technologies

MDPI

Perspective

Analysis, Evaluation, and Future Directions on Multimodal Deception Detection

Arianna D'Ulizia, Alessia D'Andrea *, Patrizia Grifoni and Fernando Ferri

Italian National Research Council-Institute for Research on Population and Social Policies (CNR-IRPPS), 00185 Rome, Italy; arianna.dulizia@irpps.cnr.it (A.D.); patrizia.grifoni@irpps.cnr.it (P.G.); fernando.ferri@irpps.cnr.it (F.F.)

* Correspondence: alessia.dandrea@irpps.cnr.it

Abstract: Multimodal deception detection has received increasing attention from the scientific community in recent years, mainly due to growing ethical and security issues, as well as the growing use of digital media. A great number of deception detection methods have been proposed in several domains, such as political elections, security contexts, and job interviews. However, a systematic analysis of the current situation and the evaluation and future directions of deception detection based on cues coming from multiple modalities seems to be lacking. This paper, starting from a description of methods and metrics used for the analysis and evaluation of multimodal deception detection on video, provides a vision of future directions in this field. For the analysis, the PRISMA recommendations are followed, which allow the collection and synthesis of all the available research on the topic and the extraction of information on the multimodal features, the fusion methods, the classification approaches, the evaluation datasets, and metrics. The results of this analysis contribute to the assessment of the state of the art and the evaluation of evidence on important research questions in multimodal deceptive deception. Moreover, they provide guidance on future research in the field.

Keywords: deception detection; multimodal; fusion methods; systematic literature review

Citation: D'Ulizia, A.; D'Andrea, A.; Grifoni, P.; Ferri, F.. Analysis, Evaluation, and Future Directions on Multimodal Deception Detection. *Technologies* **2024**, *12*, 71. https://doi.org/10.3390/technologies12050071

Academic Editor: Sikha Bagui

Received: 25 March 2024
Revised: 11 May 2024
Accepted: 17 May 2024
Published: 18 May 2024

1. Introduction

Deception detection refers to the methods of investigation used to reveal the veracity and reliability of an individual by taking into account a variety of behavioral markers in addition to more extensive contextual and situational data.

Various academic communities, such as computer vision, psychology, and language processing, have been paying more attention to deception detection lately because dishonesty affects nearly every human contact and can have expensive repercussions [1]. Furthermore, there is a global interest in identifying liars due to security reasons. For instance, seeing liars is essential in airports. Terrorists may lie to interviewers at borders and customs and withhold vital information that could endanger their lives.

To develop more advanced lie detection systems, researchers focused on the analysis of multimodal features that combine elements from multiple modalities (e.g., speech, gesture, facial expressions, and text). The extraction and analysis of multimodal features seek to avoid the human labor involved in the analysis and decision-making procedures utilized in previous techniques, as well as the uncertainty associated with the use of single modalities. By combining features from many modalities, the dataset is enhanced with data that would not be available if these modalities were employed independently. This can be seen in the overall performance and confidence level obtained by multimodal classifiers compared to the monomodal ones. Moreover, combining features from several modalities allows for the more accurate identification of a deceiver since the analysis of multimodal cues provides more information than merely observing only verbal or non-verbal behavior.

This paper provides an overview of the methods and metrics used for the analysis and evaluation of multimodal deception detection on videos and a vision of future directions

in this field. The goal is to introduce readers to the main concepts, current computational approaches for modality fusion and classification, and evaluation datasets and metrics applied in this research area. In particular, studies on deception detection systems focusing on multimodal cues from videos published over the last decade have been surveyed using a systematic literature review (SLR) protocol.

The main contributions of this paper are as follows:

1. To gather and systematically organize the scientific literature on automated multimodal deception detection from videos;
2. To provide evidence on important research questions related to (i) features extracted from different modalities, (ii) methods for the fusion of multimodal features, (iii) methods for the classification of fake and true results, and (iv) used datasets;
3. To discuss the metrics mainly used for the evaluation of multimodal deception detection;
4. To provide a vision of future research directions of multimodal deception detection.

The remainder of this paper is organized as follows. Section 2 introduces the four steps typically followed in the multimodal deception detection process. In Section 3, the methodology followed for the systematic literature review is described, while in Section 4, the results are discussed. Section 5, Section 6, and Section 7 discuss how the surveyed studies answer the questions defined in the paper. Finally, Section 8 concludes the paper.

2. Multimodal Deception Detection from Videos

In numerous real-world situations, such as airport security screening, job interviews, court trials, and personal credit risk assessment, multimodal deception detection from videos is a difficult challenge [2].

Usually, the multimodal deception detection process consists of the following four steps: feature extraction from different modalities, the fusion of multimodal features, the classification of fake and true results, and the evaluation of the classification methods.

Considering the feature extraction, cues for deception detection can come from a wide range of modalities, including speech, text, facial expressions, and gestures. Different studies showed that the analysis of multimodal features improves the performance of deception detection, as opposed to using single modalities separately, such as text analysis, voice analysis, thermal images, or visual signals [3,4]. In the analyzed papers, the multimodal features are broadly classified into the following: (i) audio features, extracted from the speech and audio modality; (ii) visual features, which can be further categorized into facial and body features and can be extracted from videos and images; (iii) textual features, extracted from the speech transcription; (iv) temporal features, which refers to the aspects related to time or sequence in multimodal data across modalities, and (v) EEG, which records electrical activity in the brain capturing the cognitive activities and emotions.

After the feature extraction, a significant task in performing a multimodal analysis consists of effectively fusing multimodal features while maintaining the integrity of modal information to minimize information loss [5,6]. For this purpose, two general types of fusion, i.e., model-independent fusion and model-dependent fusion [7], are used. Model-independent fusion avoids using specific learning models directly, as they are straightforward but less effective due to information loss during fusion. This kind of fusion can be further classified as early fusion, late fusion, and hybrid fusion [8], depending on when the fusion happens. Early fusion is the process of fusing data and features as soon as possible after the extraction of the features from the different modalities. Usually, this is conducted by performing a straightforward join operation on the features. Late fusion, often referred to as decision-level fusion, consists of combining the outputs of several models after each modality's independent model has been trained. Hybrid fusion techniques combine the advantages of late and early fusion; nevertheless, they also lead to a more intricate model structure and more challenging training. Model-dependent fusion addresses the issue of integrating diverse modalities by implementing technical and model viewpoints. It has more applications than model-independent fusion. Several frequently employed models include multi-kernel learning (MKL), graphical models (GMs), and neural networks (NNs).

For the classification of fake and true results, machine learning algorithms were used. Supervised learning and unsupervised learning are the two main classes of machine learning algorithms. In supervised learning, algorithms are trained using a training set in which proper inputs and outputs are included to perform the learning. On the contrary, in unsupervised learning, these types of data are different in terms of labeling. The datasets do not have any predefined link; thus, the outcome cannot be predicted. Compared to supervised learning, unsupervised learning requires much less human intervention.

Finally, datasets and metrics are used for training and evaluating machine learning algorithms. Datasets offer tagged images or videos with annotations—which are essential as they supply the ground truth labels required for the models to learn accurately. Evaluation metrics are numerical measurements that are used to evaluate a statistical or machine learning model's efficacy and performance. These metrics aid in the comparison of various models or algorithms and offer insights into how well the model operates.

In the following sections, we provide a detailed description of the methodology used for collecting research from the current literature and analyzing it according to the solutions proposed for the four steps of the multimodal deception detection process.

3. Materials and Methods

The process for screening and evaluating the body of the current literature is shown in this section. The following methodology is the systematic literature review (SLR) defined by the PRISMA recommendations [9]. This methodology suggests following six steps, as listed below: (1) identifying the review focus; (2) specifying the review question(s); (3) identifying studies to include in the review; (4) data extraction and study quality appraisal; (5) synthesizing the findings; and (6) reporting the results.

Concerning the first step, the identified review focuses on the systematic gathering and organization of scientific literature about automated deception detection using multimodal cues in videos. Specifically, the focus is on the multimodal extracted features, the methodologies used for fusing the multimodal features, the classification algorithms, and the datasets and metrics for the evaluation.

The questions (RQ) listed below have been specified, as indicated by the second step of the SLR process:

- RQ1: What are the multimodal features extracted for automated deception detection?
- RQ2: Which methodologies are used for the fusion of multimodal features?
- RQ3: What are the classification algorithms used for multimodal deception detection?
- RQ4: Which datasets are used for the analysis of multimodal deception detection?
- RQ5: Which metrics for the evaluation of multimodal deception detection are used? Which are the best-performing multimodal deception detection methods?
- RQ6: What are the future directions on multimodal deception detection?

The three phases of the PRISMA statement [9] (i.e., identification, screening, and inclusion), as shown in Figure 1, were followed in the third step of the SLR process to identify the studies to be included in the review.

Indexed scientific databases from the officially published literature (such as journal articles, books, and conference papers) were utilized to find an initial set of scientific works. Due to their high value in conducting bibliometric studies in the body of literature in many research fields, Web of Science (WoS) and Scopus databases were specifically used in this work as the most comprehensive sources of published scientific research [10]. The systematic review did not include grey literature since there is no gold standard method for searching grey literature, which makes it more challenging [11].

The following search strings were defined to search the scientific papers on the databases: ("deception detection" OR "lie detection" OR "deceptive behaviour*" OR "lie behaviour*" OR "detect* deception" AND "video*" AND "multimodal").

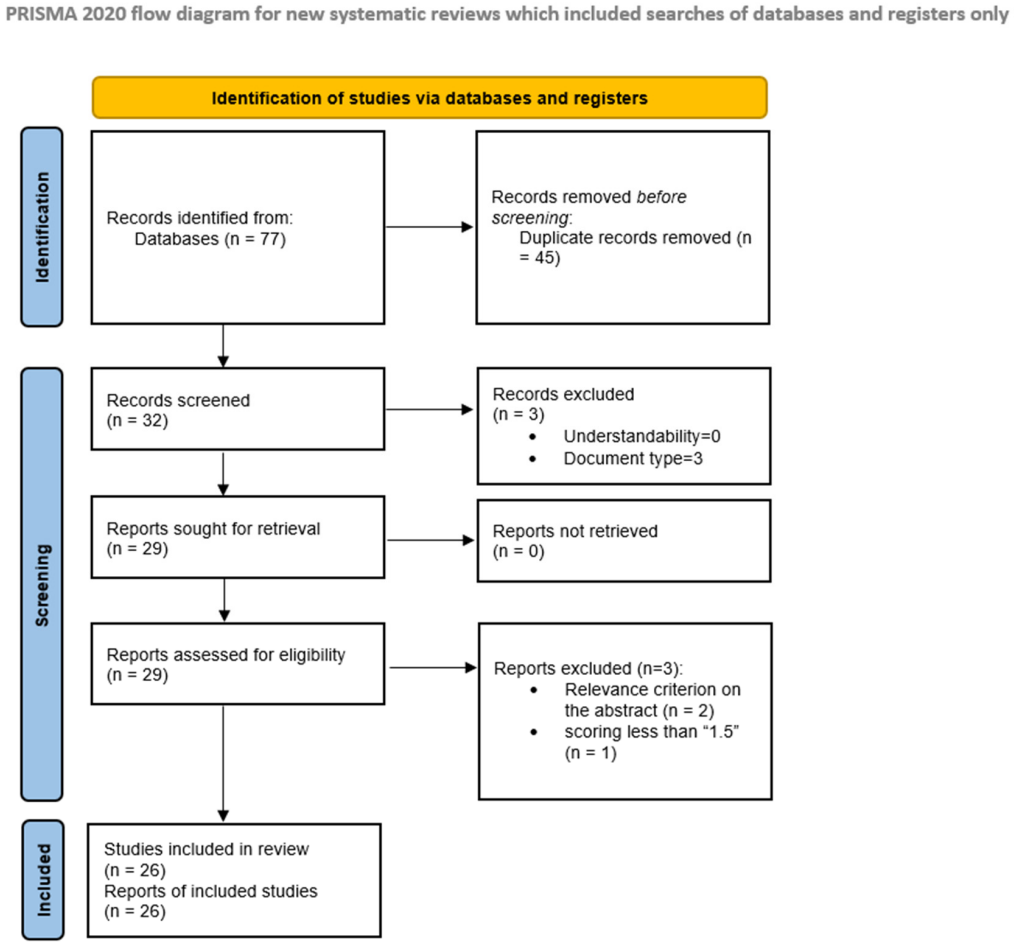

Figure 1. The three-phase flow diagram of PRISMA. Inspired by [9].

The search results were filtered throughout the screening step using the inclusion and exclusion criteria listed in Table 1.

Table 1. Exclusion and inclusion criteria used in the screening and eligibility phases.

Screening Phase	Eligibility Phase
Exclusion Criteria	
e1. Understandability criterion: - Articles that were not published in English;	
e2. Duplication criterion: - Articles with the same title and authors that were retrieved from two different databases; - Articles with the same title and authors that were retrieved from the same database.	**e3. Availability criterion:** - Studies whose full texts are not available.
Inclusion criteria	
i1. Temporal criterion: - Studies published in the period 2013–2023.	**i2. Document type criterion:** - Studies that belong to the following document types: article, review, book chapter, conference/proceedings paper. **i3. Relevance criterion on the abstract:** - Studies that are relevant to the review focus, i.e., they describe automated deception detection from videos focusing on multimodal cues; - Studies that are relevant to answer our research questions: (i) the multimodal extracted features, (ii) the methodologies used for the fusion of multimodal features, (iii) the classification algorithms, or (iv) the evaluation datasets and metrics.

The quality assessment checklist, shown in Table 2, comprises five questions and their corresponding scores, which were applied by two reviewers to analyze the full texts of the screened articles. If there was disagreement, a moderator gave the "disagreed" articles a final score after evaluating them.

Table 2. Quality assessment questions and scores formulated for the study.

Questions of the Quality Evaluation Checklist	Scores
Does the article describe the multimodal extracted features?	1—yes, the extracted multimodal features are fully described. 0.5—partially, the extracted multimodal features are just summarized without further descriptions. 0—no, the extracted multimodal features are not described.
Does the article describe the fusion methods?	1—yes, the fusion methods are fully described. 0.5—partially, the fusion methods are just summarized without further descriptions. 0—no, the fusion methods are not described.
Does the article describe the classification algorithm?	1—yes, the classification algorithm is fully described. 0.5—partially, the classification algorithm is just summarized without further descriptions. 0—no, the classification algorithm is not described.
Does the article describe the dataset(s) and metrics for evaluating the method?	1—yes, both datasets and metrics are described. 0.5—partially, only the dataset(s) or the metrics are described. 0—no, datasets and metrics are not described.
Does the article describe the future works?	1—yes, future works are described. 0—no, future works are not described.

Studies with a score of "1.5" or less were not included in the systematic review, while studies with a score of "1.5" or above were included.

Finally, the texts of the publications that were included were examined, and the following data—if any—were gathered:

- Multimodal features (e.g., visual, audio, textual, temporal, EEG);
- Multimodal fusion techniques;
- Deception detection classification approaches;
- Evaluation datasets, metrics, and scores;
- Future works.

The last two phases of the SLR process, i.e., synthesizing the findings and reporting the results, are detailed in the following sections.

4. Results of the Application of the SLR Methodology

During the identification phase, described in Section 3 and depicted in Figure 1, a total of 77 articles were returned using the two search engines (retrieved in September 2023): 47 from Scopus and 30 from WoS, respectively.

A total of 31 articles retrieved from Scopus were duplicated, while 14 duplicated articles were found from the 30 articles retrieved from WoS. Therefore, after removing duplicate records, as required by the duplication criterion, 32 studies were left that were screened according to the inclusion and exclusion criteria defined in Table 1. The understandability criterion was satisfied by all the articles (all the articles were written in English).

Removing the studies that did not satisfy the document type criterion (3 articles), a total of 29 articles resulted at the end of the screening phase.

Since all the articles were accessible in their full text (the availability criterion was satisfied by all the articles), only the articles that were not relevant (two studies for the relevance criterion) were excluded, and therefore, 27 articles were retained for a full evaluation of eligibility.

Two reviewers assessed these 27 studies according to the quality evaluation checklist shown in Table 2. One study that scored less than "1.5" was excluded, while the remaining twenty-six studies were included in the review, and the information listed in Section 3 was extracted from their full texts. Table 3 shows an overview of the selected studies, containing a short description, source title, and publication type for each study.

Table 3. Studies included in the qualitative synthesis.

Ref.	Short Description	Source Title	Publication Type
[12]	This study proposes a multimodal neural model based on a deep learning approach for multimodal deception detection.	At the 2018 International Conference on Computational Linguistics and Intelligent Text Processing	Conference paper
[13]	This study proposes an unsupervised multimodal approach for affect-aware Deep Belief Networks (DBN) to learn discriminative representations of deceptive and truthful behaviors.	From the 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition	Conference paper
[14]	This study develops a feature-level fusion approach, combining audio and video modalities to build an automated system that can help in the decision making of honesty or a lie.	15th International Conference on Computer Vision Theory and Applications 2019	Conference paper
[15]	This study proposes an ensemble-based automated deception detection framework called LiarOrNot for deception detection in group interaction videos.	From the 2019 IEEE International Conference on Multimedia and Expo	Conference paper
[16]	This study presents a benchmark multimodal dataset named Bag-of-Lies for deception detection.	From the Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops	Conference paper
[17]	This study designed a deception detection system based on a multimodal stacked Bi-LSTM model that discriminates between deceptive and truthful behaviors using text, audio, and video modalities.	From the 2023 International Conference on Innovative Data Communication Technologies and Application	Conference paper
[18]	This study investigates several multimodal fusion approaches for automatically distinguishing between deceit and truth based on audio, video, and text modalities.	Multimedia Tools and Applications	Article
[19]	This study explores the use of verbal and non-verbal modalities to build a multimodal deception detection system that aims to discriminate between truthful and deceptive statements.	From the Proceedings of the 2015 ACM on international conference on multimodal interaction	Conference paper
[20]	This study develops a multimodal deep-learning architecture for detecting deception in political debates, which combines textual and acoustic information.	From the 2019 IEEE Automatic Speech Recognition and Understanding Workshop	Conference paper
[21]	This study investigates the importance of visual, acoustic and EEG information on a human subject for a deception detection task.	From the 2022 2nd International Conference on Artificial Intelligence (ICAI)	Conference paper
[2]	This study proposes a face-focused cross-stream network (FFCSN) that induces meta learning and adversarial learning into the training process for deception detection in videos.	From the Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition	Conference Paper
[22]	This study aims to explore high-level features, extracted from different modalities, which can be interpreted by humans while being useful for the automatic detection of deception in videos.	From the Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops	Conference Paper

Table 3. *Cont.*

Ref.	Short Description	Source Title	Publication Type
[23]	This study proposes an end-to-end framework named DEV to detect DEceptive Videos automatically.	From the Proceedings of the 20th ACM international conference on multimodal interaction 2018	Conference Paper
[24]	This study presents a novel analysis of the discriminative power of the facial act for automated deception detection, along with interpretable features from visual, vocal, and verbal modalities.	From the Proceedings of the 2020 International Conference on Multimodal Interaction	Conference paper
[25]	This study presents a multimodal deception detection framework named LieNet based on a deep convolution neural network for differentiating between falsehoods and truth.	IEEE Transactions on Cognitive and Developmental Systems	Article
[26]	This study proposes a novel framework using BERT and Multiscale CNNs to perform multimodal fake news classifications.	From the 2021 2nd Global Conference for Advancement in Technology	Conference paper
[27]	This study tests the use of fully automatically extracted multimodal features for truly automated deception detection.	INTERSPEECH 2020	Conference paper
[28]	This study presents a multimodal system that detects deception in real-life trial data using verbal, acoustic, and visual modalities.	IEEE Transactions on Affective Computing	Article
[29]	This study explores the feasibility of applying AI/ML techniques to detect lies in videos using multiple datasets.	From the 2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications	Conference paper
[30]	This study introduces a novel multimodal dataset for political deception detection.	IEEE MultiMedia	Article
[31]	This study proposes an automated multimodal system named POLLY to predict whether a politician is lying in a video using visual, audio and textual features.	From the Proceedings of the 2022 International Conference on Multimodal Interaction	Conference paper
[32]	This study proposes a framework for automatic deception detection based on micro expressions, audio, and text data captured from videos.	From the 2019 IEEE Conference on Multimedia Information Processing and Retrieval	Conference paper
[33]	This study proposes a multimodal unsupervised transfer learning approach that detects real-world, high-stakes deception in videos without using high-stakes labels.	From the ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing	Conference paper
[34]	This study explores the use of multimodal real-life data for the task of deception detection.	From the Proceedings of the 2015 conference on empirical methods in natural language processing	Conference paper
[35]	This study presents a novel technique for video-based deception detection using the deep recurrent convolutional neural network.	From the 2020 Computer Vision and Image Processing: 4th International Conference	Conference paper
[36]	This study develops a multimodal neural network for lie detection by videos.	From the Companion Publication of the 2020 International Conference on Multimodal Interaction	Conference paper

The majority of the surveyed studies have been published in conference proceedings (85%–22 studies), while only 4 studies are in journals (15%).

The temporal distribution of the analysed papers, as illustrated in the graph in Figure 2, demonstrates an increasing number of papers dealing with multimodal deception detection, which began to pick up steam in 2015 and 2019. Note that in 2023, the considered period was January–September.

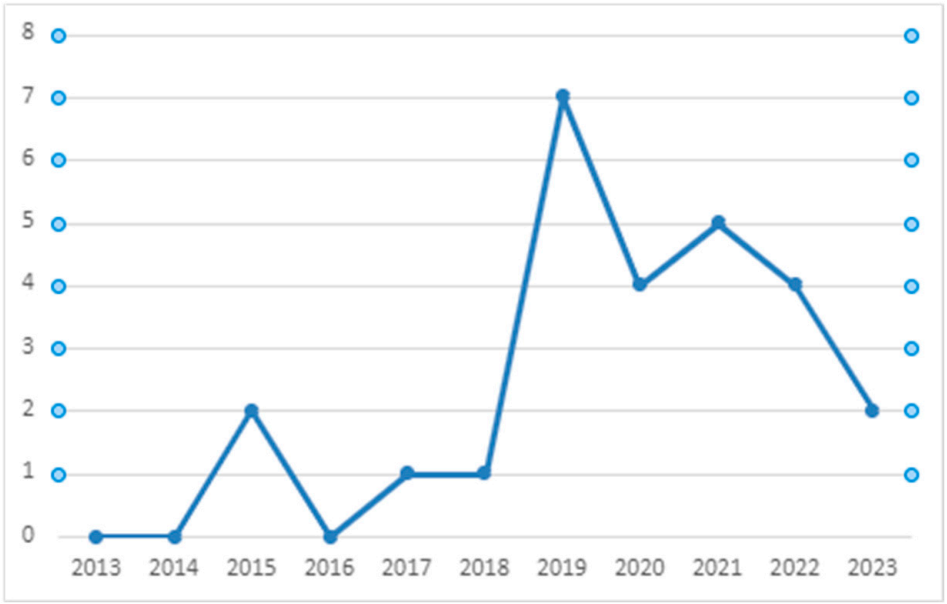

Figure 2. Temporal distribution of the articles.

5. Discussion

This section analyses how the 26 surveyed studies answered four questions introduced in Section 3 and relate to the four steps of the multimodal deception detection process. Specifically, concerning RQ1, the multimodal features extracted for automated deception detection are analyzed. To deal with RQ2, the methodologies used for the fusion of multimodal features are synthesized. Addressing RQ3, the classification algorithms used for multimodal deception detection are investigated. To answer RQ4, a discussion of the datasets used for the analysis of multimodal deception detection is provided.

5.1. Multimodal Features Extracted for Automated Deception Detection

During the multimodal feature extraction process, the unimodal raw data were transformed into a set of features that served as trustworthy indicators for deception detection.

Several types of features were extracted from the literature for detecting deception, including visual features (e.g., facial and body features), audio features, textual features, temporal features, and EEG.

Concerning the facial features, 22 papers (85%) among those surveyed (26 studies) extracted the following 10 facial features, as represented in Table 4: (A) affect features, (B) facial expression, (C) head pose, (D) eyeblink, (E) pupil size, (F) eyebrow motion, (G) mouth motion, (H) gaze motion, (I) local binary patterns (LBP), and (J) lips movement. Seven papers (27%) extracted the following three body features: (K) hand movements, (L) body language, and (M) hand trajectory. Three papers (12%) extracted visual features without specifying which kinds of features.

Concerning the audio features, 13 studies (50%) extracted those represented in Table 4 as follows: (N) pitch-based features, (O) cepstral, (P) spectral, (Q) prosodic, (R) frequency, (S) perceptual, (T) energy, (U) voice quality, (V) Mel-frequency cepstral coefficients (MFCC), (W) linear predictive coding (LPC), (X) zero crossing rate, (Y) chroma frequencies, (Z) latency period, (AA) speech fillers, (AB) speech hesitations, (AC) speech rate, and (AD) pauses. Six papers (23%) extracted audio features without specifying which kinds of features.

Table 4. Multimodal features extracted by surveyed papers.

Ref.	Multimodal Features																																						
	Visual (Not Specified)	Visual Features													Audio (Not Specified)	Audio Features																	Textual (Not Specified)	Textual Features				Temporal Representation	EEG
		Facial Features										Body Features																											
		A	B	C	D	E	F	G	H	I	J	K	L	M		N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD		AE	AF	AG	AH		
[12]			X												X																		X						
[13]		X	X	X					X								X	X	X			X	X															X	
[14]			X	X	X		X	X	X		X	X		X		X																							
[15]			X	X	X			X	X		X													X															
[16]					X	X			X	X								X						X		X	X												X
[17]	X																								X								X						
[18]			X										X					X	X		X		X					X	X	X	X	X							
[19]			X	X	X		X	X	X		X	X		X																				X	X				
[20]															X																		X	X	X				
[21]									X												X			X	X														X
[2]			X										X																									X	
[22]			X	X					X							X								X										X	X		X		
[23]	X														X																		X						
[24]			X	X					X								X	X	X				X	X										X					
[25]			X							X										X																			X
[26]	X																																X						
[27]				X	X		X	X	X		X								X															X	X		X		
[28]			X	X					X								X	X	X				X																
[29]			X	X	X		X	X	X		X	X		X																				X	X		X		
[30]			X	X	X		X	X	X		X	X		X		X																X		X	X				
[31]			X		X				X						X																		X						
[32]			X	X	X	X			X																									X	X	X	X		
[33]			X																					X											X				
[34]			X														X							X											X				
[35]			X	X								X			X																		X						
[36]			X		X				X						X																								

Regarding the textual features, 10 studies (38%) extracted those indicated in Table 4: (AE) Linguistic Inquiry and Word Count (LIWC), (AF) n-grams, (AG) semantic, and (AH) Part-of-speech (PoS) tag. Seven studies (27%) extracted textual features without specifying which kinds of features were extracted.

Finally, temporal representation was extracted by 2 studies (8%), while EEG was extracted by 3 studies (12%).

Table 5 shows the combination of modalities resulting from the analysis of the studies. The most frequent combination of modalities was visual + audio + textual (11 studies—42%), followed by visual + audio (5 studies—19%), visual + textual (4 studies—15%), and visual + audio + EEG (3 studies—12%). Only one study (4%) analyzed the combination of visual + audio + temporal, visual + temporal, and audio + textual, respectively.

Table 5. Combination of modalities resulting from the analysis of the surveyed papers.

Combination of Modalities	Ref.
Visual—audio—textual (11) (42%)	[12,17,22–24,27–29,31,32,35]
Visual—audio—temporal (1) (4%)	[13]
Visual—audio—EEG (3) (12%)	[16,21,25]
Visual—audio (5) (19%)	[14,15,18,33,36]
Visual—textual (4) (15%)	[19,26,30,35]
Visual—temporal (1) (4%)	[2]
Audio—textual (1) (4%)	[20]

Focusing on the analysis of specific modalities, almost all the studies (25 studies—96%) rely on the visual modality, followed by audio (21 studies—81%), textual (16 studies—62%), EEG (3 studies—12%), and temporal (2 studies—8%) studies. Specifically, among the visual modality, the most analyzed feature was facial expression (19 studies—73%), followed by gaze motion (15 studies—58%), head pose (12 studies—46%), eye blink (10 studies—38%), mouth motion and lips movement (6 studies—23%, respectively), eyebrow motion and hand movements (5 studies—19%, respectively), hand trajectory (4 studies—15%), pupil size, local binary patterns, and body language (2 studies—8%, respectively), and affect features (1 study—4%).

Considering the audio modality, the Mel-frequency cepstral coefficients (MFCC) were the most extracted feature (7 studies—27%), followed by spectral and prosodic features (5 studies—19%), cepstral and voice quality (4 studies—15%), pitch-based features (3 studies—12%), perceptual, linear predictive coding, and pauses. Finally, for (2 studies—8%), frequency, energy, zero crossing rate, chroma frequencies, latency period, speech fillers, speech hesitations, and speech rate (1 study—4%) were extracted most.

Finally, for the textual features, n-grams were extracted in 35% of the studies (9 studies) for linguistic inquiry and word count in 31% (8 studies), part-of-speech (PoS) TAGs in 15% (4 studies), and semantic features in 1 study (4%).

5.2. Methodologies Used for the Fusion of Multimodal Features

To integrate the multimodal features, not all the surveyed studies apply a fusion method. Specifically, five studies (19%) do not use multimodal fusion methods. Considering the remaining 21 studies, the vast majority of them (90%—19 studies) are based on model-independent fusion, while only 2 studies (10%) rely on model-dependent fusion, as represented in Table 6. In detail, among the studies based on model-independent fusion, 63% (12 studies) apply the late fusion, 47% (9 studies) the intermediate fusion, 32% (6 studies) early fusion, and 11% (2 studies) hybrid fusion.

Table 6. Types of multimodal fusion and methods applied by surveyed papers.

Ref.	What Kind of Multimodal Fusion?						Which Fusion Method?									
	Model-Independent Fusion				Model-Dependent Fusion		Belief Theory	Deep Belief Networks (DBNs)	Concatenation	Hadamard + Concatenation	Score Level	Probability Avg	Ensemble	Deep Correlation Analysis	Temporally "Informed"	Majority Voting
	Early Fusion	Intermediate Fusion	Late Fusion	Hybrid Fusion	LSTM Model with Attention	Deep Belief Networks (DBNs)										
[12]		X							X	X						
[13]						X		X								
[14]		X														
[15]			X													
[16]			X								X					
[18]		X	X				X									
[20]		X		X								X	X			
[21]			X								X					
[2]				X										X		
[22]	X		X						X				X		X	
[23]		X														
[24]	X	X							X				X			
[25]			X								X					
[26]	X	X							X							
[28]	X		X						X							
[30]		X														
[31]			X										X			
[32]			X													X
[33]	X		X						X			X				
[35]		X	X						X	X						X
[36]					X											

Late fusion is the most commonly used multimodal fusion due to its simplicity compared to the others. Indeed, it processes each modality using powerful targeted approaches specific to the unimodal input and then merges the resulting data. However, this kind of fusion also has some drawbacks related to (i) the high computation cost due to the supervised learning stage necessary for every modality and (ii) the overfitting that takes place when a model performs well with training data but badly with test data.

The second most applied kind of multimodal fusion is intermediate fusion, which has the advantage of being flexible in determining the appropriate depth and sequence of representations, even if it requires the acquisition of a large number of training samples.

Early fusion is the third most applied kind of multimodal fusion that merges all the features at once by producing an accurate representation using a single learning phase. However, the drawback of this approach is the difficulty of combining all the features into a single representation.

Finally, hybrid fusion combines the advantages of late and early fusion; however, it is not commonly used since it leads to a more intricate model structure and more challenging training.

Upon considering the methods for fusing multimodal features, 16 studies (76%) specified the use of method(s), which mainly included concatenation (33%—7 studies), followed by the score level and ensemble (19%—4 studies for each method), Hadamard + concatenation, probability avg, and majority voting (10%—2 studies for each method), and belief theory, Deep Belief Networks, deep correlation analysis, and the temporal informed method (5%—1 study for each method).

The method that was used the most was the concatenation approach, which produces a compact set of salient features that can be used after several steps of feature normalization and feature selection (or transformation). The main benefit of this strategy is the improvement of matching accuracy by removing unnecessary features. The score-level approach is applied in the late fusion and generates a posteriori probability (score) separately for each modality, indicating the likelihood that it belongs to a certain class. This method is quite common and straightforward because it is easy to obtain the scores and has enough information to distinguish between valid and not valid results. In the ensemble fusion method, each modality is first processed separately to provide decision-level results and then combined using different approaches. It offers greater flexibility in terms of feature representations and learning algorithms for different modalities, as well as greater scalability in terms of modalities.

5.3. Classification Algorithms for Multimodal Deception Detection

To automatically distinguish between deceptive and truthful videos, classification algorithms are trained using the multimodal features described in Section 5.1. as input. In particular, out of the 26 papers surveyed, 25 studies (96%) use a deception classification algorithm, as shown in Table 7. The majority of them (23 studies—92%) make use of supervised learning (referred to with an S in Table 7), which relies on labeled training data, as introduced in Section 2. Neural networks were the most often used technique in 11 studies (48%), followed by 9 studies (39%) relying on random forest, 7 studies (30%) using support vector machine (SVM), 6 studies based on K-Nearest Neighborhood (26%), and 3 studies (13%) applying Boosting algorithms and decision tree, respectively. Only 1 study (4%) applied Spectral Regression Kernel Discriminant Analysis (SRKDA), ExtraTrees, and LightGBM.

Table 7. Deception classification algorithms applied by surveyed papers.

Ref.	S vs. U	Neural Networks							K-Nearest Neighbor-hood	Support Vector Machine	Logistic Regres-sion	SRKDA	Random Forest	Boosting Algo-rithm	Decision Tree	Gaussian Naive Bayes	ExtraTrees	LightGBM
		MLP NN	Feed-Forward NN	CNN	R-CNN	Multiscale CNN	Long Short-Term Memory	Deep Belief Networks										
[12]	S	X																
[13]	U							X										
[14]	S								X									
[15]	S								X	X	X		X			X		
[16]	S								X				X					
[17]	S			X			X											
[18]	S								X									
[19]	S												X		X			
[20]	S		X								X							
[21]	S			X														
[2]	S				X													
[22]	S									X								
[24]	S									X								
[25]	S			X														
[26]	S					X												
[27]	S												X					
[28]	S		X							X			X					
[29]	S												X					
[30]	S														X			
[31]	S								X				X			X		
[32]	S						X			X		X	X	X				
[33]	U								X									
[34]	S									X			X		X			
[35]	S				X		X											
[36]	S									X	X			X		X	X	X

Unsupervised methods (referred to with a U in Table 7), which do not need manually annotated features to train the model, were used by two studies (9%). Specifically, they apply Deep Belief Networks and K-Nearest Neighborhood. Various types of neural networks have been experimented on by the surveyed papers, including convolutional neural networks (CNNs) (3 studies—13%), recurrent convolutional neural networks (R-CNNs), feed-forward neural networks, and long short-term memory (2 studies each one—9%), the multilayer perceptron neural network (MLP NN), Multiscale CNN, and Deep Belief Networks (1 study each one—4%).

5.4. Datasets for the Analysis of the Multimodal Deception Detection

The creation and application of appropriately defined assessment datasets is a crucial step for the analysis of multimodal deception detection [37].

The surveyed studies used 16 different datasets in total; Table 8 provides a summary of these datasets along with information on their availability, size, and data source. Note that in the column Size of Table 8, the letter D refers to deceptive videos, T refers to truthful videos, and HT refers to half-true videos. The real-life trial dataset [19], which is available upon request, was the most widely used (16 studies—62%). This dataset includes 121 videos, with an average duration of 28.0 s, gathered from trials held in public courts. A total of 60 trial clips are truthful, and 61 are misleading. Four studies (15%) used the Bag-of-Lies dataset [16], which is publicly available and collected in a spontaneous environment from 35 unique subjects, providing 325 annotated data points with a uniform distribution of truth (163) and lies (162). The Miami University Deception Detection Database (MU3D) [38], a free resource with 320 videos of black and white targets, of both males and females telling truths (160 videos) and lies (160 videos), was utilized in two studies (8%). Further surveyed studies (11—42%) developed their dataset mainly by downloading videos from Youtube/Twitter/Sina Weibo (3 studies—27%), recording videos from games/TV shows (2 studies—18%), using storytelling (2 studies—18%), recording videos from political debates (2 studies—18%), and the use of controlled interviews/interrogations (2 studies—18%). The larger dataset is the TRuLie dataset [36], containing 10.000 clips from 36 male and 51 female participants. Finally, the majority of the datasets are publicly available (8—50%), 5 datasets are not available (31%), and 3 datasets (19%) are available upon request.

Table 8. Datasets used by the selected studies.

Ref.	Dataset	Availability	Size	Source
[12]	Real-life trial	Available upon request	121 videos (61 D/60 T)	"The Innocence Project" website
[13]	Real-life trial	Available upon request	121 videos (61 D/60 T)	"The Innocence Project" website
[14]	Real-life trial	Available upon request	121 videos (61 D/60 T)	"The Innocence Project" website
[15]	Resistance dataset	Not available	185 videos (113 D/172 T)	5 sites of social games
[16]	Bag-of-Lies dataset	Publicly available	325 recordings (162 D/163 T)	Spontaneous environment
[17]	Real-life trial	Available upon request	121 videos (61 D/60 T)	"The Innocence Project" website
	Bag-of-Lies dataset	Publicly available	325 recordings (162 D/163 T)	Spontaneous environment
	Miami University Deception Detection Database	Available upon request	320 videos (160 D/160 T)	Storytelling about social relationships
[18]	Real-life trial	Available upon request	121 videos (61 D/60 T)	"The Innocence Project" website
[19]	Real-life trial	Available upon request	121 videos (61 D/60 T)	"The Innocence Project" website
[20]	CT-FCC-18 corpus	Publicly available	286 recordings (130 D/93 T/63 HT)	Political debates from Youtube

Table 8. *Cont.*

Ref.	Dataset	Availability	Size	Source
[21]	Bag-of-Lies dataset	Publicly available	325 recordings (162 D/163 T)	Spontaneous environment
[2]	Real-life trial	Available upon request	121 videos (61 D/60 T)	"The Innocence Project" website
[22]	Real-life trial	Available upon request	121 videos (61 D/60 T)	"The Innocence Project" website
	Novel Spanish Abortion/Best Friend Database	Not available	42 videos (21 D/21 T)	Storytelling about social relationships
[23]	Real-life trial	Available upon request	121 videos (61 D/60 T)	"The Innocence Project" website
[24]	Real-life trial	Available upon request	121 videos (61 D/60 T)	"The Innocence Project" website
[25]	Real-life trial	Available upon request	121 videos (61 D/60 T)	"The Innocence Project" website
	Bag-of-Lies dataset	Publicly available	325 recordings (162 D/163 T)	Spontaneous environment
	Miami University Deception Detection Database	Available upon request	320 videos (160 D/160 T)	Storytelling about social relationships
[26]	Twitter dataset	Publicly available	Not available	Tweets from Twitter
	Sina Weibo dataset	Publicly available	Not available	Microblogs from the authoritative news agency of China, Xinhua News Agency, and Weibo
[27]	Box of Lies corpus	Publicly available	25 videos	TV shows on Youtube
[28]	Real-life trial	Available upon request	121 videos (61 D/60 T)	"The Innocence Project" website
[29]	Real-Life Trial	Available upon request	121 videos (61 D/60 T)	"The Innocence Project" website
	Opinion dataset	Not available	Not available	Storytelling of movies
	Crime dataset	Not available	Not available	Interviewees under a controlled environment
[30]	Kamboj et al.'s dataset	Not available	180 videos	Political debates
[31]	POLLY dataset	Publicly available	146 videos (73 D/73 T)	Political speeches
[32]	Real-Life Trial	Available upon request	121 videos (61 D/60 T)	"The Innocence Project" website
[33]	Real-life trial	Available upon request	121 videos (61 D/60 T)	"The Innocence Project" website
	UR Lying Dataset	Publicly available	107 videos (63 D/44 T)	controlled game scenarios
[34]	Perez-Rosas et al.'s dataset	Publicly available	118 video clips	TV shows
[35]	Real-Life Trial	Available upon request	121 videos (61 D/60 T)	"The Innocence Project" website
[36]	TRuLie dataset	Available upon request	10.000 annotated videos	Controlled—mock crime interrogations

6. Metrics for the Evaluation of the Multimodal Deception Detection

This section analyses how the surveyed studies answered the RQ5 question introduced in Section 3 on the metrics used for the evaluation of multimodal deception detection.

To evaluate the performance of the deception detection methods, various metrics can be used, ranging from the classification accuracy (ACC) to the area under the precision–recall curve (AUC). Table 9 summarizes the metrics used in the surveyed studies. The ACC was the most commonly used metric (23 studies—88%), often quantified by a correct classification rate (CCR) and measured as the ratio between the number of correct classifications and the total number of classifications. The F1 score, which combines the precision and recall scores, was also commonly used (11 studies—42%), followed by the AUC (10 studies—38%), which represents the area under the precision–recall curve (AUC) over the test set. Nine studies measured the performance of deception detection methods

using precision and recall metrics (35% each). Precision refers to the fraction of positive results among the obtained results, while recall (also known as sensitivity) is the fraction of positive results that were retrieved. Few studies (2—8%) applied the true negative ratio (TNR) and true positive ratio (TPR), which indicates the probability that an actual negative/positive will test negative/positive.

Table 9. Evaluation metrics used by the surveyed studies.

Ref.	Evaluation Metrics	Obtained Values	Best Performing Method
[12]	ACC	96.14%	Multilayer perceptron neural network
	ROC-AUC	0.9799	
[13]	AUC	80%	Deep Belief Networks
	ACC	70%	
	Precision	88%	
[14]	ACC	97%	K-Nearest Neighborhood
	Precision	97%	
	Recall	100%	
	F1 Score	94%	
	TPR	94%	
	TNR	100%	
[15]	AUC	0.705	Logistic Regression + random forest+ linear SVM + Gaussian Naive Bayes
	F1	0.466	
	FNR	0.621	
	FPR	0.142	
	Precision	0.666	
	Recall	0.379	
[16]	ACC	66.17%	Not available
[17]	ACC	98.1%	CNN + LSTM
[18]	ACC	94%	K-Nearest Neighborhood
	Precision	88%	
	Recall	100%	
	F1 Score	94%	
	TPR	100%	
	TNR	87%	
[19]	ACC	75.2%	Decision trees
[20]	Mean Absolute Error (MAE)	0.67	Feed-forward neural network
	Macro-average Mean Absolute Error (MMAE)	0.69	
	ACC	51.04	
	Macro-average F1	45.07	
	Macro-average Recall (MAR)	47.25	
[21]	ACC	83.5%	CNN
	Precision	0.86	
	Recall	0.82	
	F1-score	0.83	
[2]	ACC	97%	R-CNN
	AUC	99.78%	
[22]	AUC ROC	0.671	Support vector machine
[23]	ACC	84.16%	N/A

Table 9. *Cont.*

Ref.	Evaluation Metrics	Obtained Values	Best Performing Method
[24]	AUC	0.91	Support vector machine
	ACC	0.84	
	F1-score	0.84	
[25]	ACC	95	CNN
	Precision	96	
	Recall	94	
	F1 Score	95	
[26]	ACC	73%	Multiscale CNN
	Precision	0.870	
	Recall	0.891	
	F1 Score	0.878	
[27]	ACC	73%	Random forest
	AUC	0.77	
	Precision	0.75	
	Recall	0.77	
	F1 Score	0.74	
[28]	ACC	83.62%	Feed-forward neural network
[29]	ACC	53%	Random forest
	AUC	0.68	
[30]	ACC	69%	Decision tree
	Recall	75%	
[31]	ACC	0.628	Not available
	AUC	0.714	
	F1-score	0.636	
[32]	ACC	97%	Linear SVM, SRKDA, LSTM, random forest.
[33]	AUC	0.64	K-Nearest Neighborhood
	ACC	0.60	
	F1-score	0.69	
[34]	ACC	82.14%	SVM
[35]	ACC	100%	R-CNN + LSTM
[36]	ACC	0.675	LightGBM
	Balanced accuracy	0.638	
	Precision	0.778	
	Recall	0.739	
	F1 Score	0.757	

Considering the obtained values of the evaluation metrics, as shown in the third column of Table 9, the best performing multimodal deception detection method (accuracy = 100%) is that proposed in [35], which integrates R-CNN with LSTM. A similar approach that integrates CNN with LSTM also provides a very good performance in terms of accuracy (98,1%), followed by the K-Nearest Neighborhood, R-CNN, and a combination of Linear SVM, SRKDA, LSTM, and random forest that obtain 97% accuracy. However, it is important to note that the surveyed studies rely on different datasets for testing the performance of the methods; therefore, a comparative analysis is not feasible.

7. Future Directions on Multimodal Deception Detection

This section analyses how the surveyed studies answer RQ6 on future directions of research.

Many directions for future studies arise from the development of multimodal deceptive detection technologies. In Table 10, an overview of the future research directions that were identified from the surveyed studies is provided. This table only shows the studies (20–77%) that contain a discussion on future work.

Table 10. Proposed future research directions emerged from the studies.

Ref.	Future Research Directions
[12]	- Creating a large multimodal dataset with a large number of subjects under various environmental conditions - Identifying deceit in a social dyadic conversational environment
[13]	- Detecting detection and other social behaviors in the wild using unsupervised, affect-aware computational methods
[14]	- Reducing complexity for real-time use by developing approaches for combining speech and video modalities
[16]	- Applying more efficient multimodal fusion techniques and building a more complex network
[17]	- Using large datasets to create a state-of-the-art model which can be used in multiple scenarios
[18]	- Reducing complexity for real-time use by developing approaches for combining speech and video modalities - Studying emotions and behaviors whose detection may be of high importance in high-stakes applications
[19]	- Using automatic gesture identification and automatic speech transcription for real-time deception detection.
[21]	- Using multiple modalities for lie detection tasks on larger and more complex datasets
[22]	- Analysis of fusion methods using the most predictive features - Tuning of hyperparameters for classifiers that can exploit the most predictive features
[23]	- Collect deceptive video datasets from real situations (not produced in laboratory settings) with high-quality videos - Developing a mechanism that offers a form of visual/vocal interpretability - Examining evidence against deception theories, including the interpersonal deception theory, derived from data
[24]	- Developing affect-aware systems for automatically detecting deception and other social behaviors, particularly those occurring in unconstrained situations in the wild
[25]	- Creation of a large multimodal database with a large number of subjects under various environmental settings based on RL videos - Developing a more robust learning system to detect deception efficiently - Extract more complex information from several modalities
[26]	- Build more robust algorithms that could be useful for real-world applications
[27]	- Evaluation of the deception detection models on real-world multimodal deception data - Develop more complex classifiers, such as recurrent neural networks, to model conversational context and time-dependent features to improve automatic deception detection

Table 10. *Cont.*

Ref.	Future Research Directions
[28]	- Use automatic gesture and facial expression identification and automated speech transcription for real-time deception detection
[30]	- Construct more advanced models to classify political statements using additional feature selection approaches - Use transfer learning to explore the feasibility of domain-specific lie detection models
[31]	- Increase the robustness of translation errors for different languages
[34]	- Use automatic gesture and facial expression identification and automated speech transcription for real-time deception detection
[36]	- Make the architecture of neural networks more complex to process more frames - Add additional features (in addition to the audio and video) to the neural network and train all of them end-to-end

Most of the research included in this survey emphasizes the necessity of developing further computational approaches (10 studies—50%), mainly for real-time deception detection [19,26,27], gesture identification [19,28,34], facial expression identification [28,34], and automatic speech transcription [19,28,34], affect-aware computing [13,24], and more efficient learning systems [25,30,36].

The extension of multimodal features is also considered relevant by six studies (30%) to better train neural networks [36], improving the accuracy of the classifiers [27,30], and extract more complex information [18,21,25], as well as the extension of the current datasets (5 studies—25%) for lie detection on larger and complex scenarios and environmental settings [12,17,21,23,25].

A look both at the development of further multimodal fusion approaches (3 studies—15%) for improving the effectiveness of current deception methods [14,16,22] and at the reduction in complexity (2 studies—10%) is also proposed by the surveyed studies [14,18].

Finally, the future research directions that look for the extension of deceptive detection scenarios [12], the optimization of classifiers [22], the interpretability of detection methods [23], the improvement of the evaluation [27], the increase in robustness [31], and the application of further deception theories [23] (1 study—5% each one) should also be mentioned.

8. Conclusions

An overview of the methods and metrics used for the analysis and evaluation of multimodal deception detection on videos and a vision of future directions in this field was given in this work.

The primary contribution involved a discussion of multimodal features, fusion methodologies, classification algorithms, and datasets.

Concerning the multimodal features, 85% of the research used facial features, 50% used audio features, 38% textual features, 27% body features, 12% EEG, and 8% temporal representation. Furthermore, 42% of the studies relied on the combination of modalities visual + audio + textual.

The vast majority of the studies (90%) were based on model-independent fusion, specifically, 63% on late fusion, 47% on intermediate fusion, 32% on early fusion, and 11% on hybrid fusion. The concatenation was the most applied method for fusing multimodal features (33%), followed by the score level and ensemble (19%).

In total, 92% of the researchers employ supervised learning techniques for their classification algorithms, with mainly neural networks accounting for 43% of the studies (43%), followed by random forests (39%), support vector machines (SVMs) (30%), and the K-Nearest Neighborhood (26%) coming in second, third, and fourth, respectively.

Additionally, the most widely used dataset for assessing how well classification algorithms perform is the real-life trial dataset [19], which was used in 62% of the studies, followed by the Bag-of-Lies dataset [16] (15%) and Miami University Deception Detection Database [38] (8%). Several studies (42%) created their dataset primarily through the use of videos from Youtube/Twitter/Sina Weibo (27%), game/TV shows (18%), storytelling (18%), political debates (18%), and controlled interviews/interrogations (18%).

Another important contribution is the analysis of the metrics used for the evaluation of multimodal deception detection. The ACC is the most commonly used metric (23 studies—88%), followed by the F1 score (11 studies—42%) and the AUC (10 studies—38%), while the best performing multimodal deception detection method (accuracy = 100%) was that proposed in [35] which integrates R-CNN with LSTM.

Finally, concerning future directions, most of the studies included in the survey emphasized the necessity of developing computational approaches further (10 studies—50%), followed by the extension of multimodal features to better train neural networks and improve the accuracy of the classifiers (6 studies—30%), as well as the extension of the current dataset (5 studies—25%) for lie detection on larger and complex scenarios and environmental settings.

In light of the analysis provided in this paper, the following important challenges in multimodal deception detection have been detected:

- Overcoming the lack of a complete multimodal dataset for use in multimodal deception detection;
- Developing more efficient fusion techniques to produce richer representations of the information;
- Developing explainable frameworks to help better understand and interpret predictions made by multimodal sensing models;
- Developing general and transferable multimodal deception detection models.

The study has a few drawbacks, mostly related to its focus on multimodal deception detection from videos. The lack of substantial real-world deception data is a major obstacle to multimodal deception detection systems performing well. In this sense, expanding the dataset and emphasizing the real-world context could significantly address the issue. Using temporal models or extracting temporal features to incorporate time-dependent data are other potential pathways for the future.

Author Contributions: Conceptualization, A.D. (Alessia D'Andrea), A.D. (Arianna D'Ulizia), P.G. and F.F.; Methodology, A.D. (Alessia D'Andrea), A.D. (Arianna D'Ulizia) and P.G.; Validation, A.D. (Alessia D'Andrea) and A.D. (Arianna D'Ulizia); Formal analysis, A.D. (Alessia D'Andrea) and A.D. (Arianna D'Ulizia); Investigation, A.D. (Alessia D'Andrea), A.D. (Arianna D'Ulizia) and P.G.; Data curation, A.D. (Alessia D'Andrea), A.D. (Arianna D'Ulizia) and F.F.; Writing—original draft preparation, A.D. (Alessia D'Andrea) and A.D. (Arianna D'Ulizia); Writing—review and editing, A.D. (Alessia D'Andrea) and A.D. (Arianna D'Ulizia); Supervision, A.D. (Arianna D'Ulizia). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. D'Ulizia, A.; D'Andrea, A.; Grifoni, P.; Ferri, F. Detecting Deceptive Behaviours through Facial Cues from Videos: A Systematic Review. *Appl. Sci.* **2023**, *13*, 9188. [CrossRef]

2. Ding, M.; Zhao, A.; Lu, Z.; Xiang, T.; Wen, J.R. Face-focused cross-stream network for deception detection in videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7802–7811.
3. Abouelenien, M.; Perez-Rosas, V.; Mihalcea, R.; Burzo, M. Detecting deceptive behavior via integration of discriminative features from multiple modalities. *IEEE Trans. Inf. Forensics Secur.* **2016**, *12*, 1042–1055. [CrossRef]
4. Wu, Z.; Singh, B.; Davis, L.S.; Subrahmanian, V.S. Deception detection in videos. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
5. D'Andrea, A.; D'Ulizia, A.; Ferri, F.; Grifoni, P. EMAG: An extended multimodal attribute grammar for behavioural features. *Digit. Sch. Humanit.* **2015**, *32*, fqv064. [CrossRef]
6. D'Andrea, A.; Caschera, M.C.; Ferri, F.; Grifoni, P. MuBeFE: Multimodal Behavioural Features Extraction Method. *JUCS J. Univers. Comput. Sci.* **2021**, *27*, 254–284. [CrossRef]
7. Yang, F.; Ning, B.; Li, H. An Overview of Multimodal Fusion Learning. In *Mobile Multimedia Communications. MobiMedia 2022*; Chenggang, Y., Honggang, W., Yun, L., Eds.; Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering; Springer: Cham, Switzerland, 2022; Volume 451. [CrossRef]
8. D'Ulizia, A. Exploring multimodal input fusion strategies. In *Multimodal Human Computer Interaction and Pervasive Services*; IGI Global: Hershey, PA, USA, 2009; pp. 34–57.
9. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Int. J. Surg.* **2021**, *88*, 105906. [CrossRef] [PubMed]
10. Pranckutė, R. Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic World. *Publications* **2021**, *9*, 12. [CrossRef]
11. Paez, A. Grey literature: An important resource in systematic reviews. *J. Evid. Based Med.* **2017**, *10*, 233–240. [CrossRef] [PubMed]
12. Krishnamurthy, G.; Majumder, N.; Poria, S.; Cambria, E. A deep learning approach for multimodal deception detection. In Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing, Hanoi, Vietnam, 18–24 March 2018; Springer Nature: Cham, Switzerland, 2018; pp. 87–96.
13. Mathur, L.; Matarić, M.J. Affect-aware deep belief network representations for multimodal unsupervised deception detection. In Proceedings of the 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), Jodhpur, India, 15–18 December 2021; IEEE: New York, NY, USA, 2021; pp. 1–8.
14. Chebbi, S.; Jebara, S.B. An Audio-Visual based Feature Level Fusion Approach Applied to Deception Detection. In Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP) 2020, Valletta, Malta, 27–29 February 2020; pp. 197–205.
15. Bai, C.; Bolonkin, M.; Burgoon, J.; Chen, C.; Dunbar, N.; Singh, B.; Subrahmanian, V.S.; Wu, Z. automatic long-term deception detection in group interaction videos. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; IEEE: New York, NY, USA, 2019; pp. 1600–1605.
16. Gupta, V.; Agarwal, M.; Arora, M.; Chakraborty, T.; Singh, R.; Vatsa, M. Bag-of-lies: A multimodal dataset for deception detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
17. Sehrawat, P.K.; Kumar, R.; Kumar, N.; Vishwakarma, D.K. Deception Detection using a Multimodal Stacked Bi-LSTM Model. In Proceedings of the International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Dehradun, India, 14–15 March 2023; IEEE: New York, NY, USA, 2023; pp. 318–326.
18. Chebbi, S.; Jebara, S.B. Deception detection using multimodal fusion approaches. *Multimed. Tools Appl.* **2021**, *82*, 13073–13102. [CrossRef]
19. Pérez-Rosas, V.; Abouelenien, M.; Mihalcea, R.; Burzo, M. Deception detection using real-life trial data. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Motif Hotel, Seattle, WA, USA, 9–13 November 2015; pp. 59–66.
20. Kopev, D.; Ali, A.; Koychev, I.; Nakov, P. detecting deception in political debates using acoustic and textual features. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; IEEE: New York, NY, USA, 2019; pp. 652–659. [CrossRef]
21. Javaid, H.; Dilawari, A.; Khan, U.G.; Wajid, B. EEG Guided Multimodal Lie Detection with Audio-Visual Cues. In Proceedings of the 2nd International Conference on Artificial Intelligence (ICAI), Islamabad, Pakistan, 30–31 March 2022; IEEE: New York, NY, USA, 2022; pp. 71–78.
22. Rill-García, R.; Jair Escalante, H.; Villasenor-Pineda, L.; Reyes-Meza, V. High-level features for multimodal deception detection in videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
23. Karimi, H. Interpretable multimodal deception detection in videos. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018; pp. 511–515.
24. Mathur, L.; Matarić, M.J. Introducing representations of facial affect in automated multimodal deception detection. In Proceedings of the 2020 International Conference on Multimodal Interaction, New York, NY, USA, 25–29 October 2020; pp. 305–314.
25. Karnati, M.; Seal, A.; Yazidi, A.; Krejcar, O. LieNet: A deep convolution neural network framework for detecting deception. *IEEE Trans. Cogn. Dev. Syst.* **2022**, *14*, 971–984. [CrossRef]

26. Raj, C.; Meel, P. Microblogs Deception Detection using BERT and Multiscale CNNs. In Proceedings of the 2nd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 1–3 October 2021; IEEE: New York, NY, USA, 2021; pp. 1–6.
27. Zhang, J.; Levitan, S.I.; Hirschberg, J. Multimodal Deception Detection Using Automatically Extracted Acoustic, Visual, and Lexical Features. In Proceedings of the INterspeech, Shanghai, China, 25–29 October 2020; pp. 359–363. [CrossRef]
28. Sen, M.U.; Perez-Rosas, V.; Yanikoglu, B.; Abouelenien, M.; Burzo, M.; Mihalcea, R. Multimodal deception detection using real-life trial data. *IEEE Trans. Affect. Comput.* **2021**, *13*, 306–319. [CrossRef]
29. Belavadi, V.; Zhou, Y.; Bakdash, J.Z.; Kantarcioglu, M.; Krawczyk, D.C.; Nguyen, L.; Rakic, J.; Thuriasingham, B. MultiModal deception detection: Accuracy, applicability and generalizability. In Proceedings of the Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), Atlanta, GA, USA, 28–31 October 2020; IEEE: New York, NY, USA, 2020; pp. 99–106.
30. Kamboj, M.; Hessler, C.; Asnani, P.; Riani, K.; Abouelenien, M. Multimodal political deception detection. *IEEE Multimed.* **2020**, *28*, 94–102. [CrossRef]
31. Bai, C.; Bolonkin, M.; Regunath, V.; Subrahmanian, V. POLLY: A multimodal cross-cultural context-sensitive framework to predict political lying from videos. In Proceedings of the 2022 International Conference on Multimodal Interaction, Bengaluru India, 7–11 November 2022; pp. 520–530.
32. Venkatesh, S.; Ramachandra, R.; Bours, P. Robust algorithm for multimodal deception detection. In Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 28–30 March 2019; IEEE: New York, NY, USA, 2019; pp. 534–537.
33. Mathur, L.; Mataric, M.J. Unsupervised audio-visual subspace alignment for high-stakes deception detection. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: New York, NY, USA, 2021; pp. 2255–2259.
34. Pérez-Rosas, V.; Abouelenien, M.; Mihalcea, R.; Xiao, Y.; Linton, C.J.; Burzo, M. Verbal and nonverbal clues for real-life deception detection. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 2336–2346.
35. Venkatesh, S.; Ramachandra, R.; Bours, P. Video based deception detection using deep recurrent convolutional neural network. In Proceedings of the Computer Vision and Image Processing: 4th International Conference, CVIP 2019, Jaipur, India, 27–29 September 2019; Revised Selected Papers, Part II. Springer: Singapore, 2020; Volume 4, pp. 163–169.
36. Karpova, V.; Popenova, P.; Glebko, N.; Lyashenko, V.; Perepelkina, O. "Was It You Who Stole 500 Rubles?"-The Multimodal Deception Detection. In Proceedings of the Companion Publication of the 2020 International Conference on Multimodal Interaction, Virtual, 25–29 October 2020; pp. 112–119.
37. D'Ulizia, A.; Caschera, M.C.; Ferri, F.; Grifoni, P. Fake news detection: A survey of evaluation datasets. *PeerJ Comput. Sci.* **2021**, *7*, e518. [CrossRef] [PubMed]
38. Lloyd, E.P.; Deska, J.C.; Hugenberg, K.; McConnell, A.R.; Humphrey, B.T.; Kunstman, J.W. Miami University deception detection database. *Behav. Res. Methods* **2018**, *51*, 429–439. [CrossRef] [PubMed]