International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 08 | August - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

A Systematic Literature Analysis on the Fake Identification

Mrs. Nirupama B K [1], Roshan Kumar [2]

[1] Assistant Professor, Department of Master of Computer Application, BMS Institute of Technology and Management, Bengaluru, Karnataka
[2]Student, Department of Master of Computer Application, BMS Institute of Technology and Management, Bengaluru, Karnataka

---

Abstract- Rapid progress in AI, machine learning, and deep learning over the last few decades has resulted in new methodologies and tools for altering multimedia. Though the technology has generally been utilized for respectable causes such as entertainment and education, rogue people have also used it for illegal or evil purposes. High-quality and realistic fake movies, photos, or audios, for example, have been generated to disseminate disinformation and propaganda, incite political division and hatred, or even harass and blackmail people. Deepfake is a term that refers to modified, high-quality, realistic videos. To address the issues posed by Deepfake, many ways have been discussed in the literature. In this study, we undertake a systematic literature review (SLR) to offer an updated overview of the research efforts in Deepfake detection, summarizing 112 relevant papers from 2018 to 2020 that provided a range of techniques. We classify them into four categories: deep learning-based approaches, traditional machine learning-based methods, statistical techniques, and blockchain-based techniques. We also compare the detection capabilities of the various algorithms across different datasets and find that deep learning-based methods outperform other methods in Deepfake detection.

Keyword: Deepfake detection, video or image manipulation, digital media forensics, systematic literature review.

1. INTRODUCTION

The significant improvements in artificial neural network (ANN)-based technologies are critical in tampering with multimedia information. AI-enabled software solutions such as FaceApp and FakeApp, for example, have been utilized for realistic-looking face swapping in photos and videos. This switching technique allows anyone to change their appearance, haircut, gender, age, and other personal characteristics. The spread of these bogus films generates numerous concerns and has earned the nickname Deepfake.

The word "Deepfake" is a combination of "Deep Learning (DL)" and "Fake," and it refers to photo-realistic video or picture material made using DL's assistance. This term was coined in late 2017 by an anonymous Reddit user who used deep learning algorithms to replace a person's face in pornographic movies with another person's face and generate photo-realistic phoney videos. To create such fake films, two neural networks were used:

- A generative network.
- A discriminative network with a FaceSwap approach.

Using an encoder and a decoder, the generative network generates fictitious visuals. The discriminative network determines the veracity of freshly produced pictures. Ian Goodfellow introduced Generative Adversarial Networks (GANs) as a mixture of these two networks.

According to an annual report in Deepfake, DL researchers achieved numerous relevant advances in generative modelling. For example, computer vision researchers devised the Face2Face approach for face re-enactment.
This technology translates facial expressions from a single person to a true digital 'avatar' in real time. CycleGAN was developed by UC Berkeley researchers in 2017 to change photos and videos into various styles.

2. LITERATURE SURVEY

Rapid progress in AI,[1] machine learning, and deep learning over the last few decades has resulted in new methodologies and tools for altering multimedia. Though the technology has generally been utilized for respectable causes such as entertainment and education, rogue people have also used it for illegal or evil purposes. High-quality and realistic fake movies, photos, or audios, for example, have been generated to disseminate disinformation and propaganda, incite political division and hatred, or even harass and blackmail people. Deepfake is a term that refers to modified, high-quality, realistic videos. To address the issues posed by Deepfake, many ways have been discussed in the literature. We perform research to give an up-to-date summary of Deepfake detection studies.

Celebrity pornography [2] is not a new phenomenon. However, in late 2017, a Reddit user entitled Deepfakes began using deep learning to create fake videos of celebrities. This sets off a fresh wave of fraudulent videos on the internet. As part of the US military, DARPA funds research on identifying false movies. Actually, using AI to make films predates Deepfakes. Face2Face and UW's "synthesizing Obama (learning lip sync from audio)" generate even more convincing false videos. Jordan Peele made the video below to alert the public about the threat. This movie was made using Adobe After Effects and FakeApp (a Deepfakes program).

We propose [3] a new framework for estimating generative models through an adversarial process in which we train two models simultaneously: a generative model G that captures the data distribution and a discriminative model D that

estimates the probability that a sample came from the training data rather than G. The training technique for G is designed to increase the likelihood of D making a mistake. This framework relates to a two-player minimax game. A unique solution occurs in the space of random functions G and D, with G recovering the training data distribution and D equal to 1 2 everywhere. Backpropagation may be used to train the whole system when G and D are specified by multilayer perceptron's.

Rapid breakthroughs [4] in AI have resulted in the technology's broad commoditization in recent years, with a wide range of good applications. However, unscrupulous actors are also using this technology for harmful purposes. This paper focuses on one specific dangerous application: the use of artificial intelligence (AI) to make damaging synthetic video, pictures, or sounds, sometimes known as 'deepfakes'.

Face2Face [5] is a unique method for real-time facial reconstruction of a monocular target video sequence (for example, a YouTube video). The source sequence is also a monocular video stream that was collected in real time with a standard camera. Our objective is to animate the target video's facial emotions using a source actor and then re-render the altered output video in a photorealistic manner. To that aim, we first use non-rigid model-based bundling to address the under-constrained problem of face identity recovery from monocular video. At runtime, we use a dense photometric consistency measure to follow the facial expressions of both the source and target footage. Reenactment is therefore accomplished by the rapid and effective transfer of deformation between source and target. The interior of the mouth that best fits the re-targeted expression is retrieved.

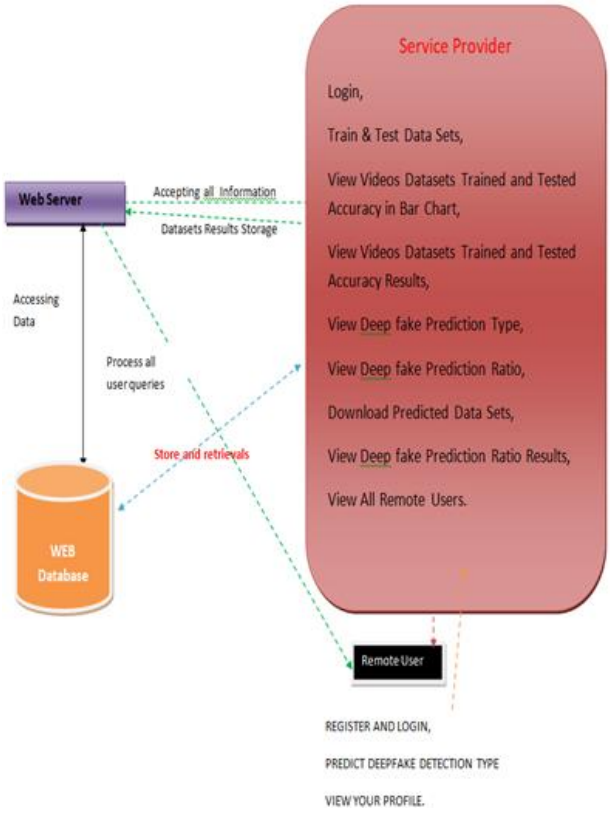

Fig. 1. Proposed Architecture

3. EXISTING WORK

The uniformity of the biological indicators, as well as the spatial and temporal orientations, are monitored in order to employ numerous landmark locations of the face as unique characteristics for certifying the authenticity of GANs generated films or photos. Deepfake films exhibit similar traits, which may be detected by approximating the 3D head posture.

In most situations, face emotions are first coupled with head movements. By leveraging visual artefacts in the face region, Habeeba et al. used MLP to identify Deepfake video with very minimal computational effort. In terms of performance, machine learning-based Deepfake algorithms have been found to achieve up to 98% accuracy in detecting Deepfakes.

When the experiment employs a comparable dataset, the research may acquire a better result by separating it into a given degree of ratio, such as 80% for a train set and 20% for a test set. The unrelated dataset reduces performance by around 50%, which is an unreasonable assumption.

Zhang et al. developed a GAN simulator that duplicates aggregate GAN-image artefacts and feeds them into a classifier to detect Deepfake. presented a network for extracting standard characteristics from RGB data, whereas advocated a comparable but general resolution. Furthermore, researchers presented a novel detection framework based on physiological measurements, such as heartbeat.

Initially, a deep learning-based solution for Deepfake video detection was proposed. To construct their suggested network, they employed two inception modules:

- Meso-4
- MesoInception-4.

The mean squared error (MSE) between the actual and anticipated labels is employed as the loss function for training in this approach. Meso-4 improvement has been proposed in as a binary coded structure, high-dimensional characteristics are not maintained. The data is not stored on a permission-based Blockchain, where the owner has complete control over its contents.

4. PROPOSED METHODOLOGY

We provide a thorough review of the available literature in the Deepfake area. By addressing specific research issues, we report on existing Deepfake detection tools, methodologies, and datasets.

We present an innovative and first-of-its-kind taxonomy that categorizes Deepfake detection algorithms into four categories, with an overview of distinct categories and related properties.

We perform a thorough examination of the experimental data from the original research. In addition, we assess the efficacy of several Deepfake detection algorithms using multiple measurement criteria.

We highlight a few observations and provide some Deepfake detection suggestions that may aid future research and practices in this area.

Establishes a proof of digital content's validity to its trusted source using a generic framework based on Blockchain technology.

The architecture and design specifics of the suggested system for controlling and administering participant interactions and transactions are presented.

Integrates essential IPFS [114]-based decentralized storage capabilities with the Blockchain-based Ethereum Name service.

5. IMPLEMENTATION

A. Service Provider

The Service Provider must login to this module using a valid user name and password. After successfully logging in, he may perform several activities such as Login, Train & Test Data Sets, and so forth. View Deep fake Prediction Type, View Deep fake Prediction Ratio, Download Predicted Data Sets, View Deep fake Prediction Ratio Results, and View All Remote Users.

B. View & Authorize Users

The admin may view a list of all registered users in this module. The admin can examine the user's data such as user name, email, and address, and the admin can authorize the users.

C. Remote User

There are a n number of users in this module. Before doing any activities, the user must first register. When a user registers, their information is saved in the database. After successfully registering, he must login using his authorized user name and password. Once logged in, the user may do actions such as REGISTER AND LOGIN, PREDICT DEEPFAKE DETECTION TYPE, and VIEW YOUR PROFILE.

D. Methodology

- Decision Tree Classifier: Decision tree classifiers are utilized successfully in a wide range of applications. The capacity to capture descriptive decision-making knowledge from provided data is their most essential attribute. Training sets can be used to construct decision trees. The following is the technique for such generation based on a collection of objects (S), each of which belongs to one of the classes C1, C2..., Ck:
  - Step 1: If all of the objects in S belong to the same class, say Ci, the decision tree for S includes a leaf labelled with this class.
  - Step 2: Alternatively, consider T to be a test with potential outcomes O1, O2..., On. Because each item in S has one result for T, the test divides S into subsets S1, S2... Sn, with each object in Si having outcome Oi for T. T becomes the decision tree's root, and for each result Oi, we create a subsidiary decision tree by applying the same technique on the set Si recursively.
- Logistic Regression: The relationship between a categorical dependent variable and a group of independent (explanatory) factors is investigated using logistic regression analysis. When the dependent variable has just two values, such as 0 and 1 or Yes and No, logistic regression is utilized. When the dependent variable contains three or more distinct values, such as Married, Single, Divorced, or Widowed, the term multinomial logistic regression is used. Although the type of data utilised for the dependent variable differs from that of multiple regression, the procedure's practical application is identical.
- Naive Bayes: The naive bayes technique is a supervised learning method that is based on a simple hypothesis: the presence (or absence) of a certain characteristic of a class is independent to the presence (or absence) of any other feature.

Nonetheless, it looks to be sturdy and efficient. Its effectiveness is equivalent to that of other supervised learning approaches. Several explanations have been suggested in the literature. In this lesson, we will focus on an explanation based on representation bias. The naive bayes classifier, like linear discriminant analysis, logistic regression, or linear SVM (support vector machine), is a linear classifier.

6. RESULTS AND PERFORMANCE EVALUATION

A. Naïve Bayes Classifier: The Naive Bayes Classifier is a simple and effective machine learning algorithm commonly used for classification tasks, including DeepFake Detection. It's based on Bayes' theorem and assumes that the features are independent, though it may not hold true for all scenarios.

Naive Bayes can be applied by calculating the probability of a video being real or fake given its features (such as facial expressions, voice characteristics, etc.).

The formula is:

$$P(Real|Features) = (P(Features|Real) * P(Real)) / P(Features)$$

Where:

- P(Real | Features) is the probability of the video being real given its features.
- P(Features | Real) is the probability of observing the features given that the video is real.
- P(Real) is the overall probability of a video being real.
- P(Features) is the probability of observing the given features.

B. Support Vector Machine (SVM): It is a powerful supervised machine learning algorithm used for classification and regression tasks, including DeepFake Detection. SVM works by finding the optimal hyperplane that best separates different classes in a high-dimensional feature space.

SVM can be applied by mapping each video's features (e.g., visual and audio cues) into a higher-dimensional space and finding the hyperplane that maximizes the margin between real and fake videos.

The formula for SVM is based on the following:

minimize: (1/2) * ||w||^2 + C * Σ(max(0, 1 - y_i * (w * x_i + b)))

Where:

- w represents the weight vector of the hyperplane.
- x_i denotes the feature vector of the i-th video sample.
- b is the bias term.
- y_i is the class label of the i-th video sample (+1 for real, -1 for fake).
- C is the regularization parameter that controls the trade-off between maximizing the margin and minimizing the classification errors.

SVM is used to find the values of w and b that minimize the expression while correctly classifying as many training samples as possible and maximizing the margin between the two classes.

C. Logistic Regression: It is a popular supervised machine learning algorithm used for binary classification tasks, including DeepFake Detection. Despite its name, it is used for classification rather than regression. Logistic Regression works by modeling the probability of an input belonging to a specific class, typically between 0 and 1.

Logistic Regression can be applied by utilizing features extracted from videos (e.g., visual and audio characteristics) to predict the probability of a video being real or fake.

The formula for Logistic Regression is as follows:

P(y = 1 | X) = 1 / (1 + e^-(w * X + b))

Where:

- P(y = 1 | X) is the probability of the input X belonging to the positive class (real video).
- w represents the weight vector for the features.
- X denotes the feature vector of the input.
- b is the bias term.

To make predictions, the probability is compared against a threshold (e.g., 0.5), and the input is classified as real or fake based on whether the probability exceeds the threshold or not.

D. Decision Tree Classifier: It is a widely used supervised machine learning algorithm for classification tasks, including DeepFake Detection. It builds a tree-like model by recursively partitioning the feature space into subsets, based on the most informative features, to make accurate predictions.

Decision Tree Classifier can be applied by utilizing various visual and audio features from videos to distinguish between real and fake content.

The formula for Decision Tree Classifier involves the following steps:

- Select the best feature that splits the data into the most distinct classes.
- Split the data based on the chosen feature.
- Repeat the process for each resulting subset until a stopping condition is met (e.g., maximum depth, minimum samples per leaf).

The final decision tree is then used to classify new videos as either real or fake based on their extracted features.

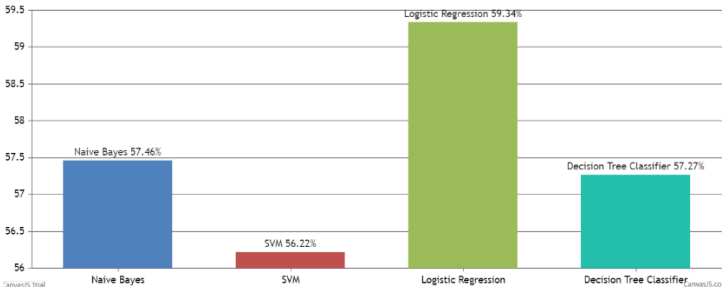

Fig. 2. Accuracy Result of Algorithms

TABLE I. Result Metrics

Sr. No.	Model	Accuracy %
1.	Naive Bayes Classifier	57.46
2.	Support Vector Machine (SVM)	56.22
3.	Logistic Regression	59.34
4.	Decision Tree Classifier	57.27

7. CONCLUSION

This SLR provides numerous cutting-edge strategies for detecting Deep Fake that were published in 112 papers between the beginning of 2018 and the end of 2020. In this paper, we introduce fundamental methodologies and discuss the usefulness of several detection models.

The whole study is summarized as follows:

- Deep learning-based algorithms are frequently employed in identifying Deep Fake. The FF++ dataset accounts for the majority of the data in the tests. Deep learning (mostly CNN) models account for a sizable portion of all models. Detection accuracy is the most often used performance measure.
- Deep learning approaches are successful in identifying Deep fake, according to the experimental data. Furthermore, it may be claimed that deep learning models outperform non-deep learning models in general. Deep fake detection still confronts numerous hurdles, despite significant advancements in underlying multimedia technology and the availability of tools and applications. We believe that this SLR will be useful to the research community in creating effective detection methods and countermeasures.

REFERENNCES

[1] G. Oberoi. Exploring DeepFakes. Accessed: Jan. 4, 2021. [Online]. Available: https://goberoi.com/exploring-deepfakes-20c9947c22d9

[2] J. Hui. How Deep Learning Fakes Videos (Deepfake) and How to Detect it. Accessed: Jan. 4, 2021. [Online]. Available: https://medium.com/how-deep-learning-fakes-videos-deepfakes-and-how-to-detect-itc0b50fbf7cb9

[3] I. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS), vol. 2. Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680.

[4] G. Patrini, F. Cavalli, and H. Ajder, "The state of deepfakes: Reality under attack," Deeptrace B.V., Amsterdam, The Netherlands, Annu. Rep. v.2.3., 2018. [Online]. Available: https://s3.eu-west2.amazonaws.com/rep2018/2018-the-state-of-deepfakes.pdf

[5] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, "Face2Face: Real-time face capture and reenactment of RGB videos," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 2387–2395, doi: 10.1109/CVPR.2016.262.

[6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Oct. 2017, pp. 2242–2251, doi: 10.1109/ICCV.2017.244.

[7] S. Suwajanakorn, S. M. Seitz, and I. K. Shlizerman, "Synthesizing Obama: Learning lip sync from audio," ACM Trans. Graph., vol. 36, no. 4, p. 95, 2017.

[8] L. Matsakis. Artificial Intelligence is Now Fighting Fake Porn. Accessed: Jan. 4, 2021. [Online]. Available: https://www.wired.com/story/gfycatartificial-intelligence-deepfakes/

[9] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics: A large-scale video dataset for forgery detection in human faces," 2018, arXiv:1803.09179.

H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, ''Deep video portraits,'' ACM Trans. Graph., vol. 37, no. 4, pp. 1–14, Aug. 2018, doi: 10.1145/3197517.3201283.