Engineering and Technology Journal e-ISSN: 2456-3358

Volume 10 Issue 04 April-2025, Page No.- 4535-4543

DOI: 10.47191/etj/v10i04.20, I.F. – 8.482

A Comprehensive Review of Deepfake Detection Pertaining to Images, Videos, Audio, and News using Deep Learning Techniques

Vakdevi Vallabhaneni[1], T. Dheeraj[2], B. Chandra Sekhar[3], Ch. Srinivas[1], Md. Ibrahim[1], S. Eswar N. V. S. P[1], Ch. Balamani[1], V.V.R Swamy[1]

[1]Department of Computer Science and Engineering, Lingayas Institute of Management and Technology, Vijayawada-521212, India

[2]Software Developer at Uzvi Service, Software Technology Parks of India, Vijayawada, Andhra Pradesh 520008, India

[3]Department of Chemical Engineering, RGUKT RK Valley, Iddupulapaya, Vempalli, YSR Kadapa, India

ABSTRACT: Deepfakes, which are synthetic media realistic in nature generated using artificial intelligence (AI); pose a significant threat to individuals and society. The rapid advancement of deepfake technology has led to the creation of highly realistic synthetic content covering images, videos, audio, and news. While deepfake applications offer creative possibilities, their misuse for misinformation, identity fraud, and cybersecurity threats necessitates robust detection methods. Deepfake crimes are rising daily, wherein deepfake media detection has become a big challenge and has high claim in digital forensics. This review explores the state-of-the-art deep learning (DL) techniques for deepfake detection of four parameters, namely images, videos, audio, and news. The ML approaches rely on handcrafted features but struggle with evolving deepfake methods. In contrast, DL techniques, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have demonstrated superior detection accuracy by learning discriminative features. Even Recurrent Neural Networks (RNNs), and Transformer-based architectures like Bidirectional encoder representations from transformers (BERT), have demonstrated superior accuracy in identifying manipulated content. Furthermore, recent advancements such as Vision Transformers (ViTs) and Explainable AI (XAI) models are enhancing detection interpretability and robustness. This review highlights the future research directions for strengthening deepfake detection mechanisms. The rapid advancements in deepfake generation necessitate continuous research and development of countermeasures.

KEYWORDS: deepfake, deep learning, misinformation, emerging technologies, dataset

1. INTRODUCTION

The proliferation of deepfake technology has raised significant concerns regarding digital media authenticity. Deepfake images and videos are two categories, generated using deep learning models like generative adversarial network (GANs) and variational autoencoders (VAEs), which produce hyper-realistic synthetic content that is difficult to distinguish from real images (Goodfellow et al., 2014). The rise of deepfakes has led to threats in areas such as political misinformation, biometric security, and privacy infringement (Nguyen et al., 2020). Therefore, developing effective detection mechanisms is crucial to mitigating their risks. The GANs consist of two neural networks: a generator that creates synthetic media and a discriminator that distinguishes between real and fake content. Through adversarial training, the generator learns to produce increasingly realistic deepfakes, while the discriminator improves its ability to detect them. Autoencoders, on the other hand, learn compressed representations of data, which can be used to generate new instances. These techniques are constantly evolving, leading to increasingly sophisticated deepfakes that are difficult to detect (Abdulreda and Obaid, 2022).

Third category, deepfake news refers to artificially generated or manipulated news content designed to deceive audiences by presenting false or misleading information as factual. The emergence of sophisticated AI models, such as GANs and autoregressive transformers, has enabled the seamless creation of fake news articles and multimedia content (Zellers et al., 2019). The proliferation of deepfake news has far-reaching consequences, including influencing political elections, manipulating public perception, and spreading misinformation in crises (Shu et al., 2020). Fourth category, deepfake audio refers to artificially generated or manipulated speech that mimics a real person's voice with high accuracy. The emergence of deep learning models, such as WaveNet, GAN-based speech synthesis, and text-to-speech (TTS) systems, has enabled the seamless generation of synthetic audio content (Oord et al., 2016). The spread of deepfake audio has far-reaching consequences, including impersonation fraud, misinformation campaigns, and potential threats to national security (Kreuk et al., 2020). Given these threats, robust detection mechanisms leveraging

ML and DL models are crucial to mitigating the risks associated with deepfake audio.

2. INSTRUMENTAL METHODS FOR DEEPFAKE DETECTION

2.1.1. Approaches based on Deep Learning

Deep learning models have significantly improved deepfake detection by automatically extracting high-dimensional features from images. Convolutional Neural Networks (CNNs) have been widely used for deepfake image classification due to their ability to detect spatial inconsistencies and subtle artifacts in synthetic images (Chollet, 2017). Popular CNN architectures include XceptionNet, which demonstrated high accuracy in deepfake detection tasks by leveraging depthwise separable convolutions (Nguyen et al., 2020), whereas ResNet and EfficientNet is employed for multi-scale feature extraction and robust classification. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) are employed to analyze sequential dependencies in deepfake images and videos. They are particularly effective in detecting temporal inconsistencies such as flickering or unnatural facial expressions (Guo et al., 2021). The GANs are primarily used to create deepfakes, they can also be leveraged for detection by training adversarial models to differentiate real and synthetic images. Adversarial training improves robustness against evolving deepfake generation techniques (Tolosana et al., 2020).

The CNNs have been adapted for text-based tasks by identifying distinguishing patterns in word embeddings (Zhang et al., 2015). The CNN-based classifiers have demonstrated strong performance in detecting textual manipulations. The RNNs and LSTMs capture sequential dependencies in text, making them well-suited for analyzing linguistic patterns in fake news articles. Transformer architectures, such as BERT and Generative Pre-training Transformer (GPT), have significantly advanced the field of fake news detection by leveraging attention mechanisms to capture contextual relationships across words (Devlin et al., 2019). Pretrained language models fine-tuned on fake news datasets have achieved state-of-the-art performance in detecting manipulated content.

The CNNs have been applied to spectrogram representations of speech, enabling the detection of anomalies in frequency patterns (Tak et al., 2021). The CNN-based classifiers have shown strong performance in detecting inconsistencies introduced by synthetic speech models. The RNNs and LSTMs are well-suited for sequential data processing and have been used to analyze temporal dependencies in speech patterns. These models help detect unnatural fluctuations in prosody and speech rhythm (Jia et al., 2018). Transformer architectures, such as wav2vec and SpeechT5, have advanced deepfake audio detection by leveraging self-attention mechanisms to capture long-range dependencies in speech signals (Baevski et al., 2020). Pretrained models fine-tuned on deepfake datasets have achieved state-of-the-art performance in synthetic speech detection.

2.1.2. Approaches based on Machine Learning

Machine learning techniques were among the earliest approaches to deepfake detection, primarily relying on feature extraction and classification models. Traditional ML techniques use handcrafted features such as texture descriptors, frequency domain analysis, and facial landmarks to identify inconsistencies in deepfake images (Agarwal et al., 2020). Handcrafted features, including lexical, syntactic, and semantic cues, have been widely used in ML-based fake news detection (Rubin et al., 2016). Term Frequency-Inverse Document Frequency (TF-IDF) measures the importance of words in a document relative to a corpus, sentiment analysis identifies emotional cues and biases in text and readability scores evaluates complexity using metrics such as Flesch-Kincaid readability tests. Other handcrafted features, such as spectral and prosodic cues, have been widely used in ML-based audio deepfake detection. Mel-Frequency Cepstral Coefficients (MFCCs) captures the spectral properties of speech signals, Linear Predictive Coding (LPC) models the vocal tract's characteristics and Pitch and Formant Analysis identifies unnatural variations in voice modulation. Commonly used feature descriptors include, Local Binary Patterns (LBP), which captures texture variations in facial images (Li et al., 2019), histogram of Oriented Gradients (HOG) detects edge-based inconsistencies and wavelet Transforms analyzes frequency domain artifacts in synthetic images. Once features are extracted, classification models such as support vector machine (SVMs), decision trees, and ensemble methods like Random Forests are applied to distinguish between real and fake images. However, handcrafted feature-based methods often struggle with generalizing to new deepfake architectures due to the rapid evolution of synthesis techniques.

2.2.3. Approaches based on Emerging Techniques (ET)

As DL models become increasingly complex, their interpretability is crucial for enhancing trust and reliability. Explainable AI (XAI) techniques, such as attention visualization and Shapley Additive Explanations (SHAP), provide insights into how AI models make predictions, improving user trust in automated fake news detection (Arrieta et al., 2020). Layer-wise relevance propagation (LRP), provide insights into model decision-making, enhancing reliability in deepfake detection (Arrieta et al., 2020). The XAI techniques help in understanding how models differentiate real from fake images, increasing trust in AI-based detection systems (Rai et al., 2022). Given that deepfake news often incorporates manipulated images and videos, multimodal detection models that integrate textual and visual features are gaining traction. Techniques such as Vision-Language Transformers (e.g., CLIP) can analyze both textual and visual cues to enhance deepfake news detection

accuracy (Radford et al., 2021). Recent research has explored Vision Transformers (ViTs) for deepfake detection, as they capture long-range dependencies in images more effectively than CNNs (Dosovitskiy et al., 2021). The ViTs have demonstrated promising performance in detecting deepfake artifacts that CNNs may overlook. Deepfake content often combines manipulated audio with video, multimodal detection models integrating speech and facial expressions are gaining traction. Models such as Audio-Visual Transformers (e.g., AV-Hubert) can analyze both speech and lip movements to enhance detection accuracy (Shi et al., 2022). Table 1 presents the overview of the few results reported on the detection of deepfake videos, image, audio and text.

Table 1. Overview of the few results reported on the detection of deepfake videos, image, audio and text.

S No	Title/description	Methods employed	Datasets used	Findings	Reference
1	Deepfake: A Survey on Facial Forgery Technique Using Generative Adversarial Network	CNN and LSTM	Face2Face, Reddit user deepfakes	95%	(Yadav and Salmani, 2019)
2	Presenting a temporal aware method for detecting automatically.	CNN and RNN	Deepfake-TIMIT and FaceForensics++	High accuracy	(Güera and Delp, 2018)
3	Using a pre-trained CNN to collect face characteristics to extract hidden features.	CNN	Celeb-DF Dataset	High accuracy	(Karandikar et al., 2020)
4	Providing a SCNN approach for detecting face modification techniques deepfakes.	CNN	Deepfake-TIMIT dataset, FaceForensics++ dataset, and DFDC Preview dataset	High accuracy	(Xu et al., 2021)
5	Developing a frequency-based CNN method.	Frequency CNN	FaceForensics++ and Celeb-DF (v2) dataset	Strong robustness	(Kohli and Gupta, 2021)
6	Proposing feature transfer, a technique based on unsupervised domain adaptation.	CNN + BP-DANN	Deepfake-TIMIT + FaceForensics++	Solved the overfitting problem	(Chen and Tan, 2021)
7	Using a binary classifier trained by a CNN.	CNN	Milborrow University of Cape Town dataset	97%	(Zhang et al., 2020)
8	Presenting an expectation–maximization method trained to identify and extract a fingerprint.	Expectation–maximization + CNN	The FACE APP, a dataset, and CELEBA images	93%	(Guarnera et al., 2020)
9	Suggesting an approach with an effective end to end false face detection pipeline that can identify fake face pictures.	GAN	HFM dataset	72.50%	(Lee et al., 2021)
10	Using the image saliency to determine the texture depth and pixel difference between actual and fake facial images.	CNN + simple linear iterative clustering	Faceforensics++ dataset	99%	(Yang et al., 2021)

11	FaceForensics++: Learning to Detect Manipulated Facial Images	1.Xception net (CNN) 2.LSTM	FaceForensics++	81%	(Rossler et al., 2019)
12	Proposing a deep CNN-based visual speaker authentication method.	Deep CNN	GRID dataset and MOBIO dataset	High accuracy	(Yang et al., 2020)
13	Using the similarities between audiovisual modalities and the similarity between affective cues.	RNN + LSTM+	TIMIT and DFDC dataset	84.40%	(Mittal et al., 2020)
14	Suggesting a system that can differentiate between actual and synthetic speech in a group conversation.	CNN + RNN	FakeOrReal dataset + the AMI Corpus dataset	NLP showed 93% accuracy in conversion. RNN model showed 80% accuracy in speaker labeling.	(Wijethunga et al., 2020)
15	Providing a collection of characteristics built on the notion of describing the speech as an auto-regressive mechanism.	LSTM	The ASVSpoof 2019 logical access audio dataset	High accuracy	(Borrelli et al., 2021)
16	Exploiting a neural attention on top of the LSTM to fuse text features, social context and image features.	Att-RNN	Two multimedia rumor datasets collected from Weibo and Twitter	High accuracy	(Jin et al., 2017)
17	Newly proposed Fake News Detection by comprehensively mining the Semantic Correlations between Text content and Images attached (FND-SCTI) is able to effectively capture the semantic correlations across multimodalities, and achieves the state-of-the-art performance.	Fake News Detection by comprehensively mining the Semantic Correlations between Text content and Images attached	Twitter and Weibo datasets	outperforms other competitive approaches	(Zeng et al., 2021)
18	This study proposed an ensembling of Deep learning and Transformer based feature to identify fake news. Proposed architecture is experimented on 3 public datasets out of which , it outperfromed on two datasets.	CNN, BiLSTM, Multilayer perception, GloVe and Transformer BERT	publicly available datasets	f1-score of 96.03% and 74.24%	(Reddy et al., 2024)
19	This work proposed a novel hybrid deep learning model that combined convolutional and recurrent neural networks for fake news classification. Further experiments on the generalization	A hybrid CNN-RNN	ISOT and FA-KES	Promising results	(Nasir et al., 2021)

	of the proposed model across different datasets, had promising results.				
20	The proposed study used CNN, Bidirectional LSTM, and ResNet combined with pre-trained word embedding, trained using four different datasets.	CNN, Bidirectional LSTM, and ResNet	Four different datasets.	The results showed that the Bidirectional LSTM architecture outperforme d CNN and ResNet on all tested datasets.	(Sastrawan et al., 2022)

As was cited in Table 1, Yadav and Salmani (2019) reported 95% accuracy in deepfake video detection making use of CNN and LSTM with Face2Face, Reddit user deepfakes datasets. In a study conducted by Guera and Delp (2018b), RNN and CNN were employed with Deepfake-TIMIT and FaceForensics++ as datasets and reported high accuracy results. Karandikar et al. (2020) reported high accuracy in detecting deepfake videos using a pre-trained CNN and Celeb-DF Dataset to collect face characteristics to extract hidden features. Kohli and Gupta (2021) reported strong robustness in deepfake video detection making use of a frequency-based CNN method FaceForensics++ and Celeb-DF (v2) dataset. A CNN and backpropagation based on domain adversarial neural network model was proposed by Chen and Tan (2021) for feature transfer, a technique based on unsupervised domain adaptation using Deepfake-TIMIT +FaceForensics++ dataset and effectively solved the overfitting problem. A binary classifier trained by a CNN was used for deepfake video detection on Milborrow University of Cape Town dataset and reported 97% accuracy (Zhang et al., 2020). An expectation–maximization+ CNN method was employed for deepfake video detection making use of the FACE APP, a dataset, and CELEBA images and reported 93% accuracy (Guarnera et al., 2020).

Lee et al. (2021) made use of GAN model with HFM dataset and proved that this approach is an effective end to end false face detection pipeline that can identify fake face pictures with 72.5% accuracy. Yang et al. (2021) employed CNN + simple linear iterative clustering and Faceforensics++ as dataset with image saliency to determine the texture depth and pixel difference between actual and fake facial images and yielded 99% accuracy. Rossler et al. (2019) made use of Xception net (CNN) and LSTM to detect manipulated facial images with FaceForensics++ as dataset and yielded 81% accuracy. A deep CNN model was employed with GRID dataset and MOBIO dataset as datasets and yielded high accuracy (Yang et al., 2020). A combination of RNN and LSTM was used with TIMIT and DFDC datasets using the similarities between audiovisual modalities and the similarity between affective cues and resulted in 84.4% (Mittal et al., 2020).

Wijethunga et al. (2020) suggested a system that differentiated between actual and synthetic speech in a group conversation using CNN and RNN with FakeOrReal dataset + the AMICorpus datasets, wherein NLP showed 93% accuracy in text conversion and RNN model showed 80% accuracy for speaker labeling. Borrelli et al. (2021) proposed a collection of characteristics built on the notion of describing the speech as an auto-regressive mechanism using LSTM with ASVSpoof 2019 logical access audio dataset and yielded high accuracy. Jin et al. (2017) proposed a novel RNN with an attention mechanism (att-RNN) to fuse multimodal features for effective rumor detection with two multimedia rumor datasets collected from Weibo and Twitter and demonstrated the effectiveness of the proposed end-to-end att-RNN in detecting rumors with multimodal contents.

Zeng et al. (2021) proposed fake news detection by comprehensively mining the semantic correlations between text content and images attached (FND-SCTI) with Twitter and Weibo datasets, which was able to effectively capture the semantic correlations across multimodalities, and achieves the state-of-the-art performance. Reddy et al. (2024) made use of CNN, BiLSTM, Multilayer perception, GloVe and Transformer BERT with publicly available datasets to identify fake news. Proposed architecture is experimented on three public datasets out of which, it outperformed on two datasets with f1-score of 96.03% and 74.24%. Nasir et al. (2021) proposed a novel hybrid CNN-RNN for fake news classification with ISOT and FA-KES datasets and yielded promising results. Sastrawan et al. (2022) used CNN, Bidirectional LSTM, and ResNet combined with pre-trained word embedding, trained with four different datasets. The results proved that the Bidirectional LSTM architecture outperformed CNN and ResNet on all tested datasets. Xu et

al. (2021) introduced a novel Set CNN (SCNN) framework for detecting facial alterations in videos. This framework treats the facial features from multiple video frames as a set for analysis. The authors demonstrated the SCNN using three example backbone networks: t-MesoNet, and two variants of t-XceptionNet. Their experiments showed that the SCNN outperforms existing methods. Future improvements, they suggest, could focus on employing stronger backbone networks and developing more effective set reduction techniques. The current set reduction methods, while effective, are relatively simple.

Sharma et al. (2024) employed a unique active forensic strategy Compact Ensemble based discriminators architecture using Deep Conditional Generative Adversarial Networks (CED-DCGAN), for ascertaining real-time deep fakes in video conferencing. The DCGAN focused on video-deep fake detection on features since technologies for creating convincing fakes are improving rapidly. The proposed model tested on publicly available datasets showed that the proposed algorithms outperform state-of-the-art methods and the proposed CED-DCGAN technique successfully detected high-fidelity deep fakes in video conferencing. Python was employed for implementing the proposed study and 98.23% accuracy was obtained (Sharma et al., (2024).

Several studies have explored modeling relationships between news posts using various neural network architectures. Ma et al. (2016) employed recurrent neural networks (RNNs) to capture sequential relationships. Yu et al. (2017) utilized convolutional neural networks (CNNs) to represent high-level semantic relationships. Bian et al. (2020) leveraged graph neural networks (GNNs) with a directed graph to model rumor propagation and dispersion. Finally, Khattar et al. (2019) proposed a multi-modal variational autoencoder (MVAE) to extract hidden representations from multimedia news.

Two approaches have been proposed to address multi-modal feature extraction for fake news detection. Xue et al. (2021) projected visual and textual features into a shared feature space using weight-sharing encoders, calculating similarity in this transformed space. However, this method struggles to capture multi-modal inconsistencies arising from the semantic gap between visual and textual data. To overcome this limitation, Ghorbanpour et al. (2021) introduced the Fake News Revealer (FNR) method. FNR leverages a Vision Transformer (Dosovitskiy et al., 2020) and BERT (Devlin et al., 2019) for independent visual and textual feature extraction, respectively. The similarity between image and text features is then determined using contrastive loss, directly addressing the challenges posed by the semantic gap.

Nguyen et al. (2021) proposed a deepfake video detection method using a 3D CNN. Their approach involves generating 3D representations of short video sequences to capture spatio-temporal characteristics. A deep 3D CNN, employing 3D convolutional kernels, learns features across both spatial and temporal dimensions. Crucially, the feature maps in successive convolutional layers are connected, allowing the network to integrate information from consecutive frames. This design enables effective learning of facial features across both space and time. The proposed model achieved over 99% accuracy on the FaceForensic and VidTIMIT datasets, demonstrating its effectiveness in detecting deepfake videos.

Jung et al. (2020) developed a deepfake detection method that analyzes variations in eye-blinking patterns. Recognizing that blinking frequency and patterns vary naturally based on factors such as gender, age, and cognitive activity, their system uses ML and heuristic rules to identify inconsistencies indicative of deepfakes. This multi-method approach, informed by prior research, achieved an 87.5% success rate in correctly identifying deepfakes in a test set of eight videos.

Yan et al. (2021) addressed the challenge of detecting deepfakes in compressed videos, a common format on social media platforms like Instagram, WeChat, and TikTok. They proposed a two-stream network architecture. A frame-level stream, employing a low-complexity network and model pruning, extracts features while mitigating compression artifacts. A separate temporal-level stream analyzes inconsistencies between frames to capture temporal characteristics. This combined approach effectively leverages both frame-level and temporal information from compressed videos. Evaluated on FaceSwap, Face2Face, NeuralTextures, and Celeb-DF datasets, the method outperformed existing techniques, demonstrating robustness to varying compression levels.

Yu et al. (2020) proposed a lip-based visual speaker authentication system (SA-DTH-Net) to detect deepfakes. This method leverages the fact that deepfake creators often lack sufficient information about a target individual's speaking habits to perfectly replicate them, particularly when speaking impromptu text. SA-DTH-Net extracts unique speaking patterns from lip movements to distinguish genuine speakers from imposters and deepfakes. By aggregating word-level authentication results, the system provides a final authentication decision. Evaluations demonstrated the system's effectiveness in rejecting deepfakes generated using various manipulation techniques, suggesting its potential for broader application in visual speaker authentication (VSA) systems.

Mittal et al. (2020) presented a learning-based deepfake detection method that analyzes both audio and visual modalities within a video. Their approach compares audio-visual consistency and extracts features related to perceived emotion to distinguish genuine from fake content. Motivated by Siamese networks and triplet loss, they developed a deep learning model that achieved Area Under the Curve (AUC) scores of 84.4% on the DFDC dataset and 96.6% on the DF-TIMIT dataset. This work is notable for its integration of both

audio and visual cues, along with emotional analysis, for deepfake detection.

Wang et al. (2020) introduced DeepSonar, a novel deepfake audio detection method that analyzes the internal activations of a voice synthesis neural network. This approach focuses on detecting subtle differences in neuron activation patterns between genuine and AI-generated speech. The method prioritizes robustness to various voice manipulations, which are often more easily disguised than image manipulations. Evaluations on three datasets (Chinese and English) demonstrated high detection rates (98.1% average accuracy) and low false alarm rates (approximately 2%), highlighting DeepSonar's effectiveness and robustness, even in noisy environments.

3. KEY MISSING GAPS AND FUTURE PROPOSALS

Despite several advancements in deepfake detection, several challenges remain, which include many detection models rely on datasets that may not generalize well to new deepfake techniques, which result in dataset biasing; deepfake generation models continue to evolve, making detection harder owing to combative attacks; existing detection methods need to improve their robustness across diverse applications and lighting conditions as it is close to real-world scenarios. There is a need for the future research focusing on developing hybrid models that integrate CNNs, RNNs, and transformers, improving model interpretability, and creating more diverse benchmark datasets to enhance generalization capabilities. Based on the above cited literature review, the future research work is proposed to making use of a combination of DL techniques, ResNext-50+MesoNet+gated recurrent unit (GRU) and ML techniques, SVM + decision tree (DT) + random forests (RF) to enhance the accuracy of the deepfake video detection. This study also proposes making use of a combination of CNN, RNN, GAN, LSTM and XAI for deepfake image detection.

4. CONCLUSION

Deepfake image detection has progressed significantly, transitioning from ML methods to sophisticated DL architectures, with CNNs and RNNs currently dominating the field. Recent advancements, including ViTs and XAI are enhancing both model accuracy and interpretability. Future progress centers on integrating XAI techniques and multimodal approaches to improve deepfake detection in images, videos, and audio. This study proposes several promising research directions, which include hybrid approach: combining DL architectures (ResNext-50, MesoNet, and GRU) with ML algorithms (SVM, DT and RF) to potentially improve deepfake detection accuracy. Another is multimodal DL integration, which include exploring the synergistic potential of CNNs, RNNs, GANs, and LSTMs alongside XAI for enhanced deepfake image detection. Addressing critical challenges, such as adversarial attacks and dataset biases, is vital for creating reliable and robust deepfake detection frameworks. There is a need for persistent research and interdisciplinary collaboration, which are essential to effectively mitigate the risks posed by deepfake technology.

REFERENCES

1. Agarwal, S., Singh, A., Singh, R. 2020. Detecting deepfake videos using frequency domain analysis. Journal of AI Research, 45(2), 112–126.
2. Ahmed S. Abdulredaa., Ahmed J. Obaida. 2022. A landscape view of deepfake techniques and detection methods. Int. J. Nonlinear Anal. Appl. 13, No. 1, 745-755. http://dx.doi.org/10.22075/ijnaa.2022.5580
3. Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, Michael Auli. 2020. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. NeurIPS. NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, Pages 12449 - 12460
4. Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images"Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1-11.
5. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012
6. Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y., Huang, J. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34, (01), pp. 549–556.
7. Borrelli, C., Bestagini, P., Antonacci, F., Sarti, A., Tubaro, S. 2021. Synthetic speech detection through short-term and long-term prediction traces. EURASIP Journal on Information Security, (1), 1–14.
8. Chen, B., Tan, S. 2021. FeatureTransfer: Unsupervised domain adaptation for cross-domain deepfake detection. Security and Communication Networks, 2021, 1–8.
9. Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. IEEE Transactions on Pattern Analysis and Machine

Intelligence, 39(12), 2551–2566. https://doi.org/10.1109/TPAMI.2016.2626660

10. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 4171–4186.
11. Digvijay Yadav, Sakina Salmani. 2019. Deepfake: A Survey on Facial Forgery Technique Using Generative Adversarial Network, Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2019). IEEE Xplore Part Number: CFP19K34-ART; ISBN: 978-1-5386-8113-8.
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16×16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations.
13. Ghorbanpour, F., Ramezani, M., Fazli, M.A., Rabiee, H.R. 2021. FNR: A similarity and transformer-based approachto detect multi-modal FakeNews in social media. arXiv preprint arXiv:2112.01131.
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. 2014. Generative adversarial nets. Advances in neural information processing systems, 27.
15. Guarnera, L., Giudice, O., Battiato, S. 2020. Fighting deepfake by exposing the convolutional traces on images. IEEE Access, 8, 165085–165098.
16. Güera, D., Delp, E. J. 2018. Deepfake video detection using recurrent neural networks. Paper presented at the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS).
17. Guo, J., Liu, X., Zhang, D. 2021. Recurrent neural networks for deepfake video detection. Neural Computing and Applications, 33(8), 4201–4215. https://doi.org/10.1007/s00521-020-05487-7
18. Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. 2021. End-to-End Audio Deepfake Detection: Waveform or Spectrogram? Interspeech.
19. Nasir, J.A., Khan O.S., Varlamis. I. 2021. Fake news detection: A hybrid CNN-RNN based deep learning approach. International Journal of Information Management Data Insights 1, 100007
20. Jaiwanth Reddy, Shikha Mundra, Ankit Mundra. 2024. Ensembling Deep Learning Models for Fake News Classification. Procedia Computer Science 235, 2766–2774.
21. Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Nguyen, P., Pang, R., Lopez Moreno, I. and Wu, Y. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. Advances in neural information processing systems, 31.
22. Jiangfeng Zeng, Yin Zhang, Xiao Ma. 2021. Fake news detection for epidemic emergencies via deep correlations between text and images. Sustainable Cities and Society 66, 102652
23. Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J., 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM International Conference on Multimedia. pp. 795–816.
24. Jung, T., Kim, S., Kim, K. 2020. DeepVision: Deepfakes detection using human eye blinking pattern. IEEE Access, 8, 83144–83154.
25. Kadek Sastrawan, I.P.A. Bayupati, Dewa Made Sri Arsa. 2022. Detection of fake news using deep learning CNN–RNN based methods, ICT Express, Volume 8, Issue 3, 2022, 396-408, https://doi.org/10.1016/j.icte.2021.10.003.
26. Kai Shu, A Amy Sliva, Suhang Wang, Jiliang Tang, Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective V 19, 22-36. https://doi.org/10.1145/3137597.3137600
27. Karandikar, A., Deshpande, V., Singh, S., Nagbhidkar, S., & Agrawal, S. 2020. Deepfake video detection using convolutional neural network. International Journal of Advanced Trends in Computer Science and Engineering, 9(2), 1311–1315.
28. Khattar, D., Goud, J.S., Gupta, M., Varma, V., 2019. Mvae: Multimodal variational autoencoder for fake news detection. In: Proceedings of the International World Wide Web Conferences. pp. 2915–2921.
29. Kohli, A., Gupta, A. 2021. Detecting DeepFake, FaceSwap and Face2Face facial forgeries using frequency CNN. Multimedia Tools and Applications, 80(12), 18461–18478.
30. Lee, S., Tariq, S., Shin, Y., Woo, S. S. 2021. Detecting handcrafted facial image manipulations and GAN-generated facial images using shallow-FakeFaceNet. Applied Soft Computing, 105, 107256.
31. Li, Y., Lyu, S. 2018. Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656.
32. Li, Y., Li, H., Wang, X. 2019. Exposing deepfake videos by detecting face warping artifacts. In Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition (pp. 11960–11969).
33. Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.F., Cha, M. 2016. Detecting rumors from microblogs with recurrent neural networks. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 3818–3824.
34. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D. 2020. Emotions don't lie: An audio-visual deepfake detection method using affective cues. Paper presented at the Proceedings of the 28th ACM International Conference on Multimedia.
35. Nguyen, X. H., Tran, T. S., Nguyen, K. D., Truong, D. T. 2021. Learning spatio-temporal features to detect manipulated facial videos created by the Deepfake techniques. Forensic Science International: Digital Investigation, 36, 301108.
36. Oord, A.V.D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K., 2016. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR.
38. Rai, A., Singh, A., Singh, R. 2022. Explainable AI in deepfake detection: Challenges and opportunities. AI Ethics, 5(3), 199–215. https://doi.org/10.1007/s43681-022-00160-7
39. Rubin, V. L., Conroy, N. J., Chen, Y. & Cornwell, S. (2016). Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. Proceedings of NAACL-HLT, p. 7–17.
40. Shi, B., Hsu, W. N., Lakhotia, K., Mohamed. A. 2022. AV-HuBERT: Self-Supervised Speech Representation Learning by Audio-Visual Multi-Modal Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence.
41. Tolosana, R., Vera-Rodriguez, R., Fierrez, J. 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131–148. https://doi.org/10.1016/j.inffus.2020.07.004
42. Wang, R., Juefei-Xu, F., Huang, Y., Guo, Q., Xie, X., Ma, L., Liu, Y. 2020. Deepsonar: Towards effective and robust detection of AIsynthesized fake voices. Paper presented at the Proceedings of the 28th ACM International Conference on Multimedia. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.
43. Wijethunga, R., Matheesha, D., Al Noman, A., De Silva, K., Tissera, M., & Rupasinghe, L. 2020. Deepfake audio detection: A deep learning based solution for group conversations. Paper presented at the 2020 2nd International Conference on Advancements in Computing (ICAC).
44. Xiang Zhang, Junbo Jake Zhao, Yann LeCun. 2015. Character-level convolutional networks for text classification. In Proc. NeurIPS, 649–657.
45. Xu, Z., Liu, J., Lu, W., Xu, B., Zhao, X., Li, B., Huang, J. 2021. Detecting facial manipulated videos based on set convolutional neural networks. Journal of Visual Communication and Image Representation, 77, 103119.
46. Xue, J., Wang, Y., Tian, Y., Li, Y., Shi, L., Wei, L., 2021. Detecting fake news by exploring the consistency of multimodal data. Inf. Process. Manage. 58 (5), 102610
47. Yang, C.-Z., Ma, J., Wang, S.-L., Liew, A. W.-C. 2020. Preventing deepFake attacks on speaker authentication by dynamic lip movement analysis. IEEE Transactions on Information Forensics and Security, 16, 1841–1854.
48. Yang, J., Xiao, S., Li, A., Lan, G., Wang, H. 2021. Detecting fake images by identifying potential texture difference. Future Generation Computer Systems, 125, 127–135.
49. Yu, F., Liu, Q., Wu, S., Wang, L., Tan, T. 2017. A convolutional approach for misinformation identification. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. pp. 3901–3907.
50. Yu, P., Xia, Z., Fei, J., Lu, Y. 2021. A survey on Deepfake video detection. IET Biometrics, 10(6) 607–624.
51. Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., Choi, Y. 2019. Defending against neural fake news. Advances in neural information processing systems, 32.
52. Zhang, W., Zhao, C., Li, Y. 2020. A novel counterfeit feature extraction technique for exposing face-swap images based on deep learning and error level analysis. Entropy, 22(2), 249.