


Detection of Frauds in Deep Fake Using Deep Learning [†]

Osipilli Aparna ¹, Pakanati Rani ¹, Tulluri Ramya ¹, Tanneru Priyanka ¹, Neela Sundari ¹, P. G. K. Sirisha ¹,
Repudi Ramesh ¹ and Dama Anand ^{2,*} 

¹ Department of Computer Science and Engineering, KKR & KSR Institute of Technology and Sciences, Guntur 522006, India; 20jr1a05d0@gmail.com (O.A.); 20jr1a05d2@gmail.com (P.R.); 20jr1a05f2.cse@gmail.com (T.R.); tannerupriyanka9@gmail.com (T.P.); ngargeya@gmail.com (N.S.); sirishapgk@gmail.com (P.G.K.S.); repudiramesh@gmail.com (R.R.)

² Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green-Fileds, Guntur 522302, India

* Correspondence: ananddama89@kluniversity.in

[†] Presented at the 5th International Conference on Innovative Product Design and Intelligent Manufacturing Systems (IPDIMS 2023), Rourkela, India, 6–7 December 2023.

Abstract: Research on DeepFake detection using deep neural networks (DNNs) has gained more attention in an effort to detect and categorize DeepFakes. In essence, DeepFakes are regenerated content made by changing particular DNN model elements. In this study, a summary of DeepFake detection methods for images and videos involving faces will be given based on their effectiveness, outcomes, methodology, and type of detection method. We will analyze and categorize the many DeepFake-generating techniques now in use into five primary classes. DeepFake datasets are frequently used to train and test DeepFake models. We will also cover the latest developments in DeepFake dataset trends that are currently accessible. We will also examine the problems in building a generalized DeepFake detection model. Lastly, the difficulties in creating and identifying DeepFakes will be covered.

Keywords: deep learning; DeepFake; CNNs; detection



Citation: Aparna, O.; Rani, P.; Ramya, T.; Priyanka, T.; Sundari, N.; Sirisha, P.G.K.; Ramesh, R.; Anand, D. Detection of Frauds in Deep Fake Using Deep Learning. *Eng. Proc.* **2024**, *66*, 48. <https://doi.org/10.3390/engproc2024066048>

Academic Editors: B. B. V. L. Deepak, M. V. A. Raju Bahubalendruni, Dayal Parhi, P. C. Jena, Gujjala Raghavendra and Aezeden Mohamed

Published: 23 September 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The necessary detection of fake documents is not a new problem. This problem has persisted for a while. In previous document legitimation processes, which involved investigation, verification, and proofreading, digital data played no significant role. Digital data are hard to ignore given the recent explosion in Internet use and applications in medical imaging, legal forensics imaging, digital marketing, and sensitive satellite image processing. Furthermore, the proliferation of digital data for a range of uses is encouraging an increase in criminal activities. In this particular context, trends point to significant risks and a decline in the reliability of digital data. These days, it is also critical to verify digital papers and ascertain whether the acquired digital data are authentic or have been falsified. Research on multimedia forensics has been conducted for at least 15 years, with contributions from government agencies, large IT companies, and academic institutions. Thanks to the methodologies used in research and benchmark datasets, important insights have been obtained. Digital media confirmation can guarantee that the digital, semantic, and physical domains are all preserved, as it is widely acknowledged that this methodology is extremely effective [1]. In fact, it is increasingly displacing most other types of technology. DeepFakes, incredibly realistic fake images and videos, may now be produced by combining deep learning with computer vision techniques like autoencoders and GANs. An image or video can be altered by an attacker or even a non-technical machine learning user by changing its content. This technique, called DeepFakes—a play on the phrases “deep learning” and “fakes”—creates a new version that is exactly the same for both computers and humans. People’s trust in digital media content has been weakened by the emergence of DeepFakes

since they can no longer trust the visuals they are seeing. Research on recognizing or detecting fabricated or modified media is regarded as traditional research when deep learning is not used [2].

2. Literature Review

A unique type of deep-learning architecture that has drawn a lot of interest in computer vision and robotics is CNN, also known as ConvNet. In 1979, Kunihiko Fukushima proposed neocognitron, a concept that would eventually be referred to as the precursor to CNNs. Additionally, Le-Cun et al. detailed the design of CNNs and subsequently presented an upgraded version. It was discovered that a CNN that had been constructed, named LeNet-5, could categorize handwritten numbers. Popular architectures from 2012 to 2015 are analyzed. As well as their fundamental elements uses are covered in convolution neural networks. Three different types of layers make up the fundamental framework of a CNN model: convolutional, pooling, and fully connected [3]. The CNN model’s fundamental structure is shown in Figure 1. Feature extraction is the convolution layer’s main function. The feature map is created during the convolutional process by applying an array of integers (kernel) across inputs (tensor). Each kernel element and the input tensor are elementwise multiplied to create a feature map, and the outputs are then added to determine the kernel element. To create the components of the feature map for that kernel, the kernel convolves over each element on the input tensor. By using several kernels to accomplish the convolution operation, an infinite number of feature maps may be produced.

Tool	Accuracy	Speed	User-friendliness	Scalability	Integration
Sensity AI	High	Fast	Easy	High	API, SDK
Trupic	High	Fast	Easy	High	API
D-DI	High	Fast	Easy	High	API
Deeptrace	High	Fast	Moderate	High	API
DeepSecure.ai	High	Fast	Easy	High	API
iProov	High	Fast	Easy	High	API
Blackbird AI	High	Fast	Easy	High	API
XRVision	High	Fast	Easy	High	API
Sentinel	High	Fast	Easy	High	API
Amber	High	Fast	Easy	High	API
FaceForensics++	High	Fast	Easy	Low	Open-source
FakeSpot	Moderate	Fast	Easy	High	Web-based

Figure 1. The CNN model’s fundamental structure.

A RNN is a neural network where the input for the subsequent phase is taken from the output of the preceding step. Typically, neural networks have independent inputs and outputs; however, there are circumstances where past words are required, like when predicting a phrase’s next word, and so, the prior words must be retained. As a result, RNNs were developed, solving the issue by using a hidden layer. Most importantly, RNNs have a hidden state that retains specific information about a sequence. Every calculation-related piece of information is stored in a RNN’s “memory.” Since this memory generates the same result by carrying out the same task on all inputs, it uses the same parameters for each input. This approach reduces the parameter complexity compared to other neural networks. Long short-term memory (LSTM) [4], which manages long-term dependencies, was proposed in 1997 by Hochreiter and Schmidhuber for cases where there is a significant gap between the pertinent input data. Since LSTM achieves almost all of the fascinating results based on RNNs, it has been the center of attention in terms of deep learning. Recurrent cells, whose states are influenced by both past states and present input via feedback connections, make up the recurrent layers, sometimes referred to as hidden layers in RNNs [5].

3. Methodology

In order to assess geographical features, increase detection effectiveness, and strengthen overall DeepFake recognition capabilities, this approach makes use of pre-existing DNN models. All of these techniques are data-driven. Nevertheless, very few research have

evaluated the durability of these DNN-based detection techniques against adversarial attacks, and none of them are immune to them. Three different kinds of studies employ DNNs to find DeepFakes [6]. Existing DNN models are refined for enhanced detection performance; artifact hints are scrutinized, and multiple dataset types are trained for enhanced generalization performance in a face-swapping-based CNN and LSTM detection technique. Frame-level features are extracted using InceptionV3 (CNN), and the CNN output is then transferred to LSTM to create a sequence descriptor for classification. The model's maximum accuracy in determining if a video is a DeepFake or clean is more than 97%. A capsule network was created to address problems related to computer vision and digital forensics. It is used to detect forged images and videos in a variety of forging scenarios, such as replay attack detection and (partial and whole) computer-generated image/video detection. It has recently been shown that hierarchical pose connections between object components can be described using a capsule network based on a dynamic routing algorithm [7].

4. Results and Comparison

To my most recent knowledge, updated in January 2022, DeepFace, which employs a deep learning Convolutional Neural Network (CNN) model, achieved impressive results in facial recognition accuracy. It claimed a high accuracy rate, even comparable to human performance in certain datasets. However, it is important to note that technology evolves rapidly, and new research or updates may have emerged since then. For the latest and most accurate information, I recommend checking recent research papers, articles, or official documentation related to DeepFace and advancements in deep learning for facial recognition [8]. DeepFace, developed by Facebook, is a facial recognition system that utilizes deep learning algorithms. Discussions around it often involve privacy concerns, ethical considerations, and the potential misuse of facial recognition technology. Additionally, there is ongoing research to improve accuracy, address biases in the algorithms, and ensure compliance with regulations. The broader discourse encompasses the balance between technological advancements and the need to safeguard individual rights and societal values [9].

5. Conclusion and Future Enhancement

An extensive investigation of DeepFake, a popular and contemporary method, is presented in this article. It describes the principles, benefits, and dangers of DeepFakes and DeepFake applications based on GANs [10]. Moreover, DeepFake detection models are also discussed. Since most deep learning-based detection methods currently in use are not transferable or generalizable, multimedia forensics may still be in its early stages [11]. Considerable interest has been exhibited by numerous important organizations and people pushing applied approaches [12]. Additional security measures are required because it still takes a lot of labor to preserve data integrity. In AI vs. AI battles where no side has the upper hand, experts also forecast a new wave of DeepFake propaganda [13].

Author Contributions: O.A., P.R., T.R., T.P., N.S., P.G.K.S., R.R.: Article preparation; D.A.: conceptualization. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Farid, H. Image forgery detection. *IEEE Signal Process. Mag.* **2009**, *26*, 16–25. [[CrossRef](#)]
2. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Proc. Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1–9.
3. Baldi, P. Autoencoders, unsupervised learning, and deep architectures. In Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop, Washington, DC, USA, 2 July 2011; pp. 37–49.
4. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 4401–4410.
5. Mirsky, Y.; Lee, W. The creation and detection of deepfakes: A survey. *ACM Comput. Surv.* **2022**, *54*, 1–41. [[CrossRef](#)]
6. Masood, M.; Nawaz, M.; Malik, K.M.; Javed, A.; Irtaza, A. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *arXiv* **2021**, arXiv:2103.00484. [[CrossRef](#)]
7. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; Ortega-Garcia, J. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Inf. Fusion* **2020**, *64*, 131–148. [[CrossRef](#)]
8. Nguyen, Q.V.H.; Nguyen, D.T.; Nguyen, D.T.; Huynh-The, T.; Nahavandi, S.; Nguyen, T.T.; Pham, Q.-V.; Nguyen, C.M. Deep learning for deepfakes creation and detection: A survey. *arXiv* **2019**, arXiv:1909.11573.
9. Verdoliva, L. Media Forensics and DeepFakes: An Overview. *IEEE J. Sel. Top. Signal Process* **2020**, *14*, 910–932. [[CrossRef](#)]
10. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **1980**, *36*, 193–202. [[CrossRef](#)] [[PubMed](#)]
11. LeCun, Y.; Boser, B.; Denker, J.; Henderson, D.; Howard, R.; Hubbard, W.; Jackel, L. Handwritten digit recognition with a back-propagation network. *Proc. Adv. Neural Inf. Process Syst.* **1989**, *2*, 396–404.
12. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [[CrossRef](#)]
13. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.