ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

A REVIEW OF MULTIMODAL DEEPFAKE FORGERY DETECTION IN AUDIO-VIDEO-TEXT

Hamna Ali
hamnaali0211@gmail.com
Department of Computer Science & Information Technology, University of Southern Punjab (USP) Multan , Punjab, Pakistan
Hamid Ghous
hamidghous@usp.edu.pk
Department of Computer Science & Information Technology, University of Southern Punjab (USP) Multan , Punjab, Pakistan
Muhammad Asim
Muhammad.aasim2024@gmail.com
Department of Artificial Intelligence
Islamia University of Bahawal Pur (IUB),
Punjab, Pakistan
Tayyaba Rashid
tayyaba.rashid@iub.edu.pk
Department of Software Engineering
Islamia University of Bahawal Pur (IUB), Punjab, Pakistan
Muhammad Naeem
phcs-f18-003@superior.edu.pk
Department of Computer Science
Islamia University BahawalPur (IUB), Punjab, Pakistan
Hafiz Muhammad Usman Akhtar
hafizusman@iub.edu.pk
Department of Artificial Intelligence
Islamia University of Bahawal Pur (IUB), Punjab, Pakistan
Corresponding Author

Abstract
Recent advancements in generative models, particularly those based on deep learning and artificial intelligence, have significantly enhanced the realism and accessibility of deepfake content. While these technologies offer numerous creative and industrial benefits, they also pose a serious threat to the authenticity and trustworthiness of digital media. The rapid evolution of deepfake techniques has made it increasingly difficult to distinguish between genuine and manipulated content, thereby raising concerns across multiple domains, including media, politics, and cybersecurity. As a result, there is a growing need for robust and advanced detection mechanisms that can effectively identify such manipulations and preserve the integrity of information in the digital ecosystem.

In this review, we present a comprehensive synthesis of existing deep learning frameworks designed for multimodal deepfake detection. Unlike traditional unimodal approaches that focus solely on visual or auditory cues, multimodal detection systems integrate auditory, visual, and textual data to provide a more holistic and accurate analysis. Special emphasis is placed on fusion techniques that combine features from these modalities, enabling models to detect subtle inconsistencies that may not be evident when analyzing a single modality in isolation.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Additionally, this study evaluates widely used public datasets and benchmark evaluation metrics, highlighting their role in training and validating detection systems. The findings demonstrate that multimodal approaches significantly improve detection performance, especially in complex scenarios involving highly sophisticated manipulations.

Furthermore, the importance of deep fake detection extends beyond technical challenges, as it has critical implications for society. Effective detection systems play a vital role in applications such as social media content moderation, judicial forensic analysis, and fraud prevention. At the same time, the deepfake of deceptive content can lead to serious psychological and social consequences, including misinformation, reputational damage, and mental health issues such as anxiety and depression. Therefore, future research should prioritize the development of lightweight, scalable, and efficient architectures that can be deployed in real-world environments. Moreover, there is a strong need for standardized evaluation protocols and interdisciplinary collaboration to ensure consistent benchmarking and improved robustness. Such efforts will be essential in building reliable systems capable of safeguarding digital content authenticity in an increasingly complex and dynamic technological landscape.

Keywords Deepfake detection, Multimodal fusion, Audio-visual-text forgery, Deep learning, Public Datasets

1. Introduction

Deepfake technology has made major strides during the past few years which allows users to create fake videos, audios and texts that are nearly indistinguishable from authentic content, both visually and audibly. The fast advancement of this field stems from deep learning approaches including Generative Adversarial Networks (GANs)[1], Variational Auto Encoders (VAEs)[2], Diffusion Models (DMs)[3] and transformer-based generative architectures[4]. The threshold for making realistic fake content has dropped significantly because of these methods which allow users to create believable face swaps[5], face reenactment[6],voice cloning[7] and speech synthesis[8]. The spread of forgery technology has created various social problems which threaten to damage both social media platforms and political communication and financial security[9], [10]. The spread of forged speech videos would create conditions for social instability to emerge while fake voice would serve criminal purposes to steal identities and run financial fraud operations. These circumstances breed substantial financial damage to individuals and their communities as well as undermining people's faith in digital media platforms. The pernicious proliferation of deepfake audio-video-text creation, opens up indirect psychological dangers by enabling impersonation [11], the dissemination of misinformation, and socially directed manipulation. Such dynamics lead to amplified cognitive stress, anxiety and depression and, therefore, highlight the necessity of detection systems that take human-centric impact into account as well as the traditional accuracy measures [12].

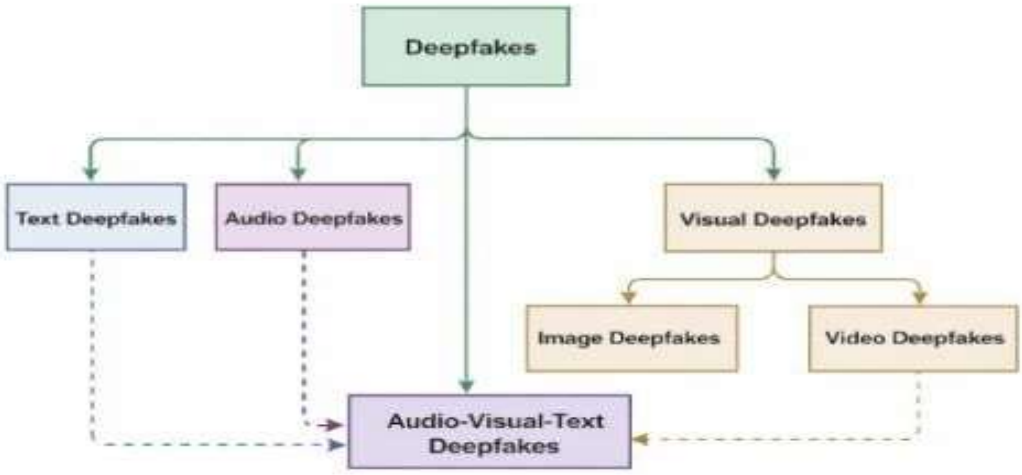
The current approaches for detecting forgery can be broadly classified into unimodal forgery detection methods and multimodal forgery detection methods. Unimodal detection approaches rely on a single modality such as analyzing video frames with convolutional neural networks (CNNs) or extracting features from audio spectrograms. While these methods[13], [14], [15], [16]perform well in certain controlled environments, they often struggle to generalize against increasingly sophisticated Deepfake techniques. Moreover, modern forgery techniques frequently combine visual and auditory streams for multimodal attacks. Research[17], [18] has shown that synchronizing forged face images with corresponding synthetic speech can produce highly

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

convincing fake videos, dramatically increase their credibility and make detection more challenging. The detection systems need to meet elevated performance standards because the current unimodal detection methods fail to identify the complex inconsistencies which appear across different modes. The fusion of video data with audio streams enables multimodal forgery detection systems[19], [20] to find additional evidence which shows manipulation activities thus they achieve better detection results and stronger system stability. The current approaches to the problem consist of feature level fusion and decision-level fusion [21], [22]and hybrid methods but each approach encounters its own set of technical obstacles[23]

The systematic review of mainstream Deepfake detection technologies which focuses on multimodal video, text and audio analysis methods. Initially, the development status explains how Deepfake technology evolved while facing various difficulties which include video manipulation, text and audio falsification techniques. Subsequently, deep learning based multimodal detection techniques are summarized including recent advances in multimodal fusion strategies by figure1. Future research directions involve the development of large-scale pre-trained models, data augmentation, and multimodal adversarial learning. Our important contribution also includes a comprehensive discussion of publicly available datasets relevant to this task. To bridge this gap, the main contributions of our work are as follows:

- To provide an unprecedented review that systematically analyzes key detection and generation methods for audio-video-text deepfakes, with a special emphasis on automatic video deepfake detection methods.
- A comprehensive overview of existing image, video and audio deepfake databases/datasets that could be utilized not only for enhancing accuracy, generalization and resilience against attacks of deepfake detection techniques but also for devising novel deepfake generation methods and tools.
- An extensive discussion of open key challenges and potential research directions in the field of audio, text and visual deepfake generation and mitigation such as the rapid evolution of forgery techniques, generalization issues in real-world scenarios and defenses against adversarial attacks.



ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Figure.1 Types of Deepfake

In sum, the review aims to be a comprehensive technical record for the academic community while at the same time driving further advancement in the field of multimodal forgery detection technologies. All in all, hope that the review will complement prior researched articles and help newcomers, researchers and engineers achieve profound comprehension and novel deepfake algorithms developments.

2. Overview of facsimile techniques:

The evolution of forgery technology has transitioned from traditional methods reliant on manual manipulation to sophisticated automated techniques powered by deep learning. Initially, video forgery involved manual cutting, splicing and tampering which were prone to inconsistency and left noticeable traces while traditional audio forgery produced low-quality outputs susceptible to detection and text forgery involves the deliberate manipulation or generation of written content such as syntactic structure, semantics or authorship style using automated or AI-driven methods to mislead readers by presenting fabricated or altered information as authentic. However, advancements in artificial intelligence particularly deep learning have revolutionized forgery methods which enhancing visual realism and behavioral naturalness to a degree that challenges human detection. Through the use of models like GANs [1], VAEs, diffusion models, and Transformers, AI-based methods produce very realistic video and audio content.

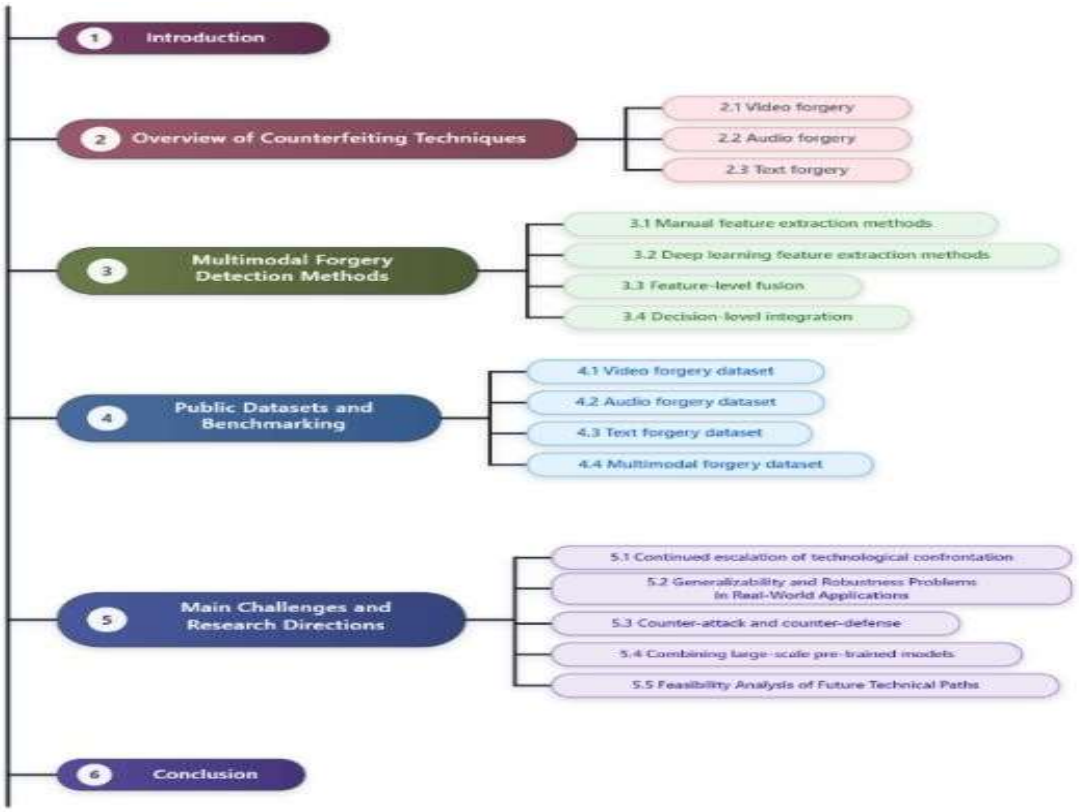


Figure.2 Overview the Forgery Survey

For instance, with the advent of the deep-fake technologies, it is now possible to achieve perfect facial substitutions and expression changes; at the same time, complex speech-synthesizers like WaveNet [7] or Transformer-based architectures can produce nearly indistinguishable sound

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

outputs compared to the recordings of the actual speakers. These changes not only make forged content better, but they also make the method of falsification faster and more subtle, which makes it much harder for detection technologies to work.

2.1. Video-Forgery

The main goal of video forgery techniques is to create false information by either manipulating existing video information or synthesis new material [24]. These approaches can be roughly classified in two main categories: Manual Forgery Techniques and Deep - Fake Technology as shown in Figure 3.

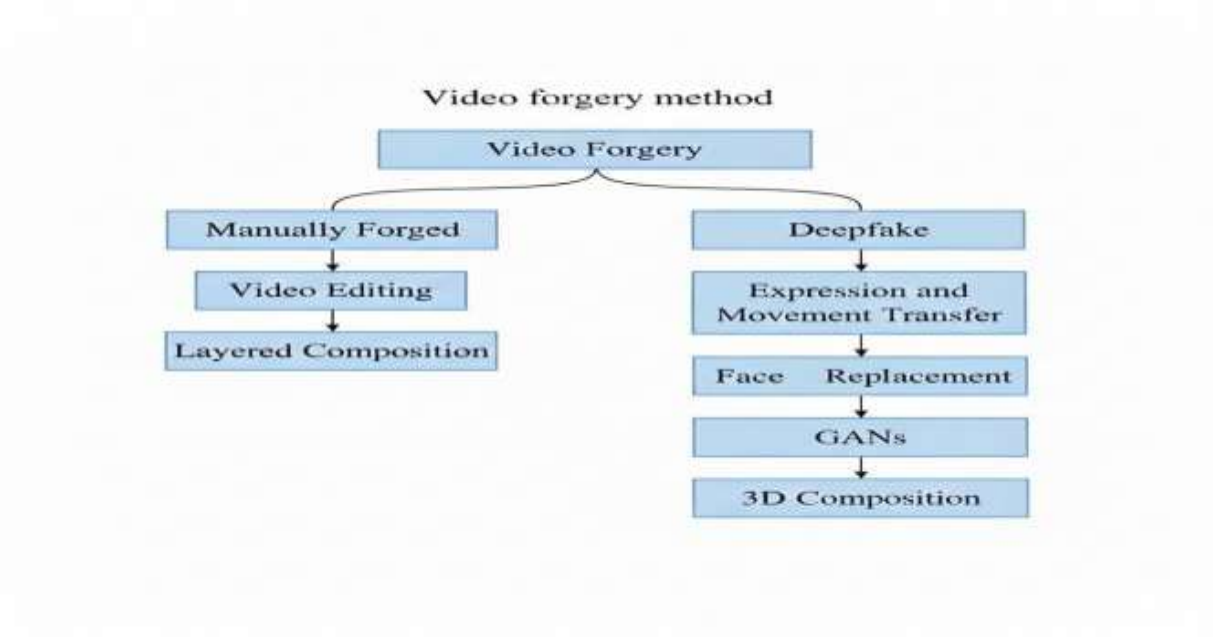


Figure.3 Video Forgery Method

Manual Forgery Techniques: Traditional methods are those based on manual operations, in which forgers use software (such as Adobe Photoshop and Adobe Premiere) to cut, splice and manipulate video segments [25]. These techniques often leave telltale signs of the tampering because they rely heavily on human intervention, and can often leave inconsistencies in the temporal continuity, lighting conditions or texture properties. Although improvements in detection technologies have reduced their overall effectiveness, such methods are still used in low resource settings or in specific cases of forgery.

Deep-Fake Technology: The advancement of deep learning models resulted in the emergence of a prominent class of forgery methods that can generate highly realistic fake videos. Key techniques are autoencoders and GAN based techniques [26]. Autoencoders are used to extract features from the source images for a face swapping operation, as shown in Figure 4. Style - GAN [27] generates high - quality facial images using adversarial training. Recent improvements, such as Talk Lip Net, better lip sync and overall visual quality. However, the increased realism attained by such technologies greatly complicates the detection of forgeries, requiring the sophisticated examination of subtle, fine-grained inconsistencies. Moreover, Deep-Fake technology is not limited to video technology alone but is also applied to the pure video-generation using generative

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

models such as GANs[28], DiT, or VAEs. This is a creation of continuous video frames using random noise as a ground or conditional inputs. The combination of VAEs and DiT (called hybrid techniques) support to create high quality videos from text prompts or images. The use of 3D composition techniques [29] also adds to increased realism with virtual modelling and lighting simulations. Nonetheless, these advancements also bring some problems in detection associated with temporal coherence, physical inconsistency, and detail deficiencies which provide opportunities for further research in enhancing forgery detection systems.

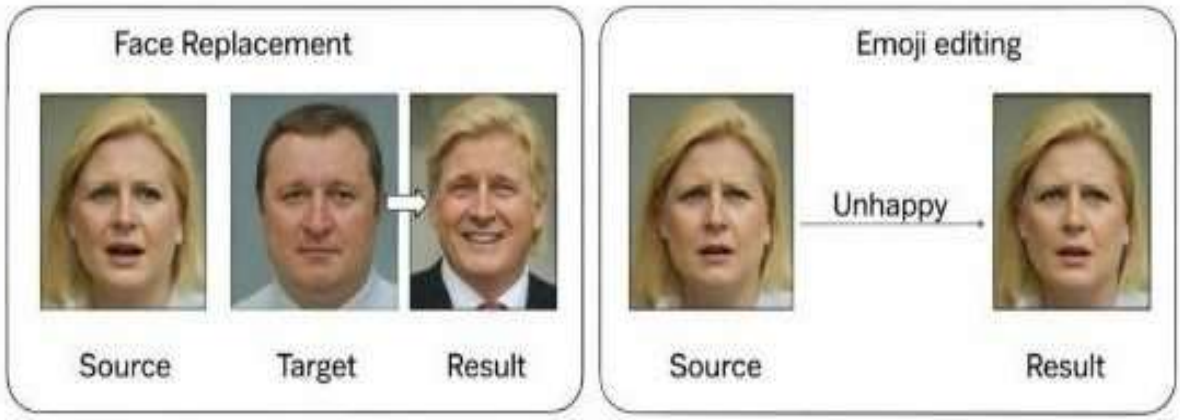


Figure.4 Artificial facial images from the Different dataset, demonstrating the uses of face swapping and expression modification[30]

2.2. Audio-Forgery

Audio forgery technologies attempt the creation or manipulation of speech content with the aim of creating falsified audio, thus requiring careful examination in modern security scenarios [31]. Along with the evolution of technological capabilities, the field of audio forgery [32] has evolved from primitive manual manipulations to high-tech automated systems based on the principles of deep learning, which have significantly improved both the generation fidelity and the efficiency of operation [33]. The major types of audio forging techniques are then described in Figure 5.

Manual Forgery Techniques: This is a traditional method of audio tampering where the main tool is human operators that perform waveform editing and segmentation and splicing operations [34]. Practitioners use readily available digital audio workstations such as Audacity [35] or Adobe Audition [36] to remove or insert sections of sound, as well as carry out frequency domain transformations (e.g. filtering, spectral re-shaping) in order to conceal evidence of manipulation. due to their necessity for bodily operation, these methods often leave telltale spectral discontinuities or temporal irregularities, making the forged material rather simple to identify.

Deepfake Technology: Modern day deepfake forgery includes a range of sophisticated techniques that involve voice cloning, speech synthesis, and voice conversion. Voice cloning is aimed to simulate the vocal idiosyncrasies of the target person using Generative Adversarial Networks (GANs) [10], Wave Net architectures [8] and Transformer-based models [37]. In the adversarial learning framework, GANs can produce realistic voice waveforms and also Wave-Net and Transformer-based architectures which output high-quality (HQ) signals by modeling temporal dependencies embedded in spoken language. Whilst these technologies can generate voice outputs that are virtually indistinguishable from real voice recordings, they at the same time create

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

significant problems for forensic acoustic detection. The Speech synthesis with models like Tacotron[38] and Fast -Speech[39] focuses on transmuting textual or spectral features into naturalistic and fluent speech audio using deep learning methods which are optimized for HQ voice synthesis. In opposition, voice conversion processes modify the vocal attributes of a speaker (e.g. gender, age, accent) whilst preserving the semantic integrity of the utterance.

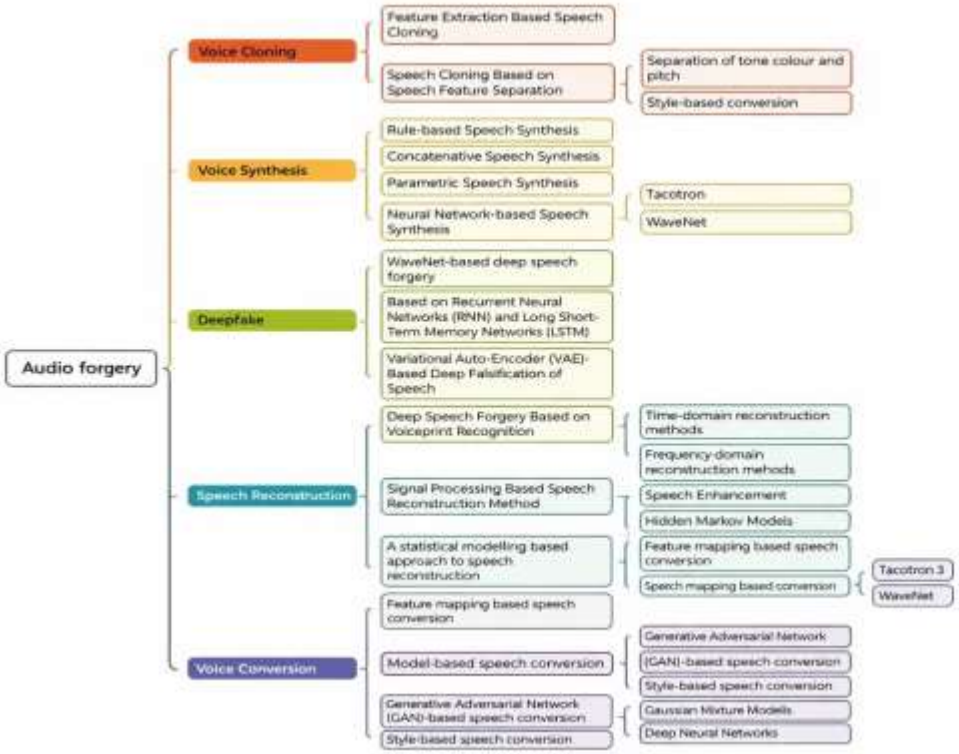


Figure.5 Audio forgery method

Despite all these improvements, subtle spectral cues and temporal inconsistencies still manage to remain and provide feasible markers for detection. Furthermore, voice reconstruction has an indispensable role in the rehabilitation, augmentation or completion of degraded or partially lost voice signals to ensure acoustic clearness and intelligibility of the voice. This process is usually a pre- or post-processing step of the deepfake process and thus improves the input quality provided to the voice cloning, speech synthesis and voice conversion processes.

2.3. Text-Forgery

Text forgery[40] is the manipulation or production of written text, from a hand edited to a text forged with artificial intelligence, and the intention to replicate or modify a source text. It may roughly be divided into manual (traditional) and deep learning based (deepfake) methods and each has a different characteristic and different challenge for detection as shown in the Figure 6.

Manual forgery Techniques: Manual Forgery Techniques In this paradigm, the forgers use simple text editors, word processors or document formatting software to insert, delete, substitute or rearrange lexical items at the word, sentence or paragraph level [41]. Such interventions may leave visible traces such as stylistic incongruity and disproportionate punctuation, grammatical anomalies and formatting irregularities or syntactic discontinuities. Elevated contextual inconsistencies as well as abrupt tonal shifts may also reveal tampering, thus providing evidence for forensic analysts. Although fairly simple compared to AI-driven procedures, manual forging is

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

still widely used in low resource or targeted manipulation situations, when the perpetrator has physical access to the original documents and wishes to make selective changes [42].

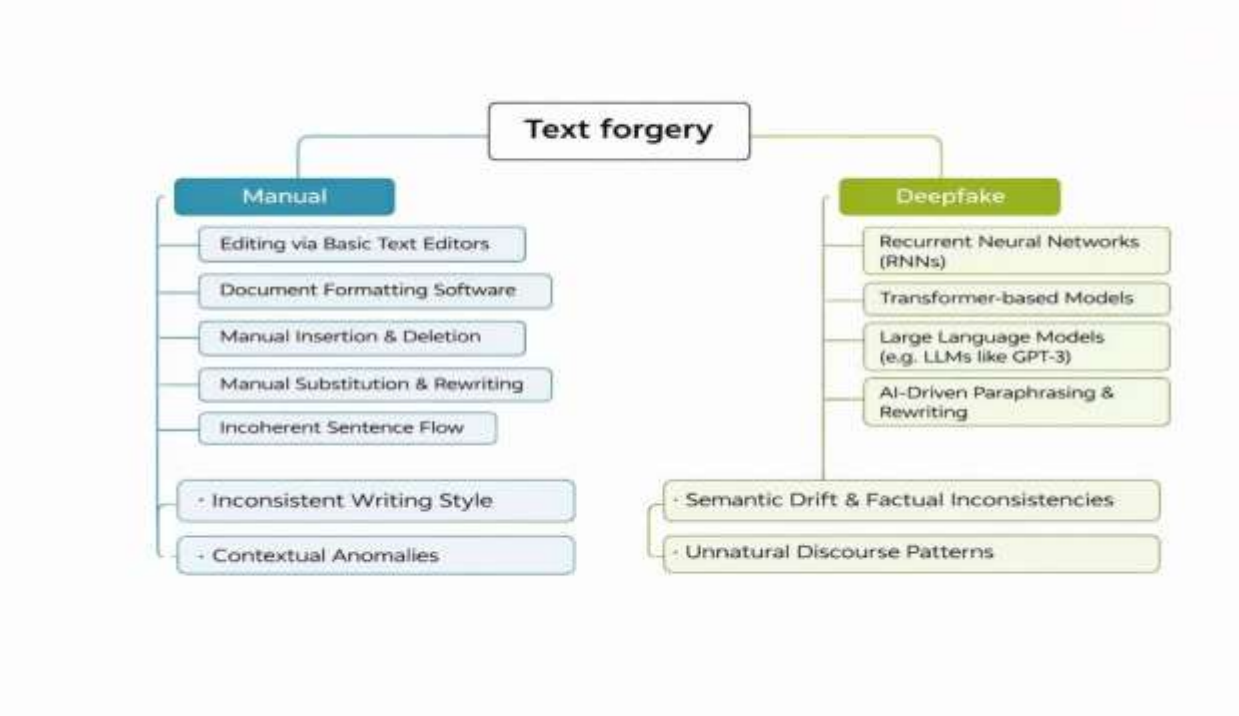


Figure.6 Text Forgery Method

Deep-fake forgery techniques: Deep-fake text forgery is a significant advancement from the old method of text manipulation, using methods of advanced natural language processing and deep learning to generate or alter text in a very precise manner. Core techniques include recurrent neural networks (RNNs) [43] , long short-term memory networks (LSTMs)[44] , Transformers [45], [46] and large language models (LLMs) [47], which are trained on large corpora in an attempt to understand complex patterns in selection, syntax and stylistic nuances with respect to particular authorship or domain [48]. Consequently, these models produce grammatically relevant, contextually appropriate and stylistically consistent text with the original source, making it difficult to distinguish between authentic and synthetic text.

Deep-fake text creation has three main approaches: automated paraphrasing of text while preserving original semantics [49] ,style transfer where linguistic peculiarities of target authors are emulated [50] ,and fully AI-assisted content generation where new text is generated in response to various prompts [51]. This triad of methodologies can produce results that are as minor as tweaks to existing documents, to completely synthetic compositions, such as articles or social-media posts that convincingly imitate the authorial voice.

Deepfake-generated texts may contain detectable markers despite their realistic appearance, such as semantic drift, factual inconsistencies, unusual discourse structures, and improbable word combinations. Q. Umer et al. [122] proposed an ensemble deep learning framework for detecting fraudulent cryptocurrency transactions, demonstrating how combining multiple models can improve detection accuracy in complex and deceptive data environments. This ensemble-based

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

strategy can be extended to deepfake text detection, where multiple linguistic and semantic models collaborate to identify subtle manipulation patterns.

In the domain of intelligent security systems, M. Malik et al. [123] developed a machine learning-based intrusion detection system for IoT environments. Their work emphasizes anomaly detection and pattern recognition in dynamic data streams, which is highly relevant to deepfake detection, as identifying abnormal linguistic patterns and inconsistencies in generated text can help distinguish between authentic and synthetic content.

Focusing directly on multimedia forgery, H. Ghous et al. [124] introduced a convolutional generative approach for detecting fake videos. Their research highlights how generative models can both create and detect manipulated content. The concepts from this work are particularly significant for multimodal deepfake detection, where insights from visual forgery detection can be integrated with textual analysis to improve overall system robustness.

Additionally, I. Rehman and H. Ghous [125] presented a structured review on market basket analysis using deep learning and association rules. Although focused on data mining, their methodology for discovering hidden patterns and associations can be adapted for deepfake text detection, where identifying unusual word relationships and co-occurrence patterns plays a crucial role in uncovering synthetic or manipulated text.

3. Multimodal forgery detection methods

Multimodal forgery detection methods substantially improve the accuracy and robustness with the integration of video-audio-text modalities to capture a complete capture of the forgery artifacts. These approaches can be divided mainly into two categories: manual feature extraction and deep learning-driven automatic feature extraction. A brief summary of some excellent studies is given in Table 1 after which the respective methodologies are analyzed in detail.

3.1 Manual feature extraction methods

In the early stages of deep fake detection research, scholars focused on manual engineered feature extraction methods that were designed to reveal forgeries through examining the physical characteristics, temporal consistency and statistical features of the audio-video-text signals. Notable research includes Singh et al.,[52] who developed a method for audio-visual forgery detection that assesses audio-video synchronization by analyzing inter-frame motion and matching waveforms with Dynamic Time Warping (DTW), revealing misalignments between lip movements and audio. Yu et al.[53] proposed a technique based on physical attribute consistency, evaluating lighting consistency to identify abnormal shadows in forged facial regions and assessing skin texture inconsistencies for AI-generated faces. In the audio domain, their method detects unusual acoustic features in AI voice cloning through the analysis of formants, incorporating audio-video synchronization to evaluate the coherence between lip movements and speech.

Raza et al. [54] proposed Multimodal Trace Method that uses Intramodality Mixer Layer (IAML) for the independent mixing of audio and video features followed by the Inter-modality Mixer Layer (IEML) for collaborative processing before it is fed to the multi-label classifier. This architecture produces better audio-video fusion and improves the deep forgery detection. Xu and co-authors [55] created a multimodal detection framework that prioritizes the identity consistency. Their approach combines the analysis of optical flow with the matching of the distribution of energy of the sounds, thus making it possible to detect video and speech forgeries by analyzing the patterns

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

of motion between frames and the patterns of anomaly in the audio. Khalid et al. [56] combined Mel Frequency Cepstral Coefficients (MFCC) as acoustic features with visual methods such as inter frame consistency analysis. Their work proves that classical signal processing strategies are still effective in multimodal detection situations.

Nevertheless, manually engineered feature-based methods have intrinsic limitations, especially in their ability to extract fine grained content in high quality forged media and to cope with the fast change of landscape technology in forgery. Despite these shortcomings, the above pioneering studies are invaluable, in fact, they have laid down some basic concepts like physical consistency checks and timeframe anomaly analysis which are now being used as guiding principles for further research in the field of deep-fake detection.

3.2 Deep learning feature extraction methods

The fast progress of the deep learning technologies has brought multimodal forgery detection into a new paradigm, where neural networks will learn complex patterns from the visual and audio modalities without any human intervention. The shift from hand-engineered feature extraction to data-driven representation learning yields quantifiable enhancements in detection accuracy, robustness and generalization, particularly when confronted with high-resolution synthesized content. Notable research includes: Boztepe et al. .[57] have created a multimodal deep learning approach utilizing VGG19 to extract visual features, Kernel Temporal Segmentation to summarize video content and MobileNetV2 along with LSTM layers to perform acoustic recognition. Middya et al. [58] used convolutional neural networks to process single frames of the video coupled with (MFCC)-based audio feature extraction, and LSTM units were used for capturing the temporal dependencies, which resulted in the effective detection of audiovisual forgeries.

Table.1 Summary of existing Multimodal detection methods

Method Category	Subcategory	Year	Authors	Core Technologies and Characteristics	Main Limitations
Manual Feature Methods	Temporal Consistency Analysis	2024	Singh et al. [52]	DTW, effective in detecting lip movement and audio misalignment	Difficult to capture subtle artifacts in HQ forgeries
	Physical Attribute Analysis	2024	Yu et al. [53]	Lighting consistency and formant detection, identifying forged facial and voice features	Insufficient cross-scenario adaptability
	Identity Consistency Analysis	2024	Xu et al. [55]	Optical flow analysis and audio energy distribution matching based on identity features	Depends on high-quality facial information
	Multi-Layer Feature Extraction	2023	Raza et al. [54]	IAML/MLM feature hybrid layers, structured multimodal feature processing	Limited feature expression capabilities

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

	Traditional Signal Processing	2021	Khalid et al. [56]	Combining audio features with inter-frame consistency analysis	Limited effectiveness against advanced forgery techniques
Deep Learning Methods	Emotional Feature Analysis	2024	Zhang et al. [59]	ResNet50 + Mel spectrogram integrating emotional features	Limited detection capabilities for non-emotional forgeries
	CNN–LSTM Architecture	2022	Boztepe et al. [57]	VGG19 + MobileNetV2 + LSTM, automatically learning complex features	Requires large-scale training data
Fusion Methods	Multimodal Temporal Fusion	2022	Middya et al. [58]	CNN + MFCC + LSTM, multimodal temporal modeling	High computational resource requirements
	Multimodal Attention Mechanism	2021	Rehman et al. [60]	3D-CNN + bidirectional gated recurrent units + attention mechanism	Complex architecture, difficult to optimize
	Spectral Temporal Analysis	2020	Lewis et al. [61]	ResNet50 + FFT + LSTM, joint time–frequency domain analysis	High training complexity
Transformer Models	Emotional Feature Analysis	2023	Davide et al. [62]	Extracts audio-visual features from input video over time using time-aware neural networks	Limited detection capabilities for non-emotional forgeries
	Multimodal Transformer	2023	Momina et al. [63]	Ensemble multimodal detector, audio-visual correspondence	Challenges in multi-source environments and high computational complexity
	Multimodal Transformer	2023	Ammarah et al. [64]	Audio-visual transformer ensemble, acoustic and visual forgeries	High data requirements and computational costs

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

	Self-Supervised Learning	2021	Akbari et al. [65]	ViT + HuBERT + BERT, self-supervised feature extraction	High pre-training costs

Zhang et al.[59] focused on facial expression and audio signal analysis enhancing emotion recognition and forgery detection through ResNet50 and Mel spectrogram analysis. Rehman et al. [60] integrated a 3D-CNN with bidirectional gated recurrent units for audio analysis, incorporating an attention mechanism to enhance multimodal feature fusion and detection accuracy for Deepfake detection. Lewis et al.[61] introduced a feature extraction technique using ResNet50 and Fast Fourier Transform for spectral analysis, also applying LSTM for time-series data modeling, facilitating simultaneous detection of forged videos and audio having extracted text.

Similarly, Davide and colleagues [62] invented a multimodal approach that fuses visual and audio features for the evaluation of video authenticity where the technology used temporal-aware neural networks deploying the analysis of audio-and video representations and modality discrepancies. Momina et al. [63] proposed a framework that uses three different and ensemble neural networks that focus on audio and video processing, with a cross-modal learning network that is designed to learn inconsistencies that can be found between the modes. Ammarah et al. [64] proposed AVTENet, an ensemble transformer-based model for image of audio-visual manipulations, it uses different transformer variants to extract salient features while making consensus predictions. In comparison with conventional CNNs, transformer methods make use of self-attention mechanisms that boost their capability to model complicated relationships in lengthy sequences and offer extra versatile feature extraction. Additionally, transformer-based multimodal learning, as shown in Akbari et al. [65] work, has exhibited impressive potential, utilizing a self-supervised framework with cross-modal feature learning and a Drop-Token strategy to enhance processing efficiency for high-resolution data, thus improving detection performance for both genuine and altered videos. However, these models necessitate large training datasets for optimal performance, exhibit quadratic computational complexity, and pose challenges in interpretation and deployment on resource-constrained devices due to their high parameter count and more demanding memory and processing requirements.

3.3 Feature-level fusion

Feature-level fusion through deep neural networks facilitates deep interaction of multimodal features, primarily categorized into early and late fusion methods. The main feature-level fusion detection methods are shown in Table 2 that early fusion integrates features from different modalities at the input layer or in low-dimensional feature spaces as illustrated in Figure 7a capturing fine-grained associations and enhancing detection

Table 2. Summary of Existing Multimodal Detection Methods

Fusion Type	Subcategory	Year	Representative Authors	Core Technologies and Characteristics	Main Limitations
Feature-Level Fusion	Early Fusion	2024	Ghadiya et al. [66]	CFA and HLGAtt for dynamically selecting and enhancing relevant	Requires large annotated training datasets

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

				audio–visual features	
		2023	Yang et al. [19]	CNN with attention mechanism for capturing fine-grained multimodal associations	Weak processing capability for asynchronous modal data
		2022	Miao et al. [20]	Res-Net + FFT, integrating audio–visual data at low-level feature stage	Susceptible to noise
	Late Fusion	2024	Yoon et al. [67]	CNN + Transformer for improved robustness with asynchronous modal data	Potentially misses fine-grained multimodal associations
		2024	Sharma et al. [68]	Feature vector splicing with fully connected networks	Limited interaction information, constrained expressiveness
		2023	Fu et al. [69]	Multi-level feature separation with hierarchical fusion	Difficult to optimize, complex model training
		2022	Concas et al. [70]	ResNet50 + CNN, modal combination at late fusion layer	Lack of deep inter-modal interaction
	Hybrid Fusion	2023	Liu et al.[71]	Meta-learning-based dynamic multimodal fusion framework (MetaMMF)	High implementation complexity
		2022	Liu et al.[72]	M2HF multi-level fusion network combining early and late fusion advantages	High model complexity, difficult to implement
Decision-Level Fusion	Score-Level Fusion	2024	Gao et al. [23]	TAD framework with MLP non-linear fusion of texture and forgery scores	Performance degrades with insufficient single-modality information

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

	Single-Level Decision Fusion	2024	Kiziltepe et al. [73]	Multiple classifiers with averaged prediction scores	May not fully utilize all feature information
	Audio–Image Fusion	2023	Shaikh et al. [74]	CNN-based feature extractors with weight assignment	Computational complexity with complex modality representations
	Multi-model Ensemble	2023	Shelke et al.[21]	CNN + RNN + SAE ensemble with voting strategies	Unable to capture deep inter-modal relationships
	Quantum Cognitive Fusion	2021	Gkoumas et al. [75]	Quantum state modeling with POVM measurement	Theoretically complex, high implementation difficulty

performance studies include Yang et al.[19]who utilized CNNs with attention mechanisms for early fusion of audio-visual data; Miao et al.[20] who combined FFT with image feature extraction in Res-Net and Ghadiya et al.[66] who proposed a weakly supervised video anomaly detection framework using an early fusion strategy with a Cross-Modal Fusion Adapter and Hyperbolic Lorentzian Graph Attention mechanism. These approaches tackle information imbalance and feature distinction challenges, achieving superior performance in violence and nudity content detection on the XD-Violence and NPDI datasets, thereby highlighting the effectiveness of early fusion in multimodal detection tasks.

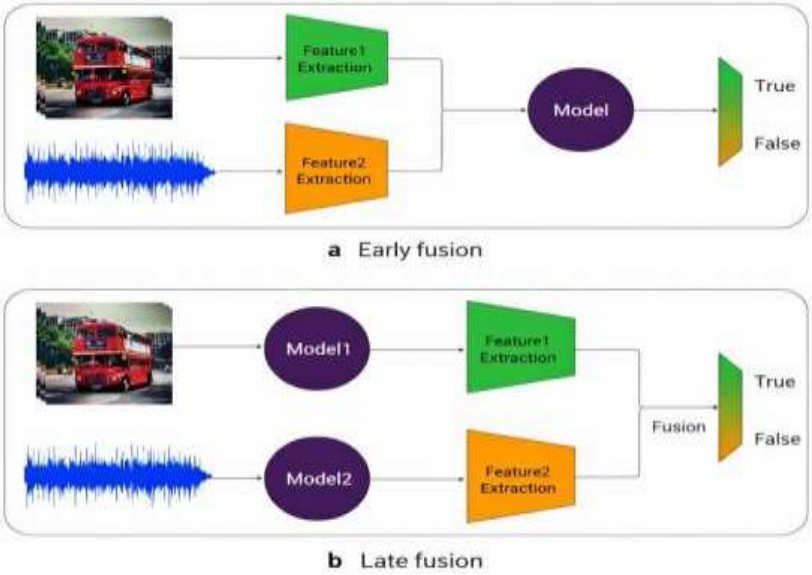


Figure.7 Feature-level fusion: a. early fusion; b. indicates late fusion

Contrasting to early fusion approaches, late fusion methodologies allow to obtain high-level features from each modality separately; afterwards they are integrated using features, e.g. feature

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

concatenation or gating mechanisms. This paradigm has been demonstrated to be superior in providing resilience when addressing asynchronous modal data as demonstrated in Figure 7b. Yoon et al. [67] used separate convolutional neural networks for audio and visual streams and used a Transformer further to mediate multimodal interactions in the late-fusion stage. Sharma et al. [68] concatenated audio and video feature with text vectors and fed the concatenated vectors into a fully connected network to obtain the accurate deep fake media detection. Fu et al. [69] proposed a multi-level feature separation scheme, perform partial fusion at multiple hierarchical levels with modalities aligned at the last fusion stage. Concas et al.[70] applied ResNet50 and CNN to extract video and audio features, respectively, merging these modalities at the late fusion layer. Liu et al.[71] proposed the MetaMMF framework, a meta-learning based dynamic fusion approach for micro-video recommendations that adapts fusion parameters dynamically for each micro-video, capturing complex visual, audio, and text interplay. The dual-structure consists of a meta-information extractor and a meta-fusion learner that respectively transform features into high-order representations and contextualize neural network parameters for effective fusion. Furthermore, Liu et al.[72] introduced the Multi-level Multimodal Hybrid Fusion Network (M2HF) that performs early fusion of visual features from the CLIP model with audio and motion features to create audio-guided and motion-guided visual features for improved modal interactions. The framework establishes relationships between fused features and textual content using late fusion to enhance text-video retrieval results.

3.4 Decision-level fusion

Decision-level fusion uses a multi-expert system architecture. With Each modal branch independently completes predictions, integrating final results through weighted voting or probabilistic fusion with main decision-level fusion detection methods are compiled in Table 2. These methods highlight the versatility of integrating various modal branches having key advantage of decision-level fusion is the ability to reuse existing single-modal detection models. The is visually illustrated in Figure.8. Shelke et al. [21]proposed a decision-level fusion method based on multimodal ensemble. The method combines prediction results from various deep learning models, including CNNs, RNNs, and SAEs. These enhancements significantly improved the accuracy and robustness of deepfake detection. Gao et al.[23] introduced the Texture and Artifact Detector (TAD) framework, employing multimodal decision-level fusion. They compared various score-level fusion methods, utilizing Multi-Layer Perceptron (MLP) for nonlinear fusion of texture and artifact scores in single-model scenarios.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

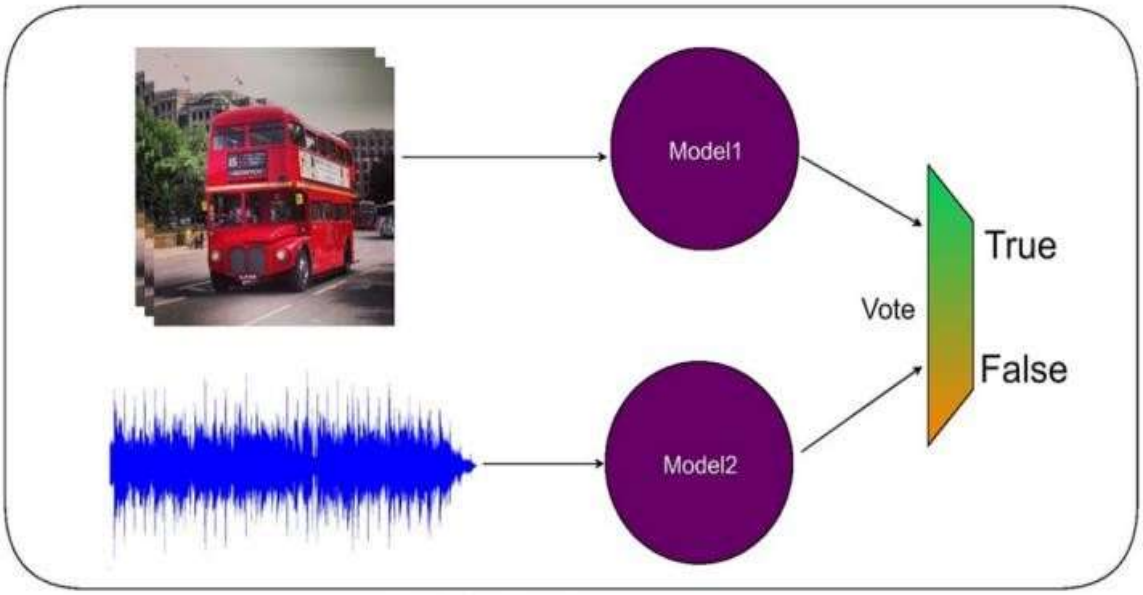


Figure.8 Decision-level convergence

For ensemble model scenarios, a two-stage fusion strategy was adopted, which consists of branch internal MLP fusion and the cross-branch averaging for several types of forgery. Kiziltepe et al.[73] suggested a hierarchical and hybrid fusion framework, which applied a multi-classifier prediction fusion method with prediction scores combination using average and the highest probability class to improve the model performance and efficiency. In the meanwhile, Shaikh et al.[74] proposed MAiVAR, a multimodal fusion framework with less computation complexity and enhanced human action recognition by fusing the audio and video characteristics and Dominant Head Fusion method. Meanwhile, for video sentiment analysis, Gkoumas et al.[75] developed a quantum cognitive decision-level fusion method which uses quantum states in an emotional Hilbert space to describe utterances. It uses training data to estimate the parameters of emotion classifiers using as a model a mutually incompatible observance. By combining these factors, the approach is used to determine the final emotional state and a probability threshold of 0.5 is used to categorize the utterances as either positive or negative. By gathering complex emotional connections between modalities this quantum technique hampers the accuracy of sentiment analysis.

But there still exist great challenges in practical implementations of fusion models. The first concern is modality imbalance. When one modality quality is low, the fusion model must be robust enough to maintain total performance. The second problem is multimodal temporal alignment, particularly when processing audio and video input at various sampling rates. Improving model interpretability remains a key research direction. Visualizing multimodal attention weights can enhance forensic investigation of forgeries.

4. Public datasets and benchmarking

Standardized evaluation measures and high-quality public datasets are essential for multimodal forgery detection[76] research. The section will include frequently used video, audio, text and multimodal datasets together with common evaluation metrics. These datasets are essential for creating algorithms and evaluating their effectiveness.

4.1 Video Forgery Datasets

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

The development of video forgery detection research is significantly connected with the development of benchmarking datasets. Starting with rudimentary early versions of synthetic samples and ending up with the latest state-of-the-art forgeries, datasets have continually been increasing in scale, variety and authenticity and therefore providing invaluable support to the improvement of detection algorithms. Table 3 provides a quick overview of some of the key attributes of several key datasets

Table3. Dataset of video forgery

Dataset	Year	# Real Videos	# Fake Videos	Dataset Link
FaceForensics++ (FF++) [77]	2019	1,000	4,000	https://github.com/ondyari/FaceForensics
DFFD [78]	2020	1,000	3,000	https://github.com/zhangxin-xd/Deepfake-Datasets
DFDC [79]	2020	23,654	104,500	https://ai.meta.com/datasets/dfdc/
Celeb-DF++[80]	2020	590	5,639	https://github.com/yuezunli/celeb-deepfakeforensics
DeeperForensics-1.0[81]	2020	50,000	10,000	https://github.com/EndlessSora/DeeperForensics-1.0
WildDeepfake[82]	2020	3,805	3,509	https://github.com/OpenTAI/wild-deepfake
FFIW_10k[83]	2021	10,000	10,000	https://github.com/tfzhou/FFIW
ForgeryNet[84]	2021	99,630	121,617	https://yinanhe.github.io/projects/forgerynet.html
DeepFakeFace[85]	2023	30,000	90,000	https://github.com/OpenRL-Lab/DeepFakeFace
DiFF[86]	2024	23,661	537,466	https://github.com/xaCheng1996/DiFF
DiffusionFace[87]	2024	30,000	600,000	https://github.com/Rapisurazurite/DiffFace
TalkingHeadBench[88]	2025	2,312	2,984	https://huggingface.co/datasets/luchaoqi/TalkingHeadBench

and the following sections will describe their technical characteristics and their importance for research in more detail.

The FaceForensics++ (FF++) [77] dataset forms a baseline dataset for deep face recognition to identify visual deep fakes by providing manipulated facial videos created using various face forgery techniques. DFFD [78]further pushes this trajectory by focusing on digitally manipulated facial content to evaluate the robustness of face manipulation detection systems. The large-scale data set of DFDC [79]covers the diversity of real-world subjects, scenes, and compression levels

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

to make it an important benchmark for generalizable deep-fake detection, and Celeb-DF++[80] is a particular focus on high-quality celebrity deep-fakes with few artifacts that are challenging for conventional detectors. DeeperForensics-1.0 [81]is focused on addressing the real-world deployment scenarios, i.e. incorporating the environmental perturbations including illumination variability, motion blur and occlusions, whereas WildDeepfake[82]corresponds to unconstrained distribution of in-the-wild deep-fake collected from the internet, thus considering the domain shift problem. Moving away from binary classification, FFIW_10k[83] proposes localization-aware deep fake detection for unconstrained scenarios with spatial annotations, and ForgeryNet[84] proposes to provide comprehensive classification, spatial, and temporal grounding annotations to achieve fine-grained forgery analysis. Attacking emerging generative threats DeepFakeFace[85], DiFF[86], and DiffusionFace[87] propose diffusion-model-based facial forgeries of different scales to enable systematic studies of the performance of detectors against diffusion synthesis of high fidelity. TalkingHeadBench[88] evaluates diffusion based talking head generation in order to support the evaluation of deep fake detection under realistic facial motion and expression dynamics.

4.2 Audio forgery dataset

Table 4 summarizes the current representative audio datasets, and the following section will provide specific introductions to the characteristics and application scenarios of each dataset.

The CodecFake dataset [89]consists of a large dataset of authentic and fake speech recordings that have been produced through codec-based text-to-speech and voice conversion pipelines. It is explicitly aimed at evaluating the robustness of the audio deepfake detection models against the emerging codec-level synthesis attack. AUDETER [90]is an open-world large-scale audio deepfake repository where millions of real and artificially generated speech samples are available which are generated using various text-to-speech systems and neural vocoders.

Table4. Dataset of Audio Forgery

Dataset	Year	# Real Audio	# Fake Audio	Dataset Link
CodecFake[89]	2024	≈42,000	≈45,000	https://github.com/roger-tseng/CodecFake
AUDETER[90]	2025	Millions	Millions	https://github.com/mike-qz-wang/AUDETER
WaveFake[91]	2021	≈13,100	≈104,885	https://github.com/RUB-SysSec/WaveFake
ASVspoof 2021 (DF Track)[92]	2021	≈18,000	≈163,000	https://github.com/asvspoof-challenge/2021

It allows us to evaluate language dependency in vulnerabilities in audio deepfake detection systems. WaveFake[91] can be used as an initial benchmark dataset for audio deepfake research and contains both real speech and a significant amount of synthetic audio generated with several architectures of neural speech synthesis. It has been widely used for baseline evaluations. Finally, the ASVspoof 2021[92] offers a standardized benchmark dataset, which includes large amounts of real and spoofed speech samples produced by text-to-speech, voice conversion, and replay attacks. It is widely used for just comparison in the fairness and performance benchmarking of audio deepfake detection methods.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

4.3 Text Forgery dataset

The body of work on textual deepfakes is based on several publicly available datasets that represent a plethora of synthetic and manipulated textual contents shown in table 5. TweepFake[93] contains a curated set of real and bot-generated tweets and thus facilitates studies of the identification of deepfake content on social media platforms such as Twitter. MAGE[94] consists of a large scale, multi-domain data set with both human and machine generated text that is designed to test the cross-domain generalization and robustness of detection methods. DeepfakeTextDetection[95] provides an in-the-wild collection of real and fake generated texts from several sources that can be used for developing and evaluating text deepfake detection systems. GeneratedTextDetection[96] introduces a collection of authentic and synthetically generated academic abstracts and thus provides a way of testing detection methods in the context of scientific text forgery. Collectively, these datasets constitute a rich resource base for the further development of the detection of machine-generated and manipulated textual content in a variety of different domains and applications.

Table5. Dataset of Text Forgery

Dataset	Year	#Real Text	#Fake Text	Dataset Link
TweepFake [93]	2021	~12,000	~12,000	https://github.com/tizfa/tweepfakedeepfake_text_detection
MAGE [94]	2023	~500,000	~500,000	https://github.com/yafuly/MAGE
DeepfakeTextDetection [95]	2023	varies	varies	https://github.com/jmpu/DeepfakeTextDetection
GeneratedTextDetection [96]	2022	~100	~100	https://github.com/vijini/GeneratedTextDetection

4.4 Multimodal Forgery dataset

The scenario of multimodal deep-fake datasets has greatly expanded for audio-visual-text forgery detection research. MAVOS DD[97] provides an extensive benchmark containing 25,195 real and 35,169 fake videos in eight languages, thus enabling the model evaluation in in-domain and open-set settings. PolyGlotFake[98] further expands this scope to offer 766 real and 14472 fake videos in 7 languages, to study cross-lingual multimodal deep fake detection. FakeAVCeleb[99], which includes around 500 real and 19,500 fake samples, focuses on high-fidelity audio - video manipulations and is conducive to the development of models that are

Table6. Dataset of Multimodal forgery

Dataset	# Real Videos	# Fake Videos	GitHub / Dataset Link
MAVOS-DD[97]	**25,195**	**35,169**	**https://github.com/CroitoruAlin/MAVOS-DD (Hugging Face)**
PolyGlotFake[98]	**766**	**14,472**	**https://github.com/tobuta/PolyGlotFake (GitHub)**

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

FakeAVCeleb[99]	**500**	**19,500**	**https://github.com/DASH-Lab/FakeAVCeleb (GitHub)**
LAV-DF[100]	**36,431**	**99,873**	**https://github.com/ControlNet/LAV-DF (HyperAI)**
AV-Deepfake1M[101]	**286,721**	**860,039**	**https://github.com/ControlNet/AV-Deepfake1M (GitHub)**
SWAN-DF[102]	**720**	**24,000**	**https://swan-df.github.io/)**

able to capture the subtleties of inconsistencies between speech, visual content and their respective transcripts. LAV- DF[100] provides 36,431 real and 99,873 fake videos with the aim of large-scale detection and localization AV Deepfake1M[101], with 286,721 real and 860,039 fake videos, is one of the largest datasets that can be used to train scalable detection architectures. Finally, SWAN-DF[102] consisting of 720 real and 24,000 fake videos, focuses on high fidelity manipulations and allows for precise assessment of the robustness for detection in controlled scenarios. As shown in Table6, collectively these datasets form a diverse foundation for the advancement of AI-driven multimodal deep fake detection.

5. Main Challenges and Research Directions

Over the past years, methods of detection of forgery with the use of multimodal video and audio have become possible due to the development of deep learning techniques, which provide impressive performance. Nevertheless, the development of the forgery methods also brings many new issues that should be studied and addressed. At the same time, such innovations as large-scale pre-training models and adversarial learning are new directions of study in the future. This part will analyze the major issues facing the current field and how it is likely to be in the future.

5.1 Progressive Technological Confrontation

Over the past few years, the Deepfake technology has shown trends of exponential growth, especially with the introduction of the next-generation generative models. These developments have continuously pushed the boundaries of forged content in visual fidelity, temporal consistency and multimodal fusion potential. As a result, the value of the traditional methods of detection is facing the threat of becoming obsolete, and it is manifested most of the time in the following ways: Deepfake ability has greatly enhanced the ability to create high-resolution content. As opposed to the poor quality of early forged videos, which were often characterized by a lower quality of images or lack thereof, modern high-resolution DeepFake videos are indistinguishable in terms of fidelity to detail, which makes it difficult to detect an anomaly with a mere visual examination of features. In addition, more sophisticated generative techniques can be used to simulate important details like change of light, expression, and texture of the skin that adds addition to the believability of forged content.

The detection ability towards forged content has significantly reduced with advancements in the breakthrough technology in audio-visual syncing[18]. The newest generation of Deepfake technologies is able to produce the accuracy of the time convergence between the audio and visual cues, which means that the videos that have been forged are significantly more realistic, and the systems that detect them face a great deal of difficulty because they are based on the principle of the temporal inconsistency analysis. This technological innovation has almost eliminated the disparity in synchronization capabilities between the artificially generated contents and the authentic images, thus bringing new challenges in authenticity verification of digital contents.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

The use of cross modal Deepfake technologies has also made detection harder. Over the past years, there has been a very fast advancement in text-based types of audio-visual generation technologies that allow attackers to automatically produce high-quality fake audio-visual content using textual descriptions[103]. As an example, a text prompt can automatically create the corresponding speech and synchronized video, even imitating the voice qualities and facial expressions of certain people. The further development of such technologies does not only increase the realism of Deepfakes but also makes the single-mode methods of detection ineffective.

5.2 Problems of generalizability and robustness in practical applications

Although recent Deepfake detection technology has made significant advances in pivotal laboratory settings, there are still significant challenges faced by these technologies when it comes to generalizability and resilience in the field. The major factors affecting the generalization capacity of a model include dispersed data distributions, content diversity, as well as the ability to adapt to cross-domain, all of them limiting the practical effectiveness of detection systems. The most dominant bottleneck is cross-domain adaptability; most of the detection models are doing well on small-scale training sets but fail disastrously when faced with non-homogeneous data distributions, e.g., differences in camera hardware, video compression codecs and light conditions. Diversity in content is a major problem to detection systems since current models often rely on well-known trends in training data to classify them. The Deepfake material in real-life situations can include unrelated subjects, languages, scenes, or different generation methods, thus causing conspicuous changes in the detection accuracy. In addition, issues of data quality in the areas of operation further undermine the strength of detection mechanisms. Numerous Deepfake detection systems can perform well with high-quality, uncompressed data, but in real-world scenarios, videos are repeatedly encoded, transmitted as coded data, or run through a social media system, and these steps can easily destroy many artefacts of forgery needed to load the system to detect them. Therefore, the three areas that should be prioritized in future studies in order to make Deepfake detection systems more generalizable include domain adaptation, data augmentation techniques, self-supervised learning, and identification of robust features. These methods are to strengthen the stability and flexibility of the detection systems under various working conditions.

5.3 Counter-attack and counter-defense

Over the last few years, the fast development of Deepfake technologies has triggered more advanced adversarial attack methods targeting detection systems, thus drawing the attention of a large number of scholars. The most sophisticated Deepfake detectors, as empirical studies showed[104], [105], can degrade significantly when subjected to carefully designed adversarial examples.

Perturbation Attacks have been given a lot of attention across the continuum of adversarial modalities due to their invisibility and effectiveness. The attack process involves adding infinitesimal perturbations to forged multimedia artifacts that are not perceivable to the eyes or ears of human subjectivity and this, in turn, undermines the judgment of the detection model but leaves the human subjective perception intact [106]. At the same time, more insidious Steganographic Attacks are emerging, taking on information-hiding methods to insert misleading properties into carrier media, and thus compromising traditional detection methods to track the forgery signatures [107].

The growing virulence of adversarial attacks has led to a broad range of defense mechanisms championed by researchers. Fundamentally, adversarial training strengthens a model by introducing it to synthetically generated adversarial instances so that the model becomes more

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

sensitive to such perturbations. Concurrently, detection methods based on spatial - temporal coherence have solved some of the shortcomings previously experienced in single - frame or fragmentary analysis. By analysing intra-frame consistency and the temporal information of acoustic signals, these techniques can be applied to detect forged material due to inconsistency[108], [109], [110], [111].

More critically, the introduction of the procedures of Explainable Artificial Intelligence (XAI) overcome the shortcomings of traditional black box models in terms of interpretability. XAI modalities not only enable the transparency of the outcomes of detection, but also provide explicit rationales for decisions that are based on inferred evidence, providing new insights into the optimization of defensive strategies [112], [113].

Even though defensive weapons now have so much more success, the old paradigms of adversarial attacks don't stay the same–that is, they are evolving–and reaching places that Deepfake detectors have not seen before. Future studies, then, ought to focus on the joint optimization of disparate defense mechanisms, by combining mechanisms like adversarial training with spatial-temporal consistency checking to give rise to more robust composite defense pipelines. Similarly, the emergence of multimodal detection systems that can simultaneously absorb audiovisual data and execute dynamically adjustable protective systems will become a key course in the field. The innovation breakthroughs on these new frontiers are expected to provide more reliable guarantees on the maintenance of authenticity and protection of digital information.

5.4 Large scale pre-trained model combinations

Over the past years, the large-scale vision-language-based deep-fake detection technologies have achieved significant advancement. The multimodal models of learning including CLIP [114], BLIP [115], and others have also shown excellent performance in detecting information compared to the traditional single-mode models of learning because they have the ability to collectively model multimodal information-text, video, and audio. These models can successfully utilize cross modal associative properties thus demonstrating better flexibility and resilience in cross modal deep-fake detection. Regarding model architecture, multimodal Transformers have been introduced, which have introduced technological evolution in deep-fake detection. These architectures are able to learn more deep forgery feature patterns with the help of multimodal feature fusion mechanisms. In particular, scientists have come up with spatial-temporal attention systems which capture semantic consistency among video frames and combine audio-feature analysis to detect minor variations in audio-visual synchronization thereby significantly enhancing the detection accuracy. As an illustration, Chen et al.[116]introduced a technique of detecting deep-fakes based on spatial-temporal variables, using 3D convolutional neural networks, which are used to capture spatial-temporal data[117] boosted detection accuracy through the attention process and the spatial-temporal consistency, especially detecting the minor and local characteristics in forged videos.

The subject of particular interest is the fact that a number of few-shot and zero-shot learning models that rely on large models have opened up new opportunities in deep-fake detection These methods leverage the powerful representation capabilities of pre-trained models, enabling effective detection with only a few labeled samples or even without specific training data. The property not only significantly lowers the annotation costs of their data, but also enables the detection systems to respond more appropriately to the quickly changing forgery technology. However, the methodologies still face major challenges, including the high consumption of computational resources, high inference latency, and high levels of data dependencies. Further studies should therefore consider more effective model architectures such as lightweight transformer structures,

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

knowledge distillation processes, as well as adaptive inference optimization processes to make certain that computational efficiency is achieved and at the same time detection abilities are improved.

5.5 Feasibility analysis of technical paths

Even though the above sections have suggested promising lines like adversarial training, multimodal detection, and pre-trained large models, more work needs to be done to streamline the viability and applicability of the technical paths, thus reducing the gap between theoretical research and practical application.

Model-level Optimization: Massive models like CLIP and BLIP are capable of representational power, but their computational complexity makes them impossible to use in practice. To improve their viability, they can be implemented with lightweight Transformer architectures - e.g. Mobile-ViT [118] and Tiny-ViT [119] -as well as knowledge-distillation methods to reduce huge detection networks without compromising performance. These approaches make real-time inference of edge devices more feasible, which is a requirement of forensic activity in the real world, such as in mobile content moderation and in-browser video verification.

Data-level Generalization: To overcome the well documented shortcomings of cross domain performance, it is wise to look into the domain adaptation training paradigms and meta learning frameworks as promising avenues. Techniques like test time adaptation (TTA) and unsupervised domain adaptation give permission to models for re-calibrate to advance distributions without the need for exhaustive re-training. Moreover, supplementing the training pipelines with synthetic data created by controllable generative models can mimic a distribution of forgery scenarios to generally improve generalizability (such as StyleGAN - 3 [120] and VALL - E).

Deployment-level Considerations: Based on an application level, detecting systems must strike an equilibrium between accuracy, inference latency and scalability. Trade-offs Pragmatic trade-offs can be provided by hybrid detection pipelines combining fast, low-level detectors e.g. eye-blink and audio-onset detectors with more accurate, although slower, multimodel detectors. At the same time, the inclusion of explainable AI (salience maps and attention visualizations) can build the trust of users and guarantee forensic traceability in a judicial or journalistic setting.

The detecting system to be built in the future must be modular, dynamical, and has the capacity to learn throughout a lifetime. Detection models can be dynamically updated by using continual-learning frameworks that exploit online feedback, including user-generated labels or correction cues, so that they can be robust to the changing forgery methods. Moreover, the implementation of such detection functions into the cloud-based content distribution systems represents an attractive implementation plan, which allows scaling and real-time protection against synthetic media threats.

6. Conclusion

Deepfake detection developed into multimodal systems, thus significantly increasing the ability to identify extremely advanced audio-visual-text forgeries. The review summary brings together recent advancements in deep learning-based detection paradigms, both traditional feature-based methods, neural networks and multimodal fusion approaches. The above methodologies show great potential in a range of use cases - ranging from content moderation in social media ecosystems to forensic judicial analyses and fraud prevention initiatives; their effectiveness, however, is limited by issues related to cross-domains generalization and the needed computational power.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Progress in this area requires a unified focus on the development of light and scalable models, in addition to the implementation of adaptive defensive strategies. Among these, the strategic use of spatial-temporal consistency analysis is outstanding as a powerful countermeasure to the ever-changing tactics of forgery. To make such approaches valid, the creation of common evaluation benchmarks is imperative, in order to ensure robust performance in a variety of real-world scenarios. In the end, the incorporation of deepfake detection into cloud-based content verification systems has the potential to protect digital media. Nonetheless, for such results to be realized, there will need to be long-term interdisciplinary collaboration between academia and industry.

Future research endeavors will want to explore human-centric multimodal deepfake detection frameworks that, in addition to the improvement of audio-visual-textual inconsistency analyses, explore the psychological consequences of deepfake exposure, namely distress, anxiety and depression that are engendered by misinformation and identity misuse [121] . Empirical research is needed to understand the relationship between the repeated exposure to deepfakes and mental health, as well as social behavior, and to design a system that can initiate early intervention to protect cognitive well-being in the digital world. By leveraging the power of spatial-temporal and cross-modal consistency analytics in combination with behavioral exposure modeling, future systems may help early identify mental degenerative synthetic content on social platforms. Such interdisciplinary scholarship will begin to bridge technological advancements with practical psychological safeguards that will lead in turn to a combination of strong detection mechanisms as well as societal resilience.

References:

[1] I. Goodfellow *et al.*, “Generative adversarial networks,” *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.

[2] C. Yankun, “Auto-Encoding Variational Bayes.”

[3] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” Dec. 2020, [Online]. Available: http://arxiv.org/abs/2006.11239

[4] A. Vaswani *et al.*, “Attention Is All You Need,” Aug. 2023, [Online]. Available: http://arxiv.org/abs/1706.03762

[5] I. Korshunova, W. Shi, J. Dambre, and L. Theis, “Fast Face-swap Using Convolutional Neural Networks,” Jul. 2017, [Online]. Available: http://arxiv.org/abs/1611.09577

[6] H. Ding, K. Sricharan, and R. Chellappa, “ExprGAN: Facial Expression Editing with Controllable Expression Intensity.” [Online]. Available: www.aaai.org

[7] Y. Jia *et al.*, “Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis,” Jan. 2019, [Online]. Available: http://arxiv.org/abs/1806.04558

[8] A. van den Oord *et al.*, “WaveNet: A Generative Model for Raw Audio,” Sep. 2016, [Online]. Available: http://arxiv.org/abs/1609.03499

[9] N. , & J. D. Diakopoulos, “Anticipating and addressing the ethical implications of deepfakes in the context of elections. New Media & Society,” vol. 23(7), 2072-2098., no. 2021, 2020.

[10] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, “Generalization of Audio Deepfake Detection,” International Speech Communication Association, May 2020, pp. 132–137. doi: 10.21437/odyssey.2020-19.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

[11] G. Yadav *et al.*, "Sciences of Conservation and Archaeology Psychological Trauma and Legal Challenges of Deep fake Technology," vol. 37, p. 1, 2025, doi: 10.48141/sci-arch-37.1.25.14.

[12] S. Ahmed, M. Masood, A. W. T. Bee, and K. Ichikawa, "False failures, real distrust: the impact of an infrastructure failure deepfake on government trust," *Front. Psychol.*, vol. 16, 2025, doi: 10.3389/fpsyg.2025.1574840.

[13] X. Li, K. Li, Y. Zheng, C. Yan, X. Ji, and W. Xu, "SafeEar: Content Privacy-Preserving Audio Deepfake Detection," in *CCS 2024 - Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, Association for Computing Machinery, Inc, Dec. 2024, pp. 3585–3599. doi: 10.1145/3658644.3670285.

[14] Z. Yang, J. Liu, Z. Wu, P. Wu, and X. Liu, "Video Event Restoration Based on Keyframes for Video Anomaly Detection," Apr. 2023, [Online]. Available: http://arxiv.org/abs/2304.05112

[15] X. Zhang, J. Yi, J. Tao, C. Wang, and C. Zhang, "Do You Remember? Overcoming Catastrophic Forgetting for Fake Audio Detection," Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.03300

[16] W. Moon, S. Hyun, S. Park, D. Park, and J.-P. Heo, "Query-Dependent Video Representation for Moment Retrieval and Highlight Detection," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.13874

[17] G. Pei *et al.*, "Deepfake Generation and Detection: A Benchmark and Survey," May 2024, [Online]. Available: http://arxiv.org/abs/2403.17881

[18] W. A. Jbara, N. A. H. K. Hussein, and J. H. Soud, "Deepfake Detection in Video and Audio Clips: A Comprehensive Survey and Analysis," *Mesopotamian Journal of CyberSecurity*, vol. 4, no. 3, pp. 233–250, Dec. 2024, doi: 10.58496/MJCS/2024/025.

[19] W. Yang et al., "'AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake,' in IEEE Transactions on Information Forensics and Security," vol. 18, 2023.

[20] Z. T. Q. C. N. Y. and G. G. C. Miao, "'Hierarchical Frequency-Assisted Interactive Networks for Face Manipulation Detection,' in IEEE Transactions on Information Forensics and Security," vol. vol.17, 2022.

[21] M. P. Shelke, N. Ranjan, A. Kharade, P. Gaikwad, S. Arakh, and A. Kalore, "Combining Computer Vision Techniques and Intraframe Noise Methods to Detect a Deepfake," in *Data Science and Intelligent Computing Techniques*, Soft Computing Research Society, 2023, pp. 471–480. doi: 10.56155/978-81-955020-2-8-43.

[22] Z. Akhtar, T. L. Pendyala, and V. S. Athmakuri, "Video and Audio Deepfake Datasets and Open Issues in Deepfake Technology: Being Ahead of the Curve," Sep. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/forensicsci4030021.

[23] J. Gao *et al.*, "Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection," *Eng. Appl. Artif. Intell.*, vol. 133, Jul. 2024, doi: 10.1016/j.engappai.2024.108450.

[24] Z. Li *et al.*, "FakeRadar: Probing Forgery Outliers to Detect Unknown Deepfake Videos," Dec. 2025, [Online]. Available: http://arxiv.org/abs/2512.14601

[25] D. Singh, P. Singh, R. Jena, and R. S. Chakraborty, "An image forensic technique based on JPEG ghosts," *Multimed. Tools Appl.*, vol. 82, no. 9, pp. 14153–14169, Apr. 2023, doi: 10.1007/s11042-022-13699-x.

[26] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "GANimation: Anatomically-aware Facial Animation from a Single Image," Aug. 2018, [Online]. Available: http://arxiv.org/abs/1807.09251

[27] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," Mar. 2020, [Online]. Available: http://arxiv.org/abs/1912.04958

[28] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation," Sep. 2018, [Online]. Available: http://arxiv.org/abs/1711.09020

[29] H. Cheng, Y. Guo, T. Wang, L. Nie, and M. Kankanhalli, "Diffusion Facial Forgery Detection," in *MM 2024 - Proceedings of the 32nd ACM International Conference on Multimedia*, Association for Computing Machinery, Inc, Oct. 2024, pp. 5939–5948. doi: 10.1145/3664647.3680797.

[30] S. Waseem, S. A. R. S. Abu Bakar, B. A. Ahmed, Z. Omar, T. A. E. Eisa, and M. E. E. Dalam, "DeepFake on Face and Expression Swap: A Review," *IEEE Access*, vol. 11, pp. 117865–117906, 2023, doi: 10.1109/ACCESS.2023.3324403.

[31] A. Di Pierno, L. Guarnera, D. Allegra, and S. Battiato, "End-to-end Audio Deepfake Detection from RAW Waveforms: a RawNet-Based Approach with Cross-Dataset Evaluation," Apr. 2025, [Online]. Available: http://arxiv.org/abs/2504.20923

[32] E. Ferrari, R. Spolaor, and V. P. Janeja, "Audio deepfakes: A survey." [Online]. Available: https://www.tampabay.com/florida-politics/buzz/////////

[33] Y. Ahmadiadli, X.-P. Zhang, and N. Khan, "Beyond Identity: A Generalizable Approach for Deepfake Audio Detection," May 2025, [Online]. Available: http://arxiv.org/abs/2505.06766

[34] E. Ozgen and S. Y. Altay, "Detecting Audio Copy-Move Forgeries on Mel Spectrograms via Hybrid Keypoint Features," *Applied Sciences (Switzerland)*, vol. 15, no. 21, Nov. 2025, doi: 10.3390/app152111845.

[35] H. F. Garcia, A. Aguilar, E. Manilow, D. Vedenko, and B. Pardo, "Deep Learning Tools for Audacity: Helping Researchers Expand the Artist's Toolkit," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2110.13323

[36] H. Fayyad-Kazan, A. Hejase, I. Moukadem, and S. Kassem-Moussa, "Verifying the Audio Evidence to Assist Forensic Investigation," *Computer and Information Science*, vol. 14, no. 3, p. 25, Jul. 2021, doi: 10.5539/cis.v14n3p25.

[37] L. Y. Gong and X. J. Li, "Deepfake Voice Detection: An Approach Using End-to-End Transformer with Acoustic Feature Fusion by Cross-Attention," *Electronics (Switzerland)*, vol. 14, no. 10, May 2025, doi: 10.3390/electronics14102040.

[38] Y. Wang *et al.*, "Tacotron: Towards end-To-end speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2017, pp. 4006–4010. doi: 10.21437/Interspeech.2017-1452.

[39] Y. Ren *et al.*, "FastSpeech: Fast, Robust and Controllable Text to Speech," Nov. 2019, [Online]. Available: http://arxiv.org/abs/1905.09263

[40] C. M. Rosca, A. Stancu, and E. M. Iovanovici, "The New Paradigm of Deepfake Detection at the Text Level," *Applied Sciences (Switzerland)*, vol. 15, no. 5, Mar. 2025, doi: 10.3390/app15052560.

[41] S. Abramova, "Detecting Copy-Move Forgeries in Scanned Text Documents."
[42] J. van Beusekom, F. Shafait, and T. M. Breuel, "Text-line examination for document forgery detection," *International Journal on Document Analysis and Recognition*, vol. 16, no. 2, pp. 189–207, Jun. 2013, doi: 10.1007/s10032-011-0181-5.
[43] DAMOYE T., "DETECTION OF AI-GENERATED TEXT THROUGH CLASSICAL MACHINE LEARNING AND DEEP LEARNING APPROACHES," *International Journal of Nature and Science Advance Research*, Nov. 2025, doi: 10.70382/mejnsar.v10i9.070.
[44] J. Blake, A. S. M. Miah, K. Kredens, and J. Shin, "Detection of AI-Generated Texts: A Bi-LSTM and Attention-Based Approach," *IEEE Access*, vol. 13, pp. 71563–71576, 2025, doi: 10.1109/ACCESS.2025.3562750.
[45] Z. Lai, X. Zhang, and S. Chen, "Adaptive Ensembles of Fine-Tuned Transformers for LLM-Generated Text Detection," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2403.13335
[46] H. U. Khan, A. Naz, F. K. Alarfaj, and N. Almusallam, "Identifying artificial intelligence-generated content using the DistilBERT transformer and NLP techniques," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-08208-7.
[47] A. Yadagiri, L. Shree, S. Parween, A. Raj, S. Maurya, and P. Pakray, "Detecting AI-Generated Text with Pre-Trained Models using Linguistic Features." [Online]. Available: https://gptzero.me/
[48] A. Kayabas, A. E. Topcu, Y. I. Alzoubi, and M. Yıldız, "A Deep Learning Approach to Classify AI-Generated and Human-Written Texts," *Applied Sciences (Switzerland)*, vol. 15, no. 10, May 2025, doi: 10.3390/app15105541.
[49] L. Yuan, H. P. Yu, J. Ren, and P. Sun, "Research of text paraphrase generation based on self-contrastive learning," *PLoS One*, vol. 20, no. 9 September, Sep. 2025, doi: 10.1371/journal.pone.0327613.
[50] M. Elgaar and H. Amiri, "Linguistically-Controlled Paraphrase Generation." [Online]. Available: https://github.com/CLU-UML/
[51] X. Hu, P.-Y. Chen, and T.-Y. Ho, "RADAR: Robust AI-Text Detection via Adversarial Learning," Oct. 2023, [Online]. Available: http://arxiv.org/abs/2307.03838
[52] Y. Singh, "MMTFD: A multimodal detector for temporal forgeries detection," 2024.
[53] C. Yu et al., "'Explicit Correlation Learning for Generalizable Cross-Modal Deepfake Detection,'" pp. 1–6, 2024.
[54] M. Anas Raza and K. Mahmood Malik, "Multimodaltrace: Deepfake Detection using Audiovisual Representation Learning," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 993–1000. doi: 10.1109/CVPRW59228.2023.00106.
[55] J. Xu, J. Chen, X. Song, F. Han, H. Shan, and Y.-G. Jiang, "Identity-Driven Multimedia Forgery Detection via Reference Assistance," in *Proceedings of the 32nd ACM International Conference on Multimedia*, in MM '24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 3887–3896. doi: 10.1145/3664647.3680622.
[56] H. Khalid, M. Kim, S. Tariq, and S. S. Woo, "Evaluation of an Audio-Video Multimodal Deepfake Dataset using Unimodal and Multimodal Detectors," in *Proceedings of the 1st Workshop on Synthetic Multimedia - Audiovisual Deepfake Generation and Detection*, in ADGD '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 7–15. doi: 10.1145/3476099.3484315.

[57] E. Beray Boztepe, B. Karakaya, B. Karasulu, and İ. Ünlü, "An Approach for Audio-Visual Content Understanding of Video using Multimodal Deep Learning Methodology," *SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES*, vol. 5, no. 2, 2022, doi: 10.35377/saucis.05.02.1139765.

[58] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, and X. Zhao, "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects," *Expert Syst. Appl.*, vol. 237, p. 121692, 2024, doi: https://doi.org/10.1016/j.eswa.2023.121692.

[59] A. I. Middya, B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowl. Based. Syst.*, vol. 244, p. 108580, 2022, doi: https://doi.org/10.1016/j.knosys.2022.108580.

[60] A.-U. Rehman, H. S. Ullah, H. Farooq, M. S. Khan, T. Mahmood, and H. O. A. Khan, "Multi-Modal Anomaly Detection by Using Audio and Visual Cues," *IEEE Access*, vol. 9, pp. 30587–30603, 2021, doi: 10.1109/ACCESS.2021.3059519.

[61] J. K. Lewis *et al.*, "Deepfake Video Detection Based on Spatial, Spectral, and Temporal Inconsistencies Using Multimodal Deep Learning," in *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2020, pp. 1–9. doi: 10.1109/AIPR50011.2020.9425167.

[62] D. Salvi *et al.*, "A Robust Approach to Multimodal Deepfake Detection," *J. Imaging*, vol. 9, no. 6, 2023, doi: 10.3390/jimaging9060122.

[63] M. Masood, A. Javed, and A. Irtaza, "Attention-based Multimodal learning framework for Generalized Audio- Visual Deepfake Detection," Oct. 11, 2023. doi: 10.21203/rs.3.rs-3415144/v1.

[64] A. Hashmi, S. A. Shahzad, C. W. Lin, Y. Tsao, and H. M. Wang, "AVTENet: A Human-Cognition-Inspired Audio-Visual Transformer-Based Ensemble Network for Video Deepfake Detection," *IEEE Trans. Cogn. Dev. Syst.*, 2025, doi: 10.1109/TCDS.2025.3554477.

[65] H. Akbari *et al.*, "VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text," Dec. 2021, [Online]. Available: http://arxiv.org/abs/2104.11178

[66] A. Ghadiya, P. Kar, V. Chudasama, and P. Wasnik, "Cross-Modal Fusion and Attention Mechanism for Weakly Supervised Video Anomaly Detection," Dec. 2024, [Online]. Available: http://arxiv.org/abs/2412.20455

[67] J. Yoon, A. Panizo-LLedot, D. Camacho, and C. Choi, "Triple-modality interaction for deepfake detection on zero-shot identity," *Information Fusion*, vol. 109, p. 102424, 2024, doi: https://doi.org/10.1016/j.inffus.2024.102424.

[68] V. K. Sharma, S. N. Singh, and Q. Caudron, "Combating Deepfakes Using an Integrated Framework forAudio and Video Deepfake Detection," Oct. 08, 2024. doi: 10.21203/rs.3.rs-4861782/v1.

[69] Z. Fu *et al.*, "Multi-level feature disentanglement network for cross-dataset face forgery detection," *Image Vis. Comput.*, vol. 135, p. 104686, 2023, doi: https://doi.org/10.1016/j.imavis.2023.104686.

[70] S. Concas *et al.*, "Analysis of Score-Level Fusion Rules for Deepfake Detection," *Applied Sciences*, vol. 12, no. 15, 2022, doi: 10.3390/app12157365.

[71] H. Liu, Y. Wei, F. Liu, W. Wang, L. Nie, and T.-S. Chua, "Dynamic Multimodal Fusion via Meta-Learning Towards Micro-Video Recommendation," *ACM Trans. Inf. Syst.*, vol. 42, no. 2, Nov. 2023, doi: 10.1145/3617827.

[72] S. Liu *et al.*, "M2HF: Multi-level Multi-modal Hybrid Fusion for Text-Video Retrieval," Aug. 2022, [Online]. Available: http://arxiv.org/abs/2208.07664

[73] R. S. Kiziltepe, J. Q. Gan, and J. J. Escobar, "Integration of Feature and Decision Fusion With Deep Learning Architectures for Video Classification," *IEEE Access*, vol. 12, pp. 19432–19446, 2024, doi: 10.1109/ACCESS.2024.3360929.

[74] M. B. Shaikh, D. Chai, S. M. S. Islam, and N. Akhtar, "Multimodal fusion for audio-image and video action recognition," *Neural Comput. Appl.*, vol. 36, no. 10, pp. 5499–5513, Apr. 2024, doi: 10.1007/s00521-023-09186-5.

[75] D. Gkoumas, Q. Li, S. Dehdashti, M. Melucci, Y. Yu, and D. Song, "Quantum Cognitively Motivated Decision Fusion for Video Sentiment Analysis," 2021. [Online]. Available: www.aaai.org

[76] S. A. Shahzad, A. Hashmi, Y.-T. Peng, Y. Tsao, and H.-M. Wang, "AV-Lip-Sync+: Leveraging AV-HuBERT to Exploit Multimodal Inconsistency for Deepfake Detection of Frontal Face Videos," Nov. 2025, doi: 10.1109/THMS.2025.3618409.

[77] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," Aug. 2019, [Online]. Available: http://arxiv.org/abs/1901.08971

[78] R. da Silva, O. S. Nakao, and J. R. D. de Felício, "A connection between the Ice-type model of Linus Pauling and the three-color problem," Jan. 2021, doi: 10.1088/1361-6404/abc89f.

[79] B. Dolhansky *et al.*, "The DeepFake Detection Challenge (DFDC) Dataset," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2006.07397

[80] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging Frequency Analysis for Deep Fake Image Recognition," Jun. 2020, [Online]. Available: http://arxiv.org/abs/2003.08685

[81] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection," Dec. 2020, [Online]. Available: http://arxiv.org/abs/2001.03024

[82] B. Zi, M. Chang, J. Chen, X. Ma, and Y. G. Jiang, "WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection," in *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, Association for Computing Machinery, Inc, Oct. 2020, pp. 2382–2390. doi: 10.1145/3394171.3413769.

[83] T. Zhou, W. Wang, Z. Liang, and J. Shen, "Face Forensics in the Wild," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.16076

[84] Y. He *et al.*, "ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis," Jul. 2021, [Online]. Available: http://arxiv.org/abs/2103.05630

[85] E. Altuncu, V. N. L. Franqueira, and S. Li, "Deepfake: definitions, performance metrics and standards, datasets, and a meta-review," 2024, *Frontiers Media SA*. doi: 10.3389/fdata.2024.1400024.

[86] S. A. R. Syed Abu Bakar, S. Waseem, Z. Omar, and Bilalashfaqahmed, "Exploring the Advancements and Challenges of Deepfake Face-swap: A Survey," *Multimed. Tools Appl.*, vol. 85, no. 1, p. 14, Jan. 2026, doi: 10.1007/s11042-026-21285-8.

[87] Z. Chen *et al.*, "DiffusionFace: Towards a Comprehensive Dataset for Diffusion-Based Face Forgery Analysis," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2403.18471

[88] X. Xiong *et al.*, "TalkingHeadBench: A Multi-Modal Benchmark & Analysis of Talking-Head DeepFake Detection," Jan. 2026, [Online]. Available: http://arxiv.org/abs/2505.24866

[89] H. Wu, Y. Tseng, and H. Lee, "CodecFake: Enhancing Anti-Spoofing Models Against Deepfake Audios from Codec-Based Speech Synthesis Systems," Jun. 2024, [Online]. Available: http://arxiv.org/abs/2406.07237

[90] Q. Wang, H. Huang, G. Pang, S. Erfani, and C. Leckie, "AUDETER: A Large-scale Dataset for Deepfake Audio Detection in Open Worlds," 2018. doi: XXXXXXX.XXXXXXX.

[91] J. Frank and L. Schönherr, "WaveFake: A Data Set to Facilitate Audio Deepfake Detection," Nov. 2021, [Online]. Available: http://arxiv.org/abs/2111.02813

[92] J. Yamagishi *et al.*, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," Sep. 2021, [Online]. Available: http://arxiv.org/abs/2109.00537

[93] T. Fagni, F. Falchi, M. Gambini, A. Martella, and M. Tesconi, "TweepFake: About detecting deepfake tweets," *PLoS One*, vol. 16, no. 5 May, May 2021, doi: 10.1371/journal.pone.0251415.

[94] Y. Li *et al.*, "MAGE: Machine-generated Text Detection in the Wild," May 2024, [Online]. Available: http://arxiv.org/abs/2305.13242

[95] Z. Guo and S. Yu, "AuthentiGPT: Detecting Machine-Generated Text via Black-Box Language Models Denoising," Nov. 2023, [Online]. Available: http://arxiv.org/abs/2311.07700

[96] V. Liyanage, D. Buscaldi, and A. Nazarenko, "A Benchmark Corpus for the Detection of Automatically Generated Text in Academic Publications," Apr. 2022, [Online]. Available: http://arxiv.org/abs/2202.02013

[97] F.-A. Croitoru, V. Hondru, M. Popescu, R. T. Ionescu, F. S. Khan, and M. Shah, "MAVOS-DD: Multilingual Audio-Video Open-Set Deepfake Detection Benchmark," May 2025, [Online]. Available: http://arxiv.org/abs/2505.11109

[98] Y. Hou, H. Fu, C. Chen, Z. Li, H. Zhang, and J. Zhao, "PolyGlotFake: A Novel Multilingual and Multimodal DeepFake Dataset," May 2024, [Online]. Available: http://arxiv.org/abs/2405.08838

[99] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2108.05080

[100] Z. Cai, S. Ghosh, A. Dhall, T. Gedeon, K. Stefanov, and M. Hayat, "Glitch in the Matrix: A Large Scale Benchmark for Content Driven Audio-Visual Forgery Detection and Localization," Jul. 2023, [Online]. Available: http://arxiv.org/abs/2305.01979

[101] Z. Cai *et al.*, "AV-Deepfake1M: A Large-Scale LLM-Driven Audio-Visual Deepfake Dataset," in *MM 2024 - Proceedings of the 32nd ACM International Conference on Multimedia*, Association for Computing Machinery, Inc, Oct. 2024, pp. 7414–7423. doi: 10.1145/3664647.3680795.

[102] P. Korshunov, H. Chen, P. N. Garner, and S. Marcel, "Vulnerability of Automatic Identity Recognition to Audio-Visual Deepfakes," Nov. 2023, [Online]. Available: http://arxiv.org/abs/2311.17655

[103] Tongyi DeepResearch Team *et al.*, "Tongyi DeepResearch Technical Report," Nov. 2025, [Online]. Available: http://arxiv.org/abs/2510.24701

[104] S. Lei, J. Song, F. Feng, Z. Yan, and A. Wang, "Deepfake Face Detection and Adversarial Attack Defense Method Based on Multi-Feature Decision Fusion," *Applied Sciences*, vol. 15, no. 12, 2025, doi: 10.3390/app15126588.

[105] P. Liu, Q. Tao, and J. T. Zhou, "Evolving from Single-modal to Multi-modal Facial Deepfake Detection: Progress and Challenges," Apr. 2025, [Online]. Available: http://arxiv.org/abs/2406.06965

[106] J. Li, R. Ji, H. Liu, X. Hong, Y. Gao, and Q. Tian, "Universal Perturbation Attack Against Image Retrieval," Sep. 2019, [Online]. Available: http://arxiv.org/abs/1812.00552

[107] W. Tang, B. Li, S. Tan, M. Barni, and J. Huang, "CNN-Based Adversarial Embedding for Image Steganography," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 8, pp. 2074–2087, 2019, doi: 10.1109/TIFS.2019.2891237.

[108] B. Kaddar, S. A. Fezza, Z. Akhtar, W. Hamidouche, A. Hadid, and J. Serra-Sagristá, "Deepfake Detection Using Spatiotemporal Transformer," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 20, no. 11, Sep. 2024, doi: 10.1145/3643030.

[109] M. A. Amin, Y. Hu, and J. Hu, "Analyzing temporal coherence for deepfake video detection," *Electronic Research Archive*, vol. 32, no. 4, pp. 2621–2641, 2024, doi: 10.3934/ERA.2024119.

[110] C. Zhu, X. Li, J. Li, and S. Dai, "Improving Robustness of Facial Landmark Detection by Defending against Adversarial Attacks," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11731–11740. doi: 10.1109/ICCV48922.2021.01154.

[111] J. Guan, Y. Zhao, X. Zhuoer, C. Meng, K. Xu, and Y. Zhao, "Adversarial Robust Safeguard for Evading Deep Facial Manipulation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 118–126, Jan. 2024, doi: 10.1609/aaai.v38i1.27762.

[112] A. Holzinger, R. Goebel, R. Fong, T. Moon, · Klaus-Robert Müller, and W. Samek, "xxAI-Beyond Explainable AI," 2020. [Online]. Available: https://link.springer.com/bookseries/1244

[113] H. and D. Y. and F. W. and Z. D. and Z. J. Xu Feiyu and Uszkoreit, "Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges," in *Natural Language Processing and Chinese Computing*, M.-Y. and Z. D. and L. S. and Z. H. Tang Jie and Kan, Ed., Cham: Springer International Publishing, 2019, pp. 563–574.

[114] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," Feb. 2021, [Online]. Available: http://arxiv.org/abs/2103.00020

[115] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," Feb. 2022, [Online]. Available: http://arxiv.org/abs/2201.12086

[116] Y. Chen, N. Akhtar, N. A. H. Haldar, and A. Mian, "Deepfake Detection with Spatio-Temporal Consistency and Attention," in *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2022, pp. 1–8. doi: 10.1109/DICTA56598.2022.10034609.

[117] I. Ganiyusufoglu, L. M. Ngô, N. Savov, S. Karaoglu, and T. Gevers, "Spatio-temporal Features for Generalized Detection of Deepfake Videos," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11844

[118] S. Mehta and M. Rastegari, "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2110.02178

[119] J. and P. H. and L. M. and X. B. and F. J. and Y. L. Wu Kan and Zhang, "TinyViT: Fast Pretraining Distillation for Small Vision Transformers," in *Computer Vision – ECCV 2022*,

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

G. and C. M. and F. G. M. and H. T. Avidan Shai and Brostow, Ed., Cham: Springer Nature Switzerland, 2022, pp. 68–85.

[120] T. Karras *et al.*, "Alias-Free Generative Adversarial Networks," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2106.12423

[121] A. Diel, T. Lalgi, F. S. Mellis, A. Teufel, and A. Bäuerle, "The harm of deepfakes: a scoping review of deepfakes' negative effects on human mind and behavior," *AI Soc.*, Dec. 2025, doi: 10.1007/s00146-025-02774-0.

[122] Umer, Q., Li, J. W., Ashraf, M. R., Bashir, R. N., & Ghous, H. (2023). Ensemble deep learning-based prediction of fraudulent cryptocurrency transactions. IEEE Access, 11, 95213-95224.

[123] Malik, M., Ghous, H., Mubeen, M., Munir, A. M., & Ahmad, N. (2024). Intelligent intrusion detection system for internet of things using machine learning techniques. International Journal of Information Systems and Computer Technologies, 3(1), 23-39.

[124] Ghous, H., Malik, M. H., Qadri, S., & Ahmad, N. (2023). Detection of fake videos using convolutional generative method. Journal of Computing & Biomedical Informatics, 4(02), 8-17.

[125] Rehman, I., & Ghous, H. (2021). Structured critical review on market basket analysis using deep learning & association rules. International Journal of Scientific & Engineering Research, 12(4), 168-183.