International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Deep Learning Framework for Detecting Synthetic Audio-Visual Media

Greeshma Chandu A.I[1], Arathi Chandran R.I[2]

[1,2]Department of Computer Science, Christ Nagar College, Maranalloor, Thiruvananthapuram, Kerala, India

Abstract:
The rapid advancement of deepfake technology has introduced significant challenges to digital forensics, especially within law enforcement agencies. Deepfakes— manipulated audio, video, and image content that appears authentic but is entirely fabricated—pose serious risks to investigations, public trust, and security. To address this growing concern, the development of an integrated software solution for deepfake detection in audio, video, and image formats is crucial for cyber police departments. This software will utilize advanced Machine Learning algorithms, Artificial Intelligence, and forensic analysis techniques to identify signs of tampering and manipulation across various forms of media. The proposed software will feature a multi-layered detection system capable of analyzing pixel abnormalities and auditory cues. Leveraging Deep Learning and Neural Networks, the software will be trained on large datasets to accurately detect deepfake patterns and differentiate them from genuine media. By using deep learning models like CNNs for visual feature extraction and RNNs for audio feature extraction, this approach improves the detection of inconsistencies in deepfake videos, making it an effective solution in the fight against misinformation. Additionally, the system will integrate with existing digital forensics tools to support police investigations, allowing officers to quickly verify the authenticity of digital evidence. By providing police departments with cutting-edge detection capabilities, this solution aims to combat the misuse of deepfake technology in criminal activities such as fraud, identity theft, blackmail, and misinformation campaigns

Keywords: Deepfake Detection, Digital Forensics, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs)

1. INTRODUCTION

The rise of deepfake technology has introduced serious risks across various domains, including politics, cybersecurity, and personal privacy. One of the most alarming threats is the spread of misinformation, where deepfake videos and audio clips can be used to fabricate speeches, manipulate public figures, and distort reality. These manipulated audio and video clips often appear highly realistic, making manual detection extremely difficult. Deep learning, a subset of machine learning that mimics the human brain through neural networks, has emerged as a powerful tool to counter this challenge.

This has the potential to undermine trust in news sources, influence elections, and incite social unrest. Moreover, deepfakes have been exploited for financial fraud, with cybercriminals using AI-

generated voices to impersonate executives in phishing scams or fake biometric authentication. On a personal level, individuals have been targeted with deepfake-based identity theft, blackmail, and non-consensual explicit content, causing significant psychological and reputational harm. The sophistication of modern deepfake models makes them increasingly difficult to detect, as they can accurately replicate facial expressions, voice tones, and subtle movements that once served as telltale signs of forgery. Additionally, the accessibility of deepfake- generating tools has lowered the barrier for malicious actors, allowing even those with limited technical expertise to create highly convincing fake media.

By leveraging Deep Learning models such as Convolutional Neural Networks (CNNs) for visual analysis and Recurrent Neural Networks (RNNs) for audio examination, the system will detect subtle inconsistencies indicative of deepfake manipulation. The software will integrate with existing digital forensic tools used by law enforcement agencies, enabling quick and efficient verification of digital evidence in criminal investigations. A user-friendly interface will allow police officers to conduct real-time deepfake analysis, providing automated detection and instant reports to aid in decision-making during time sensitive cases. This solution aims to enhance cybersecurity and public trust by equipping law enforcement with state-of-the-art tools to combat digital crimes such as fraud, identity theft, and misinformation.

2. BACKGROUND AND CONTEXT

Deepfake technology change audio and video making it look real even when it's made up or changed. The word "deepfake" comes from deep learning, a part of machine learning that uses neural networks to make very realistic but fake media. While people can use deepfakes for fun and creative things and they also bring big risks. These include spreading false information, stealing identities, and hurting people's reputations.

A. Deep Learning -An overview

Deep Learning is a subset of Machine Learning that focuses on training Artificial Neural networks to recognize patterns, make decisions, and generate predictions based on large datasets. Inspired by the structure and functioning of the human brain, Deep Learning models consist of multiple layers of artificial neurons that process and transform input data through weighted connections. These Deep Neural Networks can automatically extract meaningful features from raw data, eliminating the need for manual feature engineering.

One of the key advantages of Deep Learning is its ability to handle complex and high-dimensional data, making it highly effective for tasks such as image and speech recognition, Natural Language Processing, and autonomous decision-making.

Deep Learning architectures, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformers, have revolutionized fields like computer vision, medical diagnostics, and self-driving technology. CNNs are particularly powerful in analyzing visual data, enabling applications like facial recognition, object detection, and medical image analysis. RNNs and their advanced variants, such as long short-term memory networks, are well-suited for sequential data processing, allowing for advancements in speech recognition and language modeling. Deep Learning's success is largely driven by the availability of massive datasets, increased computational power through GPUs and TPUs, and advancements in optimization techniques such as backpropagation and gradient descent. However, despite its remarkable capabilities, Deep Learning

faces challenges such as high computational costs, data dependency, and lack of interpretability.

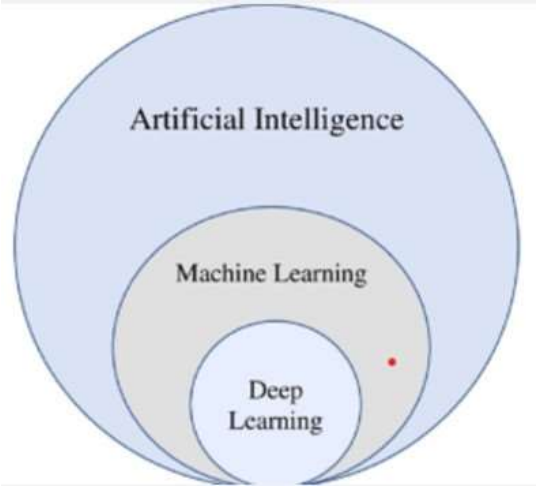

Fig 1: AI,M, DL relationship

B. Significance of Deep learning in deepfakes

Deep Learning plays a dual role in the deepfake ecosystem—it is both the driving force behind the creation of highly realistic deepfakes and the primary method for detecting them. Deepfake generation relies on deep learning models like Generative Adversarial Networks (GANs) and Autoencoders, which can synthesize hyper-realistic fake videos and audio by learning patterns from real data. These models enable seamless face swapping, voice cloning, and even lip-syncing in videos, making it increasingly difficult for humans to differentiate between real and fake media. The rapid advancements in deep learning have led to deepfakes that are almost indistinguishable from authentic content, raising significant concerns about misinformation, fraud, and digital identity theft.

On the detection side, deep learning provides powerful tools to combat the spread of deepfakes. Convolutional Neural Networks (CNNs) analyze spatial inconsistencies in deepfake videos, such as unnatural facial expressions, pixelation artifacts, and lighting mismatches. Recurrent Neural Networks (RNNs) and Transformers detect anomalies in sequential data, making them effective in identifying unnatural speech patterns in deepfake audio. Furthermore, Audio-Visual Feature Fusion (AVFF) models combine both sound and video cues to improve detection accuracy. Deep learning models can learn subtle details that human eyes and ears might miss, making them indispensable for automated deepfake detection. As deepfake technology evolves, deep learning-based detection methods must continuously improve to stay ahead of more sophisticated fake generation techniques.

Deep Learning has also enabled the development of multimodal detection techniques, which combine multiple sources of information, such as facial expressions, lip movements, and voice characteristics, to identify inconsistencies in deepfake content. For example, an Audio-Visual Feature Fusion (AVFF) model can detect mismatches between a person's speech and their lip movements, which are often present in manipulated videos. Additionally, deep learning models can be trained to recognize physiological signals, such as micro-expressions and eye-blinking patterns, that are difficult for deepfake algorithms to replicate accurately. These advanced techniques make Deep Learning a crucial tool in forensic investigations, social media moderation, and cybersecurity,

helping to identify and mitigate the risks associated with deepfake content. Despite its effectiveness, Deep Learning-based deepfake detection faces several challenges, including the continuous evolution of deepfake generation methods and the arms race between forgers and detectors. As generative models become more sophisticated, detection models must adapt quickly to recognize new manipulation techniques. Another challenge is the availability of high-quality datasets, as Deep Learning models require large, diverse, and well-annotated datasets to generalize effectively across different types of deepfakes. Moreover, Deep Learning models can sometimes struggle with generalization, meaning that a model trained on one type of deepfake might not perform well on another. To address these challenges, researchers are exploring techniques such as adversarial training, where detection models are trained alongside generative models to improve robustness, and explainable AI (XAI), which provides insights into how detection models make decisions. These innovations will help deep learning remain at the forefront of deepfake detection as the technology continues to evolve.

C. Applications of Deep Learning in deepfakes

1. Face swapping and video manipulation

Face swapping and video manipulation are among the most well-known applications of deepfake technology, powered by deep learning models like Generative Adversarial Networks (GANs) and Autoencoders. These techniques allow for seamless replacement of one person's face with another in videos while preserving facial expressions, lighting, and movements, making the swap appear highly realistic. Face-swapping is widely used in the entertainment industry for movie special effects, de-aging actors, and digital doubles, as seen in films where actors' faces are altered or recreated.

2. Synthetic voice cloning

Synthetic voice cloning, powered by deep learning models like WaveNet, Tacotron, and Transformer-based architectures, enables the precise replication of a person's voice using a small amount of recorded audio. These models analyze speech patterns, tone, and pronunciation to generate natural-sounding synthetic speech that closely mimics the original speaker. Voice cloning has numerous applications, including personalized virtual assistants, audiobook narration, dubbing in films, and helping individuals who have lost their voice due to medical conditions.

3. Lip-syncing and motion transfer

Lip-syncing and motion transfer, driven by deep learning techniques like GANs and neural motion models, enable the synchronization of lip movements with synthetic or altered speech and the transfer of facial expressions or body movements from one person to another. This technology is widely used in film dubbing, virtual avatars, and AI-powered content creation, allowing characters or digital personas to speak in multiple languages while maintaining realistic mouth movements. Motion transfer extends this capability by capturing and mapping facial expressions or full-body gestures onto another individual or animated character, making it valuable in gaming, virtual reality (VR), and digital entertainment.

4. Social media monitoring

Social media monitoring using Deep Learning-based deepfake detection is essential for identifying and mitigating the spread of manipulated content on platforms like Facebook, Twitter, TikTok, and YouTube. Platforms employ convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based models to automatically flag suspicious content, helping to combat

misinformation, identity fraud, and malicious impersonation.

D. Deepfake detection and role of deep learning

Deepfake detection is a critical field aimed at identifying manipulated media, including synthetic videos, audio, and images created using AI-driven techniques. As deepfake technology advances, traditional detection methods relying on manual inspection or simple forensic analysis have become insufficient. Deepfake detection now heavily depends on automated AI-driven solutions that can analyze minute inconsistencies in facial movements, lip synchronization, voice modulation, and even pixel-level artifacts. Given the potential misuse of deepfakes in misinformation, fraud, and cybersecurity threats, accurate and efficient detection methods are essential for law enforcement, media organizations, and online platforms to combat digital deception.

Deep Learning plays a pivotal role in deepfake detection by leveraging advanced neural networks capable of analyzing large datasets and detecting subtle patterns imperceptible to the human eye. Convolutional Neural Networks (CNNs) are widely used for image and video analysis, identifying inconsistencies in facial features, lighting, and pixel structures that may indicate manipulation. Recurrent Neural Networks (RNNs) and Transformer models are effective in detecting deepfake audio, analyzing unnatural speech patterns, pauses, and waveform distortions. Additionally, Audio-Visual Feature Fusion (AVFF) models combine both video and audio cues to improve detection accuracy. Researchers are also exploring adversarial training, where detection models are continuously trained against evolving deepfake generation techniques, ensuring their robustness against more sophisticated forgeries. As deepfake technology evolves, deep learning-based detection systems will remain essential in the ongoing battle against digital misinformation and synthetic media fraud.

3. RELATED WORK

Deepfakes have been exploited for financial fraud, with cybercriminals using AI-generated voices to impersonate executives in phishing scams or fake biometric authentication. On a personal level, individuals have been targeted with deepfake-based identity theft, blackmail, and non-consensual explicit content, causing significant psychological and reputational harm. The sophistication of modern deepfake models makes them increasingly difficult to detect, as they can accurately replicate facial expressions, voice tones, and subtle movements that once served as telltale signs of forgery

The paper [1] contains most deepfake detection models and are found to be poor in uncovering unseen or new data essentially because they are built on particular distributions that might not be identical to real-world data. The study has proposed a new stance for solving the domain generalization problem that deals with detecting fake and real sounds in a more effective way. This can be done by making a deep network to separate audio signals that are fake from those that are real.

The paper [2] is about a multi-perspective approach to detecting deepfake audios by using a lot of features from the audio recordings. A method that this study takes a step further from traditional methods which are based on a single feature and uses various morphological, acoustic, and Spectro-temporal feature extraction techniques that include spectrograms, mel-frequency cepstral coefficients (MFCC), and frequency-domain characteristics for comprehensive analysis of the audio signal. The paper [3] gives a novel deepfake detection system using the conformer-based

method, and it is proposed that with the help of hierarchical pooling and multilevel classification token aggregation, the performance of the system can be enhanced. In this conformer architecture, which is a combination of both convolutional and transformer-based models, local and global signal dependencies are covered and this is why it becomes the current best candidate for the detection of deepfake manipulations.

In paper [4] it aims at the implementation of a one-shot learning method to find out audio files made by a deepfake, that is voice biometrics security. In particular, the traditional detection methods of different types of deepfake with the absence of labeled datasets require the humans to communicate with labels but the document proposes a new technique called Siamese network and prototypical network to get higher detection accuracies with less training data. Due to the fact that one-shot learning technique lack of quantity the model is able to identify digital audio illicit of the many samples of the same file, then it is only comparing these selected samples with a very small number of authenticated audio files. Furthermore, the study of contrastive loss functions affected techniques aimed at discriminating real and synthetic speech. As it comes out from the conducted examinations, the approach of the proposed one-shot learning reaches good performance which it proves that it can be a practical during the process of detecting the deepfake in the voice in real. In paper [5] introduces an ensemble learning framework which is a combination of different detection models using weighted averaging and boosting techniques to optimize classification performance. Furthermore, the research highlights the importance of diverse training datasets to achieve the sustainability of the model in relation to diversity against completely new deepfake manifestations. Such an approach proposed, as a result of which, becomes the root cause of the change in one's approach: it is not only capable of improving accuracy but also essentially eliminates the false positives, thus making it all the more applicable in forensic and cybersecurity contexts. This study is a useful guide on how the integration of algorithms can fortify deepfake detection systems and represents a promising avenue for further research in the audio forensic field. The paper[6] reveals a way to uncover forged audio by using self-supervised learning with WAVLM (Waveform-to-Latent Model) method and a multifusion attentive classifier. The self-supervised speech representation model WAVLM captures the complicated speech patterns so it is a good option to the identification of a synthetic audio.

In paper [7] gives a full rundown of different methods used to spot fake audio deepfakes. It covers old machine learning deep learning, and mixed approaches. The paper groups detection methods by how they pull out features, like looking at sound waves modeling waveforms, and using neural networks to make embeddings. It talks about how fake audio creation has gotten better over time with text-to-speech and voice changing tech. The paper [8] examines a method to detect deepfake audio that relies on feature engineering and classic Machine Learning. It looks at various handmade audio features, like Mel-Frequency Cepstral Coefficients (MFCCs), pitch contours, spectral centroid, and energy-based features, to spot differences between real and fake voices. The paper [9] combines methods to explain AI with ways to spot fake audio making it easier to understand and trust forensic analysis. The system uses deep learning models like CNNs and RNNs to sort audio as real or fake and explain why. It uses tools such as SHAP and Grad-CAM to point out which parts of the audio led to the decision helping experts check the results. The paper [10] discusses a multimodal deepfake detection method that utilizes both visual and audio features to increase detection accuracy. Because deepfake videos tend to have longer inconsistencies in both

modalities, the combination of both offers a richer analysis. The work utilizes Convolutional Neural Networks (CNNs) for visual feature extraction and Long Short-Term Memory (LSTM) networks for audio feature analysis.The paper [11] discusses the application of Mel-Frequency Cepstral Coefficients (MFCCs) as the feature representation of choice for deepfake audio detection with Machine Learning approaches. MFCCs represent the spectral features of speech and are thus a good choice for distinguishing between real and fake voices. The paper [12] examines how Deep Learning methods affect deepfake audio detection in digital forensic examinations. The research contrasts various neural network structures, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers, in order to study fake and natural speech patterns. In paper [13] it targets deepfake audio detection in group conversations, a challenging but significant area of digital forensics and security. Employing speaker recognition models and temporal analysis, the system determines voice identity discrepancies, speech rhythm anomalies, and artificial transitions between speakers. In paper[14] it proposes a domain-invariant feature extraction method that enables the model to learn invariant representations generalizing across various datasets. By employing Contrastive Learning and Data Augmentation, the framework enhances the model's capability to identify deepfake speech in various domains. The experiments indicate dramatic accuracy gains in cross-dataset tests. Yet, the research also identifies difficulties in fine-tuning the trade-off between feature generalization and specificity. This work is vital for deepening real-world usability of deepfake detection systems. In paper [15] it examines the issue of using only one dataset or one deepfake synthesis technique for training models that would still provide strong detection performance. The research employs spectral and waveform-based features in combination with deep learning models like CNNs and Autoencoders. The findings point out that well-designed training approaches, data augmentation, and feature engineering can greatly enhance model performance, even in single-domain settings.

The paper [16] introduced as a targeted dataset aimed to enhance audio deepfake detection research. The dataset consists of speech samples synthesized via various voice synthesis and voice conversion methods. In paper [17] the research utilizes domain generalization methods, data augmentation, and adversarial training to enhance robustness. The findings show that attack-agnostic training remarkably improves detection accuracy for various types of deepfakes. The paper [18] it is a low-complexity and high-speed deepfake audio detector model that utilizes spectrogram-based neural networks for the sake of boosting speed and accessibility. Unlike regular deep learning models, which tend to be computationally costly, SpecRNet is developed with low-resource settings in mind, thus making it viable for real-time and edge computing systems. The paper [19] presents defense strategies like adversarial training, noise injection, and spectral filtering to augment resistance to such attacks. The experimental results show that adversarial training improves robustness substantially with high detection accuracy. The paper [20] introduces an original method for detecting audio deepfakes by taking advantage of stereo sound properties in speech synthesis. The research formulates a Mono-to-Stereo Conversion (MTSC) approach, in which deepfake audio—usually synthesized in mono—is processed and analyzed under the application of stereo-based models. The paper [21] presents the largest Chinese dataset to date that is specially designed for detecting fake songs. Both real and AI-generated songs are included in FSD, which spans a wide range of deepfake audio synthesis methods. The authors train CNNs and RNNs using this dataset to test their efficacy in detecting fake songs versus authentic ones.

The paper [22] The AVFF model architecture, describing how it addresses the fusion of features using a multi-stream neural network architecture. The model processes audio and visual features separately before they are fused in a single representation. The authors illustrate in extensive experiments that their method significantly surpasses standard single-stream detection models. They demonstrate that adding audio features, like speech pattern irregularities or lip movement versus sound synchronization errors, can improve the robustness and reliability of deepfake detection considerably. The paper[23] aims to minimize computational costs while ensuring high accuracy in deepfake detection. Conventional deepfake detection models tend to be computationally costly, involving large energy consumption and processing power. It suggests lightweight neural structures, knowledge distillation methods, and low-power feature extraction strategies to reduce power consumption without compromising detection performance. The paper [24] suggests an ensemble of several WavLM models to improve the system's performance in detecting artificial audio. WavLM has been shown to be particularly strong in speech representation learning, thus being very good at detecting anomalies in speech signals. The ensemble method blends various versions of WavLM, which are trained on diverse datasets and synthesis processes, to enhance generalization across various deepfake generation methods. The paper [25] presents CLAD (Contrastive Learning for Audio Deepfake detection), a new framework proposed to enhance resistance to adversarial and post-processing attacks. Numerous deepfake detectors are unsuccessful if audio is compressed, edited, or tampered with using noise, pitch manipulation, or speed modification. CLAD exploits the contrastive learning method that better enables a model to distinguish real from synthetic audio even in cases of multiple transformations.

4. COMPARISON AND RESULTS

Ref Nos.	Techniques	Merits	Demerits
Xie et al. [1]	Aggregation and separation domain generalization(ASDG) Lightweight convolutional Neural network(LCNN)	Enhanced feature discrimination. Adaptable to new deepfake technique.	Computational complexity Implementation complexity
Yang et al. [2]	Hidden-unit BERT(HUBERT) XLS-Robust(XLR-R) WavLM	Multi-view feature fusion Increased detection accuracy	The multi-view feature approach increases computational complexity and may require more resources for training
Shin et al. [3]	Hierarchical pooling Multi-level classification token aggregation (MCA)	Effective temporal modelling. Scalability	Making it less suitable for real-time or resource constrained applications
Khan et al. [4]	Uses one shot combined with metric learning technique	Quick deployment Data efficiency	It heavily relies on the quality and diversity of training dataset

Borade et al. [5]	Mel-frequency Cepstral Coefficients(MFCCs) Convolutional neural network(CNN)with attention mechanism	Combines multiple algorithms Robust detection	Method may struggle to generalize across highly diverse where deepfake techniques vary significantly
Guo et al. [6]	Feature extraction with WAVLM Multi-fusion attentive(MFA) classifier Joint training and fine tuning	Leverages self supervised learning Multi-fusion approach	It focuses on WAVLM and a specific classifier design which might limit its adaptability to future advancements.
Wang et al. [7]	Feature extraction Machine Learning classifiers Deep learning models and self supervised learning models	Bench marking competitions Dataset analysis	It lacks the empirical results or a clear unified framework for comparing different detection methods.
Iqbal et al. [8]	Feature engineering Machine learning models Classifier training	Data preprocessing Handling large datasets.	Limited ability to handle complex and diverse audio patterns effectively.
Govindu et al. [9]	Deep learning models Generative Adversarial networks Feature engineering Explainable AI(XAI)	Enhanced interpretability Increased trustworthiness	Methods may struggle to generalize to real world scenarios with diverse and noisy data, limited practical effectiveness.
Zhang et al. [10]	Multitask learning Feature fusion and score fusion backbones for feature learning	Improved accuracy Real-world applicability	Performance can degrade when detecting cross domain deepfakes or when faced with datasets that differ from those used in training
Hamza et al. [11]	MFCC Machine learning models- Semirandom forest and Neural networks preprocessing	Higher detection accuracy Practical for real time use	Models can struggle with high false positive rates and are sensitive to background noise and variation in audio quality.
Mcuba et al. [12]	Feature extraction Deep learning architectures Custom architectures	Effective for digital investigation Advanced detection techniques	Focus on specific feature types and architectures may limit the generalization of results to other audio datasets.

Wijethunga et al. [13]	Speech denoising Speaker diarization Text conversion	Improved robustness Focus on group conversations	It lacks specific accuracy metrices and does not fully evaluate the systems performance.
Xie et al. [14]	W2V2 front-end Aggregation and separation domain generalization(ASDG) Gradient Reversal layer(GRL)	Efficient and scalable Improved performance	It may require significant computational resources for training the model, which can limit its practical applicability in resource constrained environments.
Xie et al. [15]	Domain invariant Domain adversarial training Regularization technique and data augmentation SM-ASDG	Improved Detection Performance in a Controlled Domain	It focuses on detecting audio deepfake in a single domain which may limit the model's ability to generalize effectively new, unseen domains in deepfake techniques
Frank et al. [16]	WaveFake dataset Audio synthesis Data augmentation	High quality data Focused on audio deepfakes Real-world applicability	It focuses mainly on dataset creation and does not explore or evaluate specific deepfake detection models in detail
Kawa et al. [17]	Attack -Agnostic dataset creation Cross-attack generalization Stabilization of detection models Proposed model based on LCNN with LFCC and mel spectrogram front-end	Improved generalization Attack-agnostic dataset Enhanced model robustness	Without offering detailed insights into the underlying algorithms or specific improvements in detection models
Plaga et al. [18]	SpecRNet Architecture Spectrogram based feature Real time detection	Faster detection Accessibility	It focuses on speed and accessibility it may sacrifice some detection performance compared to more complex models
Syga et al. [19]	Adversarial defense techniques Data augmentation Model regularization Robust training	Adversarial attack protection Improved security Increased model reliability	It primarily focuses on defending against adversarial attacks and does address improvements in overall detection accuracy of audio deepfake.
Liu et al. [20]	Mono-_to stereo conversion Feature extraction Deep learning model	Efficient usage of audio channels Improved detection sensitivity	It relies on converting mono audio to stereo which may increase computational complexity and may not be suitable for Realtime application

Xie et al. [21]	Text to speech(TTS)& voice conversion MFCCs	Supports song manipulation detection	It focuses mainly on creating the FSD dataset and does not provide in depth analysis
Oorloff et al. [22]	Audio-Visual Feature Fusion (AVFF) - Multi-stream neural network model	Detects deepfakes more accurately by analysing both audio and video. elements. Works well on various datasets and deepfake types.	More computationally intensive due to processing both audio and video. May require more data for training compared to single-modality models.
Saha et al. [23]	ASVspoof Anti-spoofing Self supervised learning Energy efficient model Green AI framework	Faster processing Cost reduction Wider accessibility	It prioritizes energy efficiency over detection accuracy which may lead to reduced performance in deepfake detection tasks.
Combei et al. [24]	pretrained models Fine tuning Data augmentation Model ensemble	Better generalization Enhanced performance	It limit the generalizability of the results to results to real world scenarios where such conditions might not be present.
Chen et al. [25]	Contrastive learning Length loss Evaluation against manipulations	Resilience to adversarial attacks Contrastive learning Enhanced robustness	Limited scope validation complexity of Contrastive learning

5. CONCLUSION

The rapid proliferation of deepfake technology necessitates a proactive approach to detection and mitigation, particularly within law enforcement and digital forensics. The proposed deepfake detection software provides a robust solution by leveraging advanced AI techniques, including Deep Learning models for both visual and audio analysis. By integrating this software into forensic workflows, law enforcement agencies can strengthen their ability to authenticate digital media, ultimately enhancing investigative accuracy and reducing the risk of misinformation and fraudulent activities. Furthermore, this solution plays a crucial role in upholding public trust and security by equipping law enforcement with the necessary tools to combat digital deception. As deepfake technology continues to evolve, ongoing research, model updates, and data expansion will be essential to maintaining detection accuracy and reliability. By staying ahead of emerging deepfake threats, this initiative ensures that cyber police departments remain well-prepared to address the challenges posed by synthetic media, thereby contributing to a more secure and trustworthy digital landscape. The implementation of this deepfake detection system fosters collaboration among cybersecurity experts, forensic analysts, and AI researchers, further strengthening the technological defenses against digital manipulation. By promoting interdisciplinary efforts, this project not only enhances law enforcement capabilities but also sets a foundation for future advancements in digital

forensic methodologies. As the threat of deepfakes continues to grow, continuous innovation and adaptation will be key to sustaining the effectiveness of detection systems and ensuring the integrity of digital evidence in legal and investigative contexts.

REFERENCES

1. Xie, Y., Cheng, H., Wang, Y., & Ye, L. (2024). "Domain Generalization via Aggregation and Separation for Audio Deepfake Detection." IEEE Transactions on Information Forensics and Security, 19(1), 344–358. https://doi.org/10.1109/tifs.2023.3324724
2. Yang, Y., Qin, H., Zhou, H., Wang, C., Guo, T., Han, K., & Wang, Y. (2024). "A Robust Audio Deepfake Detection System via Multi-View Feature." Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, 13131– 13135. https://doi.org/10.1109/icassp48485.2024.10446560
3. Shin, H.-s., Heo, J., Kim, J.-h., Lim, C.-y., Kim, W., & Yu, H.-J. (2024). "HM-
4. CONFORMER: A Conformer-Based Audio Deepfake Detection System with Hierarchical Pooling and Multi-Level Classification Token Aggregation Methods." Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, 10581–10585. https://doi.org/10.1109/ICASSP48485.2024.10448453
5. Khan, A., & Malik, K. M. (2023). "Securing Voice Biometrics: One-Shot Learning Approach for Audio Deepfake Detection." arXiv preprint, Khan, A., & Malik, K. M. (2023). "Securing Voice Biometrics: One-Shot Learning Approach for Audio Deepfake Detection." arXiv preprint, arXiv:2310.03856. https://doi.org/10.48550/arXiv.2310.03856
6. Jayan Shah, Pratham Shah, Mustansir Godhrawala, S. B. N. J. B. P. V. K. Y. N. S. K. . (2024). Harmonizing Algorithms: An Approach to Enhancing Audio Deepfake Detection. International Journal of Intelligent Systems and Applications in Engineering, 12(3), 1297–1304. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/5520
7. Guo, Y., Huang, H., Chen, X., Zhao, H., & Wang, Y. (2024). "Audio Deepfake Detection with Self-Supervised WavLM and Multi-Fusion Attentive Classifier." Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),vol 1, 10586–10590. https://doi.org/10.1109/icassp48485.2024.10447923
8. Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C. Y., & Zhao, Y. (2023). "Audio Deepfake Detection: A Survey." arXiv preprint, arXiv:2308.14970.https://doi.org/10.48550/arXiv.2308.14970
9. Iqbal, F., Abbasi, A., Javed, A. R., Jalil, Z., & Al-Karaki, J. (2022). "Deepfake Audio Detection via Feature Engineering and Machine Learning." Proceedings of the CIKM 2022 Workshops co-located with 31st ACM International Conference on Information and Knowledge Management (CIKM 2022), 3318, 1–8.
10. Govindu, A., Kale, P., Hullur, A., Gurav, A., & Godse, P. (2023). "Deepfake Audio Detection and Justification with Explainable Artificial Intelligence (XAI)." arXiv preprint, arXiv:2303.12345. doi:10.21203/rs.3.rs-3444277/v1
11. Zhang, Y., Li, X., Wang, J., & Liu, H. (2023). "Integrating Audio-Visual Features for Multimodal Deepfake Detection." arXiv preprint, arXiv:2310.03827. doi:10.48550/arXiv.2310.03827

12. Hamza, A., Javed, A. R., Iqbal, F., Kryvinska, N., Almadhor, A. S., Jalil, Z., & Borghol, R. (2022). "Deepfake Audio Detection via MFCC Features Using Machine Learning." IEEE Access, 10, 134018–134028. https://doi.org/10.1109/access.2022.3231480
13. Mcuba, M., Singh, A., Ikuesan, R. A., & Venter, H. (2023). "The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital Investigation." Procedia Computer Science, 219, 211–219. https://doi.org/10.1016/j.procs.2023.01.283
14. Wijethunga, R. L. M. A. P. C., Matheesha, D. M. K., Noman, A. A., De Silva, K. H. V. T. A., Tissera, M., & Rupasinghe, L. (2020). "Deepfake Audio Detection: A Deep Learning-Based Solution for Group Conversations." Proceedings of the 2nd International Conference on Advancement in Computing (ICAC),vol 1192–197. https://doi.org/10.1109/ICAC51239.2020.9357161
15. Xie, Y., Cheng, H., Wang, Y., & Ye, L. (2023). "Learning a Self-Supervised Domain-Invariant Feature Representation for Generalized Audio Deepfake Detection." Proceedings of the 24th Annual Conference of the International Speech Communication Association (INTERSPEECH 2023),vol 1, 2808–2812. doi:10.21437/Interspeech.2023-1383
16. Xie, Y., Cheng, H., Wang, Y., & Ye, L. (2023). "Single Domain Generalization for Audio Deepfake Detection." CEUR Workshop Proceedings, 3597, 1–8.
17. Frank, J., & Schönherr, L. (2021). "WaveFake: A Data Set to Facilitate Audio Deepfake Detection." NeurIPS Datasets and Benchmarks 2021, url = https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/c74d97b01eae257e44aa9d5bade97baf-Paper-round2.pdf
18. Kawa, P., Plata, M., & Syga, P. (2022). "Attack Agnostic Dataset: Towards Generalization and Stabilization of Audio DeepFake Detection." Proceedings of the 23rd Annual Conference of the International Speech Communication Association (INTERSPEECH 2022), 4023–4027. https://doi.org/10.48550/arXiv.2206.13979
19. Kawa, P., Plata, M., & Syga, P. (2022). "SpecRNet: Towards Faster and More Accessible Audio DeepFake Detection." Proceedings of the 21st IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 792–799. https://doi.org/10.1109/TrustCom56396.2022.00111
20. Kawa, P., Plata, M., & Syga, P. (2023). "Defense Against Adversarial Attacks on Audio DeepFake Detection." Proceedings of the 24th Annual Conference of the International Speech Communication Association (INTERSPEECH 2023), 5276–5280. https://doi.org/10.21437/Interspeech.2023-409
21. Liu, R., Zhang, J., Gao, G., & Li, H. (2023). "Betray Oneself: A Novel Audio DeepFake Detection Model via Mono-to-Stereo Conversion." Proceedings of the 24th Annual Conference of the International Speech Communication Association (INTERSPEECH 2023), 3999–4003. https://doi.org/10.21437/Interspeech.2023-2335
22. Xie, Y., Zhou, J., Lu, X., Jiang, Z., Yang, Y., Cheng, H., & Ye, L. (2023). "FSD: An Initial Chinese Dataset for Fake Song Detection." arXiv preprint, arXiv:2309.02232. https://arxiv.org/abs/2309.02232&#8203
23. Oorloff, T., Koppisetti, S., Bonettini, N., Solanki, D., Colman, B., Yacoob, Y., Shahriyari, A., & Bharaj, G. (2024). "AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

(CVPR), 2710–2720. https://doi.org/10.1109/CVPR2024.00300
24. Saha, S., Sahidullah, M., & Das, S. (2024). "Exploring Green AI for Audio Deepfake Detection." arXiv preprint, arXiv:2403.14290. DOI:10.23919/EUSIPCO63174.2024.10715424
25. Combei, D., Stan, A., Oneață, D., & Cucu, H. (2024). "WavLM Model Ensemble for Audio Deepfake Detection." Proceedings of the Automatic Speaker Verification and Spoofing Countermeasures Challenge (ASVSpoof5). https://doi.org/10.48550/arXiv.2408.07414
26. Wu, H., Chen, J., Du, R., Wu, C., He, K., Shang, X., Ren, H., & Xu, G. (2024). "CLAD:
27. Robust Audio Deepfake Detection Against Manipulation Attacks with Contrastive Learning." arXiv preprint, arXiv:2404.15854. https://doi.org/10.48550/arXiv.2404.15854