Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061427

REVIEW

A Contemporary and Comprehensive Bibliometric Exposition on Deepfake Research and Trends

Akanbi Bolakale AbdulQudus[1], Oluwatosin Ahmed Amodu[2,3,*], Umar Ali Bukar[4], Raja Azlina Raja Mahmood[2], Anies Faziehan Zakaria[5], Saki-Ogah Queen[6] and Zurina Mohd Hanapi[2]

[1]Department of Mathematics and Computer Science, Elizade University, Ilara-Mokin, 340271, Nigeria
[2]Department of Communication Technology and Network, Universiti Putra Malaysia (UPM), Serdang, 43400, Malaysia
[3]Information and Communication Engineering Department, Elizade University, Ilara-Mokin, 340271, Nigeria
[4]Department of Computer Science, Faculty of Computing and Artificial Intelligence, Taraba State University, ATC, Jalingo, 660213, Nigeria
[5]Department of Engineering Education, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Bangi, 43600, Selangor, Malaysia
[6]Department of Applied Modelling and Quantitative Methods, Trent University, Peterborough, ON K9L 0G2, Canada

*Corresponding Author: Oluwatosin Ahmed Amodu. Email: amodu_o_a@ieee.org

Received: 24 November 2024; Accepted: 07 April 2025; Published: 09 June 2025

ABSTRACT: This paper provides a comprehensive bibliometric exposition on deepfake research, exploring the intersection of artificial intelligence and deepfakes as well as international collaborations, prominent researchers, organizations, institutions, publications, and key themes. We performed a search on the Web of Science (WoS) database, focusing on Artificial Intelligence and Deepfakes, and filtered the results across 21 research areas, yielding 1412 articles. Using VOSviewer visualization tool, we analyzed this WoS data through keyword co-occurrence graphs, emphasizing on four prominent research themes. Compared with existing bibliometric papers on deepfakes, this paper proceeds to identify and discuss some of the highly cited papers within these themes: deepfake detection, feature extraction, face recognition, and forensics. The discussion highlights key challenges and advancements in deepfake research. Furthermore, this paper also discusses pressing issues surrounding deepfakes such as security, regulation, and datasets. We also provide an analysis of another exhaustive search on Scopus database focusing solely on Deepfakes (while not excluding AI) revealing deep learning as the predominant keyword, underscoring AI's central role in deepfake research. This comprehensive analysis, encompassing over 500 keywords from 8790 articles, uncovered a wide range of methods, implications, applications, concerns, requirements, challenges, models, tools, datasets, and modalities related to deepfakes. Finally, a discussion on recommendations for policymakers, researchers, and other stakeholders is also provided.

KEYWORDS: Deepfake; bibliometric; deepfake detection; deep learning; recommendations

1 Introduction

The development of Artificial Intelligence (AI) technology in recent years has raised serious questions and concerns in various sectors, including cybersecurity, politics, and media. Recently, the World Economic Forum's 2024 Global Risks Report [1] has announced AI-powered misinformation and disinformation as the most pressing short-term global threats (refer to Fig. 1). In particular, AI technology, namely deepfakes, contributes significantly to this phenomenon. Deepfakes enable the creation of highly realistic but fabricated

content, such as images, videos, and audio recordings. This technology has been widely exploited to create false information with the intent of deceiving or misleading, such as manipulating public opinion, damaging reputations, or spreading harmful propaganda.

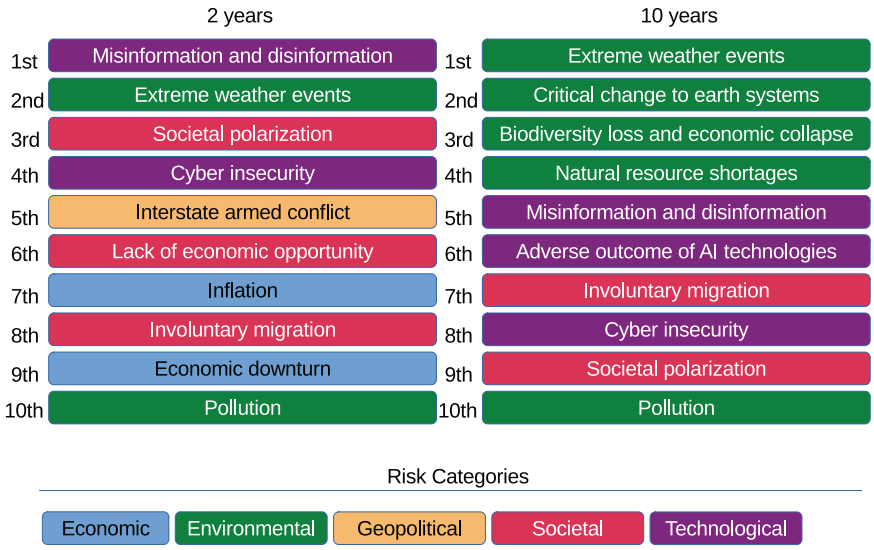

Figure 1: Top 10 risks by global risks report 2024

The word "deepfake" first appeared to the general public in 2017 when a member of the Reddit forum "deepfakes" started posting about the use of generative adversarial networks (GANs) [2] to manipulate videos of popular individuals in the society. Such algorithms can produce deceptively lifelike media, and stunning facial swaps [3]. Fig. 2 shows the number of research papers from Scopus within the last five years indicating an increasing trend in the number of publications on deepfakes.

Deepfake poses risks to individuals and organizations especially when it has been used with bad intentions, potentially damaging reputations, and causing societal harm. The potential for malicious use of deepfakes is significant, such as the creation of manipulated representations of public figures or other individuals, often leading to harm or reputation damage [4]. For instance, faces can be superimposed on explicit images, videos, or audio clips, and public figures can be fabricated as making harmful statements [5]. Besides disseminating false information, eroding confidence, and misdiagnosis, deepfakes can be used to perform cyber crimes such as fraud and security threats. They can undermine and destabilize the operations of a company via false claims [6] and are also harmful in their status as evidence which could make justice preservation quite difficult [7]. The prevalence of deepfakes makes it more challenging to filter fake news from real news, thus threatening security via the dissemination of propaganda [8]. Addressing this critical issue is paramount. This is evident from the increasing number of works in the deepfake detection research area. In particular, [9] provides statistical data from 2017 to 2024 demonstrating that research output in deepfake detection significantly exceeds that of deepfake creation. Although machine learning forms the foundation of most deepfake detection methods, challenges remain, such as the scarcity of high-quality datasets and benchmarks [10–13]. Notably, Convolutional neural networks (CNNs) have been identified as the widely used deep learning method for video deepfake detection [14].

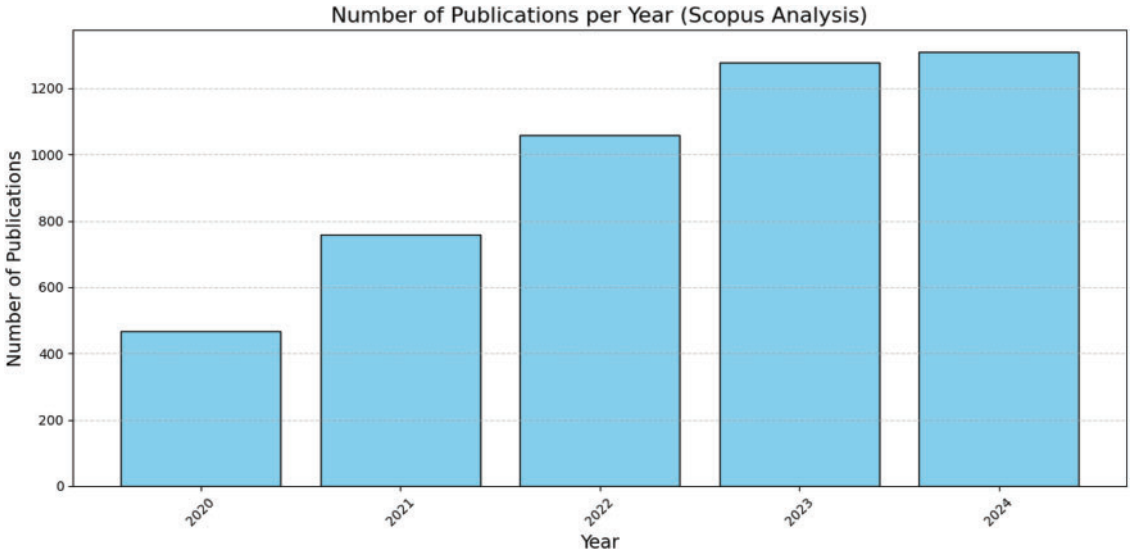

Figure 2: Published papers on deepfake within the last five years

Deepfakes can also serve benign purposes, such as enhancing photo quality for magazine covers, and may even be useful in education, fashion, marketing, and healthcare [8]. Other popular applications include interactive digital twins [15], and the deployment of digital avatars or virtual assistants within video conferencing environments [16]. Moreover, smartphone applications such as FaceApp and Facebrity, which leverage deepfake technology, have recently garnered significant public interest [17]. However, as previously discussed, the potential for malicious use of deepfakes appears to outweigh their beneficial applications, raising notable concerns for individuals, organizations, and national security.

Among ways to combat deepfakes include legislation and regulation, corporate policies and voluntary action, education, and training, as well as anti-deepfake technology. Such technologies include deepfake detection, content authentication, and deepfake prevention [8]. Enhanced detection methods help address deepfake threats by providing tools to verify content authenticity. Deepfake detection remains an active area of research, with ongoing developments aimed at improving accuracy and adapting to the evolving nature of deepfake technology [5]. Several deepfake detection strategies have been developed in response to growing concerns and garnered significant attention from specialists and academics in recent years. Deepfake detection involves several steps. The first step, data collection, involves gathering real and deepfake data for analysis. The second step, face detection, involves identifying facial regions to capture characteristics such as emotion, age, and gender. The third step, feature extraction, involves extracting distinguishing features from the face for deepfake identification. The fourth step, feature selection requires choosing the most relevant features for accurate detection. The fifth step, model selection, involves selecting a suitable model from deep learning, machine learning, or statistical approaches. The final step, model evaluation, involves assessing model performance using various metrics [18]. At the heart of these steps is feature extraction, feature selection, and model selection, where artificial intelligence plays a significant role.

Notably, to understand the depth of the literature, a useful technique for comprehending the dynamics of research output and impact across a range of topics is via bibliometric analysis [19]. Researchers can get insights that guide future research, funding choices, and policy creation by using quantitative tools to analyze academic literature. Usually, bibliometric analysis entails obtaining information from scholarly databases such as, Web of Science (WoS), and Scopus. Hence, a useful framework that can comprehend the

intricacies of deepfake research can be provided by bibliometric analysis, which is continuously evolving in this area [20,21] and various other fields [22–26]. Accordingly, a bibliometric analysis is applied to the study of scientific literature to quantify and assess research findings, patterns, and the composition of knowledge within particular fields [27].

Using bibliometric analysis, one can get insight into the evolution of research over time, pinpoint research trends, and highlight notable authors, journals, and institutions. It can also reveal the top-cited author contributions, institutions, and keyword co-occurrences. Therefore, this research aims to investigate current trends and developments in deepfake technology by analyzing publication and citation patterns, identifying key players (countries, organizations, and authors), exploring prominent themes and research interests, and identifying emerging trends within the field. This study focuses on addressing the following research questions:

- What are the distributions of publications on AI-based deepfakes geographically?
- What is the bibliographic coupling of researchers in the field of AI-based deepfakes?
- What are the most influential institutions working on AI-based deepfake research?
- What are the dominant trends from the meta-data (titles, abstracts, and keywords) on the research on AI-based deepfakes?
- Which research areas are the most prominent in the field of AI-based deepfakes and what are the top-cited papers in these areas?
- What lessons can be derived from these identified papers?
- What are the limitations, insights, and future prospects of deepfake detection research?
- What research patterns can be observed from the co-occurrence of keywords within the extensive body of deepfake research, based on a comprehensive analysis of these keywords?
- What are the trends, challenges and recommendations based on the review?
- What are the recommendations for addressing deepfakes for policymakers, researchers, and other stakeholders, such as industries and media outlets?

This paper stands out from previous bibliometric studies on deepfakes through:

- Identifying continental contributions to deepfake research and providing insights on data from WoS.
- Identifying the most prominent keywords by leading authors with the highest number of published documents on deepfakes.
- Identifying top research papers by leading authors with respect to citations based on data from WoS.
- Identifying and classifying key research areas based on the most prominent keywords into four themes: deepfake detection, feature extraction, face recognition, and forensics.
- Reviewing top cited papers that fall under these themes and their contributions.
- Discussing latest developments on global AI regulation initiatives.
- Discussing some of the trends, challenges and recommendations based on the review.
- Providing an exhaustive analysis of a more comprehensive search on deepfake based on data from Scopus.
- Providing insights based on an exhaustive keyword analysis from a comprehensive Scopus search.
- Discussing recommendations for policymakers, researchers, and practitioners.

These novel contributions provide unique insights into the rapidly advancing field of deepfakes. Accordingly, the remaining sections of the document are arranged as follows: Section 2 describes related literature on the bibliometric analysis of deepfakes, Section 3 provides details on the methodology, the results are described in Section 4, with insights into some prominent research areas, pressing concerns and key contributions are discussed in Section 5. Section 6 provides the results of the exhaustive search and analysis

of research on deepfakes. Section 7 provides recommendations for addressing deepfakes for policymakers, researchers, and practitioners.

2 Related Works

The interest in deepfake research is growing. Accordingly, we provide an overview of related works on the bibliometric analysis of deepfake research together with some of their findings.

2.1 Related Bibliometric Papers

In this section, we review the related work on deepfakes that have considered bibliometric analysis approaches to investigate trends in this area.

The work in [28] investigates misinformation in academia via network analysis of author keywords using bibliometric data. The results indicate that topics related to misinformation have increased in recent years. The work in [29] aims to select the most relevant articles on deepfakes based on data collected from Clarivate Analytics' Web of Science Core Collection. The authors show that within a period of six years (2018 to 2023), an annual growth rate of over 100% has been experienced, indicating the trends in this research area. Furthermore, the authors identify key authors, collaboration among authors, primary topics studied in research, and major keywords. In addition, the work provides potential techniques to stop the proliferation of deepfakes to ensure information trust.

In [20], using VOSviewer, the authors conduct a bibliometric analysis aimed at providing a comprehensive analysis of deepfakes and investigating influential authors and their collaboration, as well as countries and more specific institutions investing annually. Using Web of Science, they analyze top document types, source titles, publication trends, and the productivity of various countries, as well as collaborative efforts among institutions, authors, and regions. The authors also use CiteSpace to identify fundamental focal points, research directions, and shifts in citations for keywords, thus presenting an in-depth analysis.

In [30], the authors conduct a meta-analysis on deepfakes to visualize their evolution and related publications. They identify key authors, research institutions, and published papers using bibliometric data. In addition, the authors conduct a survey to test whether participants can differentiate real photos of people from fake AI-generated images. Although the study contains aspects of a bibliometric paper, it is considered a meta-research, survey, and background study. The findings of the study show that humans are falling short of keeping up with AI and must be conscious of its societal impact.

The study in [21] also aims to provide a bibliometric analysis of deepfake technology based on 217 entries spanning a range of 15 years from Scopus. The authors use VOSviewer and R-programming to perform the analysis, and the results indicate that India has the highest number of publications. There is also an emerging rise in publications on the issue of deepfakes. In addition, the authors provide insights into collaboration patterns, key contributors, and the evolving discourse, serving as a foundation for informed decision-making and further research.

The authors in [31] conduct a bibliometric analysis of articles published on deepfakes, focusing on six research questions related to the main research areas, current topics and their relationships, research trends, changes in research topics over time, contributors to deepfake research, and funders of deepfake research. Based on a study of 331 articles obtained from Scopus and Web of Science, the authors provide answers to these questions. Furthermore, they discuss emerging areas, potential development opportunities, applied methods, relationships among prominent researchers, countries conducting the research, and opportunities for practitioners interested in deepfake research.

Previous bibliometric studies have focused on specific aspects of deepfake research, such as fake news detection by Gunawan et al. [32], image anti-forensics by Lu et al. [33], and the negative effects of deepfake content by Garg and Gill [34]. However, these studies are limited in scope and do not provide a comprehensive overview of the field. For instance, Gunawan et al.'s research focuses on deepfake news detection, while Lu et al.'s study explores image anti-forensics. Garg and Gill's research, although focused on deepfake, primarily examines the negative effects of deepfake content.

Other studies have investigated related topics, such as disinformation through social media [35] and digital forensics investigation models by Ivanova and Stefanov [36]. However, these studies are restricted to specific keywords and do not provide a thorough analysis of the deepfake field. Gil et al.'s research on deepfake technology evolution and trends is based on bibliometric analysis but differentiates itself by focusing on organizations' funding deepfake research [31]. Kaushal et al.'s [20] study provides a comprehensive analysis of deepfake research but is limited to influential authors, countries, institutions, and publications.

2.2 Research Motivation

In conducting bibliometric analysis, a dataset must be acquired, typically through sources such as Web of Science (WoS) or Scopus, which have lots of bibliographic information [37], and analyzed using tools like VOSviewer, the R bibliometric package, or CiteSpace. Prior bibliometric analysis of deepfake research, as outlined in Table 1, reveals an expanding interest in the topic, but the current scope remains limited in several respects. Existing bibliometric analyses [20,21,29,31] provide valuable insights into publication trends and scholarly output, covering periods ranging from 3 to 15 years and document counts from 217 to 621. However, these analyses often lack coverage of larger datasets. This study, which analyzes 1412 documents from WoS, highlights the need for more comprehensive exploration due to the rapid development of deepfake technology. This evolution raises pressing ethical concerns, including its use in disinformation campaigns, privacy violations, and potential harm to individuals and organizations. Addressing these issues requires studies that go beyond detection and prevention to consider broader societal implications. Deepfake research remains an engaging field with few comprehensive review studies to provide insights and encourage further research. Conducting more extensive studies will support the development of policies and frameworks to address both technical and ethical challenges. The key features and analysis of existing bibliometric reviews, highlighting one of the gaps addressed by this study (dataset size and years of coverage), are presented in Table 1.

Table 1: Features and analysis of existing bibliometric reviews

Ref.	R	VOSviewer	CiteSpace	WoS	Scopus	Years coverage	No. document
[29]	✓			✓		6	584
[20]		✓	✓	✓		11	621
[21]	✓	✓			✓	15	217
[31]	✓			✓	✓	3	331

2.3 Research Contributions

Although prior works have provided different insights into countries, prominent authors, institutions, and keywords, this paper distinguishes itself from prior bibliometric studies in five ways: (1) the size of the dataset; (2) the classification of these areas into themes by grouping related concepts within a single thematic group wherever applicable; (3) the review of top-cited papers under these themes to identify research patterns

and some of the most influential research. Reviewing top-cited papers in each domain provides perspectives absent in other bibliometric papers and adds depth from an angle missing in prior works, with summaries indicating lessons learned. In addition to all these, the methodology deployed is also replicable and easy to follow, as papers selected for review are chosen based on well-defined criteria with strong relevance; (4) we provide a comprehensive analysis of a wide spectrum of keywords that were clustered using VOSviewer and we discuss the themes of each cluster. Moreover, insights into the state of deepfake research are provided from a corpus of over 8000 keywords; and (5) recommendations for addressing deepfakes for policymakers, researchers, and practitioners are provided. These contributions are unique to this paper and provide new insights into pivotal areas within the entire deepfake research domain.

This study aims to examine trends in publications and citations, countries' contributions, prominent authors, influential organizations, recurring themes, thematic elements, research interests, and emerging trends in the field of deepfake research using a distinct approach by conducting an in-depth analysis of prominent selected keywords, including detection, feature extraction, face recognition, and forensics. It then provides a review of the most cited papers in this domain, discussing some of their main contributions, motivations, and relevance. In addition, the dataset from Web of Science used in this paper is much larger than that of many existing works due to rapid advances in research in this area. Thus, many of the findings in this bibliometric analysis differ from those in prior work. Moreover, this study covers 21 research areas, with 1412 results from these areas, showing the large scope covered by the search. Furthermore, details are provided on the contributions of different continents and some of the main funding organizations in prominent countries, keywords associated with researchers with the most documents, top-cited papers by top-cited authors, and deep insights from 8790 keywords obtained from over 5,000 search entries in Scopus. The findings of this research provide valuable insights into the current state of deepfake research, identify research gaps, and offer recommendations for future studies. The study also sheds light on the negative effects of deepfake content and provides a foundation for developing strategies to mitigate these effects.

3 Research Methodology

This study utilizes a bibliometric analysis of research on deepfakes, employing VOSviewer to map and analyze the literature [38–40]. For all analyses conducted in this paper, we used the default VOSviewer settings unless stated otherwise, such as when adjusting the minimum keyword threshold. The VOSviewer provides visualization according to three bibliometric networks; a bibliographic coupling network of co-authorship (countries, researchers, and organization), a co-occurrence network of author keywords, and text analysis of title and abstract. A bibliometric network usually consists of both nodes and edges. The nodes could represent journals, publications, keywords, or researchers while the edges show the relationship between different pairs of nodes. Such relationships could be co-authorship or co-occurrence relations [40]. The primary goal is to elucidate research trends, identify influential contributors and countries, and explore critical themes in deepfake technology. The methodology outlines the data collection and visualization process.

Accordingly, the bibliometric information for this study was collected using the Web of Science, where each article's data corresponds to the theme. The research relied on comprehensive bibliographic databases, specifically Web of Science (WoS), due to their extensive coverage of peer-reviewed literature and citation information [37,41], as well as a source that favoured Natural Sciences and Engineering related disciplines [42]. As a result, this source is selected to capture a broad spectrum of foundational and recent deepfake technology studies. The search strategy involved querying terms such as "artificial intelligence" AND "deep fake" OR "deep-fake" OR "deepfake" OR "deep fakes" OR "deep-fakes" OR "deepfakes" OR "synthetic media" OR "AI-generated media". The search was conducted on August 29, 2024, and included

articles, conference papers, reviews, and proceedings, which revealed 1640 documents. The data was downloaded as a Tab delimited file from the Web of Science database. This study partially follows the PRISMA guidelines [43,44].

However, due to the limitations of PRISMA, which is a framework specifically designed for systematic and meta-analysis review [43,44], this study carefully selects the papers to meet the criteria of bibliometric review. Hence, the breakdown of the research process is presented in Fig. 3. Accordingly, the search retrieved results from various research areas, of which 21 research areas related to deepfakes were selected, yielding 1412 results, which are Computer Science: 1025; Engineering: 488; Imaging Science: 184; Telecommunications: 126; Government Law: 68; Science Technology: 52; Physics: 42; Information Science: 29; Mathematics: 27; Education Research: 22; Criminology Penology: 15; International Relations: 15; Film, Radio, Television: 13; Art: 12; Surgery: 5; Legal Medicine: 5; Theatre: 4; Medical Ethics: 3; Medical Informatics: 2; Obstetrics Gynecology: 2; Radiology, Nuclear Medicine, Imaging: 2. The choice of database, keywords and research areas as well as the use of VOSviewer for the presentation of data and visualizations helps to filter out outliers in the research on deepfakes, thus no other data cleaning process was required. Accordingly, Fig. 4 lists the top ten deepfake-related research areas in decreasing order and the number of papers from each research area.

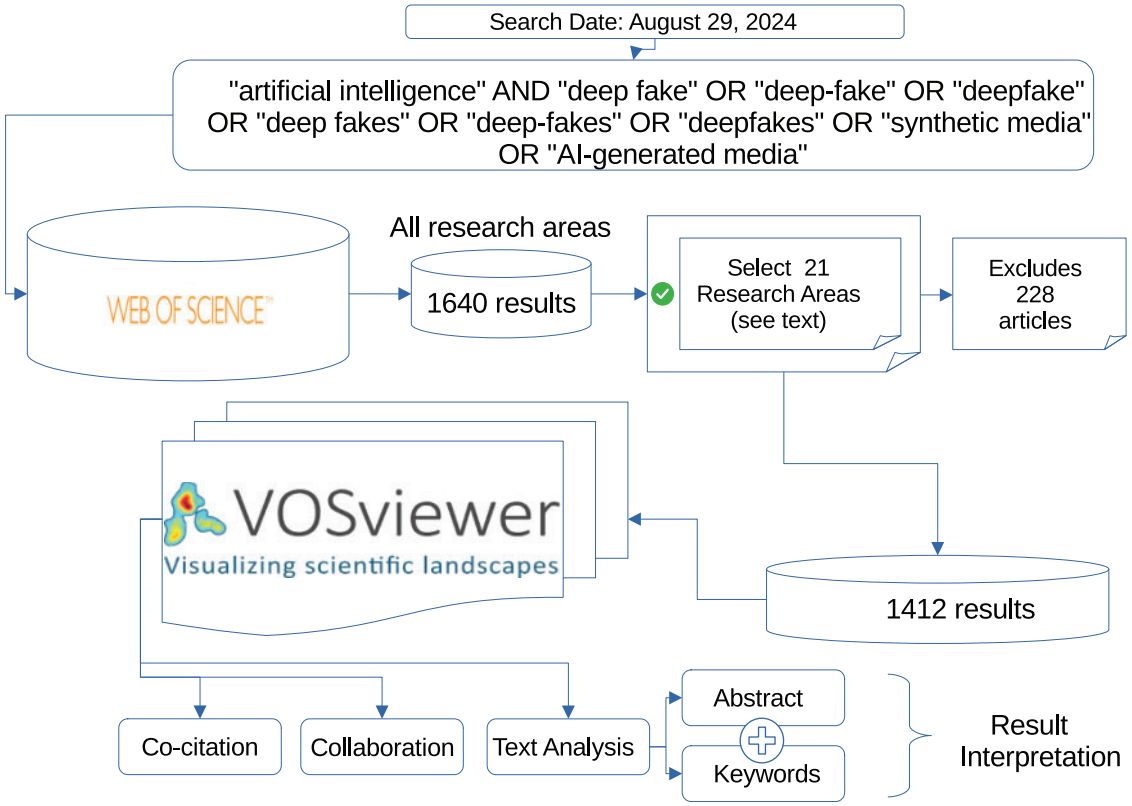

Figure 3: Research methodology

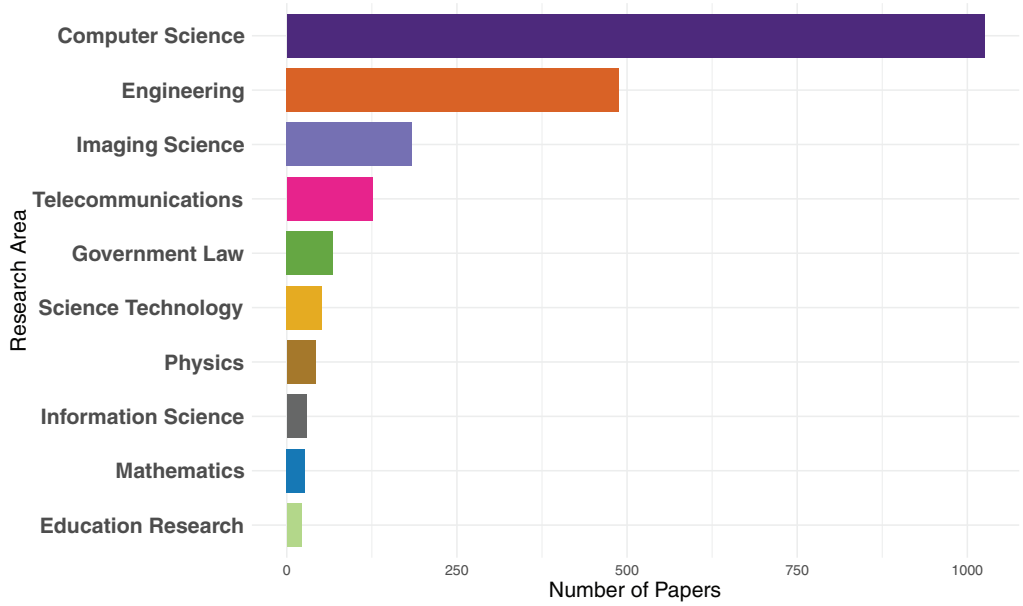

Figure 4: Top 10 deepfake-related research areas and their corresponding number of WoS indexed papers

Limiting the scope to the chosen 21 research areas helps to ensure documents not directly related to the technological or societal aspects of deepfakes were excluded. This involves 228 articles and the 21 research areas retrieved 1412 results, which were exported in tab-delimited file format for analysis. Bibliographic data for the 1412 publications was downloaded from the Web of Science database, which supports various file formats. These documents were exported in text format for analysis. The entire record was obtained for each publication.

4 Results

This section presents and analyzes the bibliographic data, focusing on key metrics such as the most prolific authors, leading countries in publication output, and other relevant trends, as observed in previous bibliometric studies [45–47]. By examining these aspects, this analysis provides a clearer view of the current research landscape, highlighting influential contributors and the regions driving advancements in deepfake research. Note that the influential contributions discussed in this section are based on the number of publications and citations.

4.1 Geographical Distribution of Publications (Citations by Country)

This study examines the geographical distribution of publications by using citations as the unit of analysis. A bibliographic map was generated based on collected data, utilizing bibliographic coupling of country co-authorship with fractional counting. The maximum number of countries per document was set to 25. To ensure a meaningful analysis, the minimum number of documents required for a country to be included in the citation analysis was set to five, which is the default value. Among the 89 countries in the dataset, 49 met this threshold. The final visualization is presented in Fig. 5.

Fig. 5 highlights regions and countries with at least five publications related to deepfake research. Each circle represents a country, where larger circles indicate higher publication counts, while smaller circles

represent countries with fewer publications. In general, the closer two countries appear in the visualization, the stronger their bibliographic coupling relationship.

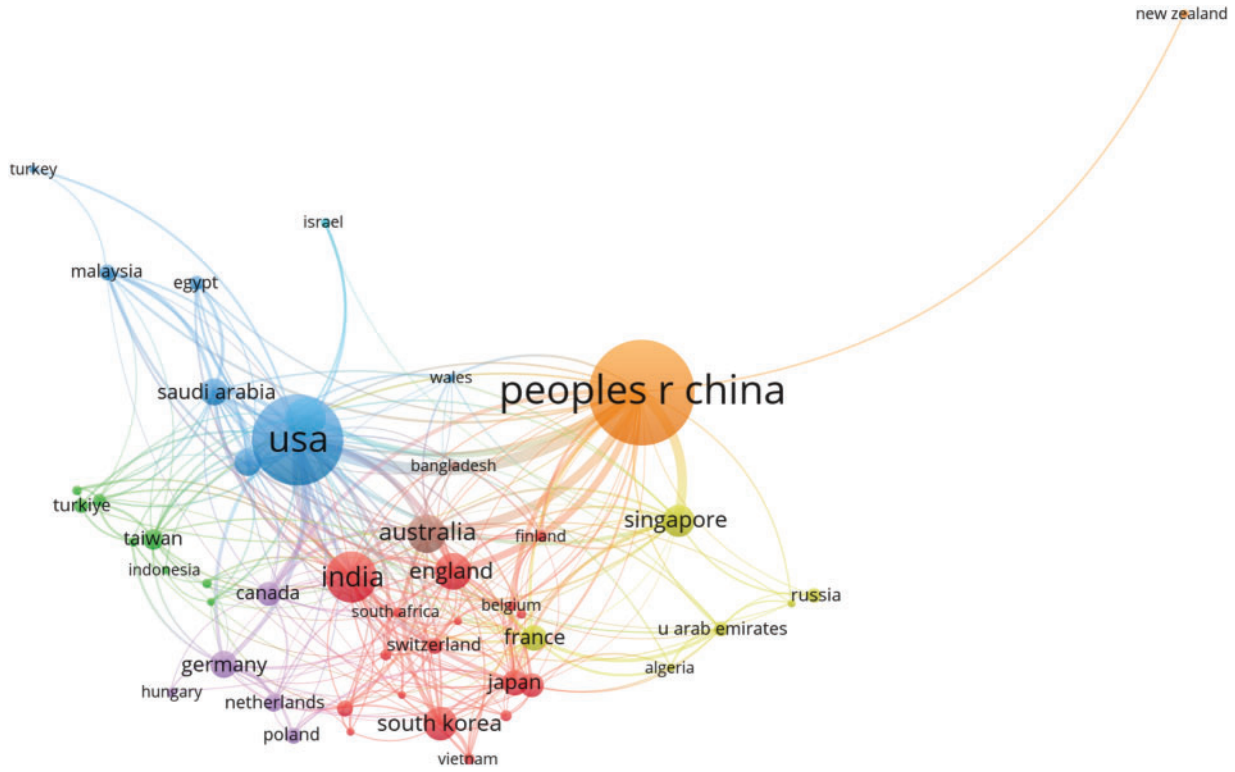

Figure 5: The visualization of the countries and regions with a minimum of a five publication threshold

4.2 Leading Countries Based on the Number of Publications

The research contributions of the top 20 countries, ranked by the number of published documents on deepfakes, are presented in Table 2(i). The minimum number of documents required for a country to be included in the visualization was set to 19.

The most prolific country in deepfake research is the People's Republic of China (PRC) or China, with 408 publications and 4403 citations, followed by the United States with 319 publications and 6259 citations, and India with 126 publications and 755 citations. This indicates that the China is the leading contributor to deepfake research, closely followed by the USA. At the lower end of the top 20 list, Norway and Switzerland each have 20 publications, while Malaysia, ranking 20th, has 19 publications.

The analysis also reveals that certain countries, such as Angola, Argentina, Bahrain, Bosnia & Herzegovina, Chile, Cyprus, Fiji, Kosovo, Lebanon, Libya, Morocco, Nepal, Somalia, Tunisia, Uzbekistan, Yemen, Ghana, and Sri Lanka, have contributed only one publication each. Similarly, Nigeria, Northern Ireland, Belarus, Colombia, Trinidad and Tobago, Estonia, and Iraq each have two publications, indicating relatively lower contributions to deepfake research.

This study finds that deepfake research is prioritized in countries such as China, the USA, India, and Italy, likely due to their strong focus on cybersecurity, emerging technologies, and digital innovation. Consequently, deepfake research is concentrated in industrialized nations with substantial public and private funding dedicated to AI and digital technologies. In contrast, countries with fewer publications in this area

may have limited access to funding and tend to focus on research addressing socio-economic priorities, such as public health, agriculture, or other pressing local concerns, rather than deepfake technology.

Table 2 presents the top 15 countries ranked by the number of publications and citations.

Table 2: Top 15 countries ranked by publications and citations

(i) By number of documents			(ii) By number of citations		
Country	**Docs**	**Cites**	**Country**	**Docs**	**Cites**
China	408	4403	U S A	319	6259
U S A	319	6259	China	408	4403
India	126	755	Italy	84	1920
Italy	84	1920	Germany	46	1432
Australia	79	590	Japan	40	1049
England	74	467	France	39	775
South Korea	63	617	India	126	755
Singapore	56	489	South Korea	63	617
Germany	46	1432	Australia	79	590
Pakistan	46	294	Singapore	56	489
Saudi Arabia	44	180	Spain	37	484
Japan	40	1049	England	74	467
France	39	775	Israel	8	335
Spain	37	484	Pakistan	46	294
Canada	37	242	U Arab Emirates	16	251

4.2.1 Leading Countries Based on Number of Citations

Considering the number of citations, the research contributions on deepfakes from the top 15 countries are presented in Table 2 (ii). The analysis was conducted by selecting a minimum of one document per country and including the top 15 countries out of 83. The USA is the most influential country in terms of citations, with 6259 citations from 319 publications, followed by the China with 4403 citations from 408 publications and Italy with 1920 citations from 84 publications. The data shows that the USA has significantly more citations on deepfakes than any other country. In contrast, countries such as Saudi Arabia, the Netherlands, and Egypt rank at the bottom of the list, with Saudi Arabia having 180 citations, the Netherlands 178 citations, and Egypt 111 citations. Moreover, the overall analysis indicates that countries such as Cyprus, Ghana, Iran, and Sri Lanka have no citations, while Slovenia, Bosnia & Herzegovina, Kosovo, and Luxembourg each have only one citation.

In summary, this section highlights the top countries contributing to deepfake research in terms of both document count and citation ranking. Specifically, China, the USA, and Italy rank in the top three for both categories (refer to Fig. 6), making them the leading contributors to deepfake research. Most countries that appear in the document ranking also appear in the citation ranking, with the exceptions of Israel, the United Arab Emirates, and Egypt. These countries are in the top 20 for document citations but not in the document ranking itself. Similarly, Norway, Switzerland, and Malaysia are in the top 20 for document rankings but not in the citation ranking.

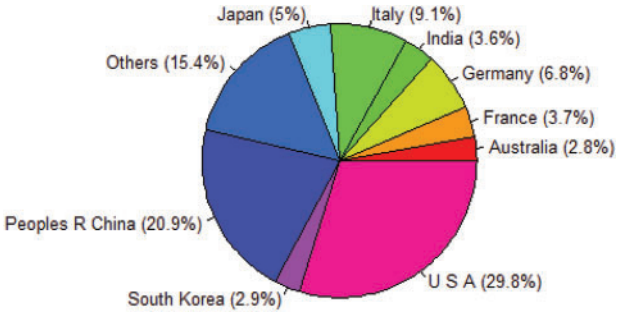

Figure 6: The top 10 countries by citations

4.2.2 Continental Insights and Recommendations

First, Asia leads in terms of the number of documents (783), with China contributing more than 52% of the total. North America follows, primarily represented by the USA (319). Europe ranks third (280), with Italy accounting for only 30% of all documents, indicating a more balanced contribution across multiple European countries. Oceania is represented solely by Australia, which has 79 documents—a significant number relative to some European countries with larger populations and more institutions. Africa's footprint is not observed in the analyzed data.

In terms of citations, North America, represented by the USA (6259), has the highest impact despite ranking second in the number of publications. Research in the USA is supported by funders such as the National Science Foundation, the Defense Advanced Research Projects Agency, and the U.S. Department of Defense. Asia follows, led by China (4403), though many other Asian countries have a lower citation-to-document ratio compared to Europe, where research, particularly from Italy and Germany, has a higher citation impact. Major funders in Europe include the European Commission and the Horizon 2020 Framework Programme. Oceania, represented by Australia (590), also contributes significantly. Africa and the Middle East do not have a notable presence in terms of citations.

Overall, Asia has the highest research volume with a strong citation count. Research in China benefits from funders such as the National Natural Science Foundation of China, the Ministry of Science and Technology of the People's Republic of China, and the National Key Research and Development Program of China. Other key funding agencies in Asia include the National Research Foundation of Korea. Meanwhile, North America (primarily the USA) produces the most impactful research overall, while Europe generates well-cited publications. Oceania also makes significant contributions, though its citation impact is lower compared to Europe and North America.

Given the low participation of some continents and countries in deepfake research, intercontinental collaboration should be encouraged. Deepfake technology is a global concern, as the internet is accessible to all. Collaboration between technologically advanced nations and developing regions would enhance the global research landscape on deepfakes, fostering more comprehensive and diverse contributions to this critical field.

4.3 Bibliographic Coupling Network of Researchers

In order to construct the visualization of researcher citations, we used bibliographic coupling based on co-authorship with fractional counting, setting a maximum of 25 countries per document. VOSviewer requires a minimum document count per country for inclusion in the citation visualization; we selected the default threshold of five publications. From our dataset, 105 authors met this criterion out of 3991 authors

with at least five publications. In the visualization shown in Fig. 7, each circle represents a researcher, with larger circles indicating researchers with many publications and smaller circles indicating those with fewer. Generally, the closer any two researchers are within the visualization, the more closely they are related in terms of bibliographic coupling. In other words, researchers positioned near each other tend to cite the same publications, whereas those further apart typically do not.

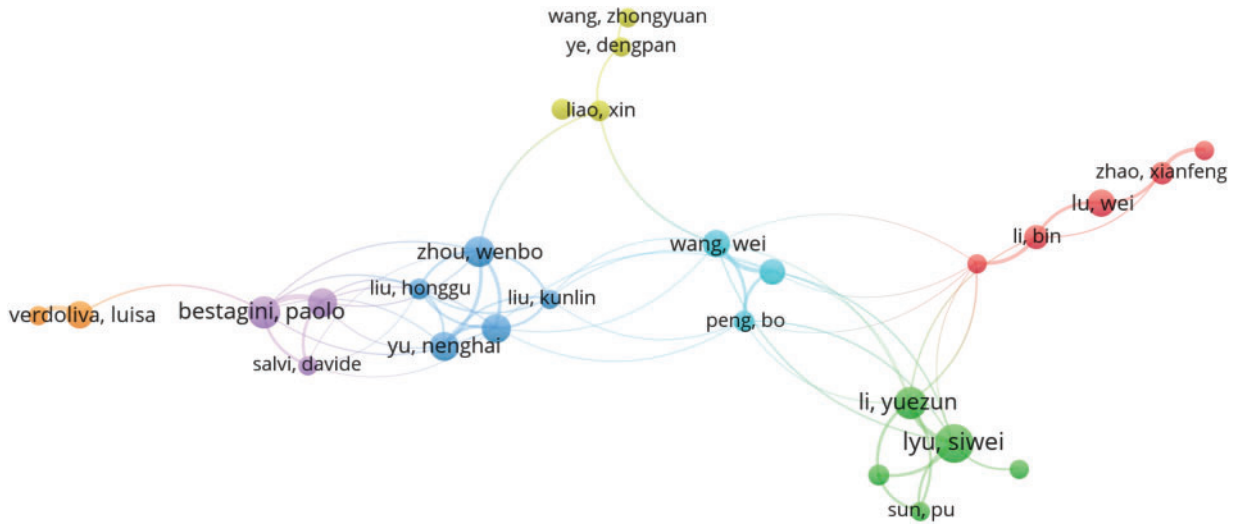

Figure 7: The visualization of the bibliographic coupling network of researchers

4.3.1 Leading Authors Based on Number of Documents

The research contributions on deepfakes by the top 20 authors, based on the number of publications, are presented in Table 3(i). This visualization was created by setting a minimum of one document per author. The most productive author is Lyu Siwei (USA), with 19 publications and 1541 citations, followed by Javed Ali (Pakistan), with 15 publications and 128 citations; and Woo Simon (South Korea), Bestagini Paolo (Italy), and Hu Yongjian (China), each with 14 publications and 209, 184, and 21 citations, respectively. The data shows that Lyu Siwei (USA) is the leading contributor to deepfake research, followed by Javed Ali (Pakistan). Authors such as Farid Hany (USA), Tariq Shahroz (Australia), Irtaza Aun (USA), Chen Yu (USA), Jin Xin (China), Jiang Qian (China), and Dong Jing (China) each have nine publications, placing them at the bottom of the top 20 list.

Table 3: Leading authors ranked by documents and citations

(i) By number of documents				(ii) By number of citations			
Author	**Docs**	**Cites**	**Country**	**Author**	**Docs**	**Cites**	**Country**
Lyu, Siwei	19	1541	USA	Lyu, Siwei	19	1541	USA
Javed, Ali	15	128	Pakistan	Li, Yuezun	13	1460	China
Woo, Simon S.	14	209	South Korea	Riess, Christian	3	1153	Germany
Bestagini, Paolo	14	184	Italy	Verdoliva, Luisa	10	1143	Italy
Hu, Yongjian	14	21	China	Yang, Xin	2	1097	USA
Li, Yuezun	13	1460	China	Cozzolino, Davide	5	860	Italy

(Continued)

Table 3 (continued)

(i) By number of documents				(ii) By number of citations			
Author	**Docs**	**Cites**	**Country**	**Author**	**Docs**	**Cites**	**Country**
Zhou, Wenbo	12	349	China	Niessner, Matthias	4	842	Germany
Yu, Nenghai	11	381	China	Thies, Justus	2	836	Germany
Zhang, Weiming	11	362	China	Roessler, Andreas	2	836	Germany
Tubaro, Stefano	11	181	Italy	Qi, Honggang	6	648	China
Liu, Beibei	11	19	China	Sun, Pu	5	647	China
Verdoliva, Luisa	10	1143	Italy	Yamagishi, Junichi	8	603	Japan
Amerini, Irene	10	209	Italy	Echizen, Isao	6	576	Japan
Wang, Wei	10	114	China	Wen, Fang	3	572	China
Lu, Wei	10	72	China	Chen, Dong	3	572	China
Zhao, Yao	10	34	China	Bao, Jianmin	3	572	China
Farid, Hany	9	254	USA	Yang, Hao	2	531	USA
Tariq, Shahroz	9	139	Australia	Li, Lingzhi	2	531	China
Irtaza, Aun	9	92	USA	Zang, Ting	2	451	China
Chen, Yu	9	32	USA	Guo, Baining	2	451	China

Additionally, the research indicates that nine of the top 20 authors are from China, four from the USA, four from Italy, and one each from Pakistan, South Korea, and Australia. The significant number of authors from China underscores their substantial contribution to deepfake research. Overall, over 200 authors have only one publication. Table 3 presents the top 20 authors, ranked by the number of publications, their citation counts, and countries.

4.3.2 Related Keywords by Authors with the Highest Number Documents

In this study, we aim to identify the research patterns represented by keywords in the works published by authors with the most documents. These keywords are mainly related to methods and techniques, as well as broader concepts and components on deepfake generation, image analysis, deepfake, and forgery detection. Similarly, text, audio, and video forgery are all evident in these keywords. These keywords include Adversarial Learning, Adversarial Networks, Audio Authenticities, Audio Forgery Detection, Data Hiding, Deepfake Detection, Deep Neural Networks, Detection Methods, Detection Models, Digital Image Forensics, Duplication Detection, Face Images, Face Recognition, Face Synthesis, Face Swapping, Facial Expressions, Facial Landmark, Fake Detection, Forgery Detections, Gait Analysis, Gait Recognition, Generalization Capability, Generative Adversarial Networks, Image Analysis, Image Classification, Image Compression, Image Enhancement, Image Features, Image Forensics, Image Matching, Image Processing, Manipulation Techniques, Media Forensics, Neural Network, Neural Networks, Object Detection, Reversible Data Hiding, Reversible Watermarking, Speech Recognition, Synthetic Data, Video Forgery Detection, Voice Replay Attack, Watermark Embedding.

4.3.3 Leading Authors Based on Number of Citations

Regarding citation count, the research contributions on deepfakes by the top 20 authors are presented in Table 3(ii). The most cited author is Lyu Siwei (USA), with 19 publications and 1541 citations, followed by Li Yuezun (China), with 13 publications and 1460 citations, and Riess Christian (Germany), with three

publications and 1153 citations. The data indicates that Lyu Siwei (USA) has the highest citation count in deepfake research, followed by Li Yuezun (China).

Authors such as Li Lingzhi (China), Yang Hao (USA), Zhang Ting (China), and Guo Baining (China) are at the lower end of the top 20 list. Li Lingzhi and Yang Hao each have 531 citations, while Zhang Ting and Guo Baining have 41 citations each across two publications.

Among the top 20 most cited authors, nine are from China, four from Germany, three from the USA, and two each from Italy and Japan. The data further indicates that over 100 authors have no citations. Lyu Siwei's leading citation count suggests that his work is highly influential and widely recognized. He is also the only author in the top three to rank highly in both publication and citation counts. Additionally, the presence of nine Chinese authors in the top 20 underscores China's significant contribution to deepfake research. Table 3 presents the top 20 authors ranked by citation count, along with their publication numbers and countries.

4.3.4 Top-Cited Papers by the Most Cited Authors

This section briefly explores the two most cited papers by the five most cited authors in deepfake research, each with over 1000 citations. Notably, these two papers collectively involve contributions from Lyu Siwei, Li Yuezun, Riess Christian, Verdoliva Luisa, and Yang Xin. Both papers highlight the importance of high-quality datasets and benchmarks for deepfake research.

The first paper, titled *"Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics"* [10], co-authored by Li, Xin Yang, and **Siwei Lyu**, along with Pu Sun and Honggang Qi, identifies a major limitation in existing deepfake datasets—their low visual quality, which makes them unrealistic compared to deepfake videos circulated online. To address this issue, the authors introduced a new large-scale deepfake video dataset containing 5639 high-quality videos featuring celebrities, generated using an improved synthesis process. A comprehensive evaluation of deepfake detection methods using this dataset demonstrates its challenges and potential impact on deepfake forensics.

The second paper, titled *"FaceForensics++: Learning to Detect Manipulated Facial Images"* [48], authored by Rossler and co-authors, including Riess Christian and Verdoliva Luisa, proposes an automated, publicly available benchmark for facial manipulation detection. This benchmark standardizes the evaluation of deepfake detection methods by incorporating prominent manipulation techniques at varying compression levels and sizes. The dataset contains over 1.8 million manipulated images, making it significantly larger than previous datasets. A thorough analysis of data-forgery detection techniques reveals that incorporating domain-specific knowledge significantly improves detection accuracy, even under strong compression, and outperforms human observers.

4.4 Most Influential Institutions

The most influential institutions are analyzed based on two criteria: the highest number of publications and the highest number of citations. The results of this analysis are presented in the following sections.

4.4.1 Influential Institutions Based on Number of Documents

An analysis of the most influential institutions reveals that the leading contributor to deepfake research is the Chinese Academy of Sciences, with 45 publications and 487 citations, followed by the University of the Chinese Academy of Sciences, with 37 publications and 984 citations, and Nanyang Technological University, with 28 publications and 393 citations (see Table 4(i)). The institution abbreviations are presented in the table as extracted from VOSviewer.

Table 4: Influential institutions ranked by documents and citations

(i) By number of documents			(ii) By number of citations		
Institution	**Docs**	**Cites**	**Institution**	**Docs**	**Cites**
Chinese Acad Sci	45	487	Suny Albany	9	1495
Univ Chinese Acad Sci	37	984	Univ Federico II Naples	9	1143
Nanyang Technol Univ	28	393	Univ Chinese Acad Sci	37	984
Wuhan Univ	27	170	Tech Univ Munich	7	849
Univ Sci & Technol China	22	452	Univ Erlangen Nurnberg	1	791
Sun Yat Sen Univ	20	122	Natl Inst Informat	11	672
Natl Univ Singapore	16	61	Microsoft Res Asia	4	590
South China Univ Technol	16	28	Peking Univ	4	575
Sungkyunkwan Univ	16	230	Ecole Ponts Paristech	1	561
Alibaba Grp	15	410	Jfli	1	561
Nanjing Univ Informat Sci & Tech	15	211	Upem	1	561
Shanghai Jiao Tong Univ	15	201	Chinese Acad Sci	45	487
Zhejiang Univ	15	113	Univ Sci & Technol China	22	452
Deakin Univ	14	207	Yale Informat Soc Project	3	427
Hunan Univ	14	237	Purdue Univ	5	423
Politecn Milan	14	184	Alibaba Grp	15	410
Shenzhen Univ	14	60	Nanyang Technol Univ	28	393
Beijing Jiaotong Univ	13	50	Univ Texas Austin	2	364
Univ Calif Berkeley	13	285	Stanford Ctr Internet & Soc	2	351
Natl Inst Informat	11	672	Microsoft Cloud AI	3	334

The data highlights the Chinese Academy of Sciences as the primary contributor to deepfake research, closely followed by the University of the Chinese Academy of Sciences. Among the top 20 institutions, the minimum publication count is 11, with four institutions meeting this threshold: Xi'an University (11 papers, 167 citations), the National Institute of Informatics (11 papers, 672 citations), SUNY Buffalo (11 papers, 93 citations), and Monash University (11 papers, 145 citations). The National Institute of Informatics ranks 20th due to its higher citation count.

Overall, more than 150 institutions have only one publication. Table 4 presents the top 20 institutions ranked by the number of documents, along with their citation counts. Additionally, some of the top 20 institutions in the network are not directly connected. The largest connected cluster consists of 18 institutions, as shown in Fig. 8.

4.4.2 Influential Institutions Based on Number of Citations

This study also examines the most influential institutions based on citation count, complementing the analysis conducted on the number of publications per institution. The results reveal that the most influential institution is SUNY Albany, with nine publications and 1495 citations, followed by the University of Federico II Naples, with nine publications and 1143 citations, and the University of the Chinese Academy of Sciences, with 37 publications and 984 citations, as shown in Table 4 (ii).

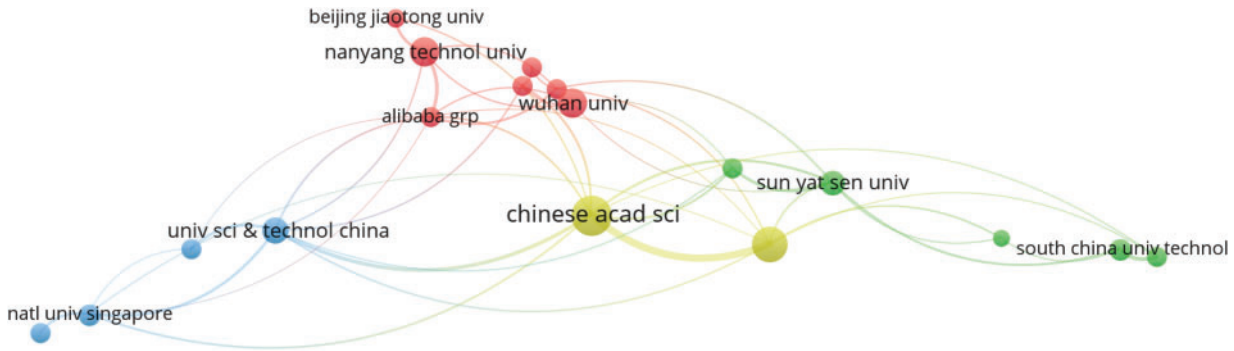

Figure 8: Visualization of connected organizations using the number of documents to rank (refer to Table 4 for details)

The findings indicate that SUNY Albany is the leading contributor to deepfake research in terms of citations, followed by the University of Federico II Naples. Notably, Dr. Siwei Lyu, the most highly cited researcher in this field, is affiliated with SUNY Albany.

Among the top 20 institutions, the lowest publication count is three, which includes Microsoft Cloud AI, with 334 citations. Additionally, the analysis reveals that more than 150 institutions have no citations. Table 4 presents the top 20 institutions ranked by citation count, along with their respective publication numbers.

It is also important to note that some of the top 20 institutions in the network are not directly connected. The largest connected cluster consists of 14 institutions, as shown in Fig. 9.

Figure 9: Visualization of connected organizations using the number of citations to rank (refer to Table 4 for details)

4.4.3 Relevance of Higher Document or Citation: Institutional

The analysis of influential institutions based on the number of publications and citations provides valuable insights into the research landscape of deepfakes. However, it is important to recognize that the number of publications produced by an institution does not necessarily correlate with the number of citations it receives. This discrepancy highlights the distinction between the quantity of research output and its quality or impact within the scientific community.

For instance, while institutions like the University of the Chinese Academy of Sciences have a high number of publications (37), it is institutions such as SUNY Albany that lead in citations (1495 citations from just 9 publications). This suggests that although SUNY Albany has fewer publications, its research has a greater impact, receiving significant attention and citations from other scholars. Conversely, an institution with a higher publication count may not necessarily receive a proportional number of citations, as seen with Microsoft Cloud AI (3 publications, 334 citations), which, despite its smaller output, has a relatively high citation rate.

This comparison underscores the importance of not relying solely on the number of publications as a measure of an institution's research influence. Citations often provide a more accurate reflection of the quality,

relevance, and impact of research, as they indicate how frequently other researchers reference and build upon that work. Therefore, institutions like SUNY Albany, with fewer but highly cited papers, may contribute more significantly to the field than institutions with a larger number of publications but fewer citations.

4.5 Co-occurrence Network of Keywords

In this analysis, we present a visualization of the keywords used by authors. Specifically, we use bibliographic coupling to analyze the co-occurrence of author keywords, applying the fractional counting option. Out of 2809 keywords, we set the minimum occurrence threshold at 25, resulting in a selection of 23 keywords. In the visualization, each circle represents a keyword, with closer proximity indicating a stronger relationship between keywords. The co-occurrence of keywords in publications was analyzed to determine their interconnectedness. The results show that the primary keyword is "deepfake detection," with prominent related keywords including deepfake, deep learning, artificial intelligence, feature extraction, machine learning, faces, and generative adversarial networks. Additionally, synonyms such as deepfake, deepfakes, and deepfake are collectively represented as "Deepfake" in the visualization, as shown in Fig. 10.

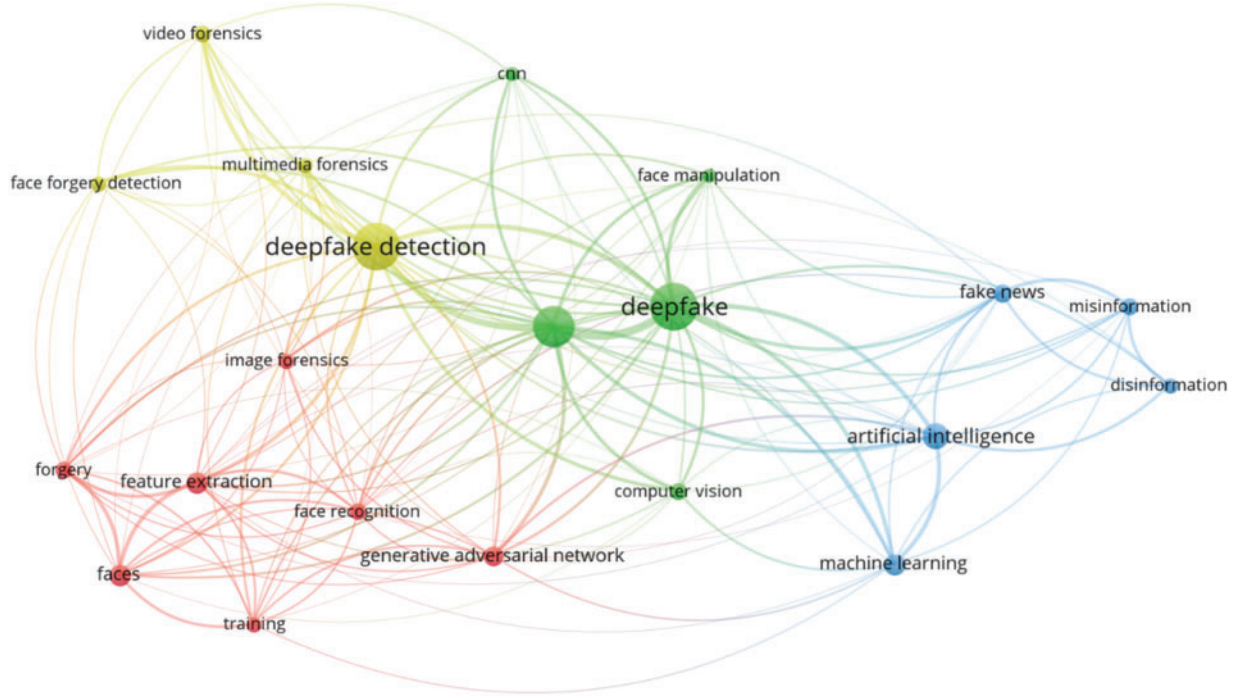

Figure 10: Visualization of author keywords

In the visualization (see Fig. 10), four clusters are identified. The first cluster, shown in red, focuses on technologies and processes related to deepfake creation and detection. This cluster contains seven keywords, making it the largest, which suggests that researchers in this field prioritize this aspect more than others. Clusters 2 and 3 each contain five keywords. Cluster 2, represented in green, is centered on deepfake generation techniques and technologies, while Cluster 3, in blue, relates to the impact of deepfakes on information integrity and misinformation. Cluster 4, shown in yellow, consists of four keywords and is associated with deepfake detection and forensic analysis, offering opportunities for further research contributions. Cluster 3 has the highest number of connections in the visualization, linking it to most of the other keywords.

The emphasis on deepfake detection in research suggests that deepfakes are becoming increasingly common, necessitating the development of effective detection systems. Researchers are actively working to improve detection methods capable of accurately identifying deepfake content. Given that deepfakes can be highly realistic, they pose risks such as misinformation, security threats, and reputation damage to individuals and organizations. Consequently, many researchers focus on developing deepfake detection solutions. Table 5 presents the 23 keywords identified (including variant forms such as 'Deepfake', 'Deepfakes', 'deepfake'), along with their occurrence counts based on the analysis.

Table 5: Most frequent occurrence of authors keywords

Keyword	Occurrences	Keyword	Occurrences
Deepfake Detection	276	Computer Vision	33
Deepfake	267	Face Forgery Detection	32
Deep Learning	198	Multimedia Forensics	29
Deepfakes	194	Training	29
Artificial Intelligence	87	Image Forensics	28
Feature Extraction	55	Disinformation	27
Machine Learning	54	CNN	26
Faces	53	Face Manipulation	26
Generative Adversarial Networks	46	deepfake	25
Fake News	43		
Forgery	38		
Face Recognition	36		
Misinformation	34		
Video Forensics	34		

4.6 Text Analysis of Titles and Abstracts

This study also performs a text analysis of titles and abstracts to identify common research themes and focus areas. The minimum number of occurrences for a term is set at 160 out of 25,799 items, with 29 terms meeting this threshold and being considered for visualization. According to the extracted data, *model*, most likely referring to detection models, is mentioned frequently in the literature, followed by other commonly used terms such as *image*, *deepfake*, *video*, *dataset*, *detection*, *feature*, *technique*, and *face*.

The network visualization presented in Fig. 11 reveals four clusters. The first cluster, shown in red, focuses on evaluating the accuracy and effectiveness of deepfake detection. This cluster contains 14 keywords, making it the largest, which suggests that researchers in this field place significant emphasis on assessing detection performance. Cluster 2, shown in green, consists of 11 keywords and is centered on advancements in deepfake technology. Clusters 3 and 4 each contain two keywords. Cluster 3, shown in blue, focuses on techniques and deep learning algorithms used in deepfake research, while Cluster 4, shown in yellow, pertains to *deepfake* and *authentic videos*. Notably, Cluster 3 has the most connections in the visualization, as it is closely linked to other clusters, with *deepfake* serving as a key term connected to most other keywords.

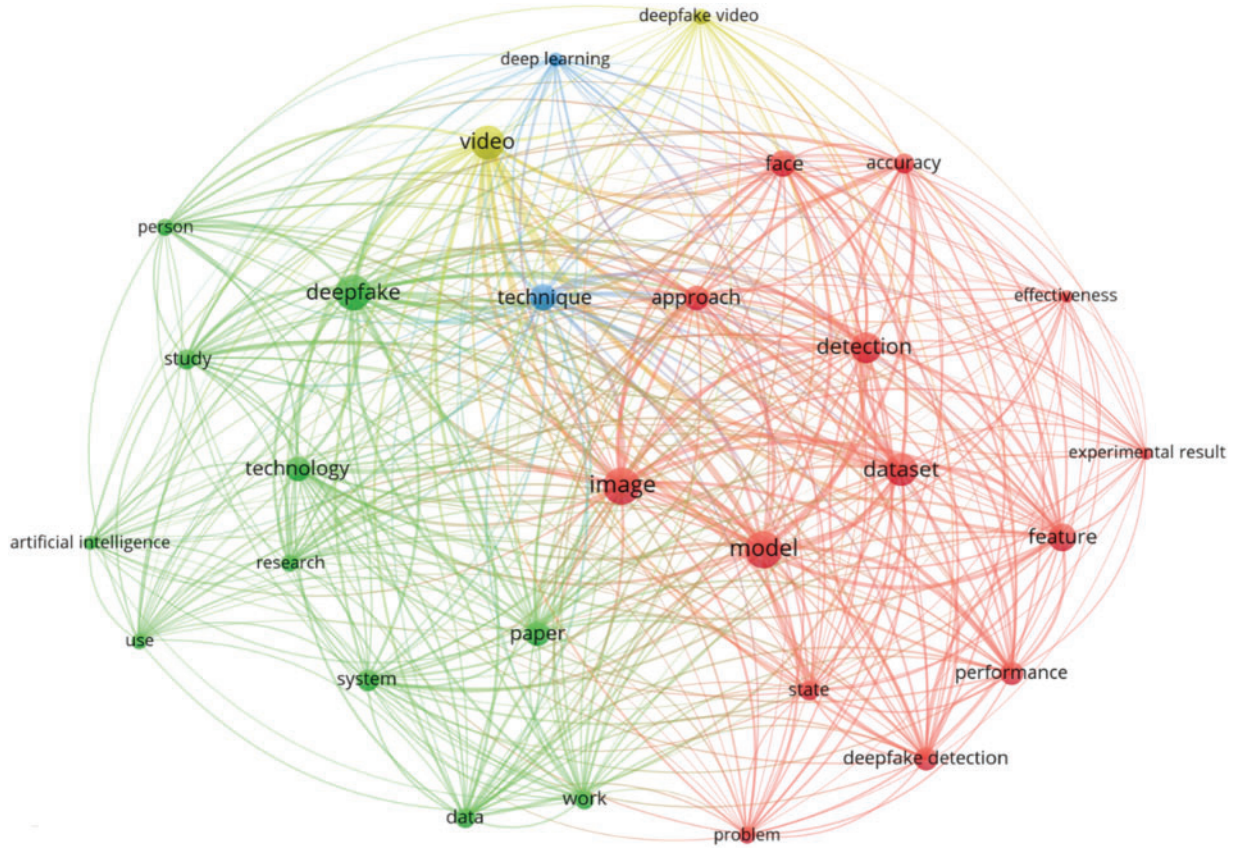

Figure 11: Network visualization of text analysis using the title and abstract considering full counting

According to the data, seven of the 29 identified terms represent different aspects of deepfake research: *deepfake detection*, *models*, *features*, *accuracy*, *effectiveness*, *fields*, and *studies*. These terms highlight the central focus on deepfake detection and the associated challenges. Based on these findings, several key observations can be made:

- Deepfake detection is applied to manipulated *images*, *audio*, or *videos*, where *faces* can be easily altered.
- The rapid advancements in *GANs* have made the creation of deepfakes more accessible.
- Extensive research has been dedicated to improving *models* capable of detecting deepfake media.
- Researchers are examining deepfake *features*, such as inconsistencies in facial movements or lighting, to identify fake content.
- Efforts are being made to assess the *effectiveness* of detection systems in accurately identifying deepfakes.
- Studies explore the challenges and solutions related to deepfake technology, with significant contributions focused on enhancing detection methods.

Table 6 presents the frequency of keyword occurrences based on titles and abstracts.

Table 6: Occurrence of keywords in title and abstract

Keyword	Occurrences	Keyword	Occurrences	Keyword	Occurrences
Model	1523	Image	1505	deepfake	1419
Video	1291	Dataset	1130	Detection	1007

(Continued)

Table 6 (continued)

Keyword	Occurrences	Keyword	Occurrences	Keyword	Occurrences
Feature	888	Technique	756	Face	750
Approach	712	Paper	647	Technology	643
Performance	592	System	513	Study	484
Data	474	Accuracy	467	State	447
Work	385	Research	347	Problem	335
Person	325	deepfake Video	283	Use	255
Deep Learning	242	Artificial Intelligence	215	Effectiveness	192
Experimental Result	183				

5 Discussion

The findings of this study further highlight the importance of deepfake detection. Therefore, we take a closer look at the keywords associated with deepfake detection. In this analysis, we use the following keyword query: (detect OR detection OR detecting (Title) AND "Deep fake" OR deepfake OR deep-fake OR "Deep fakes" OR deepfakes OR deep-fakes OR "face forgery" OR "face-forgery" OR "face manipulation" OR "face-manipulation" (Topic)), retrieved from the Web of Science (WoS) database on 1 November, 2024. The results are summarized in Table 7, which clearly demonstrates that deepfakes, deepfake detection, and related processes and analyses constitute some of the most frequently used keywords. These include *deepfake(s), face recognition, face forgery detection, face manipulation, detectors, feature fusion, forgery detection, deepfake video detection, multimedia forensics, image forensics, face forensics, media forensics, forensics, face swapping, face forgery, face manipulation detection, face swap, and audio deepfakes*.

Table 7: Keywords and occurrences For deepfake detection (1/11/2024)

Keyword	Occurrences	Keyword	Occurrences
deepfake detection	235	detection	20
deep learning	128	machine learning	20
deepfakes	88	computer vision	19
deepfake	131	information integrity	7
feature extraction	39	transformer	13
faces	36	artificial intelligence	16
forgery	33	detectors	7
face recognition	21	generalization	12
face forgery detection	59	fake news	11
visualization	14	data augmentation	10
generative adversarial networks	23	multimedia forensics	18
transformers	15	image forensics	11
face manipulation	21	feature fusion	12
training	12	attention mechanism	16
task analysis	10	contrastive learning	12

(Continued)

Table 7 (continued)

Keyword	Occurrences	Keyword	Occurrences
convolutional neural networks	16	databases	5
video forensics	25	forgery detection	12
convolutional neural network	18	neural networks	15
videos	9	deepfake video detection	25
cnn	21	self-supervised learning	6
frequency-domain analysis	8	robustness	7
gan	15	face forensics	9
vision transformer	19	media forensics	8
face swapping	5	transfer learning	12
face forgery	6	face manipulation detection	8
forensics	8	accuracy	5
cybersecurity	6	domain generalization	8
face swap	8	self-attention	5
audio deepfakes	7	convolutional neural network (cnn)	5
deep fake detection	7	deepfakes detection	8
metric learning	5	wavelet transform	5
adversarial attacks	6	deepfake dataset	5
optical flow	6	anti-spoofing	6

Additionally, machine learning models, modeling approaches, and detection techniques are commonly referenced. In this context, *generative adversarial networks (GANs)* emerge as the most prevalent, followed by *transformers, artificial intelligence, neural networks*, and *convolutional neural networks (CNNs)*. Other relevant techniques include the *attention mechanism, contrastive learning, self-supervised learning*, and *transfer learning*.

Furthermore, several desired features of deepfake detection solutions are evident, including *generalization, robustness*, and *accuracy*. The importance of security is also reflected in the presence of keywords such as *cybersecurity, adversarial attacks*, and *anti-spoofing*. Ethical concerns surrounding deepfakes are also noticeable, with terms such as *forgery, information integrity*, and *fake news* appearing frequently.

Similarly, the significance of databases and deepfake datasets is evident from keywords like *database* and *deepfake dataset*. Finally, various modeling techniques and analytical mechanisms are represented by keywords such as *task analysis, training, feature extraction, feature fusion, frequency-domain analysis, metric learning, contrastive learning, wavelet transform*, and optical flow.

5.1 Prominent Areas and Key Contributions

In this section, we discuss some of the prominent areas and key contributions to deepfake research. Prior to that, a background on how these papers were selected is provided. First, we identified four themes based on the visualization produced by VOSviewer of keywords in Fig. 10. This categorization of major themes is presented in Fig. 12. In addition, we conducted a new search on WoS for each of these categories as provided in Table 8. Note that the results in WoS are sensitive to plural, and hence, the plurals for deepfake have also been included. Based on the above finding, we provide a discussion of the top-cited papers and top contributions in this area based on published papers between 2021 and 2024.

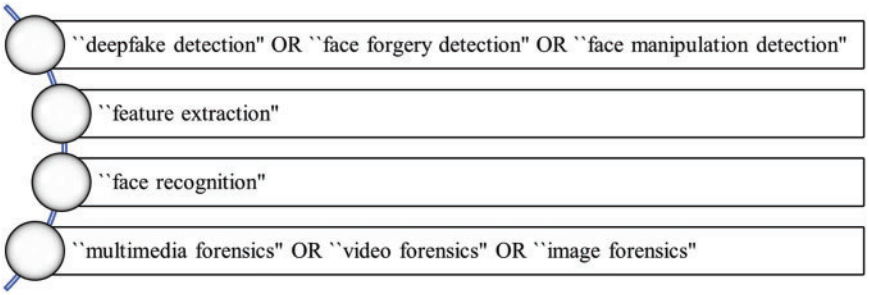

Figure 12: Four key themes based on the VOSviewer visualization

Table 8: Search queries conducted on WoS

No.	Search query	Results
1	detect OR detection OR detecting (Title) AND “deepfake” OR “deep fake” OR “deep-fake” OR “deepfakes” OR “deep fakes” OR “deep-fakes” AND “face forgery” OR “face-forgery” OR “face manipulation” OR “face-manipulation” (Title)	749
2	“feature extraction” OR “extract feature” OR “extract features” OR “extracting feature” OR “extracting features” (Title) AND “deep fake” OR “deepfake” OR “deepfake” (Topic: Abstract, Keywords, Title)	8
3	“face recognition” OR “recognize face” OR “recognize faces” OR “recognise face” OR “recognise faces” (Title) AND “deep fake” OR “deepfake” OR “deep-fake” (Topic: Abstract, Keywords, Title)	8
4	“deep fake” OR “deepfake” OR “deep-fake” OR “deep fakes” OR “deepfakes” OR “deep-fakes” (Topic) AND “multimedia forensic” OR “multimedia forensics” OR “video forensic” OR “video forensics” OR “image forensic” OR “image forensics” (Title)	8

5.1.1 Deepfake Detection

In this section, we provide the top-cited articles on deepfake detection or face forgery detection, refer to Table 9. From the search using the keywords “deepfake detection” OR “face forgery detection” OR “face manipulation detection”, we obtained 49 results, which included review papers and technical articles. The top three reviews in this category are presented below, followed by the top 15 technical papers (most cited) are discussed. For each of the categories discussed, we present the motivation of these works and then provide a summary of the main issues covered.

Deepfakes are of utmost concern due to the threat they pose to modern society [1]. These reviews have highlighted some of the key factors driving the advancement of deepfake technology. The first [49] identifies the technical advancement that led to the availability of deepfakes as the easy access to audio-visual content on social media, the availability of modern machine learning tools and libraries, and open-source trained models, coupled with the rapid development of deep learning. Particularly, the availability of generative adversarial networks (GANs) has led to the proliferation of disinformation. The second [14] emphasizes the threats of deepfakes to national security and confidentiality, and highlights that it is becoming difficult to distinguish real and fake content with the naked eye, which can lead to several societal challenges, such as deceiving public opinion or the use of doctored evidence in court. Similar to the previous study, the third [50] also highlights the advancement of deep learning techniques, and the existence of large multimedia databases

makes it much easier to manipulate or generate realistic facial images even by common people with malicious intentions. The following is a summary of the top three reviews on deepfakes.

Table 9: Most cited papers on deepfake detection

Ref.	Cited	Title
[49]	73	Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward
[51]	34	F2Trans: High-Frequency Fine-Grained Transformer for Face Forgery Detection
[52]	33	AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio-visual deepfakes detection
[53]	33	ISTVT: Interpretable Spatial-Temporal Video Transformer for Deepfake Detection
[54]	28	Masked Relation Learning for DeepFake Detection
[14]	28	Deepfake detection using deep learning methods: A systematic and comprehensive review
[55]	26	Implicit Identity Driven Deepfake Face Swapping Detection
[56]	25	DeepFake detection algorithm based on improved vision transformer
[57]	22	Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization
[58]	22	ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild
[59]	21	AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake
[60]	20	Dynamic Graph Learning with Content-guided Spatial-Frequency Relation Reasoning for Deepfake Detection
[61]	18	Learning Features of Intra-Consistency and Inter-Diversity: Keys Toward Generalizable Deepfake Detection
[50]	18	Deepfakes Generation and Detection: A Short Survey
[62]	16	FAMM: Facial Muscle Motions for Detecting Compressed Deepfake Videos Over Social Networks
[63]	16	Artifacts-Disentangled Adversarial Learning for Deepfake Detection
[64]	16	FedForgery: Generalized Face Forgery Detection With Residual Federated Learning
[65]	15	DeepFake detection with multi-scale convolution and vision transformer

Considering the ease of access to content on social media, the availability of tools such as Keras or TensorFlow, open-source trained models, and cheaper computing infrastructure, deep learning methods have rapidly evolved. Generative adversarial networks (GANs) now make it possible to generate deepfake media, which can be used to disseminate misinformation and facilitate other social vices, such as financial fraud, hoaxes, and disruptions to government functioning. Thus, the work in [49] provides a comprehensive review of tools and ML approaches for deepfake generation and detection in audio and video, covering manipulation methods, public datasets, performance standards, and results. In addition, it discusses challenges and future directions.

The evolution of deep learning for solving various challenges in academia, industry, and healthcare has been well utilized. However, it has also been used to pose threats to confidentiality, national security, and other areas. Problems such as deepfakes, creating fake images, videos, and speech that are difficult to distinguish from real ones, have become a significant menace. At times, even humans cannot differentiate between false and authentic content, hence posing a serious threat to public opinion and court evidence.

This motivates the work in [14], which assesses deepfake detection strategies using deep learning, categorizing methods by application (video, image, audio, and hybrid multimedia detection). It provides insights into deepfake generation, detection developments, weaknesses of existing methods, and areas for further investigation, noting that CNNs are the most widely used approach.

Considering the advancements in deep learning techniques and the availability of large databases that can be freely accessed, the layman can now generate or manipulate facial samples for different purposes some of which are malicious. This motivates the work in [50], which provides an overview of deepfake and face manipulation techniques and discusses identity swap, face reenactment, attribute manipulation, and entire-face synthesis, along with current challenges and future research directions.

Apart from the above highly cited reviews on deepfake and face forgery detection, several top-cited contributions to deepfake detection are provided in this section.

Although face forgery detectors have become popular and performed impressively well, they struggle with the problem of generalization and robustness. To address these issues in face forgery detection, the authors [51] propose a high-frequency fine-grained transformer network with two components: CDA, which captures invariant manipulation patterns, and HWS, which filters out low-frequency components to focus on high-frequency forgery cues. Experiments on benchmarks demonstrate the model's robustness.

Existing methods for detecting deepfakes often focus on visual or audio modalities alone, with low accuracy in multimodal approaches. To improve this, the authors in [52] propose a unified framework for detecting manipulations in audio-visual streams of deepfake videos. The dense Swin transformer network (AVFakeNet) shows robustness across varied illumination and ethnicity, with experiments confirming its efficiency and generalization.

For robust deepfake detection, researchers explore joint spatial-temporal information, but these models often lack interpretability. Thus, the authors in [53] propose an interpretable spatial-temporal video transformer (ISTVT) to capture spatial artifacts and temporal inconsistencies. Extensive experiments validate its effectiveness and provide visualization-based insights.

Most deepfake detection approaches treat it as a binary classification task, ignoring relationships across regions. This motivates the study in [54], which formulates detection as a graph classification problem, where facial regions are vertices. To reduce redundancy, the authors use masked relation learning, achieving a 2% improvement over state-of-the-art methods.

Face swapping is aimed at replacing the target face with the source face and generating a fake face difficult for humans to tell whether it is fake or genuine. Thus, the authors in [55] aim to look at the problem of face-swapping detection from the perspective of face identity. Thus, they propose an implicit identity-driven framework, utilizing differences between explicit and implicit identities to detect fakes. This method generalizes well against other solutions, as shown by experiments and visualizations.

CNNs can identify deepfakes but often suffer from overfitting and struggle to connect local and global features, leading to misclassification. Thus, the authors in [56] propose an efficient vision transformer model that combines CNN and patch-based positioning, showing improved generalization and performance, accurately detecting 2313 out of 2500 fake videos.

Unexpected learned identity representations on images hinder the generalization of binary classifiers for detection. This is the observation made by the authors in [57] who analyzed binary classifiers' generalization performance in deepfake detection, finding that implicit identity leakage limits generalization. They propose a method to reduce this effect, outperforming other methods in both in-dataset and cross-dataset evaluations.

Benchmarking is crucial in enabling meaningful comparisons of solutions to popular problems in language and speech processing. Benchmark evaluations can demonstrate the transition from laboratory conditions to scenarios observed in the real world. In this context, ASVspoof is a challenge focused on spoofing and deepfake detection. The paper in [58] summarizes the ASVspoof 2021 challenge, presenting the results of 54 participating teams that concentrated on deepfake and spoofing detection. The results show robustness in countermeasures to logical access tasks and robustness for physical access tasks in real physical spaces. Similarly, it was observed that detection generalization for deepfake target detection solutions for manipulated compressed speech is resilient to compression effects but not generalizable across different source datasets. The paper also reviews top-performing systems and challenges and provides a roadmap for the future of ASVspoof development.

Prior research on deepfake detection mostly captures intra-modal artifacts, but real-world deepfakes involve both audio and visual elements. Thus, the authors in [59] propose a joint audio-visual detection method that leverages inconsistencies between modalities. For evaluation, the authors built a new benchmark that focuses on more than one modality and can cover more forgery methods. The proposed method shows a superior performance over other methods in experiments.

Existing face forgery methods using frequency-aware information combined with CNN lack adequate information interaction with image content, thus limiting the generalizability. Hence, the work in [60] proposes a spatial-frequency dynamic graph method to capture relation-aware features in spatial and temporal domains via dynamic graph learning, achieving performance improvements over state-of-the-art methods.

Several deepfake detection approaches attempt to learn discriminative features between real and fake faces using an end-to-end trained DNN. However, most of those works suffer from poor generalization among different data sources, forgery methods, and post-processing operations. To address these generalization issues, the authors in [61] propose a transformer-based self-supervised learning method and data augmentation strategy, enhancing the model's ability to distinguish subtle differences in real and fake images. Experiments validate its superior generalization ability on unseen forgery methods and untrained datasets.

Most detection methods do not perform detection sufficiently well on compressed videos, which are common on social media uploads. Thus, the authors in [62] propose a facial muscle-motion framework based on residual federated learning for face forgery detection. The proposed framework detects compressed deepfake videos, demonstrating strong performance and resilience to compression effects. Also, results from theoretical analysis show that compression does not affect facial muscle motion feature construction, and differences in features exist between deepfake and real videos.

Effective extraction of forgery artifacts is crucial for deepfake detection. However, features extracted by a supervised binary classifier often contain irrelevant information. Moreover, existing algorithms experience performance degradation when there is a mismatch between training and testing datasets. Thus, the study in [63] proposes an artifact-disentangled adversarial learning framework to isolate artifact features, outperforming other methods on benchmark datasets.

Existing face forgery detection methods rely on publicly shared or centralized data for training, overlooking privacy and security concerns when personal data cannot be shared in real-world scenarios. Additionally, variations in artifact types can negatively impact detection accuracy due to differences in data distribution. Thus, authors in [64] propose a federated learning model (FedForgery) that enhances detection generalization across decentralized data without compromising privacy. Experiments were conducted on a publicly available face forgery detection dataset, and the result proves the superiority of the performance of the proposed Fedforgery.

The authors in [65] note that while existing methods perform well on high-quality datasets, their performance on low-quality and cross-validation datasets is often unsatisfactory. To address this, the authors propose a new CNN-based method for deepfake detection. The proposed CNN-based model is combined with a vision transformer for improved detection of deepfake artifacts at different scales, achieving better detection performance across datasets of different quality levels and good generalization across cross-datasets.

In summary, one of the primary challenges faced by face forgery detectors is achieving good generalization and robustness, despite their growing popularity. To address this, models capable of effective generalization are essential. Many existing proposals tend to focus exclusively on either visual or audio modalities, often neglecting the comprehensive detection of multimodal deepfakes, a task that presents significant challenges. Another critical consideration is the need for interpretable deepfake detection models, as many current approaches lack interpretability. Additionally, effective detection methods should account for inter-relationships across regions in deepfakes to improve performance. Avoiding overfitting is also a crucial aspect of designing robust deepfake detection algorithms and frameworks. In addition, the use of advanced learning architectures to improve deepfake detection accuracy is another major aspect that needs to be well considered. Given the prevalence of low-quality datasets and compressed videos on social media, detection methods must perform reliably under such conditions. Effective extraction of forgery artifacts is essential, and classification algorithms must demonstrate strong performance across diverse datasets for successful deepfake detection. Moreover, many forgery detection methods rely heavily on publicly shared or centralized data, raising significant security and privacy concerns. Variations in artifact types due to data distribution further complicate detection accuracy. Finally, benchmarking and comparing solutions to address common challenges in language and speech processing, especially those related to deepfakes, is vital. Organizing competitions in this domain can help drive innovation and establish standardized evaluation criteria.

5.1.2 Feature Extraction

Feature extraction plays a pivotal role in the detection of AI-generated media, especially in the context of deepfakes and other forms of synthetic content. As generative models, such as DeepFake, DALL-E, and various voice synthesis technologies, continue to advance, they produce hyper-realistic images, videos, and audio that challenge traditional authentication and detection systems. Feature extraction techniques help address these challenges by identifying unique patterns, artifacts, and inconsistencies that can distinguish authentic content from manipulated or artificially generated media [66]. Effective feature extraction captures critical details within the data that may not be visually or audibly apparent but are essential for classification and detection. For example, in image-based deepfake detection, methods such as Error Level Analysis (ELA) and Photo Response Non-Uniformity (PRNU) have been employed to highlight compression artifacts or sensor noise patterns that differ between real and synthetic images. In audio deepfake detection, Mel spectrograms and Gammatone spectrograms can reveal subtle frequency anomalies introduced during synthetic generation, while advanced feature extraction through modified neural networks like ResNet enhances the identification of these anomalies.

Furthermore, in the selected articles, we found that various cutting-edge feature extraction techniques were designed to improve the robustness and accuracy of deepfake detection across media types. These techniques leverage deep learning architectures, optimized spectrograms, and innovative neural network structures to enhance the granularity and relevance of the extracted features, thus facilitating more precise differentiation between real and manipulated content. For example, a study that uses Face-Swap Detection with ELA and Convolutional Neural Network (CNN). A novel technique combines deep learning and error

level analysis (ELA) to detect these manipulations. By identifying differences in image compression ratios between the fake and original areas, the ELA method exposes counterfeit traces. A Convolutional Neural Network (CNN) is trained to extract these counterfeit features and classify images as real or fake. This approach offers significant advantages in terms of accuracy, efficiency, and computational cost reduction, making it a powerful tool for detecting DeepFake-generated images [67]. The work in [68] introduces a novel deep neural network architecture to extract robust lip features for speaker authentication, particularly in the face of deepfake attacks. To mitigate the impact of static lip information and enhance the representation of dynamic talking habits, the proposed model incorporates two innovative units: Diffblock and DRblock. Experimental results on the GRID dataset demonstrate the effectiveness of the proposed approach, surpassing state-of-the-art methods in both human and CG imposter scenarios. The proposed network incorporates two innovative units: the Feature-level Difference block (Diffblock) and the Pixel-level Dynamic Response block (DRblock). These units effectively mitigate the impact of static lip information and capture dynamic talking habits. Experimental results using the GRID dataset demonstrate the superior performance of the proposed method in accurately distinguishing between genuine and forged lip presentations, outperforming state-of-the-art visual speaker authentication techniques. It is worth noting that recent years have witnessed a surge in audio impersonation attacks, posing a significant threat to voice-based authentication systems and speech recognition applications [66].

To counter the above-mentioned attacks, robust detection methods are imperative. This paper introduces a novel approach to enhance front-end feature extraction for audio impersonation attack detection, specifically focusing on the Hindi language. The proposed model leverages a combination of Gammatone spectrogram, Mel spectrogram, and Ternary Pattern Audio Features (TPAF) spectrogram, followed by an optimized ResNet27 for feature extraction. Subsequently, four different binary classifiers (XGboost, Random Forest, K-Nearest Neighbors, and Naive Bayes) are employed to classify audio samples as genuine or spoofed. The proposed method demonstrates superior performance, achieving a 0.9% Equal Error Rate (EER) for impersonation attacks on the Voice Impersonation Corpus in Hindi Language (VIHL) dataset, outperforming existing techniques [66].

Besides, reference [69] used Gammatone spectrograms and a ResNet27 model; this method detects Hindi-language audio impersonation attacks with high accuracy, surpassing existing techniques in robustness and accuracy. Reference [70] has improved a deep learning approach with multi-phase feature extraction (including Gabor Filter and RN50MHA) that accurately detects deep fake images, achieving high detection rates across various datasets. Another study has leveraged Photo Response Non-Uniformity (PRNU) and Error Level Analysis (ELA), this method trains CNNs to differentiate photorealistic AI images from real photos, achieving over 95% accuracy [71]. In addition, MSFRNet, a multi-scale feature extraction framework, addresses feature omission and redundancy in detecting deep fake images, outperforming standard binary classifiers through a multi-scale prediction network [72]. Another study uses Rotation-Invariant Local Binary Pattern in Fog Computing (VRLBP), a secure fog computing protocol for rotation-invariant local binary pattern (RI-LBP) feature extraction, enhances privacy in outsourced deepfake detection, achieving accuracy close to RI-LBP with reduced computational overhead [73].

The advancement of generative models, including DeepFake, DALL-E, and various voice synthesis technologies, has enabled the production of synthetic content with a level of realism that complicates conventional authentication and detection efforts. Feature extraction techniques are essential for isolating subtle artifacts, inconsistencies, and patterns, such as compression irregularities or sensor-specific noise, that serve as distinguishing markers between authentic and manipulated content. Recent scholarly efforts underscore the significance of developing advanced feature extraction methods tailored to diverse media modalities. In

image-based deepfake detection, techniques such as Error Level Analysis (ELA) and Photo Response Non-Uniformity (PRNU) have proven effective in highlighting compression artifacts and sensor noise anomalies. Similarly, in audio-based detection, spectrogram-based approaches, including Mel and Gammatone spectrograms, integrated with advanced neural networks such as ResNet, have demonstrated efficacy in identifying subtle frequency aberrations induced by synthetic generation. Innovative methodologies have further enhanced the robustness and precision of deepfake detection. Notable examples include convolutional neural networks (CNNs) trained with ELA, which effectively classify manipulated images based on compression disparities, and multi-scale feature extraction frameworks like MSFRNet, which address feature omission and redundancy to improve detection performance. Additionally, models employing novel components such as Diffblock and DRblock for dynamic lip feature extraction have achieved superior accuracy in detecting visual manipulations, while optimized spectrogram-based techniques have demonstrated high efficacy in audio impersonation detection. Despite significant advancements, the persistent evolution of deepfake technologies underscores the critical need for continued innovation in feature extraction methodologies. The development of more sophisticated and computationally efficient techniques is imperative to maintain detection accuracy and reliability in the face of increasingly sophisticated synthetic media. Such efforts are vital for ensuring the integrity of authentication systems across diverse applications and domains.

5.1.3 Face Recognition

The rapid evolution of deep learning and generative models has significantly impacted fields such as computer vision, natural language processing, and multimedia processing, introducing both groundbreaking opportunities and complex challenges. One of the most contentious applications of these advancements is the creation of deepfakes- highly realistic, AI-generated images, videos, or audio clips that convincingly replicate the likeness of real individuals. Enabled by generative adversarial networks (GANs) and other sophisticated deep learning algorithms, deepfakes are increasingly indistinguishable from authentic content and pose serious implications for privacy, security, and ethical standards. Consequently, the field of deepfake detection has gained immense attention in both academic research and industry applications, particularly as public concerns over misuse and manipulation grow.

While many researchers have developed algorithms to identify deepfake content, current literature reveals several persistent challenges in detection methods. Existing techniques often struggle with generalizability across diverse datasets, maintaining efficiency in computationally constrained environments, and effectively handling nuanced presentation attacks like morphing and impersonation [74]. Additionally, there are growing concerns about ethical implications, such as racial bias in face recognition systems, which may be exacerbated by deepfake manipulations [75]. These issues underscore the need for advanced feature extraction techniques, novel neural network architectures, and robust evaluation methodologies to improve the accuracy, efficiency, and fairness of deepfake detection systems.

From novel applications of the Fisherface algorithm combined with Local Binary Pattern Histogram (FF-LBPH) for image analysis [76] to the use of advanced contrastive learning frameworks for video detection, these studies illustrate the breadth of techniques being developed to tackle the deepfake problem [77]. Furthermore, research into the cognitive and neural responses to deepfake stimuli highlights new frontiers in detection that leverage human perceptual differences [78], while analyses of racial bias in face recognition APIs underscore the importance of ethical considerations in deploying detection systems. By systematically summarizing and analyzing these diverse approaches, this review aims to provide a comprehensive overview of the state of deepfake detection research, identify key trends and challenges, and suggest directions for future investigation.

Besides, recent breakthroughs in deep generative models have enabled the creation of highly realistic fake faces, known as deepfakes. To combat this growing threat, this research paper explores the effectiveness of various state-of-the-art loss functions commonly used in face recognition for deepfake detection. By conducting extensive experiments on challenging deepfake datasets, the authors provide a comprehensive evaluation of these loss functions and their generalization capabilities across different deepfake datasets. The findings highlight the potential of face recognition-based approaches in accurately distinguishing between real and fake faces, offering a promising avenue for robust deepfake detection [79]. Tariq et al. [80] investigated the robustness of face recognition and verification APIs against deepfake impersonation attacks. By subjecting these APIs to a series of controlled experiments using deepfake-generated celebrity faces, the authors assess their ability to accurately identify real individuals from their fabricated counterparts. The study highlights the potential vulnerabilities of these APIs to deepfake attacks and underscores the need for robust security measures to mitigate such threats. Furthermore, the work in [81] proposes a novel deepfake detection method combining Fisherface with Local Binary Pattern Histograms (FF-LBPH) and Deep Belief Networks (DBN) with Restricted Boltzmann Machines (RBM). By leveraging the dimensionality reduction capabilities of FF-LBPH and the powerful feature extraction of DBN-RBM, the proposed method aims to accurately identify deepfake images from real ones. The effectiveness of the approach is evaluated on publicly available datasets, demonstrating its potential to mitigate the risks associated with deepfake technology.

The rapid evolution of deep learning and generative models has enabled the creation of hyper-realistic deepfakes that pose significant threats to privacy, security, and ethical standards, particularly in face recognition systems. Despite advances in detection algorithms, persistent challenges such as limited generalizability, computational inefficiencies, and the complexity of detecting sophisticated attacks like impersonation and morphing remain unresolved. Moreover, ethical concerns, including racial biases in face recognition systems, amplify the need for equitable solutions. Recent breakthroughs, such as integrating Fisherface with Local Binary Pattern Histograms (FF-LBPH) and Deep Belief Networks (DBN), have demonstrated promising results in improving detection accuracy. Studies also reveal vulnerabilities in widely used face recognition APIs, underscoring the urgent need for enhanced security measures. To combat the escalating risks posed by deepfake technology, future research needs to prioritize innovative, fair, and robust methodologies that not only advance technical performance but also address ethical implications, ensuring reliable and equitable protection against these threats.

5.1.4 Deepfake Forensics

Forged images and videos have become widespread in the last few years due to the availability of powerful and easy-to-use media editing tools. Moreover, to make matters worse, social media has provided a convenient platform to share and spread these forged multimedia files or deepfake content easily. Multimedia forensics focuses on analyzing digital multimedia content to produce evidence and detect deepfakes. The following paragraphs discuss some of the top-cited articles on deepfake forensics. In this section, we provide the top-cited review articles and technical contributions to deepfake forensics.

The authors in [82] emphasize image forgery problems, which include misleading public opinion and the usage of doctored proof in court. In the paper, they present a comprehensive literature review of image forensics techniques, focusing on deep-learning-based methods. Specifically, they discuss the image forensics challenges including the detection of routine image manipulations, detection of intentional image falsifications, camera identification, classification of computer graphics images, and detection of emerging Deepfake images. They also provide a review of the available image databases and recent anti-forensic methods, and finally, some proposals on a few effective ways to curb the spread of doctored images.

Due to the rise of fake multimedia content that leads to many undesirable incidents, such as ruining the image of a public figure, or criminal activities such as terrorist propaganda and cyberbullying, there is a need for multimedia forensics. In this survey paper [83], the authors investigate the latest trends and deep learning-based techniques used in the field of multimedia forensics, regarding deepfake detection. First, they examine the manipulations of images and videos produced with editing tools, as well as the deep-learning approaches to counter these attacks. Secondly, they discuss the challenges of source camera model and device identification, including monitoring image and video sharing on social media. Thirdly, they present methods to identify deepfakes by showing the existence of traces left in deepfake content. The commonly used metrics and datasets are also discussed in this paper.

In this survey paper [84], the authors discuss the various approaches for tampering detection in multimedia data using deep learning models. They provide a comprehensive list of tampering clues and the commonly used deep learning architectures. They then discuss the available tampering detection methods, including their strengths and weaknesses, by classifying them into deepfake detection methods, splice tampering detection methods, and copy-move tampering detection methods. A detailed analysis of publicly available benchmark datasets for malicious manipulation detection is also provided. Finally, they discuss their findings, the research gaps, and the future direction of multimedia data tampering detection works.

In this digital era, images and videos have been edited with malicious intentions. Hence, the need for effective defense instruments that are able to detect such alterations has increased. The advent of deep-learning-based techniques has benefited both content manipulation (deepfakes) as well as provided effective detection solutions. This work in [85] provides a comprehensive study on the evolution of the various kinds of manipulations, as well as focuses on the diverse multimedia forensic techniques and approaches. Some lessons learned and future research challenges are also presented, together with an analysis of the solutions provided.

Apart from the above reviews on deepfake forensics, several technical papers have been identified as being the most cited. The top four contributions of these papers are provided in what follows.

The authors in [86] highlight the lack of interpretability in the feature extraction and analysis processes during the neural network model training phase. Hence, they propose an interpretable DeepFake video detection method using facial textural disparities in multi-color channels. This includes using statistical disparities of the real and fake frames in each color channel and a co-occurrence matrix in constructing a low-dimensional set of features for detection. The proposed method, when evaluated on video and frame levels, outperforms the benchmark methods. In particular, it performs better than the machine learning-based detectors and is comparable to some of the deep learning-based detectors when used on FaceForensics++ and Celeb-DF datasets. The proposed method performs well in face compression attacks and is time-efficient compared to some deep learning-based detection methods.

Due to the rise of multimedia manipulations, the authors in [87] focus on multimedia forensics, whereby they present reliable methods for detecting manipulated images and source identification. In digital integrity, the main techniques for forgery detection and localization, starting from methods that rely on camera-based and format-based artifacts, are presented. The results of their proposed deep learning-based approaches using challenging datasets and realistic scenarios are presented, showing robustness to adversarial attacks. In identifying image and video source attribution, the device used for its acquisition is studied from different viewpoints, including detecting the device function, the device's make and model, as well as the use of a specific device. Results of both exploited model-based and data-driven techniques solutions are presented using standard datasets.

Recently, Generative Adversarial Networks (GANs) have been misused to facilitate deceptive content creation, including deepfakes, image tampering, and information hiding. Authors in [88] propose a detection model that employs a spatial-frequency joint dual-stream convolutional neural network. They leverage the learnable frequency-domain filtering kernels and frequency-domain networks to thoroughly learn and extract frequency-domain features. These two sets of traits are then combined to identify GAN-created faces effectively. The proposed model outperforms other recent methods when tested using various datasets, in terms of detection accuracy on high-quality created datasets as well as generalization across datasets.

Many forged video detection works focus on exploring frame-level cues, thus lacking in investigating the affluent temporal information, such as the spatiotemporal features. Thus, authors in [89] propose a Channel-Wise Spatiotemporal Aggregation (CWSA) module to fuse deep features of continuous video frames without any recurrent units. They crop the face region with some background remaining, which transforms the learning objective from manipulations to the difference between pristine and manipulated pixels. Then, a deep convolutional neural network (CNN) with skip connections that are conducive to the preservation of detection-helpful low-level features is implemented to extract the frame-level features. The CWSA module then decides by aggregating deep features of the frame sequence. Using FaceForensics++, Celeb-DF, and DeepFake Detection Challenge Preview datasets, their proposed method outperforms other recent methods.

To reiterate, multimedia forensics focuses on analyzing digital multimedia content to produce evidence and detect deepfakes. However, numerous challenges exist in conducting forensic analysis on voice, images, and videos to determine their authenticity. In image forensics, key concerns include detecting routine image manipulations, identifying intentional falsifications, camera identification, classifying computer-generated images, and source attribution. The lack of interpretability due to the black-box nature of deep learning models further complicates forensic analysis. To enhance interpretability, some researchers propose leveraging facial textural disparities in multi-color channels, while others suggest detecting traces left in deepfake content as evidence. More advanced techniques, such as spatial-frequency joint dual-stream convolutional neural networks and spatiotemporal feature analysis, have also been introduced. Additionally, a comprehensive list of tampering clues has been compiled to aid in the detection of fabricated multimedia data. By addressing these challenges and refining existing techniques, the research community can enhance the accuracy and reliability of multimedia forensics in detecting and analyzing manipulated content.

5.2 Insights, Limitations, and Way Forward

Significant advancements in artificial intelligence have led to the rise of deepfakes in modern society. This has prompted the development of numerous solutions to detect deepfakes, given their societal impact, including misinformation, the spread of fake news, political influence, decision-making challenges, mistrust, and ethical concerns. Deepfakes also pose serious cybersecurity threats and raise issues related to fact-checking, copyright, intellectual property, and fraud. For these reasons, mitigating deepfakes and improving detection methods are of critical importance. However, there are limitations in detecting deepfakes-related research, in particular the dataset used, which is among keywords with the highest number of occurrences in our bibliometric study (refer to Table 6 for keywords *data* and *dataset*).

One of the biggest challenges in deepfake detection is the availability of diverse public datasets. Some of the commonly used deepfakes datasets, namely FaceForensics++ (FF++), Deepfake Detection Challenge (DFDC), Celeb-DF, Deepfake Detection (DFD), WildDeepfake, and DeeperForensics-1.0, lack diversity in attributes within the datasets, which can hinder the development of robust and generalizable deepfake detection models [9,84]. As such, in 2021, the Korean DeepFake Detection Dataset (KoDF) [90] was created to address the underrepresentation of Asian subjects in existing deepfake datasets. However, it still does not encompass the full diversity of Asian appearances. Such limitation needs to be addressed accordingly,

as detection performance relies heavily on diverse datasets for training [91]. Besides insufficient diverse public datasets, the quality and the size of the available datasets are another key challenge in developing accurate models [9,10,49,92]. Such a situation leads to the inability of the models to effectively generalize unseen data or in a real-world setting [10]. Important issues such as quality, fairness, and trust of deepfake datasets (to overcome biased and imbalanced data) are also discussed in [9,92]. Evidently, larger and more diverse datasets are required to represent a wider range of demographics to improve detection. By exploring advanced data augmentation techniques, the size and diversity of the deepfake datasets can be increased. Data cleaning is also important to improve data quality, which minimizes inconsistencies in data. By addressing these dataset quality issues, the accuracy and reliability of the detection models can be improved, hence combating the fabricated contents.

Deepfake detection requires robust algorithms, with AI playing a central role. Future research should prioritize accuracy, accessibility of tools, ease of use, multi-modal support, and enhanced detection performance. In addition to popular algorithms like multilayer neural networks CNNs, GANs, LSTM, and RNNs, frameworks such as federated learning and transfer learning can be refined to boost detection capabilities. Techniques, including ensemble learning, vision transformers, attention mechanisms, and decision trees, can also be leveraged. Feature extraction is crucial, and exploring unique or multi-scale features can enhance detection accuracy. Efficient algorithms capable of handling large datasets and the curation of high-quality datasets are essential for advancing research in this area.

Additionally, beyond AI, blockchain technology offers a decentralized approach to verifying media authenticity through tamper-proof public records [18]. Combining AI with blockchain could provide robust solutions for detecting deepfakes and verifying authenticity with high confidence. Furthermore, the study also highlights that fewer research papers address deepfakes in medicine, despite their significant implications in medical imaging and decision-making. For instance, a fake X-ray could result in incorrect medical advice. Similarly, more research is needed on the societal impacts of deepfakes, particularly in criminology and law, where fake media presented as evidence could lead to erroneous judgments.

Finally, to combat AI-based disinformation and its threats to national security, countries must update and modernize their laws and regulations. EU, the United Kingdom, the United States, China, Canada, and Korea have introduced several initiatives to address these threats [93]. The DEEP FAKES Accountability Act was introduced in 2019 by the US government and focuses on preventing the distribution of deepfakes during an election. In the same year, China introduced laws mandating individuals and organizations to disclose when they have used deepfake technology in videos and other media. In December 2024, the South Korean AI Basic Act, aligned with the EU AI Act, was announced to provide rules for AI governance, which include ethical AI usage [94]. South Korea was one of the first countries to invest in AI regulatory exploration, due to its strong AI technological advancement. Apart from a modernized regulatory framework, the UK government funds research into deepfake detection technologies as well as collaborating with industry and academic institutions to develop best practices to detect and respond to deepfakes [93]. Many governments around the world are beginning to acknowledge the importance of mitigating the potential risks associated with AI advancement, with many having started proposing similar initiatives.

6 Exhaustive Search and Analysis

To fully comprehend the diverse work around deepfake research more, we conducted another Scopus Search on 5/01/2025 using only the words *deepfake OR deep fake OR deepfake OR deepfakes OR deep fakes OR deepfakes*, without *artificial intelligence*. The result was uploaded to VOSviewer with a minimum keyword frequency of 5. Out of 8790 keywords, 512 met the minimum threshold, with 511 being fully connected (used to generate Fig. 13). First, we listed out all 512 keywords and observed keywords with more than

one variant. For example, the following keywords represent the same concept: *convolution neural network, convolution neural network (CNN), convolution neural networks, convolutional networks, convolutional neural network, convolutional neural network (CNN), convolutional neural networks, convolutional neural network, and convolutional neural networks (cnns)*. Based on the VOSviewer results, different keywords representing the same concept may appear across multiple clusters and will be discussed within their respective contexts. Second, many of these keywords indeed have a huge influence in the deepfake research area. Particularly, it can be observed that deep learning is the most popularly deployed technique for detecting deepfakes, which is also evident in Fig. 13. Also, keywords like *deep learning (DL), deep learning algorithms, deep learning methods, deep learning model, deep learning models, and deep learning techniques* all have an aggregate occurrence and link strength of 1732 and 4613, respectively, the highest in the entire dataset. This excludes neural network-related keywords. Besides, the role of convolutional neural networks as well as GAN in deepfake research is also evidently observed. Particularly, convolutional neural network-related keywords have a total occurrence (frequency) of 399 and a link strength of 1049, while keywords related to GAN have a total frequency of 370 and a link strength of 713. Other keywords with multiple variants include *LSTM*, *SVM*, *RNN*, vision transformer, and Graph neural networks.

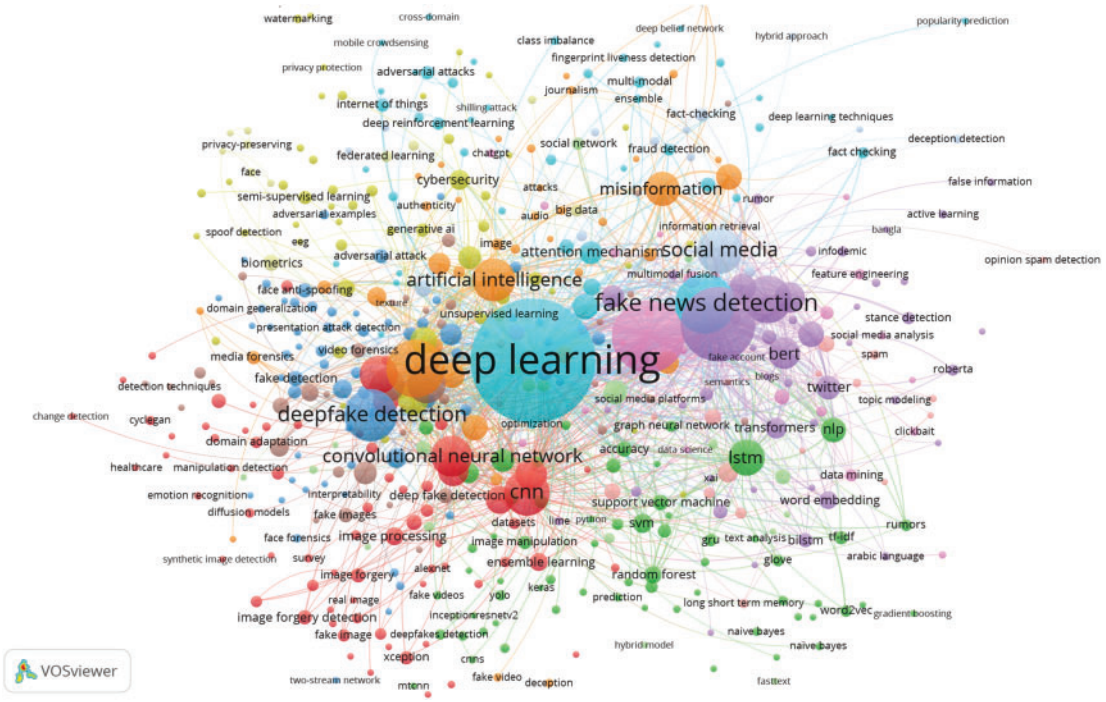

Figure 13: A comprehensive exposition of keywords on Deepfakes

Finally, VOSviewer tool was used to cluster the keywords. The results produced 13 clusters with their associated keywords (refer to Figs. 13–16) that may be useful for readers to further explore deepfake research areas. The full meanings of acronyms related to these clusters are provided in Table 10. To understand these keywords better and their level of presence within the research landscape, the number of occurrences of these keywords and their total link strength are provided for each cluster. The occurrences show the level of presence of these terms in the deepfake research landscape, and the link strength shows how interconnected they are with several other terms within the entire research landscape. These clusters and the exhaustive

list show the diverse methods, implications, applications, concerns, requirements, challenges, models, tools, datasets, and forms of deepfakes.

Cluster 1

Adversarial Learning, Alexnet, Artificial Neural Network, Audio Deepfakes, Augmentation, Change Detection, Classification, CNN, Convolutional Neural Network, Copy-Move Forgery, Cyclegan, Data Augmentation, Datasets, Deep Convolutional Generative, Deep Convolutional Neural Network, Deep Learning Model, Densenet, Diffusion Models, Digital Image Forensics, Domain Adaptation, Edge Computing, EfficientNetB0, ELA, Emotion Recognition, Ensemble Learning, Error Level Analysis, Face Manipulation Detection, Fake Currency, Fake Face Detection, Fake Image, Fake Image Detection, Forgery, Generalizability, Generative Adversarial Network, Healthcare, Image Augmentation, Image Authentication, Image Forensic, Image Forgery, Image Forgery Detection, Image Processing, Image-to-Image Translation, Inception V3, IoT, Manipulation Detection, MobileNet, Multimedia, Object Detection, Pre-trained Model, Preprocessing, Real Image, Remote Sensing, Resnet, Resnet50, Segmentation, Semantic Segmentation, Signal Processing, Splicing, Style Transfer, Survey, Synthetic Image Detection, Synthetic Media, Transfer Learning, U-Net, VGG, VGG-16, VGG16, Xception

Cluster 2

Accuracy, ANN, Bangla Fake News, Bidirectional LSTM, Binary Classification, Celeb-DF, CNNs, Confusion Matrix, Convolution Neural Network (CNN), Convolutional Neural Network (CNN), Convolutional Neural Networks (CNN), Decision Tree, Deep Learning Algorithm, Deepfake Video, Deepfake Video Detection, Deepfakes Detection, DFDC, Encoder, Face detection, Faceforensics++, Fake Videos, GANs, Glove, Gradient Boosting, Graph Neural Network, GRU, HOG, Image Manipulation, Imbalanced Dataset, InceptionResnetv2, Kaggle, Keras, KNN, Logistic Regression, Long short term memory (lstm), Long Short-term Memory (lstm), LSTM, Machine Learning Algorithms, Model Training, MTCNN, Näive Bayes, Naive Bayes, Natural Language Processing, Naive Bayes, NLP, Prediction, Python, Random Forest, Recurrent Neural Network, ResNext, Rumors, Support vector machine, Support vector machine (SVM), SVM, Tensorflow, Text Analysis, Text Preprocessing, TF-IDF, Video Analysis, Vision Transformer, Word2Vec, Xceptionnet, XGBoost, YOLO.

Cluster 3

Add Challenge, Adversarial Attack, Adversarial Examples, Adversarial Training, Anti-spoofing, ASVspoof, Audio Deepfake, Audio Deepfake Detection, Automatic Speaker Verification, CNN-LSTM, Contrastive Learning, Convolution Neural Network, Dataset, Deep Neural Networks, DeepFake Audio, Deepfake Detection, Deepfake Dataset, DNN, Domain Generalization, Ensemble Model, Face Anti-Spoofing, Face Forensics, Face Forgery, Face Forgery Detection, Face Liveness Detection, Face Recognition, Face Spoofing Detection, Faceswap, Fake Audio Detection, Fake Detection, Fake Face, Fake Speech, Fake Speech Detection, Feature Selection, Few-Shot Learning, Fine-Tuning, Fingerprint, Frequency Domain, Generalization, Information Security, Interpretability, Knowledge Distillation, Liveness Detection,Local Binary Pattern, Meta-Learning, MFCC, Multi-Task Learning, Phishing Attacks, Presentation Attack Detection, Residual Network, Robustness, Self-Attention, Self-Supervised Learning, Speech Synthesis, Spoofing, Spoofing Detection, Synthetic Speech Detection, Triplet Loss, Two-Stream Network, VGG19, Voice Conversion

Figure 14: Clusters 1–3 from VOSviewer on keywords from Deepfakes

Cluster 4

Adversarial Machine Learning, Android, Anomaly Detection, Classifier, Component, Conditional GAN, Convolutional Networks, Cosine Similarity, Cyber Security, Cybersecurity, Data Security, Deep Convolutional generative adversarial networks, Deep Fake, Deep Neural Network (dnn), Deep-Learning, Discriminator, EEG, Face Swap, Face Synthesis, Facial Recognition, Fake Reviews Detection, Fault Diagnosis, GAN, Generative Adversarial Network (GAN), Generative Adversarial Networks (GAN), Generative AI, Generative Model, Generator, Graph Convolutional Net, Image Generation, Image Restoration, Imbalanced Data, Impersonation, Indoor Localization, Information Retrieval, Intrusion Detection, Malware, Medical Imaging, Phishing, Phishing Attack, Phishing Detection, Privacy Protection, QR Code, Radiology, Security, Semi-Supervised Learning, Social Engineering, Spoof Detection, StyleGAN, Supervised Learning, Synthetic Data, Unsupervised Learning, Watermarking, Zero-Shot Learning.

Cluster 5

Active Learning, Arabic Language, Bag Of Words, Bangla, BERT, BiLSTM, Blogs, Capsule Network, COVID-19, Cyberbullying, Data Analysis, Deep Learning (DL), Deep Neural Network, DistilBERT, Explainability, Explainable AI,Fake Account, Fake News, Fake News Classification, Fake News Identification, False Information, Hate Speech, Hate Speech Detection, Infodemic, Lime, Linguistic Features, Machine Learning (ML), Misinformation Detection, Multimodal Fusion, Natural Language Processing, Opinion Mining, Pandemic, Pre-trained Models, Roberta, Rumor, Sentiment, Sentiment Analysis, Shap, Social networking (online), Stance Detection, Text Classification, Topic Modeling, Transformer Models, Transformers, Tweets, Twitter, Visualization, Word Embedding, Word Embeddings, XAI

Cluster 6

Adversarial Attacks, Adversarial Networks, Attack Detection, Attention Mechanism, Biometric, Black-Box Attacks, Class Imbalance, Collaborative Filtering, Convolution, Convolution Neural Network, Coronavirus, Cross-Domain, Deep Belief Network, Deep Learning, Deep Learning Methods, Deep Learning Technique, Deep Reinforcement Learning, Ensemble, Fact Checking, Fake Information, Fake News Detection, Feature Fusion, Fingerprint Liveness Detection, Fraud Detection, Genetic Algorithm, Graph Neural Networks, Identification, Internet of Things, Knowledge Graph, Language Model, Mobile Crowdsensing, Multi-Modal, Multi-Modal Fusion, Multimodal Fake News Detection, Neural Networks, Optimization, Poisoning Attack, Popularity Prediction, Recognition, Recommender System, Recommender Systems, Reinforcement Learning, Rumor Detection, Shilling Attack, Transformer.

Cluster 7

Artificial Intelligence, Attacks, Audio, Authentication, Authenticity, Autoencoders, Benchmark, Blockchain, Cheapfakes, Computer Vision, Deception, Deep Fakes, Deepfake, Detection, Disinformation, Facial Manipulation, Fake, Fake Media, Fake Video, Forensics, Generative Models, Hoax, Image, Journalism, Manipulation, Media, Media Forensics, Metaverse, Misinformation, Multimodal, Neural Network, News, Post-Truth, Smart Contracts, Synthetic Speech, Technology, Trust, Verification, Video, Video Manipulation.

Figure 15: Clusters 4–7 from VOSviewer on keywords from Deepfakes

Cluster 8

Attention, Auto-Encoder, Autoencoder, Bidirectional Encoder Representations from Transformer, Classification Algorithms, Convolutional Neural Networks, Deep Convolutional Neural Networks, Deep Fake Detection, Deep Generative Model, Deepfake Generation, Deepfake Videos, Deepfakes, Detection Techniques, Digital Forensics, Digital Media Forensics, EfficientNet, Face Manipulation, Face Swapping, Fake Images, Forgery Detection, Fusion, Generative Adversarial Networks, Hierarchical Attention Networks, Image Forensics, Image Recognition, Image Reconstruction, Image Splicing, Image Synthesis, News Classification, Optical Flow, Semi-Supervised, Social Media Platforms, Super Resolution, Swin Transformer, Synthetic Images, Texture, Video Forensics, Video Forgery, Video Forgery Detection

Cluster 9

AI, Arabic Fake News, Artificial Intelligence (AI), Chatgpt, Clickbait, Contextualized Text Representations, Cybercrime, Data Mining, Data Science, Deep-Fake, Emotions, Ethics, Fake Accounts, Fake Review, Fake Review Detection, Feature Engineering, Internet, Large Language Models, Machine Learning, Opinion Spam, Social Media Analysis, Social Networks, Text Mining

Cluster 10

Credibility, E-commerce, Fake Reviews, Features, Gated Recurrent Unit, Long Short Term Memory, Long Short-Term Memory, MLP, Multilayer Perceptron, Multimodality, Opinion Spam Detection, PCA, Recurrent neural network, Recurrent neural networks, RNN, Semantics, Spam, Spam Review Detection, Spam Reviews.

Cluster 11

Bi-LSTM, DCGAN, DCNN, Deep Learning Models, Deepfake Images, FastText, Feature Extraction, Fingerprint Recognition, Hybrid Model, Image Classification, Iris Recognition, Multimodal Learning, Pre-processing, Representation Learning, Social Network, Spoof Attacks, Twitter Data, Vision Transformers.

Cluster 12

Audio Forensics, Bot Detection, Deception Detection, Disinformation Detection, Fact-Checking, Fake Content, Fake Faces, Fake Profile, Hybrid Approach, Information Disorder, Multimedia Forensics, Online Social Network, Online Social Networks, Social Media, Social Network Analysis

Cluster 13

Algorithms, Big Data, Biometrics, Democracy, Explainable, Face, Fair, Federated Learning, Gesture, Privacy, Privacy-Preserving

Figure 16: Clusters 8–13 from VOSviewer on keywords from Deepfakes

Table 10: Acronyms and their meanings

Abbreviation	Full meaning
ANN	Artificial Neural Network
ASV	Automatic Speaker Verification
ASVspoof	ASV Spoofing and Countermeasures Challenge
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
ChatGPT	Chat Generative Pre-trained Transformer
CNN	Convolutional Neural Network
CycleGAN	Cycle Generative Adversarial Network
DCGAN	Deep Convolutional Generative Adversarial Network
DCNN	Deep Convolutional Neural Network
DFDC	Deepfake Detection Challenge
DL	Deep Learning
DNN	Deep Neural Network
DRL	Deep Reinforcement Learning
ELA	Error Level Analysis
EEG	Electroencephalogram
GAN	Generative Adversarial Network
GNN	Graph Neural Network
GRU	Gated Recurrent Unit
HAN	Hierarchical Attention Networks
HOG	Histogram of Oriented Gradients
IoT	Internet of Things
LIME	Local Interpretable Model-agnostic Explanations
LSTM	Long Short-Term Memory
MLP	Multilayer Perceptron
MFCC	Mel-frequency Cepstral Coefficients
ML	Machine Learning
MTCNN	Multi-task Cascaded Convolutional Neural Networks
NLP	Natural Language Processing
PCA	Principal Component Analysis
QR Code	Quick Response Code
ResNet	Residual (Neural) Network
RNN	Recurrent Neural Network
RoBERTa	Robustly Optimized BERT-Pretraining Approach
SHAP	SHapley Additive exPlanations
StyleGAN	Style-Based Generative Adversarial Network
SVM	Support Vector Machine
Swin Transformer	Shifted Window Transformer
TF-IDF	Term Frequency-Inverse Document Frequency
U-Net	Convolutional Network for Image Segmentation
VGG	Visual Geometry Group
VGG-16	Visual Geometry Group 16-Layer Model

(Continued)

Table 10 (continued)

Abbreviation	Full meaning
VGG19	Visual Geometry Group 19-Layer Model
Word2Vec	Word to Vector (a Word Embedding Model)
XAI	Explainable Artificial Intelligence
XceptionNet	Extreme Inception Network
XGBoost	Extreme Gradient Boosting
YOLO	You Only Look Once (Object Detection Algorithm)

From this extensive list in Figs. 14–16, a lot of insights can be derived. For instance, many of the keywords in clusters 1, 2, 3, 4, and 8 relate to deepfakes detection for media in different formats and their associated methods. Also, in clusters 5 and 7, many of the keywords are associated with misinformation and forgery showing the importance of detecting misinformation, the issues of fake news and the importance of its classification, the spread of rumours, sentiment analysis, the effect of deepfakes on social networking, and media, the menace of fake media, and the impact of deepfake on journalism, the importance of trust and verification of information to prevent deception due to the presence of deepfakes. Furthermore, some of the keywords in clusters 5, 7, 9, 12, and 13 indicate the social aspects of deepfakes, such as false information/misinformation, disinformation, online deception, social networks, and fake profiles. We provide some highlights of the lessons that we have derived from them.

6.1 Cluster 1

Table 11 presents the keywords in Cluster 1, which span techniques, models, tools, and requirements related to deepfake types, their generation, detection, and mitigation solutions. Notably, the versatility of CNNs is evident in Fig. 17, where they are linked to various applications. On the left, CNNs are connected to *fake image detection, deepfake video analysis, image forensics, fake news, social media,* and *spoof detection*. Similarly, on the right subfigure, *CNNs* are associated with *digital forensics, biometrics, cybersecurity, fake image analysis*, and *spam review detection*, among others.

Table 11: Keywords in Cluster 1 ordered by total link strength, i.e. Lk. Str. (descending)

Keyword	Link strength	Frequency	Keyword	Link strength	Frequency
cnn	708	226	convolutional neural network	476	184
classification	330	106	transfer learning	329	104
generative adversarial network	226	130	data augmentation	141	72
ensemble learning	107	36	image processing	93	38
fake image detection	73	25	image forgery	68	23
image forgery detection	58	24	xception	60	20
forgery	52	16	image forensic	49	15

(Continued)

Table 11 (continued)

Keyword	Link strength	Frequency	Keyword	Link strength	Frequency
datasets	43	15	resnet50	43	17
digital image forensics	38	16	fake face detection	38	18
survey	38	9	synthetic media	35	14
iot	34	10	splicing	31	7
vgg16	30	11	adversarial learning	29	15
domain adaptation	28	14	alexnet	28	9
segmentation	28	11	fake currency	27	9
real image	25	6	artificial neural network	24	8
remote sensing	23	9	vgg	21	5
synthetic image detection	20	5	copy-move forgery	19	5
vgg-16	19	6	augmentation	18	8
face manipulation detection	17	8	manipulation detection	17	8
image authentication	17	5	cyclegan	16	10
healthcare	16	6	style transfer	15	8
audio deepfakes	14	6	inception v3	14	5
signal processing	13	5	emotion recognition	13	7
image-to-image translation	12	6	deep convolutional neural network	15	9
object detection	34	17	image augmentation	10	5
semantic segmentation	10	5	generalizability	9	5
pre-trained model	8	6	u-net	6	5

Several CNN models for deepfake detection, such as *EfficientNetB0, VGG-16, DenseNet, MobileNetV2, ResNet50, InceptionV3*, and *Xception* [95], are observed within this cluster. Other CNN-based models, including *CycleGAN* and *AlexNet* [96,97], are also present. Additionally, technologies fundamental to deepfake creation, such as *diffusion models* [96] and *generative adversarial networks (GANs)*, are part of this cluster.

Classification is another prominent keyword, playing a crucial role in identifying fake news and misinformation, as shown in Fig. 18. Similarly, *transfer learning* is essential for enhancing the adaptability and generalization of deepfake detection solutions. The significance of *data augmentation* is highlighted by its strong link strength within this cluster's keyword corpus.

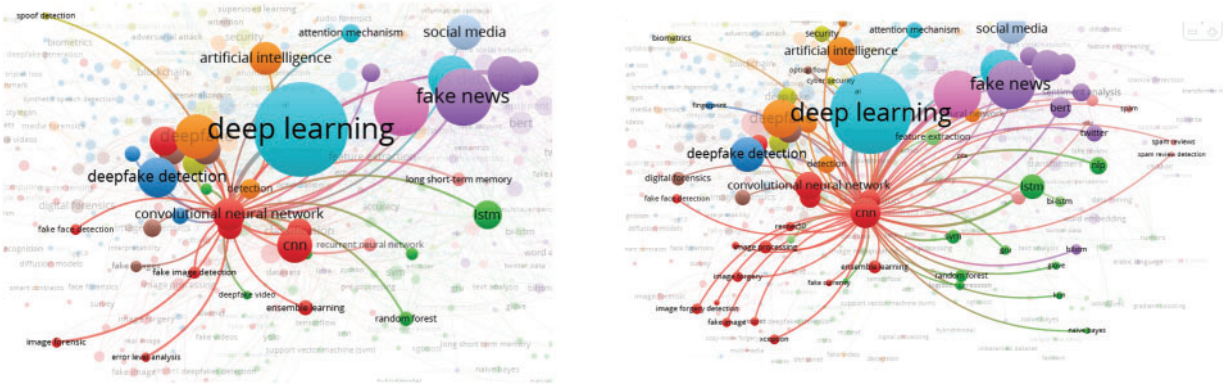

Figure 17: Different variants of CNN representations in cluster 1

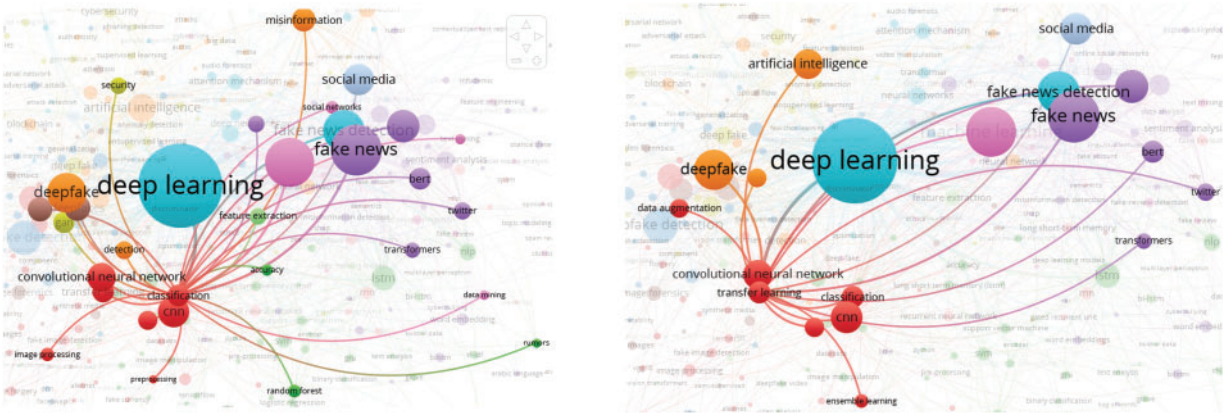

Figure 18: Classification (left) and transfer learning (right) in cluster 1

Ensemble learning also emerges as one of the most widely used techniques for deepfake detection. Several studies [98–100] have explored ensemble models for this purpose. Furthermore, keywords related to the *generalizability* of deepfake detection solutions, such as *transfer learning* and *pre-trained models*, are notable. *Image detection* and forensic-related terms, including *image forensics*, *image forgery detection*, and *fake image detection*, are also observed, with fake image detection and image forgery detection standing out due to their high link strengths.

Finally, this cluster includes keywords representing application domains relevant to deepfake research, such as *digital image forensics*, *fake currency detection*, *remote sensing*, *image authentication, healthcare, emotion recognition*, and *object detection*.

6.2 Cluster 2

Table 12 presents the associated keywords for Cluster 2, organized based on their link strength and frequency of occurrence. This cluster primarily includes deep learning models, with LSTM being the most popular, as shown in Fig. 19. LSTM has been widely used for deepfake detection across various studies [101–104] and is particularly effective in detecting misinformation, fake news, and deepfake videos.

Table 12: Keywords in Cluster 2 ordered by total link strength (descending)

Keyword	Link strength	Frequency	Keyword	Link strength	Frequency
lstm	532	143	svm	164	40
random forest	129	32	gans	127	48
recurrent neural network	109	30	accuracy	87	26
convolution neural network (cnn)	81	13	support vector machine	82	25
tf-idf	80	19	long short-term memory (lstm)	68	24
rumors	67	15	vision transformer	60	22
natural language processing (nlp)	59	19	glove	60	14
word2vec	56	15	image manipulation	54	20
deepfake video detection	45	21	knn	45	10
tensorflow	45	10	graph neural network	38	15
resnext	37	10	naive bayes	37	7
bidirectional lstm	34	8	yolo	32	12
fake videos	25	8	prediction	25	9
model training	25	5	deepfake video	24	11
ann	24	7	text analysis	23	7
xgboost	23	7	binary classification	21	8
hog	21	6	inceptionresnetv2	21	6
gradient boosting	20	5	python	20	5
video analysis	20	5	xceptionnet	19	5
cnns	18	7	long short term memory (lstm)	17	5
bangla fake news	17	5	celeb-df	17	5
mtcnn	17	6	dfdc	26	9
deep learning algorithms	15	5	kaggle	15	5
imbalanced dataset	8	5			

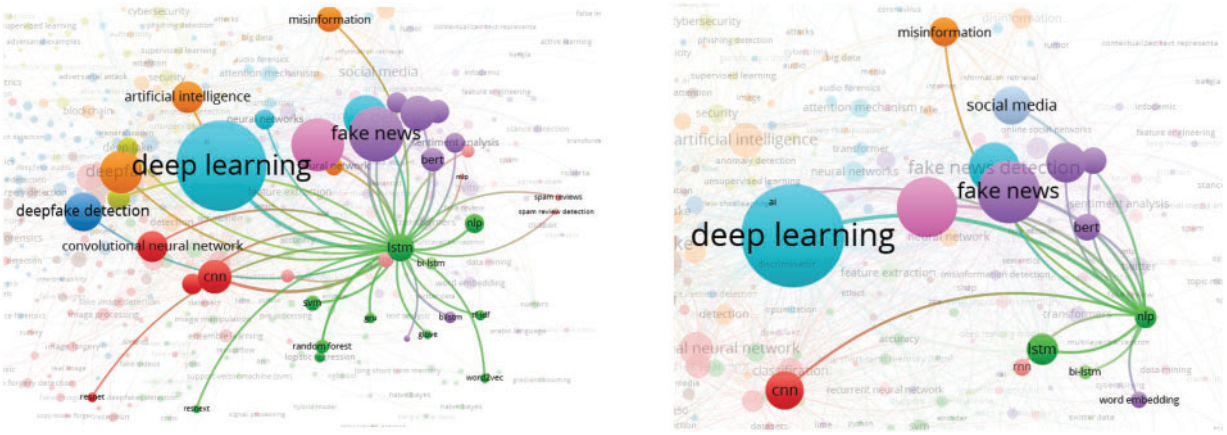

Figure 19: Links for LSTM (left) and NLP (right) in cluster 2

Overall, Cluster 2 encompasses terms related to different forms of deepfakes and misinformation, such as *rumors, fake videos*, and *fake news*, along with several deep learning algorithms employed for their detection. For instance, *CNN* [101,103,105–107], *RNN* [101,103,105,107], *InceptionResNetV2* [108], *Xception-Net* [109], and *Vision Transformer* [106,110,111] can all be leveraged for various types of deepfake data.

Additionally, the presence of supervised learning algorithms, such as *random forest and support vector machines*, highlights their relevance to deepfake detection, as illustrated in Fig. 20. Techniques related to natural language processing, which are crucial for fake news detection, are also present in this cluster. These include *TF-IDF* [112,113], *Word2Vec* [112,113], and *GloVe* [114–116].

Furthermore, algorithms for face detection, such as *MTCNN* [117–120], and techniques instrumental in eye detection, such as the histogram of oriented gradients *(HOG)* [121], are also observed in this cluster. Similarly, datasets and benchmarks for deepfake detection evaluation, including *Celeb-DF, DFDC*, and *FaceForensics++* [10,122,123], are present. Evaluation metrics, such as *accuracy*, are also included.

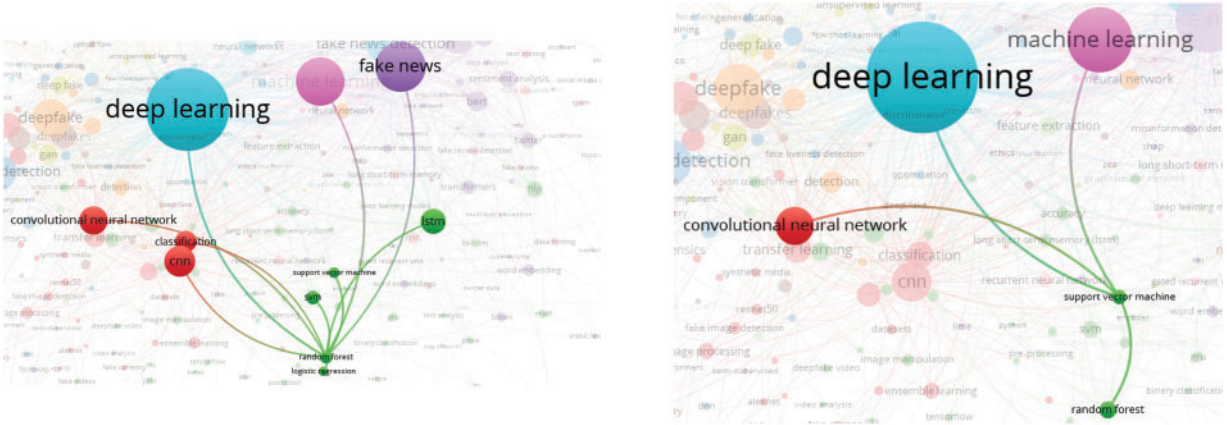

Figure 20: Links for random forest (left) and SVM (right) in cluster 2

The cluster further includes tools and frameworks essential for deepfake research, such as *Python, Keras*, and *TensorFlow*, which are crucial for implementing and training AI models. Additionally, the presence of

keywords related to fake news in languages like *Bangla* indicates the widespread impact of misinformation resulting from deepfakes and underscores the need for deep-learning solutions to address this issue.

6.3 Cluster 3

Table 13 presents the associated keywords for Cluster 3, organized based on their link strength and frequency of occurrence. Deepfake detection is a prominent focus in this cluster, as illustrated in Fig. 21 (left), where it is linked to multiple keywords both within and outside the cluster. Notably, deepfake detection is associated with CNNs, face forgery detection, contrastive learning, and generalization. Similarly, face forgery detection is linked to contrastive learning within the cluster and to convolutional neural networks outside the cluster.

Table 13: Keywords in Cluster 3 ordered by total link strength (descending)

Keyword	Link strength	Frequency	Keyword	Link strength	Frequency
deepfake detection	577	294	face recognition	125	45
deep neural networks	117	52	fake detection	76	25
face forgery detection	71	40	dataset	62	19
contrastive learning	41	21	generalization	41	16
feature selection	43	16	liveness detection	40	17
face anti-spoofing	40	22	fingerprint	40	14
self-attention	35	12	adversarial attack	35	22
adversarial training	31	14	spoofing	31	10
convolution neural network (cnn)	31	13	audio deepfake	30	11
presentation attack detection	29	13	fine-tuning	27	9
anti-spoofing	26	20	faceswap	24	9
face liveness detection	23	14	knowledge distillation	22	9
adversarial examples	21	11	self-supervised learning	21	14
binary classification	21	8	robustness	19	11
phishing attacks	19	7	cnn-lstm	18	9
face forensics	17	9	spoofing detection	17	7
domain generalization	17	9	frequency domain	16	8
vgg19	15	5	triplet loss	14	6
few-shot learning	14	5	asvspoof	14	8
fake face	14	7	residual network	13	6

(Continued)

Table 13 (continued)

Keyword	Link strength	Frequency	Keyword	Link strength	Frequency
face forgery	13	5	audio deepfake detection	13	13
interpretability	12	6	speech synthesis	12	8
information security	11	6	phishing attacks	19	7
synthetic speech detection	10	6	face spoofing detection	10	6
fake audio detection	8	7	automatic speaker verification	8	6
fake speech detection	8	7	fake speech	8	5
voice conversion	8	6	add challenge	7	5
two-stream network	6	5			

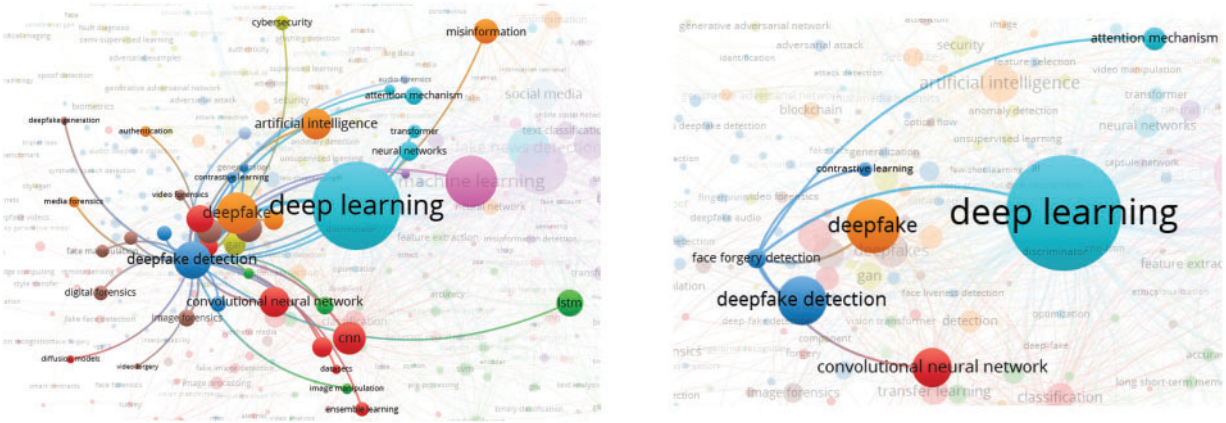

Figure 21: Links for deepfake detection (left) and face forgery detection (right) in cluster 3

This cluster also includes keywords related to various forms of deepfake detection, such as *spoofing, liveness detection, presentation attack detection, spoofing detection, audio deepfake detection, phishing attacks, synthetic speech detection, fake speech detection, fake speech, fake audio detection, fake face, face forgery*, and *speech synthesis*. These keywords indicate a focus on deepfake detection across different modalities, including audio, video, and facial images. For instance, techniques such as *self-supervised learning* have been used to detect synthetic and imitated voices [124,125].

Several algorithms for deepfake detection appear in this cluster, including *VGG19* and *CNNs*, as well as datasets such as the *Automatic Speaker Verification (ASV) Spoof dataset* [126]. Additionally, face digital manipulation techniques, such as *FaceSwap*, pose significant challenges for automated face recognition systems. Potential countermeasures, such as *face anti-spoofing*, are also observed in this cluster. These methods help prevent unauthorized access to facial recognition systems by detecting presentation attacks that try to impersonate legitimate users [127].

Furthermore, this cluster highlights techniques essential for deepfake detection, including *triplet loss*, which is commonly used in *face recognition* [128–130]. *Generalization* is another key focus, with techniques such as *self-supervised learning* and *few-shot learning* playing a crucial role, thus improving the *robustness* of deepfake detection models [131–133].

Finally, the concepts of *adversarial attacks* and *adversarial training* underscore the presence of adversarial techniques designed to evade deepfake detection systems [134].

6.4 Cluster 4

Table 14 presents the associated keywords for Cluster 4, organized based on their link strength and frequency of occurrence. *GAN* is the most prominent keyword in this cluster, with a link strength of 264 and a frequency of 100. Additionally, the keywords *deepfake, cybersecurity*, and *security* have the highest link strength. This cluster is primarily related to the techniques involved in deepfake generation, as well as cybersecurity and the detection of various forms of cyberattacks and anomalies. GANs also play a crucial role in cybersecurity research [135].

Table 14: Keywords in Cluster 4 ordered by total link strength (descending)

Keyword	Link strength	Frequency	Keyword	Link strength	Frequency
gan	264	100	deep fake	205	77
security	169	60	cybersecurity	138	43
generative adversarial network (GAN)	72	34	anomaly detection	63	24
phishing	61	19	generator	58	18
cyber security	55	15	generative ai	51	19
social engineering	46	12	supervised learning	46	15
unsupervised learning	42	20	generative adversarial networks (GAN)	33	15
semi-supervised learning	32	19	phishing detection	32	11
image generation	34	16	synthetic data	29	10
imbalanced data	27	9	information retrieval	25	6
adversarial machine learning	24	11	StyleGAN	22	8
impersonation	19	5	spoof detection	18	7
component	18	8	data security	18	5
classifier	17	6	phishing attack	16	5
indoor localization	15	6	generative model	15	7
radiology	14	5	android	12	5
convolutional networks	12	5	cosine similarity	12	5

(Continued)

Table 14 (continued)

Keyword	Link strength	Frequency	Keyword	Link strength	Frequency
graph convolutional networks	12	6	privacy protection	12	5
watermarking	12	8	intrusion detection	9	6
conditional gan	9	5	zero-shot learning	9	5
image restoration	10	5	qr code	6	5

Fig. 22 illustrates the connections between GANs and cybersecurity. From this figure, it is evident that GANs are central to deepfake research, particularly in the context of fake images, face manipulation, and image forensics. Additionally, misinformation, fake news, and phishing detection are connected to cybersecurity and deepfake-based scams, such as impersonation. Keywords such as *phishing, social engineering*, and *phishing attacks* indicate various cybersecurity threats, while terms such as *spoof detection, data security, privacy protection, intrusion detection,* and *watermarking* highlight techniques used to mitigate these cybersecurity challenges.

Figure 22: Links for GAN (left) and cybersecurity detection (right) in cluster 4

Similarly, the presence of keywords like *radiology* and *medical imaging* provides insights into domains where detecting fake images is critical. It is important to note that GANs are not only used for deepfake generation but also play a significant role in deepfake detection and the classification of images as real or fake [136–138].

Another notable issue in this cluster is the challenge of *imbalanced data*, which can negatively impact deepfake detection performance. In particular, deepfake detection backbone models trained on biased or imbalanced datasets may yield inaccurate detection results, leading to concerns about security, fairness, and generalizability [139].

The concept of *security* in relation to deepfakes is a key focus of this cluster. As shown in Fig. 23 (left), deepfake is directly linked to security, while Fig. 23 (right) shows security connected to other terms within the cluster, such as *privacy* and biometrics.

Additionally, this cluster includes keywords related to artifact classification, such as *graph networks* [140,141] and *cosine similarity*, which can be used to measure inter-sample relationships [142].

6.5 Cluster 5

Table 15 presents the associated keywords for Cluster 5, organized based on their link strength and frequency of occurrence. The link strengths indicate that *fake news* and *natural language processing (NLP)*, particularly for detecting fake news, are highly prominent in this cluster. Additionally, the prevalence of fake news during the *COVID-19* pandemic and the widespread use of *BERT* models for *fake news detection* are also significant themes. Numerous studies have utilized BERT for this purpose [143–147], with particular emphasis on detecting COVID-19-related misinformation [148–150]. More broadly, NLP remains a well-established approach for fake news detection [151].

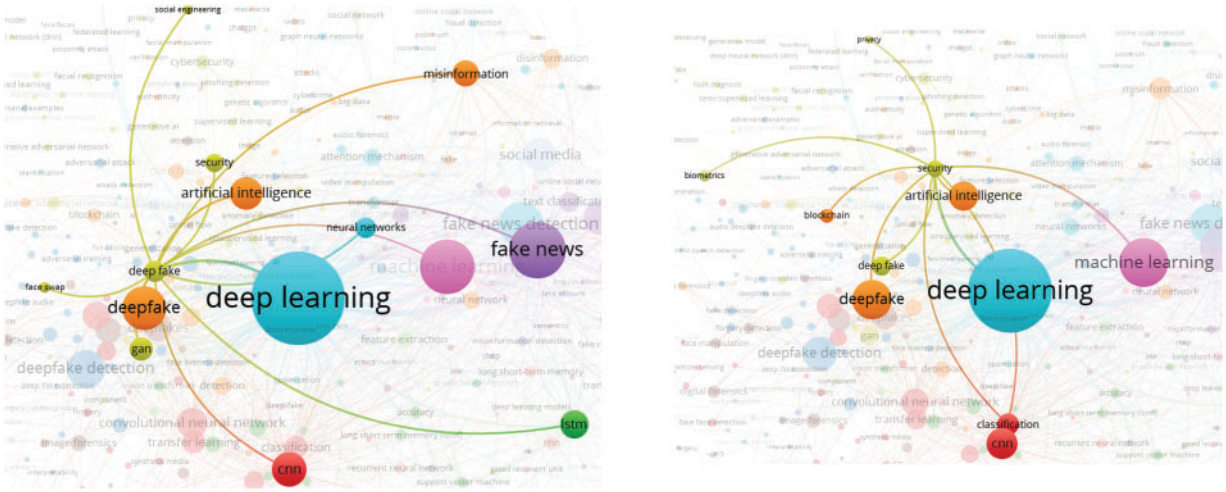

Figure 23: Links for deepfake (left) and security (right) in cluster 4

Table 15: Keywords in Cluster 5 ordered by total link strength (descending)

Keyword	Link strength	Frequency	Keyword	Link strength	Frequency
Fake News	1830	597	Natural Language Processing	849	243
COVID-19	385	104	BERT	376	106
Twitter	246	63	Text Classification	258	77
Sentiment Analysis	236	75	Transformers	173	49
Deep Neural Network	141	59	Word Embedding	124	36
Misinformation Detection	70	21	Word Embeddings	68	22
Deep Learning (DL)	60	25	Fake News Classification	62	22
RoBERTa	54	16	Hate Speech	49	14
Stance Detection	47	17	Infodemic	45	13
Social Networking	41	8	Rumor	41	12
Explainable AI	40	13	Machine Learning (ML)	38	12

(Continued)

Table 15 (continued)

Keyword	Link strength	Frequency	Keyword	Link strength	Frequency
Multimodal Fusion	34	11	Explainability	35	12
Capsule Network	30	13	Data Analysis	28	9
Arabic Language	28	8	Topic Modeling	27	9
Tweets	26	5	Opinion Mining	26	8
DistilBERT	25	6	Fake News Identification	25	9
Bag of Words	24	5	False Information	24	7
SHAP	24	7	Visualization	24	6
Fake Account	22	5	Pre-trained Models	22	6
Sentiment	22	6	Pandemic	21	7
Hate Speech Detection	20	5	Transformer Models	19	5
Bangla	19	5	XAI	18	6
Linguistic Features	18	6	Active Learning	17	8
Cyberbullying	17	5			

Fig. 24 illustrates the connections between fake news and natural language processing. The significance of fake news in this cluster is evident from its strong links with multiple keywords, including *false information, infodemic, stance detection, sentiment analysis, data analysis, BERT, Twitter, RoBERTa, word embeddings*, and *transformers*. Additionally, fake news is linked to keywords from other clusters, such *as fact-checking, information disorder, deception detection, rumors, recurrent neural networks (RNNs), synthetic media, CNNs, detection, ensemble learning, accuracy, support vector machines (SVMs), feature extraction, attention mechanisms, cybersecurity, journalism, social networks, neural networks, fake news detection, disinformation*, and *text mining*. This highlights the widespread concern over fake news and the extensive range of machine learning-based methods proposed to address it.

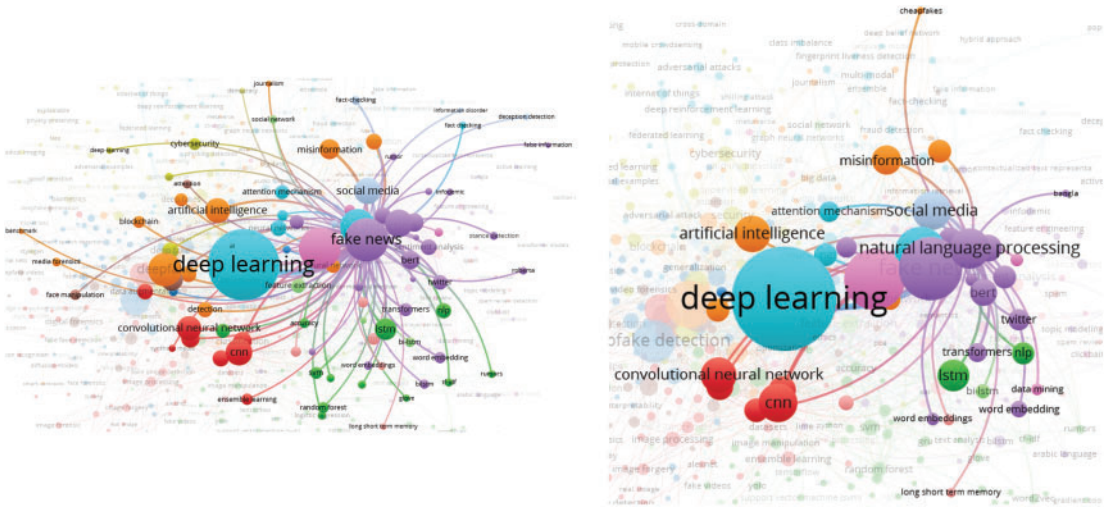

Figure 24: Links for fake news (left) and natural language processing (right) in cluster 5

Since natural language processing is strongly connected to fake news, they share numerous related keywords, including *sentiment analysis, transformers, word embeddings, BERT*, and *Twitter* within the same cluster. Additionally, keywords from other clusters, such as *text mining, attention mechanisms, social media, misinformation, disinformation, LSTMs*, and *CNNs*, are also linked to NLP and fake news. Furthermore, this cluster reveals the presence of research addressing fake news in different languages, such as *Bangla* and *Arabic*, as well as the role of social media platforms like Twitter in its dissemination.

The prominence of BERT and research related to COVID-19 is further illustrated in Fig. 25. Notably, keywords connected to COVID-19 include pandemic, text classification, sentiment analysis, Twitter, misinformation detection, misinformation, and disinformation, reflecting the surge of research on identifying misinformation during the pandemic. Other notable keywords in this cluster include hate speech, sentiment, and cyberbullying, which may be linked to misinformation. In particular, deepfakes have been used as tools for cyberbullying, resulting in social, educational, and psychological consequences [152].

Additionally, this cluster includes *multimodal fusion*, a technique that enhances fake news detection [153]. The presence of *explainable artificial intelligence (XAI)*, along with explainability mechanisms such as *SHapley Additive exPlanations (SHAP)* [154], suggests that XAI is widely utilized for misinformation and deepfake detection [155,156]. Furthermore, pre-trained models such as *RoBERT*a and *BERT* [157] are frequently used to detect the widespread dissemination of misinformation on social media platforms like Twitter [158].

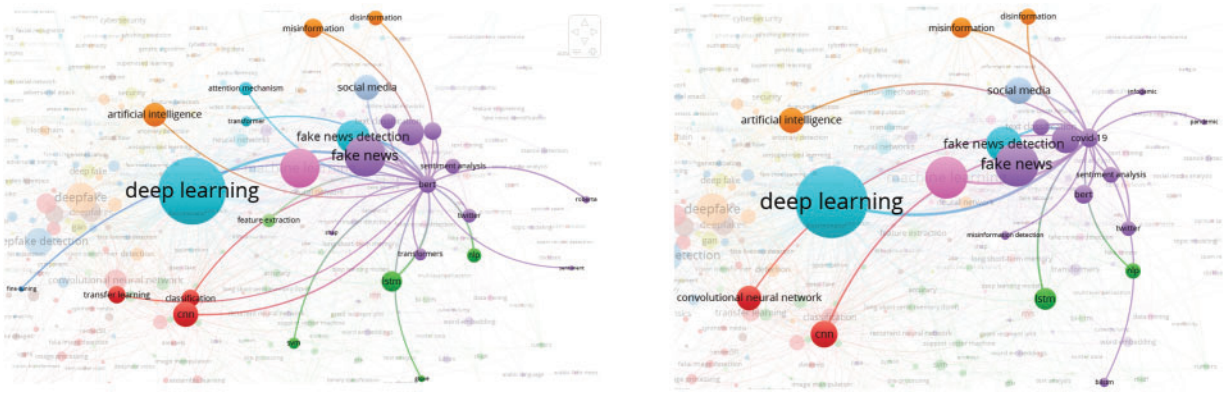

Figure 25: Links for BERT (left) and Covid-19 (right) in cluster 5

6.6 Cluster 6

Table 16 presents the associated keywords for Cluster 6, organized based on their link strength and frequency of occurrence. At the core of this cluster, and the broader research landscape of deepfakes is *deep learning*, as evident from Fig. 26 (left) and its high link strength of 4472. Similarly, *neural networks*, the backbone of deep learning, are also prominent in this cluster, with a link strength of 209. An even more dominant keyword than *neural networks* in this cluster is *fake news detection*, as shown in Fig. 26 (right). Additionally, *multimodal fake news detection* is observable in this cluster, emerging as a recent research hotspot [159–162].

Table 16: Keywords in Cluster 6 ordered by total link strength (descending)

Keyword	Link strength	Frequency	Keyword	Link strength	Frequency
Deep Learning	4472	1672	Fake News Detection	1007	390
Neural Networks	209	74	Attention Mechanism	146	57
Convolution Neural Networks	81	244	Transformer	90	33
Rumor Detection	52	22	Feature Fusion	50	22
Internet of Things	51	20	Adversarial Attacks	44	19
Fact Checking	43	14	Fraud Detection	38	14
Optimization	33	12	Multi-Modal	33	15
Knowledge Graph	32	10	Ensemble	24	9
Coronavirus	23	6	Genetic Algorithm	23	7
Class Imbalance	22	8	Graph Neural Networks	22	12
Recommender System	17	10	Reinforcement Learning	16	11
Biometric	16	6	Multi-Modal Fusion	16	5
Language Model	16	6	Recognition	15	6
Fake Information	15	6	Black-Box Attacks	15	6
Fingerprint	14	9	Mobile Crowdsensing	14	5
Liveness Detection	13	8	Deep Learning Methods	12	6
Deep Reinforcement Learning	12	5	Popularity Prediction	12	5
Adversarial Networks	11	6	Identification	11	6
Deep Learning Technique	10	5	Poisoning Attack	10	6
Collaborative Filtering	10	5	Cross-Domain	9	5
Deep Belief Network	9	5	Attack Detection	7	6
Shilling Attack					

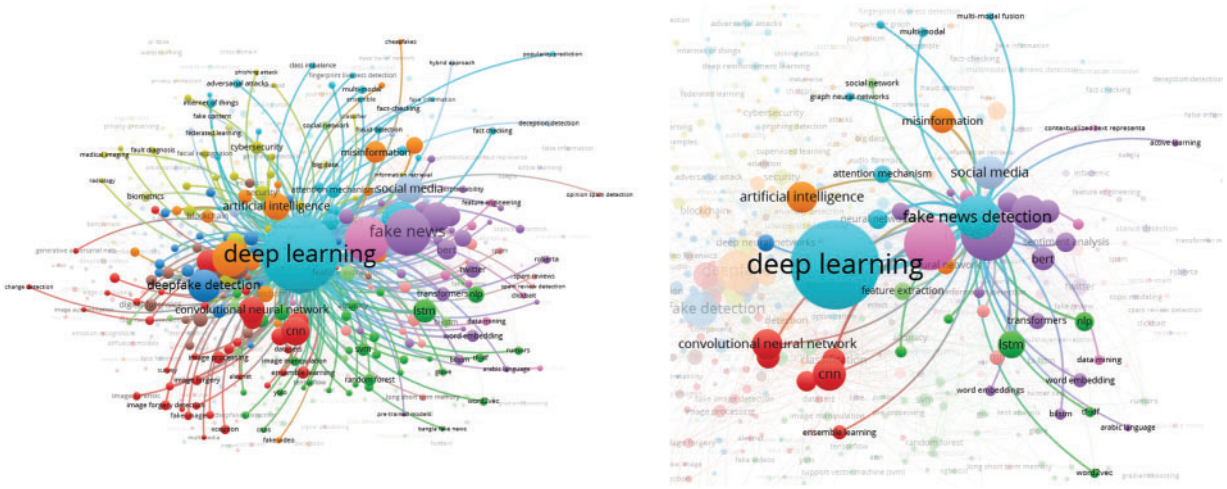

Figure 26: Links for fake news (left) and natural language processing (right) in cluster 6

Furthermore, *attention mechanisms* and *transformers* are among the most widely used deep learning-based methods in deepfake research. This cluster includes various machine learning algorithms and frameworks employed in the detection of deepfakes and fake news, such as *adversarial networks, attention mechanisms, convolution, convolutional neural networks (CNNs) [163], deep belief networks, deep learning, deep learning methods [164,165], deep learning techniques [166], deep reinforcement learning, ensemble learning [167,168], feature fusion [165,169], graph neural networks [170], knowledge graphs, language models [171], neural networks [164], optimization, reinforcement learning*, and *transformers* [172,173].

Similarly, various types of attacks and security threats are evident in this cluster, including *adversarial attacks, black-box attacks*, and *poisoning attacks* [174–177]. These attacks, particularly adversarial and poisoning attacks are designed to prevent fake news detection models from correctly identifying misinformation [174,177]. The cluster also includes mechanisms for attack identification and detection, as indicated by keywords such as *attack detection, fake news detection, fact-checking, liveness detection, fraud detection, identification, recognition*, and *rumor detection*, all of which contribute to identifying deepfakes and fake news.

Researchers have also explored the integration of *biometrics* into deepfake detection [178]. Detecting deepfake modifications in biometric images using neural networks and other technologies is crucial for ensuring the security of biometric authentication systems [179]. Additionally, the presence of keywords such as *fingerprint liveness detection* highlights an advanced method that differentiates real fingerprints from artificial replicas, which pose a security threat to fingerprint-based biometric systems [180].

Other notable keywords in this cluster include *collaborative filtering, recommendation systems*, and *shilling attacks*. Collaborative filtering (CF) is a widely used recommendation system technique that suggests content based on users' preferences. However, CF systems are vulnerable to *shilling attacks* (also known as profile injection attacks), where attackers alter recommendation results by injecting fake user profiles [181,182].

Additionally, *ensemble learning* and *genetic algorithms* appear in this cluster, as both methods can be deployed for detecting fake imagery [183]. Another critical challenge identified in this cluster is *class imbalance*, which must be carefully addressed to ensure datasets are balanced and free from bias. A well-balanced dataset is essential for creating a fair training environment, preventing deepfake detection models from producing inaccurate results [184].

6.7 Cluster 7

Table 17 presents the associated keywords for Cluster 7, organized based on their link strength and frequency of occurrence. The most prominent keywords in this cluster include *deepfake, artificial intelligence, misinformation, computer vision, detection, neural networks*, and *blockchain*. Fig. 27 (left) illustrates the connections between *deepfake* and various keywords both within and outside its cluster. Within its own cluster, *deepfake* is linked to terms such as *misinformation, artificial intelligence, manipulation, detection, authentication, authenticity*, and *fake media*. Outside the cluster, it is connected to keywords from other clusters, including *CNN, digital forensics, forgery detection, fake image detection, cybersecurity, LSTM*, and *GAN*, among others.

Table 17: Keywords in Cluster 7 ordered by total link strength (descending)

Keyword	Link strength	Frequency	Keyword	Link strength	Frequency
Deepfake	870	362	Artificial Intelligence	589	196
Misinformation	451	125	Computer Vision	219	74
Detection	209	71	Neural Network	184	60
Deep Fakes	130	51	Blockchain	110	40
Multimodal	93	34	Forensics	86	25
Authentication	55	21	Fake	57	21
Image	51	14	Autoencoders	44	14
Manipulation	40	11	Generative Models	37	16
News	34	9	Journalism	29	9
Benchmark	29	7	Video	28	9
Cheapfakes	26	6	Trust	25	9
Authenticity	23	10	Verification	23	6
Fake Video	23	9	Video Manipulation	18	8
Hoax	18	5	Technology	17	5
Audio	16	8	Smart Contracts	16	5
Attacks	15	7	Synthetic Speech	15	6
Fake Media	14	6	Media	14	6
Deception	13	9	Post-Truth	12	5
Facial Manipulation	9	6	Metaverse	11	5

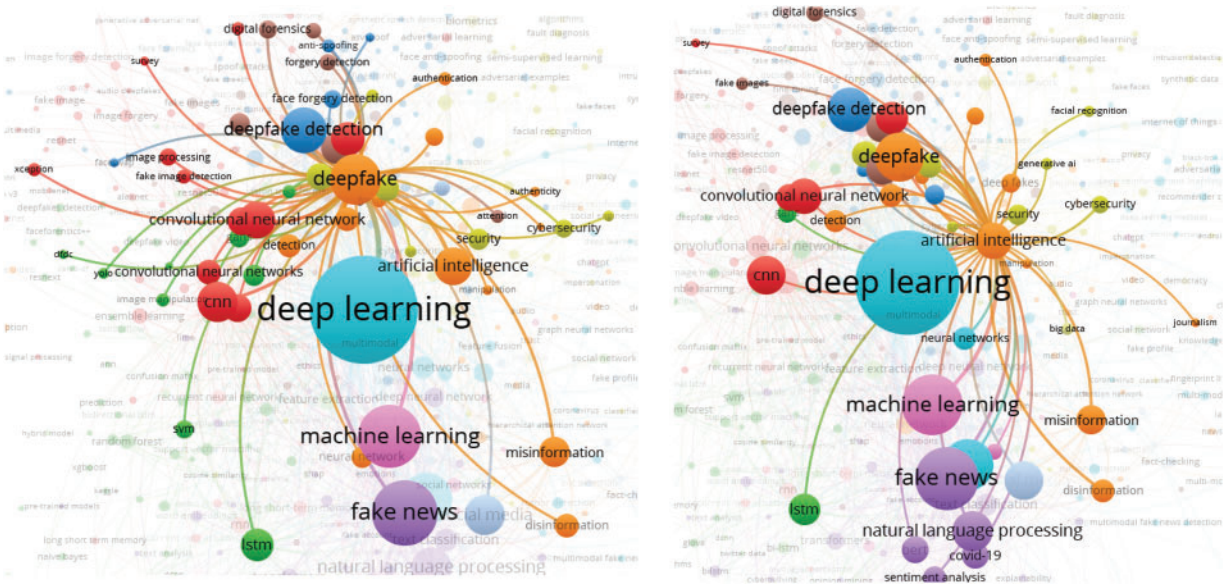

Figure 27: Links for deep fake (left) and artificial intelligence (right) in cluster 7

Similarly, *artificial intelligence* is linked to *deepfake* as well as to other terms within its cluster, such as *journalism, authentication, misinformation, disinformation, detection, blockchain*, and *forensics*. Additionally, it is connected to keywords from other clusters, including *CNN, LSTM, fake news, natural language processing, COVID-19, generative AI*, and *face recognition*. These connections highlight the extensive role of artificial intelligence in deepfake detection research.

Fig. 28 illustrates the connections associated with *misinformation* and *computer vision*. As shown in the left side of the figure, *misinformation* is linked to *cheapfakes*, *news*, *disinformation*, *deepfakes*, and *artificial intelligence* within its cluster. Additionally, it is connected to *fact-checking*, *fake news*, *rumors*, *social media*, *Twitter*, *COVID-19*, *machine learning*, *BERT*, *natural language processing*, *transformer*, *LSTM*, *classification*, and *deepfake detection*. These linked keywords highlight various ways misinformation can spread and the methods used to detect it. In the right side of the figure, *computer vision* is closely linked to *face recognition, digital forensics, fake news, misinformation*, and various machine learning-related terms, including *deep learning, artificial intelligence, natural language processing, transfer learning*, and *machine learning*. The keyword *benchmark* is also featured in this cluster, emphasizing the importance of standardized benchmarks in deepfake detection research for ensuring fair performance comparisons and accurate results [185].

Several keywords in this cluster reflect concerns regarding deepfakes and their various forms, including *deepfake, fake video, video manipulation, facial manipulation, fake media*, and *cheapfakes*. These different forms of deepfakes span multiple modalities, including *video, image, audio*, and *synthetic speech*. Additionally, some keywords highlight the potential consequences of deepfakes, such as *attacks, deception, disinformation, misinformation, hoaxes, post-truth*, and *trust*.

This cluster also includes keywords that indicate the areas most vulnerable to deepfakes, such as *media, journalism, news*, and the *metaverse*. Furthermore, various technical tools used for both the generation and detection of deepfakes are present in this cluster, including *artificial intelligence, neural networks, generative models, autoencoders*, and *computer vision*. Generative models, particularly *generative adversarial networks* (GANs), are widely used for deepfake generation [186]. However, they have also been employed for detecting deepfakes, particularly in cases involving social media images and voice manipulation [186,187].

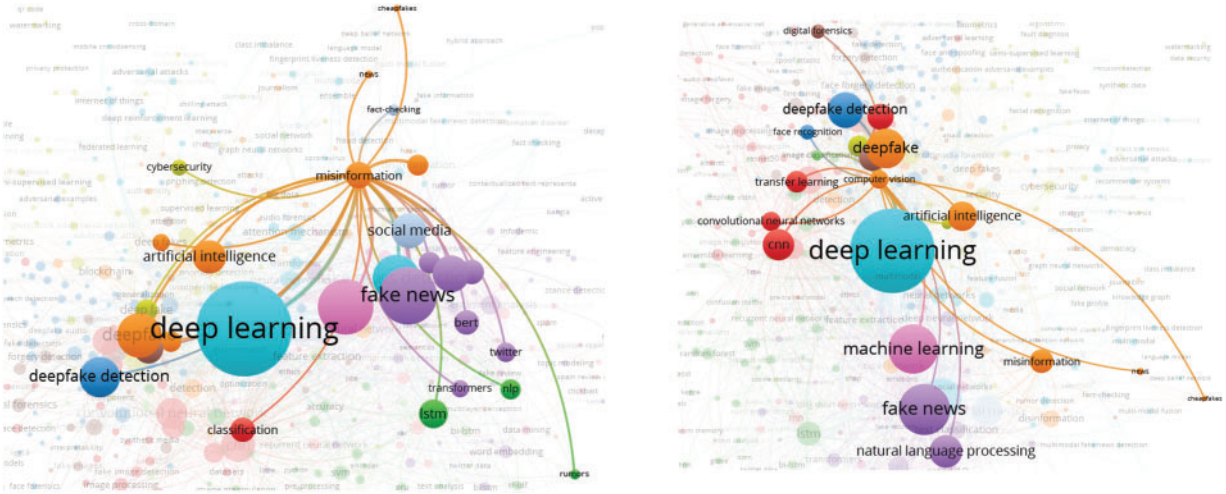

Figure 28: Links for misinformation (left) and computer vision (right) in cluster 7

Finally, keywords such as *blockchain* and *smart contracts* are associated with the security aspects of deepfake research. These technologies play a crucial role in ensuring authentication and trust, offering potential solutions for addressing deepfake-related concerns.

6.8 Cluster 8

Table 18 presents the associated keywords for Cluster 8, organized based on their link strength and frequency of occurrence. The most predominant keywords in this cluster are *deepfakes, generative adversarial networks* (GANs), *image forensics*, and *digital forensics*, as indicated by their link strength and frequency. This prominence is also evident in Figs. 29 and 30. Among these, *generative adversarial networks* have the highest link strength in this cluster. GANs are connected to *digital forensics* and *autoencoders* within the cluster, while they are linked to *adversarial training, semi-supervised learning, convolutional neural networks* (CNNs), and the *attention mechanism* in other clusters.

Table 18: Keywords in Cluster 8 ordered by total link strength (descending)

Keyword	Link strength	Frequency	Keyword	Link strength	Frequency
Deepfakes	407	154	Generative Adversarial Networks	306	156
Convolutional Neural Networks	244	81	Image Forensics	134	56
Digital Forensics	133	40	Face Manipulation	105	29
Forgery Detection	90	34	Fake Images	70	22
Deep Fake Detection	59	27	Attention	58	20
Video Forensics	47	22	Optical Flow	30	12
Image Synthesis	31	13	Face Swapping	32	11
Deepfake Videos	28	14	EfficientNet	28	10

(Continued)

Table 18 (continued)

Keyword	Link strength	Frequency	Keyword	Link strength	Frequency
Social Media Platforms	24	9	Deepfake Generation	45	9
News Classification	17	5	Deep Convolutional Neural Networks	20	12
Fusion	19	5	Video Forgery	20	5
Classification Algorithms	16	7	Hierarchical Attention Networks	15	6
Synthetic Images	15	5	Image Recognition	15	5
Super Resolution	14	6	Bidirectional Encoder Representations	14	5
Auto-Encoder	13	5	Swin Transformer	13	5
Image Reconstruction	12	6	Texture	12	5
Image Splicing	11	5	Detection Techniques	10	6
Digital Media Forensics	9	6	Semi-Supervised	9	5
Video Forgery Detection	8	6	Deep Generative Model	8	6

GANs serve multiple roles in deepfake generation and digital forensics. They can enhance image and video quality, generate synthetic data, and create realistic deepfakes. In digital forensics, GANs contribute to improving machine learning algorithms, making them more robust in scenarios where specialized training data is lacking or prior knowledge about attacks is unavailable. Specifically, GANs can be used to augment existing datasets with synthetic samples, enhancing the generalizability of forensic classifiers [188].

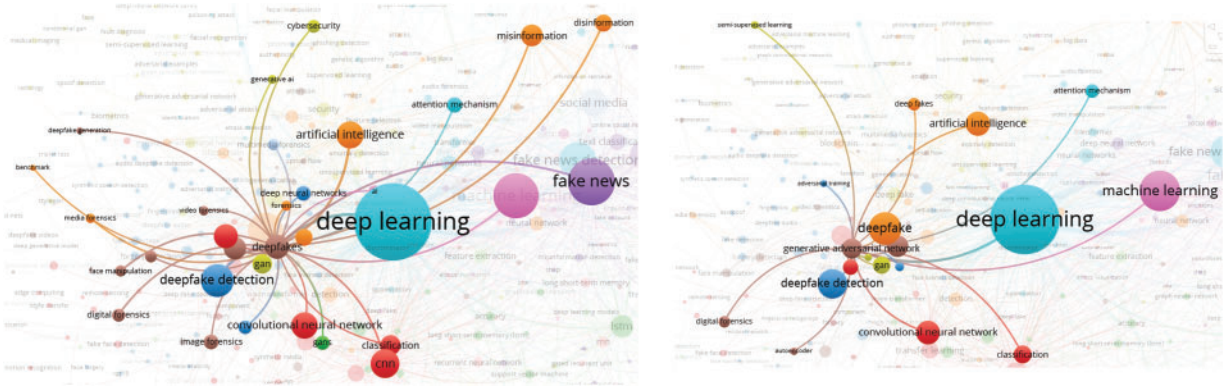

Figure 29: Links for deep fake (left) and generative adversarial networks (right) in cluster 8

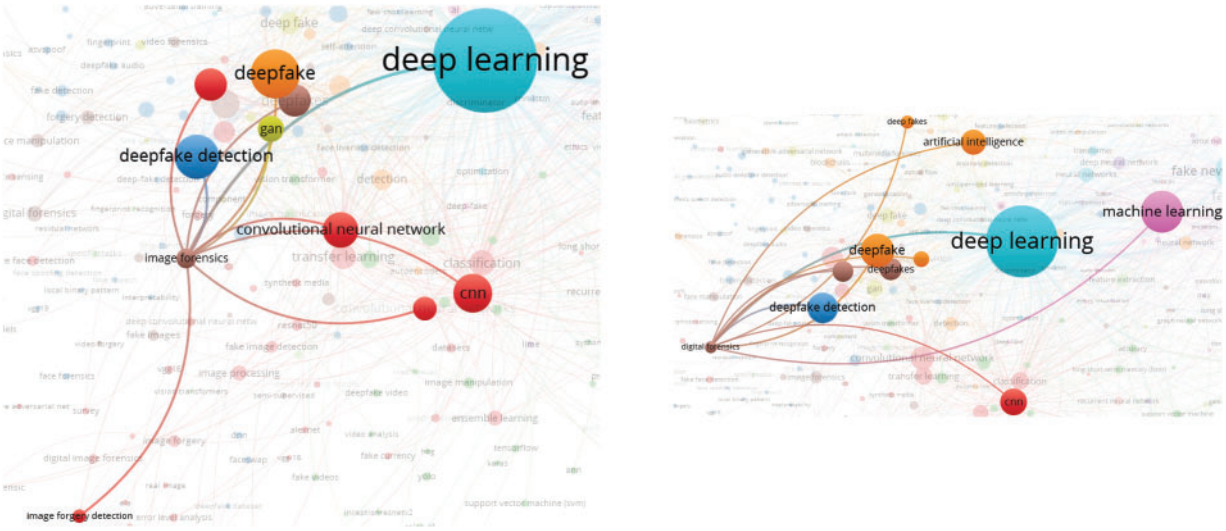

Figure 30: Links for image forensics (left) and digital forensics (right) in cluster 8

The process of deepfake creation involves the use of neural network architectures such as *autoencoders* and *GANs*, which learn and replicate facial features and expressions. Once trained, these models facilitate *face swapping, blending*, and *post-processing*, resulting in highly realistic deepfakes [189]. Additionally, deep generative models, when combined with deep neural networks, have been extensively used for generating deepfakes [190].

Several algorithms used in deepfake detection are also highlighted in this cluster, including *EfficientNets* [191,192], *convolutional neural networks* [193], and *bidirectional encoder representations* for fake news detection [194] or news classification. Furthermore, *hierarchical attention networks* have been employed for multimodal detection in social networks and social media platforms [195].

Beyond deepfake generation and its associated processes, this cluster includes numerous keywords related to deepfake detection and forensics. These include *deepfake detection, detection techniques, forgery detection, image forensics, image recognition, video forensics, video forgery detection, digital forensics*, and *digital media forensics*. Various techniques and algorithms for deepfake detection are also present in this cluster, such as *fusion* [123], *texture analysis* [196], *optical flow* and *optical flow CNN* [197,198], and the *Swin Transformer* [52].

Additionally, several keywords in this cluster pertain to image manipulation, including *face manipulation* and various techniques for altering visual content, such as *splicing, image splicing*, and *face swapping* [199]. The growing concern over deepfakes in social media is also evident in this cluster, as indicated by the presence of keywords such as *social media platforms, fake images, face swapping*, and *face manipulation*.

6.9 Cluster 9

Table 19 presents the keywords associated with Cluster 9, organized based on their link strength and frequency of occurrence. Machine learning is the most prominent keyword in this cluster, with a link strength of 1794 and a frequency of 562. Its presence is also illustrated in Fig. 31 (left).

Table 19: Keywords in Cluster 9 ordered by total link strength (descending)

Keyword	Link strength	Frequency	Keyword	Link strength	Frequency
Machine learning	1794	562	Social networks	112	33
AI	98	29	Text mining	89	20
Fake review detection	55	25	Data mining	56	17
Feature engineering	39	12	Fake review	37	13
Clickbait	34	9	Large language models	30	10
Cybercrime	29	9	Artificial Intelligence (AI)	26	10
Fake accounts	24	7	Data science	24	5
Social media analysis	22	7	Deep-Fake	21	9
Emotions	19	7	Arabic fake news	16	7
ChatGPT	16	8	Ethics	16	7
Internet	12	5	Contextualized text representations	13	5
Opinion spam	10	6	–	–	–

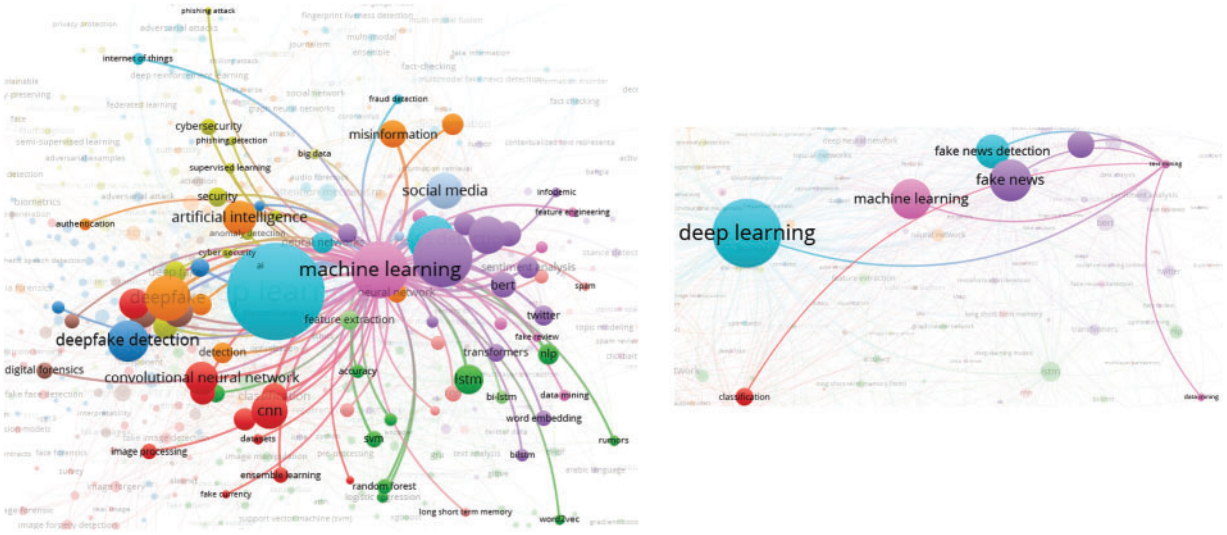

Figure 31: Links for machine learning (left) and text mining (right) in cluster 9

Within the cluster, machine learning is closely connected to keywords such as data mining, text mining, feature engineering, social networks, fake review, fake review detection, and AI. These keywords highlight the applications of machine learning in social network analysis, text and data mining, and fake review detection. Notably, feature engineering plays a crucial role in performing sentiment analysis on social media text, as demonstrated in the Twitter use case reported in [200]. In particular, sentiment analysis is essential for determining the emotional tone of text [201].

Furthermore, machine learning is utilized for distinguishing machine-generated (or deepfake) text from human-generated text, as well as for detecting social media spam [202,203]. Beyond its cluster, machine learning is linked to a wide range of keywords, including IoT, phishing attack, cybersecurity, big data, fraud detection, misinformation, social media, infodemics, sentiment analysis, BERT, Twitter, Transformer, word embedding, BiLSTM, rumor, word2vec, LSTM, feature extraction, accuracy, SVM, random forest, logistic regression, ensemble learning, fake currency, dataset, image processing, digital forensics, authentication, cybersecurity, anomaly detection, supervised learning, phishing detection, and spam. These keywords indicate various applications of machine learning in deepfake research, different machine learning algorithms, in addressing social challenges such as cybersecurity threats.

Overall, this cluster primarily focuses on the deployment of machine learning and data science techniques for detecting fake news, fake reviews, cybercrime, opinion spam, fake accounts, and other forms of manipulation and attacks on social media. Given the internet's critical role in the spread of fake news, deepfakes, and cybercrime, social media analysis remains a significant research area. Additionally, the application of text mining in fake news detection is evident in this cluster, as illustrated in Fig. 31. Keywords such as AI, feature engineering, and clickbait further highlight the role of artificial intelligence in detecting misleading content, as exemplified in [204].

6.10 Cluster 10

Table 20 presents the associated keywords for clusters 10–13, organized based on their link strength and frequency of occurrence. In cluster 10, *Recurrent Neural Networks* (RNN) and *Long Short-Term Memory* (LSTM) emerge as predominant keywords, as illustrated in Fig. 32.

Table 20: Keywords in Cluster 10–13 ordered by total link strength (descending)

Cluster	Cluster Keywords Statistics			Cluster Keywords Statistics		
	Keyword	**Link strength**	**Frequency**	**Keyword**	**Link strength**	**Frequency**
10	RNN	183	43	Long short-term memory	103	30
	Long short term memory	49	11	Recurrent neural networks	49	16
	Fake reviews	89	30	Gated recurrent unit	56	14
	Spam	55	11	Spam reviews	35	7

(Continued)

Table 20 (continued)

Cluster	Cluster Keywords Statistics			Cluster Keywords Statistics		
	Keyword	Link strength	Frequency	Keyword	Link strength	Frequency
11	Spam review detection	29	5	PCA	25	6
	Features	23	9	Credibility	16	6
	Semantics	11	5			
	Feature extraction	206	61	bi-LSTM	102	27
	Image classification	63	23	Social network	42	18
	Pre-processing	39	13	Spoof attacks	28	9
	Dcgan	26	13	Representation learning	25	12
	Twitter data	21	5	Fasttext	19	5
	Multimodal learning	18	7	Iris recognition	18	6
	Hybrid model	17	5	Vision transformers	13	6
	DCNN	11	5	Fingerprint recognition	10	6
12	Social media	691	222	Fact-checking	74	20
	Social network analysis	50	16	Bot detection	27	8
	Audio forensics	25	12	Fake content	23	9
	Deception detection	20	8	fake profile	15	8
	Information disorder	14	5	hybrid approach	13	5
	Fake faces	12	5			
13	Biometrics	91	28	big data	48	17
	Privacy	35	12	federated learning	32	14
	Face	32	9	algorithms	29	7
	Privacy-preserving	25	8	fair	21	5
	Gesture	20	6			

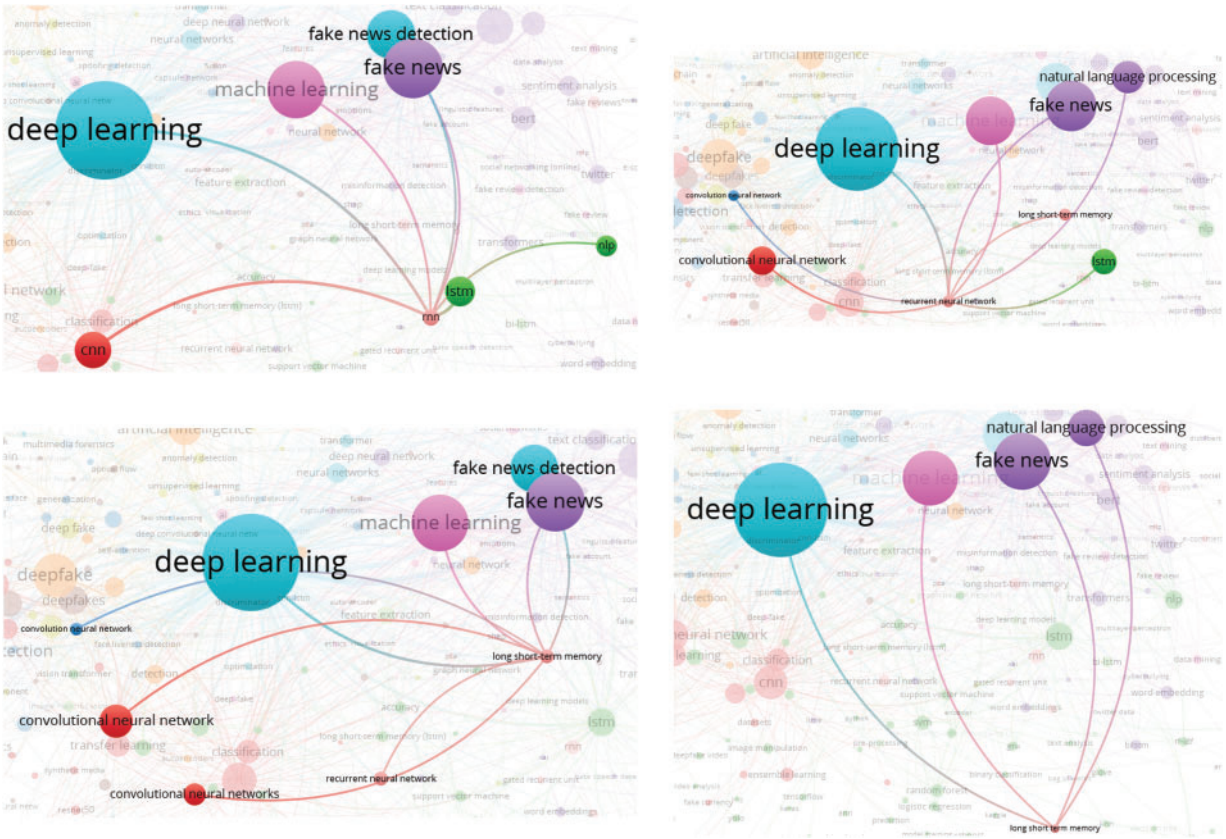

Figure 32: Links for recurrent neural networks (top) and long short term memory in cluster 10

Spam reviews, which pose a significant threat to e-commerce platforms by misleading consumers into poor decisions, necessitate the use of deep learning models for spam detection. Notably, models such as *LSTM* and *GRU*, particularly when hybridized, as demonstrated in [205], are effective in detecting spam reviews. Similarly, these deep learning models, including *LSTM* and *RNN*, can be leveraged for spam email classification [206]. Furthermore, *RNN* and *LSTM* are valuable for sentiment analysis, which is crucial in identifying fake or spam reviews [201].

To apply these models effectively for spam detection, capturing semantic meaning and contextual information is essential, as demonstrated for SMS messages in [207]. Additionally, *feature extraction* techniques are required to represent messages effectively before deploying classifiers [208]. Feature dimensionality reduction also plays a critical role in identifying the minimal optimal set of features necessary for spam email detection. In this regard, *Principal Component Analysis* (PCA) is among the key feature selection techniques used in this domain [209]. Thus, the deployment of these algorithms is instrumental in filtering and ensuring credible reviews.

6.11 Cluster 11

Table 20 also presents the associated keywords for cluster 11, organized based on their link strength and frequency of occurrence. In this cluster, *feature extraction* and *Bi-LSTM* exhibit the highest link strength. Notably, *feature extraction* is closely linked with *pre-processing* within its cluster, as illustrated in Fig. 33. *Feature extraction* is a fundamental pre-processing technique essential for *deepfake* detection. Beyond its cluster, it is also associated with keywords such as *attention mechanism*, indicating the role of attention

mechanisms in deepfake research involving feature extraction. Additionally, *feature extraction* is connected to *classification*, *social media*, *social networking*, *fake news*, and *fake news detection*, further highlighting its significance in identifying deepfakes and misinformation.

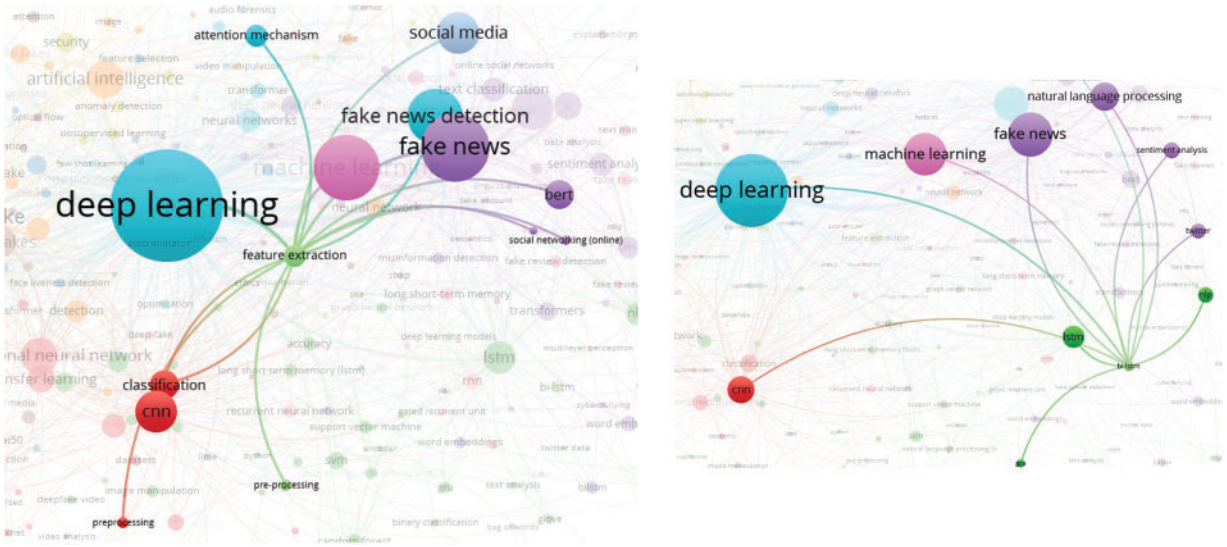

Figure 33: Links for feature extraction (left) and bi-long short term memory in cluster 11

Another predominant keyword in this cluster, as shown in Fig. 33, is *Bi-LSTM*, which is linked to *LSTM*, *CNN*, *GRU*, *natural language processing*, *sentiment analysis*, *fake news*, and *social media*. This indicates that *Bi-LSTM* and other deep learning algorithms play a crucial role in sentiment analysis and the detection of fake news and misinformation on social media. Moreover, *Bi-LSTM* is a deep learning model useful for detecting *deepfakes* [101] and can also serve as a classifier in multimodal biometric systems, providing protection against spoofing attacks [210].

The keywords in cluster 11 underscore the importance of *feature extraction* in *deepfake* research. Several models used for feature extraction, such as the *Swin Transformer*, have been identified as effective deep learning models for multi-modal deepfake detection [52]. Additionally, *vision transformers* can be employed for the classification of extracted features, particularly in *face recognition* [211]. Similarly, Twitter data, which often includes both text and images, necessitates multimodal approaches for detecting fake tweets [212].

Another noteworthy model, *DCGAN*, is a combination of *GAN* and *CNN* that is used to generate high-quality photorealistic images. It also has applications in *face recognition* and advancements in biometric system authentication [213]. Additionally, *DCGAN*, as a deep learning model, can be deployed for detecting *deepfakes*, particularly for voice recognition [214].

Deep CNNs play a crucial role in multimedia deepfake detection by analyzing facial features, speech patterns, and contextual information to identify manipulated videos [215]. Similarly, models such as *FastText* are useful for *sentiment analysis*, enabling the classification of text on social media platforms such as Twitter [216,217].

In the domain of biometric security, *iris recognition*-based systems are susceptible to breaches such as spoof attacks [218,219]. Therefore, *anti-spoofing* mechanisms are essential to determine whether an iris trait is genuine or fake. The adoption of machine learning has significantly improved spoof detection, as models can learn from training samples to assess the liveliness of an image [220]. The process of building

such models involves *pre-processing*, *feature extraction*, and a *classifier*. Generally, *deep learning* plays a vital role in *iris recognition* [221], and efficient *pre-processing* enhances prediction accuracy in iris recognition applications [220].

6.12 Cluster 12

The keywords in cluster 12 are organized based on their link strength and frequency of occurrence, as shown in Table 20. The most prominent keyword in this cluster is *social media*, with a link strength of 691, due to its strong interconnections with several keywords across other clusters as shown in Fig. 34 (left). For instance, *social media* is linked to *fact-checking*, *multi-modal*, *disinformation*, *misinformation*, *attention*, *AI*, *feature extraction*, *classification*, *fake news*, *BERT*, *transformers*, *Twitter*, *NLP*, *BI-LSTM*, *text classification*, and *COVID-19*. This highlights the various processes, algorithms, and concerns associated with the intersection of deepfakes and social media.

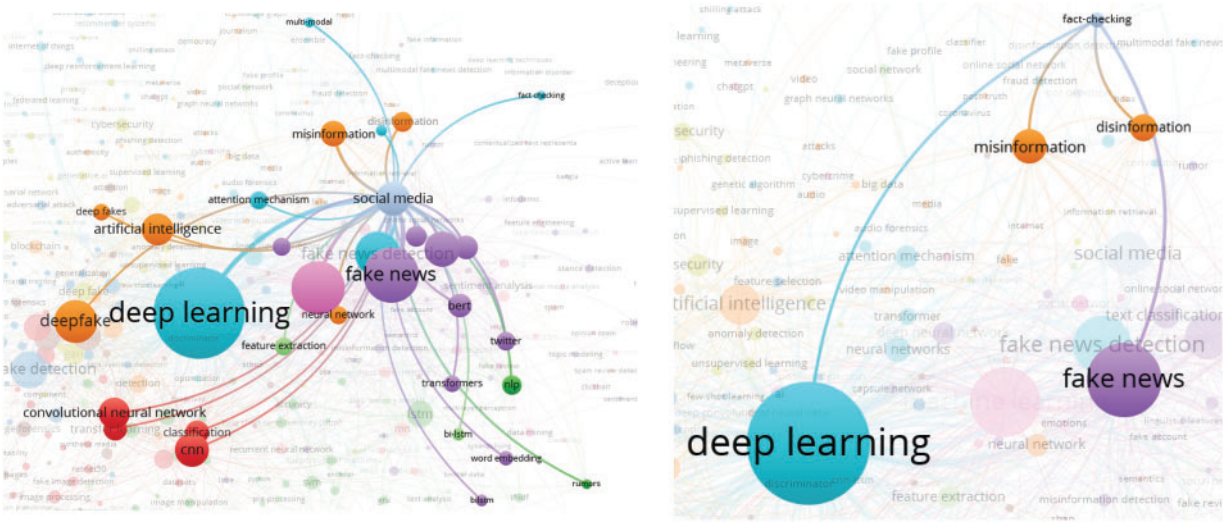

Figure 34: Links for social media (left) and fact-checking in cluster 12

Other keywords in this cluster include *deception detection*, *information disorder*, *fake faces*, *fake profiles*, *fake content*, *fact-checking*, and *social network analysis*, all of which relate to the presence and detection of deepfakes on social media. Similarly, keywords such as *audio forensics* pertain to detecting deepfakes, including manipulated audio content that may be uploaded to social media platforms.

Detecting *multimodal deepfakes* requires robust solutions, often necessitating hybrid approaches. Additionally, deep learning algorithms play a crucial role in detecting *social bots* within online social networks [222].

Another prominent keyword in cluster 12 is *fact-checking*, which is closely linked to *misinformation*, *disinformation*, *fake news*, and *deep learning* as shown in Fig. 34 (right). This connection underscores the importance of deep learning techniques in *fact-checking*, helping to prevent the spread of false or misleading information.

6.13 Cluster 13

The keywords in cluster 13 are organized based on their link strength and frequency of occurrence, as shown in Table 20. The most prominent keyword in this cluster is *biometrics*. As illustrated in Fig. 35,

biometrics is linked to *gesture*, *face*, *algorithms*, and *security* within the same cluster. Additionally, it is connected to keywords from other clusters, such as *fingerprint*, *liveliness detection*, *spoofing*, *authentication*, *deep learning*, and *CNN*.

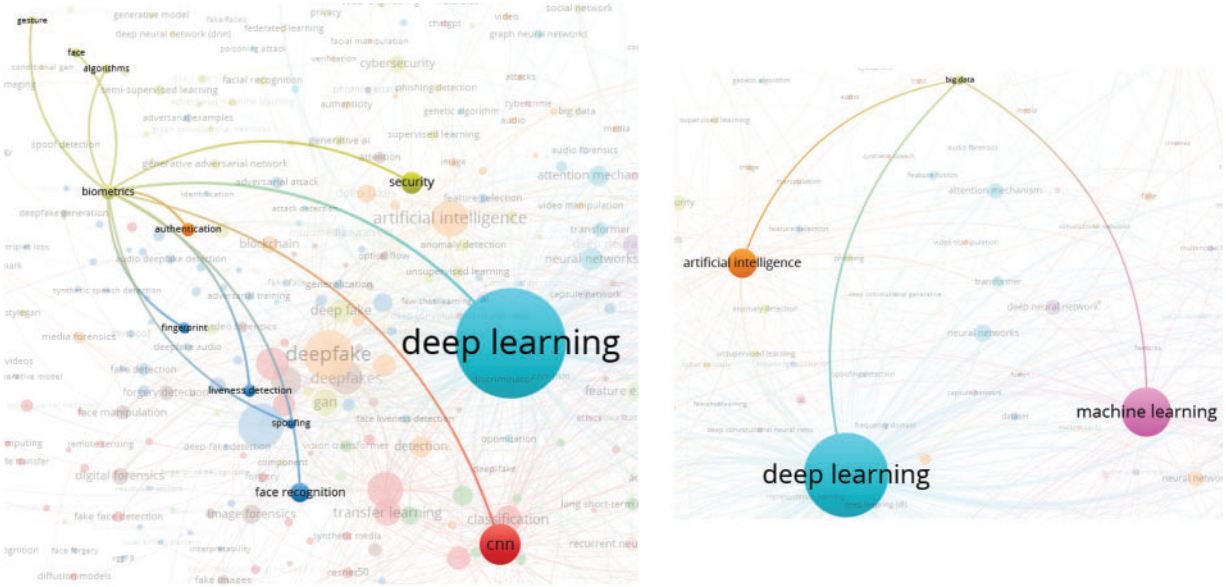

Figure 35: Links for biometrics (left) and big data in cluster 13

Other notable keywords in this cluster include *democracy*, *fair*, *federated learning*, *privacy*, and *privacy-preserving*. Notably, some of these keywords highlight the social implications of deepfake prevalence, particularly *privacy*, *privacy-preserving*, *democracy*, and *fairness*. Similarly, several technological keywords are associated with deepfakes, including *algorithms*, which are essential for deepfake detection, and *big data*, as large datasets are often required to evaluate the effectiveness of deepfake detection models.

One approach to achieving *privacy preservation* in deepfake research is through *federated learning [223]* which can be used to create a secure training strategy that protects local data privacy [224]. This concept and the discussed keywords show the intersection between deepfake research, security and ethics in the modern society.

6.14 Summary, Trends, Challenges, and Recommendations Based on the Review

Deepfake detection, generation, and their social implications constitute a large portion of the technical discussion of deepfakes in literature. This technical discussion involves audio, video, and textual deepfakes as well as multi-modal deepfakes. Particularly, deep learning, neural network, and their associated methods are most prominent in deepfake detection. Also, the area of computer vision is one of the most relevant areas to deepfake research. For a deepfake generation, the use of generative adversarial networks and their architectural variants is also prominent. Convolutional neural network is one of the most popular algorithms used for deepfake detection and classification. Similarly, domain adaptation methods, emotion recognition, gated recurrent units, contrastive learning, EfficientNet, the Triplet Loss approach, XGBoost, autoencoders, and attention mechanisms are all techniques used in deepfake detection, each contributing from different technical perspectives.

The process of deepfake detection and improving detection methods requires comprehensive datasets, proper feature extraction, data augmentation, big data processing, addressing data and class imbalance, and model fine-tuning. Similarly, proper segmentation, classification, and change detection are all required for deepfake detection. Technical Requirements of deepfake models that are desired include transferability, explainability, generalization, improved dataset utilization and data analysis, robustness, efficient preprocessing, efficient text mining, and accuracy. Proper benchmarking is also required for high-quality results.

Deepfake has a lot of implications as keywords such as fake media, fake account(s), fake content, fake currency, fake face, fake image, fake information, fake news, fake profile, fake account, fake reviews, fake speech, fake video, false information, impersonation, information disorder, cybercrime all indicate the negative implications of deepfakes which needs to be urgently addressed. Other concerns about deepfakes span across privacy, journalism, its impact on emotions, security, hate speech, cybersecurity, cyberbullying, the credibility of information, the authenticity of information, trust, presence of click baits, bot detection, spam messages, and fake currency. All these represent concerns that need to be addressed especially with the prevalence of deepfakes.

Deepfake challenges in investigation and forensics all require the improvement of deepfake detection solutions, cross-disciplinary efforts as well as advancements in computer vision and artificial intelligence. Such advancements should involve advanced model architectures such as transformers, multitask and multimodal learning, pre-trained architectures, federated learning, zero-shot and meta-learning improvements in explainable AI, semi-supervised learning, and deep learning architectures. Also, fusion methods, leveraging hybrid solutions, feature engineering, and feature extraction methods (e.g., bag of words and NLP model) are all vital.

Technical Models peculiar to deepfake research (such as deepfake detection and image recognition) include: ANN, CNN, Deep DCNN, ResNet/ResNet50/ResNext, VGG/VGG16/VGG19 EfficientNet/EfficientNetB0, Xception/XceptionNet, Inception V3/InceptionResNetV2, Recurrent Neural Networks (RNN), LSTM/Gated Recurrent Unit (GRU), BiLSTM, Multilayer Perceptron (MLP), Transformer Models/Transformers.

Emerging trends in deepfake research and challenges include the following:

- Cheapfakes: These are low-cost fake media that are easy to create and are gaining attention due to their accessibility.
- Edge Computing: Leveraging edge computing to provide decentralized and efficient computational facilities for running deepfake detection models.
- IoT and Deepfake Research: Exploring the intersection between the Internet of Things (IoT) and deepfake research, such as IoT-based security systems to mitigate deepfake threats.
- Detection evasion: Addressing attacks like black-box attacks that aim to evade deepfake image detection models.
- Social Media Bot Detection: Enhancing bot detection techniques on social media platforms using AI-driven approaches.
- Resource-Constrained Environments: Tackling the challenges of detecting and managing deepfakes in environments with limited computational resources.
- Application-Specific Solutions: Developing targeted solutions for specific applications, such as detecting deepfakes in medical images, including radiographic imaging.
- Implications for the Metaverse: Investigating how the rise of the metaverse could escalate deepfake challenges if these issues are not adequately addressed.
- Social Media Misinformation: Reducing the impact of deepfakes and fake information on widely available social media platforms.

- Fake Reviews: Addressing fraud stemming from fake reviews on social media platforms, which poses a significant concern for consumers and businesses alike.

Recommendations for addressing the challenges posed by deepfakes include the following:

- Dataset Availability: Improving the availability of high-quality deepfake datasets for research purposes across all modalities, including text, audio, and video.
- Detection Tools: Developing and distributing tools capable of detecting deepfakes with a high level of accuracy, including leveraging freely available large language models (LLMs) for deepfake detection.
- Social Media and Misinformation: (1) Preventing social media attacks and the spread of misinformation through stricter enforcement of policies. (2) Creating attack-proof social media platforms and encouraging content verification to discourage the intersection of AI-generated and deepfake content with social media.
- Research Funding: Increasing funding for research into deepfake detection and related methods such as active learning, adversarial training, attention mechanisms, capsule networks, contrastive learning, deep belief networks, reinforcement learning, representation learning, self-supervised learning, semi-supervised learning, supervised learning, unsupervised learning, and transformer models.
- Interdisciplinary Research: Promoting research in complementary areas such as data analytics, statistics, and machine learning.
- Broader Detection Efforts: Advancing research on the detection of fake information, phishing, media forensics, malware, social engineering, and cybercrime.
- Computational Resources: Ensuring the availability of high computational power for solving complex deepfake-related problems.
- Multilingual Focus: Addressing the emergence of deepfakes in non-English languages, such as Bangla and Arabic, which are gaining traction. With the increasing popularity of deepfakes, their spread in other languages, especially in the context of fake text, is likely.

Finally, we have uploaded all the keywords to a text analysis tool (Voyant Tools) to examine the frequency of the respective keywords. We only include the words with a frequency of over 100 to show the most popular research aspects and concerns in deepfake research. The results confirm the prevalence of deep learning, neural networks, CNN, adversarial networks, adversarial attacks, adversarial learning, adversarial training, concerns with respect to media and forensics, issues of forgery, misinformation, and attacks. Also, popular methods other than CNN include GAN, LSTM, BERT. deep learning (2397), detection (1914), fake (1790), news (1202), network (1001), neural (904), adversarial (606), generative (521), social (483), face (464), convolutional (462), media (402), classification (398), CNN (374), processing (370), language (366), data (336), analysis (319), forensic (296), artificial (265), LSTM (261), video (235), forgery (232), feature (230), GAN (217), model (217), text (200), recognition (193), attention (182), misinformation (166), attack (165), security (163), information (158), audio (149), manipulation (145), digital (144), vision (136), transfer (135) models (135), BERT (128), AI (128), COVID (123), Images (120), Generation (119), dataset (113), attacks (112), transformer (110), graph (110), spoofing (108), speech (105), ensemble (104), sentiment (102), features (102), extraction (101).

On the other side of the spectrum, we study the keywords with a frequency between (5) and (10) which gives an idea of some of the most emergent trends, methods, and concerns. These include conspiracy (5), counterfeiting (5), contract (5), generalizability (5), justice (5), metaverse (5), oversampling (5), outlier (5), RCNN (5), Vgg19 (5), android(6), cyberbullying (6), DCNN (6), googlenet (6), keras (6), offensive (6), sarcasm (6), unet (6), confusion (7), discrimination (7), DNNs (7), encryption (7), epidemic (7), ethical (7), forged (7), fraudulent (7), impersonation (7), misleading (7), mobilenet (7), xceptionnet (7), mtcnn (7), pretrained (7), spoofed (7), uncertainty (7), belief (8), chatgpt (8), integrity (8), normalization (8), spammer

(8), StyleGAN (8), violence (8), bots (9), crime (9), confidence (9), densenet (9), faceswap (9), eye (9), finger (9), gait (9), policy (9), regulation (9), vectorizer (9), XGBoost (10), celebrity (10), cybercrime (10), cycleGAN (10), banknote (10), evidence (10), threat (10), and interpretability (10).

In all, the research on deepfakes spans various domains and contributions, and concerted efforts are required in the fields of Computer Science, Engineering, Mathematics, Data Science, Decision Science, Social Sciences, as well as multidisciplinary research encompassing other disciplines.

7 Recommendations for Addressing Deepfakes

The spread of deepfakes poses a significant threat to society, and thus it is important to recommend policies for addressing their harmful effects. This has been discussed with respect to policymakers, researchers, and other practitioners, such as tech industries and media outlets.

7.1 Recommendations for Policymakers

In this section, we discuss policy recommendations for controlling and regulating deepfakes and their harmful consequences.

7.1.1 Development of Regulatory Frameworks, Media Literacy, and International Cooperation

The integration of AI into communication systems has garnered significant interest across various fields, including journalism, marketing, and diplomacy. Although AI could offer opportunities for improving diplomatic processes, the widespread of deepfakes and their misuse threaten and undermine trust in diplomatic engagements. Therefore, policymakers should promote media literacy initiatives to counter the influence of deepfakes, encourage technological advancements in deepfake detection tools, and provide a comprehensive regulatory framework for deepfakes. Additionally, enhanced international cooperation is required to combat the threats posed by deepfakes across borders and to empower individuals in discerning deepfake propaganda. By implementing these recommendations, policymakers can protect the integrity of diplomatic processes and mitigate the risks associated with AI misuse [225].

Developing laws to control deepfakes is crucial, as even small but strategic changes in legal regimes can yield effective protection against unauthorized deepfakes [226]. Thus, proposing regulatory tools that consider the rights of all entities involved in deepfake creation and dissemination should be prioritized. This includes protecting individuals whose original artifacts have been used in the generation of deepfakes. Detailed and well-implemented legal enactments can go a long way in regulating the misuse of AI technologies.

7.1.2 Utilization of State-of-the-Art Detection Technologies for Law Enforcement

Law enforcement must have access to advanced, state-of-the-art technologies for accurately detecting deepfakes. This need is especially pressing given the rapid evolution of deepfake generation algorithms [227]. Therefore, the development of explainable forensic algorithms that integrate human expertise into the detection loop is highly desirable.

Deepfake-generated media is multifaceted in both nature and impact. Since its applications span technological, social, economic, and political domains, state-of-the-art detection mechanisms for all deepfake types are essential. A holistic approach, integrating technical solutions, public awareness, and legislative action is necessary. Furthermore, unified, real-time, adaptable, and generalized solutions for deepfake detection are critical as the challenges posed by deepfakes continue to intensify [228].

7.1.3 Promotion of Public Awareness on Digital Literacy

Education plays a crucial role in helping the public, particularly youth, develop resilience against malicious deepfakes and counter disinformation. Therefore, more targeted educational programs on deepfakes for young people are highly recommended. Educators, curriculum developers, and policymakers should leverage these programs to ensure that both current and future generations are well-equipped to protect society from the plague of disinformation [229].

Information asymmetries are on the rise, imposing significant societal costs across different demographics [230]. Consequently, actionable policymaking recommendations and educational strategies are necessary to address the spread of harmful deepfake content. Policies should ensure an equitable distribution of authentic information and promote media literacy. Moreover, stakeholders must navigate the ethical dilemmas posed by deepfakes while ensuring equitable access to digital information to enhance discernment, decision-making, and awareness.

Policymakers should also recognize the importance of increasing access to advanced information technologies while addressing their repercussions. Efforts to disseminate knowledge about deepfakes should particularly target individuals with limited or no access to information and communication infrastructures. Learning from past successes and failures will help shape more effective strategies to counter deepfake-related challenges [230].

Additionally, addressing information asymmetry is critical due to disparities in how different age groups are exposed to and affected by disinformation. Research indicates that the likelihood of falling for disinformation increases with age. Therefore, policymakers, social scientists, and technology companies all have significant roles to play in mitigating these risks.

7.1.4 Identification of Risks and Development of Adaptive Policies and Regulations

Addressing gaps in our understanding of deepfakes is essential for facilitating timely and effective regulatory action. Deepfakes have the potential to amplify existing societal problems, such as disinformation, making supervision, enforcement of rules, and necessary policy adjustments vital. Consequently, further research is required to examine the societal challenges posed by deepfakes and the need for adaptive policies [231].

Regulations should also account for text-based deepfakes, as advancements in natural language processing and large language models have increased the potential for manipulating textual content, shaping online discourse, and spreading misinformation [228].

Furthermore, disclosure policies regarding the use of synthetic media are critical, as transparency significantly impacts public perception and credibility [232]. Researchers, policymakers, and practitioners involved in deepfake-related synthetic media should be well-informed about its implications. Laws should be enacted to address the negative consequences of such media.

7.1.5 Implementation of Comprehensive Deepfake Regulation

Relative to the number of countries in the world, very few regulatory frameworks are available. Although the EU's Artificial Intelligence Act introduces regulations on deepfakes, it should be amended to better prevent deepfake-associated risks such as blackmail, abusive content, misinformation, and emotional or financial harm. Swift action is needed to facilitate deepfake detection by classifying AI systems intended for deepfake creation as high-risk. In addition to clear definitions and resilient safeguards, these measures would ensure more effective deepfake regulation. Policymakers should adopt these amendments for the betterment of society [233].

Laws to prevent the unchecked harms of AI are crucial, as these harms include cultural anxiety, racial polarization, and cyberattacks, particularly as synthetic video and audio content gain increasing public attention [234]. Therefore, policymakers, activists, and technology companies must act swiftly to regulate AI. Other countries should collaborate to establish a unified AI Act that mitigates the harmful effects of deepfakes.

Deriving lessons from high-profile deepfake incidents in the past, researchers, practitioners, and policymakers must engage in continuous innovation to counter the rapidly evolving deepfake landscape [235]. In addition, the establishment of clear guidelines for reporting AI abuse in an evidence-based manner is essential for ensuring that penalties can be effectively implemented [236].

7.2 Recommendations for Researchers

In this section, we provide recommendations for researchers about addressing deepfakes.

7.2.1 Prioritizing Evidence-Based Research

The generation of fake textual, audio, and visual content poses a significant societal threat to trust, political stability, and information integrity [228]. Addressing deepfakes requires solutions that span technological, economic, social, and political domains. Therefore, comprehensive research on deepfakes is essential to propose integrative solutions, enhance public awareness, and inform legislative actions.

7.2.2 Advancing Scientific Research in Deepfake Detection

From a scientific standpoint, current research limitations in deepfake detection include challenges in cross-modality detection. Researchers should prioritize innovations in this area to counter the rapidly evolving landscape of deepfakes [235]. Additionally, robust detection algorithms capable of identifying even minor artifacts introduced by generative algorithms must be developed [227].

Furthermore, explainable forensic techniques, which integrate human judgment into the detection loop, can enhance accurate decision-making [227]. As deepfake generation technologies continue to advance, malicious actors increasingly weaponize the internet. Unfortunately, existing tools to detect, measure, and mitigate these threats remain insufficient. Therefore, developing advanced tools to prevent and protect against deepfake threats should be a research priority [237].

Researchers must also analyze the strengths and weaknesses of current deepfake detection techniques, evaluate their effectiveness, and monitor their evolution over time. Such efforts will provide policymakers with a clearer understanding of the current technological landscape and highlight areas requiring further development [238].

Addressing the privacy and security challenges inherent in generative AI and deepfakes is crucial. Consequently, improvements in AI architectures, model designs, security strategies, and sustainable solutions must involve collaboration between developers, institutions, policymakers, and users [239]. Additionally, to enhance the effectiveness of deepfake detection systems, it is important to investigate why specific content is flagged as deepfake and how detection mechanisms can be refined.

Although much research has focused on deepfake detection, mitigating the dissemination and propagation of deepfake content is equally vital. A multidisciplinary approach, encompassing expertise from machine learning, computer vision, cybersecurity, and media forensics is necessary to comprehensively address these challenges [240].

7.2.3 Enhancing Legal, Policy and Social Science Research on Deepfakes

From a legal perspective, regulatory responses to deepfakes must be critically assessed at a global level. This involves a thorough analysis of policy and legal documents [231]. This way best practices can be adopted and improvements can be made.

For social science researchers, research on deepfakes should prioritize rigorous, evidence-based analysis to accurately assess their impact. Research should focus on demographic factors such as age, gender, ethnicity, and ideology that influence an individual's susceptibility to misinformation [241]. Understanding these factors can help in designing more targeted awareness campaigns and educational initiatives.

Moreover, social cynicism plays a crucial role in how people perceive the credibility of deepfake sources [232]. Studies indicate that the public holds negative perceptions of deepfakes across both social and non-social media platforms [242]. Policymakers and other stakeholders can leverage this awareness to further educate the public about the harms of deepfakes and implement preventive measures.

7.3 Recommendations for Practitioners (Tech and Media Industry)

Practitioners such as those involved in the Tech and Media industry have a large role to play in the mitigation of the spread of deepfakes and their negative consequences. This involves developing robust deepfake detection tools useful for media practitioners and enhancing reporting mechanisms.

7.3.1 Develop Robust AI and Content Authentication Tools

Developers of social media platforms and news agencies should create robust deepfake detection mechanisms to safeguard against the spread of misinformation and harmful online content [237]. Furthermore, social media platforms, policymakers, and governments must recognize the potential risks posed by the widespread propagation of deepfakes. Understanding these threats requires an analysis of the actors involved, their motives, and the varied responses necessary to combat them [237]. Consequently, it is crucial to develop models that track the origin, spread, virality, and effects of deepfakes on targeted individuals and society at large [237].

7.3.2 Enhance Social Media Moderation and Reporting Mechanisms

To prevent the weaponization of the internet for spreading misinformation and harmful content, stronger platform moderation policies must be established [228]. Additionally, collaboration between social media companies, fact-checkers, and independent watchdog organizations is essential to enhance the accuracy and speed of misinformation detection.

7.3.3 Adopt Transparent Disclosure Policies

Organizations, particularly media and marketing bodies that use synthetic media, should transparently disclose their use of AI-generated content [232]. Moreover, tech companies should develop and adhere to industry-wide best practices regarding the ethical use of AI-generated media [243]. Establishing clear guidelines will encourage responsible innovation and reduce the risks associated with misinformation.

Since multiple factors (such as social, political, and economic) influence the adoption of new technologies such as generative AI, policymakers, media professionals, and the general public must be informed about the potential risks of deepfakes. Therefore, responsible innovation should be a central theme in media discourse to ensure ethical AI development and deployment [243].

7.3.4 Implement Robust Cybersecurity Solutions

The misuse of AI presents a significant cybersecurity threat, highlighting the need for finance leaders and cybersecurity professionals to develop adaptive strategies for mitigating AI-driven scams and cyberattacks. AI is increasingly being exploited in cybercrime, including enhanced phishing and Business Email Compromise (BEC) attacks, automated hacking strategies, and the proliferation of black-market AI tools on the dark web. To effectively combat these threats, enhanced cybersecurity strategies and international cooperation are required. Finance leaders, cybersecurity professionals, policymakers, and researchers must deepen their understanding of the cybersecurity challenges posed by generative AI and explore the most effective ways to mitigate these risks [244].

7.3.5 Train Media Professionals and Other Stakeholders

The rapid advancement of deepfake technology underscores the urgency for policymakers and tech companies to implement stronger moderation practices for synthetic media content. Studies show that individuals are highly susceptible and may likely not recognize fake videos [245]. For this reason, it is crucial to provide support for media personnel and other stakeholders to ensure they are equipped with the necessary information needed to use tools and discern real from fake media.

7.4 Current Work Limitations

Several potential constraints could affect the comprehensiveness, and generalizability of this study, including database selection, search keywords, and time frame. While we used WoS, a well-known and trusted database, other peer-reviewed papers indexed in reputable databases like Scopus may not have been captured. Identifying precise search keywords also poses challenges, as efforts to align keywords with the study's objectives may not encompass the entire scope, particularly when researchers use uncommon or technical terms. Additionally, two different search dates (29th August and 1st November 2024) were used for reporting bibliometric findings and investigating top cited works, respectively, reflecting the state of the WoS database at those times. The selection criteria may have excluded papers offering valuable insights, and focusing solely on academic literature may have overlooked sources like white papers and technical reports. Finally, while this study utilizes the PRISMA framework, designed for systematic and meta-analyzes [43,44], its application to bibliometric reviews presents challenges. PRISMA's checklist is tailored for systematic reviews, making some items less relevant for bibliometric studies. This highlights the need for PRISMA guidelines specifically adapted to bibliometric reviews.

8 Conclusion

This paper provides a bibliometric analysis of deepfake technology by providing a comprehensive exposition of leading countries, leading authors, research collaborations, most influential institutions, and key themes associated with deepfake research. Using VOSviewer visualization tool on data extracted from WoS database, we take a closer look into some of the most popular keywords associated with deepfake research. These keywords are mapped into four discussed themes: deepfake detection, feature extraction, face recognition, and deepfake forensics. Based on the results of the analysis, artificial intelligence-based algorithms have proven to be the predominant tool used in deepfake detection studies. This is evident from the various machine learning models, and detection techniques been identified. For instance, the popularity of artificial intelligence and neural networks, and their derivatives such as generative adversarial networks (GANs), transformers, convolutional neural networks (CNNs), self-supervised learning, and transfer learning have also been observed. The importance of security in the deepfake research area can also be observed with the presence of keywords such as cybersecurity, adversarial attacks, and anti-spoofing. In addition,

the ethical concerns associated with deepfakes are also evident with keywords such as forgery, information integrity, and fake news. The results also show the important features expected in proposed solutions such as generalization, robustness, and accuracy. Similarly, the significance of databases and deepfake datasets can also be identified from the keywords, database, and deepfake dataset. This research shows that while several other efforts have been made to review prior works on deepfakes, the fast advancement in this area and growing international collaboration among, academics, institutions, and nations warrants continuous efforts at understanding the trends, challenges, and solutions in this area.

Using VOSviewer, we have mapped the major themes and provided insights into the interconnections between key concepts in the deepfake literature. Furthermore, we also performed a more comprehensive search on the Scopus database to further explore deepfake research areas and provided an analysis of the findings including methods, implications, applications, concerns, requirements, challenges, models, tools, datasets, and modalities of deepfakes. We hope that researchers, legislators, and industry stakeholders can get valuable insight from the discussions and recommendations to effectively navigate the moral, societal, and technological issues raised by deepfake technology. This paper seeks to provide readers with a broad understanding of deepfake research, highlighting its societal importance and effects, alongside its technical complexity, research trends, and challenges.

Acknowledgement: The authors wish to express their appreciation to the reviewers for their helpful suggestions which greatly improved the presentation of this paper.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm their contribution to the paper as follows: Study Conception and Design: Oluwatosin Ahmed Amodu, Akanbi Bolakale AbdulQudus, Umar Ali Bukar, Raja Azlina Raja Mahmood; Data Collection: Akanbi Bolakale AbdulQudus, Oluwatosin Ahmed Amodu, Raja Azlina Raja Mahmood, Anies Faziehan Zakaria; Analysis and Interpretation of Results: Akanbi Bolakale AbdulQudus, Oluwatosin Ahmed Amodu; Draft Manuscript Preparation: Akanbi Bolakale AbdulQudus, Oluwatosin Ahmed Amodu, Raja Azlina Raja Mahmood; Review and Editing: Oluwatosin Ahmed Amodu, Raja Azlina Raja Mahmood, Umar Ali Bukar, Anies Faziehan Zakaria, Zurina Mohd Hanapi; Illustrations: Akanbi Bolakale AbdulQudus, Oluwatosin Ahmed Amodu, Raja Azlina Raja Mahmood, Saki-Ogah Queen; Supervision: Oluwatosin Ahmed Amodu; Funding: Oluwatosin Ahmed Amodu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The authors confirm that the data supporting the findings of this study are available within the article.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Forum WE. Global risks 2024: disinformation tops global risks 2024 as environmental threats intensify. [Internet]. [cited 2025 Apr 6]. Available from: https://www.weforum.org/press/2024/01/global-risks-report-2024-press-release/.
2. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. arXiv:1406.2661. 2014.
3. Chesney B, Citron D. Deep fakes: a looming challenge for privacy, democracy, and national security. Calif L Rev. 2019;107:1753.
4. Pantserev KA. The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability. In: Jahankhani H, Kendzierskyj S, Chelvachandran N, Ibarra J, editors. Cyber defence in the

age of AI, smart societies and augmented humanity. advanced sciences and technologies for security applications. Cham: Springer; 2020. doi:10.1007/978-3-030-35746-7_3.
5. Patel Y, Tanwar S, Gupta R, Bhattacharya P, Davidson IE, Nyameko R, et al. Deepfake generation and detection: case study and challenges. IEEE Access. 2023;11(1):143296–323. doi:10.1109/ACCESS.2023.3342107.
6. Pashentsev E. Introduction: the malicious use of artificial intelligence—Growing Threats, delayed responses. In: Pashentsev E, editor. The Palgrave handbook of malicious use of AI and psychological security. Cham: Palgrave Macmillan; 2023. doi:10.1007/978-3-031-22552-9_1.
7. Brown SD. Virtual unreality: potential implications of deepfake technology for the course of justice. ERA Forum. 2023;24(4):501–18. doi:10.1007/s12027-024-00780-1.
8. Westerlund M. The emergence of deepfake technology: a review. Technol Innov Manag Rev. 2019;9(11):40–53. doi:10.22215/timreview/1282.
9. Edwards P, Nebel JC, Greenhill D, Liang X. A review of deepfake techniques: architecture, detection and datasets. IEEE Access. 2024;12:154718–42. doi:10.1109/ACCESS.2024.3477257.
10. Li Y, Yang X, Sun P, Qi H, Lyu S. Celeb-df: a large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020; Seattle, WA, USA. p. 3207–16.
11. Zi B, Chang M, Chen J, Ma X, Jiang YG. Wilddeepfake: a challenging real-world dataset for deepfake detection. In: Proceedings of the 28th ACM International Conference on Multimedia; 2020; Seattle, WA, USA. p. 2382–90.
12. Nadimpalli AV, Rattani A. On improving cross-dataset generalization of deepfake detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022; New Orleans, LA, USA. p. 91–9.
13. Zhang T. Deepfake generation and detection, a survey. Multimed Tools Appl. 2022;81(5):6259–76. doi:10.1007/s11042-021-11733-y.
14. Heidari A, Jafari Navimipour N, Dag H, Unal M. Deepfake detection using deep learning methods: a systematic and comprehensive review. Wiley Interdiscip Rev: Data Mini Knowl Discov. 2024;14(2):e1520. doi:10.1002/widm.1520.
15. Caporusso N. Deepfakes for the good: a beneficial application of contentious artificial intelligence technology. In: Ahram T, editor. Advances in artificial intelligence, software and systems engineering. Cham: Springer International Publishing; 2021. p. 235–41.
16. Wang TC, Mallya A, Liu MY. One-shot free-view neural talking-head synthesis for video conferencing. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020; Nashville, TN, USA. p. 10034–44.
17. Gambín ÁF, Yazidi A, Vasilakos A, Haugerud H, Djenouri Y. Deepfakes: current and future trends. Artif Intell Rev. 2024;57(3):64. doi:10.1007/s10462-023-10679-x.
18. Rana MS, Nobi MN, Murali B, Sung AH. Deepfake detection: a systematic literature review. IEEE Access. 2022;10:25494–513. doi:10.1109/ACCESS.2022.3154404.
19. Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: an overview and guidelines. J Bus Res. 2021;133(5):285–96. doi:10.1016/j.jbusres.2021.04.070.
20. Kaushal A, Kumar S, Kumar R. A review on deepfake generation and detection: bibliometric analysis. Multimed Tools Appl. 2024;83:87579–619.
21. Bisht V, Taneja S. A decade and a half of deepfake research: a bibliometric investigation into key themes. In: Lakhera G, Taneja S, Ozen E, Kukreti M, Kumar P, editors. Navigating the world of deepfake technology. Hershey, PA, USA: IGI Global Scientific Publishing; 2024. p. 1–25. doi:10.4018/979-8-3693-5298-4.ch001.
22. Zhang X, Chen H, Wang W, Ordóñez de Pablos P. What is the role of IT in innovation? A bibliometric analysis of research development in IT innovation. Behav Inf Technol. 2016;35(12):1130–43. doi:10.1080/0144929X.2016.1212403.
23. Luo J, Hu Y, Bai Y. Bibliometric analysis of the blockchain scientific evolution: 2014–2020. IEEE Access. 2021;9:120227–46. doi:10.1109/ACCESS.2021.3092192.
24. Firdaus A, Razak MFA, Feizollah A, Hashem IAT, Hazim M, Anuar NB. The rise of "blockchain": bibliometric analysis of blockchain study. Scientometrics. 2019;120(3):1289–331. doi:10.1007/s11192-019-03170-4.

25. Ozyurt O, Ayaz A. Twenty-five years of education and information technologies: insights from a topic modeling based bibliometric analysis. Educ Inform Technol. 2022;27(8):11025–54. doi:10.1007/s10639-022-11071-y.
26. Donthu N, Kumar S, Pandey N, Gupta P. Forty years of the international journal of information management: a bibliometric analysis. Int J Inf Manag. 2021;57(5):102307. doi:10.1016/j.ijinfomgt.2020.102307.
27. Moral-Muñoz JA, López-Herrera AG, Herrera-Viedma E, Cobo MJ. Science mapping analysis software tools: a review. In: Springer handbook of science and technology indicators. Cham, Switzerland: Springer; 2019. p. 159–85. doi:10.1007/978-3-030-02511-3
28. Lee S, Nah K. A counterattack of misinformation: how the information influence to human being. In: Advances in intelligent systems and computing. Cham, Switzerland: Springer; 2020. Vol. 1131, p. 600–4. doi:10.1007/978-3-030-39512-4.
29. Domenteanu A, Tătaru GC, Crăciun L, Molănescu AG, Cotfas LA, Delcea C. Living in the age of deepfakes: a bibliometric exploration of trends, challenges, and detection approaches. Information. 2024;15(9):525. doi:10.3390/info15090525.
30. Pocol A, Istead L, Siu S, Mokhtari S, Kodeiri S. Seeing is no longer believing: a survey on the state of deepfakes, AI-generated humans, and other nonveridical media. In: Advances in computer graphics. Cham: Springer; 2024. Vol. 14496.
31. Gil R, Virgili-Gomà J, López-Gil JM, García R. Deepfakes: evolution and trends. Soft Comput. 2023;27(16):11295–318. doi:10.1007/s00500-023-08605-y.
32. Gunawan B, Ratmono BM, Abdullah AG, Sadida N, Kaprisma H. Research mapping in the use of technology for fake news detection: bibliometric analysis from 2011 to 2021. Indones J Sci Technol. 2022;7(3):471–96. doi:10.17509/ijost.v7i3.51449.
33. Lu Y, Liu J, Zhang R. Current status and trends in image anti-forensics research: a bibliometric analysis. arXiv: 240811365. 2024.
34. Garg D, Gill R. A bibliometric analysis of deepfakes: trends, applications and challenges. EAI Endorsed Trans Scalable Inf Syst. 2024;11(6). doi:10.4108/eetsis.4883.
35. Akram M, Nasar A, Arshad-Ayaz A. A bibliometric analysis of disinformation through social media. Online J Commun Media Technol. 2022;12(4):e202242. doi:10.30935/ojcmt/12545.
36. Ivanova M, Stefanov S. Digital forensics investigation models: current state and analysis. In: 2023 8th International Conference on Smart and Sustainable Technologies (SpliTech); 2023; Split/Bol, Croatia. p. 1–4.
37. Pranckutė R. Web of Science (WoS) and Scopus: the titans of bibliographic information in today's academic world. Publications. 2021;9(1):12. doi:10.3390/publications9010012.
38. Bukar UA, Sayeed MS, Razak SFA, Yogarayan S, Amodu OA, Mahmood RAR. A method for analyzing text using VOSviewer. MethodsX. 2023;11(1):102339. doi:10.1016/j.mex.2023.102339.
39. Eck N, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics. 2010;84(2):523–38. doi:10.1007/s11192-009-0146-3.
40. Van Eck NJ, Waltman L. Visualizing bibliometric networks. In: Measuring scholarly impact: Methods and practice. Cham: Springer; 2014. p. 285–320.
41. Aksnes DW, Sivertsen G. A criteria-based assessment of the coverage of scopus and web of science. J Data Inf Sci. 2019;4(1):1–21. doi:10.2478/jdis-2019-0001.
42. Singh VK, Singh P, Karmakar M, Leta J, Mayr P. The journal coverage of Web of Science, Scopus and Dimensions: a comparative analysis. Scientometrics. 2021;126(6):5113–42. doi:10.1007/s11192-021-03948-5.
43. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71.
44. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2007;6(7):e1000097.
45. Bukar UA, Sayeed MS, Razak SFA, Yogarayan S, Amodu OA. An exploratory bibliometric analysis of the literature on the age of information-aware unmanned aerial vehicles aided communication. Informatica. 2023;47(7):91–114. doi:10.31449/inf.v47i7.4783.

46. Vaio A, Hassan R, Alavoine C. Data intelligence and analytics: a bibliometric analysis of human-Artificial intelligence in public sector decision-making effectiveness. Technol Forecast Soc Change. 2022;174(2):121201. doi:10.1016/j.techfore.2021.121201.
47. Xu Z, Ge Z, Wang X, Skare M. Bibliometric analysis of technology adoption literature published from 1997 to 2020. Technol Forecast Soc Change. 2021;170(1):120896. doi:10.1016/j.techfore.2021.120896.
48. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Niessner M. FaceForensics++: learning to detect manipulated facial images. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019; Seoul, Republic of Korea. p. 1–11.
49. Masood M, Nawaz M, Malik KM, Javed A, Irtaza A, Malik H. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. Appl Intell. 2023;53(4):3974–4026. doi:10.1007/s10489-022-03766-z.
50. Akhtar Z. Deepfakes generation and detection: a short survey. J Imaging. 2023;9(1):18. doi:10.3390/jimaging9010018.
51. Miao C, Tan Z, Chu Q, Liu H, Hu H, Yu N. F^2trans: high-frequency fine-grained transformer for face forgery detection. IEEE Trans Inf Forensics Secur. 2023;18:1039–51. doi:10.1109/TIFS.2022.3233774.
52. Hafsa I, Ali J, Mahmood MK. AVFakeNet: a unified end-to-end Dense Swin Transformer deep learning model for audio-visual deepfakes detection. Appl Soft Comput J. 2023;136:32–3.
53. Zhao C, Wang C, Hu G, Chen H, Liu C, Tang J. ISTVT: interpretable spatial-temporal video transformer for deepfake detection. IEEE Trans Inf Forensics Secur. 2023;18(1):1335–48. doi:10.1109/TIFS.2023.3239223.
54. Yang Z, Liang J, Xu Y, Zhang XY, He R. Masked relation learning for deepfake detection. IEEE Trans Inf Forensics Secur. 2023;18:1696–708. doi:10.1109/TIFS.2023.3249566.
55. Huang B, Wang Z, Yang J, Ai J, Zou Q, Wang Q, et al. Implicit identity driven deepfake face swapping detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023; Vancouver, BC, Canada. p. 4490–9.
56. Heo YJ, Yeo WH, Kim BG. Deepfake detection algorithm based on improved vision transformer. Appl Intell. 2023;53(7):7512–27. doi:10.1007/s10489-022-03867-9.
57. Dong S, Wang J, Ji R, Liang J, Fan H, Ge Z. Implicit identity leakage: the stumbling block to improving deepfake detection generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023; Vancouver, BC, Canada. p. 3994–4004.
58. Liu X, Wang X, Sahidullah M, Patino J, Delgado H, Kinnunen T, et al. Asvspoof 2021: towards spoofed and deepfake speech detection in the wild. IEEE/ACM Trans Audio, Speech, Lang Process. 2023;31:2507–22. doi:10.1109/TASLP.2023.3285283.
59. Yang W, Zhou X, Chen Z, Guo B, Ba Z, Xia Z, et al. AVoiD-DF: audio-visual joint learning for detecting deepfake. IEEE Trans Inf Forensics Secur. 2023;18(2):2015–29. doi:10.1109/TIFS.2023.3262148.
60. Wang Y, Yu K, Chen C, Hu X, Peng S. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023; Vancouver, BC, Canada. p. 7278–87.
61. Chen H, Lin Y, Li B, Tan S. Learning features of intra-consistency and inter-diversity: keys toward generalizable deepfake detection. IEEE Trans Circuits Syst Video Technol. 2022;33(3):1468–80. doi:10.1109/TCSVT.2022.3209336.
62. Liao X, Wang Y, Wang T, Hu J, Wu X. FAMM: facial muscle motions for detecting compressed deepfake videos over social networks. IEEE Trans Circuits Syst Video Technol. 2023;33(12):7236–51. doi:10.1109/TCSVT.2023.3278310.
63. Li X, Ni R, Yang P, Fu Z, Zhao Y. Artifacts-disentangled adversarial learning for deepfake detection. IEEE Trans Circuits Syst Video Technol. 2022;33(4):1658–70. doi:10.1109/TCSVT.2022.3217950.
64. Liu D, Dang Z, Peng C, Zheng Y, Li S, Wang N, et al. FedForgery: generalized face forgery detection with residual federated learning. IEEE Trans Inf Forensics Secur. 2023;18:4272–84. doi:10.1109/TIFS.2023.3293951.
65. Lin H, Huang W, Luo W, Lu W. DeepFake detection with multi-scale convolution and vision transformer. Digit Signal Process. 2023;134:103895. doi:10.1016/j.dsp.2022.103895.

66. Chakravarty N, Dua M. A lightweight feature extraction technique for deepfake audio detection. Multimed Tools Appl. 2024;83(26):67443–67. doi:10.1007/s11042-024-18217-9.
67. Zhang W, Zhao C, Li Y. A novel counterfeit feature extraction technique for exposing face-swap images based on deep learning and error level analysis. Entropy. 2020;22(2):249. doi:10.3390/e22020249.
68. Ma J, Wang S, Zhang A, Liew AWC. Feature extraction for visual speaker authentication against computer-generated video attacks. In: 2020 IEEE International Conference on Image Processing (ICIP); 2020; Anchorage, AK, USA: IEEE. p. 1326–30.
69. Chakravarty N, Dua M. An improved feature extraction for Hindi language audio impersonation attack detection. Multimed Tools Appl. 2024;83(25):66565–90. doi:10.1007/s11042-023-18104-9.
70. Sekar RR, Rajkumar TD, Anne KR. Deep fake detection using an optimal deep learning model with multi head attention-based feature extraction scheme. Vis Comput. 2025;41:2783–800.
71. Martin-Rodriguez F, Garcia-Mojon R, Fernandez-Barciela M. Detection of AI-created images using pixel-wise feature extraction and convolutional neural networks. Sensors. 2023;23(22):9037. doi:10.3390/s23229037.
72. Yu M, Zhang J, Li S, Lei J. MSFRNet: two-stream deep forgery detector via multi-scale feature extraction. IET Image Process. 2023;17(2):581–96. doi:10.1049/ipr2.12657.
73. Bian M, Liu J, Sun S, Zhang X, Ren Y. Verifiable privacy-enhanced rotation invariant LBP feature extraction in fog computing. IEEE Trans Ind Inform. 2023;19(12):11518–30. doi:10.1109/TII.2023.3246992.
74. Agarwal A, Singh R, Vatsa M, Noore A. MagNet: detecting digital presentation attacks on face recognition. Front Artif Intell. 2021;4:643424. doi:10.3389/frai.2021.643424.
75. Tariq S, Jeon S, Woo SS. Evaluating trustworthiness and racial bias in face recognition apis using deepfakes. Computer. 2023;56(5):51–61. doi:10.1109/MC.2023.3234978.
76. Ramachandran S, Nadimpalli AV, Rattani A. An experimental evaluation on deepfake detection using deep face recognition. In: 2021 International Carnahan Conference on Security Technology (ICCST); 2021; Hatfield, UK: IEEE. p. 1–6.
77. Alshehri M. Deep fake video face recognition using supervised contrastive learning for scalability and interpretability. Arab J Sci Eng. 2024;19(6):101115. doi:10.1007/s13369-024-09676-1.
78. Tarchi P, Lanini MC, Frassineti L, Lanatà A. Real and deepfake face recognition: an EEG study on cognitive and emotive implications. Brain Sci. 2023;13(9):1233. doi:10.3390/brainsci13091233.
79. Liu YC, Chang CM, Chen IH, Ku YR, Chen JC. An experimental evaluation of recent face recognition losses for deepfake detection. In: 2020 25th International Conference on Pattern Recognition (ICPR); 2021; Online: IEEE. p. 9827–34.
80. Tariq S, Jeon S, Woo SS. Am I a real or fake celebrity? Evaluating face recognition and verification APIs under deepfake impersonation attack. In: Proceedings of the ACM Web Conference 2022; 2022; Lyon, France. p. 512–23.
81. Suganthi S, Ayoobkhan MUA, Bacanin N, Venkatachalam K, Štěpán H, Pavel T, et al. Deep learning model for deep fake face recognition and detection. PeerJ Comput Sci. 2022;8(2):e881. doi:10.7717/peerj-cs.881.
82. Castillo Camacho I, Wang K. A comprehensive review of deep-learning-based methods for image forensics. J Imaging. 2021;7(4):69. doi:10.3390/jimaging7040069.
83. Amerini I, Anagnostopoulos A, Maiano L, Celsi LR. Deep learning for multimedia forensics. Foundations Trends® Comput Graph Vis. 2021;12(4):309–457. doi:10.1561/0600000096.
84. Yadav A, Vishwakarma DK. Datasets, clues and state-of-the-arts for multimedia forensics: an extensive review. arXiv:2401.06999. 2024.
85. Caldelli R. Multimedia Forensics versus disinformation in images and videos: lesson learnt and new challenges. In: Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation; 2023; Thessaloniki, Greece. p. 2. doi:10.1145/3592572.3596489.
86. Xia Z, Qiao T, Xu M, Zheng N, Xie S. Towards DeepFake video forensics based on facial textural disparities in multi-color channels. Inf Sci. 2022;607(2):654–69. doi:10.1016/j.ins.2022.06.003.
87. Verdoliva L, Bestagini P. Multimedia forensics. In: Proceedings of the 27th ACM International Conference on Multimedia; 2019; Nice, France. p. 2701–2.

88. Byeon H, Shabaz M, Shrivastava K, Joshi A, Keshta I, Oak R, et al. Deep learning model to detect deceptive generative adversarial network generated images using multimedia forensic. Comput Electr Eng. 2024;113:109024. doi:10.1016/j.compeleceng.2023.109024.
89. Lu Y, Liu Y, Fei J, Xia Z. Channel-wise spatiotemporal aggregation technology for face video forensics. Secur Commun Netw. 2021;2021(8):1–13. doi:10.1155/2021/8388480.
90. Kwon P, You J, Nam G, Park S, Chae G. KoDF: a large-scale korean deepfake detection dataset. arXiv:2103.10094. 2021.
91. Suratkar S, Kazi F. Deep fake video detection using transfer learning approach. Arab J Sci Eng. 2023;48(8):9727–37. doi:10.1007/s13369-022-07321-3.
92. Stroebel L, Llewellyn M, Hartley T, Ip TS, Ahmed M. A systematic literature review on the effectiveness of deepfake detection techniques. J Cyber Secur Tech. 2023;7(2):83–113. doi:10.1080/23742917.2023.2192888.
93. Institute RAI. A look at global deepfake regulation approach [Internet]. [cited 2025 Apr 6]. Available from: https://www.responsible.ai/a-look-at-global-deepfake-regulation-approaches/.
94. AI B. South Korea unveils unified AI act [Internet]. [cited 2025 Apr 6]. Available from: https://babl.ai/south-korea-unveils-unified-ai-act/.
95. Raveena, Chhikara R, Punyani P. Exploring deepfake detection: a comparative study of CNN models. In: 2024 International Conference on Intelligent Systems for Cybersecurity (ISCS); 2024; Gurugram, India. p. 1–6.
96. Xie K, Liu J, Zhu M, Ding G, Liu Z, Chen H, et al. Unveiling universal forensics of diffusion models with adversarial perturbations. In: 2024 International Joint Conference on Neural Networks (IJCNN); 2024; Yokohama, Japan. p. 1–8.
97. Xie D, Chatterjee P, Liu Z, Roy K, Kossi E. DeepFake detection on publicly available datasets using modified AlexNet. In: 2020 IEEE Symposium Series on Computational Intelligence (SSCI); 2020; Canberra, ACT, Australia: IEEE. p. 1866–71.
98. Zhang J, Cheng K, Sovernigo G, Lin X. A Heterogeneous feature ensemble learning-based deepfake detection method. In: Proceedings of ICC 2022—IEEE International Conference on Communications; 2022; Seoul, Republic of Korea. p. 2084–9.
99. Samrouth K, Beuve N, Deforges O, Bakir N, Hamidouche W. Ensemble learning model for face swap detection. In: 2024 12th International Symposium on Digital Forensics and Security (ISDFS); 2024; San Antonio, TX, USA. p. 1–5.
100. Keerthika S, Santhiya S, Jayadharshini P, Ruthranayaki J, Hariarasu R, Naveenan M. An innovative method for identifying deepfake videos through ensemble learning. In: 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT); 2024; Mandi, India. p. 1–10.
101. Tantawy O, Elshafee A. Advancements in deepfake detection: leveraging Bi-LSTM-CNN architecture for robust identification. In: 2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES); 2024; Giza, Egypt. p. 525–8.
102. Patel S, Chandra SK, Jain A. DeepFake videos detection and classification using resnext and LSTM neural network. In: 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON); 2023; Bangalore, India. p. 1–5.
103. Satpute R, Onwe CP. CNN-LSTM model for deepfake image detection. In: 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI); 2024; Wardha, India. p. 1–6.
104. Dagar D, Vishwakarma DK. A Hybrid Xception-LSTM model with channel and spatial attention mechanism for deepfake video detection. In: 2023 3rd International Conference on Mobile Networks and Wireless Communications (ICMNWC); 2023; Tumkur, India. p. 1–5.
105. Al-Dhabi Y, Zhang S. Deepfake video detection by combining convolutional neural network (CNN) and recurrent neural Network (RNN). In: 2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE); 2021; Beijing, China. p. 236–41.

106. Kaddar B, Fezza SA, Hamidouche W, Akhtar Z, Hadid A. HCiT: deepfake video detection using a hybrid model of CNN features and vision transformer. In: 2021 International Conference on Visual Communications and Image Processing (VCIP); 2021; Munich, Germany. p. 1–5.
107. Chitale M, Dhawale A, Dubey M, Ghane S. A hybrid CNN-LSTM approach for deepfake audio detection. In: 2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT); 2024; Vellore, India. p. 1–6.
108. Guefrechi S, Jabra MB, Hamam H. Deepfake video detection using InceptionResnetV2. In: 2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP); 2022; Sfax, Tunisia. p. 1–6.
109. V A, Joy PT. Deepfake detection using XceptionNet. In: 2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE); 2023; Kerala, India. p. 1–5.
110. Liu D, Lin Y. A deepfake face detection method using vision transformer with a convolutional module. In: 2024 17th International Conference on Advanced Computer Theory and Engineering (ICACTE); 2024; Hefei, China. p. 180–4.
111. Doshi A, Venkatadri A, Kulkarni S, Athavale V, Jagarlapudi A, Suratkar S, et al. Realtime deepfake detection using video vision transformer. In: 2022 IEEE Bombay Section Signature Conference (IBSSC); 2022; Mumbai, India. p. 1–6.
112. Zhang S, Wang Y, Tan C. Research on text classification for identifying fake news. In: 2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC); 2018; Jinan, China. p. 178–81.
113. Lubis AR, Prayudani S, Hamzah ML. Classification of text on social media data using the TF-IDF approach, Word2Vec and transfer learning. In: 2023 10th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI); 2023; Palembang, Indonesia. p. 282–7.
114. Betul Polat S, Cankurt S. Fake news classification using BLSTM with glove embedding. In: 2023 17th International Conference on Electronics Computer and Computation (ICECCO); 2023; Kaskelen, Kazakhstan. p. 1–5.
115. Saini K, Jain R. A hybrid LSTM-BERT and glove-based deep learning approach for the detection of fake news. In: 2023 3rd International Conference on Smart Data Intelligence (ICSMDI); 2023; Trichy, India. p. 400–6.
116. Kishwar A, Zafar A. Predicting fake news using GloVe and BERT embeddings. In: 2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM); 2021; Preveza, Greece. p. 1–6.
117. Bao Y, Dang R. Face detection under non-uniform low light based on improved MTCNN. In: 2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE); 2021; Hangzhou, China. p. 704–7.
118. Kütükçü YE, Polat H. BYOL yaklasimini kullanarak MTCNN ile Yakalanan Yüzlerde Deepfake Tespiti Deepfake detection on faces captured with MTCNN, by using the BYOL approach. In: 2024 Innovations in Intelligent Systems and Applications Conference (ASYU); 2024; Ankara, Türkiye. p. 1–6.
119. Bahmeie A, Shakiba M. Comparing MTCNN and viola-jones algorithm in face recognition. In: 2024 19th Iranian Conference on Intelligent Systems (ICIS); 2024; Sirjan, Iran. p. 68–72.
120. Yang Z, Ge W, Zhang Z. Face recognition based on MTCNN and integrated application of FaceNet and LBP method. In: 2020 2nd International Conference on Artificial Intelligence and Advanced Manufacture (AIAM); 2020; Manchester, UK. p. 95–8.
121. Syarif H, Zainuddin Z, Tahir Z. Evaluation of concentration levels in learning using zoom meeting by applying the histogram of oriented gradients (HOG) method and support vector machine (SVM) classification. In: 2022 8th International Conference on Education and Technology (ICET); 2022; Malang, Indonesia. p. 207–12.
122. Trabelsi A, Pic MM, Dugelay JL. Improving deepfake detection by mixing top solutions of the DFDC. In: 2022 30th European Signal Processing Conference (EUSIPCO); 2022; Belgrade, Serbia. p. 643–7.
123. Tolosana R, Romero-Tapiador S, Vera-Rodriguez R, Gonzalez-Sosa E, Fierrez J. DeepFakes detection across generations: analysis of facial regions, fusion, and performance evaluation. Eng Appl Artif Intell. 2022;110(4):104673. doi:10.1016/j.engappai.2022.104673.
124. Almutairi ZM, Elgibreen H. Detecting fake audio of arabic speakers using self-supervised deep learning. IEEE Access. 2023;11:72134–47. doi:10.1109/ACCESS.2023.3286864.

125. Anthony P, Ay B, Aydin G. A review of face anti-spoofing methods for face recognition systems. In: 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA); 2021; Kocaeli, Turkey. p. 1–9.
126. Almutairi Z, Elgibreen H. A review of modern audio deepfake detection methods: challenges and future directions. Algorithms. 2022;15(5):155. doi:10.3390/a15050155.
127. Kittur P, Pasha A, Joshi S, Kulkarni V. A review on anti-spoofing: face manipulation and liveness detection. In: 2023 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI); 2023; Gwalior, India: IEEE. p. 1–6.
128. Beuve N, Hamidouche W, Deforges O. DmyT: dummy triplet loss for deepfake detection. In: Proceedings of the 1st Workshop on Synthetic Multimedia—Audiovisual Deepfake Generation and Detection. 2021; Virtual. p. 17–24. doi:10.1145/3476099.3484316.
129. Beuve N, Hamidouche W, Déforges O. Hierarchical learning and dummy triplet loss for efficient deepfake detection. ACM Trans Multimed Comput Commun Appl. 2023;20(3):89. doi:10.1145/3626101.
130. Liang B, Wang Z, Huang B, Zou Q, Wang Q, Liang J. Depth map guided triplet network for deepfake face detection. Neural Netw. 2023;159(3):34–42. doi:10.1016/j.neunet.2022.11.031.
131. Chen L, Zhang Y, Song Y, Liu L, Wang J. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022; New Orleans, LA, USA. p. 18710–9.
132. Korshunov P, Marcel S. Improving generalization of deepfake detection with data farming and few-shot learning. IEEE Trans Biometr, Behav, Identity Sci. 2022;4(3):386–97. doi:10.1109/TBIOM.2022.3143404.
133. Huang D, Zhang Y. Learning meta model for strong generalization deepfake detection. In: 2024 International Joint Conference on Neural Networks (IJCNN); 2024; Yokohama, Japan: IEEE. p. 1–8.
134. Ruiz N, Bargal SA, Sclaroff S. Disrupting deepfakes: adversarial attacks against conditional image translation networks and facial manipulation systems, Workshops. In: Computer Vision–ECCV 2020 Workshops; 2020 Aug 23–28; Glasgow, UK. p. 236–51.
135. Ayyaz S, Malik SM. A comprehensive study of generative adversarial networks (GAN) and generative pre-trained transformers (GPT) in cybersecurity. In: 2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS); 2024; Marrakech, Morocco: IEEE. p. 1–8.
136. Huang Y, Juefei-Xu F, Guo Q, Liu Y, Pu G. FakeLocator: robust localization of GAN-based face manipulations. IEEE Trans Inf Forensics Secur. 2022;17(1):2657–72. doi:10.1109/TIFS.2022.3141262.
137. MeenaPrakash R, Kamali B, Vimala M, Madhuvandhana K, Krishnaleela P. A DenseNet-Enhanced GAN model for classification of medical images into original and fake. In: 2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI); 2025; Erode, India. p. 1540–5.
138. Safwat S, Mahmoud A, Eldesouky Fattoh I, Ali F. Hybrid deep learning model based on GAN and RESNET for detecting fake faces. IEEE Access. 2024;12:86391–402. doi:10.1109/ACCESS.2024.3416910.
139. Xu Y, Terhörst P, Pedersen M, Raja K. Analyzing fairness in deepfake detection with massively annotated databases. IEEE Trans Technol Soc. 2024;5(1):93–106. doi:10.1109/TTS.2024.3365421.
140. Saif S, Tehseen S, Ali SS. Fake news or real? Detecting deepfake videos using geometric facial structure and graph neural network. Technol Forecas Soc Change. 2024;205(5):123471. doi:10.1016/j.techfore.2024.123471.
141. Yin Q, Lu W, Cao X, Luo X, Zhou Y, Huang J. Fine-grained multimodal deepfake classification via heterogeneous graphs. Int J Comput Vis. 2024;132(11):5255–69. doi:10.1007/s11263-024-02128-1.
142. Sun Z, Chen S, Yao T, Yin B, Yi R, Ding S, et al. Contrastive pseudo learning for open-world deepfake attribution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023; Paris, France. p. 20882–92.
143. Souto Moreira L, Machado Lunardi G, de Oliveira Ribeiro M, Silva W, Paulo Basso F. A study of algorithm-based detection of fake news in brazilian election: is BERT the best. IEEE Lat Am Trans. 2023;21(8):897–903. doi:10.1109/TLA.2023.10246346.
144. Angizeh LB, Keyvanpour MR. Detecting fake news using advanced language models: BERT and RoBERTa. In: 2024 10th International Conference on Web Research (ICWR); 2024; Tehran, Iran. p. 46–52.

145. Koru GK, Uluyol C. Detection of Turkish fake news from tweets with BERT models. IEEE Access. 2024;12(5):14918–31. doi:10.1109/ACCESS.2024.3354165.
146. Martino AI, Lhaksmana KM. Classification of fake news on social media using BERT. In: 2024 International Conference on Data Science and Its Applications (ICoDSA); 2024; Kuta, Indonesia. p. 225–9.
147. Ramzan A, Ali RH, Ali N, Khan A. Enhancing fake news detection using BERT: a comparative analysis of logistic regression, RFC, LSTM and BERT. In: 2024 International Conference on IT and Industrial Technologies (ICIT); 2024; Bristol, UK. p. 1–6.
148. Omrani P, Ebrahimian Z, Toosi R, Akhaee MA. Bilingual COVID-19 fake news detection based on LDA topic modeling and BERT transformer. In: 2023 6th International Conference on Pattern Recognition and Image Analysis (IPRIA); 2023; Bristol, UK. p. 1–6.
149. Pavlov T, Mirceva G. COVID-19 fake news detection by using BERT and RoBERTa models. In: 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO); 2022; Opatija, Croatia. p. 312–6.
150. Shrivastava P, Sharma DK. COVID-19 fake news detection using pre-tuned BERT-based transfer learning models. In: 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART); 2022; Moradabad, India. p. 64–8.
151. Mahmud MAI, Talukder AAT, Sultana A, Bhuiyan KIA, Rahman MS, Pranto TH, et al. Toward news authenticity: synthesizing natural language processing and human expert opinion to evaluate news. IEEE Access. 2023;11:11405–21.
152. Laczi SA, Póser V. Impact of deepfake technology on children: risks and consequences. In: 2024 IEEE 22nd Jubilee International Symposium on Intelligent Systems and Informatics (SISY); 2024; Pula, Croatia. p. 215–20.
153. Wu L, Long Y, Gao C, Wang Z, Zhang Y. MFIR: multimodal fusion and inconsistency reasoning for explainable fake news detection. Inf Fusion. 2023;100(1):101944. doi:10.1016/j.inffus.2023.101944.
154. Venkateswarulu S, Srinagesh A. DeepExplain: enhancing deepfake detection through transparent and explainable AI model. Informatica. 2024;48(8). doi:10.31449/inf.v48i8.5792.
155. Gong Y, Shang L, Wang D. Integrating social explanations into explainable artificial intelligence (XAI) for combating misinformation: vision and challenges. IEEE Trans Comput Soc Syst. 2024;11(5):6705–26.
156. Mansoor N, Iliev AI. Explainable AI for deepfake detection. Appl Sci. 2025;15(2):725. doi:10.3390/app15020725.
157. Pu J, Sarwar Z, Abdullah SM, Rehman A, Kim Y, Bhattacharya P, et al. Deepfake text detection: limitations and opportunities. In: 2023 IEEE Symposium on Security and Privacy (SP); 2023; San Francisco, CA, USA: IEEE. p. 1613–30.
158. Folino F, Folino G, Guarascio M, Pontieri L, Zicari P. Towards data-and compute-efficient fake-news detection: an approach combining active learning and pre-trained language models. SN Comput Sci. 2024;5(5):470.
159. Abdali S, Shaham S, Krishnamachari B. Multi-modal misinformation detection: approaches, challenges and opportunities. ACM Comput Surv. 2024;57(3):1–29.
160. Jing J, Wu H, Sun J, Fang X, Zhang H. Multimodal fake news detection via progressive fusion networks. Inf Process Manag. 2023;60(1):103120. doi:10.1016/j.ipm.2022.103120.
161. Comito C, Caroprese L, Zumpano E. Multimodal fake news detection on social media: a survey of deep learning techniques. Soc Netw Anal Min. 2023;13(1):101. doi:10.1007/s13278-023-01104-w.
162. Tufchi S, Yadav A, Ahmed T. A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities. Int J Multimed Inf Retr. 2023;12(2):28. doi:10.1007/s13735-023-00296-3.
163. Abe T, Yoshida S, Muneyasu M. Dynamic Graph convolutional network with time series-aware structural feature extraction for fake news detection. ITE Trans Media Technol Appl. 2025;13(1):106–18.
164. Meel P, Raj C, Bhawna. A review of web infodemic analysis and detection trends across multi-modalities using deep neural network. Int J Data Sci Anal. 2025. doi:10.1007/s41060-025-00727-w.
165. Pontorno O, Guarnera L, Battiato S. DeepFeatureX Net: deep features extractors based network for discriminating synthetic from real images. In: Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and lecture notes in bioinformatics). Cham, Switzerland: Springer; 2025. Vol. 15321, p. 177–93.

166. Kareem MAR, Abdulrahman AA. Advanced text vectorization and deep learning models for enhanced fake news detection on social medi. Commun Comput Inf Sci. 2025;2329:151– 71. doi:10.1007/978-3-031-81065-7.
167. Nikumbh D, Thakare A. Ensemble-inspired multi-modal fusion of features for fake news detection on social media. Lecture Notes Netw Syst. 2025;1144:97–109. doi:10.1007/978-981-97-7839-3.
168. Toor MS, Shahbaz H, Yasin M, Ali A, Fitriyani NL, Kim C, et al. An optimized weighted-voting-based ensemble learning approach for fake news classification. Mathematics. 2025;13(3):449.
169. Kumar A, Singh D, Jain R, Jain DK, Gan C, Zhao X. Advances in DeepFake detection algorithms: exploring fusion techniques in single and multi-modal approach. Inf Fusion. 2025;118(1):102993. doi:10.1016/j.inffus.2025.102993.
170. Mahdi AS, Shati NM. Utilizing graph neural networks for the detection of fake news through analysis of relationships among various social media entities. Commun Comput Inf Sci. 2025;2329:172–85. doi:10.1007/978-3-031-81065-7.
171. Roumeliotis KI, Tselikas ND, Nasiopoulos DK. Fake news detection and classification: a comparative study of convolutional neural networks, large language models, and natural language processing models. Fut Internet. 2025;17(1):28. doi:10.3390/fi17010028.
172. LekshmiAmmal HR, Madasamy AK. A reasoning based explainable multimodal fake news detection for low resource language using large language models and transformers. J Big Data. 2025;12(1):46.
173. Usmani S, Kumar S, Sadhya D. Spatio-temporal knowledge distilled video vision transformer (STKD-VViT) for multimodal deepfake detection. Neurocomputing. 2025;620(4):129256. doi:10.1016/j.neucom.2024.129256.
174. Wang L, Wang Z, Wu L, Liu AA. Bots shield fake news: adversarial attack on user engagement based fake news detection. In: CIKM '24: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management; 2024. p. 2369–78.
175. Si J, Wang Y, Hu W, Liu Q, Hong R. Making strides security in multimodal fake news detection models: a comprehensive analysis of adversarial attacks. In: lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). Cham, Switzerland: Springer; 2025. Vol. 15521, p. 296–309.
176. Ain QU, Javed A, Irtaza A. DeepEvader: an evasion tool for exposing the vulnerability of deepfake detectors using transferable facial distraction blackbox attack. Eng Appl Artif Intell. 2025;145(1):110276. doi:10.1016/j.engappai.2025.110276.
177. Liang J, Zhang X, Shang Y, Guo S, Li C. Clean-label poisoning attack against fake news detection models. In: 2023 IEEE International Conference on Big Data (BigData); 2023; Sorrento, Italy: IEEE. p. 3614–23.
178. He Q, Peng C, Liu D, Wang N, Gao X. GazeForensics: deepFake detection via gaze-guided spatial inconsistency learning. Neural Netw. 2024;180(2):106636. doi:10.1016/j.neunet.2024.106636.
179. Dudykevych V, Yevseiev S, Mykytyn H, Ruda K, Hulak H. Detecting Deepfake modifications of biometric images using neural networks. In: Workshop on Cybersecurity Providing in Information and Telecommunication Systems II (CPITS 2024); 2024; Kyiv, Ukraine. p. 391–7.
180. Cheniti M, Akhtar Z, Chandaliya PK. Dual-model synergy for fingerprint spoof detection using VGG16 and ResNet50. J Imaging. 2025;11(2):42. doi:10.3390/jimaging11020042.
181. Sundar AP, Li F, Zou X, Hu Q, Gao T. Multi-armed-bandit-based shilling attack on collaborative filtering recommender systems. In: 2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS); 2020; Delhi, India: IEEE. p. 347–55.
182. Patel K, Thakkar A, Shah C, Makvana K. A state of art survey on shilling attack in collaborative filtering based recommendation system. In: Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems; 2016; Ahmedabad, India. p. 377–85. doi:10.1007/978-3-319-30933-0.
183. Dutta H, Pandey A, Bilgaiyan S. EnsembleDet: ensembling against adversarial attack on deepfake detection. J Electron Imaging. 2021;30(6):063030.
184. Arshed MA, Mumtaz S, Ibrahim M, Dewi C, Tanveer M, Ahmed S. Multiclass ai-generated deepfake face detection using patch-wise deep learning model. Computers. 2024;13(1):31. doi:10.3390/computers13010031.

185. Yan Z, Zhang Y, Yuan X, Lyu S, Wu B. DeepfakeBench: a comprehensive benchmark of deepfake detection. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23; 2023; New Orleans, LA, USA. p. 4534–65.
186. Sharma P, Kumar M, Sharma HK. GAN-CNN ensemble: a robust deepfake detection model of social media images using minimized catastrophic forgetting and generative replay technique. Procedia Comput Sci. 2024;235(1):948–60. doi:10.1016/j.procs.2024.04.090.
187. Song D, Lee N, Kim J, Choi E. Anomaly detection of deepfake audio based on real audio using generative adversarial network model. IEEE Access. 2024;12(5):184311–26. doi:10.1109/ACCESS.2024.3506973.
188. Veksler M, Akkaya K. Good or Evil: generative adversarial networks in digital forensics. In: Adversarial Multimedia Forensics; 2023; Cham, Switzerland: Springer. p. 55–91.
189. Amerini I, Barni M, Battiato S, Bestagini P, Boato G, Bruni V, et al. Deepfake Media Forensics: status and future challenges. J Imaging. 2025;11(3):73. doi:10.3390/jimaging11030073.
190. Agarwal H, Singh A, Rajeswari D. Deepfake detection using svm. In: 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC); 2021; Coimbatore, India: IEEE. p. 1245–9.
191. Pokroy AA, Egorov AD. Efficientnets for deepfake detection: comparison of pretrained models. In: 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus); 2021; St. Petersburg, Moscow, Russia: IEEE. p. 598–600.
192. Ain QU, Javed A, Malik KM, Irtaza A. Regularized forensic efficient net: a game theory based generalized approach for video deepfakes detection. Multimed Tools Appl. 2024;53(4):1–44. doi:10.1007/s11042-024-20268-x.
193. Suratkar S, Kazi F, Sakhalkar M, Abhyankar N, Kshirsagar M. Exposing deepfakes using convolutional neural networks and transfer learning approaches. In: 2020 IEEE 17th India Council International Conference (INDICON); 2020; New Delhi, India: IEEE. p. 1–8.
194. TS SM, Sreeja PS. Adam Adadelta Optimization based bidirectional encoder representations from transformers model for fake news detection on social media. Multiagent Grid Syst. 2023;19(3):271–87. doi:10.3233/MGS-230033.
195. Wang K, Xiong Q, Wu C, Gao M, Yu Y. Multi-modal cyberbullying detection on social networks. In: 2020 International Joint Conference on Neural Networks (IJCNN); 2020; Glasgow, UK: IEEE. p. 1–8.
196. Bonomi M, Pasquini C, Boato G. Dynamic texture analysis for detecting fake faces in video sequences. J Vis Commun Image Represent. 2021;79:103239. doi:10.1016/j.jvcir.2021.103239.
197. Amerini I, Galteri L, Caldelli R, Del Bimbo A. Deepfake video detection through optical flow based cnn. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops; 2019; Seoul, Republic of Korea. p. 1205–7.
198. Caldelli R, Galteri L, Amerini I, Del Bimbo A. Optical Flow based CNN for detection of unlearnt deepfake manipulations. Pattern Recognit Lett. 2021;146(10):31–7. doi:10.1016/j.patrec.2021.03.005.
199. Verdoliva L. Media forensics and deepfakes: an overview. IEEE J Sel Top Signal Process. 2020;14(5):910–32. doi:10.1109/JSTSP.2020.3002101.
200. Khalid M, Raza A, Younas F, Rustam F, Villar MG, Ashraf I, et al. Novel sentiment majority voting classifier and transfer learning-based feature engineering for sentiment analysis of deepfake tweets. IEEE Access. 2024;12:67117–29.
201. Gupta D, Bhargava A, Agarwal D, Alsharif MH, Uthansakul P, Uthansakul M, et al. Deep learning-based truthful and deceptive hotel reviews. Sustainability. 2024;16(11):4514. doi:10.3390/su16114514.
202. Chong ATY, Chua HN, Jasser MB, Richard Wong TK. Bot or human? detection of deepfake text with semantic, emoji, sentiment and linguistic features. In: 2023 IEEE 13th International Conference on System Engineering and Technology (ICSET); 2023; Shah Alam, Malaysia. p. 205–10.
203. Rao S, Verma AK, Bhatia T. A review on social spam detection: challenges, open issues, and future directions. Expert Syst Appl. 2021;186(1):115742. doi:10.1016/j.eswa.2021.115742.
204. Ma YW, Chen JL, Chen LD, Huang YM. Intelligent clickbait news detection system based on artificial intelligence and feature engineering. IEEE Trans Eng Manag. 2024;71:12509–18.
205. Alghaligah A, Alotaibi A, Abbas Q, Alhumoud S. Optimized hybrid deep learning for enhanced spam review detection in E-commerce platforms. Int J Adv Comput Sci Appl. 2025;16(1). doi:10.14569/issn.2156-5570.

206. Chen H. Research on spam classification based on traditional machine learning and deep learning. AIP Conf Proc. 2024;3194(1):040016. doi:10.1063/5.0225580.
207. Kalyani VV, Rama Sundari MV, Neelima S, Satya Prasad PS, PattabhiRama Mohan P, Lakshmanarao A. SMS spam detection using NLP and deep learning recurrent neural network variants. In: 2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC—ROBINS); 2024; Coimbatore, India. p. 92–6.
208. Zhang M. Ensemble-based text classification for spam detection. Informatica. 2024;48(6):71–80. doi:10.31449/inf.v48i6.5246.
209. Bahgat EM, Rady S, Gad W, Moawad IF. Efficient email classification approach based on semantic methods. Ain Shams Eng J. 2018;9(4):3259–69. doi:10.1016/j.asej.2018.06.001.
210. Jha K, Jain A, Srivastava S. Feature-level fusion of face and speech based multimodal biometric attendance system with liveness detection. AIP Adv. 2024;14(11):115007. doi:10.1063/5.0234430.
211. Rasool A, Katarya R. Seeing through the lies: a vision transformer-based solution. Lecture Notes Electr Eng. 2024;1191:373–87. doi:10.1007/978-981-97-2508-3.
212. Das A, Pal S, Das B, Kaur P. FakeTweet busters: a combination of BERT and Deepfake detection to resolve the spreading of fake AI generated news. In: 2024 IEEE Region 10 Symposium (TENSYMP); 2024; New Delhi, India. p. 1–6.
213. Jenkins J, Roy K. Exploring deep convolutional generative adversarial networks (DCGAN) in biometric systems: a survey study. Discov Artif Intell 2024;4(1):42. doi:10.1007/s44163-024-00138-z.
214. Lapates JM, Gerardo BD, Medina RP. Spectrogram-based analysis and detection of deepfake audio using enhanced DCGANs for secure content distribution. In: 2024 15th International Conference on Information and Communication Technology Convergence (ICTC); 2024; Jeju Island, Republic of Korea: IEEE. p. 1851–6.
215. Martin N, Nambayil SR, Devikrishna U, Fasila K, et al. Multimodal deepfake detection using deep-convolutional neural networks and mel-frequency cepstral coefficients. In: 2024 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES); 2024; Kottayam, India: IEEE. p. 1–6.
216. Qoiriah A, Putri DO, Yamasari Y, Suartana I, Putra RE, Nurhidayat AI. Utilizing hybrid CNN-SVM and FastText word embedding for twitter cyberbullying classification. In: 2024 Seventh International Conference on Vocational Education and Electrical Engineering (ICVEE); 2024; Malang, Indonesia: IEEE. p. 317–22.
217. Reiki MKA, Sibaroni Y. Improving feature extraction for sentiment analysis on Indonesian election 2024 using term weighting with fastText. In: 2024 International Conference on Data Science and Its Applications (ICoD SA); 2024; Kuta, Indonesia: IEEE. p. 481–6.
218. Galbally J, Marcel S, Fierrez J. Biometric antispoofing methods: a survey in face recognition. IEEE Access. 2014;2:1530–52. doi:10.1109/ACCESS.2014.2381273.
219. Akhtar Z, Kale S, Alfarid N. Spoof attacks on multimodal biometric systems. In: International Conference on Information and Network Technology; 2011; Barcelona, Spain. p. 46–51.
220. Verma P, Selwal A, Sharma D. An exploration of pre-processing approaches for iris spoof detectors. In: 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES); 2022; Greater Noida, India: IEEE. p. 271–7.
221. Devi PJ, Sonali A, Karthik BNS, Sindhuja A, Reddy AAK. Deep learning for iris recognition: an integration of feature extraction and clustering. In: 2023 4th IEEE Global Conference for Advancement in Technology (GCAT); 2023; Bangalore, India: IEEE. p. 1–12.
222. Fazil M, Sah AK, Abulaish M. A hierarchical attention-based neural network model for socialbot detection in OSN. In: 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT); 2020; Melbourne, VIC, Australia: IEEE. p. 954–9.
223. Lian Z, Zhang C, Su C, Dharejo FA, Almutiq M, Memon MH. Find: privcy-enhanced federated learning for intelligent fake news detection. IEEE Trans Comput Soc Syst. 2023;11(4):5005–14. doi:10.1109/TCSS.2023.3304649.
224. Gautam V, Kaur G, Malik M, Pawar A, Singh A, Singh KK, et al. FFDL: feature fusion-based deep learning method utilizing federated learning for forged face detection. IEEE Access. 2025;13:5366–79.
225. Ikenga FA, Nwador AF. The intersection of artificial intelligence, deepfake, and the politics of international diplomacy. Ianna J Interdiscip Stud. 2024;6(2):53–71.

226. Pavis M. Rebalancing our regulatory response to Deepfakes with performers' rights. Convergence. 2021;27(4):974–98. doi:10.1177/13548565211033418.
227. Silva SH, Bethany M, Votto AM, Scarff IH, Beebe N, Najafirad P. Deepfake forensics analysis: an explainable hierarchical ensemble of weakly supervised models. Forensic Sci Int: Synergy. 2022;4(1):100217. doi:10.1016/j.fsisyn.2022.100217.
228. Mubarak R, Alsboui T, Alshaikh O, Inuwa-Dutse I, Khan S, Parkinson S. A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. IEEE Access. 2023;11(3):144497–529. doi:10.1109/ACCESS.2023.3344653.
229. Naffi N, Charest M, Danis S, Pique L, Davidson AL, Brault N, et al. Empowering youth to combat malicious deepfakes and disinformation: an experiential and reflective learning experience informed by personal construct theory. J Contr Psychol. 2025;38(1):119–40. doi:10.1080/10720537.2023.2294314.
230. Matli W. Extending the theory of information poverty to deepfake technology. Int J Inf Manag Data Insights. 2024;4(2):100286. doi:10.1016/j.jjimei.2024.100286.
231. Birrer A, Just N. What we know and don't know about deepfakes: an investigation into the state of the research and regulatory landscape. New Media Soc. 2024. doi:10.1177/14614448241253138.
232. Powers G, Johnson JP, Killian G. To tell or not to tell: the effects of disclosing deepfake video on US and Indian consumers' purchase intention. J Interact Advert. 2023;23(4):339–55. doi:10.1080/15252019.2023.2260399.
233. Romero Moreno F. Generative AI and deepfakes: a human rights approach to tackling harmful content. Int Rev Law, Comput Technol. 2024;38(3):297–326. doi:10.1080/13600869.2024.2324540.
234. Overton S. Overcoming racial harms to democracy from artificial intelligence. Iowa Law Rev. 2025;110(2):805–66.
235. Qureshi SM, Saeed A, Almotiri SH, Ahmad F, Ghamdi MAA. Deepfake forensics: a survey of digital forensic methods for multimodal deepfake identification on social media. PeerJ Comput Sci. 2024;10(2):1–40. doi:10.7717/peerj-cs.2037.
236. Döring N, Le TD, Vowels LM, Vowels MJ, Marcantonio TL. The impact of artificial intelligence on human sexuality: a Five-year literature review 2020–2024. Curr Sex Health Rep. 2025;17(1):1–39.
237. McLoughlin I. The weaponisation of the internet—effect models. In: 2023 6th International Conference on Applied Computational Intelligence in Information Systems (ACIIS); 2023; Bandar Seri Begawan, Brunei Darussalam. p. 1–6.
238. Kaushal A, Mina A, Meena A, Babu TH. The societal impact of Deepfakes: advances in detection and mitigation. In: 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT); 2023; Delhi, India: IEEE. p. 1–7.
239. Golda A, Mekonen K, Pandey A, Singh A, Hassija V, Chamola V, et al. Privacy and security concerns in generative AI: a comprehensive survey. IEEE Access. 2024;12:48126–44. doi:10.1109/ACCESS.2024.3381611.
240. Atlam ES, Almaliki M, Elmarhomy G, Almars AM, Elsiddieg AMA, ElAgamy R. SLM-DFS: a systematic literature map of deepfake spread on social media. Alex Eng J. 2025;111(2):446–55. doi:10.1016/j.aej.2024.10.076.
241. Caramancion KM. The demographic profile most at risk of being disinformed. In: 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS); 2021; Toronto, ON, Canada. p. 1–7.
242. Vandana, Chaturvedi K. Illusion or reality: analyzing sentiments on deepfakes. In: 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC); 2024; Coimbatore, India. p. 1207–10.
243. Pramod D, Patil KP, Kumar D, Singh DR, Singh Dodiya C, Noble D. GenerativeAI and deep fakes inmedia industry —an innovation resistance theory perspective. In: 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT); 2024; Greater Noida, India. p. 1–5.
244. Zandi G, Yaacob NA, Tajuddin M, Rahman NKNA. Artificial intelligence and the evolving cybercrime paradigm: current threats to businesses. J Inf Technol Manag. 2024;16(4):162–70.
245. Lewis A, Vu P, Duch RM, Chowdhury A. Deepfake detection with and without content warnings. R Soc Open Sci. 2023;10(11):1753. doi:10.1098/rsos.231214.