Image and Vision Computing 163 (2025) 105738

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Explainable deepfake detection across different modalities: An overview of methods and challenges[☆]

MD Sarfaraz Momin [a,d], Abu Sufian [b,e,*], Debaditya Barman [a], Marco Leo [b], Cosimo Distante [b], Naser Damer [c,f]

[a] Visva-Bharati, Santiniketan, 731235, India
[b] National Research Council of Italy, Institute of Applied Sciences and Intelligent Systems, Lecce, 73100, Italy
[c] Fraunhofer Institute for Computer Graphics Research IGD, Fraunhoferstr. 5, 64283 Darmstadt, Germany
[d] Kaliachak College, Kaliachak, Malda, 732201, India
[e] University of Gour Banga, English Bazar, 732103, India
[f] Department of Computer Science, TU Darmstadt, Karolinenpl. 5, 64289 Darmstadt, Germany

ARTICLE INFO

Keywords:
Machine learning
Deep learning
Generative AI
Deepfake
Explainable AI

ABSTRACT

The increasing use of deepfake technology enables the creation of realistic and deceptive content, raising concerns about several serious issues, including biometric authentication, misinformation, politics, privacy, and trust. Many Deepfake Detection (DD) models are entering the market to combat the misuse of deepfakes. With these developments, one primary issue occurs in ensuring the explainability of the proposed detection models to understand the rationale of the decision. This paper aims to investigate the state-of-the-art explainable DD models across multiple modalities, including image, video, audio, and text. Unlike existing surveys that focus on detection methodologies with minimal attention to explainability and limited modality coverage, this paper directly focuses on these gaps. It offers a comprehensive analysis of advanced explainability techniques, including Grad-CAM, LIME, SHAP, LRP, Saliency Maps, and Anchors, for detecting deceptive content across the modalities. It identifies the strengths and limitations of existing models and outlines research directions to enhance explainability and interpretability in future works. By exploring these models, we aim to enhance transparency, provide deeper insights into model decisions, and bridge the gap between detection accuracy with explainability in DD models.

1. Introduction

Rapid developments in deep learning (DL) [1–3], specifically in generative artificial intelligence (GenAI) [4–6] have recently transformed numerous sectors, including healthcare [7–9], biometrics [10–12], finance [13–15], and autonomous systems [16,17]. These technological breakthroughs are powered by AI models - mathematical frameworks and algorithms designed to learn patterns from data and make predictions. These models have demonstrated exceptional performance in many domains, often surpassing human-level accuracy and achieving remarkable results [18,19].

However, these advancements come with unintended consequences, such as the proliferation of deepfake technology [20–22]. Deepfakes are synthetic or manipulated media, including images, videos, audio, and text - generated using AI algorithms, primarily DL methods like generative adversarial networks (GANs) [23,24] and diffusion models [25,26]. These technologies enable the creation of realistic fabricated content that is often indistinguishable from real ones [27–31]. As the number of such intentions grows, detecting and mitigating deepfakes has become a critical priority. Consequently, numerous Deepfake Detection (DD) AI models have emerged in the market [32–34]. But one of the major challenges in these models is the “black box” nature of these [35,36]. This lack of transparency in the model’s inference makes it difficult to understand the decision-making processes. This opacity raises significant concerns, particularly in high-stakes applications like law enforcement, media verification, and cybersecurity, where trust, accountability, and regulatory compliance are very important [37,38]. To address these issues, researchers have turned to explainable DD, an emerging field focused on not only identifying deepfakes but also providing interpretable insights into the underlying decision-making processes [39–42]. Consequently, numerous explainable detection models have emerged [32,

[☆] This article is part of a Special issue entitled: ‘XAISynData’ published in Image and Vision Computing.

* Corresponding author at: National Research Council of Italy, Institute of Applied Sciences and Intelligent Systems, Lecce, 73100, Italy.
E-mail addresses: mdsarfarazn3@gmail.com (M.S. Momin), abusufian@cnr.it, sufian@ieee.org (A. Sufian), debadityabarman@gmail.com (D. Barman), marco.leo@cnr.it (M. Leo), cosmio.distante@cnr.it (C. Distante), naser.damer@igd.fraunhofer.de (N. Damer).

https://doi.org/10.1016/j.imavis.2025.105738
Received 31 March 2025; Received in revised form 24 August 2025; Accepted 10 September 2025
Available online 16 September 2025

Table 1
Comparative analysis of SOTA survey papers on explainable DD.

Papers	Modalities covered	Explainability techniques discussed	DD method used	Multimodal fusion	Future directions highlighted	Novelty
Edwards et al. [44]	Image, video	Grad-CAM	CNN, multi-modal transformers	Audio-visual fusion	Explainability, temporal consistency, dataset bias	Explains multi-modal approach
Wang et al. [32]	Image, video	Noise-based methods, heatmaps	CNN, vision transformer, frequency analysis	No	Robustness, transferability	Proposes a reliability evaluation metric
Sharma et al. [45]	Image, video, audio	None	CNN, RNN, GAN, vision transformers	No	Dataset diversity, generalization, performance metrics	Systematic review of datasets and detection methods
Pei et al. [46]	Image, video, audio	None	CNN, vision transformers, diffusion models	No	Improving multimodal fusion, Robust detection frameworks	Comprehensive benchmark of deepfake generation and detection methods
Heidari et al. [47]	Image, video, audio	None	CNN, RNN, GAN, transfer learning	No	Addressing dataset bias, enhancing detection methods	Systematic literature review on DD across multiple modalities
Abdullah et al. [48]	Image	None	CNN, vision transformers, foundation models	No	Improving generalization, adversarial training	Evaluation of DD under adversarial attacks using foundation models
Our proposed Paper	Image, video, audio, text	Grad-CAM, LIME, LRP, FLRP, SHAP, anchors, saliency maps	Various AI, especially DL techniques with explainability	Audio-Visual-Text fusion	Comprehensive with explainability focus	First review paper to cover explainable DD across image, video, audio, and text modalities

43], necessitating a comprehensive survey and analysis, which this paper aims to provide.

Despite numerous surveys and review studies on DD models, most have overlooked the explainability of these models and placed less emphasis on multiple modalities. Table 1 presents a comparative overview of recent surveys and highlights how our survey uniquely addresses explainability (or interpretability, depending on context and referenced papers) in recent DD models across modalities.

1.1. Contributions of this paper

- We present insights into popular explainable techniques, including Grad-CAM, LIME, LRP, FLRP, SHAP, ANCHORS, and Saliency Maps, in the context of DD.
- We provide a taxonomy and a comprehensive survey of state-of-the-art (SOTA) explainable DD models across four modalities: image, video, audio, and text.
- We also present a comparison of existing survey papers with ours to show the uniqueness of this work.
- Our paper includes information about the datasets designed for DD methods.
- We also provide a summary of DD types along with their associated papers in tables for each of the four modalities discussed.

1.2. Organization of the paper

The rest of the paper is structured as follows: Section 2 introduces the methodology with technical components. Section 3 presents a comprehensive literature review that examines SOTA explainable DD models across four modalities: image, video, audio, and text. In Section 4, we discuss future research directions and opportunities in this domain. Finally, Section 5 concludes the paper with a summary of the key findings and implications.

2. Methodology with technical components

We selected the Systematic Literature Review (SLR) methodology [49,50] to ensure comprehensive inclusion of relevant studies. In contrast, bibliometric analysis is less suited for an in-depth exploration of explainability techniques across multiple modalities. The flowchart in Fig. 1 outlines our research methodology for explainable DD across four modalities: image, video, audio, and text. The initial phase involved a comprehensive search across several academic databases, including IEEE Xplore, Scopus, Google Scholar, and PubMed, using a variety of keywords related to explainable DD. This process yielded a total of 147 papers, out of which 19 duplicates were removed. Subsequently, 42 papers were selected based on predefined inclusion criteria, including their relevance to DD methods, explainability, and coverage of one or more of the four modalities targeted in our study. The final phase focused on analyzing and synthesizing the methods and techniques presented in these studies. Following the SLR approach, we ensured clarity and repeatability and established a strong foundation for the analysis.

The following points provide detailed explanations of the key components and concepts to maintain clarity and uniformity across the study. These terms facilitate a straightforward understanding for readers with varying levels of expertise. These include terms related to explainability techniques, DD methodologies, and some other terms associated with the four modalities.

2.1. Deepfake

Recent advancements in AI have enabled the creation of highly realistic multimedia content across various domains, offering significant advantages in real-world applications [51,52]. For example, in the entertainment business, changing faces can save the overhead of finding body-double actors for stunts [53,54]. In the education sector, it supports the development of learning tools [55,56]. However, these technical developments have also brought up several serious challenges, especially regarding security and privacy. One of the most

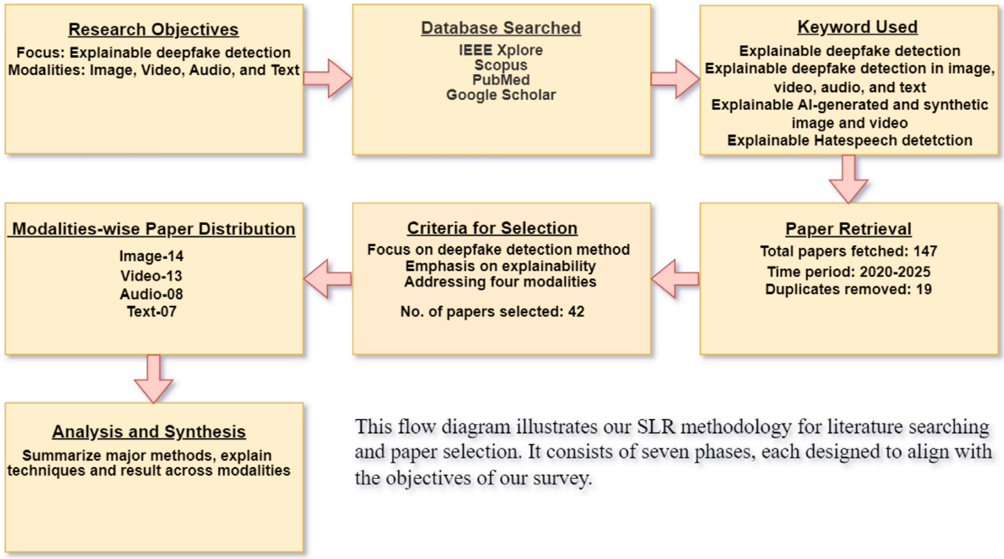

Fig. 1. The overview of our literature searching and paper selection SLR methodology.

pressing challenges is the creation of realistic, forged content, commonly known as deepfake, which seriously threatens digital media's credibility [27,57,58]. The term deepfake is derived from the combination of *deep learning* and *fake*. It refers to synthetic media generated using AI algorithms that mimic real-world data with higher accuracy. Beyond security and privacy, the potential abuse of deepfake technology presents several other serious issues, including a decline in public confidence in the media [59,60], a threat to democratic processes from manipulated political content [59,61], and an increase in cyberbully or harassment due to the production of explicit, non-consensual content [62,63]. Deepfakes can take various forms, including morphed faces, altered emotional states, voice modulation, etc. [64,65]. They may appear in multiple formats, such as images, videos, audio clips, or a combination of these media types, raising critical concerns about their misuse in digital fraud, misinformation campaigns, and invasions of privacy [66,67]. Instead of deepfake categorization based on target functionality, we have categorized deepfakes based on four modalities in the following points.

Deepfake Image: Deepfake image technology leverages advanced AI techniques to manipulate visual content, allowing the creation of highly realistic but artificially altered images. This technology raises critical privacy and ethical concerns as it can be used to fabricate or distort visual data for various malicious purposes. Several sophisticated techniques are utilized for this, including face swapping [68,69], where one person's face is seamlessly replaced with another's, and facial expression alteration [70,71], which can convincingly modify emotions or gestures in images.

Deepfake Video: Modern AI-powered technologies have revolutionized the creation and manipulation of video content, making it possible to generate synthetic videos and alter existing footage with remarkable precision. These manipulations include face swapping [72], face reenactment [73], which involves altering facial movements to mimic another person's expressions or speech patterns, and emotional change [74], where a subject's emotional expressions are modified to convey a different mood or sentiment. These capabilities pose significant ethical, privacy, and security challenges. They can be used for malicious purposes, such as spreading misinformation, creating non-consensual content, or fabricating evidence.

Deepfake Audio: Deepfake audio technology allows the creation of fraudulent audio that closely mimics a specific individual's voice. This raises significant ethical and security concerns. Moreover, it can convincingly replicate vocal characteristics, making it difficult to distinguish fake audio from genuine audio. Audio deepfakes are categorized into three primary types based on their approach: (1) Imitation-based: It replicates the vocal trait of a specific person by analyzing and modeling their voice patterns, such as pitch, tone, and speaking style [75]. It typically requires a target individual's recorded speech dataset to produce convincing results. (2) Synthetic-based: It artificially generates entire speech from text input [76], enabling the creation of highly realistic synthetic audio even without direct recordings of the person. (3) Replay-based: It uses pre-recorded audio segments to create manipulated conversations. It can fabricate new statements or contexts by altering existing audio clips while retaining the authenticity of the original voice [77]. Audio deepfake technology has diverse applications. It enhances virtual assistants, facilitates dubbing in entertainment, and creates realistic simulations for education and training. However, its misuse poses severe risks such as identity theft, fraud, misinformation, and privacy violations.

Deepfake Text: The term 'deepfake text' refers to artificially generated or modified textual content intended to mislead or disseminate false information [78,79]. Unlike visual or audio deepfakes, deepfake text leverages advanced Natural Language Processing (NLP) models to generate highly realistic yet fabricated written content. These contents are often indistinguishable from genuine text. This form of manipulation is becoming increasingly prevalent, and it raises serious ethical and societal concerns. Examples of deepfake text include misleading and false articles designed to influence public opinion, manipulate political narratives, or damage reputations [80].

2.2. Explainable AI (XAI)

Despite the advancements in DL, its intrinsic 'black box' nature raises concerns about the transparency and trustworthiness of its decision-making processes [36,81]. To address these concerns, XAI has emerged, offering tools and frameworks to provide transparency and interpretability to complex ML models, including DL networks [82]. XAI helps users to understand and trust the outputs of ML models by explaining how decisions are made. This explainability is critical in various high-stakes domains such as healthcare, law enforcement, and legal systems. In these fields, the consequences of decisions can significantly impact human lives and society [83]. In the following sections, we discuss some of the most popular XAI techniques, highlighting their

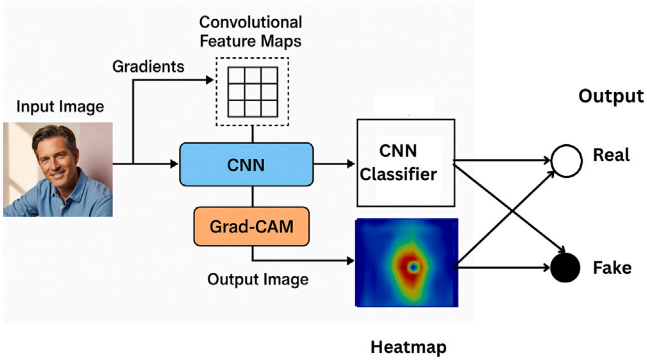

Fig. 2. Heatmap generation using Grad-CAM.

roles in improving the interpretability and usability of advanced AI systems in DD.

Grad-CAM: Grad-CAM (Gradient-weighted Class Activation Mapping) is a technique used to visualize the most critical regions of an input that influence the decision-making process, especially in Convolutional Neural Networks (CNNs). It achieves this by leveraging the gradients of the target class to identify areas of interest [84]. Grad-CAM is widely applied in various domains, including object detection [85], medical imaging [86], autonomous driving [87], adversarial detection [88], etc. In DD, Grad-CAM plays an essential role across multiple modalities [89]. It helps localize changes in photos and videos by creating heatmaps that emphasize modified facial areas (such as lips or eyes). In audio DD, it can be applied to spectrograms (e.g., Mel spectrograms) to highlight suspicious frequency distortions that aid in classification [90].

To interpret the model's DD process, Grad-CAM usually generates heatmaps that highlight manipulated regions in the input images, as shown in Fig. 2. This figure illustrates an explainable DD system that combines a CNN with Grad-CAM. CNN processes the input image to identify features and determine if it is authentic or not. Grad-CAM improves interpretability by producing a heatmap that highlights the significant areas that affected the model's decision.

LIME: LIME (Local Interpretable Model-agnostic Explanations) explains predictions by locally approximating the model's behavior around a specific instance through perturbations [91]. This technique has applications in multiple domains, including healthcare [92], model debugging [93], customer insights [94], etc. In DD, LIME aids in decision interpretation across various modalities [95]. In audio, LIME alters portions of a spectrogram or waveform to identify which frequency elements are most critical in recognizing synthetic speech [96]. In text, it modifies or substitutes essential words and assesses their effect on classification, aiding in revealing linguistic patterns typically present in AI-generated content. Through frame perturbation and classification impact measurement, it detects important altered regions in images and videos, as shown in Fig. 3. In this Figure, LIME generated multiple perturbed images from the input image and then generated altered regions of the image with multiple colors to show if it is a deepfake or not, with the possibility mentioned. Red and yellow color in the processed image indicates a strong influence, while blue and green color indicates a weak influence on deepfake detection and thus provide explainability.

SHAP: SHAP (Shapley Additive Explanations) is a unified framework for interpreting ML models' outputs by assigning important values (Shapley values) to each feature, such as pixels, frames, or phonemes [97]. SHAP is widely used in feature engineering [98], finance [99], customer analytics [100], etc. SHAP provides a multimodal explanation of classification decisions in DD [101]. In audio, it assigns significance ratings to different frequency elements. This reveals irregularities, such as robotic sounds in deepfake voices. In text, SHAP identifies the words or sentence structures that play a crucial role in distinguishing AI-generated from human-written content. In images and videos, SHAP measures the contribution of various pixels or frames that help to detect the artificial modifications. SHAP provides visualization by generating feature importance maps that resemble heatmaps, as shown in Fig. 4. In the processed image, the red and yellow color region indicates a high contribution, while the green and blue regions indicate a low contribution of the features in terms of both explainability and prediction.

LRP and FLRP: LRP (Layer-Wise Relevance Propagation) explains DL models' decisions by backpropagating relevance scores to input features. It highlights the elements that contributed most to a given classification [102]. It is used in various fields, such as NLP [103], medical imaging [104], financial risk analysis [105], etc. LRP enhances explainability across multiple modalities in DD [106]. It helps to detect face swaps and artificial textures in images and video deepfakes by highlighting pixel regions that affect the model's choice. In audio, LRP assigns relevance values to different frequency bands, revealing artifacts produced by generative models [107]. In the text, it highlights key phrases that differ significantly from natural language patterns, providing insights into AI-generated material. Focused LRP (FLRP) [108], also known as Selective LRP, further refines explanations by concentrating on specific layers.

Anchors: Anchors is an interpretable ML method that generates human-understandable decision rules for model predictions. It identifies key conditions that strongly influence classification [109]. It has been applied in text classification [110], fraud detection [111], recommendation systems [112], etc. Anchors create interpretable "if-then" rules for DD in a variety of modalities [113]. For images and videos, it may suggest that the presence of specific facial asymmetries indicates the content is likely inauthentic. In audio, an anchor rule might suggest that if certain unnatural pitch alterations are found, the sample is AI-generated. In text, Anchors can formulate classification guidelines by identifying unusual phrase structures typically present in deepfake content.

Saliency map: A saliency map is a visual representation that highlights key input areas that influence a model's output [114]. It is used in various fields, including image classification [115], medical imaging [116], attention mechanisms [117], etc. Saliency maps enhance interpretability across several modalities in DD [118,119]. For images and videos, it emphasizes altered facial areas, revealing artifacts caused by generative networks. In audio, they can be utilized on spectrograms to uncover significant frequency distortions, highlighting artificial speech patterns. In text-based DD, saliency maps can assist transformer-based models in highlighting essential words or phrases that influence classification.

XAI techniques operate based on different principles. For example, Grad-CAM and saliency maps utilize gradients to estimate pixel importance, while LIME adopts a perturbation method and analyzes how the model's outputs change when parts of the input are modified. SHAP works by computing Shapley values and measures each input feature's contribution to the model's decision, whereas LRP and FLRP highlight the most influential regions.

3. State-of-the-art explainable DD models

This section presents the latest advancements in explainable DD models. It also highlights the current methods employed to enhance the interpretability and transparency of DD models. This review analyzes various techniques for explaining model behavior and categorizes them based on their approach. It also evaluates their strengths and limitations.

This study explores four primary categories of explainable DD models: image, video, audio, and text. It examines research works published since 2020 and selected from several reputable databases. This scoping review highlights the SOTA methodologies and identifies their capabilities and areas where more research and development are needed to enhance the interpretability and effectiveness of DD systems.

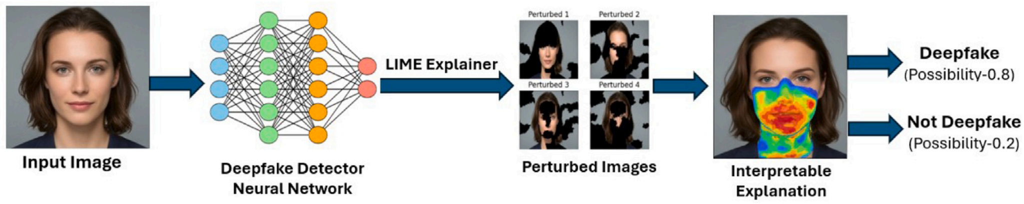

Fig. 3. Heatmap generation using LIME. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

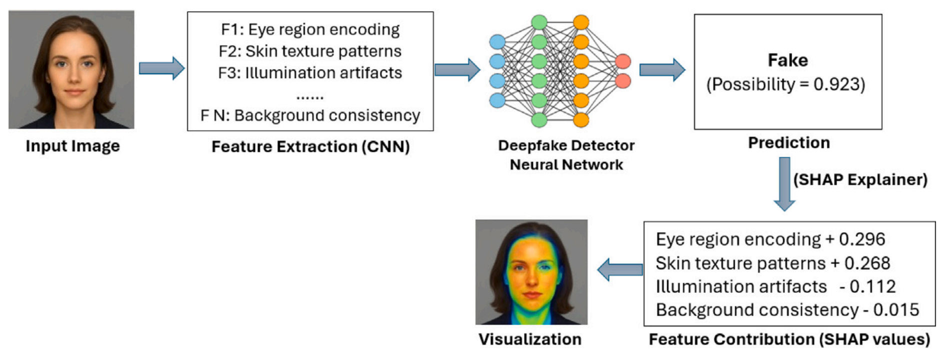

Fig. 4. Visualization using SHAP. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.1. Explainable deepfake image forgery detection

The increasing prevalence of deepfake image forgeries has prompted significant research efforts toward developing effective detection models in recent years. Several explainable deepfake image forgery detection models have emerged, utilizing innovative techniques to balance performance and interpretability. Explainability in this context refers to the ability of a model to provide transparent reasoning for its decisions, enabling users to understand how and why a particular image is classified as real or fake. This attribute is very important for fostering trust and ensuring ethical deployment of detection systems in sensitive applications such as media authentication, digital forensics, and online content moderation.

3.1.1. Grad-CAM integrated approaches for image modality

A learning-based approach [120] utilizing a CNN architecture was proposed for detecting copy-move and inpainting-based forgeries. The method incorporated the Grad-CAM technique to enhance model interpretability by generating heatmaps that highlight forged regions within an image. The model was trained on a custom dataset derived from the COCO dataset [121]. Another approach [122] utilized a DL-based method to detect manipulated images, leveraging the Mesoinception Net (Meso-Net) [123] architecture integrated with Grad-CAM to enhance interpretability by visualizing distinctions between authentic and forged images. Furthermore, an AI bias assessment was conducted to identify and mitigate potential biases present in the dataset. Similarly, another study [124] proposed using a CNN-based approach for distinguishing AI-generated images from real ones, incorporating the Grad-CAM technique for explainability that highlights the most influential regions of an image. Additionally, the study contributed by introducing the CIFAKE dataset [124] that comprises 120 K images and is publicly available. A comprehensive investigation [125] was conducted using CNN and Vision Transformers (ViTs) models [126] to differentiate authentic and AI-generated fake images. The investigation was carried out using the CIFAKE dataset [124] to evaluate model performance. To enhance interpretability, the study incorporated the Grad-CAM technique. The results demonstrated that the CNN-based approach outperformed other models and achieved high accuracy when integrated with Grad-CAM for explainability. Similarly, another approach [127] integrated Grad-CAM for interpretable deepfake face detection, utilizing ResNet [128] models, including ResNet50V2, ResNet101V2, and ResNet152V2. The models were evaluated using the FFHQ dataset [129], and their effectiveness in detecting deepfake facial manipulations was demonstrated. Another study [130] on AI-generated image detection utilized the SOTA DL model named DenseNet121 for distinguishing real and fake images and incorporated Grad-CAM for the explainability. The research used the ‘Real and Fake Faces’ dataset from Kaggle, which comprises 140k facial images [131]. A novel approach, named ExplaNET [132], employed Grad-CAM to enhance model interpretability. The method integrated Multi-task Cascaded CNN (MTCNN) [133] in the pre-processing phase for face detection and alignment, while DenseNet-121 with additional layers served as the model’s backbone. Extensive experiments were conducted, including cross-corpora evaluations on benchmark datasets such as Face-Forensics++ (FF++) [134], Celeb-DF [135], and DFDC-P [136]. A Grad-CAM-integrated approach, MEXFIC (Meta Ensemble eXplainable Fake Image Classifier) [137], was proposed for detecting real and synthetic images. The method leveraged transfer learning with pre-trained SOTA classifiers to enhance detection performance. Additionally, the study introduced two new datasets, FakeGPT and PFake, designed explicitly for model testing. Moreover, the CIFAKE dataset [124] was utilized for further evaluation, ensuring a comprehensive assessment of the model’s effectiveness.

3.1.2. Miscellaneous technique integrated approaches for image modality

Apart from Grad-CAM, some researchers utilized various miscellaneous methods for their model interpretability, which are discussed here. Feature focus [138], a transparent DL-based face morphing detector, was constructed on a modified VGG-A [139] architecture and emphasized the importance of explainability by integrating FLRP. They collected 2 K face images from different public and in-house datasets. The method highlighted manipulated regions of the faces using FLRP and performed exceptionally well across several key metrics like APCER (Attack Presentation Classification Error Rate), BPCER (Bona-fide Presentation Classification Error Rate), and EER (Equal-Error-Rate). A human prompt-based DD framework, E-AANet [140], also emphasized interpretability by utilizing an artefact localization module that predicts possible forged regions of the image. They also used ‘Feature

Abstraction Blocks' (FABs) to enhance feature extraction and used an 'artifacts enhancement module' to enhance the attention on image artifacts. Attentive PixelHop (A-PixelHop) [141] is an efficient fake image detector, designed to offer robust explainability by focusing on high-frequency components in local regions of the images. It achieved this by combining edge/texture block selection, multiple filter banks, and an ensemble of binary classifiers. It ensures the detection process is mathematically transparent and interpretable. Another study [142] introduced an explainable DD technique using advanced CNNs such as ResNet-50, Inception V3, and VGG-16. They incorporated a 'network dissection algorithm' for the model interpretability that provides relevant labels to the individual hidden units of the CNN model. It also explains the classification of real and fake images. The model was trained on the Celeb-DF dataset [135] and tested on the deepfake fact dictionary generated from the CelebAMask-HQ dataset [143]. Another DD approach [144] that integrated ViTs [126] combined with Support Vector Machines (SVM) [145] leveraged the interpretability of classical ML to provide insights into the decision process. The method was tested on a custom dataset created using diffusion models from open datasets like ImageNet, FFHQ, and Oxford-IIIT Pet [146,147]. ForgeryGPT [148] was proposed to enhance the Image Forgery Detection and Localization (IFDL) task with interpretability. It integrated visual and textual modalities, leveraging ViTs and CNNs for feature extraction. A key contribution to this model is the mask-aware forgery extractor, comprising the Forgery Localization Expert (FL-Expert) and the mask encoder, which together generate interpretable heatmaps to highlight manipulated regions. The model was trained on a custom dataset derived from CASIAv2 [149] and the Fantastic-Reality dataset [150], ensuring robustness against various forgery types such as copy-move, splicing, and inpainting.

A detailed summary of the image-based DD approaches, including their methodologies, XAI techniques, datasets, performance metrics with results, pros, cons, and code availability, is presented in Table 2.

3.2. Explainable deepfake video forgery detection

The emergence of deepfake videos has raised significant concerns due to their potential misuse in spreading misinformation, defamation, and other malicious activities [21,154]. As a result, researchers have been actively developing advanced methods for detecting deepfake video, focusing on enhancing the explainability and interpretability of these detection models. These attributes are crucial for building trust in detection systems, especially in high-stakes applications such as law enforcement, media verification, and cybersecurity.

3.2.1. Grad-CAM integrated approaches for video modality

A notable approach [155] combined CNN, CapsuleNet [156], and Long Short-Term Memory (LSTM) [157] networks was specifically designed to enhance explainability by overcoming the limitations of conventional CNNs, such as information loss and difficulty in recognizing rotating features. The method integrated a CNN for feature extraction, CapsuleNet to preserve spatial relationships, and LSTM to capture temporal dependencies across video frames. It offers a more interpretable detection process by using the Grad-CAM technique. An ensemble-based method [40] was proposed to improve generalizability and explainability by using a hierarchical ensemble of weakly supervised models, including human input in the loop. The model combined standard and attention-based data-augmented networks, with Grad-CAM employed to explain its decisions. The inclusion of human interpretability within the model evaluation process, tested on the DFDC dataset [136], resulted in good performance and demonstrated the effectiveness of integrating human insight. Similarly, DFP-Net [158] stands out as an explainable framework for DD by utilizing interpretable prototypes and Grad-CAM that explain the model's reasoning. The method achieved SOTA performance on several benchmarking datasets like FF++ [134], Celeb-DF [135], and DFDC-P [57] and prioritized clarity in its decision-making process. Additionally, they integrated MTCNN [133] during preprocessing to extract the facial region from the video. A framework named AUFF-Net [159] was explicitly designed to detect face swap (FS) and face reenactment (FR) deepfakes, integrating Inception-Swish-ResNet-v2 with Bi-LSTM to capture spatial and temporal features. For explainability, AUFF-Net provides heatmap visualizations using the Grad-CAM technique that depict the areas of the video that contributed most to the detection decision. This transparent approach, along with rigorous cross-validation on unseen data, led to impressive accuracy for FS and FR on the FF++ dataset [57]. Another DeepExplain model [160] combined CNNs with LSTM, leveraged Grad-CAM, and SHAP values. These explainability mechanisms provide insights into the model's decision-making process, allowing users to visualize which features influenced the predictions and thereby enhancing trust in the model's outputs. The DD challenge (DFDC) dataset [136] demonstrated the model's effectiveness.

3.2.2. Saliency map integrated approaches for video modality

ConvNext-PNet [161] is a prototype-based learning framework designed to enhance interpretability and explainability for deepfake video detection. They introduced ConvNext as the framework's base architecture. In the preprocessing phase, MTCNN was employed for facial feature extraction. Post-hoc analysis techniques, including saliency map visualization, activation maximization, and deconvolution, were used for model transparency. The proposed model's performance was evaluated on FF++ [134], CelebDF [135], DFDC-P [136], and DeepFake-Face (DFF) [162] datasets. Another approach [163] based on learning patch-channel correspondence for face forgery detection incorporated a feature rearrangement layer within a DL and emphasized explainability. The method was rigorously evaluated on FF++ [134], Celeb-DF(v2) [135], and DFDC datasets [57]. Additionally, interpretability was checked on three evaluations: visualization of patch-channel correspondence, channel-wise explanations, and saliency explanations.

3.2.3. Miscellaneous technique integrated approaches for video modality

Apart from these explainable methods, researchers integrated various miscellaneous methods for their model interpretability. A study [164] presented a novel DD framework that integrated XAI and Adversarial Robustness Training (ART) to enhance the model's interpretability and resilience against adversarial attacks. They integrated LIME and SHAP techniques to interpret the model's decisions, which provide visual and textual explanations of the input. A traditional CNN-based approach [165] using the Xception network was trained on a DeepFakeDetection (DFD) dataset [166], a sub-component of the FF++ dataset [134], to detect deepfake videos. To enhance visual interpretability, authors employed XAI techniques, such as LIME and LRP. A notable approach [109] utilized the 'Anchors' XAI method and an EfficientNet architecture, which excelled in visualizing suspected parts of video frames to be forged. The method outperformed LIME in terms of explainability by providing more precise and user-friendly visualizations. They employed DFDC [136] and FF++ [134] for the model training. The Dynamic Prototype Network (DPNet) [167] focused heavily on interpretability by capturing dynamic features such as unnatural movements and temporal artifacts, which are common indicators of deepfake content. Post-hoc analysis via gradient and perturbation-based methods was used for the explainability. The model was evaluated on DeepFakeDetection(DFD) [168], DeeperForensics [169], FF++ [134], and Celeb-DF datasets [57] and showed significant improvements in AUC scores. The supervised contrastive (SupCon) learning approach [39] offered a generalizable and explainable method for deepfake video detection. It performed well on the FF++ dataset [134], and the accuracy improved when fused with the Xception network. It enhanced explainability using the heatmap that provides better visualization. Moreover, it used Uniform Manifold Approximation and Projection (UMAP) [170] to get topological insights into the learned feature.

Table 2
Summary table of SOTA deepfake image detection techniques.

Type of DD	Methodology and XAI technique used	Dataset used	Performance metrics and results	Pros	Cons	Code provided
Morphed face in 2021 [138]	Modified VGG architecture and FLRP	Custom dataset	Accuracy in EER: 0.6%, APCER: 1.2%, BPCER: 0.5%	Higher accuracy and lightweight model	Limited coverage of features and computational burden	No
Deepfake image in 2022 [141]	Parallel subspace learning (PSL), edge/texture block, multiple filter banks, and an ensemble of binary classifiers	Custom dataset	Accuracy: 98.7% (average) with 114K–171K parameters	Low computational complexity and improved accuracy	Manual parameter tuning	No
Copy-move and inpainting forgery in 2022 [120]	In-house CNN and Grad-CAM	Synthetic dataset using COCO [121]	Accuracy in inpainting: 80%, copy-Move: 70%, combined: 70%	Enhanced interpretability and performance	Limited to a few forgery types, under-perform due to basic CNN architecture	No
AI-generated images in 2023 [125]	CNN, ViTs Model, and Grad-CAM	CIFAKE [124]	Accuracy in CNN: 96.31% and ViT: 87.09%	High accuracy and used hybrid approach	Limited resolution handling and dataset dependency	No
Deepfake images in 2023 [122]	MesoNet and Grad-CAM	DFIM-HQ (own dataset) [122]	Accuracy: 99.87%, F1-score: 0.9986, AUC: 0.99973 (validation)	Bias analysis and mitigation	Limited model variety and dataset dependency	No
Deepfake images in 2023 [144]	ViTs named ViT-L-32, Explainable Deep Neural Network (xDNN), and SVM	Custom dataset using Oxford-IIIT Pet [151], ImageNet [146], and FFHQ [129]	Accuracy: 99.7%, 99.5%, 99.4% (ViT+MLP fine-tuned models)	Prototype-based explainability and improved performance	Lack of dataset generalization and high resource demand	Yes[a]
AI-generated synthetic images in 2024 [124]	CNN and Grad-CAM	CIFAKE (own dataset) [124]	Accuracy: 93.55%, F1-score: 0.936	Introduced CIFAKE dataset and provide high accuracy	Over-focused on background artifacts and solely relies on CNN	No
Deepfake face images in 2024 [127]	ResNet50V2, ResNet101V2, ResNet152V2, and Grad-CAM	FFHQ [152]	Precision: 90%, Recall: 93%, F1-score: 91%	Enhanced accuracy and interpretability	Biased dataset and high computational cost	No
AI-generated images in 2024 [130]	DenseNet121 and Grad-CAM	140k Real and Fake Faces [153]	Accuracy: 93.42%	Robust training strategy and high accuracy	Limited dataset diversity with false-negative issues	No
Deepfake images in 2024 [140]	E-AANet with human prompts and artifact localization module	FF++ [134]	Accuracy: 96.34% and AUC: 99.12%,	Human-guided attention for better interpretability and effective localization of manipulations	Dependency on human annotations and performance decrease on lower quality data	No
Deepfake images in 2024 [132]	DenseNet-121, MTCNN, and Grad-CAM	FF++ [134], Celeb-DF [135], DFDC-P [136]	Accuracy: 98.2% on FF++ and 96.05% on Celeb-DF	Prototype-based interpretability and improved decision visualization	Computational overhead and limited performance on complex manipulations	No
Copy-move, splicing, and inpainting images in 2024 [148]	CNN, ViTs, Forgery localization expert, and mask encoder	CASIAv2 [149], Fantastic-Reality [150]	Accuracy: 81.60% (average)	Multimodal detection and explanation with precise localization	Computational overhead and dataset dependency with overfitting	No
Deepfake images in 2024 [142]	ResNet-50, Inception V3, and VGG-16 and Network dissection algorithm	Celeb-DF [135], CelebAMask-HQ [143]	Accuracy: 86.2% on VGG-16 and F1-score up to 0.9	Multimodal evaluation with enhanced interpretability	Lack of dataset diversity and higher computational cost	No
Synthetic images in 2024 [137]	Transfer learning and Grad-CAM	FakeGPT and PFake (own dataset), CIFAKE [124]	Accuracy: 94% on CIFAKE and 96.61% on PFake	High performance with meta-ensemble learning and robust cross-dataset evaluation	Longer inference	Yes[b]

[a] https://github.com/lira-centre/DeepfakeDetection.
[b] https://github.com/tanvirnwu/MEXFIC.

A generalized Graph Neural Network (GNN)-based DD model [171] was developed on a pyramid ResNet structure. It achieved high AUC scores across multiple datasets while improving both performance and interpretability. The model utilized the PReLU activation function to fine-tune parameters and prevent gradient disappearance.

A detailed comparison of DD techniques in the video modality, including their methodologies, XAI approaches, datasets, performance metrics, advantages, disadvantages, and code repository, is provided in Table 3.

3.3. Explainable deepfake audio forgery detection

Explainability and interpretability are critical components in deepfake audio forgery detection, as they help to ensure transparency and build trust in the models' decision-making processes. The rapid

Table 3
Summary table of SOTA deepfake video detection techniques.

Type of DD	Methodology and XAI technique used	Dataset used	Performance metrics and results	Pros	Cons	Code provided
Deepfake videos in 2020 [165]	Xception network, LRP, and LIME	Subcomponent of FF++ [134]	Accuracy: 94.33% and 90.17% on 1.3x and 2.0x image scale	Improved model interpretability with robustness to noise and transformations	Computational overhead with limited focus on accuracy	No
Face forgery in 2021 [167]	DPNet, posthoc analysis via gradient, and perturbation-based methods	DFD [168], Deeper forensics [169], FF++ [134], and Celeb-DF [135]	AUC: 99.2% on FF++ and 90.8% on DFD	Improved visual and temporal explanation	Computational overhead and limited accuracy	No
Forensic deepfake videos in 2022 [109]	EfficientNet architecture and anchors XAI	DFDC [136] and FF++ [134]	Accuracy: 91.92%, AUC: 0.931 and anchor affinity score: 70.23%	Improved explainability and accuracy with benchmarking against CNN models	Computational overhead and limited dataset	No
Forensic deepfake videos in 2022 [40]	Weakly supervised models and Grad-CAM	DFDC [136] and Celeb-DF [135]	Accuracy: 92.2%, AUC: 0.975 on DFDC and 93.6% on CelebDF	High forensic accuracy	Complex model structure and dependence on weak supervision quality	No
Deepfake images and videos in 2022 [39]	Fusion of supervised contrastive learning, Xception network, heatmap visualization, and UMAP	FF++ [134]	Accuracy (TOSC): 78.74% and CSC up to 96.57% on FF++	Improved explainability and generalizability, including robust performance in unseen data	Computational overhead, moderate accuracy, and dependency on high-quality dataset	Yes[a]
Synthetic facial images in 2023 [171]	GNN, feature map, and use of KNN	DFDC [136], FF++ [134], Celeb-DF [135], and World Leader Dataset [172]	Accuracy: 76.10% on Celeb-DF dataset	Robust feature representation and enhanced generalizability	Computational complexity, overfitting and limited performance	No
Deepfake videos in 2023 [158]	Prototype-based learning and Grad-CAM	FF++ [134], Celeb-DF [135], and DFDC-P [136]	Accuracy: 96.05% on Celeb-DF and 92.53% on DFDC-P	High accuracy and improved interpretability	Computational burden and lower performance due to overfitting	No
Face forgery in 2023 [163]	Xception, resNet18, denseNet121, visualization of patch-channel, channel-wise explanations, and saliency explanations	FF++ [134], Celeb-DF(v2) [135], and DFDC [136]	Accuracy: 99.79% on FF++, 99.03% on Celeb-DF, and 99.52% on DFDC	High accuracy and enhanced saliency-based explanation	Computational overhead and limited generalization	Yes[b]
FaceSwap (FS) and Face Reenactment (FR) in 2024 [159]	Inception-Swish-ResNet-v2 with Bi-LSTM and Grad-CAM	FF++ [134]	Accuracy: 99.21%, and 98.32% for FS, or FR	Novelty with high accuracy and robust performance	Complex model, computational burden, and difficulty in handling low-quality samples	No
Deepfake videos in 2024 [161]	ConvNext architecture based on ResNet, saliency map visualization, activation maximization, and de-convolution	FF++ [134], CelebDF [135], DFDC-P [136], and DFF [162]	Accuracy: 98.70% on FF++, 97.09% on CelebDF, and 90.87% on DFDC-P	Enhanced performance with robust feature learning	Limited generalization with computational burden	No
Deepfake videos in 2024 [160]	CNN, LSTM, Grad-CAM, and SHAP	DFDC [136]	Accuracy: 0.95, precision: 0.82, recall: 0.97, F1-score 0.94, and AUROC 0.9 on LSTM	Hybrid model and enhanced accuracy with robust performance	Computational complexity and performance drop in noisy condition	No
Deepfake videos in 2024 [155]	CNN, capsuleNet with LSTM and Grad-CAM	DFDC [136]	Accuracy: 88.00% and recall: 88.30%	Hybrid model with improved accuracy	Limited dataset variations with decreased performance	No
Deepfake video in 2024 [164]	Adversarial robustness training, SHAP, and LIME	DFDD [173]	Accuracy: 97.5%, recall: 96.8%, and F1-score: 97.5%	Hybrid approach with high accuracy	Computational overhead, challenges in balancing robustness and interpretability during training	No

[a] https://github.com/xuyingzhongguo/deepfake_supcon.
[b] https://github.com/Jae35/IFFD.

advancements in DL have enabled the development of highly accurate detection models, but their inherent complexity often makes them difficult to understand. To address this challenge, researchers have focused on creating models that achieve high detection performance and provide insights into how and why specific decisions are made.

3.3.1. SHAP integrated approaches for audio modality

One such approach [174] for deepfake speech detection proposed using raw waveform signals and spectrograms as fused input features. To improve interpretability, the approach employed SHAP to analyze and visualize the feature contributions of different acoustic components. The model named Wav+Spec-Res-TSSDNet combined the strengths of waveform-based and spectrogram-based detection to enhance classification accuracy while reducing computational complexity. Experimental evaluations on the ASVspoof 2019 dataset [175] demonstrated that the fusion of time-domain and frequency-domain features significantly improved the performance of DD. Another study [176] proposed a lightweight multi-feature ML-based framework for spoofed audio detection, utilizing the SHAP XAI method to enhance interpretability. To enhance the accuracy, authors employed a diverse set of hand-crafted audio features, including spectral, temporal, chroma, and frequency-domain. Extensive testing validates its effectiveness across three popular datasets: ASVSpoof2019 [177], FakeAVCelebV2 [178], and In-The-Wild database [179]. Another study [180] for deepfake audio detection employed GANs to generate fake audio samples from the Fake or Real (FoR) dataset. The quality of these synthetic samples was evaluated using the Fréchet Audio Distance (FAD) metric. For the classification task, they utilized VGG16, MobileNet, InceptionV3, and a custom CNN model [181]. Among these, VGG16 achieved the highest accuracy. To enhance interpretability, the authors incorporated XAI techniques, including LIME, SHAP, and Grad-CAM, which provide insights into the model's decision-making process by highlighting key frequency bands influencing the classification. The study's findings are particularly relevant for addressing cybersecurity threats such as phishing calls and voice impersonation attacks.

3.3.2. Miscellaneous technique integrated approaches for audio modality

Apart from SHAP, some other methods have also been used in the latest research that are discussed here. An innovative approach [182] proposed for voice spoofing detection utilized complex-valued Neural Networks (CVNNs) [183]. The model transformed the audio taken as input into complex-valued spectrograms using the constant-Q transform. For explainability, Saliency maps and SmoothGrad [184] were used to visualize the detection process. The approach was tested on the ASVspoof 2019 [175] and In-the-Wild (ITW) [179] datasets. It outperformed SOTA methods on the ITW dataset. The study [185] applied XAI techniques, particularly LIME and spectral band masking, to identify which frequency bands contribute most to synthetic audio classification. Experiments using the ASVspoof 2019 dataset [175] revealed that synthetic speech artifacts are primarily concentrated in the lowest (below 1000 Hz) and the highest (above 8000 Hz) frequency bands. The study further validated these findings by evaluating different models, such as ResNet and RawNet2 [186]. Another model [107] utilized a combination of CNN and LSTM networks to maintain interpretability at the human perception level. This approach incorporated several XAI methods, including LRP, Deep Taylor decomposition, and integrated gradients, to provide insights into the decision-making process. In addition to methods, datasets play a crucial role in explainable deepfake audio detection. The HateXplain benchmark dataset [187] was introduced for explainable hate speech detection. It provides human-level explanations using the concept of rationales. The explainability was measured using plausibility and faithfulness within the ERASER benchmark framework [188]. This dataset contains 20K samples, sourced from platforms like Twitter and Gab and annotated by Amazon Mechanical Turk. It categorizes data into three classes such as: hateful, offensive, and normal, thereby contributing to more interpretable and transparent detection processes. In [189], the authors introduced an attention roll-out mechanism that enhanced the interpretability of transformer-based audio deepfake classifiers. The authors also proposed a generalizability benchmark by training models on the ASVspoof −5 dataset [190] and evaluated them on FakeAVCeleb [178]. Additionally, the study explored various XAI techniques, including occlusion-based feature masking and attention visualization [191], to better understand the decision-making process.

An overview of audio DD methods is given in Table 4, including information on the detection types, applicable XAI methods, datasets used, performance outcomes, related strengths and limitations with code accessibility.

3.4. Explainable deepfake textual forgery detection

In the domain of deepfake textual content detection, explainability and interpretability are crucial factors for ensuring the reliability of the detection system. With the rising prevalence of text-based deepfakes, including fabricated news articles and manipulated conversations, the need for effective detection models is increasing. These models must not only achieve high accuracy but also offer clear insights into their decision-making processes. Explainability and interpretability aspects are particularly critical for effectively combating misinformation and fostering trust in the deployed systems.

3.4.1. LIME integrated approaches for text modality

In [194], the authors developed a model using an advanced DL-based framework for Arabic fake news classification, leveraging CNN-LSTM ensemble learning combined with ELMo. The framework combined the strengths of CNNs in feature extraction and LSTMs in sequence modeling. The study also employed LIME to enhance model transparency by identifying key linguistic patterns influencing predictions. By conducting thorough experiments and utilizing 5-fold cross-validation, the study showed that the suggested method outperformed previous approaches and benchmarked against SOTA transformer architectures, including BERT and RoBERTa. MFIR (Multimodal Fusion and Inconsistency Reasoning) [195] is a novel approach proposed for detecting fake news by leveraging both textual and visual content. The model combined multimodal data via a cross-modal infiltration fusion mechanism, facilitating richer semantic interaction while preserving contextual coherence. It further enhanced interpretability through an explanation reasoning (ER) layer to create user-friendly interpretations. The study also employed LIME to enhance model interpretability. Through extensive experiments and 5-fold cross-validation, the research demonstrated that the proposed approach surpassed existing Arabic fake news detection models. A surrogate-based explainability approach was proposed in [196]. It focused on enhancing the explainability of BERT-based fake news detectors using LIME and Anchors. These XAI techniques were applied to models trained on genuine news articles from Reuters and fake news from Kaggle. The effectiveness of the method was tested on two distinct models: a BERT-based fake news detector and a bidirectional LSTM memory classifier, each trained on different datasets.

3.4.2. Miscellaneous technique integrated approaches for text modality

Apart from LIME, some other techniques are also leveraged for the explainability that are discussed here. Quantitative argumentation-based automated explainable decision-making system (QA-AXDS) [197] is entirely a data-driven approach and is designed to combat fake news on social media. Unlike traditional ML techniques, QA-AXDS is more transparent and interpretable, as it captures human-level knowledge from unstructured text data and expresses it within a quantitative argumentation framework. This approach demonstrated superior performance to existing ML algorithms, as evidenced by experiments on several publicly available datasets such as Twitter-2017, Twitter-2019 [198], and Reddit-2019 [197]. In [199], a novel approach for interpretable fake news detection was presented by leveraging external commonsense knowledge. The method integrates knowledge graphs to provide explainable reasoning. It also gathered commonsense information from news articles and contrasted it with a commonsense knowledge graph by employing a random walks-based graph matching algorithm. By recognizing inconsistencies in the attributes and relationships of entities, the model assessed whether the news was fabricated and offered a rationale for its conclusion. Another approach [200] for fake news classification was developed by integrating topic modeling with deep variational autoencoders (VAE). The model, called LDAVAE, combined a Bayesian topic modeling technique (Latent Dirichlet Allocation - LDA) with a bidirectional LSTM variational

Table 4
Summary table of SOTA deepfake audio detection techniques.

Type of DD	Methodology and XAI technique used	Dataset used	Performance metrics and results	Pros	Cons	Code provided
Hate speech in 2021 [187]	CNN-GRU, BiRNN, BiRNN-Attention, BERT, concept of rationales, and measuring plausibility and faithfulness	HateXplain (Own dataset) [187]	Accuracy: 0.698, macro F1-score: 0.687, and AUROC: 0.851	Benchmark dataset and enhanced bias mitigation	Limited generalization	Yes[a]
Deepfake voice in 2022 [107]	CNN, CNN-LSTM, CNN-LSTM-permuted, deep taylor, integrated gradients, and LRP	ASVspoof 2021 [192] and LJSpeech [193]	Accuracy: 99.97% on both ASV multiple TTS and LjSpeech	Multimodal interpretability and effective feature extraction	Computational burden and difficult to detect sophisticated deepfake voices	No
Voice anti-spoofing in 2023 [182]	Complex-valued CQT spectrograms, complex-valued neural networks, saliency maps, and Smooth Grad	ASVspoof 2019 [175] and ITW [179]	EER: 3% and 10%	High accuracy	Higher computational cost and limited generalizability	No
Deepfake audio in 2023 [180]	VGG16, mobile net, inceptionV3, custom CNN, LIME, SHAP, and Grad-CAM	Custom dataset	Accuracy: 93.37% on VGG-16	Robust detection and effective evaluation of fake audio	Limited performance on noisy samples	No
Synthetic speech detection in 2023 [185]	ResNet, rawNet2 and LIME	ASVspoof 2019 [175]	Accuracy: 78% in S1 frequency for real speech	Enhanced accuracy	Limited generalization and computational burden	No
Spoofed audio in 2024 [176]	Multi-feature ML-based framework and SHAP	ASVSpoof2019 [175], FakeAVCeleb V2 [178], and an ITW [179]	Accuracy: 89% on ASVSpoof2019, 94.5% on FakeAVCelebV2, and 94.67% on ITW	Lightweight ML approach and high generalization	Performance drop in handling noisy data and unseen attacks	No
Deepfake speech in 2024 [174]	Ensemble of spectograms, waveforms, Res-TSSDNet model, and SHAP	ASVspoof 2019 [175]	EER: 3.39% for evaluation	Enhanced accuracy and explainability with lightweight model	Computational overhead and limited generalization	No
Deepfake audio in 2024 [189]	Transformer based audio deepfake classifier, occlusion-based feature masking, and attention visualization	ASVspoof-5 [190] and FakeAVCeleb [178]	Precision: 0.97, recall: 0.63, and F1-score: 0.77 on Wav2Vec model	Robust performance and new benchmark generalization	High computational demand	No

[a] https://github.com/punyajoy/HateXplain.

autoencoder to improve both classification accuracy and interpretability. Experiments conducted on three real-world datasets, including ISOT [201], COVID-19 [202], and Twitter [198], demonstrated that LDAVAE achieved competitive performance compared to SOTA fake news detection models while retaining interpretability. Explainable and Context-Enhanced Network (ECENet) [203] for fact verification was developed by integrating multi-clue feature extraction, multi-level reasoning, and explanation generation within a single framework. ECENet also introduced an improved Coarse and Fine-Grained Attention Network (CFgAN) to understand contextual information better. Additionally, a novel justification generation module was introduced that leveraged deep reinforcement learning to select and rewrite key sentences from evidence documents, all without requiring additional labels. A recent study in [204] employed Bloom filters and recursive feature elimination alongside classical ML methods to detect deepfake news.

Table 5 shows a comparative summary of text-based DD strategies, with attention to methodology, explainability, datasets, evaluation metrics, pros, cons, and code availability.

4. Discussion and future scope

This section highlights the main takeaways from the analysis, assessing the advantages, shortcomings, and gaps in current methods. Furthermore, it explores potential directions for prospective research works to tackle new challenges and improve the efficacy of explainable multimodal DD systems.

4.1. Discussion

The domain of explainable DD, which encompasses image, video, audio, and textual content, highlights the critical need for transparency and interpretability in detection models. Various techniques have been employed to enhance accuracy while providing insights into the model's decision-making. Integrating XAI methods, such as Grad-CAM, LIME, and SHAP, has made these models more trustworthy and interpretable. By visualizing manipulated regions, identifying key features, and leveraging humanly understandable frameworks, the goal remains to develop effective and transparent models. These advancements improve the reliability of DD and contribute to broader efforts in combating misinformation. It also ensures ethical AI deployment. Enhancing explainability helps bridge the gap between complex black-box AI models and human understanding of their decision-making processes. This improvement can contribute to the development of more robust, reliable, and trustworthy detection systems.

As discussed, deepfakes have raised significant privacy and security concerns, prompting researchers to propose various detection techniques. However, the rapid advancement of deepfake generation models continues to challenge the trustworthiness and generalizability of existing methods. Therefore, developing explainable and adaptable detection models is crucial. This paper focuses on explainable DD across multiple modalities, including image, video, audio, and text. Over the years, numerous explainable DD approaches have been introduced, as discussed in Section 3. We systematically reviewed these models across four primary modalities and compiled a tabular summary

Table 5
Summary table of SOTA deepfake text detection techniques.

Type of DD	Methodology and XAI technique used	Dataset used	Performance metrics and results	Pros	Cons	Code provided
Fake news in 2021 [196]	BERT-based fake news detector, LIME, and anchors	Fake-and-real-news-dataset from kaggle [205]	Precision: 98%, recall: 99%, and F1-score: 98% for fake news	Plug and play design and support for short texts	Limited local explanation and incomplete anchors coverage	No
Fake news on social media in2022 [197]	Quantitative argumentation-based automated eXplainable decision system with xgboost classifier and quantitative argument tree	Twitter-2017 [198], Twitter-2019, and Reddit-2019 [206]	Precision: 0.353, recall: 0.600, and F1-scoreL 0.444 using fidelity accuracy method for false news on Reddit-2019	High performance with automated knowledge acquisition	Less intuitive due to large structure and difficulty in managing unverified sample	No
Fake news in 2023 [199]	Random walks graph matching algorithm, lexical analysis of Chinese (LAC), the universal information extraction (UIE) framework, and utilizing commonsense knowledge	Custom dataset	Accuracy: 91.07%, precision: 85.00%, and recall: 89.47%	High performance with commonsense knowledge integration	Dependency on knowledge graph	No
Fact verification in 2023 [203]	Improved-coarse, fine-grained attention network (ImprovedCFgAN), resNet-50 with a feature pyramid network (FPN) and reinforcement learning agent	NewsCLIPpings [207] and FACTIFY [208]	Accuracy: 0.877 and F1-score: 0.875 using ECENet method	Multi-clue feature extraction with enhanced context understanding	Computational burden and limitation in sentence extraction	No
Fake news in 2023 [200]	Bayesian topic modeling, LDA with a BiLSTM-based VAE, and Bayesian admixture model	Information security and object technology (ISOT) [201], COVID-19 [209], and Twitter [210]	Accuracy: 0.99, F1-score: 0.99, FPR: 0.02, and FNR: 0.01 on COVID-19 using RNN	Enhanced performance and interpretability with dual embedding approach	Topic misalignment on noisy distribution	No
Fake news in 2023 [195]	Feature representations, CIF layer, MIL layer, ER layer, task learning, and Explanation reasoning (ER) layer	Twitter [210], Weibo [211], and Politifact [212]	Accuracy: 0.954, precision: 0.913, recall: 0.910, and F1-score: 0.911 for fake news on politifact	Cross-modal infiltration fusion and multi-modal inconsistency learning	Dependency on data quality and limited interpretability on single-modality data	No
Arabic fake news in 2024 [194]	Combined CNN and LSTM, ELMO word embedding technique, and LIME	Arabic fake news dataset (AFND) [213]	Accuracy: 91.57%, precision: 91.56%, recall: 91.63%, F1-score: 91.59%, and inference time 1.2 s	Hybrid ensemble model that outperforms traditional model using contextual word embeddings	Feature overfitting risk and language dependency	No

categorizing each study based on detection type, methodology, XAI techniques, datasets, performance metrics, with results, pros, cons, and code availability.

After quantifying the related literature, we draw the pie-chart presented in Fig. 5. The chart indicates that the Grad-CAM explainability technique has been widely used due to its good visual interpretability. However, LIME and SHAP have also been widely adopted for explainability. For image and video-based DD, again, Grad-CAM and LRP are preferred for their ability to visually interpret spatial and temporal features effectively. In contrast, LIME and SHAP are better suited for audio and textual deepfakes, offering insights into specific features and segments. Additional techniques such as saliency maps, integrated gradients, and spectrogram-based feature visualization also potentially improved explainability. Among the miscellaneous techniques presented in the pie chart, many studies employed multiple XAI methodologies to enhance explainability in DD, which we categorized under ‘miscellaneous’. In this category, some works utilized multiple classical ML methods, which are generally more interpretable. Additionally, several studies explored neuro-symbolic AI, a hybrid approach that combines classical symbolic AI with modern neural network-based models such as DL.

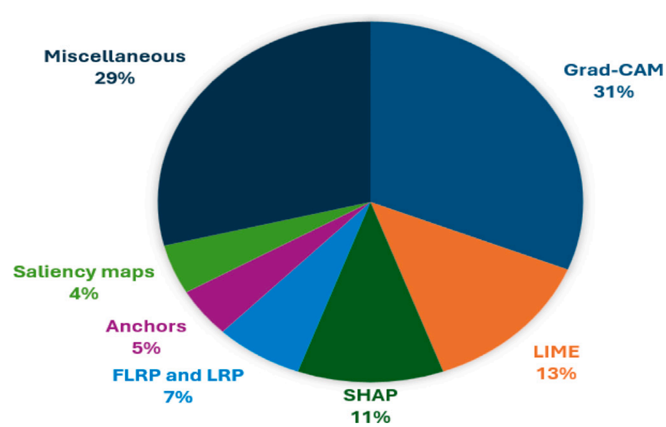

Fig. 5. Percentage distribution of various XAI techniques used in SOTA DD models.

4.2. Future scope

Through our exploration of SOTA models for explainable and interpretable DD, we have observed a clear need and substantial opportunity for further advancements in this area. Below, we outline several key directions for future research:

- **Advancing Multimodal Explainability Techniques:** Develop novel methods that provide integrated explanations across multiple modalities, such as image, video, audio, and text patterns, to enhance the interpretability of DD systems.
- **Real-Time Explainable DD:** Design lightweight models capable of detecting deepfakes in real time with low-resource devices, while maintaining a high degree of explainability.
- **Robustness in Low-Resource and Noisy Environments:** Create resilient detection methods that preserve both accuracy and explainability when dealing with low-resolution or noisy data.
- **User-Friendly Visualization:** Develop intuitive and accessible tools, such as interactive heatmaps to visually spot the manipulation, to help non-technical users understand model predictions.
- **Neuro-symbolic AI:** Neuro-symbolic AI can enhance explainable DD by combining the transparency of symbolic reasoning with automatic feature extraction capabilities of DL. This hybrid approach enables more interpretable decision-making while maintaining high detection performance.
- **Development of Datasets:** Encourage the development and open availability of curated multimodal datasets that capture diverse and emerging types of deepfakes, facilitating more effective training and evaluation of explainable models, and aligning with the needs of multimodal foundation models.
- **Ensuring Fairness Across Demographic Groups:** Investigate and mitigate potential biases in DD methods and datasets to ensure fair and equitable performance of detection methods across different demographic groups.

5. Conclusion

This study provided a comprehensive, multi-dimensional overview of SOTA explainable DD models across image, video, audio, and text modalities. Existing surveys in the literature were primarily focused on detection methods while overlooking the explainability of the models and the different modalities of deepfake content. Our study aims to bridge this gap by evaluating advanced detection models and offering a thorough examination of prominent explainability methods such as Grad-CAM, LIME, LRP, FLRP, SHAP, Anchors, and Saliency Maps, emphasizing their suitability across various deepfake modalities.

Our research also underscores the societal challenges posed by deepfake technology, including its implications for trust, privacy, and security. We identified significant limitations in current detection models, particularly regarding explainability. The key contribution of our paper lies in its structured approach, integrating SOTA detection methods with explainability and interpretability techniques. The findings of this study offer significant insights that can inform and direct future research toward the development of more reliable DD systems.

CRediT authorship contribution statement

MD Sarfaraz Momin: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Abu Sufian:** Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Debaditya Barman:** Writing – review & editing, Validation, Supervision, Resources, Formal analysis. **Marco Leo:** Writing – review & editing, Validation, Supervision, Resources, Methodology, Formal analysis. **Cosimo Distante:** Writing – review & editing, Validation, Supervision, Resources, Funding acquisition. **Naser Damer:** Writing – review & editing, Validation, Supervision, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was partially supported by the project Future Artificial Intelligence Research, FAIR CUP B53C22003630006, grant number PE0000013, financed through NextGenerationEU.

Data availability

No data was used for the research described in the article.

References

[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nat. 521 (7553) (2015) 436–444.
[2] S. Dong, P. Wang, K. Abbas, A survey on deep learning and its applications, Comput. Sci. Rev. 40 (2021) 100379.
[3] S.J. Prince, Understanding Deep Learning, MIT Press, 2023.
[4] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P.S. Yu, L. Sun, A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt, 2023, arXiv preprint arXiv:2303.04226.
[5] S.S. Sengar, A.B. Hasan, S. Kumar, F. Carroll, Generative artificial intelligence: a systematic review and applications, Multimedia Tools Appl. (2024) 1–40.
[6] B. Fu, A. Hadid, N. Damer, Generative AI in the context of assistive technologies: Trends, limitations and future directions, Image Vis. Comput. 154 (2025) 105347.
[7] M.Y. Shaheen, Applications of Artificial Intelligence (AI) in healthcare: A review, Sci. Prepr. (2021).
[8] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare, Nature Med. 25 (1) (2019) 24–29.
[9] M. Bakator, D. Radosav, Deep learning and medical diagnosis: A review of literature, Multimodal Technol. Interact. 2 (3) (2018) 47.
[10] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, D. Zhang, Biometrics recognition using deep learning: A survey, Artif. Intell. Rev. 56 (8) (2023) 8647–8695.
[11] C. Tucci, A. Della Greca, G. Tortora, R. Francese, Explainable biometrics: a systematic literature review, J. Ambient. Intell. Humaniz. Comput. (2024) 1–20.
[12] A. Golda, K. Mekonen, A. Pandey, A. Singh, V. Hassija, V. Chamola, B. Sikdar, Privacy and security concerns in generative AI: a comprehensive survey, IEEE Access (2024).
[13] L. Cao, Ai in finance: challenges, techniques, and opportunities, ACM Comput. Surv. 55 (3) (2022) 1–38.
[14] J.B. Heaton, N.G. Polson, J.H. Witte, Deep learning for finance: deep portfolios, Appl. Stoch. Models Bus. Ind. 33 (1) (2017) 3–12.
[15] J. Heaton, N.G. Polson, J.H. Witte, Deep learning in finance, 2016, arXiv preprint arXiv:1602.06561.
[16] W. Totschnig, Fully autonomous AI, Sci. Eng. Ethics 26 (5) (2020) 2473–2485.
[17] Z. Peng, X. Li, F. Yan, An adaptive deep learning model for smart home autonomous system, in: 2020 International Conference on Intelligent Transportation, Big Data & Smart City, ICITBS, IEEE, 2020, pp. 707–710.
[18] K.-B. Ooi, G.W.-H. Tan, M. Al-Emran, M.A. Al-Sharafi, A. Capatina, A. Chakraborty, Y.K. Dwivedi, T.-L. Huang, A.K. Kar, V.-H. Lee, et al., The potential of generative artificial intelligence across disciplines: Perspectives and future directions, J. Comput. Inf. Syst. 65 (1) (2025) 76–107.
[19] D. Ha, Y. Tang, Collective intelligence for deep learning: A survey of recent developments, Collect. Intell. 1 (1) (2022) 26339137221114874.
[20] F. Boutros, V. Struc, J. Fierrez, N. Damer, Synthetic data for face recognition: Current state and future prospects, Image Vis. Comput. 135 (2023) 104688.
[21] A. Sufian, AI-generated videos and deepfakes: A technical primer, Authorea Prepr. (2024) http://dx.doi.org/10.36227/techrxiv.172348990.01007128/v1.
[22] S. Alanazi, S. Asif, A. Caird-daley, I. Moulitsas, Unmasking deepfakes: a multidisciplinary examination of social impacts and regulatory responses, Hum.-Intell. Syst. Integr. (2025) 1–23.
[23] K. Remya Revi, K. Vidya, M. Wilscy, Detection of deepfake images created using generative adversarial networks: A review, in: Second International Conference on Networks and Advances in Computational Technologies: NetACT 19, Springer, 2021, pp. 25–35.
[24] J. Gui, Z. Sun, Y. Wen, D. Tao, J. Ye, A review on generative adversarial networks: Algorithms, theory, and applications, IEEE Trans. Knowl. Data Eng. 35 (4) (2021) 3313–3332.

[25] F.-A. Croitoru, V. Hondru, R.T. Ionescu, M. Shah, Diffusion models in vision: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 45 (9) (2023) 10850–10869.
[26] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, M.-H. Yang, Diffusion models: A comprehensive survey of methods and applications, ACM Comput. Surv. 56 (4) (2023) 1–39.
[27] M. Westerlund, The emergence of deepfake technology: A review, Technol. Innov. Manag. Rev. 9 (11) (2019).
[28] M. Sharma, M. Kaur, A review of Deepfake technology: an emerging AI threat, in: Soft Computing for Security Applications: Proceedings of ICSCS 2021, Springer, 2022, pp. 605–619.
[29] S. Thompson, TikTok, AI-Generated Voices, and the Rise of Disinformation, The New York Times, 2023, (Accessed 08 March 2024).
[30] Y. Patel, S. Tanwar, R. Gupta, P. Bhattacharya, I.E. Davidson, R. Nyameko, S. Aluvala, V. Vimal, Deepfake generation and detection: Case study and challenges, IEEE Access (2023).
[31] S. Kopecky, Challenges of deepfakes, in: Science and Information Conference, Springer, 2024, pp. 158–166.
[32] T. Wang, X. Liao, K.P. Chow, X. Lin, Y. Wang, Deepfake detection: A comprehensive survey from the reliability perspective, ACM Comput. Surv. 57 (3) (2024) 1–35.
[33] F. Abbas, A. Taeihagh, Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence, Expert Syst. Appl. (2024) 124260.
[34] M. Li, Y. Ahmadiadli, X.-P. Zhang, A survey on speech deepfake detection, ACM Comput. Surv. (2025).
[35] E. Marcus, J. Teuwen, Artificial intelligence and explanation: How, why, and when to explain black boxes, Eur. J. Radiol. (2024) 111393.
[36] F. He, D. Tao, Deep learning: A (currently) Black-Box model, in: Foundations of Deep Learning, Springer, 2025, pp. 1–13.
[37] R. Matulionyte, A. Hanif, A call for more explainable AI in law enforcement, in: 2021 IEEE 25th International Enterprise Distributed Object Computing Workshop, EDOCW, IEEE, 2021, pp. 75–80.
[38] B. Sahoh, A. Choksuriwong, The role of explainable Artificial Intelligence in high-stakes decision-making systems: a systematic review, J. Ambient. Intell. Humaniz. Comput. 14 (6) (2023) 7827–7843.
[39] Y. Xu, K. Raja, M. Pedersen, Supervised contrastive learning for generalizable and explainable deepfakes detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 379–389.
[40] S.H. Silva, M. Bethany, A.M. Votto, I.H. Scarff, N. Beebe, P. Najafirad, Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models, Forensic Sci. Int.: Synerg. 4 (2022) 100217.
[41] I.U. Haq, K.M. Malik, K. Muhammad, Multimodal neurosymbolic approach for explainable deepfake detection, ACM Trans. Multimed. Comput. Commun. Appl. 20 (11) (2024) 1–16.
[42] B. Gowrisankar, V.L. Thing, An adversarial attack approach for explainable AI evaluation on deepfake detection models, Comput. Secur. 139 (2024) 103684.
[43] K. Tsigos, E. Apostolidis, S. Baxevanakis, S. Papadopoulos, V. Mezaris, Towards quantitative evaluation of explainable ai methods for deepfake detection, in: Proceedings of the 3rd ACM International Workshop on Multimedia AI Against Disinformation, 2024, pp. 37–45.
[44] P. Edwards, J.-C. Nebel, D. Greenhill, X. Liang, A review of deepfake techniques: Architecture, detection and datasets, IEEE Access (2024).
[45] V.K. Sharma, R. Garg, Q. Caudron, A systematic literature review on deepfake detection techniques, Multimedia Tools Appl. (2024) 1–43.
[46] G. Pei, J. Zhang, M. Hu, Z. Zhang, C. Wang, Y. Wu, G. Zhai, J. Yang, C. Shen, D. Tao, Deepfake generation and detection: A benchmark and survey, 2024, arXiv preprint arXiv:2403.17881.
[47] A. Heidari, N. Jafari Navimipour, H. Dag, M. Unal, Deepfake detection using deep learning methods: A systematic and comprehensive review, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 14 (2) (2024) e1520.
[48] S.M. Abdullah, A. Cheruvu, S. Kanchi, T. Chung, P. Gao, M. Jadliwala, B. Viswanath, An analysis of recent advances in deepfake image detection in an evolving threat landscape, in: 2024 IEEE Symposium on Security and Privacy, SP, IEEE, 2024, pp. 91–109.
[49] H.A. Mohamed Shaffril, S.F. Samsuddin, A. Abu Samah, The ABC of systematic literature review: the basic methodological guidance for beginners, Qual. Quant. 55 (2021) 1319–1346.
[50] R. Van Dinter, B. Tekinerdogan, C. Catal, Automation of systematic literature reviews: A systematic literature review, Inf. Softw. Technol. 136 (2021) 106589.
[51] L.A. Passos, D. Jodas, K.A. Costa, L.A. Souza Júnior, D. Rodrigues, J. Del Ser, D. Camacho, J.P. Papa, A review of deep learning-based approaches for deepfake content detection, Expert Syst. 41 (8) (2024) e13570.
[52] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. Yu, L. Sun, A survey of ai-generated content (aigc), ACM Comput. Surv. 57 (5) (2025) 1–38.
[53] G. Murphy, D. Ching, J. Twomey, C. Linehan, Face/Off: Changing the face of movies with deepfakes, PLoS One 18 (7) (2023) e0287503.
[54] E. Lundberg, P. Mozelius, The potential effects of deepfakes on news media and entertainment, AI Soc. (2024) 1–12.
[55] N. Komasawa, M. Yokohira, Simulation-based education in the artificial intelligence era, Cureus 15 (6) (2023).
[56] J. Roe, M. Perkins, L. Furze, Deepfakes and higher education: A research agenda and scoping review of synthetic media, J. Univ. Teach. Learn. Pr. 21 (10) (2024) 1–22.
[57] J.W. Seow, M.K. Lim, R.C. Phan, J.K. Liu, A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities, Neurocomputing 513 (2022) 351–371.
[58] M.S. Rana, M.N. Nobi, B. Murali, A.H. Sung, Deepfake detection: A systematic literature review, IEEE Access 10 (2022) 25494–25513.
[59] C. Vaccari, A. Chadwick, Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news, Soc. Media+ Soc. 6 (1) (2020) 2056305120903408.
[60] A.D. Samuel-Okon, O.I. Akinola, O.O. Olaniyi, O. Olateju, S.A. Ajayi, Assessing the effectiveness of network security tools in mitigating the impact of deepfakes AI on public trust in media, 2024, Available At SSRN 4908098.
[61] M. Oancea, AI and deep fake-video and audio manipulation techniques capable of altering the political process, Rev. Ştiinţe Politice. Rev. Sci. Politiques • No 81 (2024) 70–82.
[62] S.A. Laczi, V. Póser, Impact of deepfake technology on children: Risks and consequences, in: 2024 IEEE 22nd Jubilee International Symposium on Intelligent Systems and Informatics, SISY, IEEE, 2024, pp. 215–220.
[63] S. Alexander, Deepfake cyberbullying: The psychological toll on students and institutional challenges of AI-driven harassment, Clear. House: J. Educ. Strat. Issues Ideas 98 (2) (2025) 36–50.
[64] J. Kietzmann, L.W. Lee, I.P. McCarthy, T.C. Kietzmann, Deepfakes: Trick or treat? Bus. Horiz. 63 (2) (2020) 135–146.
[65] A. Chadha, V. Kumar, S. Kashyap, M. Gupta, Deepfake: an overview, in: Proceedings of Second International Conference on Computing, Communications, and Cyber-Security: IC4S 2020, Springer, 2021, pp. 557–566.
[66] H. Hao, E.R. Bartusiak, D. Güera, D. Mas Montserrat, S. Baireddy, Z. Xiang, S.K. Yarlagadda, R. Shao, J. Horváth, J. Yang, et al., Deepfake detection using multiple data modalities, in: Handbook of Digital Face Manipulation and Detection: From DeepFakes To Morphing Attacks, Springer International Publishing Cham, 2022, pp. 235–254.
[67] Y. Cai, P. Chen, J. Tian, J. Liu, J. Dai, X. Wang, S. Jia, S. Lyu, J. Han, Modality-agnostic deepfakes detection, in: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, 2025, pp. 12–23.
[68] C. Xu, J. Zhang, M. Hua, Q. He, Z. Yi, Y. Liu, Region-aware face swapping, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7632–7641.
[69] B. Huang, Z. Wang, J. Yang, J. Ai, Q. Zou, Q. Wang, D. Ye, Implicit identity driven deepfake face swapping detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4490–4499.
[70] G. Mazaheri, A.K. Roy-Chowdhury, Detection and localization of facial expression manipulations, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 1035–1045.
[71] S. Waseem, S.A.R.S.A. Bakar, B.A. Ahmed, Z. Omar, T.A.E. Eisa, DeepFake on face and expression swap: A review, IEEE Access 11 (2023) 117865–117906.
[72] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C.S. Facenheim, L. RP, J. Jiang, et al., DeepFaceLab: Integrated, flexible and extensible face-swapping framework, 2020, arXiv preprint arXiv:2005.05535.
[73] P. Kumar, M. Vatsa, R. Singh, Detecting face2face facial reenactment in videos, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 2589–2597.
[74] B. Hosler, D. Salvi, A. Murray, F. Antonacci, P. Bestagini, S. Tubaro, M.C. Stamm, Do deepfakes feel emotions? A semantic approach to detecting deepfakes via emotional inconsistencies, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1013–1022.
[75] Y. Rodríguez-Ortega, D.M. Ballesteros, D. Renza, A machine learning model to detect fake voice, in: International Conference on Applied Informatics, Springer, 2020, pp. 3–13.
[76] X. Tan, T. Qin, F. Soong, T.-Y. Liu, A survey on neural speech synthesis, 2021, arXiv preprint arXiv:2106.15561.
[77] Z. Khanjani, G. Watson, V.P. Janeja, How deep are the fakes? focusing on audio deepfake: A survey, 2021, arXiv preprint arXiv:2111.14203.
[78] J. Pu, Z. Sarwar, S.M. Abdullah, A. Rehman, Y. Kim, P. Bhattacharya, M. Javed, B. Viswanath, Deepfake text detection: Limitations and opportunities, in: 2023 IEEE Symposium on Security and Privacy, SP, IEEE, 2023, pp. 1613–1630.
[79] A. Uchendu, S. Venkatraman, T. Le, D. Lee, Catch me if you gpt: Tutorial on deepfake texts, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts), 2024, pp. 1–7.
[80] J. Botha, H. Pieterse, Fake news and deepfakes: A dangerous threat for 21st century information security, in: ICCWS 2020 15th International Conference on Cyber Warfare and Security. Academic Conferences and Publishing Limited, 2020, p. 57.
[81] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, A. Hussain, Interpreting black-box models: a review on explainable artificial intelligence, Cogn. Comput. 16 (1) (2024) 45–74.

[82] D. Becking, M. Dreyer, W. Samek, K. Müller, S. Lapuschkin, Ecq x: explainability-driven quantization for low-bit and sparse DNNs, in: International Workshop on Extending Explainable AI beyond Deep Models and Classifiers, Springer, 2020, pp. 271–296.

[83] P. Gohel, P. Singh, M. Mohanty, Explainable AI: current status and future directions, 2021, arXiv preprint arXiv:2107.07045.

[84] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2020) 336–359.

[85] D. Dworak, J. Baranowski, Adaptation of Grad-CAM method to neural network architecture for LiDAR pointcloud object detection, Energies 15 (13) (2022) 4681.

[86] M. Xiao, L. Zhang, W. Shi, J. Liu, W. He, Z. Jiang, A visualization method based on the grad-CAM for medical image segmentation model, in: 2021 International Conference on Electronic Information Engineering and Computer Science, EIECS, IEEE, 2021, pp. 242–247.

[87] S. Kolekar, S. Gite, B. Pradhan, A. Alamri, Explainable AI in scene understanding for autonomous vehicles in unstructured traffic environments on Indian roads using the inception U-Net model with grad-CAM visualization, Sens. 22 (24) (2022) 9677.

[88] T. Chakraborty, U. Trehan, K. Mallat, J.-L. Dugelay, Generalizing adversarial explanations with Grad-CAM, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 187–193.

[89] S. Mamarasulov, L. Chen, C. Chen, Y. Li, C. Wang, Data augmentation with attention framework for robust deepfake detection, Vis. Comput. 41 (7) (2025) 4779–4798.

[90] M.K.Z. Bajwa, A. Castiglione, C. Pero, Mel spectrogram-based CNN framework for explainable audio deepfake detection, in: International Conference on Advanced Information Networking and Applications, Springer, 2025, pp. 407–416.

[91] S. Bramhall, H. Horn, M. Tieu, N. Lohia, Qlime-a quadratic local interpretable model-agnostic explanation approach, SMU Data Sci. Rev. 3 (1) (2020) 4.

[92] Y. Wu, L. Zhang, U.A. Bhatti, M. Huang, Interpretable machine learning for personalized medical recommendations: A LIME-based approach, Diagn. 13 (16) (2023) 2681.

[93] J. Oh, C.J. Hughes, G. Venkataramani, M. Prvulovic, LIME: A framework for debugging load imbalance in multi-threaded execution, in: Proceedings of the 33rd International Conference on Software Engineering, 2011, pp. 201–210.

[94] M.O. Faruk, R.B. Reza, S.H. Sourav, M. Hasan, M. Fazle Rabbi, M.A. Marjan, An explainable AI enable approach to reveal feature influences on social media customer purchase decisions, in: International Conference on Big Data, IoT and Machine Learning, Springer, 2023, pp. 79–94.

[95] S. Parmar, C. Mala, Trustworthiness in deepfake detection using explainability, in: Digital Image Security, CRC Press, 2024, pp. 159–177.

[96] E. Soltandoost, R. Plesh, S. Schuckers, P. Peer, V. Štruc, Extracting local information from global representations for interpretable deepfake detection, in: Proceedings of the Winter Conference on Applications of Computer Vision, 2025, pp. 1629–1639.

[97] Y. Nohara, K. Matsumoto, H. Soejima, N. Nakashima, Explanation of machine learning models using shapley additive explanation and application for real data in hospital, Comput. Methods Programs Biomed. 214 (2022) 106584.

[98] Y. Zhu, W.Z. Taffese, G. Chen, Enhancing FRP-concrete interface bearing capacity prediction with explainable machine learning: A feature engineering approach and SHAP analysis, Eng. Struct. 319 (2024) 118831.

[99] K. Lin, Y. Gao, Model interpretability of financial fraud detection by group SHAP, Expert Syst. Appl. 210 (2022) 118354.

[100] A. Nguyen, S. Foerstel, T. Kittler, A. Kurzyukov, L. Schwinn, D. Zanca, T. Hipp, S. Da Jun, M. Schrapp, E. Rothgang, et al., System design for a data-driven and explainable customer sentiment monitor using IoT and enterprise data, IEEE Access 9 (2021) 117140–117152.

[101] W. Ge, J. Patino, M. Todisco, N. Evans, Explaining deep learning models for spoofing and deepfake detection with shapley additive explanations, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2022, pp. 6387–6391.

[102] T. Huber, D. Schiller, E. André, Enhancing explainability of deep reinforcement learning through selective layer-wise relevance propagation, in: KI 2019: Advances in Artificial Intelligence: 42nd German Conference on AI, Kassel, Germany, September 23–26, 2019, Proceedings 42, Springer, 2019, pp. 188–202.

[103] L. Arras, F. Horn, G. Montavon, K.-R. Müller, W. Samek, Explaining predictions of non-linear classifiers in NLP, 2016, arXiv preprint arXiv:1606.07298.

[104] B.C. Case, C. Shea, R. Torguson, C. Zhang, C. Yerasi, G.A. Medranda, K.O. Kuku, H.M. Garcia-Garcia, G.S. Mintz, R. Waksman, Impact of baseline imaging of non-culprit coronary lesions on adverse events: Insight from LRP study, Cardiovasc. Revascularization Med. 39 (2022) 1–5.

[105] K. Nakagawa, T. Uchida, T. Aoshima, Deep factor model: Explaining deep learning decisions for forecasting stock returns with layer-wise relevance propagation, in: Workshop on Mining Data for Financial Applications, Springer, 2018, pp. 37–50.

[106] M. David, M.M. William, N. Bridget, N.H. Raudha, R. Nakibuule, G. Marvin, Layer-wise relevance backward propagation for fake face detection and reconstruction, in: 2024 5th International Conference on Image Processing and Capsule Networks, ICIPCN, IEEE, 2024, pp. 586–594.

[107] S.-Y. Lim, D.-K. Chae, S.-C. Lee, Detecting deepfake voice using explainable deep learning techniques, Appl. Sci. 12 (8) (2022) 3926.

[108] C. Seibold, A. Hilsmann, P. Eisert, Focused LRP: Explainable AI for face morphing attack detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 88–96.

[109] K. Jayakumar, N. Skandhakumar, A visually interpretable forensic deepfake detection tool using anchors, in: 2022 7th International Conference on Information Technology Research, ICITR, IEEE, 2022, pp. 1–6.

[110] G. Lopardo, F. Precioso, D. Garreau, A sea of words: An in-depth analysis of anchors for text data, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2023, pp. 4848–4879.

[111] B. Raufi, C. Finnegan, L. Longo, A comparative analysis of SHAP, LIME, ANCHORS, and DICE for interpreting a dense neural network in credit card fraud detection, in: World Conference on Explainable Artificial Intelligence, Springer, 2024, pp. 365–383.

[112] O. Barkan, R. Hirsch, O. Katz, A. Caciularu, N. Koenigstein, Anchor-based collaborative filtering, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 2877–2881.

[113] R. Xu, X. Liu, D. Nagothu, Q. Qu, Y. Chen, Detecting manipulated digital entities through real-world anchors, in: International Conference on Advanced Information Networking and Applications, Springer, 2025, pp. 450–461.

[114] X. Lu, A. Tolmachev, T. Yamamoto, K. Takeuchi, S. Okajima, T. Takebayashi, K. Maruhashi, H. Kashima, Crowdsourcing evaluation of saliency-based XAI methods, in: Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part V 21, Springer, 2021, pp. 431–446.

[115] G. Sharma, F. Jurie, C. Schmid, Discriminative spatial saliency for image classification, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3506–3513.

[116] N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J. Patel, M. Gidwani, et al., Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging, Radiol.: Artif. Intell. 3 (6) (2021) e200267.

[117] M.T. Nguyen, P. Siritanawan, K. Kotani, Saliency detection in human crowd images of different density levels using attention mechanism, Signal Process., Image Commun. 88 (2020) 115976.

[118] Q. Li, M. Gao, G. Zhang, W. Zhai, Defending deepfakes by saliency-aware attack, IEEE Trans. Comput. Soc. Syst. (2023).

[119] G. Cartella, V. Cuculo, M. Cornia, M. Papasidero, F. Ruozzi, R. Cucchiara, Pixels of faith: Exploiting visual saliency to detect religious image manipulation, in: European Conference on Computer Vision, Springer, 2025, pp. 229–245.

[120] A. Katiyar, A. Bhavsar, Image forgery detection with interpretability, 2022, arXiv preprint arXiv:2202.00908.

[121] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.

[122] S. Mathews, S. Trivedi, A. House, S. Povolny, C. Fralick, An explainable deepfake detection framework on a novel unconstrained dataset, Complex Intell. Syst. 9 (4) (2023) 4425–4437.

[123] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: a compact facial video forgery detection network, in: 2018 IEEE International Workshop on Information Forensics and Security, WIFS, IEEE, 2018, pp. 1–7.

[124] J.J. Bird, A. Lotfi, Cifake: Image classification and explainable identification of ai-generated synthetic images, IEEE Access (2024).

[125] M.Z. Hossain, F.U. Zaman, M.R. Islam, Advancing AI-generated image detection: Enhanced accuracy through CNN and vision transformer models with explainable AI insights, in: 2023 26th International Conference on Computer and Information Technology, ICCIT, IEEE, 2023, pp. 1–6.

[126] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2020.

[127] B. Sugiantoro, Deepfake face images: Explainable detection using deep neural networks and class activation mapping, in: 2024 IEEE International Symposium on Consumer Technology, ISCT, IEEE, 2024, pp. 86–90.

[128] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[129] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

[130] P.K. Rangarajan, M. Sukesh, D. Abinandhini, Y. Jaikanth, et al., Detecting AI-generated images with CNN and interpretation using explainable AI, in: 2024 IEEE International Conference on Contemporary Computing and Communications, InC4, Vol. 1, IEEE, 2024, pp. 1–6.

[131] C. Lab, Real and fake face detection, 2021, https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection. (Accessed 29 June 2025).
[132] F. Khalid, A. Javed, K.M. Malik, A. Irtaza, ExplaNET: A descriptive framework for detecting deepfakes with interpretable prototypes, IEEE Trans. Biom. Behav. Identity Sci. (2024).
[133] J. Xiang, G. Zhu, Joint face detection and facial expression recognition with MTCNN, in: 2017 4th International Conference on Information Science and Control Engineering, ICISCE, IEEE, 2017, pp. 424–427.
[134] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1–11.
[135] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df: A large-scale challenging dataset for deepfake forensics, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3207–3216.
[136] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C.C. Ferrer, The deepfake detection challenge (dfdc) dataset, 2020, arXiv preprint arXiv:2006.07397.
[137] M.T. Islam, I.H. Lee, A.I. Alzahrani, K. Muhammad, MEXFIC: A meta ensemble explainable approach for AI-synthesized fake image classification, Alex. Eng. J. 116 (2025) 351–363.
[138] C. Seibold, A. Hilsmann, P. Eisert, Feature focus: Towards explainable and transparent deep face morphing attack detectors, Comput. 10 (9) (2021) 117.
[139] A. Sengupta, Y. Ye, R. Wang, C. Liu, K. Roy, Going deeper in spiking neural networks: VGG and residual architectures, Front. Neurosci. 13 (2019) 95.
[140] X. Ma, J. Tian, Z. Li, Y. Chai, L. Zang, J. Han, Explainable deepfake detection with human prompts, in: 2024 27th International Conference on Computer Supported Cooperative Work in Design, CSCWD, IEEE, 2024, pp. 3023–3029.
[141] Y. Zhu, X. Wang, H.-S. Chen, R. Salloum, C.-C.J. Kuo, A-pixelhop: A green, robust and explainable fake-image detector, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2022, pp. 8947–8951.
[142] N. Mansoor, A.I. Iliev, Explainable AI for DeepFake detection, Appl. Sci. 15 (2) (2025) 725.
[143] C.-H. Lee, Z. Liu, L. Wu, P. Luo, Maskgan: Towards diverse and interactive facial image manipulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5549–5558.
[144] A. Aghasanli, D. Kangin, P. Angelov, Interpretable-through-prototypes deepfake detection for diffusion models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 467–474.
[145] S. Suthaharan, S. Suthaharan, Support vector machine, in: Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, Springer, 2016, pp. 207–235.
[146] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.
[147] A. Rougetet, Flickr-faces-HQ dataset (FFHQ), 2019, https://www.kaggle.com/datasets/arnaud58/flickrfaceshq-dataset-ffhq. (Accessed 29 June 2025).
[148] J. Li, F. Zhang, J. Zhu, E. Sun, Q. Zhang, Z.-J. Zha, ForgeryGPT: Multimodal large language model for explainable image forgery detection and localization, 2024, arXiv preprint arXiv:2410.10238.
[149] J. Dong, W. Wang, T. Tan, Casia image tampering detection evaluation database, in: 2013 IEEE China Summit and International Conference on Signal and Information Processing, IEEE, 2013, pp. 422–426.
[150] V.V. Kniaz, V. Knyaz, F. Remondino, The point where reality meets fantasy: Mixed adversarial generators for image splice detection, Adv. Neural Inf. Process. Syst. 32 (2019).
[151] O.M. Parkhi, A. Vedaldi, A. Zisserman, C. Jawahar, Cats and dogs, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3498–3505.
[152] NVlabs:, Ffhq-dataset, 2025, Available at-https://github.com/NVlabs/ffhq-dataset?tab=readme-ov-file. (Accessed 31 January 2025).
[153] G. Dutta, 140K real and fake faces, 2025, Available at-https://www.kaggle.com/datasets/gauravduttakiit/140k-real-and-fake-faces. (Accessed 31 February 2025).
[154] H. Farid, Creating, using, misusing, and detecting deep fakes, J. Online Trust. Saf. 1 (4) (2022).
[155] G.H. Ishrak, Z. Mahmud, M. Farabe, T.K. Tinni, T. Reza, M.Z. Parvez, Explainable deepfake video detection using convolutional neural network and CapsuleNet, 2024, arXiv preprint arXiv:2404.12841.
[156] N. Van Quang, J. Chun, T. Tokuyama, CapsuleNet for micro-expression recognition, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2019, IEEE, 2019, pp. 1–7.
[157] K. Greff, R.K. Srivastava, J. Koutník, B.R. Steunebrink, J. Schmidhuber, LSTM: A search space odyssey, IEEE Trans. Neural Netw. Learn. Syst. 28 (10) (2016) 2222–2232.
[158] F. Khalid, A. Javed, K.M. Malik, A. Irtaza, DFP-Net: An explainable and trustworthy framework for detecting deepfakes using interpretable prototypes, in: 2023 IEEE International Joint Conference on Biometrics, IJCB, IEEE, 2023, pp. 1–9.
[159] M. Nawaz, A. Javed, A. Irtaza, A deep learning model for FaceSwap and face-reenactment deepfakes detection, Appl. Soft Comput. (2024) 111854.
[160] S. Venkateswarulu, A. Srinagesh, DeepExplain: Enhancing DeepFake detection through transparent and explainable AI model, Inform. 48 (8) (2024).
[161] H. Ilyas, A. Javed, K.M. Malik, ConvNext-PNet: An interpretable and explainable deep-learning model for deepfakes detection, in: 2024 IEEE International Joint Conference on Biometrics, IJCB, IEEE, 2024, pp. 1–9.
[162] H. Song, S. Huang, Y. Dong, W.-W. Tu, Robustness and generalizability of deepfake detection: A study with diffusion models, 2023, arXiv preprint arXiv:2309.02218.
[163] Y. Hua, R. Shi, P. Wang, S. Ge, Learning patch-channel correspondence for interpretable face forgery detection, IEEE Trans. Image Process. 32 (2023) 1668–1680.
[164] R. Uma Maheshwari, B. Paulchamy, Securing online integrity: a hybrid approach to deepfake detection and removal using explainable AI and Adversarial robustness training, Autom.: Çasopis Autom. Mjer. Elektron., Računarstvo Komun. 65 (4) (2024) 1517–1532.
[165] B. Malolan, A. Parekh, F. Kazi, Explainable deep-fake detection using visual interpretability methods, in: 2020 3rd International Conference on Information and Computer Technologies, ICICT, IEEE, 2020, pp. 289–293.
[166] Google Research Blog, Contributing data to deepfake detection research, 2019, https://research.google/blog/contributing-data-to-deepfake-detection-research/. (Accessed 29 June 2025).
[167] L. Trinh, M. Tsang, S. Rambhatla, Y. Liu, Interpretable and trustworthy deepfake detection via dynamic prototypes, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1973–1983.
[168] P. Gupta, K. Chugh, A. Dhall, R. Subramanian, The eyes know it: Fakeet-an eye-tracking database to understand deepfake perception, in: Proceedings of the 2020 International Conference on Multimodal Interaction, 2020, pp. 519–527.
[169] L. Jiang, R. Li, W. Wu, C. Qian, C.C. Loy, Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2889–2898.
[170] L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: Uniform manifold approximation and projection, J. Open Source Softw. 3 (29) (2018).
[171] F. Khalid, A. Javed, H. Ilyas, A. Irtaza, et al., DFGNN: An interpretable and generalized graph neural network for deepfakes detection, Expert Syst. Appl. 222 (2023) 119843.
[172] H.E. Goemans, K.S. Gleditsch, G. Chiozza, Introducing Archigos: A dataset of political leaders, J. Peace Res. 46 (2) (2009) 269–283.
[173] B. Zi, M. Chang, J. Chen, X. Ma, Y.-G. Jiang, Wilddeepfake: A challenging real-world dataset for deepfake detection, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2382–2390.
[174] N. Yu, L. Chen, T. Leng, Z. Chen, X. Yi, An explainable deepfake of speech detection method with spectrograms and waveforms, J. Inf. Secur. Appl. 81 (2024) 103720.
[175] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K.A. Lee, et al., Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech, Comput. Speech Lang. 64 (2020) 101114.
[176] C. Bisogni, V. Loia, M. Nappi, C. Pero, Acoustic features analysis for explainable machine learning-based audio spoofing detection, Comput. Vis. Image Underst. 249 (2024) 104145.
[177] A. Nautsch, X. Wang, N. Evans, T.H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, K.A. Lee, ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech, IEEE Trans. Biom. Behav. Identity Sci. 3 (2) (2021) 252–265.
[178] H. Khalid, S. Tariq, M. Kim, S.S. Woo, FakeAVCeleb: A novel audio-video multimodal deepfake dataset, Adv. Neural Inf. Process. Syst. (2021).
[179] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, K. Böttinger, Does audio deepfake detection generalize? in: Proc. Interspeech 2022, 2022, pp. 2783–2787.
[180] A. Govindu, P. Kale, A. Hullur, A. Gurav, P. Godse, Deepfake audio detection and justification with explainable artificial intelligence (XAI), Appl. Sci. (2023).
[181] F. Sultana, A. Sufian, P. Dutta, Advancements in image classification using convolutional neural network, in: 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks, ICRCICN, IEEE, 2018, pp. 122–129.
[182] N.M. Müller, P. Sperl, K. Böttinger, Complex-valued neural networks for voice anti-spoofing, in: Proc. Interspeech 2023, 2023, pp. 3814–3818.
[183] A. Hirose, et al., Complex-Valued Neural Networks, vol. 32, Wiley Online Library, 2006.
[184] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, Smoothgrad: removing noise by adding noise, 2017, arXiv preprint arXiv:1706.03825.
[185] D. Salvi, P. Bestagini, S. Tubaro, Towards frequency band explainability in synthetic speech detection, in: 2023 31st European Signal Processing Conference, EUSIPCO, IEEE, 2023, pp. 620–624.
[186] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, A. Larcher, End-to-end anti-spoofing with rawnet2, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 6369–6373.
[187] B. Mathew, P. Saha, S.M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, Hatexplain: A benchmark dataset for explainable hate speech detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 14867–14875.

[188] J. DeYoung, S. Jain, N.F. Rajani, E. Lehman, C. Xiong, R. Socher, B.C. Wallace, ERASER: A benchmark to evaluate rationalized NLP models, Trans. Assoc. Comput. Linguist. (2020).
[189] G. Channing, J. Sock, R. Clark, P. Torr, C.S. de Witt, Toward robust real-world audio deepfake detection: Closing the explainability gap, 2024, arXiv preprint arXiv:2410.07436.
[190] X. Wang, H. Delgado, H. Tak, J.-w. Jung, H.-j. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T.H. Kinnunen, et al., ASVspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale, in: Proc. ASVspoof 2024, 2024, pp. 1–8.
[191] Y. Li, J. Zeng, S. Shan, X. Chen, Occlusion aware facial expression recognition using CNN with attention mechanism, IEEE Trans. Image Process. 28 (5) (2018) 2439–2450.
[192] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K.A. Lee, T. Kinnunen, N. Evans, et al., Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection, in: ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge, 2021, p. n.p..
[193] K. Ito, L. Johnson, The LJ speech dataset, 2025, https://keithito.com/LJ-Speech-Dataset/. (Accessed 14 January 2025).
[194] O. Ibrahim Aboulola, M. Umer, Novel approach for Arabic fake news classification using embedding from large language features with CNN-LSTM ensemble model and explainable AI, Sci. Rep. 14 (1) (2024) 1–13.
[195] L. Wu, Y. Long, C. Gao, Z. Wang, Y. Zhang, MFIR: Multimodal fusion and inconsistency reasoning for explainable fake news detection, Inf. Fusion 100 (2023) 101944.
[196] M. Szczepański, M. Pawlicki, R. Kozik, M. Choraś, New explainability method for BERT-based model in fake news detection, Sci. Rep. 11 (1) (2021) 23705.
[197] H. Chi, B. Liao, A quantitative argumentation-based automated explainable decision system for fake news detection on social media, Knowl.-Based Syst. 242 (2022) 108378.
[198] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G.W.S. Hoi, A. Zubiaga, SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours, 2017, arXiv preprint arXiv:1704.05972.
[199] X. Gao, W. Chen, L. Lu, Y. Cui, X. Dai, L. Dai, K. Wang, J. Shen, Y. Wang, S. Wang, et al., An interpretable fake news detection method based on commonsense knowledge graph, Appl. Sci. 13 (11) (2023) 6680.
[200] M. Hosseini, A.J. Sabet, S. He, D. Aguiar, Interpretable fake news detection with topic and deep variational models, Online Soc. Netw. Media 36 (2023) 100249.
[201] H. Ahmed, I. Traore, S. Saad, Detecting opinion spams and fake news using text classification, Secur. Priv. 1 (1) (2018) e9.
[202] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M.S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Fighting an infodemic: Covid-19 fake news dataset, in: Combating Online Hostile Posts in Regional Languages During Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Springer, 2021, pp. 21–29.
[203] F. Zhang, J. Liu, Q. Zhang, E. Sun, J. Xie, Z.-J. Zha, ECENet: Explainable and context-enhanced network for muti-modal fact verification, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 1231–1240.
[204] M. Al-Naeem, M. Hafizur Rahman, A. Banerjee, A. Sufian, BSM-DND: Bias and sensitivity-aware multilingual deepfake news detection using bloom filters and recurrent feature elimination, IEEE Access (2025).
[205] clmentbisaillon:, Fake-and-real-news-dataset, 2025, Available at-https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset. (Accessed 14 Januray 2025).
[206] G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, L. Derczynski, Semeval-2019 task 7: Rumoureval 2019: Determining rumour veracity and support for rumours, in: Proceedings of the 13th International Workshop on Semantic Evaluation: NAACL HLT 2019, Association for Computational Linguistics, 2019, pp. 845–854.
[207] G. Luo, T. Darrell, A. Rohrbach, NewsCLIPpings: Automatic generation of out-of-context multimodal media, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 6801–6817.
[208] S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A.N. Reganti, P. Patwa, A. Das, T. Chakraborty, A.P. Sheth, A. Ekbal, et al., FACTIFY: A multi-modal fact verification dataset, in: DE-FACTIFY@ AAAI, 2022, p. n.p..
[209] S. Banik, COVID fake news dataset, 2025, Available at-doi:10.5281/zenodo.4282522. (Accessed 14 January 2025).
[210] C. Boididou, S. Papadopoulos, M. Zampoglou, L. Apostolidis, O. Papadopoulou, Y. Kompatsiaris, Detection and visualization of misleading content on Twitter, Int. J. Multimed. Inf. Retr. 7 (1) (2018) 71–86.
[211] Z. Jin, J. Cao, H. Guo, Y. Zhang, J. Luo, Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 795–816.
[212] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media, Big Data 8 (3) (2020) 171–188.
[213] M. Muneer, Arabic-fake-news-dataset-afnd, 2025, Available at-https://https://www.kaggle.com/datasets/murtadhayaseen/arabic-fake-news-dataset-afnd. (Accessed 01 February 2025).