xxx (xxxx) xxx

Contents lists available at ScienceDirect

Data and Information Management

journal homepage: www.journals.elsevier.com/data-and-information-management

DeepFake video detection: Insights into model generalisation — A Systematic review

Ramcharan Ramanaharan, Deepani B. Guruge *, Johnson I. Agbinya

Melbourne Institute of Technology, 288 La Trobe Street, Melbourne, VIC 3000, Australia

ARTICLE INFO

Keywords:
DeepFake
Detection
Generalisability
Systematic review
Machine learning

ABSTRACT

Deep learning generative models have progressed to a stage where distinguishing fake images and videos has become difficult, posing risks to personal integrity, potentially leading to social instability, and disrupting government functioning. Existing reviews have mainly focused on the approaches used to detect DeepFakes, and the data sets used for those approaches. However, challenges persist when attempting to generalise detection techniques to identify previously unseen datasets. The purpose of this systematic review is to explore state-of-the-art frameworks for DeepFake detection and provide readers with an understanding of the strengths and weaknesses of current approaches, as well as the generalisability of existing detection techniques. The study indicates that generalising DeepFake detection remains a challenge that requires further research. Moreover, 46.3% of the selected publications agreed that DeepFake detection techniques could be generalised to identify various types of DeepFakes. A key limitation in achieving generalisation is the tendency of models to overfit to available data datasets, reducing their effectiveness in adapting to new or unseen types of DeepFakes. This review emphasises the need for the development of extensive and diverse datasets that more accurately reflect the wide range of DeepFake manipulations encountered in real-world applications. Lastly, the paper explores potential advancements that could pave the way to the next generation of solutions against DeepFakes.

1. Introduction

Advancements in the field of artificial intelligence (AI) are progressing at an unprecedented pace, presenting complex challenges that demand careful consideration. A prominent example is the generation of synthetic audiovisual content that is nearly indistinguishable from authentic material (Jada & Mayayise, 2024; Lomnitz, Hampel-Arias, Sandesara, & Hu, 2020). This advancement is largely powered by cutting-edge AI technologies, including auto-encoders and generative adversarial networks (GANs) (Gong, Goh, Kumar, Ye, & Chi, 2020; Ismail, Elpeltagy, Zaki, Eldahshan, Kamal, 2021a). The development and dissemination of DeepFakes (DF) have become a critical area of concern. DeepFakes are fake videos, audio, or images created using deep learning algorithms, often with malicious intent to harm individuals. These falsified media can severely damage someone's reputation, influence public opinion on significant topics, influence election results (Agarwal, Farid, El-Gaaly, & Lim, 2020; Deng, Suo, Li, et al., 2022), posing substantial threats to personal privacy and public safety (Singh, Saimbhi, Singh, & Mittal, 2020).

Furthermore, widely available and user-friendly apps such as FakeApp and DeepFaceLab can lead to increased abuse, which is quite alarming (Liu et al., 2023). We have reached a point where it is essential to develop generalisable DF detection techniques as well as to establish strict guidelines or regulations to mitigate the creation and use of DeepFake images, audio, or video (Gambini, Fagni, Falchi, & Tesconi, 2022). Researchers and technology experts strive to develop reliable methods to detect DFs to mitigate this threat (Bondi, Cannas, Bestagini, & Tubaro, 2020; Suratkar, Kazi, Sakhalkar, Abhyankar & Kshirsagar, 2020; Zi, Chang, Chen, Ma, & Jiang, 2020). Detecting DFs has become an increasingly challenging task due to the rapid advancements in Generative AI (GenAI) algorithms, which malicious users exploit to produce more authentic DFs (Bansal et al., 2023a; Coccomini, Caldelli, Falchi, Gennaro, & Amato, 2022; Li et al., 2020; Zhao, Zhang, Ding, & Cui, 2021). Most current research relies heavily on well-known data sets, including FF++, the DeepFake Detection Challenge (DFDC), and Celeb-DF (Du, Pentyala, Li, & Hu, 2020). Although these data sets are valuable, they do not cover the full range of DFs encountered in real-world scenarios (Du et al., 2020; Malik, Kuribayashi, Abdullahi, & Khan, 2022). This reliance on a limited number of datasets causes a significant challenge in developing robust detection methods. This systematic review of the literature aims to investigate the generalisability of existing DeepFake detection techniques in published research studies.

* Corresponding author.
E-mail addresses: mit225094@stud.mit.edu.au (R. Ramanaharan), dguruge@mit.edu.au (D.B. Guruge), jagbinya@mit.edu.au (J.I. Agbinya).

https://doi.org/10.1016/j.dim.2025.100099
Received 15 October 2024; Received in revised form 12 March 2025; Accepted 13 March 2025

This paper is organised as follows: Section 2 reviews related systematic reviews published between 2022 February and 2024 August. Section 3 outlines the methodology used in this systematic literature review (SLR), while Section 4 covers the findings of this SLR. Section 5 is a discussion, and Section 6 concludes and highlights future work.

2. Background and related work

Researchers have conducted comprehensive systematic reviews to assess the current state of DeepFake detection research, while some studies have focused on conducting surveys of DeepFake detection techniques, providing a broader exploration of the methodologies and advancement in this domain. For example, a survey conducted by Masood et al. discusses the primary challenges associated with DeepFake detection and proposes potential countermeasures to address them. The authors categorised visual DF into five distinct categories based on the level of manipulation: face swap, lip syncing, puppet-mastery, complete face synthesis, and facial attribute manipulation. They also provide insights into the challenges that arise from post-processing operations such as noise effects, compressions, and lighting variations (Masood et al., 2023). Researchers in Kaur, Noori Hoshyar, Saikrishna, Firmin, and Xia (2024), Rana, Nobi, Murali, and Sung (2022) organise DeepFake research into four groups: methods based on machine learning, deep learning, statistical measurement, and blockchain technologies. At the conclusion of the review conducted by Kaur et al. (2024) highlighted hybrid methods hold significant potential to achieve high classification accuracy for fake videos in real-time. These categories reflect the diverse techniques used to create and detect DF, emphasising the complexity of addressing this evolving technology. The review in Malik et al. (2022) provides insights into challenges in generalising existing DF detection techniques. These include; a lack of comprehensive DeepFake datasets, unknown types of attacks, inter-frame temporal consistency issues, and additional efforts required for labelling (or adding scores corresponding to the type of forgery in datasets). Mirsky and Lee (2021) proposed unpaired self-supervised training techniques to reduce the need for large training datasets in deep learning algorithms. Addressing these challenges will be crucial to developing more effective and generalisable DeepFake detection techniques.

Table 1 summarises eight previous systematic reviews on DeepFake detection techniques published between 2022 February and 2024 August. These reviews focus primarily on the effectiveness of existing DeepFake detection techniques (Passos et al., 2024; Rana et al., 2022; Stroebel, Llewellyn, Hartley, Ip, & Ahmed, 2023), rather than on their generalisability. A systematic review conducted by Stroebel et al. (2023) explored the challenges associated with detecting DF across various modalities, including audio, images, and videos, and did not delve deeply into video-specific DeepFake detection techniques. It highlighted that the combination (CNN, DNN and LSTM) producing ensembled and multi-attentional architectures will strengthen the detection techniques. A review by Rana et al. (2022) categorise detection methods into three primary types: naive detectors, spatial detectors, and frequency detectors. In particular, their review reveals that 77% of the studies used deep learning-based approaches for detecting DF. Misirlis and Munawar (2023) reviewed 41 documents and discussed the risks, threats, and ethical considerations associated with DeepFake technology. However, their discussion lacks an analysis of the generalisation of DeepFake models, and the major issues and drawbacks in the existing detection models.

The authors in Sharma, Garg, and Caudron (2024) reviewed the datasets and features used to detect DF and proposed six prominent DeepFake detection techniques; XceptionNet, ResNet-50, VGG16, Capsule Networks, 3D-CNN, EfficientNet-V2. Passos et al. (2024) evaluated the detection of DF using deep learning-based approaches. The authors emphasised that the combination of supervised and unsupervised learning techniques will support handling rapidly evolving complex DF.

Existing systematic reviews in the field (Table 1) of DeepFake detection exhibit significant limitations. These reviews fail to critically evaluate the main features, strengths, and weaknesses of the proposed techniques in the selected publications. Additionally, most do not investigate the authors views on the adaptability of their proposed models across diverse datasets and strengths and weaknesses toward developing them to generalisable models. As a result, the generalisability of the current models and their best practices require further investigation to provide clearer insights for the DeepFake research community.

This systematic review aims to address these gaps by offering a comprehensive evaluation of existing video DeepFake detection models. It will analyse their strengths, distinctive features, and limitations while identifying the challenges associated with generalising these models to unseen manipulation types. Such insights will contribute to a more holistic understanding of the field, providing valuable information on the key features and algorithms of tested models in a single, consolidated table, facilitating to design more robust and effective algorithms.

Finally, Section 5 offers recommendations for future research, addressing challenges and incorporating the latest technological advancement in video DeepFake detection research. These recommended approaches aim to contribute to the development of more robust and generalisable detection models in future.

3. Systematic literature review (SLR) methodology

This systematic review aims to explore existing DF detection techniques and provide the reader with an understanding of the strengths and weaknesses of current approaches, as well as the possibility of generalising these existing approaches. In particular, the contribution of this systematic review is threefold:

Investigation of detection techniques: An in-depth examination of the various algorithms used for the detection of DF is conducted to understand how these techniques can be applied in various contexts. These algorithms are evaluated with respect to their strengths, weaknesses, and accuracy in detecting DF, as shown in Table 3.

Analysis and classification: Investigation and categorisation of the approaches identified in the reviewed literature to identify the most widely used deep fake detection technique, as shown in Fig. 6.

Generalisability of detection algorithms: The algorithms applied in the reviewed literature are analysed in terms of the potential to generalise across different datasets and real-life applications. This assessment is critical to understanding the gaps in existing research and guiding researchers towards solutions that can be applied in a variety of external environments. Finally, the paper discusses potential future advances that could lead to the next generation of solutions against DF.

To conduct a thorough and systematic investigation, we followed established principles and procedures for systematic reviews. The proposed SLR process is illustrated in Fig. 1 and includes four essential steps: Formulate research questions, establish a systematic review process, review and evaluate literature, as well as data analysis and presentation of results.

3.1. Research questions

The formulation of research questions plays a vital role in determining the success of any systematic literature review (SLR). These questions are the general framework of the review as they determine how the available studies are selected, appraised, and summarised. Therefore, in the context of the dynamically evolving and highly competitive area of DeepFake detection, it is becoming crucial that the review to focus on the core questions that need to be addressed.

As for the questions of this review, we have posed specific questions that aim to identify what has been done in the DeepFake detection field, how well the existing solutions work, and whether the approaches are uniform across different domains. These questions are intended to help

Table 1
Related systematic reviews conducted between 2022–2024.

Paper	Reviewed	Published	Advantages	Limitations
Rana et al. (2022)	2018 Jan–2020 Dec	2022 Feb	Discussed Widely used DDT, features, and datasets utilised	There is no discussion about strengths and weakness, and generalisability of the DF detection models
Stroebel et al. (2023)	2021-Aug 2022	2023 Mar	Discussed current DDT and proposed hybrid model (CNN, DNN, LTSM)and ensemble and multi-attentional as best architectures and uniform rating system to validate these techniques	There is no discussion about strengths and weakness, and generalisability of the DF detection models
Sharma et al. (2024)	2017–2023	2024 Aug	Evaluates DeepFake detection methods by discussing manipulations, optimisations, and enhancements and proposed 6 prominent DDT: XceptionNet, ResNet-50, VGG16, Capsule Networks, 3DCNN, EfficientNet-V2, proposed for future research	There is no discussion about strengths and weakness, and generalisability of the DF detection models
Whittaker, Mulcahy, Letheren, Kietzmann, and Russell-Bennett (2023)	2017 June 2021	2023 May	Establishing firmer conceptual understanding of DF and implications for innovation	There is no discussion about strengths and weakness, and generalisability of the DF detection models
Heidari, Jafari Navimipour, Dag, and Unal (2024)	2021–2020	2023 Nov	How DF are generated, identified, latest developments, areas requiring more investigation	There is no discussion about generalisability of the DF detection models, a brief discussion about challenges and advantages of detection models
Misirlis and Munawar (2023)	2018–2019	2023	Risks and threats of DF	Generalisation is not discussed
Vasist and Krishnan (2022)	Nov 2021–Apr 2022	2022 Aug	Gaps in DF definitions, theoretical foundations,	generalisation is not discussed
Passos et al. (2024)	2018–2024	2023 Oct	Current deep learning based DDT& future directions for further studies	There is no detailed discussion about generalisability of the DF detection models

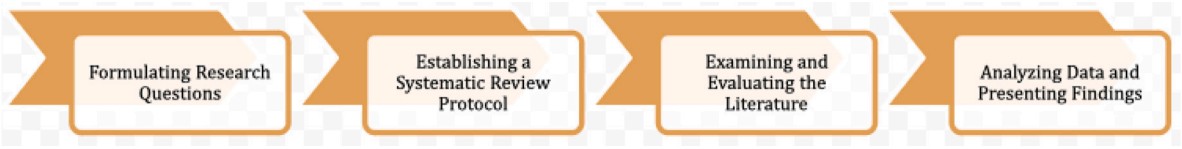

Fig. 1. Systematic literature review process.

gain a broad perspective on the subject and to assess the effectiveness of the currently used techniques. In this way, we want to provide recommendations for future advances that may benefit the development of more effective and comprehensive systems for DeepFake detection.

Table 2 provides a list of research questions aiming to cover every aspect of DeepFake detection, including methodological approaches and practical applications. These questions confirm the extent of the review and the crucial aspects of the objective of detection strategies.

3.2. Establishing a systematic review protocol

There has been a significant increase in interest in DeepFake technology in recent times, which has resulted in the rise of a large number of research papers. This comprehensive review aims to compile the most recent and innovative work published between January 2018 and February 2024. The primary goal of this review is to identify the latest trends in DeepFake creation and detection, which can provide valuable insights for future researchers. The review objectives have been defined and the appropriate search terms and publication selection criteria have been identified, as discussed in the following sections.

3.2.1. Data sources

This systematic review was conducted using publications from three databases of scientific articles covering the period from January 2018 to February 2024: Google Scholar (accessed October 2023), IEEE Xplore (accessed October 2023), ACM Digital Library (October 2023). The inclusion of Google Scholar was based on the recognition in the research literature that it is essential to ensure comprehensive and efficient coverage.

The distribution of publication types in the reviewed literature highlights the predominance of journal articles over conference papers. As shown in Fig. 2, 54.2%, of the publications are journal articles, while 45.8% are conference articles. This distribution suggests a substantial focus on DeepFake detection research within the academic community, with a significant portion of findings being disseminated through peer-reviewed journals. The balance between journal and conference publications also indicates the dynamic nature of the field, where ongoing developments are frequently shared at conferences before being published in more extensive journal articles.

3.2.2. Search strategy

In order to conduct a thorough search, various sources were consulted instead of relying on one or two sources. Due to the wide range

Table 2
Research questions and their purposes.

	Research Question	Purpose
RQ1	Which state-of-the-art methods are often used in the current literature on detecting DeepFake videos?	To identify and evaluate the most effective techniques currently being employed for detecting DeepFake videos.
RQ2	What are the main issues and constraints experienced by researchers and technologists when enhancing videos DeepFake detection algorithms?	To understand the challenges and limitations faced in improving the accuracy and efficiency of video DeepFake detection methods.
RQ3	What are the data sources utilised to evaluate the video DeepFake detection techniques?	To analyse the datasets and benchmarks used to test and validate the performance of detection algorithms.
RQ4	Can video DeepFake detection be generalised?	To assess the adaptability and robustness of detection algorithms across different datasets and real-world conditions.

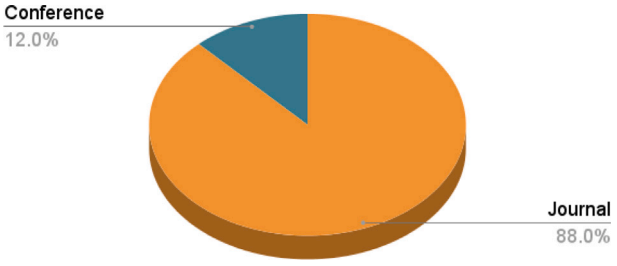

Fig. 2. Type of the documents included in SLR.

of digital repositories, the search was limited to three main sources, namely Google Scholar, ACM Digital Library, and IEEE Xplore Digital Library. These were chosen because they contain extensive collections on digital forensics and machine learning.

The next step is to choose the appropriate search terms. To gather as much relevant material as possible, a comprehensive approach was taken. The objective was to ensure that N significant research was overlooked. Therefore, a wide range of search terms were used to minimise potential biases, effectively combining keywords using Boolean search terms like 'AND' and 'OR'. The initial search setup was broad, incorporating combinations such as: (DeepFake OR FaceSwap OR Video manipulation OR Fake face/image/video) AND (detection OR detect) OR (Facial Manipulation OR Digital Media Forensics).

Throughout the search process, the terms used to investigate DeepFake research were refined to better align with emerging trends and specific areas of interest. Initially, the search terms included "DeepFake detection AND (digital forensics OR detection algorithms OR detection methods) AND deep learning". As the search progressed, more specific terms were added, such as "DeepFake detection AND (digital forensics OR detection algorithms OR detection methods) AND facial DF learning OR deep learning OR video". Finally, the search concluded with a highly targeted search string:

"DeepFake AND (generalisation in video forensics OR detection algorithms OR detection methods) AND Deep learning or Machine learning",

which focused specifically on studies that address the generalisation capabilities within deep learning frameworks for DeepFake detection. To ensure that the reviewed research was the most recent and innovative and to capture significant advancements in the field, the search was limited to the period of January 2018 to February 2024. This strategic approach allowed a thorough analysis of the current state of DeepFake detection, providing a solid foundation to comprehend the progress and ongoing challenges in the field.

3.3. Examining and evaluating the literature

3.3.1. Inclusion criteria

The following criteria were applied to include publications in the selected collection.

- Publications from 2018 until February 2024
- Conference proceedings and journals
- Research focused on DeepFake detection techniques

3.3.2. Exclusion criteria

A set of exclusion criteria is also established to omit studies that may not be relevant to this review.

- Studies that did not have precise descriptions of the machine learning or deep learning models used in their research.
- Review and survey papers: A discussion on the review papers and survey papers is included in the related research section but excluded from the systematic literature review document collection to maintain clear focus on primary studies and original research contributions to DeepFake detection.
- Research involving machine learning or deep learning solutions to problems not related to DeepFake detection.
- Technical reports, Articles without a full paper
- Master and PhD research thesis.

3.3.3. Study selection

Fig. 3 shows the various phases of the document selection process in this systematic review, developed according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Page et al., 2021). In the early stages of the literature review, a wide range of articles were obtained from various academic databases. 200 articles were retrieved by a Google Scholar search. Other sources included the ACM and IEEE Digital Libraries, each of which retrieved 100 articles, for a total of 400 articles. These articles were reviewed with respect to the main algorithm used, evaluation metrics, and the strengths and weaknesses discussed in the paper. The review identified and removed 36 duplicate entries, reducing the total to 364. These 256 articles were excluded from the review because they did not strictly focus on DeepFake detection. Hence, the selection was made into 108 articles that were directly related to the subject matter. This careful selection ensured that the literature used was relevant, forming a strong foundation for a detailed analysis.

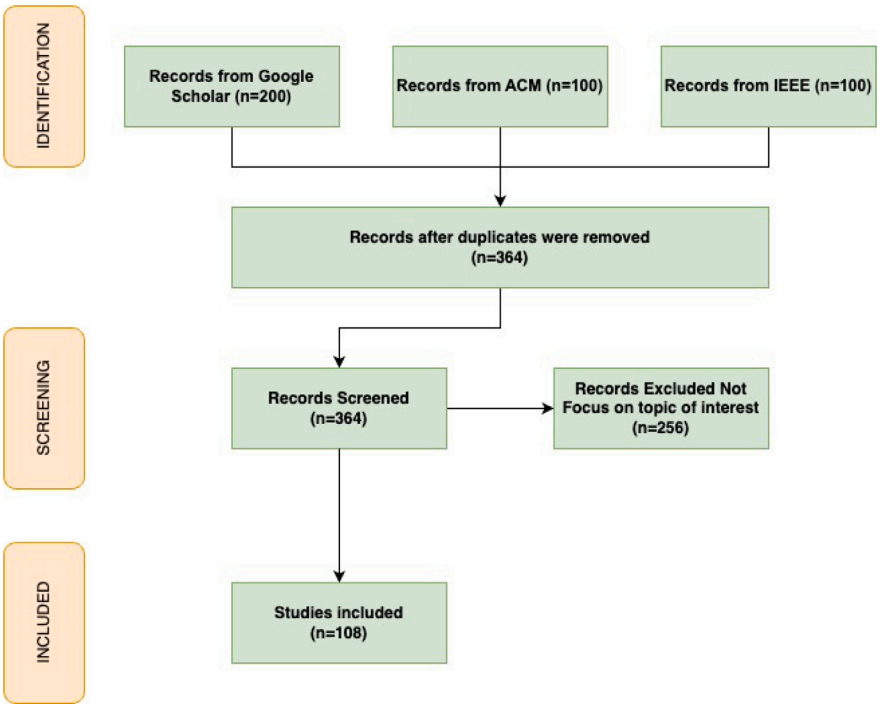

Fig. 3. PRISMA (Page et al., 2021).

4. Background on video deepFake detection techniques

This section provides an overview of the key techniques and approaches used in video DeepFake detection, specially focusing on manipulations within video DeepFakes. Detection methods are organised into three categories: spatial methods (frame-level sensing), temporal methods (sequence-level sensing), and multimodal methods (cross-stream integration). While some methods focus on mitigating spatial biases in individual frames, others aim to smooth temporal artifacts across the entire video sequence. Further improvements are achieved by using various advanced architectures, such as transformers and hybrid frameworks.

4.1. Spatial techniques: Frame-level detection

Spatial techniques form the backbone of DeepFake detection, where individual frames are analysed to find some visual inconsistencies or artifacts introduced in the manipulation. These methods make use of static anomalies, such as blending errors, inconsistencies in texture, or unnatural lighting that usually arise due to the limitations of DeepFake generation models. CNNs may be used for spatial analysis since they can extract features at the pixel level effectively. CNN-based techniques detect abnormalities, such as unnatural textures, shading mismatches, or compression artifacts, with strong detection performances for frame-level anomalies (Chang, Wu, Yang, & Feng, 2020; Liu, Boongoen, & Iam-On, 2024; Liu, Zhu, Lu, Luo, & Zhao, 2021; Taeb & Chi, 2022).

For improved spatial detection, edge and texture analysis can find increased use when analysing manipulated images, helping to highlight the finer details. Methods such as Local Binary Patterns (LBP) investigate local textures to identify oversmoothing or slight textural anomalies properties associated with manipulated face areas. These techniques are superior for the detection of anomalies that are not visible to the human eye (Abdullah & Ali, 2023; Karanwal & Diwakar, 2023; Sedaghatjoo, Hosseinzadeh, & Bigham, 2024). In a similar vein, edge detection algorithms also known as Sobel and Canny filters are utilised to detect sharp discrepancies or transitions at the boundary between a tampered region and an untampered region. It was discovered that these transitions are strong indicators of frames that have undergone DeepFake manipulation, especially for the blend artifacts frames (Chang et al., 2020; Khalil & Maged, 2021; Siegel, Kraetzer, Seidlitz, & Dittmann, 2021).

The second most important cue in spatial analysis is to notice discrepancies in colour and lighting. In distorted regions, manipulated frames usually show unnatural illumination gradients, synthetic shadows, or some reflections that are not consistent. Real and manipulated datasets train models to recognise these artificial effects while avoiding the effects of natural lighting conditions (Bondi et al., 2020; Coccomini, Caldelli, Falchi, & Gennaro, 2023; Siegel et al., 2021). DeepFake content, on the other hand, is often detected near manipulated boundaries by shading mismatches and significant differences in intensities of the pixels adjacent to the boundaries.

Another significant inconsistency identified by spatial methods is misaligned facial features. When face-swapping algorithms do use manipulation content but are unable to seamlessly integrate it into the original frame, alignments of eyes, lips, or facial proportions often experience errors. These spatial inconsistencies are particularly strong in early generation DeepFakes (Chang et al., 2020; Taeb & Chi, 2022). Furthermore, unrealistic textures, such as too smooth skin or checkerboard artifacts, are common artifacts generated by generative models, especially GANs, when generating DeepFake images (Karanwal & Diwakar, 2023; Liu et al., 2021).

Although spatial methods are highly successful in detecting frame-level anomalies, they are less effective against high-def generated DeepFakes. As generation techniques become more sophisticated, artifacts in individual frames become increasingly subtle, making it more challenging for spatial methods to detect manipulations. Furthermore, these methods do not account for motion inconsistencies or sequence-level anomalies, necessitating the use of temporal or multimodal approaches to achieve exhaustive detection (Chang et al., 2020; Myvizhi & Pamila, 2022; Zhang, Wu, Li, Zhu, & Sheng, 2022). Some of these problems can be solved by including temporal analysis with spatial techniques to identify dynamic inconsistencies throughout video sequences,leading to a firmer detection framework (Lewis et al., 2020; Myvizhi & Pamila, 2022).

4.2. Temporal techniques: Sequence-level analysis

Temporal methods are aimed at discovering inconsistencies in motion or continuity between video frames, and generally rely on the temporal ordering of video streams to analyse DeepFake distortions. Unlike spatial methods that focus on specific frames, temporal techniques leverage the continuity and interrelation between adjacent frames to identify anomalies that static analysis may overlook, making them especially powerful in detecting artificial motion, lip synchronisation discrepancies, or sudden frame transitions, phenomena which naturally occur from the constraints of DeepFake generation models.

Motion dynamics is one of the most common methods used for temporal analysis. Modelling temporal dependencies in video sequences is primarily achieved using RNNs, particularly LSTMs, which excel at capturing subtle motion patterns and identifying abnormal patterns such as jittery head movements or unexpected frame-to-frame transitions. Likewise, optical flow-based methods track the displacement of facial features and objects between frames, enabling the detection of inconsistencies in motion patterns. This approach is particularly valuable for attack detection, as it draws inspiration from anomaly detection techniques by identifying poor frame alignments followed by interpolation (Myvizhi & Pamila, 2022; Zhang, Wu, et al., 2022).

Another essential aspect of sequence-level analysis is temporal consistency. DeepFake videos usually have some inconsistencies in expression or face movement such as blinking of the eye, tilt of head, movement of lip, etc. Temporal coherence models are trained to identify such subtle differences between frames that are challenging to achieve in generated content naturally. The temporal convolutional networks (TCN) and hybrid CNN-RNN architectures have demonstrated great potential to discover such discrepancies (Lewis et al., 2020; Wang & Dantcheva, 2020).

One of the important aspects of temporal analysis is audio–visual synchronisation, as DeepFake videos do not always synchronise speech and lip movements accurately. Cross-modal evaluation models to assess the timeliness and coherence of audio and visual streams while measuring the differences between them in terms of synchronisation. This method is especially useful for spotting discrepancies, such as lagging lip movements or non-matching speech in manipulated audio–visual content (Lomnitz et al., 2020; Suratkar, Kazi, et al., 2020).

Temporal techniques can capture motion dynamics and sequence-level inconsistencies, but they are also limited. As these advanced models can produce high-quality DeepFakes which can also minimise these artifacts, thus making detection more difficult. For instance, algorithmically temporal methods usually need far more numerical branches compared to parallel techniques due to processing lumps of frames that have not yet been processed independently of each other. In spite of these issues, such temporal techniques are often still necessary components in DeepFake video detection pipelines, and most thorough detection frameworks combine spatial or other multimodal methods with a temporal component (Taeb & Chi, 2022; Zhang, Wu, et al., 2022).

4.3. Multimodal techniques: Cross-stream analysis

Multimodal approaches combine diverse modalities, including visual, audio, and physiological signals, improving DeepFake detection accuracy and robustness. This method characterises inter-modal relationships, identifying discrepancies that may be missed in uni-modal processing. These methods are particularly useful for solving complex video DeepFakes manipulations that can break cross-stream correlation (between modalities, such as audio or visual).

Multimodal techniques have been extensively used in audio–visual tasks. This approach addresses detection of temporal misalignment between speech and lip movements, which is another potential bug in DeepFake videos. These approaches use discrepancies between audio and visual data streams to detect differences in timing, facial expression, and behaviour. Additionally, advanced models such as audio–visual transformers and cross-modal attention mechanisms have also shown remarkable performance in recognising lip-sync errors and are very effective in identifying audio–visual mismatches (Lomnitz et al., 2020; Suratkar, Kazi, et al., 2020).

Physiological signal analysis is another key component of multimodal detection. However, DeepFake videos fail to capture subtle biometric cues that unconsciously exist in original videos, like heart rate change, skin tone change, and micro-expressions. Remote photoplethysmography (rPPG) and ISP are examples of the approach that study facial colour variations to estimate the heart rate, making it possible to assess the manipulation. It would be quite challenging to convincingly synthesise these physiological signals, which makes them a good indicator of whether tampering occurred (Coccomini et al., 2023; Taeb & Chi, 2022).

Hybrid feature fusion often merges spatial, temporal, and cross-modal features to offer a comprehensive detection framework. Most action recognition models consist of a spatial CNN to extract spatial features for single frames and an RNN or other temporal model to model dependencies across frames. In combination with audio features, these approaches yield strong multimodal models that can detect complex anomalies cross-modally. By combining the complementary interactions among visual, audio, and physiological data, multimodal transformers and CNN-RNN hybrid frameworks have achieved considerable advancements in detection accuracy (Lomnitz et al., 2020; Muppalla, Jia, & Lyu, 2023).

Although multimodal methods are powerful, they are also computationally complex, often requiring both streams to be temporally aligned for implementation. However, they cannot detect discrepancies across different modalities and thus become ineffective against high-quality DeepFakes. Nevertheless, the combination of spatial, temporal and multimodal approaches enhances the robustness of the detection systems, making them resilient to increasingly complex manipulations (Muppalla et al., 2023; Suratkar, Kazi, et al., 2020).

4.4. Hybrid frameworks and advanced architectures

As DeepFake technologies continue to evolve, researchers are attempting to combat them by implementing more advanced architectures and hybrid frameworks. These methods enable fusion of multiple detection paradigms for cooperative analysis of spatio-temporal and multimodal data streams. The section reviews cutting-edge approaches that involve hybrid models and sophisticated architectures that improve precision in detection and robustness to attacks.

4.4.1. Hybrid frameworks

Hybrid approaches exploit the strengths of multiple architectures to improve DeepFake detection. These models utilise complementary approaches to achieve strong accuracy in detecting spatio-temporal inconsistencies, which are essential for spotting manipulations in videos. Such frameworks are indeed good for frame-level spatial features and sequence-level temporal anomalies, which contributes to a global detection method for video.

A hybrid framework that fused CNNs and RNNs is presented in the literature. CNNs extract spatial features from each video frame, including textures and facial artifacts, while RNNs learn temporal dependencies from sequential frames. CNN-RNN hybrids, which combine these architectures, successfully catch spatio-temporal anomalies, such as unnatural lip-sync, abrupt transitions, and jerky movements. These frameworks react significantly better to the inconsistencies that static or sequential models alone could miss (Khormali & Yuan, 2022; Lomnitz et al., 2020).

An alternative approach which is widely adopted is a hybrid method by combining CNNs with transformers. As depicted in Fig. 4, dense neural networks (CNN) take care of the main feature extraction of the spatial patterns; whereas, transformers analyse the temporal patterns

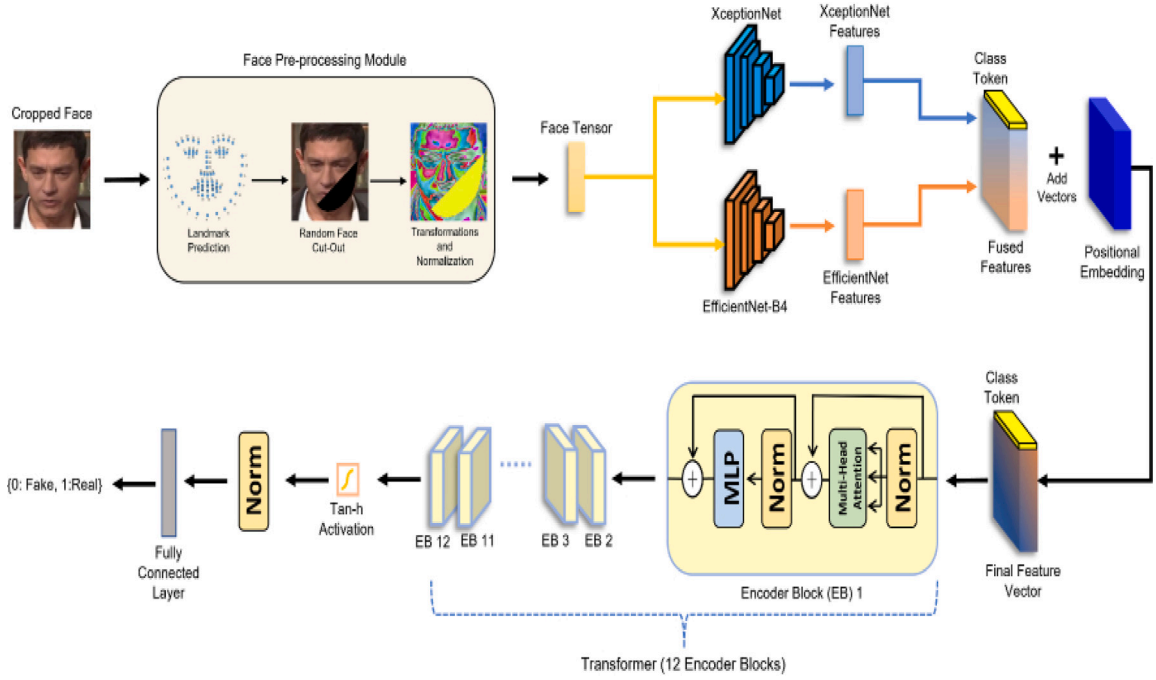

Fig. 4. Hybrid framework (Khan & Dang-Nguyen, 2022).

by means of the self-attention mechanism. This multi-grained union strongly learns both short-distance and long-distance dependencies in the video data. Facial landmark detection, normalisation and augmentation, All these preprocessing steps contribute to making these models even more robust. CNN architectures such as XceptionNet and EfficientNet-B4 extract features, which are then fused with positional embeddings and fed through a series of transformer encoder blocks. It is capable of reliable detection of advanced deceptions such as non-matching facial expressions, or slight speed inconsistencies (Bondi et al., 2020; Gani-yusufoglu, Ngô, Savov, Karaoglu, & Gevers, 2020).

Hybrid frameworks are not limited to CNNs and transformers alone. Alternate techniques, such as CNNs with optical flow processing, emphasise motion sequences and temporal continuity for smaller detection of DeepFake alterations. These frameworks successfully identify artifacts that can be caused by inconsistent movement between frames, ensuring increased accuracy (Bondi et al., 2020; Khormali & Yuan, 2021).

Although hybrid frameworks take advantage of billions of parameters from multiple architectures, they naturally incur higher computational costs. However, hybrid frameworks consistently outperform single-architecture models in terms of accuracy and generalisation on benchmark datasets (Gani-yusufoglu et al., 2020; Khormali & Yuan, 2021).

CNNs and transformers can provide spatial and temporal analysis as shown in Fig. 4, combined into the hybrid framework. In the front-end, a pre-processing module such as facial cropping and normalisation is followed by CNN-based spatial feature extraction. The transformer encoder blocks model captures temporal dependencies and long-range interactions among frames. When combined, these methods can successfully overcome existing spatio-temporal inconsistencies, laying the foundation for a unified method of DeepFake detection (Khormali & Yuan, 2022; Lomnitz et al., 2020).

4.4.2. Advanced architectures

DeepFake detection techniques that use complex architectures are cutting-edge advances capable of handling ever more sophisticated manipulations. Deep learning based architectures can extract complex patterns and anomalies in manipulated content which can be used for detection purpose. The major approaches are GANs, transformers, capsule networks, disentangled representation learning, unsupervised domain adaptation, and improved feature extraction.

Generative adversarial networks (GANs)

In the context of DeepFake detection, GANs play a dual role. Originally proposed to generate realistic synthetic data, GANs have recently been modified to detect the presence of manipulations — this is achieved by leveraging their adversarial training process. Generative Adversarial Network detection models comprise a generator and a discriminator, with the generator generating the most realistic DeepFakes possible and the discriminator learning to differentiate between genuine and fake content. This process running iteratively in an adversarial setting further improves the discriminator to identify finer anomalies.

The GAN structure is depicted in Fig. 5 as a multi-generator and multi-discriminator framework (Coccomini et al., 2023; Ding et al., 2021). This multi-task framework trains three generators (H1, H2 and H3) to generate DeepFake photos across different object classes under real scenario and logically inconsistent outputs. These condensed and deceptive images force the discriminators to continually reconvene, improving their potential to find minor adjustments in the lot. This design pushes towards the generation of more and more advanced DeepFakes, forcing the discriminator to learn and improve detection techniques (Ding et al., 2021; Giudice, Guarnera, & Battiato, 2021; Gong, Kumar, Goh, Ye, & Chi, 2021).

The architecture uses six components, all designated as J1 to J6—to serve as discriminators for any image based on the layer of the network it goes through. Specifically, odd-numbered discriminators (J1, J3, and J5) are trained to assess real images and even-numbered ones (J2, J4, and J6) are supposed to discriminate fake images. The staged evaluation provides insight into subtle inconsistencies, and thus guarantees that both original and re-crafted DeepFakes are examined in detail. The model enforces multiple layers of scrutiny, making it more powerful in detecting sophisticated fakes more effectively (Coccomini et al., 2023; Ding et al., 2021; Giudice et al., 2021).

One of the most powerful aspects of GAN-based models is that they are amenable to adversarial training. This creates a feedback loop where the discriminator must consistently improve its ability

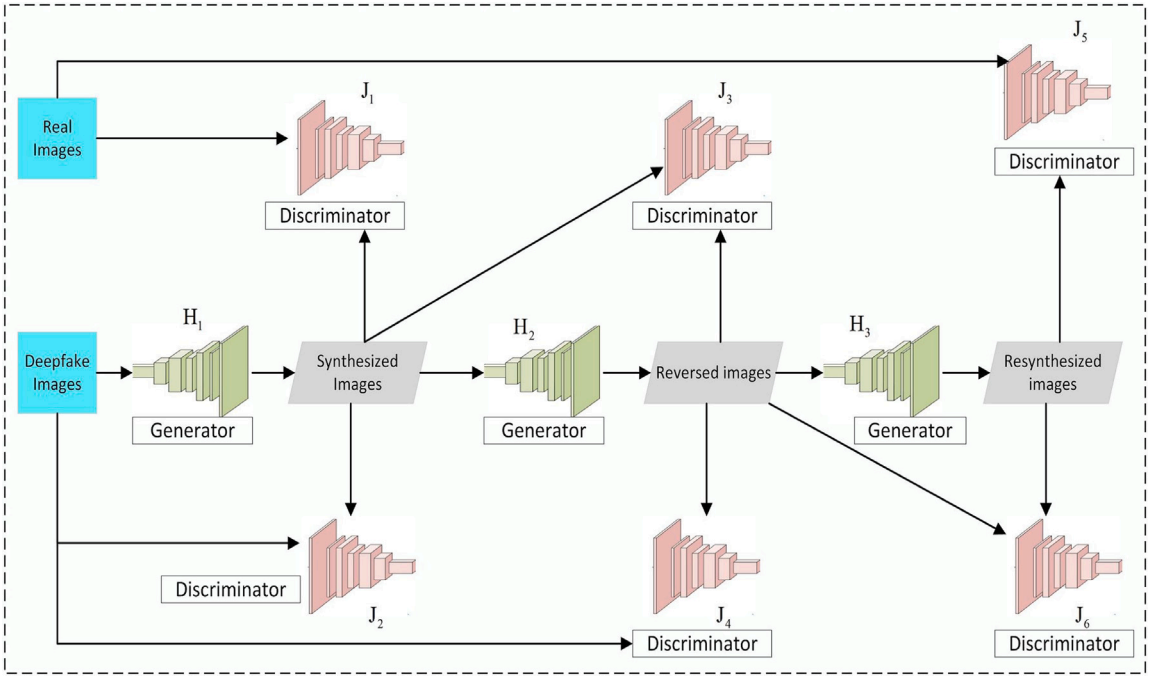

Fig. 5. GAN framework (in Abbas and Taeihagh (2024) based on Ding et al. (2021)).

to differentiate between real and generated samples, as it is pitted against progressively better generators. Because generation techniques are evolving, adversarial nature keeps GANs effective against sophisticated DeepFake strategies. This competitive nature generates strong and generalisable detection features (Coccomini et al., 2023; Giudice et al., 2021; Gong et al., 2021).

Transformers

The introduction of Transformers, specifically ViTs, led to an inspiring and robust architecture already being used widely in DeepFake detection. The study proposes an alternative approach based on vision transformer (ViTs) for modelling video + depth sequences, a decision motivated by the use of self-attention mechanisms for modelling spatial and temporal dependencies. Traditional CNNs have a limited scope, being confined to what they can learn by focusing on localised features, while transformers process entire sequences, and therefore suit themselves to long-range dependencies.

ViTs segment video frames into distinct associated patches and consider each patch as a token in a sequence. This method then allows the sensitive detection of minor abnormalities such as waddling or lip-sync discrepancies. Through the addition of positional encodings, transformers capture temporal order, making them more suitable for detecting temporal anomalies, such as unnatural blinking or sudden change of appearances (Li et al., 2023; Vaswani, 2017).

Multimodal transformers generalise this as they can deal with several data streams — for example, visual and audio input. To improve detection performance, these models analyse the relations of cross-modal elements to identify the synchronisation errors of speech and lip movements (Muppalla et al., 2023; Wang et al., 2022). This has made them critical for complex manipulations, despite their computational heavyweight nature.

Capsule networks

A unique advantage of capsule networks is the information they retain about spatial hierarchies within a frame of a video. Whereas CNNs tend to lose spatial relationships through pooling, capsule networks preserve the relative positions of features, such as facial landmarks. This property allows capsule networks to identify detailed alterations to textures and structures in facial components, even if the DeepFakes are low quality or compressed (Choudhary, Saurav, Saini, & Singh, 2023; Khalil, Youssef, & Saleh, 2021).

Capsule networks are particularly useful for detecting DeepFakes with subtle distortions, such as changes in skin texture or facial geometry. Their ability to capture part-whole relationships makes them a powerful tool for addressing complex manipulations (Stanciu & Ionescu, 2022).

Disentangled representation learning

Disentangled representation learning involves extracting domain-invariant features so that models can distinguish manipulated content from genuine content. This enables to train detection models with greater test performance by learning representations not specific to certain manipulation techniques. The disentangled features are more robust against various datasets and manipulation types, which serve as a good candidate for DeepFake detection under unseen conditions.

This method is particularly valuable in scenarios where DeepFakes are generated using novel or unknown techniques. By focusing on fundamental differences between real and fake content, disentangled representation learning ensures consistent detection performance (Ding et al., 2021; Jia, Cheng, Lu, & Zhang, 2022).

Unsupervised domain adaptation

The purpose of an unsupervised domain adaptation is to align the feature distributions from the training dataset with the feature distributions of the testing dataset, so that models can generalise from different environments. This is important in practice, where testing data typically differ quite a bit from training data.

Methods such as domain adversarial neural networks (DANNs) are typically employed to align the two. This allows unsupervised domain adaptation to generalise well to manipulations not seen in the training data, as they tend to minimise the gap between the source and target domains. This is significant when combating the DeepFake techniques which aim to escape detection by utilising biases in models (Bondi et al., 2020; Chen & Tan, 2021).

Enhanced feature extraction

Some advanced feature extraction methods seek to analyse particular features of video content to detect hidden anomalies. Fourier transforms and spectral analysis are among methods that focus on anomalies in frequency elements and texture characteristics. These approaches outperform on strong manipulations such as high-quality DeepFakes (Giudice et al., 2021; Khormali & Yuan, 2021).

For example, high-frequency details can escape conventional types of detections, while spectral analysis isolates them. Since DeepFake synthesis introduces unique signatures, advanced feature extraction methods integrated into detection frameworks provide robustness (Giudice et al., 2021; Zhao et al., 2021).

5. Discussion

The selected 108 articles were reviewed to evaluate the applied DDTs, the data sets used, their strengths and weaknesses with respect to the generalisability of the proposed models across different datasets. The analysis also included examining the authors' perspectives on why the proposed methods struggle to generalise effectively. Furthermore, insights on how generalisability might be achieved using the proposed algorithms were gathered. These details have been systematically analysed and summarised in Table 3 to provide a comprehensive overview of the current state of the research in this domain. The heatmap shown in Fig. 6 provides an overview of the different types of DDTs used and the categories of the models applied. The categorisation is based on the algorithms applied in 108 selected papers which will be discussed in Section 5.1. This comprehensive evaluation aims to provide a valuable guide for researchers by highlighting the challenges and potential solutions in developing robust and adaptable DeepFake detection algorithms.

Numerous enhancements have been proposed in the reviewed articles, showcasing innovative approaches to overcome challenges in detecting DeepFakes. Hybrid and multimodal models effectively blend the best aspects of various methods to improve accuracy and robustness. GAN-based models have emerged as a double-edged sword, playing a pivotal role in both DeepFake generation and detection. Their ability to generate highly realistic content necessitates the development of more sophisticated detection mechanisms to identify and exploit subtle artifacts in the generated media.

Transformers have been used to handle sequential data and capture long-range dependencies in videos and have become increasingly popular for video DF analysis tasks. Many novel approaches, such as capsule networks or disentangled representations, have been introduced to enhance the re-usability and adaptability of both generation and detection models. Capsule networks aim to better capture spatial hierarchies and relationships in data, while disentangled representations facilitate the isolation of specific features, improving model interpretability and tranferability.

In the following sections, we have delved into the key research questions that are essential to understand the state-of-the-art methods often applied in current literature and their generalisability in detecting DF. We have analysed various detection techniques identified through our systematic literature review, evaluating their effectiveness and the datasets they utilised. In addition, we have examined recent advancements in detection technologies, highlighted current challenges and limitations, and discussed potential future trends that could influence the development of more generalisable DeepFake detection methods.

5.1. RQ1: What state-of-the-art methods are often used in the current literature on detecting DeepFake videos?

Convolutional Neural Networks (CNNs) have gained significant prominence in recent years due to their ability to capture spatial–temporal dependencies from adjacent frames in sequences of frames in a video, such as small inconsistencies in facial details or textures that alter the semantics of objects (Bondi et al., 2020; Coccomini et al., 2022; Mcuba, Singh, Ikuesan, & Venter, 2023). However, detecting temporal manipulations in videos requires models capable of reasoning about changes over time. This has led to recent advancements in the use of hybrid models that combine; Recurrent Neural Networks (RNNs) (Khormali & Yuan, 2022) and Vision Transformers (ViTs) (Saikia, Dholaria, Yadav, Patel, & Roy, 2022; Zhang, Zhao, & Li, 2020) for video-based DeepFake detection. The combination of CNNs with RNNs and ViTs enables models to overcome spatial–temporal gaps, ultimately improving detection performance.

CNNs, RNNs, ViTs and Generative Adversarial Networks (GANs) are designed to create DF, but they can also be utilised in adversarial training to improve the robustness of detection (Ding et al., 2021; Jung, Kim, & Kim, 2020a). Some of the most commonly used DDT models in the reviewed documents are discussed below.

Hybrid

One of the noteworthy hybrid approaches involves using CNNs together with RNNs, which have proven to work wonders in video based detection tasks, as we discussed in 4.4.1. Five prominent studies investigating CNN-RNN hybrids were Chintha et al. (2020), Cho et al. (2023), Jiang et al. (2021), Lewis et al. (2020), Yadav, Bommareddy, and Vishwakarma (2022) as depicted in Fig. 7. Studies consistently show that hybrid CNN-RNN models outperform standalone CNN or RNN architectures, especially in tasks that involve spatial and temporal features. CNNs are well known for capturing spatial features, making them an excellent choice for detecting anomalies in an individual frame. Since DeepFake images alter individual frames of the generated videos, RNNs (particularly Long Short-Term Memory (LSTM) networks) are combined with CNNs to analyse video frame sequences, detecting temporal inconsistencies in frames such as jerky movements or scene switch markers (Afchar, Nozick, Yamagishi, & Echizen, 2018; Ismail, Elpeltagy, Zaki, ElDahshan, A, 2021b; Saikia et al., 2022). Hence, CNN-RNN hybrid architectures can capture higher-level spatio-temporal interactions in video sequences, such as small temporal deformations such as facial expressions shifts or lip-sync errors that occur across multiple frames. These models have achieved exceptionally high performance with accuracy as high as 98% and precision rates as high as 81%.

However, more advanced techniques are needed to effectively identify manipulations across multiple frames, such as, subtle facial changes or unnatural transitions that occur over time. For example, CNN-RNN hybrid networks combined with auto-encoders, have been proposed as a promising approach to address these challenges (Agarwal et al., 2020; Kirn et al., 2022; Wang & Dantcheva, 2020). In addition, CNN-RNN hybrid networks have been extended to other hybrid approaches, such as fusing audio–visual data and utilising person-based techniques. These models incorporate multiple modalities in a fused form, enhancing their detection capabilities. Audio–visual attention models, for instance, detect discrepancies between audio and visual data, such as discordance between speech and lip movements (Ge et al., 2022; Muppalla et al., 2023). Additionally, identity-referenced models focus on assessing the compatibility of observed facial attributes with previously known identity information to improve detection accuracy (Lewis et al., 2020; Wang, Jiang, Jin & Cui, 2022). Since hybrid fakes combine visual, auditory, and identity-referenced, it becomes fairly impossible for DeepFake developers to manipulate all aspects of the entire media file simultaneously.

Ensemble

An emerging trend in the field of DeepFake detection is the adoption of ensemble methods to address challenges encountered during the learning process of deep learning algorithms. Ensemble models involve training and combining multiple baseline models or similar models to enhance the robustness, realiability and accuracy of detection systems (Mohammed & Kora, 2023). The most widely applied ensemble techniques include averaging, bagging, stacking, random forest, and boosting. For example, the fusion of models like YOLO-CNN-XGBoost (Rana & Sung, 2020) and the integration of models such as

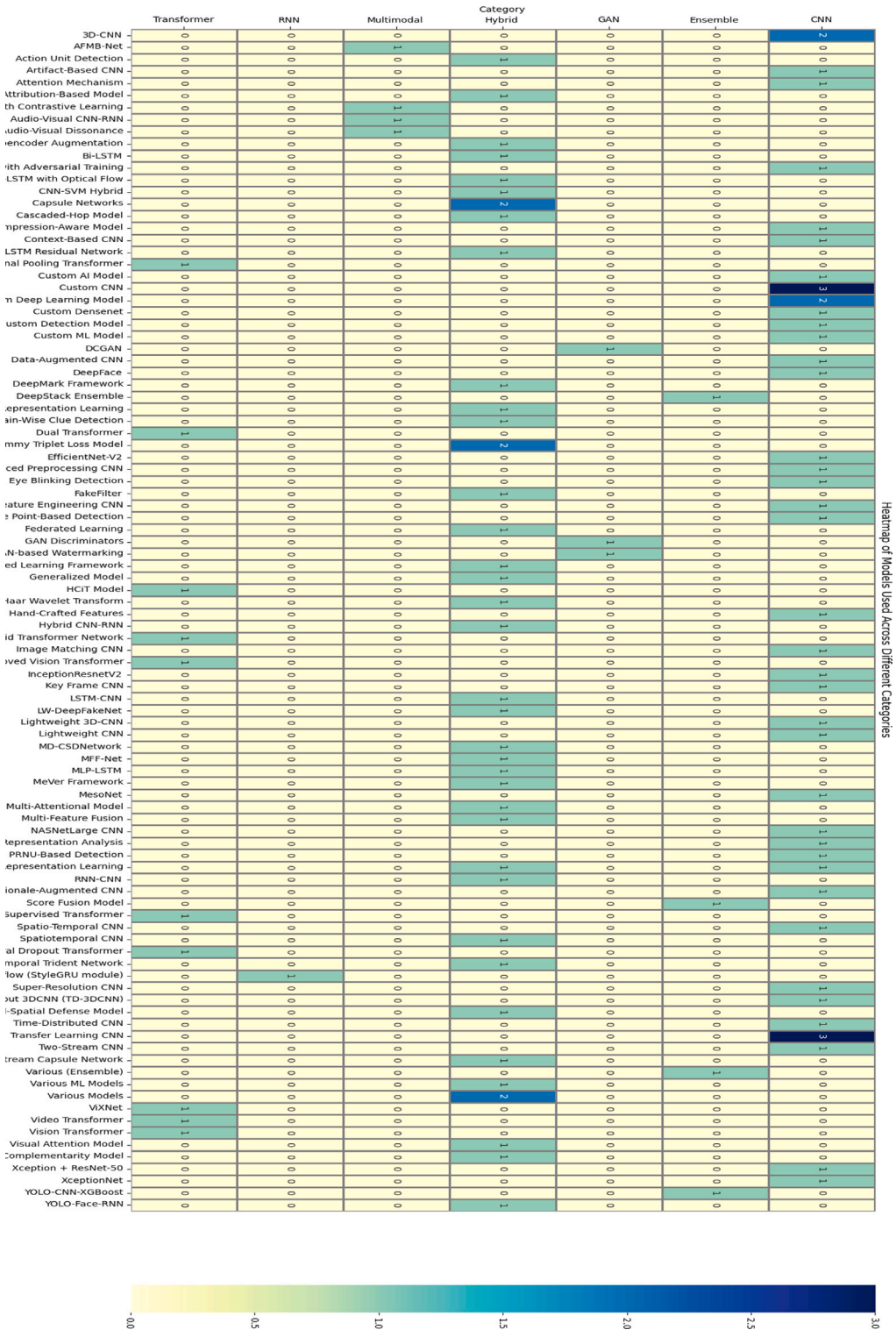

Fig. 6. Heatmap of models used across different categories.

XceptionNet, LSTM, and CNN (Mcuba et al., 2023) demonstrate how ensemble approaches can effectively combine strengths of different models. Ensemble methods address the limitation of single-model approaches in DeepFake detection by combining models specialised in detecting various types of DeepFake manipulations.

However, the studies indicate that ensemble methods can achieve high accuracy (ranging from 88% to 97%) depending on the models used in the ensemble models, and the applied dataset (Lewis et al., 2020; Mcuba et al., 2023; Rana & Sung, 2020). Ensemble models combine several detection mechanisms to capture discrepancies in both spatial and temporal features across various types of DF, leading to improved performance. In addition, they are especially helpful in circumstances where one detection method may be insufficiently performing on its own. Efficiently designed ensemble models have the ability to recognise a broader range of potential DeepFake alterations and unseen manipulations.

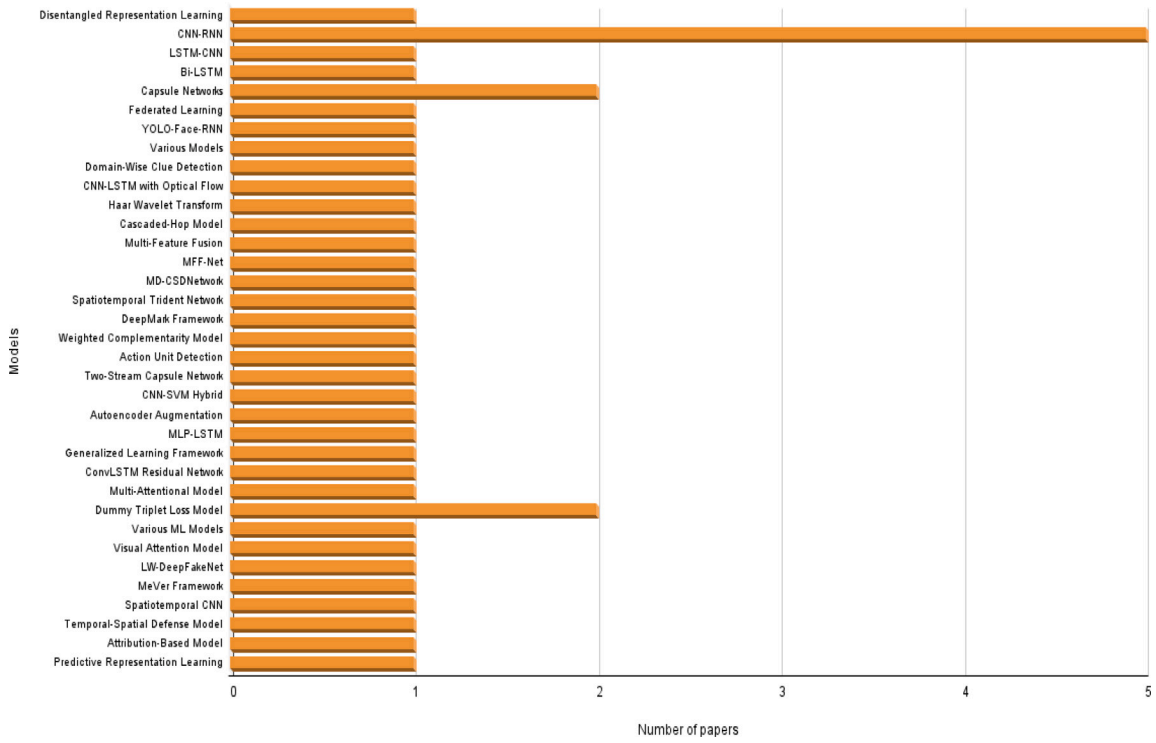

Fig. 7. Variety of hybrid models used by researchers.

CNN-based techniques

CNNs are crucial for detecting spatial abnormalities in frames of an image or video, recognising the contortion in facial components, skin patterns, and suboptimal illumination conditions triggered by facial spoofing-attackers (Bondi et al., 2020; Coccomini et al., 2022; Mcuba et al., 2023). Additionally, with temporal information support, these models help to analyse spatial and temporal features of videos (Coccomini et al., 2022; Saikia et al., 2022).

CNN-based methods, such as 3D-CNNs, have proven to be effective in capturing motion-related artifacts across multiple frames. As illustrated in Fig. 8, numerous studies have demonstrated the effectiveness of 3D-CNNs for DeepFake detection. The 3D-CNNs are capable of learning fine-grained details from interpolative frames, particularly in detecting changes in facial expression and unrealistic interchanges, as emphasised in Bondi et al. (2020), Liu et al. (2021), Saikia et al. (2022). The utilisation of other CNN variations, besides 3D-CNNs, such as time-distributed CNN, and lightweight 3D-CNN, are also being explored. These models are particularly useful for handling large-scale datasets where computational efficiency is crucial. The time-distributed CNNs, process each video frame independently while preserving temporal connections. These models are ideal candidates for detecting both spatial and temporal inconsistencies in video frame sequences (Liu et al., 2021; Zhang et al., 2020). Furthermore, lightweight 3D-CNNs offer competitive computational costs compared to traditional 3D-CNNs, making them well-suited for real-time detection tasks. Generally, CNN-based methods achieve an accuracy rate ranging from 77% to 88%, positioning them as strong options for DeepFake detection (Bondi et al., 2020; Liu et al., 2021; Wang, Cheng, Chow, & Nie, 2023).

Other variety of models

Although CNNs and hybrid approaches remain dominant in the field, other models, including multimodal have demonstrated significant potential for DeepFake detection. These models integrate data from various streams, such as audio and video to enhance detection capabilities. As illustrated in Fig. 9, Multi-modal visual and audio models have been studied in 4 research articles (Chen, Kumar, Nagarsheth, Sivaraman, & Khoury, 2020; Ge et al., 2022; Zhang, Lin, & Xu, 2024) showing their effectiveness in handling complex DeepFake manipulations.

Moreover, the Transformers are among the powerful tools for dealing with sequential data over time, making them highly effective for video-based detection. This review identifies five key papers that showcase the effectiveness of Vision Transformers (ViTs) in detecting DF across video frames by analysing both spatial and temporal features (Coccomini et al., 2022; Khan & Dai, 2021; Khormali & Yuan, 2022; Zhang et al., 2024, 2020). As transformers capable of learning long dependencies between frames, they can discover slight anomalies that might be overlooked in the frame-by-frame detection.

GANs serve a dual purpose in DF as we discussed in 4.4.2. This adversarial setup allows models to create new types of DeepFake manipulations but also encourages the use of GAN-based methods to detect DF (Ding et al., 2021; Hu, Wang, & Li, 2021; Jung, Kim, & Kim, 2020a). Specifically, these models achieve detection accuracies between 88% and 96%, showcasing their versatility and effectiveness on a variety of datasets and tasks (Ding et al., 2021; Hu et al., 2021; Jung, Kim, & Kim, 2020a).

Hybrid I3D ViViT models

An effective approach to improving generalisation in DFD is through the development of hybrid models. One promising combination is the integration of I3D (Inflated 3D ConvNet) and ViViT(Video Vision Transformers) architectures. I3D is highly effective at capturing spatial–temporal features in video data, enabling it to analyse motion and appearance across frames. On the other hand ViViT utilises its intrinsic transformer architecture to model long-range dependencies. Therefore, by integrating these strengths, this hybrid model could effectively detect manipulations across different video sequences. Although this hybrid combination is not explored in the selected literature, this integration has the potential to offer significant improvements over current DeepFake detection models by improving over current DeepFake detection models, particularly to improve the generalisability of existing detection techniques across various datasets and DeepFake types (Coccomini et al., 2022; Khormali & Yuan, 2022; Zhang et al., 2020)

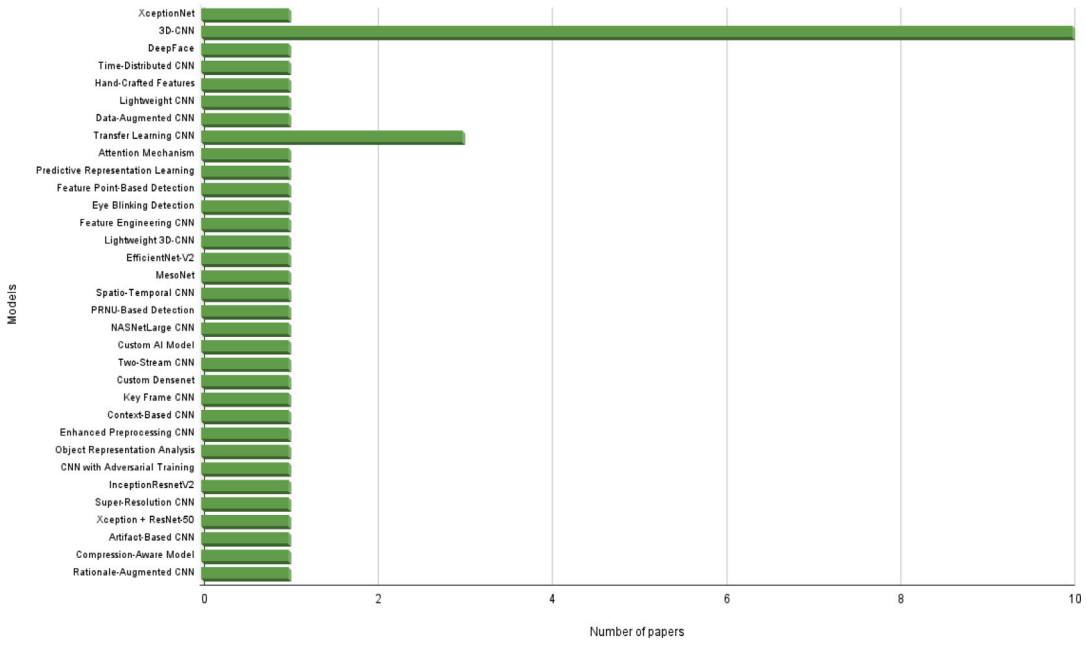

Fig. 8. Variety of CNN models.

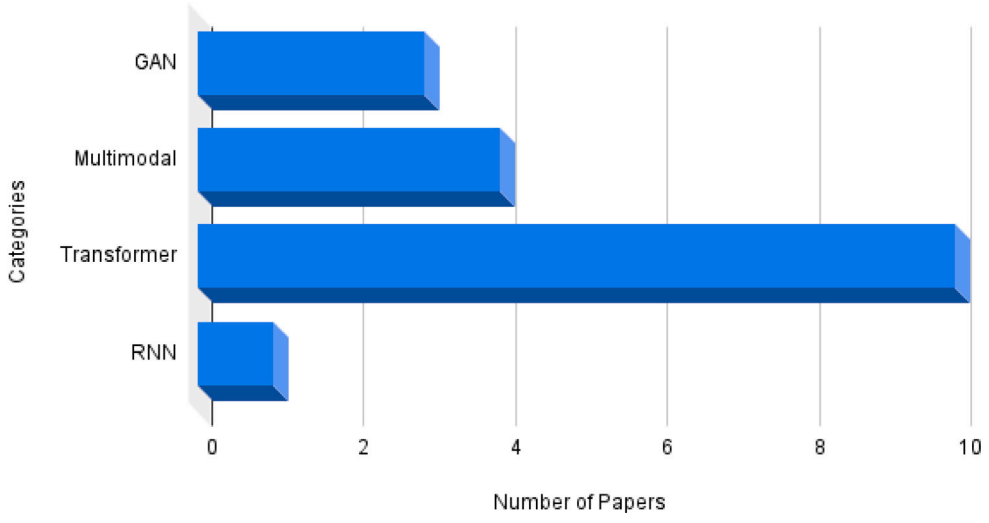

Fig. 9. Other variety of models.

5.2. RQ2: What are the main issues and constraints experienced by researchers and technologists when enhancing video DeepFake detection algorithms?

Although significant advances have been made in DeepFake detection, substantial challenges persist.

Generalisation to new datasets The biggest challenge with DeepFake detection models is generalisation. A model may be well-trained on some specific dataset but may still fail to produce accurate results, even if the test dataset is similar in category to the training dataset. This issue arises mainly due to the deficiencies in current datasets, such as FF++, DFDC, and Celeb-DF which do not encompass the full spectrum of real-world DeepFake manipulations (Ge et al., 2022; Liu et al., 2021). As a result, detection algorithms frequently overfit to available data distributions, making them less effective at adapting to new or unseen types of DF. Generalising well across datasets remains an open problem.

Availability of datasets & types One of the key limitations is the prevalence of homogeneous datasets. Though they provide a large variety of DeepFake demonstrations, datasets such as FF++ and Celeb-DF, they do not include every type of manipulation in existence, especially the newer and more complicated techniques (Chen & Tan, 2021; Hu et al., 2021). It is much harder to train and test models across the full spectrum of DeepFake manipulations. Therefore, there is a growing demand for more comprehensive real-world datasets that are supported in favour of more generalisable DeepFake detection systems.

Real-time detection and the speed of computation The computational complexity of most state-of-the-art DeepFake detection models is a significant challenge for DeepFake detection systems. CNN-RNN

Table 3
Summary of DDT used in Reviewed Literature.

Title	Datasets	Model Used	Strengths	Weaknesses	Generalisability
Coccomini et al. (2023)	FF++, DFDC	XceptionNet	High detection accuracy for manipulated videos by leveraging depth-wise separable convolutions in controlled environment.	Limited scalability & poor adaptability to unseen datasets; resource-intensive, require high-end GPUs for inference.	No
Gani-yusufoglu et al. (2020)	Celeb-DF, DFDC	3D-CNN	Captures spatiotemporal features, improving detection accuracy on manipulated video sequences.	Overfitting risk on small datasets, computationally expensive due to 3Dconvolutional operations.	No
Hu et al. (2021)	FF++, Celeb-DF	Disentangled Representation Learning	Enhances generalisation by isolating domain-invariant features, reducing dataset bias.	Complex training, high computational overhead, sensitive to hyperparameter setting.	Yes
Ismail, Elpeltagy, Zaki, Eldahshan, Kamal (2021a)	FF++, UADFV	YOLO-CNN-XGBoost	Efficient feature extraction, classification pipeline provide faster training, inference times.	Depend on quality of extracted features; struggles with capturing temporal inconsistencies.	Not Mentioned
Gong et al. (2020)	DFDC, DeepFake Detection Challenge	DCGAN	Generates adversarial examples to train detection models more robustly against diverse manipulations.	High computational cost for generating adversarial samples and convergence issues during GAN training.	No
Lewis et al. (2020)	DFDC, Celeb-DF	Hybrid CNN-RNN	Leverages multimodal inconsistencies, improving robustness against high-quality DeepFakes.	Requires extensive training data and struggles with high-resolution videos due to computational bottlenecks.	Yes
Nguyen, Tran, Nguyen, Truong, et al. (2021)	Celeb-DF, FF++	LSTM-CNN	Effectively models temporal inconsistencies in videos, enhancing detection of manipulated motion.	High inference latency and difficulty handling longer sequences.	No
Guarnera et al. (2022)	DFDC, Google DeepFake	Various (Ensemble)	Provides a comprehensive benchmark for evaluating detection models across datasets.	Inconsistent results across models; ensemble methods require high computational resources.	Not Mentioned
Agarwal et al. (2020)	FF++, Celeb-DF	Bi-LSTM	Integrates behavioural and visual cues, improving detection accuracy for subtle manipulations.	Limited performance on datasets with low variability in behaviours or static images.	Yes
Rama chandran, Nadimpalli, and Rattani (2021)	UADFV, DFDC	DeepFace	Leverages pre-trained face embeddings, reducing training time and improving accuracy on facial regions.	Overfits to datasets focused on facial features, limiting general applicability to non-facial manipulations.	No
Rana and Sung (2020)	DFDC, FF++	DeepStack Ensemble	Combines diverse model outputs, making it more robust to adversarial examples.	Computationally expensive ensemble setup, requiring large-scale parallel infrastructure.	Not Mentioned
Singh et al. (2020)	FF++, DFDC	Time-Distributed CNN	Efficiently captures temporal relationships in manipulated videos, enhancing sequential analysis.	Performance drops on long video sequences due to fixed input size constraints.	No
Siegel et al. (2021)	FF++, UADFV	Hand-Crafted Features	Lightweight and interpretable, suitable for low-resource environments.	Limited performance against high-quality manipulations; lacks scalability for diverse datasets.	Not Mentioned
Bansal et al. (2023a)	Celeb-DF, DFDC	Lightweight CNN	Optimised for real-time detection with minimal latency.	Lower accuracy on high-resolution fakes; lacks robustness to sophisticated manipulations.	No
Khalil et al. (2021)	FF++, DFDC	Capsule Networks	Models hierarchical spatial relationships, improving accuracy on subtle facial manipulations.	High computational demand for training, making it unsuitable for low-resource settings.	Yes
Bondi et al. (2020)	DFDC, FF++	Data-Augmented CNN	Improves robustness against unseen manipulations through data augmentation techniques.	Performance heavily dependent on quality and variety of augmentations.	Not Mentioned
Suratkar, Kazi, et al. (2020)	Celeb-DF, FF++	Transfer Learning CNN	Reduces training time by leveraging pretrained networks, maintaining high accuracy.	Requires significant fine-tuning for new datasets; struggles with domain shifts.	Yes

(*continued on next page*)

Table 3 (*continued*).

Title	Datasets	Model Used	Strengths	Weaknesses	Generalisability
Khormali and Yuan (2021)	DFDC, Celeb-DF	Attention Mechanism	Enhances focus on critical features, reducing false positives in detection.	Computationally expensive on larger datasets; not optimised for low-resource environments.	No
Liu et al. (2023)	DFDC, Google DeepFake	Federated Learning	Preserves privacy while improving generalisation across diverse devices.	High communication overhead during training; requires synchronised device participation.	Yes
Ismail, Elpeltagy, Zaki, ElDahshan, A (2021b)	FF++, Celeb-DF	YOLO-Face-RNN	Combines accurate face localisation with sequential analysis for improved detection.	High inference latency for large datasets; struggles with non-facial manipulations.	No
Ge et al. (2021)	FF++, Celeb-DF	Predictive Representation Learning	Learns representations that enhance generalisation across datasets.	High training complexity and dependence on large-scale datasets for effectiveness.	Yes
Wang, Jiang, et al. (2022)	DFDC, Celeb-DF	Feature Point-Based Detection	Detects facial inconsistencies efficiently using feature points, reducing computational cost.	Limited accuracy for subtle manipulations or high-quality fakes.	No
Vinay et al. (2022)	Celeb-DF, UADFV	AFMB-Net	Integrates heart rate analysis, leveraging physiological differences between real and fake videos.	Sensitive to video quality and fails in cases of occlusion or low frame rates.	No
Wang and Dantcheva (2020)	DFDC, FF++	3D-CNN	Evaluates multiple 3D-CNN architectures, identifying optimal configurations for DeepFake detection.	Computationally expensive and overfits on smaller datasets.	No
Jung, Kim, and Kim (2020b)	UADFV, Celeb-DF	Eye Blinking Detection	Utilises eye blinking inconsistencies to detect fake videos, requiring minimal computation.	Ineffective against high-quality DeepFakes that accurately mimic blinking patterns.	No
Zi et al. (2020)	Wild-DeepFake	CNN-RNN	Introduces a real-world dataset, improving robustness testing for detection models.	Dataset diversity limits generalisation for extreme edge cases.	Not Mentioned
Burroughs, Gokaraju, Roy, and Khoa (2020)	DFDC, UADFV	Feature Engineering CNN	Employs hand-crafted features to reduce computational cost while maintaining reasonable accuracy.	Limited scalability for high-quality DeepFake datasets.	No
Liu et al. (2021)	DFDC, Celeb-DF	Lightweight 3D-CNN	Optimised for lower latency and reduced resource requirements, supporting real-time detection.	Lower accuracy on high-resolution manipulations.	Yes
Chintha et al. (2020)	Celeb-DF, UADFV	RNN-CNN	Combines audio and video cues to enhance detection accuracy.	Requires high-quality audio and video synchronisation, limiting applicability.	Yes
Muppalla et al. (2023)	DFDC, FF++	Audio-Visual CNN-RNN	Multimodal integration improves robustness against diverse manipulations.	Computationally expensive; relies on high-quality multimodal data.	Yes
Lee, Lee, and Yoo (2023)	FF++, Celeb-DF	Domain-Wise Clue Detection	Identifies domain-specific artifacts, increasing detection reliability.	Struggles with unseen domains or datasets with minimal artifacts.	Yes
Khormali and Yuan (2022)	DFDC, Celeb-DF	Vision Transformer	Leverages attention mechanisms for robust detection of subtle manipulations.	High computational cost and memory requirements.	Yes
Saikia et al. (2022)	FF++, Celeb-DF	CNN-LSTM with Optical Flow	Utilises motion inconsistencies, enhancing temporal detection.	High computational demands for optical flow estimation.	Yes
Deng et al. (2022)	Celeb-DF, DFDC	EfficientNet-V2	Achieves high accuracy with lower computational cost compared to traditional CNNs.	Limited performance on low-resolution videos.	No
Younus and Hasan (2020)	DFDC, UADFV	Haar Wavelet Transform	Reduces computational cost by extracting efficient spatial features.	Struggles with temporal inconsistencies or high-resolution manipulations.	No
Afchar et al. (2018)	Celeb-DF, UADFV	MesoNet	Compact architecture suitable for low-resource environments.	Limited robustness against high-quality manipulations.	No
Zhang, Wu, et al. (2022)	DFDC, FF++	Cascaded-Hop Model	Iteratively improves detection through cascaded layers, enhancing precision.	High latency due to multiple cascaded processing steps.	No

(*continued on next page*)

Table 3 (*continued*).

Title	Datasets	Model Used	Strengths	Weaknesses	Generalisability
Lai, Wang, Feng, Hu, and Xu (2022)	DFDC, Celeb-DF	Multi-Feature Fusion	Combines spatial, temporal, and frequency features for robust detection.	Computationally expensive; requires extensive preprocessing.	Yes
Gani-yusufoglu et al. (2020)	FF++, DFDC	Spatio-Temporal CNN	Captures temporal artifacts effectively in video sequences.	High memory requirements; struggles with high-resolution videos.	No
Zhao et al. (2021)	DFDC, Celeb-DF	MFF-Net	Combines multimodal features, improving robustness against diverse manipulations.	Requires significant computational resources for training.	Yes
Agarwal, Agarwal, Sinha, Vatsa, and Singh (2021)	FF++, Celeb-DF	MD-CSDNetwork	Cross-domain stitching improves feature sharing across domains, enhancing detection robustness.	High complexity due to multi-domain processing; requires large datasets for training.	Yes
Zhang et al. (2024)	DFDC, Celeb-DF	Audio-Visual Attention with Contrastive Learning	Combines multimodal data with contrastive learning, improving generalisation and robustness.	Resource-intensive; performance depends on the quality of synchronised audio–visual data.	Yes
Lin et al. (2023)	DFDC, FF++	Spatiotemporal Trident Network	Efficiently captures spatial and temporal features with multi-branch processing.	Requires extensive computational resources due to multi-branch architecture.	No
Tang et al. (2024)	Celeb-DF, UADFV	DeepMark Framework	Scalable framework capable of integrating diverse detection methods.	Performance varies with the integrated models; requires tuning for specific datasets.	Yes
Wang et al. (2023)	DFDC, Celeb-DF	Convolutional Pooling Transformer	Combines convolutional pooling and transformer attention, improving detection of subtle manipulations.	High computational cost and memory requirements.	Yes
Zhang et al. (2022)	ADD Challenge Dataset	Score Fusion Model	Aggregates multiple model outputs, improving overall accuracy for the challenge.	Overfitting risk due to reliance on specific dataset characteristics.	No
Khan and Dang-Nguyen (2022)	FF++, Celeb-DF	Hybrid Transformer Network	Combines CNN and Transformer features for enhanced robustness.	High resource requirements for training; sensitive to hyperparameter selection.	Yes
Xiao, Zhang, Yang, Wen, and Li (2023)	DFDC, Celeb-DF	Transformer	Detects forgery by focusing on invariant features and enhanced details.	Limited performance on low-resolution videos.	No
Nadimpalli and Rattani (2023)	FF++, Celeb-DF	GAN-based Watermarking	Adds visible watermarks, making DeepFakes easier to detect.	Watermarking can be removed by advanced forgery methods.	No
Zhang et al. (2022)	DFDC, Celeb-DF	Spatiotemporal Dropout Transformer	Efficiently captures spatiotemporal inconsistencies with dropout regularisation.	Requires large-scale datasets to prevent overfitting.	Yes
Lugstein, Baier, Bachinger, and Uhl (2021)	DFDC, FF++	PRNU-Based Detection	Leverages Photo-Response Non-Uniformity (PRNU) for robust detection of camera inconsistencies.	Limited to detecting manipulations visible in camera artifacts.	No
Jaleel and Hadi (2022)	Celeb-DF, DFDC	Action Unit Detection	Exploits inconsistencies in facial action units for detection, improving robustness.	Struggles with high-quality fakes mimicking natural expressions.	No
İlhan, Balı, and Karaköse (2022)	FF++, UADFV	NASNetLarge CNN	High accuracy with reduced computational cost due to efficient neural architecture search.	Resource-intensive for training and fine-tuning.	No
Joseph and Nyirenda (2021)	DFDC, Celeb-DF	Two-Stream Capsule Network	Combines spatial and temporal streams for improved forgery detection.	High computational demands and complex architecture.	Yes
Yadav et al. (2022)	FF++, Celeb-DF	Generalised Model	Enhances robustness by addressing overfitting with domain adaptation techniques.	Limited by dataset diversity; struggles with extreme edge cases.	Yes
Garde, Suratkar, and Kazi (2022)	DFDC, Celeb-DF	3D-CNN	Utilises advanced AI algorithms for forgery detection.	Lacks transparency in methodology and generalisability metrics.	No
Ranjan, Patil, and Kazi (2020)	FF++, UADFV	Transfer Learning CNN	Reduces training time while improving performance across datasets.	Requires fine-tuning for diverse datasets.	Yes
Zhao, Wang, and Lu (2020)	DFDC, Celeb-DF	Two-Stream CNN	Effectively captures global and local inconsistencies in video sequences.	High computational cost and sensitivity to noise in the data.	Yes

(*continued on next page*)

Table 3 (*continued*).

Title	Datasets	Model Used	Strengths	Weaknesses	Generalisability
Stephen and Mantoro (2022)	Celeb-DF, DFDC	lightweight 3D-CNN	Specialises in detecting face-swapping manipulations.	Limited to specific forgery types; lacks flexibility for other manipulations.	No
Cho et al. (2023)	Wild-DeepFake	Multimodal	Provides insights into real-world DeepFake detection challenges	Model performance varies significantly depending on datasets used	Not Mentioned
Pasupuleti, Tathireddy, Dontagani, and Rahim (2023)	DFDC, FF++	Custom Densenet	High accuracy with feature reuse, reducing the computational footprint.	Overfits to specific datasets; struggles with unseen manipulations.	No
Mitra, Mohanty, Corcoran, and Kougianos (2021)	Social Media Dataset	Key Frame CNN	Reduces computational overhead by analysing key video frames.	Limited performance on high-quality and temporally inconsistent fakes.	No
Pryor, Dave, Vanamala, et al. (2023)	Hybrid Dataset	CNN-SVM Hybrid	Combines feature extraction (CNN) and effective classification (SVM).	High dependency on feature selection quality; scalability issues.	No
Stanciu and Ionescu (2023)	Celeb-DF, DFDC	Autoencoder Augmentation	Enhances training data diversity, improving robustness against unseen manipulations.	Performance highly dependent on augmentation quality.	Yes
Nirkin, Wolf, Keller, and Hassner (2021)	DFDC, FF++	Context-Based CNN	Detects inconsistencies between facial features and their context in videos.	Requires high-resolution video data for effective detection.	No
Jiang et al. (2021)	FF++, Celeb-DF	CNN-RNN	Adapts to new domains effectively using domain adaptation techniques.	Struggles with datasets containing minimal artifacts.	Yes
Abdulhamid and Hashim (2023)	DFDC, Celeb-DF	Enhanced Preprocessing	Improved feature extraction tech. to enhance overall detection accuracy.	Relies heavily on preprocessing techniques; computationally expensive.	No
Mallet, Krueger, Dave, and Vanamala (2023)	DFDC, Celeb-DF	MLP-LSTM	Captures sequential features efficiently, improving temporal detection accuracy.	High memory consumption due to sequential processing.	Yes
Li et al. (2023)	FF++, Celeb-DF	Self-Supervised Transformer	Enhances detection by leveraging spatio-temporal inconsistencies with self-supervised learning.	Resource-intensive, requiring extensive training data.	Yes
Aduwala, Arigala, Desai, Quan, and Eirinaki (2021)	DFDC, Celeb-DF	GAN Discriminators	Identifies manipulation artifacts effectively by leveraging adversarial features.	Susceptible to adversarial attacks and requires high computational resources.	No
Stanciu and Ionescu (2022)	FF++, Celeb-DF	Capsule Networks	Effectively captures hierarchical spatial relationships for improved detection.	High computational cost and training complexity.	Yes
Heo, Yeo, and Kim (2023)	DFDC, Celeb-DF	Improved Vision Transformer	Leverages transformer attention for subtle manipulation detection.	Requires large-scale datasets and high memory for training.	Yes
Bhaumik and Woo (2023)	DFDC, Celeb-DF	Object Representation Analysis	Focuses on object-level inconsistencies for improved detection.	Limited robustness against highly realistic manipulations.	No
Bomma-reddy, Samyal, and Dahiya (2023)	DFDC, FF++	CNN with Adversarial Training	Robust against adversarial examples; improves detection generalisation.	Resource-intensive; requires careful adversarial training setup.	Yes
Liu, Li, Duan, and Huang (2022)	FF++, Celeb-DF	Dual Transformer	Combines spatial and temporal transformers for comprehensive detection.	Computationally expensive and challenging to optimise.	Yes
Dong, Wang, Liang, Fan, and Ji (2022)	Celeb-DF, DFDC	Image Matching CNN	Utilises image matching techniques to explain detection decisions.	Limited performance on high-resolution videos.	No
Mehta, Gupta, Subramanian, and Dhall (2021)	Video Confer-encing Dataset	3D-CNN	Designed for real-time video conferencing scenarios with low latency.	Struggles with high-quality and low-resolution fakes.	No
Chen, Lin, Li, and Tan (2022)	FF++, Celeb-DF	Generalised Learning Framework	Enhances generalisation by capturing intra-consistency and inter-diversity.	High sensitivity to noise in datasets; complex training pipeline.	Yes
Guefrachi et al. (2023)	DFDC, UADFV	3D-CNN	Achieves reasonable accuracy with lightweight architecture.	Limited robustness against advanced manipulations.	No
Rahman et al. (2022)	Low-Resolution Dataset	3D-CNN	Optimised for low-resolution DeepFake detection, reducing computational cost.	Limited applicability for high-resolution and complex manipulations.	No

(*continued on next page*)

Table 3 (*continued*).

Title	Datasets	Model Used	Strengths	Weaknesses	Generalisability
Tariq, Lee, and Woo (2020)	DFDC, Celeb-DF	ConvLSTM Residual Network	Effectively models temporal inconsistencies with residual learning.	High memory requirements and computational demands.	Yes
Guefrechi, Jabra, and Hamam (2022)	FF++, Celeb-DF	InceptionResnetV2	High accuracy due to advanced architecture and residual connections.	Computationally expensive and overfits on small datasets.	No
Yang, Chen, and Zhong (2023)	Celeb-DF, DFDC	Multi-Attentional Model	Combines spatial and channel information for robust detection.	High computational cost and complex model tuning.	Yes
Hongmeng, Zhiqiang, Lei, Xiuqing, and Yuehan (2020)	Compressed Video Dataset	Super-Resolution CNN	Detects artifacts in hard-compressed videos with super-resolution techniques.	Struggles with high-quality uncompressed videos.	No
Arini, Bahaweres, and Al Haq (2022)	FF++	Xception + ResNet-50	Combines pretrained models for fast and efficient classification.	Limited by preprocessing steps and binary pattern extraction.	No
Beuve, Hamidouche, and Deforges (2021)	DFDC, Celeb-DF	Dummy Triplet Loss Model	Introduces triplet loss to improve robustness against adversarial examples.	Computationally intensive and sensitive to hyperparameter tuning.	Yes
Chugh, Gupta, Dhall, and Subramanian (2020)	DFDC, Celeb-DF	Audio-Visual Dissonance	Focuses on audio–visual inconsistencies, enhancing multimodal detection.	Requires high-quality and synchronised multimodal data.	Yes
Bansal et al. (2023b)	DFDC, FF++	Artifact-Based CNN	Targets manipulation artifacts for accurate detection.	Limited to artifacts visible in specific datasets.	No
Maksutov, Morozov, Lavrenov, and Smirnov (2020)	Celeb-DF, DFDC	Various ML Models	Provides insights into the performance of machine learning techniques for detection.	Model performance varies significantly by dataset.	Not Mentioned
Khan and Dai (2021)	FF++, DFDC	Video Transformer	Leverages incremental learning to adapt to new datasets and manipulations.	Requires large-scale resources for effective training.	Yes
Adnan and Abdulbaqi (2022)	Celeb-DF, DFDC	3D-CNN	Simplifies detection by focusing on core convolutional features, improving processing speed.	Struggles with detecting subtle manipulations in high-quality fakes.	No
Mira (2023)	FF++, Celeb-DF	RNN	Achieves reasonable detection accuracy with minimal computational requirements.	Limited robustness to diverse manipulations across datasets.	No
Ganguly, Mohiuddin, Malakar, Cuevas and Sarkar (2022)	Celeb-DF, DFDC	Visual Attention Model	Enhances detection by focusing on forgery-critical regions using attention mechanisms.	Computationally intensive and sensitive to attention weights.	Yes
Masud, Sadiq, Masood, Ahmad, and Abd El-Latif (2023)	DFDC, FF++	LW-DeepFakeNet	Optimised for real-time detection with low computational overhead.	Reduced accuracy for complex manipulations in high-resolution videos.	No
Baxevana-kis, Kordopatis-Zilos, Galopoulos, Apostolidis, Levacher, Baris Schlicht, Teyssou, Kompatsiaris, and Papadopoulos (2022)	Wild-DeepFake	MeVer Framework	Practical insights into real-world deployment, improving scalability.	Limited generalisability due to dataset-specific optimisations.	No
Humidan, Abdullah, and Halin (2022)	Compressed Video Dataset	Compression-Aware Model	Handles artifacts in compressed videos, improving detection accuracy.	Struggles with uncompressed high-quality fakes.	No
Kaddar, Fezza, Hamidouche, Akhtar, and Hadid (2021)	DFDC, Celeb-DF	HCiT Model	Combines CNN and Transformer features for robust detection.	Requires extensive tuning and high computational resources.	Yes
Mitra, Mohanty, Corcoran, and Kougianos (2020)	Social Media Dataset	3D-CNN	Tailored for social media platforms, optimising for low-resolution videos.	Struggles with high-resolution manipulations and diverse datasets.	No
Gu et al. (2021)	FF++, DFDC	Spatiotemporal CNN	Captures temporal inconsistencies effectively, enhancing sequential detection.	Requires extensive training data for accurate predictions.	Yes
Suratkar, Johnson, Variyambat, and Panchal and Kazi (2020)	Celeb-DF, UADFV	Transfer Learning CNN	Reduces training time while improving generalisation across datasets.	Requires significant fine-tuning for new manipulations.	Yes
Asha, Vinod, and Menon (2023)	DFDC, Celeb-DF	Temporal–Spatial Defense Model	Robust against adversarial attacks by leveraging both spatial and temporal features.	High computational cost and training complexity.	Yes

(*continued on next page*)

Table 3 (*continued*).

Title	Datasets	Model Used	Strengths	Weaknesses	Generalisability
Jain, Korshunov, and Marcel (2021)	Celeb-DF, DFDC	Attribution-Based Model	Focuses on model attribution to improve generalisation across datasets.	Requires careful labelling and large-scale datasets for training.	Yes
Ahmed and Sonuç (2023)	FF++, DFDC	Rationale-Augmented CNN	Enhances detection accuracy by integrating rationale-based features.	Computationally demanding and dependent on quality of rationale extraction.	No
Ganguly, Ganguly, Mohiuddin, Malakar and Sarkar (2022)	Celeb-DF, DFDC	ViXNet	Combines Vision Transformer & Xception to enhance spatial–temporal analysis.	Requires extensive computational resources and memory.	Yes
Beuve, Hamidouche, and Déforges (2023)	DFDC, Celeb-DF	Dummy Triplet Loss Model	Improves robustness to adversarial attacks with hierarchical learning.	Sensitive to hyperparameter tuning and training configuration.	Yes
Ge et al. (2021)	Celeb-DF, FF++	Predictive Representation Learning	Learns latent patterns effectively for robust forgery detection.	High training complexity and computational demands.	Yes
Zhang, Li, Lin, Zeng, and Ge (2021)	Celeb-DF, FF++, DFDC	Temporal Dropout 3DCNN (TD-3DCNN)	Used 3DCNN and 3D Inception Modules to extract features and Temporal Dropout to leverage inconsistent cues in video frames.	High training complexity and tested only on selected datasets.	Yes
Choi, Kim, Jeong, Baek, and Choi (2024)	Celeb-DF, FF++, DFD	Style latent flow (StyleGRU module) extracted from consecutive frames of a video is used as a cue.	Proposed temporal changes in style latent vector to generalise DF video detection.	High data preprocessing time.	Yes

hybrid, ViTs, and GAN-based approaches require substantial computing resources, particularly for real-time detection of high-resolution videos. This complexity can also limit the scalability of these models, making it difficult to deploy them in real-world applications where both timeliness and efficiency are critical (Ge et al., 2022). The balance between accuracy and computational efficiency remains a key area for future research.

5.3. RQ3: What data sources are utilised to evaluate video DeepFake detection techniques?

The choice of datasets in DeepFake detection benchmarks has a significant impact on the ability and resilience of these techniques. More importantly, it is crucial in determining the effectiveness and generalisability of the detection techniques. The dataset used in the reviewed literature are included in the second column of the Table 3;

Common datasets include:

FaceForensics++ (FF++) One of the most popular datasets for DeepFake detection is FF++, which has a large collection of fake videos that consist of different resolutions with both high and low quality deep fakes. It forms an important benchmark that is very useful to evaluate detection algorithms (Liu et al., 2021).

DeepFake detection challenge (DFDC) With a wide range of DF, the DFDC dataset is one of the largest and most extensive datasets available. Frequently used for assessing the detection accuracy of subtle facial manipulations and lip-sync in videos (Ge et al., 2022).

Celeb-DF

Celeb-DF A high-quality dataset over facial reenactment techniques. This makes it an excellent testbed for testing sophisticated manipulation methods as well (Chen & Tan, 2021), particularly when evaluating the capacity of the models to detect subtle manipulations.

Cross-dataset evaluation: For generalisability, a common strategy is to evaluate models on separate datasets; what Ming Liu et al. referred to as cross-dataset evaluation in the following citation: This method is used to determine whether a detection model can effectively adapt across distribution changes and transformation techniques (Hu et al., 2021). The research is valid, as DeepFake mechanisms are getting better and we need systems that can generalise to the real world so in the future models built on those data will be more applicable.

5.4. RQ4: Can video DeepFake detection be generalised?

The generalisation is probably, one of the major challenges in DeepFake detection models.

Generalisation potential

Most of the systematic review papers, as highlighted in Section 2, failed to discuss strengths and weaknesses of the existing DeepFake detection methods concerning their ability to generalise on new or unseen manipulations. Lack of such analysis raises difficulties in assessment of real-world applicability, especially when confronting DF different from training datasets. These are gaps in the literature that set limits on how deeply one would understand the performance and robustness of these detecting techniques in a practical way. As shown in Fig. 10 around 46.3% of researchers believe that DeepFake detection is generalisable across different types of DF and datasets. However, state-of-the-art models still degrade in performance when tested with novel or previously unobserved manipulations (Chen & Tan, 2021). CNN-RNN hybrids and ViTs perform quite well, but due to their training on specific datasets, they are somewhat limited in generalisation.

5.5. Recommendations

Importance of standadised datasets with proper scoring system

After reviewing these studies, we emphasise the need for a standardise dataset with an appropriate scoring system. This scoring system should reflect the types of DeepFakes in the data set and their complexities.

Support cross-study analysis

These scoring systems need a proper framework to support the analysis of detection models' performances across different datasets, ensuring that variations in DeepFake types and complexities. By standardising the evaluation process, these scoring systems will enable a consistent and reliable evaluation of the performance of new DeepFake detection models.

In addition, these systems will provide a common platform for cross-study comparisons, facilitate the identification of best practices, and enhance reproducibility in DeepFake research. Ultimately, this approach will contribute to the development of more robust and generalisable detection models.

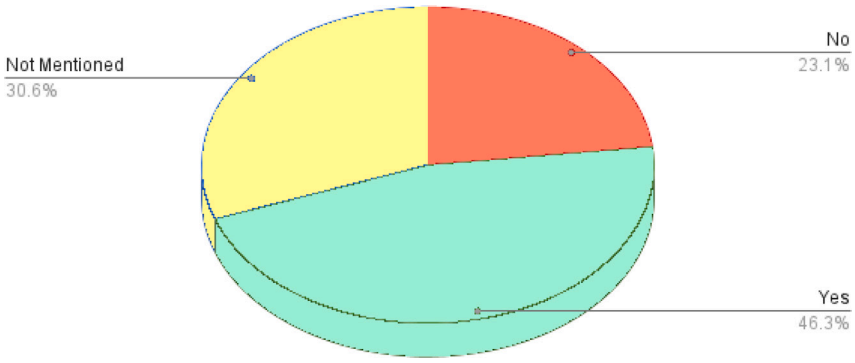

Fig. 10. Researchers discuss the potential for generalisability in DeepFake.

6. Conclusion

This research provides a comprehensive analysis of DeepFake detection techniques, highlighting their strengths and weaknesses, as well as their generalisability, that have not been adequately addressed in the existing literature. The heatmap and Table 3 offer an overview of the primary detection models recently applied in the video DeepFake detection domain, emphasising their strengths, limitations, and challenges associated with model generalisation. These insights serve as a valuable foundation for future research, providing a clear understanding of the current state of the field. Researchers can find key features and algorithms of tested models in a single compiled table, facilitating the design of more robust and effective algorithms.

The findings indicate that detection models designed for specific datasets face significant limitations, particularly in generalisation of the findings to unseen or real-world situations. CNN-based models have shown strong performance, especially in identifying spatial inconsistencies in images and videos; their effectiveness tends to deteriorate when confronted with a DeepFake from a different environment. Furthermore, the tendency of these models to overfit to specific subsets of data has been identified as a critical weakness, severely limiting their applicability and effectiveness in real-world applications. It is important to note that this review has certain limitations and may not include studies published before 2018 or after February 2024.

Furthermore, current research supports multimodal and mixed approaches, where different detection methods are integrated to improve performance. Examples include combining CNN-RNN models or using audiovisual models that take advantage of both visual and auditory cues. These models have demonstrated significant improvements in robustness due to their ability to capture spatial and temporal anomalies that are crucial for identifying complex manipulations in video DeepFakes. However, as with other advanced models, these models come with substantial computational costs that can hinder their usability in real-time applications, especially in environments with restricted resources. For example, deploying such models in edge devices or latency-sensitive systems remains a significant challenge.

The current literature on DeepFake detection exhibits limitations, primarily due to the reliance on a limited set of benchmark datasets, including FF++, DFDC, and Celeb-DF. These datasets fail to encompass the full spectrum of DF manipulations. Hence, models trained on these datasets often biased toward specific types of DF manipulations and often fail to understand new forms of manipulation. This lack of diversity in training data poses a critical challenge and highlights the need to develop extensive and diverse datasets that mimic a greater variety of the DeepFake manipulations present in real-world applications.

Moreover, sophisticated detection methods, such as transformers, GAN-based approaches, and capsule networks, are available but are computationally intensive and difficult to deploy. This limitation is especially significant in cases where real-time detection or response is necessary, making large-scale deployment difficult at present.

The challenge of generalising DeepFake detection models remains significant and requires further research. The survey revealed that 46.3% of the selected publications recognised that, proposed DeepFake detection techniques could be generalised across various types of DF and datasets. It emphasises the need for future research to focus on developing more adaptable detection models maintain high accuracy rates across different testing environments and various datasets. Researchers must seek more sustainable and generalisable solutions that are not susceptible to overfitting. This includes enhancing data augmentation methods and developing more accurate datasets.

To overcome these challenges, it is imperative to direct new research efforts to develop standardised datasets accompanied by an appropriate scoring system. These systems should accurately reflect the types of DeepFakes within the dataset and their associated complexities. Furthermore, a well-defined framework is essential to support for cross-study comparisons, assist in identifying best practices, and improve reproducibility in DeepFake research. Ultimately, these approaches will contribute towards a development of more robust and generalisable detection models.

Abbreviations

The following abbreviations are used in this manuscript:

DF	DeepFake
DDT	DeepFake Detection Techniques
SLR	Systematic Literature Review
ASARM	Adaptive-Support ARM
DL	Deep Learning
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
VGG	Video
FF++	Faceforensic ++
NM	Not-Mentioned
Y	Yes
N	No
DFDC	DeepFake Detection Challenge
CNNs	Convolutional Neural Networks
VGG	Visual Geometry Group
GAN	generative adversarial networks
ViTs	Vision Transformers
DCGAN	Deep convolutional generative adversarial networks
CL	Contrastive Learning
Long-Short-Term-Memory	LSTM

CRediT authorship contribution statement

Ramcharan Ramanaharan: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Deepani B. Guruge:** Writing – review & editing, Supervision, Resources, Methodology, Formal analysis, Conceptualization. **Johnson I. Agbinya:** Writing – review & editing, Supervision, Resources, Project administration, Methodology.

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.

Acknowledgements

The authors acknowledge the MIT administration for providing equipment, work space, and other administrative supports.

References

Abbas, F., & Taeihagh, A. (2024). Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems with Applications*, Article 124260.

Abdulhamid, M. A., & Hashim, A. N. (2023). Enhanced preprocessing stage for feature extraction of deepfake detection based on deep learning methods. In *2023 7th international symposium on innovative approaches in smart technologies* (pp. 1–6). IEEE.

Abdullah, M. T., & Ali, N. H. M. (2023). DeepFake detection improvement for images based on a proposed method for local binary pattern of the multiple-channel color space. *International Journal of Intelligent Engineering & Systems*, *16*(3).

Adnan, S. R., & Abdulbaqi, H. A. (2022). Deepfake video detection based on convolutional neural networks. In *2022 international conference on data science and intelligent computing* (pp. 65–69). IEEE.

Aduwala, S. A., Arigala, M., Desai, S., Quan, H. J., & Eirinaki, M. (2021). Deepfake detection using GAN discriminators. In *2021 IEEE seventh international conference on big data computing service and applications (bigDataService)* (pp. 69–77). IEEE.

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security* (pp. 1–7). IEEE.

Agarwal, A., Agarwal, A., Sinha, S., Vatsa, M., & Singh, R. (2021). MD-csdnetwork: Multi-domain cross stitched network for deepfake detection. In *2021 16th IEEE international conference on automatic face and gesture recognition (FG 2021)* (pp. 1–8). IEEE.

Agarwal, S., Farid, H., El-Gaaly, T., & Lim, S.-N. (2020). Detecting deep-fake videos from appearance and behavior. In *2020 IEEE international workshop on information forensics and security* (pp. 1–6). IEEE.

Ahmed, S. R. A., & Sonuç, E. (2023). RETRACTED ARTICLE: Deepfake detection using rationale-augmented convolutional neural network. *Applied Nanoscience*, *13*(2), 1485–1493.

Arini, A., Bahaweres, R. B., & Al Haq, J. (2022). Quick classification of xception and resnet-50 models on deepfake video using local binary pattern. In *2021 international seminar on machine learning, optimization, and data science* (pp. 254–259). IEEE.

Asha, S., Vinod, P., & Menon, V. G. (2023). A defensive framework for deepfake detection under adversarial settings using temporal and spatial features. *International Journal of Information Security*, *22*(5), 1371–1382.

Bansal, N., Aljrees, T., Yadav, D. P., Singh, K. U., Kumar, A., Verma, G. K., et al. (2023a). Real-time advanced computational intelligence for deep fake video detection. *Applied Sciences*, *13*(5), 3095.

Bansal, S., et al. (2023b). Artifact based deepfake detection methods. In *2023 second international conference on informatics* (pp. 1–6). IEEE.

Baxevana-kis, S., Kordopatis-Zilos, G., Galopoulos, P., Apostolidis, L., Levacher, K., Baris Schlicht, I., et al. (2022). The mever deepfake detection service: Lessons learnt from developing and deploying in the wild. In *Proceedings of the 1st international workshop on multimedia AI against disinformation* (pp. 59–68).

Beuve, N., Hamidouche, W., & Deforges, O. (2021). Dmyt: Dummy triplet loss for deepfake detection. In *Proceedings of the 1st workshop on synthetic multimedia-audiovisual deepfake generation and detection* (pp. 17–24).

Beuve, N., Hamidouche, W., & Déforges, O. (2023). Hierarchical learning and dummy triplet loss for efficient deepfake detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, *20*(3), 1–18.

Bhaumik, K. K., & Woo, S. S. (2023). Exploiting inconsistencies in object representations for deepfake video detection. In *Proceedings of the 2nd workshop on security implications of deepfakes and cheapfakes* (pp. 11–15).

Bomma-reddy, S., Samyal, T., & Dahiya, S. (2023). Implementation of a deepfake detection system using convolutional neural networks and adversarial training. In *2023 3rd international conference on intelligent technologies* (pp. 1–6). IEEE.

Bondi, L., Cannas, E. D., Bestagini, P., & Tubaro, S. (2020). Training strategies and data augmentations in cnn-based deepfake video detection. In *2020 IEEE international workshop on information forensics and security* (pp. 1–6). IEEE.

Burroughs, S. J., Gokaraju, B., Roy, K., & Khoa, L. (2020). Deepfakes detection in videos using feature engineering techniques in deep learning convolution neural network frameworks. In *2020 IEEE applied imagery pattern recognition workshop* (pp. 1–4). IEEE.

Chang, X., Wu, J., Yang, T., & Feng, G. (2020). Deepfake face image detection based on improved VGG convolutional neural network. In *2020 39th Chinese control conference* (pp. 7252–7256). IEEE.

Chen, T., Kumar, A., Nagarsheth, P., Sivaraman, G., & Khoury, E. (2020). Generalization of audio deepfake detection. In *Odyssey* (pp. 132–137).

Chen, H., Lin, Y., Li, B., & Tan, S. (2022). Learning features of intra-consistency and inter-diversity: Keys toward generalizable deepfake detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(3), 1468–1480.

Chen, B., & Tan, S. (2021). Featuretransfer: Unsupervised domain adaptation for cross-domain deepfake detection. *Security and Communication Networks*, *2021*, 1–8.

Chintha, A., Thai, B., Sohrawardi, S. J., Bhatt, K., Hickerson, A., Wright, M., et al. (2020). Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE Journal of Selected Topics in Signal Processing*, *14*(5), 1024–1037.

Cho, B., Le, B. M., Kim, J., Woo, S., Tariq, S., Abuadbba, A., et al. (2023). Towards understanding of deepfake videos in the wild. In *Proceedings of the 32nd ACM international conference on information and knowledge management* (pp. 4530–4537).

Choi, J., Kim, T., Jeong, Y., Baek, S., & Choi, J. (2024). Exploiting style latent flows for generalizing deepfake video detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1133–1143).

Choudhary, S., Saurav, S., Saini, R., & Singh, S. (2023). Capsule networks for computer vision applications: a comprehensive review. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *53*(19), 21799–21826.

Chugh, K., Gupta, P., Dhall, A., & Subramanian, R. (2020). Not made for each other-audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 439–447).

Coccomini, D. A., Caldelli, R., Falchi, F., & Gennaro, C. (2023). On the generalization of deep learning models in video deepfake detection. *Journal of Imaging*, *9*(5), 89.

Coccomini, D. A., Caldelli, R., Falchi, F., Gennaro, C., & Amato, G. (2022). Cross-forgery analysis of vision transformers and cnns for deepfake image detection. In *Proceedings of the 1st international workshop on multimedia AI against disinformation* (pp. 52–58).

Deng, L., Suo, H., Li, D., et al. (2022). Deepfake video detection based on EfficientNet-V2 network. *Computational Intelligence and Neuroscience*, *2022*.

Ding, F., Zhu, G., Li, Y., Zhang, X., Atrey, P. K., & Lyu, S. (2021). Anti-forensics for face swapping videos via adversarial training. *IEEE Transactions on Multimedia*, *24*, 3429–3441.

Dong, S., Wang, J., Liang, J., Fan, H., & Ji, R. (2022). Explaining deepfake detection by analysing image matching. In *European conference on computer vision* (pp. 18–35). Springer.

Du, M., Pentyala, S., Li, Y., & Hu, X. (2020). Towards generalizable deepfake detection with locality-aware AutoEncoder. In *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 325–334). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3340531.3411892.

Gambini, M., Fagni, T., Falchi, F., & Tesconi, M. (2022). On pushing DeepFake tweet detection capabilities to the limits. In *Proceedings of the 14th ACM web science conference 2022* (pp. 154–163). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3501247.3531560.

Ganguly, S., Ganguly, A., Mohiuddin, S., Malakar, S., & Sarkar, R. (2022). ViXNet: Vision transformer with xception network for deepfakes based video and image forgery detection. *Expert Systems with Applications*, *210*, Article 118423.

Ganguly, S., Mohiuddin, S., Malakar, S., Cuevas, E., & Sarkar, R. (2022). Visual attention-based deepfake video forgery detection. *Pattern Analysis and Applications*, *25*(4), 981–992.

Gani-yusufoglu, I., Ngô, L. M., Savov, N., Karaoglu, S., & Gevers, T. (2020). Spatio-temporal features for generalized detection of deepfake videos. arXiv preprint arXiv:2010.11844.

Garde, A., Suratkar, S., & Kazi, F. (2022). AI based deepfake detection. In *2022 IEEE 1st international conference on data, decision and systems* (pp. 1–6). IEEE.

Ge, S., Lin, F., Li, C., Zhang, D., Tan, J., Wang, W., et al. (2021). Latent pattern sensing: Deepfake video detection via predictive representation learning. In *Proceedings of the 3rd ACM international conference on multimedia in Asia* (pp. 1–7).

Ge, S., Lin, F., Li, C., Zhang, D., Wang, W., & Zeng, D. (2022). Deepfake video detection via predictive representation learning. *ACM Trans. Multimedia Comput. Commun. Appl.*, *18*(2s), http://dx.doi.org/10.1145/3536426.

Giudice, O., Guarnera, L., & Battiato, S. (2021). Fighting deepfakes by detecting gan dct anomalies. *Journal of Imaging*, *7*(8), 128.

Gong, D., Goh, O. S., Kumar, Y. J., Ye, Z., & Chi, W. (2020). Deepfake forensics, an ai-synthesized detection with deep convolutional generative adversarial networks. *International Journal*, *9*(3).

Gong, D., Kumar, Y. J., Goh, O. S., Ye, Z., & Chi, W. (2021). DeepfakeNet, an efficient deepfake detection method. *International Journal of Advanced Computer Science and Applications*, *12*(6), 201–207.

Gu, Z., Chen, Y., Yao, T., Ding, S., Li, J., Huang, F., et al. (2021). Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 3473–3481).

Guarnera, L., Giudice, O., Guarnera, F., Ortis, A., Puglisi, G., Paratore, A., et al. (2022). The face deepfake detection challenge. *Journal of Imaging*, *8*(10), 263.

Guefrachi, S., Jabra, M. B., Alsharabi, N. A., Othman, M. T. B., Alharabi, Y. O., Alkholidi, A., et al. (2023). Deep learning based DeepFake video detection. In *2023 international conference on smart computing and application* (pp. 1–8). IEEE.

Guefrechi, S., Jabra, M. B., & Hamam, H. (2022). DeepFake video detection using InceptionResnetV2. In *2022 6th international conference on advanced technologies for signal and image processing* (pp. 1–6). IEEE.

Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *14*(2), Article e1520.

Heo, Y.-J., Yeo, W.-H., & Kim, B.-G. (2023). Deepfake detection algorithm based on improved vision transformer. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *53*(7), 7512–7527.

Hongmeng, Z., Zhiqiang, Z., Lei, S., Xiuqing, M., & Yuehan, W. (2020). A detection method for deepfake hard compressed videos based on super-resolution reconstruction using CNN. In *Proceedings of the 2020 4th high performance computing and cluster technologies conference & 2020 3rd international conference on big data and artificial intelligence* (pp. 98–103).

Hu, J., Wang, S., & Li, X. (2021). Improving the generalization ability of deepfake detection via disentangled representation learning. In *2021 IEEE international conference on image processing* (pp. 3577–3581). IEEE.

Humidan, A. S., Abdullah, L. N., & Halin, A. A. (2022). Detection of compressed DeepFake video drawbacks and technical developments. In *2022 5th international conference on signal processing and information security* (pp. 11–16). IEEE.

İlhan, I., Balı, E., & Karaköse, M. (2022). An improved deepfake detection approach with nasNetLarge CNN. In *2022 international conference on data analytics for business and industry* (pp. 598–602). IEEE.

Ismail, A., Elpeltagy, M., S. Zaki, M., & Eldahshan, K. (2021a). A new deep learning-based methodology for video deepfake detection using xgboost. *Sensors*, *21*(16), 5413.

Ismail, A., Elpeltagy, M., Zaki, M., & ElDahshan, K. A. (2021b). Deepfake video detection: YOLO-face convolution recurrent approach. *PeerJ Computer Science*, *7*, Article e730.

Jada, I., & Mayayise, T. O. (2024). The impact of artificial intelligence on organisational cyber security: An outcome of a systematic literature review. *Data and Information Management*, *8*(2), Article 100063.

Jain, A., Korshunov, P., & Marcel, S. (2021). Improving generalization of deepfake detection by training for attribution. In *2021 IEEE 23rd international workshop on multimedia signal processing* (pp. 1–6). IEEE.

Jaleel, Q., & Hadi, I. (2022). Facial action unit-based deepfake video detection using deep learning. In *2022 4th international conference on current research in engineering and science applications* (pp. 228–233). IEEE.

Jia, M., Cheng, X., Lu, S., & Zhang, J. (2022). Learning disentangled representation implicitly via transformer for occluded person re-identification. *IEEE Transactions on Multimedia*, *25*, 1294–1305.

Jiang, J., Li, B., Wei, B., Li, G., Liu, C., Huang, W., et al. (2021). FakeFilter: A cross-distribution deepfake detection system with domain adaptation. *Journal of Computer Security*, *29*(4), 403–421.

Joseph, Z., & Nyirenda, C. (2021). Deepfake detection using a two-stream capsule network. In *2021 IST-africa conference (IST-africa)* (pp. 1–8). IEEE.

Jung, T., Kim, S., & Kim, K. (2020a). DeepVision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, *8*, 83144–83154. http://dx.doi.org/10.1109/ACCESS.2020.2988660.

Jung, T., Kim, S., & Kim, K. (2020b). Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, *8*, 83144–83154.

Kaddar, B., Fezza, S. A., Hamidouche, W., Akhtar, Z., & Hadid, A. (2021). Hcit: Deepfake video detection using a hybrid model of CNN features and vision transformer. In *2021 international conference on visual communications and image processing* (pp. 1–5). IEEE.

Karanwal, S., & Diwakar, M. (2023). Triangle and orthogonal local binary pattern for face recognition. *Multimedia Tools and Applications*, *82*(23), 36179–36205.

Kaur, A., Noori Hoshyar, A., Saikrishna, V., Firmin, S., & Xia, F. (2024). Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review*, *57*(6), 1–47.

Khalil, H. A., & Maged, S. A. (2021). Deepfakes creation and detection using deep learning. In *2021 international mobile, intelligent, and ubiquitous computing conference* (pp. 1–4). IEEE.

Khalil, S. S., Youssef, S. M., & Saleh, S. N. (2021). Icaps-dfake: An integrated capsule-based model for deepfake image and video detection. *Future Internet*, *13*(4), 93.

Khan, S. A., & Dai, H. (2021). Video transformer for deepfake detection with incremental learning. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 1821–1828).

Khan, S. A., & Dang-Nguyen, D.-T. (2022). Hybrid transformer network for deepfake detection. In *Proceedings of the 19th international conference on content-based multimedia indexing* (pp. 8–14). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3549555.3549588.

Khormali, A., & Yuan, J.-S. (2021). Add: Attention-based deepfake detection approach. *Big Data and Cognitive Computing*, *5*(4), 49.

Khormali, A., & Yuan, J.-S. (2022). DFDT: an end-to-end deepfake detection framework using vision transformer. *Applied Sciences*, *12*(6), 2953.

Kirn, H., Anwar, M., Sadiq, A., Zeeshan, H. M., Mehmood, I., & Butt, R. A. (2022). Deepfake tweets detection using deep learning algorithms. *Engineering Proceedings*, *20*(1), 2.

Lai, Z., Wang, Y., Feng, R., Hu, X., & Xu, H. (2022). Multi-feature fusion based deepfake face forgery video detection. *Systems*, *10*(2), 31.

Lee, E. G., Lee, I., & Yoo, S.-B. (2023). ClueCatcher: Catching domain-wise independent clues for deepfake detection. *Mathematics*, *11*(18), 3952.

Lewis, J. K., Toubal, I. E., Chen, H., Sandesera, V., Lomnitz, M., Hampel-Arias, Z., et al. (2020). Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning. In *2020 IEEE applied imagery pattern recognition workshop* (pp. 1–9). IEEE.

Li, X., Lang, Y., Chen, Y., Mao, X., He, Y., Wang, S., et al. (2020). Sharp multiple instance learning for DeepFake video detection. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 1864–1872). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3394171.3414034.

Li, M., Li, X., Yu, K., Deng, C., Huang, H., Mao, F., et al. (2023). Spatio-temporal catcher: A self-supervised transformer for deepfake video detection. In *Proceedings of the 31st ACM international conference on multimedia* (pp. 8707–8718).

Lin, K., Han, W., Li, S., Gu, Z., Zhao, H., & Mei, Y. (2023). Detecting deepfake videos using spatiotemporal trident network. *ACM Trans. Multimedia Comput. Commun. Appl.*, http://dx.doi.org/10.1145/3623639, Just Accepted.

Liu, J. J., Boongoen, T., & Iam-On, N. (2024). Improved detection of transient events in wide area sky survey using convolutional neural networks. *Data and Information Management*, *8*(3), Article 100035.

Liu, D., Dang, Z., Peng, C., Zheng, Y., Li, S., Wang, N., et al. (2023). FedForgery: generalized face forgery detection with residual federated learning. *IEEE Transactions on Information Forensics and Security*.

Liu, C., Li, J., Duan, J., & Huang, H. (2022). Video forgery detection using spatio-temporal dual transformer. In *Proceedings of the 2022 11th international conference on computing and pattern recognition* (pp. 273–281).

Liu, J., Zhu, K., Lu, W., Luo, X., & Zhao, X. (2021). A lightweight 3D convolutional neural network for deepfake detection. *International Journal of Intelligent Systems*, *36*(9), 4990–5004.

Lomnitz, M., Hampel-Arias, Z., Sandesara, V., & Hu, S. (2020). Multimodal approach for deepfake detection. In *2020 IEEE applied imagery pattern recognition workshop* (pp. 1–9). IEEE.

Lugstein, F., Baier, S., Bachinger, G., & Uhl, A. (2021). PRNU-based deepfake detection. In *Proceedings of the 2021 ACM workshop on information hiding and multimedia security* (pp. 7–12).

Maksutov, A. A., Morozov, V. O., Lavrenov, A. A., & Smirnov, A. S. (2020). Methods of deepfake detection based on machine learning. In *2020 IEEE conference of Russian Young researchers in electrical and electronic engineering (eIConRus)* (pp. 408–411). IEEE.

Malik, A., Kuribayashi, M., Abdullahi, S. M., & Khan, A. N. (2022). DeepFake detection for human face images and videos: A survey. *Ieee Access*, *10*, 18757–18775.

Mallet, J., Krueger, N., Dave, R., & Vanamala, M. (2023). Hybrid deepfake detection utilizing MLP and LSTM. In *2023 3rd international conference on electrical, computer, communications and mechatronics engineering* (pp. 1–5). IEEE.

Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *53*(4), 3974–4026.

Masud, U., Sadiq, M., Masood, S., Ahmad, M., & Abd El-Latif, A. A. (2023). LW-DeepFakeNet: a lightweight time distributed CNN-LSTM network for real-time DeepFake video detection. *Signal, Image and Video Processing*, *17*(8), 4029–4037.

Mcuba, M., Singh, A., Ikuesan, R. A., & Venter, H. (2023). The effect of deep learning methods on deepfake audio detection for digital investigation. *Procedia Computer Science*, *219*, 211–219.

Mehta, V., Gupta, P., Subramanian, R., & Dhall, A. (2021). Fakebuster: a deepfakes detection tool for video conferencing scenarios. In *Companion proceedings of the 26th international conference on intelligent user interfaces* (pp. 61–63).

Mira, F. (2023). Deep learning technique for recognition of deep fake videos. In *2023 IEEE IAS global conference on emerging technologies (globConET)* (pp. 1–4). IEEE.

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, *54*(1), 1–41.

Misirlis, N., & Munawar, H. B. (2023). From deepfake to deep useful: risks and opportunities through a systematic literature review. arXiv preprint arXiv:2311.15809.

Mitra, A., Mohanty, S. P., Corcoran, P., & Kougianos, E. (2020). A novel machine learning based method for deepfake video detection in social media. In *2020 IEEE international symposium on smart electronic systems (ISES)(formerly iNiS)* (pp. 91–96). IEEE.

Mitra, A., Mohanty, S. P., Corcoran, P., & Kougianos, E. (2021). A machine learning based approach for deepfake detection in social media through key video frame extraction. *SN Computer Science*, *2*(2), 98.

Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University-Computer and Information Sciences*, *35*(2), 757–774.

Muppalla, S., Jia, S., & Lyu, S. (2023). Integrating audio-visual features for multimodal deepfake detection. arXiv preprint arXiv:2310.03827.

Myvizhi, D., & Pamila, J. (2022). Extensive analysis of deep learning-based deepfake video detection. *Journal of Ubiquitous Computing and Communication Technologies*, *4*(1), 1–8.

Nadimpalli, A. V., & Rattani, A. (2023). ProActive DeepFake detection using GAN-based visible watermarking. *ACM Trans. Multimedia Comput. Commun. Appl.*, http://dx.doi.org/10.1145/3625547, Just Accepted.

Nguyen, X. H., Tran, T. S., Nguyen, K. D., Truong, D.-T., et al. (2021). Learning spatio-temporal features to detect manipulated facial videos created by the deepfake techniques. *Forensic Science International: Digital Investigation*, *36*, Article 301108.

Nirkin, Y., Wolf, L., Keller, Y., & Hassner, T. (2021). Deepfake detection based on discrepancies between faces and their context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 6111–6121.

Page, M., Moher, D., Bossuyt, P., Boutron, I., Hoffmann, T., Mulrow, C., et al. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, *372*, n160. http://dx.doi.org/10.1136/bmj.n160.

Passos, L. A., Jodas, D., Costa, K. A., Souza Júnior, L. A., Rodrigues, D., Del Ser, J., et al. (2024). A review of deep learning-based approaches for deepfake content detection. *Expert Systems*, *41*(8), Article e13570.

Pasupuleti, V. R., Tathireddy, P. R., Dontagani, G., & Rahim, S. A. (2023). Deepfake detection using custom densenet. In *2023 14th international conference on computing communication and networking technologies* (pp. 1–5). IEEE.

Pryor, L., Dave, R., Vanamala, M., et al. (2023). Deepfake detection analyzing hybrid dataset utilizing CNN and SVM. arXiv preprint arXiv:2302.10280.

Rahman, A., Siddique, N., Moon, M. J., Tasnim, T., Islam, M., Shahiduzzaman, M., et al. (2022). Short and low resolution deepfake video detection using cnn. In *2022 IEEE 10th region 10 humanitarian technology conference (r10-HTC)* (pp. 259–264). IEEE.

Rama chandran, S., Nadimpalli, A. V., & Rattani, A. (2021). An experimental evaluation on deepfake detection using deep face recognition. In *2021 international carnahan conference on security technology* (pp. 1–6). IEEE.

Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access*, *10*, 25494–25513.

Rana, M. S., & Sung, A. H. (2020). Deepfakestack: A deep ensemble-based learning technique for deepfake detection. In *2020 7th IEEE international conference on cyber security and cloud computing (cSCloud)/2020 6th IEEE international conference on edge computing and scalable cloud (edgeCom)* (pp. 70–75). IEEE.

Ranjan, P., Patil, S., & Kazi, F. (2020). Improved generalizability of deep-fakes detection using transfer learning based CNN framework. In *2020 3rd international conference on information and computer technologies* (pp. 86–90). IEEE.

Saikia, P., Dholaria, D., Yadav, P., Patel, V., & Roy, M. (2022). A hybrid CNN-LSTM model for video deepfake detection by leveraging optical flow features. In *2022 international joint conference on neural networks* (pp. 1–7). IEEE.

Sedaghatjoo, Z., Hosseinzadeh, H., & Bigham, B. S. (2024). Local binary pattern (LBP) optimization for feature extraction. arXiv preprint arXiv:2407.18665.

Sharma, V. K., Garg, R., & Caudron, Q. (2024). A systematic literature review on deepfake detection techniques. *Multimedia Tools and Applications*, 1–43.

Siegel, D., Kraetzer, C., Seidlitz, S., & Dittmann, J. (2021). Media forensics considerations on deepfake detection with hand-crafted features. *Journal of Imaging*, *7*(7), 108.

Singh, A., Saimbhi, A., Singh, N., & Mittal, M. (2020). DeepFake video detection: a time-distributed approach. SN comput sci 1: 212.

Stanciu, D. C., & Ionescu, B. (2022). Uncovering the strength of capsule networks in deepfake detection. In *Proceedings of the 1st international workshop on multimedia AI against disinformation* (pp. 69–77).

Stanciu, D.-C., & Ionescu, B. (2023). Autoencoder-based data augmentation for deepfake detection. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation* (pp. 19–27).

Stephen, D., & Mantoro, T. (2022). Usage of convolutional neural network for deepfake video detection with face-swapping technique. In *2022 5th international conference of computer and informatics engineering (IC2IE)* (pp. 22–28). IEEE.

Stroebel, L., Llewellyn, M., Hartley, T., Ip, T. S., & Ahmed, M. (2023). A systematic literature review on the effectiveness of deepfake detection techniques. *Journal of Cyber Security Technology*, *7*(2), 83–113.

Suratkar, S., Johnson, E., Variyambat, K., Panchal, M., & Kazi, F. (2020). Employing transfer-learning based CNN architectures to enhance the generalizability of deepfake detection. In *2020 11th international conference on computing, communication and networking technologies* (pp. 1–9). IEEE.

Suratkar, S., Kazi, F., Sakhalkar, M., Abhyankar, N., & Kshirsagar, M. (2020). Exposing deepfakes using convolutional neural networks and transfer learning approaches. In *2020 IEEE 17th India council international conference* (pp. 1–8). IEEE.

Taeb, M., & Chi, H. (2022). Comparison of deepfake detection techniques through deep learning. *Journal of Cybersecurity and Privacy*, *2*(1), 89–106.

Tang, L., Ye, Q., Hu, H., Xue, Q., Xiao, Y., & Li, J. (2024). DeepMark: A scalable and robust framework for DeepFake video detection. *ACM Trans. Priv. Secur.*, *27*(1), http://dx.doi.org/10.1145/3629976.

Tariq, S., Lee, S., & Woo, S. S. (2020). A convolutional lstm based residual network for deepfake video detection. arXiv preprint arXiv:2009.07480.

Vasist, P. N., & Krishnan, S. (2022). Deepfakes: An integrative review of the literature and an agenda for future research. *Communications of the Association for Information Systems*, *51*(1), 14.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Vinay, A., Bhat, N., Khurana, P. S., Lakshminarayanan, V., Nagesh, V., Natarajan, S., et al. (2022). Afmb-net: Deepfake detection network using heart rate analysis. *TehniČki Glasnik*, *16*(4), 503–508.

Wang, T., Cheng, H., Chow, K. P., & Nie, L. (2023). Deep convolutional pooling transformer for deepfake detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, *19*(6), 1–20.

Wang, Y., & Dantcheva, A. (2020). A video is worth more than 1000 lies. Comparing 3DCNN approaches for detecting deepfakes. In *2020 15Th IEEE international conference on automatic face and gesture recognition (FG 2020)* (pp. 515–519). IEEE.

Wang, G., Jiang, Q., Jin, X., & Cui, X. (2022). FFR_FD: Effective and fast detection of DeepFakes via feature point defects. *Information Sciences*, *596*, 472–488.

Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Jiang, Y.-G., et al. (2022). M2tr: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 international conference on multimedia retrieval* (pp. 615–623).

Whittaker, L., Mulcahy, R., Letheren, K., Kietzmann, J., & Russell-Bennett, R. (2023). Mapping the deepfake landscape for innovation: A multidisciplinary systematic review and future research agenda. *Technovation*, *125*, Article 102784.

Xiao, S., Zhang, Z., Yang, J., Wen, J., & Li, Y. (2023). Forgery detection by weighted complementarity between significant invariance and detail enhancement. *ACM Trans. Multimedia Comput. Commun. Appl.*, http://dx.doi.org/10.1145/3605893, Just Accepted.

Yadav, S., Bommareddy, S., & Vishwakarma, D. K. (2022). Robust and generalized Deep-Fake detection. In *2022 13th international conference on computing communication and networking technologies* (pp. 1–6). IEEE.

Yang, T., Chen, K., & Zhong, S. (2023). Deepfake detection using fusion channel information in a multi-attentional model. In *Proceedings of the 2023 Asia conference on artificial intelligence, machine learning and robotics* (pp. 1–5).

Younus, M. A., & Hasan, T. M. (2020). Effective and fast deepfake detection method based on haar wavelet transform. In *2020 international conference on computer science and software engineering* (pp. 186–190). IEEE.

Zhang, D., Li, C., Lin, F., Zeng, D., & Ge, S. (2021). Detecting deepfake videos with temporal dropout 3Dcnn.. In *IJCAI* (pp. 1288–1294).

Zhang, D., Lin, F., Hua, Y., Wang, P., Zeng, D., & Ge, S. (2022). Deepfake video detection with spatiotemporal dropout transformer. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 5833–5841). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3503161.3547913.

Zhang, Y., Lin, W., & Xu, J. (2024). Joint audio-visual attention with contrastive learning for more general deepfake detection. *ACM Trans. Multimedia Comput. Commun. Appl.*, *20*(5), http://dx.doi.org/10.1145/3625100.

Zhang, Y., Lu, J., Wang, X., Li, Z., Xiao, R., Wang, W., et al. (2022). Deepfake detection system for the ADD challenge track 3.2 based on score fusion. In *Proceedings of the 1st international workshop on deepfake detection for audio multimedia* (pp. 43–52). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3552466.3556528.

Zhang, D., Wu, P., Li, F., Zhu, W., & Sheng, V. S. (2022). Cascaded-hop for deepfake videos detection. *KSII Transactions on Internet and Information Systems (TIIS)*, *16*(5), 1671–1686.

Zhang, W., Zhao, C., & Li, Y. (2020). A novel counterfeit feature extraction technique for exposing face-swap images based on deep learning and error level analysis. *Entropy*, *22*(2), 249.

Zhao, Z., Wang, P., & Lu, W. (2020). Detecting deepfake video by learning two-level features with two-stream convolutional neural network. In *Proceedings of the 2020 6th international conference on computing and artificial intelligence* (pp. 291–297).

Zhao, L., Zhang, M., Ding, H., & Cui, X. (2021). MFF-net: Deepfake detection network based on multi-feature fusion. *Entropy*, *23*(12), 1692.

Zi, B., Chang, M., Chen, J., Ma, X., & Jiang, Y.-G. (2020). Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 2382–2390).