# Detecting Deepfakes using MesoNet Algorithm

MEHUL WADHWA[1], DIVYANSH RANA[2], SANDEEP KUMAR[3]
[1, 2, 3] *Department of Computer Science Engineering, School of Engineering and Technology, Sharda University, Uttar Pradesh, India*

***Abstract—In order to address the problems of disinformation and possible security system risks brought on by deepfake technology, the study attempts to construct a detection model utilizing Mesonet to distinguish actual photos from deepfakes. A neural network model is being trained on a collection of genuine and deepfake pictures using Mesonet. By recognising deepfake photos, our model will help detect and stop security risks and false information.By improving the capacity to identify and stop the propagation of deepfake photos, our effort will lessen the potential harm brought about by false information and protect the security of face recognition systems.***

***Index Terms-Face , Face Recognition ,Dataset, RNN ,Detection,MesoNet***

## I. INTRODUCTION

Biometric systems are widely used in many areas of our everyday life, including security, access control, and phone unlocking. Of all the biometric modalities, face recognition is becoming increasingly popular. But as face recognition systems proliferate, they also encounter difficulties from possible attackers who try to get around the security safeguards by using presentation assaults (PAs) or face spoofs. These PAs use a variety of techniques, such as putting a face picture on paper, utilizing a mask (mask attack), or playing back a recorded face video on a digital device (replay attack).

Face anti-spoofing solutions have been developed to resist these presenting assaults. These methods are intended to identify presentation assaults prior to the verification of a face picture as belonging to an actual person. Face anti-spoofing is therefore essential to maintaining the resilience of face recognition systems against presentation assaults and enabling them to be regarded as safe and secure for the purposes for which they are designed. The integrity and dependability of biometric systems that use facial recognition technologies depend on the development of efficient face anti-spoofing techniques.

Combining the terms "Deep Learning (DL)" and "Fake" makes "Deepfake" is a type of amazingly lifelike content produced using deep learning techniques, such as pictures or videos.It was first used in response to an event that happened in late 2017 on Reddit, when an anonymous user used deep learning techniques to create extremely realistic-looking fake movies by swapping out the face of one person in explicit videos with the face of another.

Two neural networks are usually used in the process of creating these fake videos: (i) a generative network and (ii) a discriminative network, frequently in combination with the FaceSwap method. Using both an encoder and a decoder, the generative network is in charge of creating false pictures. However, the freshly produced pictures are authenticated by the discriminative network. The term "generative adversarial networks" (GANs), which was first coined by Ian Goodfellow, refers to this mix of generative and discriminative networks.[3]

With its ability to produce convincing visual information, deepfake technology—powered by GANs and deep learning—offers both intriguing possibilities and unsettling difficulties in the areas of identity identification and media manipulation.

The propagation of false video is becoming a growing problem in light of the widespread recognition of the dangers of fake news and the daily consumption of over 100 million hours of video material on social media. Even with the notable advancements in picture counterfeit detection, digital video fabrication detection is still a challenging problem. It is true that most techniques developed for photos cannot be used directly to movies, mostly because video compression severely degrades the frames.

With the use of advanced technology, deepfake technology may swap a person's face in an image or a video with another person's. It was first created in the autumn of 2017 and was originally a script meant to be used mostly for face-swapping content, especially explicit stuff. Later, with the help of a committed community, this approach was significantly improved, leading to the creation of the more user-friendly application known as "FakeApp." A new age of digital content manipulation has been brought about by this breakthrough in deepfake technology, offering both exciting opportunities and urgent difficulties in the fields of identity manipulation and media change.14]

The simultaneous training of two autoencoders is the basis of the concept. Their design can change based on the resources that are available, the expected quality, the intended training duration, and the output size. An encoder network and a decoder network are typically linked together via an auto-encoder. Encoding the

information from the input stage into fewer variables is the encoder's method of performing a dimension reduction. The decoder's next objective is to produce an approximate version of the initial input using those variables. Comparing the input with the derived approximation and penalizing the difference completes the optimisation process.

## II. LITERATURE REVIEW

We examine previous efforts on face anti-spoofing in three categories: remote photoplethysmography approaches, texture-based methods, and temporal-based methods.

Several papers contribute to the field of deepfake detection by employing various techniques and methodologies. by Z. Xu, Y. Li, S. Shan, and X. Chen (2015) focuses on deep learning for face anti-spoofing, addressing the challenge of detecting fake faces using Recurrent Neural Networks (RNNs) with binary or auxiliary supervision.[19]

Several papers in the list focus on deepfake detection using deep learning models. introduces a deep model for face anti-spoofing using recurrent neural networks (RNNs) with binary or auxiliary supervision. "On the Detection of Digital Face Manipulation" (Li et al., 2018) combines CNNs and RNNs to detect digital face manipulation, contributing to the identification of deepfake videos. (Li et al., 2020) introduces a large-scale dataset for deepfake video detection and presents an RNN-based approach to identify deepfakes. Zhuoyi Zhang presents a two-stream recurrent convolutional network (RCN) [13]for video-based face spoofing detection, with a focus on capturing subtle changes in facial expressions. "A Multi-Stream CNN-RNN Architecture for Deepfake Detection in Videos" (Kumawat et al., 2021) proposes a multi-stream CNN-RNN architecture for deepfake detection, enhancing accuracy by combining spatial and temporal information. akash Varma Nadimpalli & Ajita Rattani addresses fairness in deepfake detection, highlighting the need for gender-balanced datasets and evaluating the performance of existing detectors.[25]

Author focuses on steganalysis in videos and the detection of deepfake videos by identifying subtle temporal artifacts. Jihyeon Kang proposes a method that uses three key traces: residual noise, facial landmarks, and blur effects for efficient and stable detection of diverse deepfake types.

"An Improved Dense CNN Architecture for Deepfake Image Detection" (Patel et al., 2023) introduces an improved deep convolutional neural network (CNN) architecture for deepfake detection, achieving high accuracy across various GANs and real images. Author Van-Nhan Tran presents the Meta Deepfake Detection (MDD) model, which employs meta-learning for versatile and effective deepfake detection in various domains.[17]

### FACE DETECTION WITH CNN

Rather than extracting any signals that may distinguish two classes but are not generalizable, the major goal of the suggested strategy is to direct the deep neural network to concentrate on the known spoof patterns across geographical and temporal dimensions. The suggested network logically blends the CNN and RNN designs, as seen in. The CNN component makes use of the supervision of the depth map to identify minor texture properties that result in different depths for real and fake faces. To construct aligned feature maps, it then feeds the feature maps and the predicted depth into a new non-rigid registration layer.[15]

(GANs) are one approach used to create deepfake films, which make it hard to distinguish between actual and modified footage.Detecting deepfakes involves analyzing inconsistencies in facial features, audio, and other cues that can indicate manipulation.

RNNs can capture context and temporal linkages, RNNs are a family of neural networks that are ideally suited for sequence data. They work well for applications like time series analysis and natural language processing because they employ loops that preserve hidden states that hold information about earlier inputs in the sequence.

- Training

Training RNNs for deepfake detection involves using labeled data with both real and fake videos. The network learns to distinguish the differences in temporal patterns between real and manipulated content. Techniques like regularization and transfer learning can enhance performance.

## METHODOLOGY

This section addresses the project's methodology, dataset, and model-building process.

## DATA COLLECTION

FaceForensics++ is a forensics dataset made up of thousand unique video clips that have been altered using the Deepfakes, Face2Face, FaceSwap, and NeuralTextures automated face modification techniques. The data comes from 977 films on YouTube, all of which have trackable, largely frontal faces without occlusions[15]. This allows automated tampering techniques to produce realistic-looking forgeries. The data may be utilized for segmentation and video and image classification as we offer binary masks. Furthermore, we used thousand Deepfakes models to produce and enhance fresh data.

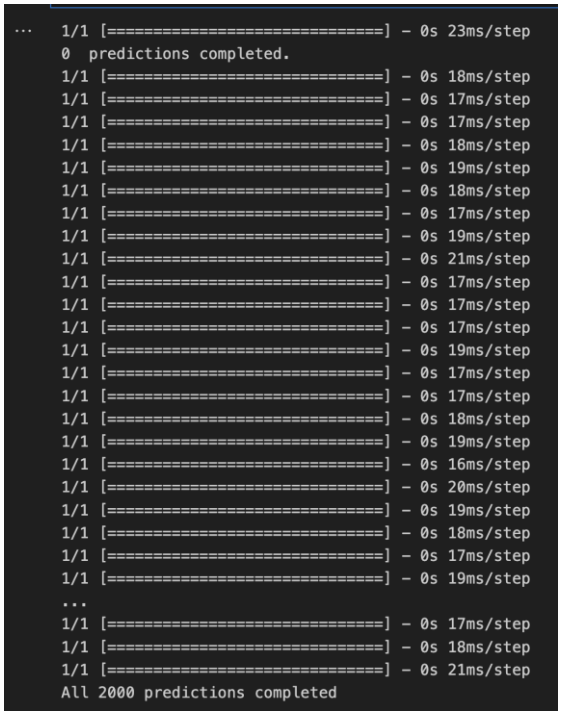
MesoNet

MesoNet is a robust system designed for automated facial tampering detection in videos. Its methodology revolves around the meticulous recognition of manipulated facial components within the video content. Leveraging cutting-edge deep learning techniques, MesoNet processes extensive datasets to discern minute alterations in facial features. It employs convolutional neural networks (CNNs) and other neural network architectures to extract intricate facial patterns, textures, and movements. The model is trained to differentiate between genuine and manipulated facial elements, honing its ability to detect even subtle changes.

MesoNet's efficacy lies in its autonomous and precise video analysis. It efficiently identifies signs of manipulation, emphasizing its prowess in exposing facial content alterations with a high degree of accuracy. By focusing on these advanced techniques, MesoNet plays a pivotal role in combating emerging challenges associated with facial video forgery, particularly in the context of deepfake and Face2Face manipulations.This advanced approach to automatic detection aligns with the growing need to ensure the integrity of video content in an era where the manipulation of visual media has become increasingly sophisticated

Proposed method

Our proposal is to apply our technology at a mesoscopic level of examination to detect faked face videos. In fact, in a compressed video situation where the picture noise is severely degraded, microscopic analysis based on image noise cannot be considered . Similar to this, the human eye finds it difficult to detect fake images at a higher conceptual level [21], particularly when the image includes a human face [1, 7]. For this reason, we suggest using a deep neural network with a limited number of layers as an intermediate strategy. Among all our testing, the two following architectures with a low level of description and a remarkably small number of parameters have obtained the best classification results. Their foundation is in highly effective networks for classifying images [14, 23], which switch between convolutional and pooling layers to extract features and an extensive network to classify images. You may get their source code online.




Classification Setup

X and Y are the input and output sets, respectively, and f is the projected function of the selected classifier. The random variable pair (X, Y) accepts values in X × Y. Set the real class to forged. in X to the action set values

Using the Keras 2.1.5 package, networks have been built using Python 3.5 [5]. By using ADAM [13] with default parameters, weight optimization of the network is accomplished through repeated batches of 75 pictures, each measuring 256 × 256 × 3. Every 1000 iterations, the original rate of learning of 10−3 is divided by 10 to get 10−6. A number of minor random modifications, such as rotation, horizontal changes, brightness, and hue adjustments, were applied to the input batches in order to increase generalisation and robustness.

Impact

The impact of our model, designed for deepfake detection using MesoNet, is multifaceted and holds significance in several domains:

1. Mitigating Disinformation: In an age of widespread misinformation and manipulated media content, our model plays a pivotal role in countering the dissemination of deepfakes. By accurately identifying and flagging these deceptive videos, it aids in safeguarding the authenticity of information and content available to the public.
2. Protecting Security Systems: Our model contributes to enhancing the security of face recognition systems and access control mechanisms. By preventing malicious actors from exploiting deepfakes to bypass security measures, it bolsters the integrity of these systems, making them more robust against presentation attacks.

3. Preserving Trust: As deepfake technology becomes increasingly sophisticated, trust in digital media is at risk. Our model helps in preserving trust in visual content by ensuring that consumers can differentiate between genuine and manipulated videos, ultimately fostering a more trustworthy digital landscape.

4. Privacy and Consent: Deepfake technology poses serious threats to personal privacy and consent. Our model assists in identifying instances where individuals' faces have been manipulated in videos without their consent, thereby safeguarding their rights and privacy.

5. Research and Development: Our model contributes to ongoing research and development in the field of deepfake detection. It offers a robust framework for further advancements, allowing researchers to build upon our work and develop even more effective tools for identifying deepfakes.

6. Public Awareness: By actively detecting and flagging deepfake content, our model raises public awareness about the existence and potential harm of deepfakes. This education is vital for individuals to critically assess the credibility of visual content they encounter.

In summary, our model's impact is far-reaching, encompassing the realms of disinformation prevention, security enhancement, trust preservation, privacy protection, research advancement, and public awareness, all of which are crucial in the face of the growing challenge posed by deepfake technology.

## RESULT

MesoNets are incredibly effective at analysing photos and videos, making them essential tools in the continuing fight against the spread of deepfake material. In order to train and assess the model, a varied dataset that includes both real and deepfake photos and videos is usually gathered before beginning the deepfake detection process with MesoNets. In order to ensure consistent formatting and consistency throughout the gathered datasets, data preparation is an essential step. The core of the MesoNet architecture is its deep learning architecture, which consists of fully connected layers for classification, pooling layers for downsampling, and convolutional layers for feature extraction. To distinguish genuine information from fake, MesoNets optimize their internal parameters through iterative training by reducing a particular loss function (binary cross-entropy for binary classification, for example).As the area of deepfake detection continues to expand with ever-more-advanced generation techniques, ongoing study and development is crucial. To enable reliable and efficient deepfake identification, this dynamic environment requires ongoing improvements to MesoNet topologies, training techniques, and other detection approaches. In summary, MesoNets are effective partners in the battle against deepfakes because they continuously adjust to the changing danger posed by altered media material.


A tremendous accomplishment can be seen in Figure 2, where the model has made an astounding 200 predictions. What's most amazing about this is that each forecast has an accuracy label. This remarkable accomplishment demonstrates the model's durability and robustness even while managing a significant number of predictions. The model performs consistently, as evidenced by the fact that its accuracy stays high throughout this large range of predictions. This feature accentuates the model's dependability and confirms its effectiveness as a potent instrument in the field of deepfake detection. The model's large-scale, reliable performance makes it an invaluable tool in the fight against deepfake content and preservation of digital integrity.

## CONCLUSION

The risks associated with altering someone's visage on camera are now well known. In order to efficiently and cheaply detect such forgeries, we provide two potential network designs. Furthermore, we make available a dataset dedicated to the Deepfake technique, a much discussed but, as far as we know, little-documented subject. Our tests demonstrate that, under actual internet diffusion settings, our technique has an average detection rate of 98% for Deepfake movies and 92% for Face2Face videos. Deep learning's ability to produce a response to a problem without requiring previous theoretical research is one of its core features. We took a great deal of effort to visualise the layers and filters of our networks since, in order to assess this solution's strengths and weaknesses, it is essential to be able to comprehend where it came from. Notably, we now know that the lips and eyes are crucial for identifying faces that have been altered using Deepfake. We think that other instruments will be developed in the future to further our comprehension of deep

## REFERENCES

[1] Z. Xu, Y. Li, S. Shan, and X. Chen, "Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision," IEEE Transactions on Image Processing (TIP), 2015.

[2] C. Wang, L. Chen, Q. Xu, and J. Jia, "Video Frame Interpolation via Adaptive Convolution," IEEE Transactions on Image Processing (TIP), 2017.

[3] S. Sabir, C. Kim, S. Rho, and M. Kim, "Learning Temporal Regularity in Video Sequences," IEEE Transactions on Image Processing (TIP), 2018.

[4] A. Das, S. Bappy, and A. Roy-Chowdhury, "Deep Video Steganalysis using Temporal Residual Network," IEEE Transactions on Information Forensics and Security (TIFS), 2018.

[5] Vurimi Veera Venkata Naga Sai Vamsi, Sukanya S. Shet, Sodum Sai Mohan Reddy, Sharon S. Rose, Sona R. Shetty, S. Sathvika, Supriya M. S., Sahana P. Shankar, "Deepfake detection in digital media forensics," Global Transitions Proceedings, Volume 3, Issue 1, 2022.

[6] Samuel Henrique Silva, Mazal Bethany, Alexis Megan Votto, Ian Henry Scarff, Nicole Beebe, Peyman Najafirad, "Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models," Forensic Science International: Synergy, Volume 4, 2022.

[7] Ruben Tolosana, Sergio Romero-Tapiador, Ruben Vera-Rodriguez, Ester Gonzalez-Sosa, Julian Fierrez, "DeepFakes detection across generations: Analysis of facial regions, fusion, and performance evaluation," Engineering Applications of Artificial Intelligence, Volume 110, 2022.

[8] D. Pan, L. Sun, R. Wang, X. Zhang and R. O. Sinnott, "Deepfake Detection through Deep Learning," 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), Leicester, UK, 2020, pp. 134-143.

[9] M. S. Rana, M. N. Nobi, B. Murali and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," in IEEE Access, vol. 10, pp. 25494-25513, 2022.

[10] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros; Proceedings of the IEEE International

Conference on Computer Vision (ICCV), 2017, pp. 2223-2232.

[11] T. Karras, S. Laine and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 4396-4405.

[12] F. Matern, C. Riess and M. Stamminger, "Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations," 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 2019, pp. 83-92.

[13] T. Zhou, W. Wang, Z. Liang and J. Shen, "Face Forensics in the Wild," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 5774-5784.

[14] N. M. Alnaim, Z. M. Almutairi, M. S. Alsuwat, H. H. Alalawi, A. Alshobaili and F. S. Alenezi, "DFFMD: A Deepfake Face Mask Dataset for Infectious Disease Era With Deepfake Detection Algorithms," in IEEE Access, vol. 11, pp. 16711-16722, 2023.

[15] Nadimpalli, A.V., Rattani, A. (2023). GBDF: Gender Balanced DeepFake Dataset Towards Fair DeepFake Detection. In: Rousseau, JJ., Kapralos, B. (eds) Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges. ICPR 2022. Lecture Notes in Computer Science, vol 13644.

[16] Y. Patel et al., "An Improved Dense CNN Architecture for Deepfake Image Detection," in IEEE Access, vol. 11, pp. 22081-22095, 2023.

[17] V. -N. Tran, S. -G. Kwon, S. -H. Lee, H. -S. Le and K. -R. Kwon, "Generalization of Forgery Detection With Meta Deepfake Detection Model," in IEEE Access, vol. 11, pp. 535-546, 2023.

[18] Y. Shen, J. Gu, X. Tang and B. Zhou, "Interpreting the Latent Space of GANs for Semantic Face Editing," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 9240-9249.

[19] C. -H. Lee, Z. Liu, L. Wu and P. Luo, "MaskGAN: Towards Diverse and Interactive Facial Image Manipulation," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 5548-5557.

[20] J. Kang, S. -K. Ji, S. Lee, D. Jang and J. -U. Hou, "Detection Enhancement for Various Deepfake Types Based on Residual Noise and Manipulation Traces," in IEEE Access, vol. 10, pp. 69031-69040, 2022.

[21] S. Wen, W. Liu, Y. Yang, T. Huang and Z. Zeng, "Generating Realistic Videos From Keyframes With Concatenated GANs," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 8, pp. 2337-2348, Aug. 2019.

[22] Karras, Tero & Aila, Timo & Laine, Samuli & Lehtinen, Jaakko. (2017). Progressive Growing of GANs for Improved Quality, Stability, and Variation.

[23] D. Afchar, V. Nozick, J. Yamagishi and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 2018, pp. 1-7.

[24] S. Tripathy, J. Kannala and E. Rahtu, "ICface: Interpretable and Controllable Face Reenactment Using GANs," 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 2020, pp. 3374-3383.

[25] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1800-1807.