International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

DEEPFAKE DETECTION TECHNIQUES AND METHODOLOGIES

Kanak Rajan Verlekar[1], Aditya Rambabu Soni[2], Sumit Ghanshyam Verma[3], Sushant Sanjay Pawar[4]

ASM Institute of Management and Computer Studies University of Mumbai
C–4,Wagle Industrial Estate, Near Mulund Check Naka, Thane West, Opp. Aplab, Mumbai - 400604, Maharashtra, India.

ABSTRACT :

Significant advancements in AI, machine learning, and deep learning over the past few years have led to the development of new tools, approaches, and technologies for manipulating multimedia. While the majority of uses of this technology have been in the fields of education, humour, entertainment, etc., hackers have also taken advantage of them for illegal and criminal activities. Let's take an example where realistic-looking and high-quality false audio, photographs, and videos are being produced with the intention of spreading hate speech, political unrest, and misinformation in addition to harassing and blackmailing individuals. Deepfake is the term used to describe these highly manipulated, realistic-looking, and high-quality material.

The research has outlined a number of methods and strategies for handling issues like these brought on by Deepfake. In this paper, we undertake a systematic literature review (SLR) to provide an updated overview of the research efforts in Deepfake security. We include important publications that present a variety of approaches from 2018 to 2024. We categorise them into four main groups for analysis: deep learning techniques, statistical approaches, blockchain-based usages, and traditional machine learning approaches. We also examined the performance and detection capability of several approaches in relation to various data/datasets and came to the conclusion that different security-based solutions and deep learning-based methods were best for detecting deepfakes.

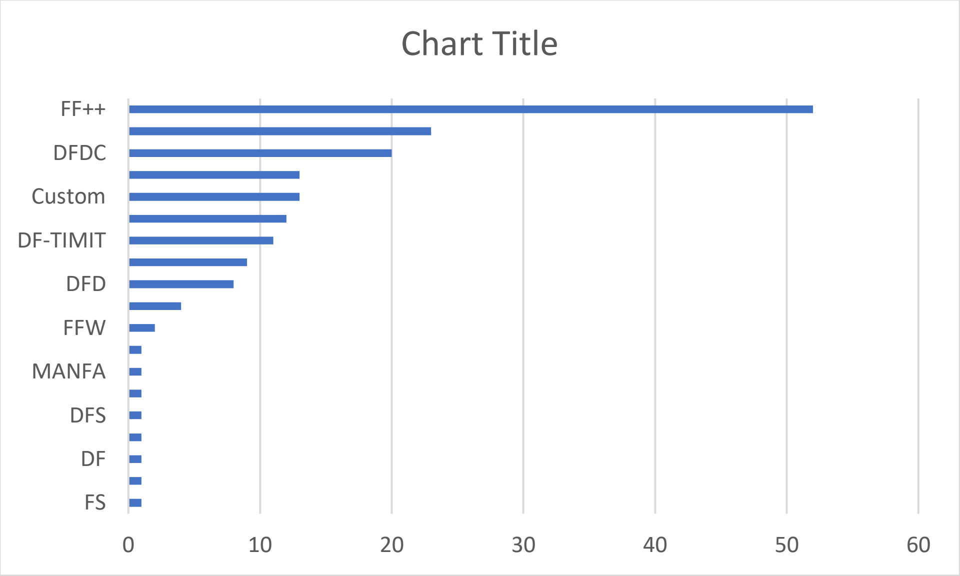
INTRODUCTION :

Prominent advancements in artificial neural network (ANN) technology are fundamental to the transformation of multimedia content. For instance, realistic-looking face swapping in images and videos has been made possible by AI-enabled software applications like FaceApp and FakeApp. With this swapping tool, anyone can change their age, sex, haircut, and other personal characteristics. These phoney recordings are the source of many conflicts and are now widely appreciated on Deepfake. Derived from the words "Deep Learning (DL)" and "Fake," "Deepfake" refers to specific photo-realistic video or picture content created using DL's assistance. This was called after a Reddit user who, in late 2017, used deep learning techniques to replace a person's face in pornographic films with another person's. To create such videos, two particular neural networks: (1) generative network and (2) discriminative network with a FaceSwap method were used. This generative network creates deepfake images using an encoder and a decoder.

The discriminative network proves the authenticity of the newly generated images. The combination of these two neural networks is known as Generative Adversarial Networks (GANs) which Ian Goodfellow had proposed. In recent years, analysts have made critical strides in handling this issue, utilizing different methods such as computer vision, facial recognition, and deep learning. These approaches use both visual and audio cues, scrutinizing components like facial expressions, lip reading, and irregularities in pixel-level details to pick out signs of fake media.

Furthermore, some strategies analyse metadata and relevant data to evaluate the authenticity of the video source. While many of the existing detection algorithms may have proven effective, the battle against deepfakes remains a never-ending challenge. Adversarial networks constantly evolve, creating more sophisticated and real-life-looking manipulations. Therefore, researchers must continuously keep themselves updated and enhance their detection techniques to keep up with the rapidly advancing Deepfake technology.

MAJORLY USED DATASETS:



FF: FaceForensics,
FF++:FaceForensics++,
DFD: Deepfake Detection,
CELEB-A: DeepFake Forensics V1,
CELEB-DF: DeepFake Forensics V2,
DFDC: Deepfake Detection Challange,
DF-TIMIT: Deepfake-TIMIT,
DF: Deepfake,
DF-1.0: DeeperForensics-1.0,
DFS: Deep Fakes,
FFD: Fake Faces in the Wild,
FE: FakeET,
FS: Face Shifter,
WDF: Wild Deepfake,
SMFW: SwapMe and FaceSwap,
SFD: Swapped Face Detection,
UADFV: Inconsistent Head Poses,
MANFA: Tampered Face,
Other: Authors' Custom datasets

PROBLEM STATEMENT:

Deepfake technology has created a tense environment regarding its potential for misuse and deception. To combat this growing threat, deep learning-based detection methods have emerged as a crucial defence mechanism.

By taking it up a notch advanced neural networks aim to accurately recognise manipulated content and separates it from the authentic media. Detecting deepfakes is essential for preserving trust, maintaining the integrity of digital media, and safeguarding individuals from malicious exploitation.

Through ongoing research and development, deep learning-based detection offers a promising solution to tackle the alarming proliferation of deepfakes and protect society from their harmful consequences.

PROPOSED METHODOLOGY:

1) MACHINE LEARNING BASED METHODS

Traditional machine learning (ML) techniques help understand the rationale behind any choice that can be described in human words. Such approaches are appropriate for the Deepfake domain because they provide a more complete understanding of the data and processes. Furthermore, hyperparameter tweaking and model design changes are considerably easier to handle. Tree-based ML techniques, such as Decision Tree, Random Forest, Extremely Randomized Trees, and so on, depict the decision process as a tree structure. Thus, a tree-based technique has no explainability concerns.

GANs are used to automatically train a generative model by treating an unsupervised problem as supervised and producing photorealistic synthetic faces in photos or videos. Some ML-based algorithms seek to reveal certain abnormalities identified in such GANs created fake images and videos.

Machine learning-based Deepfake detection algorithms have achieved up to 98% accuracy. However, performance depends on the dataset type, feature selection, and alignment between train and test sets. Using a similar dataset with a certain ratio, such as 80% for a train set and 20% for a test set, can lead to better results. Using unrelated datasets reduces performance by approximately 50%, which is an unjustified assumption.

2) STATISTICAL MEASUREMENTS BASED METHODS

Using various statistical metrics, such as average normalized cross-correlation scores, to compare original and suspected data helps to assess the data's originality. Using photo response nonuniformity (PRNU) to detect Deepfakes in video frames.

PRNU is a distinctive noise pattern in digital images caused by faults in the camera's light-sensitive sensors. Because of its uniqueness, it is also known as the fingerprint of digital photographs. The study creates a series of frames from input films and saves them in chronologically organized directories. Each video frame is cropped with the same pixel range to preserve and clarify the PRNU sequence. The frames are then separated into eight equal groups. It then applies the second-order FSTV algorithm to each frame to generate the conventional PRNU pattern. It then correlates them by computing the differences between the normalized cross-correlation scores and the mean correlation score for each frame.

3) BLOCKCHAIN BASED METHODS

Blockchain technology offers a variety of characteristics for verifying the legality and provenance of digital material in a highly trustworthy, secure, and decentralized manner. In public Blockchain technology, everyone has immediate access to every transaction, log, and tamper-proof record. For deepfake detection, public Blockchain is regarded as one of the most acceptable technology options for confirming the authenticity of videos or images in a decentralized manner. When videos or photos are flagged as questionable, users must often investigate their origin.

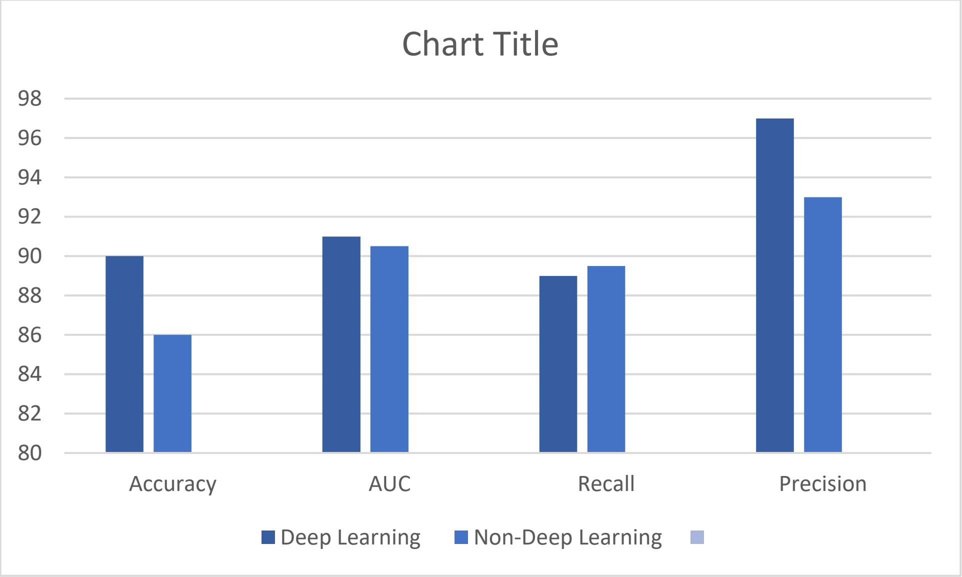
- Presents the proposed solution’s architecture and design details to control and administrate the interactions and transactions among participants.
- Integrates the critical features of IPFS based decentralized storage ability to Blockchain-based Ethereum Name service.
- Demonstrates a general framework built on Blockchain technology by establishing a means of authenticating digital content to its reliable source.

This suggested method creates discriminating features by using multiple LSTM networks as a deep encoder. The features are then compressed and used to hash the transaction.

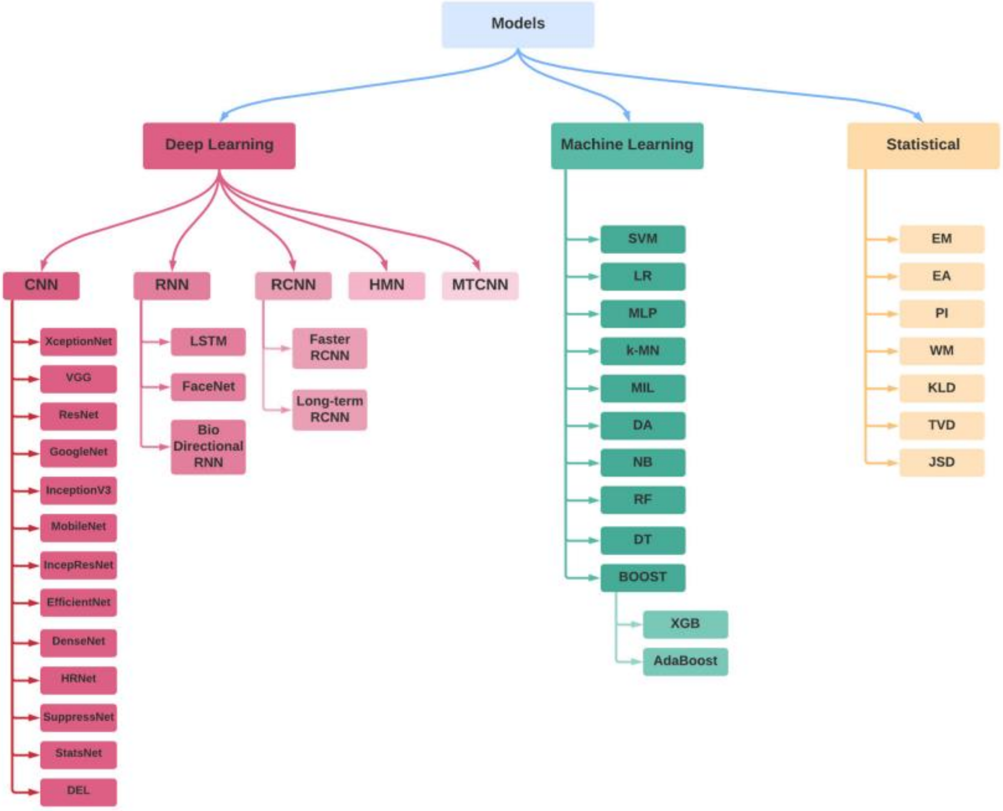
4) DEEP LEARNING BASED METHODS

Many studies have used deep learning-based techniques to identify particular artefacts produced by their creation pipeline in the context of deepfake identification in photos. In order to detect collective GAN-image artefacts as Deepfake, we built a GAN simulator that replicates them and feeds the data to a classifier.

Biological signals, phoneme-viseme mismatches, facial expression and motions (i.e., 2D and 3D facial landmark positions, head attitude, and facial action units), etc. are some of the fundamental methodologies accessible for identifying Deepfake. We group them together under two main categories of techniques: digital media forensics and facial manipulation.



DETECTION MODELS:



PROPOSED ALGORITHMS/USED METHODS:

- Support Vector Machine (SVM)
- Convolutional Neural Network (CNN)
- Recurrent Neural Network (RNN)
- Multi-attentional methods
- Ensemble methods
- Face alignment
- Gauss-Newton optimization
- Hybrid CNN and RNN with particle swarm optimization (PSO) algorithm

Category	Model	#Studies	PCT (%)
	CNN	71	78%
Deep Learning	RNN	12	13%
	RCNN	2	2%
	SVM	11	12%
	k-MN	4	4%
	LR	3	3%
	MLP	3	3%
Machine Learning	BOOST	2	2%
	RF	I	1%
	DT	I	1%
	DA	I	1%

	NB	I	1%
	MIL	I	1%
Statistical	EM	1	1%
	TV. KL, JS	1	1%

STEPS USED FOR DEEPFAKE DETECTION:

1. Data collection: Collecting original and Deepfaked data (images or videos) is done in this initial phase.
2. Face detection: Identifying what parts of an image or video need to be focused on to reveal characteristics like age, gender, race, color , emotions, etc., using facial expressions fall under this stage.
3. Feature extraction: Extracting various facial features as candidate features for the detector.
4. Feature selection: Selecting from the studied data of features those that are most useful for Deepfake detection.
5. Model selection: Finding a model that will be beneficial from a vast variety of available models for classification. These models include deep learning-based models, statistical models & machine-learning models.
6. Model evaluation: Finally, studying the performance of the selected models using various measurement metrics.

INFORMATION GATHERING :

Our goal was to gather as many relevant works as we could for our study questions. To prevent bias, we made an effort to include every possible combination of relevant search terms or keywords when compiling Deepfake detection research.

The main concept behind combining those search phrases with "AND" or "OR" utilizing Boolean terminology. The main keywords for the search terms are (Deepfake OR FaceSwap OR Video manipulation OR Fake face/image/video) AND (detection OR detect) OR (Facial Manipulation OR Digital Media Forensics).

- Web of Science
- IEEE Xplore Digital Library

The repositories include journals, conferences, and archives

FILTERING DIFFERENT FORMS OF INFORMATION:

Certain works focused primarily on Deepfake without including relevant keywords in the abstract, title, or keywords. In this situation, we search various sections of the book for the relevant keywords. If we come across such works, we include them.

A small number of research studies are concurrently published in journals and conferences. To prevent duplications, we took into consideration the most accurate one in this situation.

We sorted between the research that focuses on particular transformation techniques in Deepfake detection.

QUALITY ASSURANCE:

Interpreting the results of a poorly executed study should be done so with caution because biases in the research technique can skew the results. These studies ought to be explicitly omitted from the systematic review or identified as such. It's also crucial to choose the right criteria to assess the quality of the evidence and any ingrained biases in each study.

We have used these criteria to validate the chosen research and have applied the requirements to these studies' review. Furthermore, a cross-checking methodology has been used during the evaluation of these chosen research to guarantee uniformity among various conclusions.

DATA MINING & SUPERVISION:

The process of creating the systems needed to extract data from the research is covered in this step. We conducted a thorough search of popular libraries to locate potentially pertinent materials, we identified the various methods based on the feature analysis that are applied for detecting Deepfakes, datasets that are used by the authors, features used for analysis, measurement metrics used by different authors for their studies.

DATA VISUALIZATION :

In this process, the data extracted from the extraction process is used to visually display and present it in a better way. Histograms, pie charts, graphs, etc.

PERFORMANCE OF MODELS:

Category	Metrics	#Studies	Min	Max	Mean	STD

Deep Learning	Accuracy	50	63.15	100.0	89.73	10.08
	AUC	37	0.572	1.000	0.917	0.114
	Recall	5	82.74	100.0	89.47	12.88
	Precision	6	90.55	100.0	88.89	4.948
Machine Learning	Accuracy	12	85.00	91.07	86.86	11.04
	AUC	12	0.531	1.000	0.909	0.127
	Recall	2	82.74	92.11	89.92	10.15
	Precision	2	90.55	96.40	93.48	4.137

CONCLUSION :

We used basic techniques and discuss different detection models' efficacy in this work.
We brief out the overall study as follows:

- The deep learning-based methodologies are widely used in detecting Deepfake.
- In the experiments, the FF++ dataset contains the largest proportion.
- The deep learning (primarily CNN) models hold a large percentage of all the models.
- The most widely used performance metric is detection accuracy.
- The experimental results demonstrate that deep learning techniques are quite effective in detecting Deepfake. Further, it can be stated that the deep learning models outperform the non-deep learning models.

REFERENCE:

1. FaceApp. Accessed: Jan. 4, 2021. [Online]. Available: https://www.faceapp.com/.
2. X. Zhang, S. Karaman, and S.-F. Chang, ‘‘Detecting and simulating artifacts in GAN fake images,’’ in Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS), Dec. 2019, pp. 1–6.
3. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, ‘‘Densely connected convolutional networks,’’ in Proc. IEEE Conf. Comput. Vis.
4. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243
5. Md Shohel Rana; Mohammad Nur Nobi; Beddhu Murali; Andrew H. Sung, 24 February 2022, doi: https://doi.org/10.1109/ACCESS.2022.3154404.
6. Contributing Data to Deepfake Detection Research. Accessed: Jan. 4, 2021. [Online]. Available: https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html