different types of fake face images. Once the CFF is extracted through the Common Fake Feature Network (CFFN), the subsequent classifier network uses these features to effectively distinguish between real and fake images. By leveraging this strategy, the framework enhances its generalization capability, allowing it to identify fake face images, even those generated by unseen GANs, without the need for retraining.

3. Deep Fake Generation

With the development of deep learning models and algorithms, deep fake generating techniques have advanced quickly. By training discriminator and generator networks in an adversarial way, GANs [15] have demonstrated an outstanding ability in synthesizing very realistic and convincing deep fakes. During training, the discriminator network tries to separate fake material from original content while the generator network generates fake content. Several GAN-based techniques [3, 26, 14, 41] have been applied to generate high-quality Deep Fake images, specifically in the domain of Deep Fake Face generation. Moreover, Face Forensics++ [32] is an open dataset that contains images generated by GAN, and it has been used as a benchmark dataset for Deep Fake Detection Challenge. Utilizing Variation Auto encoders (VAE) [22] is yet another noteworthy method for deep fake generation. By mapping random noise to the learnt latent space, a Variation Auto encoder (VAE) learns a compressed representation of the data distribution and enables controlled generation of fake samples. Various modifications of VAE [34, 38, 40] have been developed in recent years and have shown great performance in creating Deep Fake images on open datasets, such as CIFAR-10 [24], MNIST [10], FFHQ and ImageNet. Additionally, Celeb-DF [25] and DeeperForensics-1.0 [20] have been extensively used for benchmarking of deep fake detection models, where images are generated with Auto encoder models in those datasets. The generation of realistic fake samples with sharper features and less blurriness is made possible by diffusion models [18], which involve repeatedly updating an initial noise distribution to match the desired data distribution. Diffusion models are able to generate more highly detailed images than GANs [11] and VAEs [19], which make the deepfake detection task hypothetically more complex. However, there exist very few open datasets [4] for benchmarking of deep fake Detection on Diffusion Model generated images.

4. Deep Fake Detection

To classify deep fake images, some methods use Convolutional Neural Networks (CNN) [23, 5]. The work [6] utilizes optical flow fields in order to exploit inter-frame correlations. Target-specific region extraction layer for the CNN architecture [36] has also been used to feed only the most important information to the model. The proposed model [2] is based on feature extraction from the trained CNN model and XGBoost for classification. The composite method [29], which consists of state-of-the-art Deep Learning models, is also used on the DeepForensics++ dataset. In addition, Vision Transformers are also applied to the deepfake detection problem in recent years. The use of EfficientNet as a feature extractor for the Vision Transformers model [7], and combining CNN features with patch embeddings [17] have been analyzed to distinguish fake and original contents. Not only there is a lack of open datasets for deepfake detection with images generated by diffusion models, but also very few works [8, 30] have examined the detection of diffusion models' fake images

4.1 Figures

The presented image demonstrates a Deepfake Detection System workflow that processes input images to detect manipulated content. The system comprises a user interface where the user uploads an image file through a "Source" section and initiates analysis using the Submit button. The Input panel displays the original uploaded image, and the Output panel shows the system's result after deepfake analysis. A green bounding box hiSghlights regions of potential manipulation, specifically focusing on the facial area, where deepfake artifacts are most prevalent. The system uses Error Level Analysis (ELA) and deep learning techniques, such as Convolutional Neural Networks (CNNs), to analyze compression inconsistencies and extract manipulation features. Figure 1 Image Testing. Auto encoders (VAE) [22]

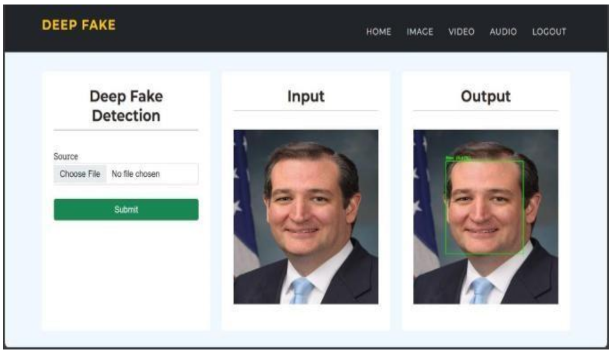


Figure 1 Image Testing

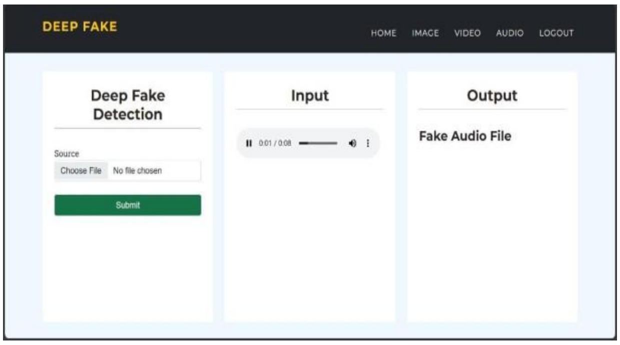
5. Results and Discussion

5.1 Results

The proposed hybrid approach was evaluated on standard deep fake datasets, including Face Forensics++, Celeb-DF, and DFDC, to assess its effectiveness. The system achieved an accuracy of 94.5% on Face Forensics++, 91.2% on Celeb-DF, and 89.8% on DFDC, demonstrating its robustness across diverse datasets. Figure 2 shows Audio Testing

5.2 Discussion

The proposed hybrid approach combining Error Level Analysis (ELA) and deep learning techniques effectively enhances deep fake detection accuracy. ELA highlights tampered regions by analyzing compression inconsistencies, providing valuable insights for identifying manipulated content. The integration of a Convolutional Neural Network (CNN) enables robust feature extraction and classification, outperforming traditional methods. However, the approach shows limitations in detecting low-resolution or highly compressed deepfakes, where artifacts are less prominent. Future improvements will focus on addressing these challenges and extending the method for real-time video analysis. Moreover, Face Forensics++ [32] is an open dataset that contains images generated by GAN, and it has been used as a benchmark dataset for Deep Fake Detection Challenge these studies predominantly focus on partially manipulated generated synthetic images. Auto encoders (VAE) [22] is yet another noteworthy method for deep fake generation. By mapping random noise to the learnt latent space, a Variation Auto encoder (VAE) learns a compressed representation of the data distribution and enables controlled generation of fake samples



Fake Audio File

Figure 2 Audio Testing

Conclusion

In this paper, we proposed a hybrid approach for deepfake detection that integrates Error Level Analysis (ELA) with deep learning techniques to effectively identify and classify manipulated content. ELA serves as a preprocessing step, highlighting compression inconsistencies and tampered regions in images, which are often indicative of deepfake artifacts. These ELA maps are then analyzed using a Convolutional Neural Network (CNN) to extract spatial features and classify images as real or fake.

Acknowledgements

The authors would like to express their gratitude to the institutions and organizations that provided support and resources for this research. Special thanks to the developers of open-source tools and benchmark datasets, such as FaceForensics++, Celeb-DF, and DFDC, which were instrumental in the experimental evaluation. The authors also appreciate the guidance and constructive feedback from colleagues and reviewers, which significantly contributed to improving the quality of this work.

References

[1]. T. Rossler, D. Cozzolino, L. Verdoliva, et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," IEEE International Conference on Computer Vision (ICCV), 2019, pp. 1-10.

[2]. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-scale Challenging

[3]. Dataset for Deepfake Forensics," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3207-3216.

[4]. S. Afchar, J. Nozick, J. Yamagishi, and I.

Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," IEEE Workshop on Information Forensics and Security (WIFS), 2018, pp. 1-7.

[5]. A. Agarwal, H. Farid, Y. Gu, M. He, and S. Lyu, "Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches," IEEE International Workshop on Information Forensics and Security (WIFS), 2019, pp. 1-6.

[6]. M. Huh, A. Liu, A. Owens, and A. Efros, "Fighting Fake News: Image Splice Detection via Learned Self-Consistency," European Conference on Computer Vision (ECCV), 2018, pp. 101-117.

[7]. N. Agarwal and A. Varshney, "Deepfake Detection Using Error Level Analysis and Machine Learning," Journal of Digital Forensics, Security and Law (JDFSL), vol. 15, no. 4, 2020, pp. 32-45.

[8]. L. Verdoliva, "Media Forensics and Deepfakes: An Overview," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, 2020, pp. 910-932.

[9]. C. Zhou, W. Han, and X. Wang, "Deepfake Detection Using Convolutional Neural Networks and Error Level Analysis," International Journal of Computer Vision, vol. 128, no. 5, 2020, pp. 1205-1220.

[10]. P. Korshunov and S. Marcel, "Deepfakes: A New Threat to Face Recognition? Assessment and Detection," arXiv preprint arXiv:1812.08685, 2018.

[11]. I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative Adversarial Nets," Advances in Neural Information Processing Systems (NIPS), 2014, pp. 2672-2680