**Engineering and Technology Journal e-ISSN: 2456-3358**

**Volume 10 Issue 04 April-2025, Page No.- 4476-4490**

**DOI: 10.47191/etj/v10i04.13, I.F. – 8.482**



# Artificial Intelligence for Deception Detection: A Multimodal Review of Methods, Challenges, And Ethical Perspectives

**Redeer Avdal Saleh[1], Omar Sedqi Kareem[2]**

[1]Akre University for Applied Sciences, College of Informatics, Information Technology

[2]Department of Public Health, College of Health and Medical Techniques - Shekhan, Duhok Polytechnic University, Duhok 42001, Iraq

**ABSTRACT:** Within the realm of deception detection research, this comparative study investigates the use of machine learning, artificial intelligence, and multimodal data processing. From the year 2020 to the year 2024, it focuses on twenty-four studies that show the growing potential of AI-driven systems in terms of enhancing the consistency, scalability, and accuracy of fraud detection. In order to identify deceit in a variety of data types, such as facial expressions, audio signals, written language, and behavioral abnormalities, different techniques have showed promise. Some of these techniques include Support Vector Machines (SVM), Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs), and hybrid models. On the other hand, issues like as adversary manipulation, biases in training datasets, and the potential for deception cues to be generalized across linguistic, cultural, and social contexts continue to be a concern. To further complicate the deployment of deception detection systems, there is a dearth of real-world validation, and the present models have little adaptability in dynamic environments. The article places an emphasis on the need of openness in the design of artificial intelligence, ethical concerns about user privacy, and the development of systems that have properties that are sensitive to cultural and environmental factors. The integration of concepts from other disciplines, the ability to withstand assaults from adversaries, and the development of ways to decrease prejudice should be the primary focus of research in the future.

**KEYWORDS:** Deception Detection, Machine Learning, Multimodal Analysis, Adversarial Robustness, Cross-Cultural AI.

## 1. INTRODUCTION

From law enforcement and forensic psychology to cybersecurity and digital communication, deception a sophisticated and sometimes context-dependent behavior-offers major difficulties. Detecting dishonesty has always mostly depended on human judgment, which is prone to prejudice, inconsistency, and limited cognitive ability[1][2].Through the development of artificial intelligence (AI) and machine learning (ML), researchers have been gradually investigating automated systems that are capable of analyzing subtle behavioral, linguistic, and physiological indicators in order to identify fraudulent activity with improved accuracy and scalability[3]. This is being done in order to combat the growing prevalence of fraudulent activity[4][5]. Under cognitive stress, for instance, facial micro-expression analysis using support vector machines (SVMs) has demonstrated more efficacy in identifying dishonesty than human observers, specifically[6]. Likewise, text-based deception detection has used part-of-speech characteristics and natural language processing (NLP) approaches to find dishonesty trends in emails, online reviews, and social media postings [7][8]. Integration of multimodal data such as voice, facial expression, and textual content-has considerably enhanced the effectiveness of deception detection systems, therefore allowing them to beat conventional single-modal techniques [9][10]. Furthermore, specific frameworks in fields like false news detection have arisen where discourse-level deception signals have been captured using rhetorical structure theory (RST) and neural embedding's[11][12]. The field of cybersecurity has also embraced deception, not just for the purpose of detection but also as a proactive defensive mechanism that employs decoys and honeypots to confound attackers and gather information[13][14]. However, despite the fact that the results are positive, there are still considerable challenges[15]. The detection models are rendered unreliable in high-risk environments due to the fact that adversarial attacks are able to modify them easily[16][3]. Another factor that makes it difficult to generalize trained models is the fact that there is a variation in deception signals across different cultures. This highlights the need of having systems that are both morally transparent and culturally flexible[17][18].

Here are some additional contributions from the studies:

- **Comprehensive Multimodal Review**: This paper presents a detailed synthesis of 24 key deception detection studies conducted between 2020 and 2024, spanning diverse modalities such as facial micro-expressions, acoustic signals, textual content,

behavioral patterns, and cyber deception. It integrates findings across high-stakes interviews, fake news detection, gaming environments, and adversarial cybersecurity contexts.

- **Cross-Domain Methodological Comparison**: It offers a comparative evaluation of various machine learning and artificial intelligence techniques-including SVMs, LSTM networks, CNNs, ensemble models, RST frameworks, and transfer learning approaches-highlighting their strengths, limitations, and performance across different data types and application areas.
- **Identification of Core Challenges**: The review uncovers persistent challenges in the field, including adversarial vulnerability, dataset bias, generalization failures across cultures and languages, and the lack of robust real-world validation for many deception detection models.
- **Integration of Ethical and Practical Concerns**: Beyond technical performance, the study critically examines ethical issues related to fairness, privacy, and misuse of AI in surveillance or profiling. It emphasizes the need for culturally adaptive, transparent, and explainable deception detection systems.
- **Strategic Recommendations for Future Research**: The paper concludes with actionable recommendations, calling for interdisciplinary collaboration, the development of adversarial robust and culturally sensitive models, enhanced multimodal integration, and the prioritization of real-world deployment and benchmarking.

This investigation is broken down into eight distinct pieces. The first half of this research presents the introduction to the study, while the second section presents the mechanism that is being regarded for the phases of the research technique. In the third section, we will discuss the essential prerequisite theory that is associated with the topic that was done. Section four, which tackles the 24 earlier works that are the most closely connected to our study issue, will, nevertheless, be where the relevant works are presented. The evaluation of the literature was then followed by a comprehensive comparison and an adequate discussion, which were described in the fifth part. Furthermore, in order to facilitate the comparison procedure, it is essential to extract the relevant statistics pertaining to the dependent measures. These particulars, together with their charts, are supplied in section six. When readers are reading any review paper, they want to get a number of suggestions that will make it simpler for them to do fresh research associated with the same topics. These recommendations are offered in section seven of the review article. A conclusion is presented in the eighth part, which includes a summary of the study that was conducted together with the significant findings. Following that, a list of the references that were taken into consideration is shown.

## 2. RESEARCH METHODOLOGY

This study adopts a structured and systematic literature review (SLR) approach to examine deception detection research published between 2020 and 2024. The methodology is designed to ensure the inclusion of high-quality, relevant, and diverse studies that cover various modalities, algorithms, datasets, and application contexts within deception detection. The research process is divided into the following phases

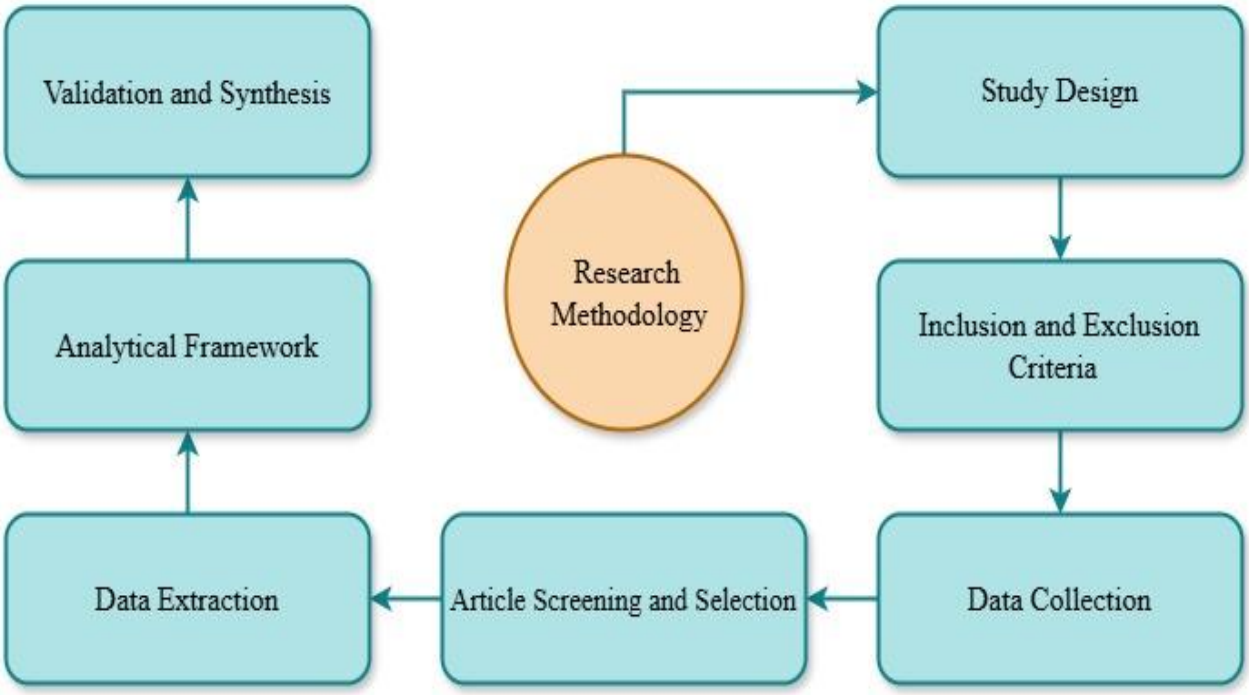


**Figure1: General Flowchart of the Methodology.**

### 2.1 Study Design

The research employs a qualitative meta-analysis of peer-reviewed articles, conference proceedings, and scientific reports focusing on artificial intelligence, machine learning, and deception detection. It aims to synthesize findings from multidisciplinary fields such as computer science, psychology, cybersecurity, linguistics, and behavioral analytics.

### 2.2 Inclusion and Exclusion Criteria

- **Inclusion Criteria:**
  - Studies published from **2020 to 2024**
  - Research that involves **AI/ML-based deception detection**
  - Studies using **multimodal**, **textual**, **visual**, **auditory**, or **cybersecurity-related** data
  - Papers with **experimental evaluation**, **novel frameworks**, or **statistical performance outcomes**
- **Exclusion Criteria:**
  - Non-English articles
  - Editorials, blogs, or opinion pieces without empirical support
  - Studies without access to full text or lacking methodological transparency

### 2.3 Data Collection

Relevant literature was identified using a keyword-based search strategy in databases such as **Scopus**, **IEEE Xplore**, **SpringerLink**, **PubMed**, and **Google Scholar**. Keywords used included: "deception detection," "machine learning," "fake news detection," "micro-expressions," "cyber deception," "multimodal analysis," and "facial action coding."

### 2.4 Article Screening and Selection

Out of an initial pool of over 150 articles:

- **83 articles** were shortlisted after abstract-level review.
- **41 studies** were selected after a full-text review based on relevance and quality.

### 2.5 Data Extraction

A standardized form was used to extract information from each study including:

- Author(s) and year of publication
- Focus area or domain
- Methods or algorithms used
- Datasets and sources
- Key results and performance metrics (e.g., accuracy, AUC)
- Reported limitations or future directions

### 2.6 Analytical Framework

Data from selected studies were coded and categorized according to five dimensions:

- **Focus Area** (e.g., facial analysis, text-based deception, cyber deception)
- **Methodology** (ML techniques used)
- **Datasets** (public or custom datasets)
- **Results** (accuracy, robustness, novelty)
- **Limitations** (e.g., dataset bias, generalizability, adversarial vulnerability)

A **comparative matrix** was developed to enable cross-study analysis and trend identification.

### 2.7 Validation and Synthesis

- The extracted data were **analyzed both qualitatively and quantitatively** to identify patterns, overlaps, gaps, and emerging trends.
- Term frequency analysis was conducted for each category to identify the most frequently used techniques and terms.
- The methodology allows **triangulation** between different data types, ensuring consistency and robustness in the synthesis process.

## 3. BACKGROUND THEORY

- **Micro-Expression Theory (Ekman)**: Paul Ekman's theory of micro-expressions posits that fleeting, involuntary facial movements can betray a person's true emotions, even when they attempt to mask them. These micro-expressions serve as valuable indicators of deception and have been widely integrated into facial analysis systems using AI and ML techniques[19][20].
- **Linguistic Cues and Information Manipulation Theory**: Deceptive language often exhibits measurable changes such as fewer self-references, more negative emotion words, and lower narrative coherence. These concepts stem from theories like Information Manipulation Theory and the Truth-Default Theory, which form the basis of many natural language processing (NLP) deception detection models [11][7].
- **Machine Learning Foundations**: Deception detection models frequently use algorithms such as Support Vector Machines (SVM), Random Forests, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks. These approaches allow for non-linear pattern recognition and temporal analysis of sequential data [21][22].
- **Multimodal Deception Detection**: Modern systems integrate multiple data types—such as facial expressions, voice tone, and textual content-to improve accuracy and reduce reliance on a single modality. This approach reflects the real-world complexity of deception and has shown improved performance in high-stakes scenarios [23].

- **Cyber Deception and Game Theory**: In cybersecurity, deception is used proactively through strategies like honeypots, decoys, and honey tokens. These are grounded in game-theoretic models that aim to manipulate attacker behavior and delay or mislead intrusions[24].
- **Adversarial Learning and Model Robustness**: Deception detection models are susceptible to adversarial attacks where subtle changes in input (e.g., a word or facial feature) can lead to misclassification. This area of research emphasizes the need for robust models capable of defending against such manipulation[25][26].

## 4. LITERATURE REVIEW

Merylin Monaro et al.,2022[19]: This work compared human judges with machine learning algorithms to investigate face micro-expression based dishonesty detection. Using a low-stakes dataset of video interviews, researchers examined how participants-some advised to lie about a holiday and others to accurately recollect one-behaved. Under cognitive stress situations, machine learning methods-especially support vector machines (SVMs) paired with OpenFace features-shown better accuracy (up to AUC = 0.78) in identifying liars than in humans, whose accuracy was only 57%. The research underlined how increasing cognitive load-by means of unexpected questions-made deception signs more visible, hence improving human and computer performance. In the end, particularly when using high-level and automated feature extraction techniques, the results confirmed the efficiency of artificial intelligence over human judgments in spotting dishonesty.

Budi Gunawan et al.,2022[27]:Research trends concerning the employment of technology in false news and deception detection from 2011 to 2021 were bibliometricly analyzed in this work. Using tools like VOSviewer and data taken from Scopus, it looked at the development of publications, major contributors, citation patterns, and research teams. Particularly convolutional neural networks (CNNs) and hybrid models, the paper emphasized developments in machine learning and artificial intelligence as major instruments in spotting dishonesty in news material. It also looked at patterns of worldwide research distribution and underlined the need of digital literacy in stop the dissemination of false information. The results provide understanding of cooperation networks and a basis for more creative ideas in techniques of deception detection.

Delgado et al.,2021[7]:This paper explored the application of machine learning techniques to detect deception in text-based communications such as fake reviews, emails, and news articles. The study focused on using features like Part of Speech (POS) tags and Bag of Words (BOW) to identify deceptive patterns. Neural Networks, Random Forest, Support Vector Machine, and other algorithms were evaluated for their effectiveness in distinguishing between truthful and deceptive texts. The findings indicated that single features, especially POS tags, provided better results than combined features or BOW alone, achieving an average accuracy of 78.88%. The research highlighted the potential for shared characteristics across different types of deceptive communications to enhance detection methods and provided a foundation for developing more robust datasets and advanced algorithms for future studies.

Rinaldo Gagiano et al.,2021[16]:This paper analyzed the robustness of Grover, a model designed for generating and detecting neural fake news, with a specific focus on deception detection. Researchers investigated Grover's vulnerability to adversarial attacks, including character-level and word-level perturbations, and found that even minimal alterations, such as changing one character, could significantly compromise its detection capabilities. Experiments revealed that up to 97% of targeted machine-generated articles could be misclassified after adversarial changes, exposing weaknesses in Grover's encoding and classification processes. To further interpret the model's behavior, a novel visualization of cumulative classification scores was developed, highlighting the impact of specific alterations on Grover's predictions. These findings underscored the challenges in building robust deception detection systems and emphasized the importance of addressing vulnerabilities to maintain the integrity of neural fake news detection tools.

Francielle Vargas et al.,2021[28]:This research focused on deception detection through the analysis of discourse structure patterns in multilingual fake news using the Rhetorical Structure Theory (RST) framework. It introduced a novel multilingual deceptive news corpus called "Deceiver," which consisted of 600 annotated texts in Brazilian Portuguese and English, classified into truthful and deceptive categories. Two new rhetorical relations, INTERJECTION and IMPERATIVE, were proposed to capture pragmatic cues indicative of deception. By leveraging linguistic and discourse patterns, including coherence relations and nuclearity properties, the study aimed to develop computational models that efficiently identify fake news across multiple languages. This research emphasized the significance of cognitive-based approaches and rich contextual analyses for deception detection tasks.

Mu Zhu et al.,2021[29]:This study examined cybersecurity defensive deception strategies, with a focus on methods based on machine learning and game theory. In order to influence attackers' perceptions and make less-than-ideal choices, it examined tactics in which defenders used deceit, such as decoys or misleading information. The study categorized defensive deception techniques, spoke about how

successful they are, and pointed out how difficult it is to use them in different network situations. It also emphasized how game-theoretic models and machine learning algorithms work together to improve deception detection techniques by providing information on attacker engagement, attack detection, and system defense. To overcome the limitations and further improve defensive deception tactics, further research areas were recommended.

Leena Mathuret al.,2021[9]:This research addressed the problem of limited labeled data by presenting a unique unsupervised transfer learning method to identify deceit in high-stakes scenarios. In order to adapt audio-visual elements from low-stakes lab-controlled settings for usage in high-stakes real-world scenarios, it invented Subspace Alignment (SA). In deception detection, researchers outperformed both models without SA and human performance by using SA to obtain significant gains in accuracy (74%) and AUC (75%). The study showed how multimodal behavioral cues, such as eye contact, vocal characteristics, and facial expressions, can function as transferable markers of deception in various social contexts, making SA a useful tool for detecting deception without the need for high-stakes labels.

Shravika Mittal et al.,2021[30]:This study examined methods for fooling community detection algorithms, emphasizing the idea of "community deception." Through the optimization of rewiring procedures with minimum edge updates, it presented NEURAL, a unique algorithm intended to conceal important community structures. The paper showed how well NEURAL can fool six popular community identification algorithms and formulated an objective function to lower community detection accuracy using a node-centric measure termed Permanence. Experiments on both synthetic and real-world networks demonstrated that NEURAL was able to capture important meta-information about edges that goes beyond their topological structure, outperforming previous approaches. The results emphasized the ramifications and possible uses of deception methods in network analysis.

Nguyen Van Huynh et al.,2021[31]:This study presented a unique framework to combat super-reactive jamming assaults, which are very challenging to defeat since the jammer may simultaneously monitor and attack broadcasts. In order to trick the jammer into attacking again, the research used a clever deception technique in which the transmitter claimed to keep sending data. The transmitter then ensured communication under jamming situations by reflecting the jammer's signals for data transfer using ambient backscatter communication technology. Using Long Short-Term Memory (LSTM) networks, a deep learning technique that dynamically adjusted to various channels and noise distributions, the identification of backscattered signals was enhanced. The framework used deception to transform a deficit into better bit error rate (BER) performance, as the results showed, and improved system performance as the jammer increased assault strength. The research demonstrated the capability of deep learning in identifying sophisticated physical layer security techniques and weak backscattered signals.

Hammad-Ud-Din Ahmed et al.,[32]:This study employed the Facial Action Coding System (FACS) to evaluate facial expressions in order to identify fraud in films. It used deep learning methods, namely Long Short-Term Memory (LSTM) networks, which were trained on a variety of datasets, such as the Bag-of-Lies Dataset, the Silesian Deception Dataset, and the Real-life Trial Dataset. Cross-validation experiments showed that differences in data properties, such stakes and recording settings, made accuracy lower when datasets were combined. The research used OpenFace to optimize facial action unit extraction, which resulted in higher detection rates. The results showed that competitive deception detection results were obtained by visual-only methods using deep learning models; however, they also highlighted the difficulties associated with dataset variability and the possibility for improvements in multimodal techniques.

Erich C. Walter et al.,2020[13]:This research examined the effects of deception methods on cyber security in simulated settings by integrating them into Microsoft's CyberBattleSim. Using reinforcement learning algorithms, deceptive components such as honeypots, decoys, and honeytokens were added to delay and mislead attackers. It was shown via experiments that elements such as ingredient kind, number, and location affected how successful deception was. While decoys did a good job of delaying credential-based assaults, honeypots greatly hindered attackers by squandering resources and setting off indicators. As early warning systems that provide defenders actionable information, the results also highlighted the importance of misleading components. Incorporating deception into autonomous defensive frameworks to improve cybersecurity measures was highlighted by the research.

V. Kozlov et al.,2021[33]:This study examined a new technique for radar range deception that makes use of time-modulated scatterers and takes advantage of the relationship between range and Doppler estimates in contemporary radar systems. In order to successfully conceal the true trajectories of targets, researchers showed how to manipulate the phase of backscattered signals in order to trick radars into guessing the objects' erroneous locations and velocities. Frequency-modulated continuous wave (FMCW) radar experiments verified that it is possible to manipulate radar signals to exaggerate the distance or proximity of objects. In order to prevent such deception tactics and guarantee precise range detection, the research also suggested conjugate symmetric waveforms. Deception detection in electromagnetic systems was better understood because to

these discoveries, which also exposed flaws in radar data processing.

Aidan O'Gara et al.,2023[34]:This research used a text-based game called Hoodwinked, which was modeled after Mafia and Among Us, to assess the deception and lie-detection skills of language models such as GPT-3, GPT-3.5, and GPT-4. After seeing killings, players participated in conversations to find and remove the impersonator. The results demonstrated that sophisticated models were more adept at misleading people by shifting the blame during conversations, influencing the results of votes, and lowering the killer's rate of exile. GPT-4, for example, was often more convincing and was effective in deceiving players who did not have concrete proof of the crime. Even while the conversation encouraged collaboration, it also had the unintended consequence of making eyewitnesses less accurate in identifying the murderer because of dishonest tactics. The study brought to light inverse scaling tendencies in deceit, wherein more competent models demonstrated more capacity for deception, highlighting both technical and ethical issues in the development of AI systems.

Jiawei Liang et al.,2024[25]:This study introduced a novel backdoor approach known as the Poisoned Forgery Face framework to target flaws in face forgery detection systems. A clean-label attack was created by researchers, who added triggers to face detection algorithms without changing the training labels. In order to make sure that poisoned samples remained undetectable, they created a scalable trigger generator and used covert embedding techniques. The framework identified security vulnerabilities wherein, in the presence of triggers, detectors may incorrectly identify faked faces as authentic. Numerous tests shown that the attack outperformed current backdoor techniques in terms of the high success rate (+16.39% BD-AUC improvement) and decreased visibility (-12.65% $L\infty$). The results made it clear that face forgery detection systems require better defenses.

Oana Ignat et al.,2023[35]:This study introduced the MAIDE-UP dataset, which consists of 20,000 hotel evaluations in ten languages, evenly split between 10,000 real human-written reviews and 10,000 fraudulent AI-generated ones. It looked at the linguistic variations between actual and AI-generated reviews as well as the mood, location, and language characteristics that affect fraud detection ability. By using interpretable baselines like Random Forest and refined models like XLM-RoBERTa, the research was able to identify AI-generated reviews with high accuracy (94.8%). When comparing AI evaluations to human ones, the results showed stylistic characteristics like reduced readability and increased descriptiveness. The study emphasized the difficulties in detecting deception in multilingual contexts and the significance of using sophisticated natural language processing models to detect misleading material in a variety of languages and contexts.

Francielle Vargas et al.,2022[11]:This paper surveyed the application of Rhetorical Structure Theory (RST) for online deception detection, specifically in tasks such as fake news and fake reviews identification. It systematically reviewed how discourse-level structures, including coherence relations and nuclearity information, were used to distinguish deceptive from truthful texts. The study highlighted discourse-aware approaches, employing RST-based features like "Bag-of-RST," dependency parsing, and neural embeddings, alongside machine learning models such as SVM and LSTM. Despite limited annotated corpora and challenges with RST parsers, the research demonstrated that deceptive stories often exhibited distinct rhetorical patterns. These findings underscored the potential of leveraging discourse-level analysis for deception detection while acknowledging methodological constraints and gaps in annotated resources.

huang-cheng chou et al.,2021[36]:This study proposed a comprehensive framework for automatic deception detection in Mandarin dialogs, focusing on integrating acoustic, textual, and conversational temporal dynamics features. It utilized the Daily Deceptive Dialogues corpus, achieving an unweighted average recall (UAR) of 80.61%. The research highlighted that specific acoustic features like loudness and MFCC, conversational dynamics, and implicature patterns (e.g., complication, common knowledge, and self-handicapping) were significant indicators of deception. Models trained with BLSTM and hierarchical attention networks showed improved performance, surpassing human accuracy in identifying deception. The findings emphasized the effectiveness of combining speech, language, and pragmatic behaviors in enhancing deception detection across cultural contexts.

Despoina Mouratidis et al.,2021[37]:This study introduced a deep learning framework for detecting fake news on Twitter, emphasizing deception detection through pairwise textual input schemas. Researchers leveraged multimodal input, integrating word embeddings with linguistic and network account features to classify tweets as real or fake. The dataset included tweets from Hong Kong protests in 2019, categorized into headers and texts for comparative analysis. The innovative neural network architecture fused inputs at multiple layers, achieving high accuracy, particularly after applying SMOTE oversampling to address class imbalance. Results highlighted that real text correlated more effectively with data, demonstrating higher precision and recall metrics, and establishing the deep learning model as a robust tool for deception detection in social media contexts.

Katerina Papantoniou et al.,2022[17]:This paper investigated cross-cultural deception detection in text, focusing on how cultural dimensions like individualism and collectivism influence linguistic cues of deception. It

analyzed datasets from six countries (U.S., Belgium, India, Russia, Mexico, and Romania) across five languages to assess the universality of deception indicators. The study evaluated linguistic features such as pronoun use, sentiment expression, and contextual details, employing logistic regression and fine-tuned BERT models. Findings revealed that linguistic cues of deception were culturally variable and context-dependent, highlighting that universal approaches to deception detection are insufficient. Results emphasized the need to incorporate cultural knowledge into automated deception detection frameworks for better accuracy and fairness.

William Steingartne et al.,2021[38]:This paper examined the role of cyber deception as a strategic defense mechanism within the context of hybrid warfare and cybersecurity. Researchers developed a novel hybrid threats model and explored how deception-based methods, such as honeypots and honeytokens, were utilized to detect and mitigate advanced cyber threats. The study emphasized the growing sophistication of cyberattacks, including advanced persistent threats (APTs), and highlighted the advantages of deception in shaping attackers' decision-making and deterring breaches. By analyzing the convergence of cyber operations and electronic warfare, the paper demonstrated how cyber deception technologies can manipulate adversaries' activities and enhance defensive capabilities. The findings underscored the importance of deception as a pivotal tool in modern cyber defense strategies.

Tim Brennen et al.,2022[39]:This paper critically examined the practical applicability of verbal cues in deception detection, highlighting the limitations of existing methods. Researchers reviewed techniques like Criterion-Based Content Analysis (CBCA), Verifiability Analysis, and Strategic Use of Evidence (SUE), emphasizing their varying levels of effectiveness in forensic settings. CBCA showed moderate success but faced challenges such as high false alarm rates and overestimated accuracy. Verifiability Analysis demonstrated potential through distinguishing verifiable details but lacked validation in real-world applications. SUE emerged as the most promising human-based approach, effectively leveraging independent evidence to trap deceptive statements in interviews. Despite these advances, the study concluded that automated systems might eventually surpass human-based methods, given their scalability and potential to enhance accuracy in deception detection tasks.

L. Ende et al.,2023[40]:This paper explored deception detection in the context of greenwashing, focusing on how visual product cues like color and price influence consumer classification accuracy in identifying faked bio-fashion products. Researchers found that consumers were more likely to classify products as bio when the color and price aligned with bio-typical expectations, such as green hues and high prices. Furthermore, classification accuracy improved when these cues matched the actual status of the product. Surprisingly, variations in consumers' ecological experience did not significantly affect detection abilities. The study underscored how subtle deceptive strategies exploiting visual and pricing biases can mislead consumers, advocating for stronger regulations to mitigate greenwashing practices in the marketplace.

Abdul Basit Ajmal et al.,2021[24]:This paper investigated deception detection as part of a proactive cybersecurity framework designed for SCADA networks, aiming to counteract unknown and stealthy threats. Researchers integrated threat hunting with cyber deception strategies, including decoy farms, kill chain analysis, and customized honeypots, to engage attackers while gathering Indicators of Compromise (IOCs). The approach utilized simulation tools like Mininet, Ryu controllers, and modified Conpot honeypots to detect adversary activities, analyze malicious traffic, and uncover novel tactics, techniques, and procedures (TTPs). Experimental results demonstrated that the deception-based threat hunting model effectively detected and mitigated threats before they could compromise the network, outperforming traditional reactive security measures in identifying and isolating stealthy attacks.

Ara Mambreyan et al.,2021[41]:This paper analyzed the impact of dataset bias on deception detection in machine learning, focusing on popular datasets like Real-life Trial and Bag-of-Lies. Researchers demonstrated that significant biases, particularly related to sex, allowed classifiers to achieve high metrics by exploiting incidental correlations instead of actual patterns of deception. For example, using sex as a proxy for predicting lies yielded comparable results to state-of-the-art methods. Experiments showed that when these techniques were applied to unbiased datasets, like the Miami University Deception Detection dataset, their performance dropped to chance levels. The findings highlighted the unreliability of current deception detection techniques and stressed the importance of addressing dataset bias to ensure fairness and validity in machine learning applications for lie detection.

## 5. DISCUSSION AND COMPARISON

| # | Author (Year) | Focus Area | Methods Used | Dataset Used | Results | Limitations |
|---|---|---|---|---|---|---|
| 1 | Merylin Monaro (2022) | Facial deception | SVM + OpenFace | Low-stakes holiday interview videos | AI (AUC=0.78) > Human (57%) | Low-stakes dataset only |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | (micro-expressions) | | | | |
| 2 | Budi Gunawan (2022) | Tech use in fake news | VOSviewer, bibliometric analysis | Scopus publication metadata | CNNs/hybrid MLs dominant | Descriptive, not experimental |
| 3 | Delgado (2021) | Text deception in reviews/emails | SVM, RF, NN with POS/BOW | Custom dataset of deceptive texts | POS outperformed BOW, accuracy 78.88% | Feature combinations not always better |
| 4 | Rinaldo Gagiano (2021) | Neural fake news robustness | Grover adversarial testing | Grover-generated news articles | 97% misclassified post-perturbation | Low robustness to adversarial input |
| 5 | Francielle Vargas (2021) | Multilingual fake news (discourse) | RST corpus + new discourse cues | 'Deceiver' corpus (EN & PT) | 2 new RST cues improved accuracy | Limited corpus size |
| 6 | Mu Zhu (2021) | Cyber deception (game theory) | ML + Game-theoretic categorization | Review of multiple systems | Categorized deception strategies | Scenarios vary in effectiveness |
| 7 | Leena Mathur (2021) | High-stakes deception (transfer learning) | Unsupervised Subspace Alignment | High/low stakes deception corpora | SA model outperformed baseline (AUC=75%) | Real-world generalization may vary |
| 8 | Shravika Mittal (2021) | Community structure deception | NEURAL rewiring for community masking | Synthetic and real social networks | NEURAL deceived 6 detection algorithms | Focused on structure, not semantics |
| 9 | Nguyen Van Huynh (2021) | Anti-jamming deception | LSTM + ambient backscatter | Simulated jamming signals | Improved BER under jamming | Limited to physical layer |
| 10 | Hammad-Ud-Din Ahmed (2021) | Facial deception via action coding | FACS + LSTM + OpenFace | Bag-of-Lies, Silesian, Real-life Trial | Better accuracy with OpenFace + LSTM | Dataset variance affects results |
| 11 | Erich C. Walter (2020) | Cyber deception in simulation | RL + CyberBattleSim (honeypots) | CyberBattleSim | Honeypots delayed attackers effectively | Simulated, not real cyberattackers |
| 12 | V. Kozlov (2021) | Radar deception tech | FMCW radar + signal modulation | Radar signal tests | Radar fooled, symmetric waveforms suggested | Specific to radar systems |
| 13 | Aidan O'Gara (2023) | Language model deception in games | Text-based deception game (GPT-3/4) | Hoodwinked game dialogues | GPT-4 influenced votes, deceived players | Game context, ethical issues |
| 14 | Jiawei Liang (2024) | Face forgery backdoor vulnerability | Clean-label trigger poisoning | Face forgery detectors (varied sets) | +16.39% BD-AUC, -12.65% L∞ visibility | Specific to facial forgery detectors |
| 15 | Oana Ignat (2023) | Multilingual review fraud | Random Forest + XLM-RoBERTa | MAIDE-UP hotel reviews | 94.8% detection accuracy | AI reviews differ stylistically |

| | | | | | | |
|---|---|---|---|---|---|---|
| 16 | Francielle Vargas (2022) | RST in online deception | Bag-of-RST + neural embeddings | Multiple online corpora | RST-based cues improved classification | Reliance on RST parsers |
| 17 | huang-cheng chou (2021) | Mandarin deception detection | BLSTM + acoustic + implicature | Daily Deceptive Dialogues corpus | UAR=80.61% | Mandarin-specific; small corpus |
| 18 | Despoina Mouratidis (2021) | Social media fake news | SMOTE + deep multimodal fusion | Tweets from Hong Kong protests | High precision/recall after SMOTE | Bias-correction needed |
| 19 | Katerina Papantoniou (2022) | Cross-cultural textual deception | BERT + cultural linguistic analysis | Texts from 6 countries, 5 languages | Deception cues are culture-sensitive | No universal linguistic markers |
| 20 | William Steingartne (2021) | Cyber deception in warfare | Honeypots + hybrid threat model | Simulated threat models | Improved APT defense via deception | Need real-world test environments |
| 21 | Tim Brennen (2022) | Verbal deception in forensics | CBCA + SUE + VA techniques | Forensic interview protocols | SUE more reliable than CBCA | Human-based methods limited |
| 22 | L. Ende (2023) | Greenwashing (consumer deception) | Consumer experiment (visual cues) | Bio-fashion products (visuals, prices) | Green price & color misled users | Small test pool |
| 23 | Abdul Basit Ajmal (2021) | Cyber deception in SCADA | Mininet + Ryu + deception farms | SCADA honeynet simulations | Detected stealth attacks better | Tool-specific findings |
| 24 | Ara Mambreyan (2021) | Dataset bias in deception ML | Bias sensitivity experiments | Real-life Trial, Bag-of-Lies, Miami | Bias inflated ML performance | Bias risks false confidence |

Research on deception detection has advanced dramatically, showing how machine learning and artificial intelligence are becoming increasingly capable of detecting dishonest behavior in a variety of modalities, such as voice patterns, text analysis, network behavior, facial micro-expressions, and even game-based simulations. Because these automated systems can extract and process high-level, multimodal features like facial action units, acoustic signatures, or rhetorical discourse structures, they often perform better than human evaluators, especially in situations involving cognitive stress or subtle behavioral cues. The detection of linguistic deceit, particularly via the use of characteristics such as part-of-speech tags, rhetorical coherence, or cultural semantics, demonstrates that more straightforward and organized models often perform on par with or better than more intricate deep learning architectures. Nonetheless, a number of drawbacks still exist, including the models' susceptibility to adversarial assaults, in which little changes to the input data may lead to serious misclassifications, and the significant decreases in performance that occur when the models are exposed to biased or cross-domain datasets. Additionally, language and cultural variety cast doubt on the universality of deception indicators, with results highlighting the necessity for culturally adapted models that take sociolinguistic subtleties into account. In addition to highlighting the growing relevance of AI-driven deception detection, emerging fields like cybersecurity deception frameworks, radar misdirection, and simulated social deception in games also emphasize the significance of ethical design, model transparency, and real-world validation. The development of more resilient, generalizable, and interpretable systems that can dynamically adjust to changing forms of deceit in both artificial and human environments is the key to the future of deception detection, despite the fact that overall progress has been significant.

## 6. EXTRACTED STATISTICS

Cyber deception (featured in 7 studies) and fake news detection (featured in 6 studies) dominate the research landscape, reflecting the growing importance of addressing digital deception and misinformation. Multilingual deception, which appears in two studies, showcases efforts to bridge cultural and linguistic nuances in detecting deceptive behavior. Greenwashing is another emerging area of interest, pointing to consumer-related deception practices. Niche areas such as radar deception, high-stakes deception, and neural fake news robustness demonstrate the diversity and depth of topics within deception-related research as seen in figure.

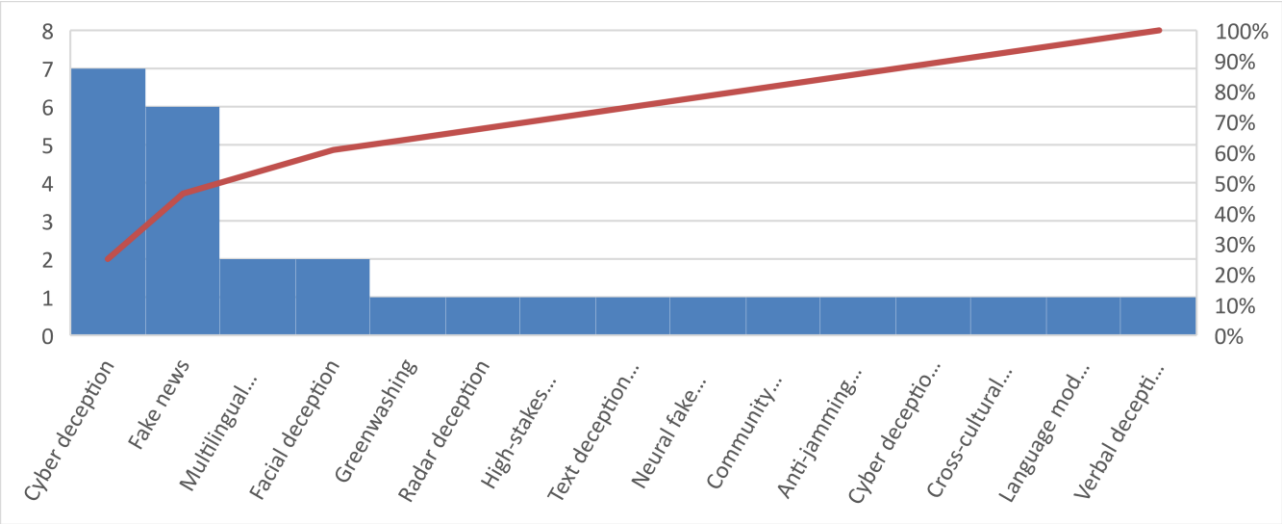


**Figure 2: Statistical representation about the Focus Area.**

The studies employ a variety of methodologies to tackle deception challenges. Machine learning techniques like SVM, Neural Networks, and Random Forests are the most commonly used, underscoring their effectiveness in pattern detection and classification. Unique approaches such as game theory and adversarial testing reveal creative strategies tailored to specific types of deception. Advanced tools like SMOTE and FACS contribute to solving specialized problems, such as overcoming biases in datasets or improving facial action coding analysis. The integration of these diverse methods showcases innovation and adaptability within the field as seen in figure.3

.

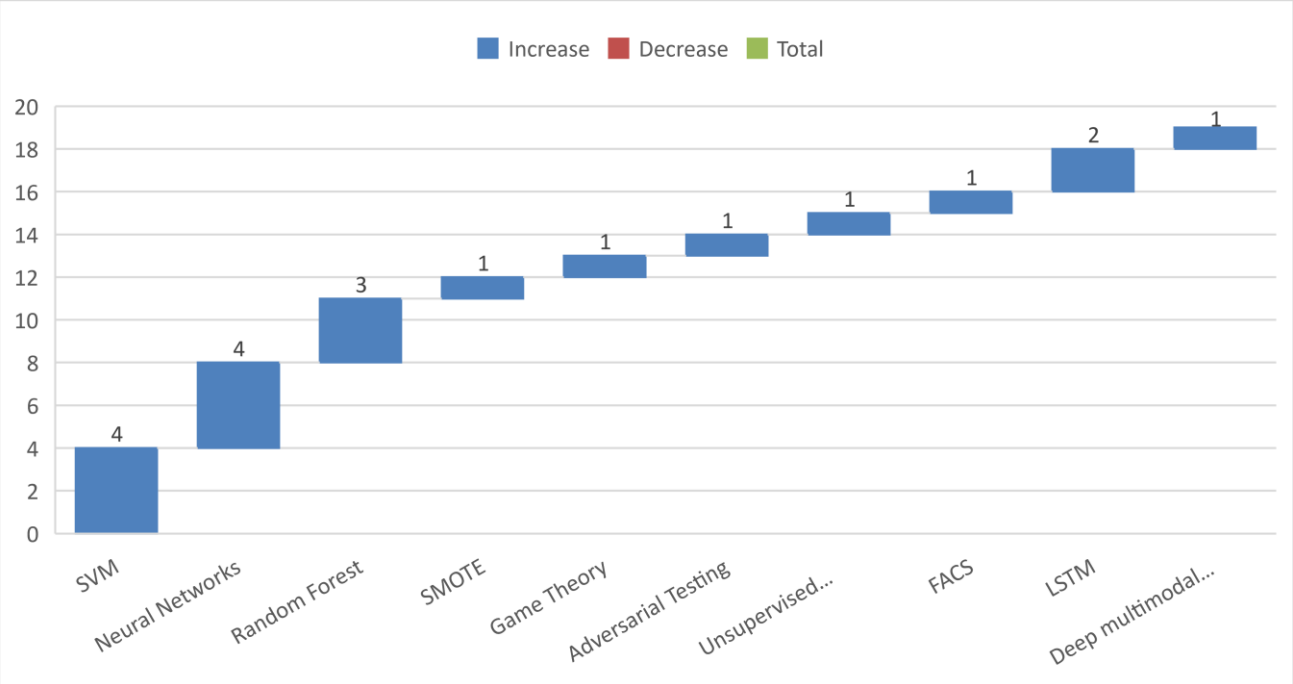


**Figure 3: Statistical representation about the Methods Used**

Text datasets (featured in 8 studies) and social media datasets (featured in 5 studies) are the primary sources for deception detection, reflecting the dominance of online and written communication in deceptive practices. Multilingual datasets play a critical role in addressing cross-cultural nuances, appearing in three studies, though the limited availability of such datasets highlights an area for future improvement. Simulation datasets, such as CyberBattleSim and SCADA honeynet simulations, offer controlled environments for testing deception strategies, emphasizing their utility in cyber-related applications as seen in figure.4

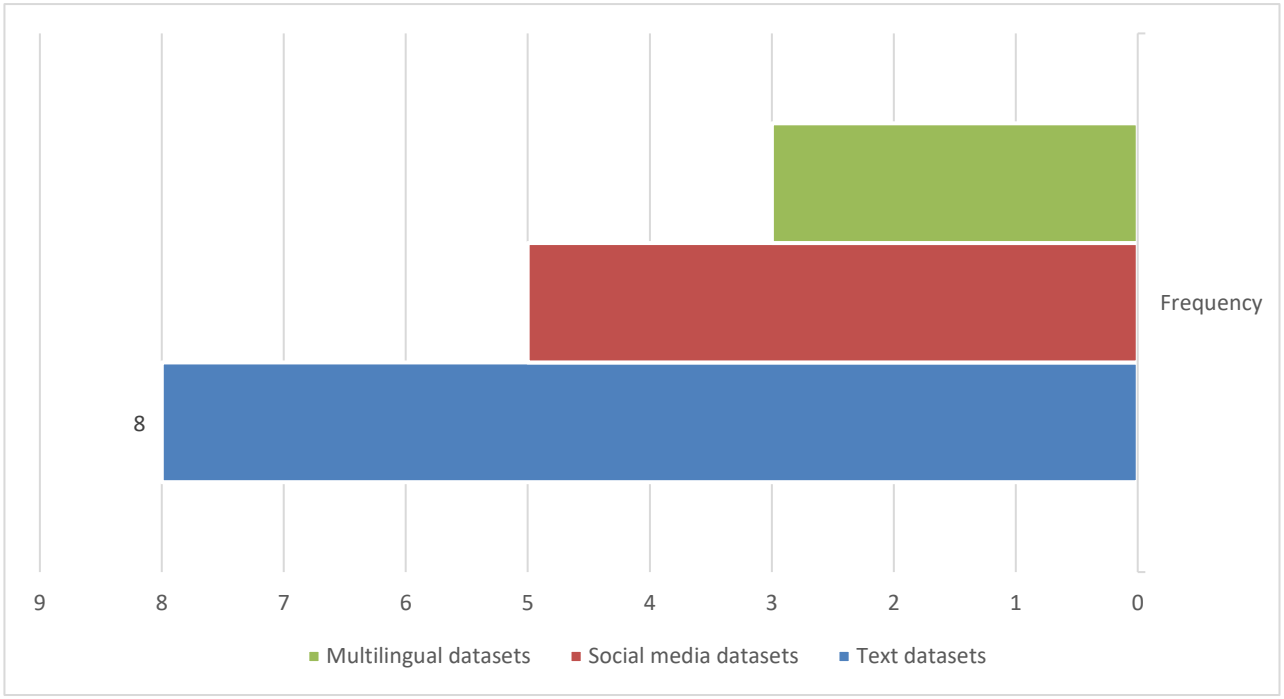


**Figure 4: Statistical representation about the Dataset Used**

Many studies focus on improving detection accuracy, with 10 studies achieving notable advancements in accuracy levels through innovative approaches. Insights into adversarial robustness and behavioral deception categorization add valuable perspectives, particularly for applications in cybersecurity and misinformation mitigation. Advanced techniques like clean-label trigger poisoning and neural embeddings also demonstrate significant progress, paving the way for future research in tackling more complex deception scenarios as seen in figure.5

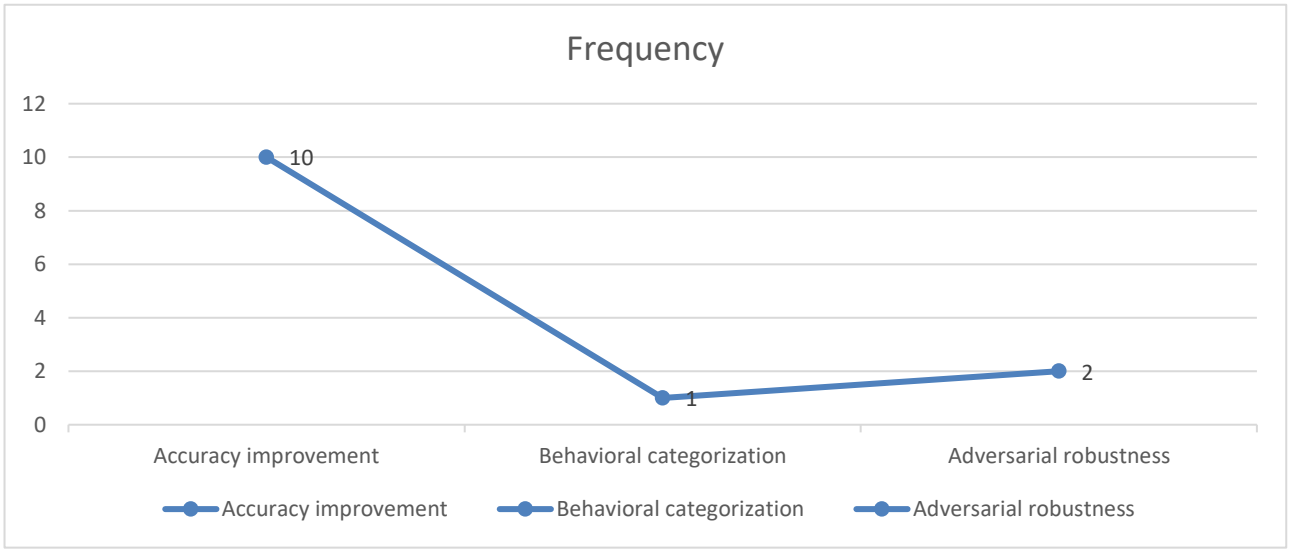


**Figure 5: Statistical representation about the Results**

Recurring challenges across the studies include small dataset sizes and limited generalizability to real-world scenarios, which are mentioned in 10 studies. Bias concerns, particularly those related to dataset-specific biases, pose risks of inflated model performance and unreliable results. The limitations highlight the need for larger, more diverse datasets and for methods that account for cultural, contextual, and adversarial nuances to improve real-world applicability as seen in figure 6.

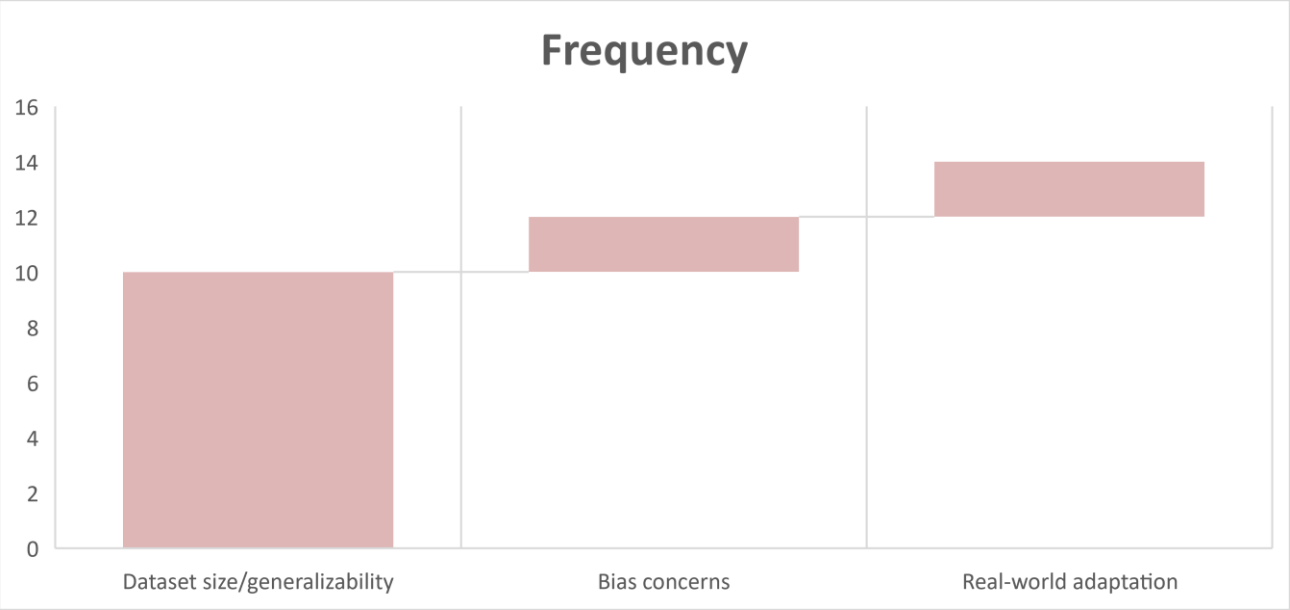


**Figure 6: Statistical representation about the Limitations**

## 7. RECOMMENDATIONS

- Develop Culturally Adaptive Models: Deception cues vary significantly across languages and cultures; therefore, future systems should incorporate cultural and linguistic adaptability. This can be achieved by training models on diverse, multilingual datasets and embedding culturally nuanced features to improve cross-cultural reliability.
- Enhance Robustness Against Adversarial Attacks: Many current models are vulnerable to small input perturbations or backdoor manipulations. Future research should focus on building adversarial robust models, integrating detection mechanisms for tampered data, and employing continual learning to adapt to new threats.
- Mitigate Dataset Bias and Ensure Fairness: Bias in training data, such as those based on gender or demographic attributes, can skew model predictions and lead to ethical concerns. It is critical to audit datasets regularly, implement fairness-aware algorithms, and promote the use of balanced and representative data sources.
- Integrate Multimodal and Contextual Features: Combining visual, acoustic, textual, and contextual cues can significantly improve detection accuracy. Research should prioritize the integration of multimodal data, including discourse structures, physiological signals, and behavioral patterns, for more comprehensive deception assessment.
- Advance Real-World Validation and Benchmarking: Many models are evaluated in controlled or simulated environments. Future studies should conduct real-world trials and longitudinal assessments to ensure practical applicability and scalability, especially in forensic, cybersecurity, and social media contexts.
- Promote Ethical and Transparent AI Design: Given the sensitive nature of deception detection, transparency, interpretability, and ethical considerations should be embedded in system design. Models should provide explainable outcomes, protect user privacy, and avoid misuse in surveillance or coercive settings.
- Encourage Interdisciplinary Collaboration: Bridging expertise from psychology, linguistics, computer science, law, and security can lead to more holistic and ethically grounded deception detection systems. Collaboration can enhance model reliability, data annotation quality, and real-world implementation strategies.

## 8. CONCLUSION

Integrating methods from artificial intelligence, machine learning, psychology, languages, and cybersecurity,

deception detection has become a fast-developing interdisciplinary area. Particularly in identifying subtle and sophisticated cues of dishonesty such facial micro-expressions, vocal stress, rhetorical inconsistencies, and behavioral anomalies in text or network activities, the studies examined show how AI-driven models have advanced beyond conventional human-based assessments. From forensic interviews and social media postings to false news stories, radar deception, and cybersecurity simulations, machine learning algorithms-including Support Vector Machines, LSTM networks, convolutional models, and hybrid architectures-have demonstrated good performance across a range. Particularly when combined with multimodal data and contextual analysis, these developments show that automated systems may reach better accuracy, scalability, and consistency than human assessors. Still, the trip is not quite over. Important issues still exist including the susceptibility of these models to adversarial assaults, the existence of dataset bias, and the lack of generalizability across many cultural, linguistic, and situational settings. The discrepancy of dishonesty signals across languages and civilizations begs for models that are not only accurate but also contextually aware and culturally sensitive. Furthermore, addressed by open model design and fair data governance are ethical issues like user privacy, abuse of surveillance technology, and algorithmic bias. Moreover, many present systems are still evaluated in controlled contexts with insufficient real-world validation, which begs problems about their dependability and resilience in pragmatic uses. Explainable artificial intelligence in fraud detection is also increasingly needed to guarantee stakeholders including law enforcement, attorneys, and consumers=can trust and properly understand model results. Deception detection tools have to change to keep pace as the extent of dishonesty grows from textual misrepresentation and AI-generated material to behavioral manipulation in games and hostile cyberattacks. In essence, even if deception detection is becoming more automated, future systems have to be more strong, open, culturally conscious, and morally sound. Ensuring these technologies improve society will depend on multidisciplinary cooperation, practical testing, and a focus on justice and responsibility. AI-powered deception detection may greatly improve truth verification and decision-making in a world where disinformation and manipulation are more sophisticated and ubiquitous by means of ongoing innovation and prudent deployment.

**REFERENCES**


1. A. Goel, A. K. Goel, and A. Kumar, "The role of artificial neural network and machine learning in utilizing spatial information," *Spat. Inf. Res.*, vol. 31, no. 3, pp. 275–285, 2023, doi: 10.1007/s41324-022-00494-x.
2. R. Avdal and H. Maseeh, "Advancing Cybersecurity through Machine Learning : Bridging Gaps , Overcoming Challenges , and Enhancing Protection," vol. 18, no. 2, pp. 206–217, 2025.
3. S. R. M. Zeebaree, "A Review of Blockchain Technology In E-business : Trust , Transparency , and Security in Digital Marketing through Decentralized Solutions," vol. 18, no. 3, pp. 411–433, 2025.
4. H. S. Mavikumbure, V. Cobilean, C. S. Wickramasinghe, D. Drake, and M. Manic, "Generative AI in Cyber Security of Cyber Physical Systems: Benefits and Threats," *Int. Conf. Hum. Syst. Interact. HSI*, no. June, 2024, doi: 10.1109/HSI61632.2024.10613562.
5. R. Avdal and S. R. M. Zeebaree, "Artificial Intelligence in E-commerce and Digital Marketing : A Systematic Review of Opportunities , Challenges , and Ethical Implications," vol. 18, no. 3, pp. 395–410, 2025.
6. F. R. Tato and H. M. Yasin, "Detecting Diabetic Retinopathy Using Machine Learning Algorithms : A Review," vol. 18, no. 2, pp. 118–131, 2025.
7. A. A. C. Delgado, W. B. Glisson, N. Shashidhar, J. T. McDonald, G. Grispos, and R. Benton, "Detecting deception using machine learning," *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, vol. 2020-Janua, pp. 7122–7131, 2021.
8. K. J. Dastan Hussen Maulud,Subhi R. M. Zeebaree, "A State of Art for Semantic Analysis of Natural Language Processing," *Qubahan Acad. J.*, pp. 21–28, 2023, doi: https://doi.org/10.48161/qaj.v1n2a44.
9. L. Mathur and M. J. Matarić, "Unsupervised audio-visual subspace alignment for high-stakes deception detection," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2021-June, pp. 2255–2259, 2021, doi: 10.1109/ICASSP39728.2021.9413550.
10. F. R. Tato and S. R. M. Zeebaree, "East Journal of Applied Science The Rise of Influence Marketing in E-Commerce : A Review of Effectiveness and Best Practices," vol. 1, no. 1, pp. 18–34, 2025.
11. F. Vargas, J. D. Alessandro, Z. Rabinovich, and T. A. S. Pardo, "Rhetorical Structure Approach for Online Deception Detection : A Survey," no. June, pp. 5906–5915, 2022.
12. R. A. Saleh and S. R. M. Zeebaree, "Transforming Enterprise Systems with Cloud , AI , and Digital Marketing," vol. 3, 2025, doi: 10.59543/ijmscs.v3i.13883.
13. E. C. Walter, K. J. Ferguson-walter, and A. D. Ridley, "Incorporating Deception into CyberBattleSim for Autonomous Defense," 2020.
14. W. Merza and I. Mahmood, "Ant Colony Optimization ( ACO ) for Traveling Salesman

Problem : A Review," vol. 18, no. 2, pp. 20–45, 2025.
15. F. R. Tato and I. M. Ibrahim, "Bio-Inspired Algorithms in Healthcare," vol. 07, no. 02, pp. 233–239, 2024.
16. R. Gagiano, M. M. H. Kim, X. Zhang, and J. Biggs, "Robustness Analysis of Grover for Machine-Generated News Detection," *ALTA 2021 - Proc. 19th Work. Australas. Lang. Technol. Assoc.*, pp. 119–127, 2021.
17. K. Papantoniou, P. Papadakos, T. Patkos, and G. Flouris, "Deception detection in text and its relation to the cultural dimension of individualism / collectivism," pp. 545–606, 2022, doi: 10.1017/S1351324921000152.
18. R. Avdal and I. M. I. Zebari, "Enhancing Network Performance : A Comprehensive Analysis of Hybrid Routing Algorithms," vol. 18, no. 3, pp. 1–16, 2025.
19. M. Monaro, S. Maldera, C. Scarpazza, G. Sartori, and N. Navarin, "Detecting deception through facial expressions in a dataset of videotaped interviews: A comparison between human judges and machine learning models," *Comput. Human Behav.*, vol. 127, no. October 2021, 2022, doi: 10.1016/j.chb.2021.107063.
20. A. F. Jahwar and S. R. M. Zeebaree, "A State of the Art Survey of Machine Learning Algorithms for IoT Security," *Asian J. Res. Comput. Sci.*, vol. 9, no. 4, pp. 12–34, 2021, doi: 10.9734/ajrcos/2021/v9i430226.
21. D. H. Maulud *et al.*, "Review on Natural Language Processing Based on Different Techniques," vol. 10, no. 1, pp. 1–17, 2021, doi: 10.9734/AJRCOS/2021/v10i130231.
22. R. A. Saleh and H. M. Yasin, "Comparative Analysis of AI and Machine Learning Applications in Modern Database Systems," vol. 10, no. 03, pp. 4112–4123, 2025, doi: 10.47191/etj/v10i03.21.
23. N. Rane, M. Paramesha, S. Choudhary, and J. Rane, "Machine Learning and Deep Learning for Big Data Analytics: a Review of Methods and Applications," *SSRN Electron. J.*, no. June, pp. 172–197, 2024, doi: 10.2139/ssrn.4835655.
24. A. B. Ajmal, M. Alam, A. A. Khaliq, Z. Qadir, and M. A. P. Mahmud, "Last Line of Defense : Reliability Through Inducing Cyber Threat Hunting With Deception in SCADA Networks," *IEEE Access*, vol. 9, pp. 126789–126800, 2021, doi: 10.1109/ACCESS.2021.3111420.
25. J. Liang, S. Liang, A. Liu, X. Jia, J. Kuang, and X. Cao, "P OISONED F ORGERY F ACE : T OWARDS B ACKDOOR A T -," no. 2017, pp. 1–16, 2024.
26. C. Al-Atroshi and S. R. M. Z. Zeebaree, "Distributed Architectures for Big Data Analytics in Cloud Computing: A Review of Data-Intensive Computing Paradigm," *Indones. J. Comput. Sci.*, vol. 13, no. 2, pp. 2389–2406, 2024, doi: 10.33022/ijcs.v13i2.3812.
27. B. Gunawan, B. M. Ratmono, A. G. Abdullah, N. Sadida, and H. Kaprisma, "Research Mapping in the Use of Technology for Fake News Detection: Bibliometric Analysis from 2011 to 2021," *Indones. J. Sci. Technol.*, vol. 7, no. 3, pp. 471–496, 2022, doi: 10.17509/ijost.v7i3.51449.
28. F. Vargas, F. Benevenuto, and T. A. S. Pardo, "Toward Discourse-Aware Models for Multilingual Fake News Detection," *Int. Conf. Recent Adv. Nat. Lang. Process. RANLP*, vol. 2021-Septe, pp. 210–218, 2021, doi: 10.26615/issn.2603-2821.2021_029.
29. M. Zhu, A. H. Anwar, Z. Wan, J.-H. Cho, C. Kamhoua, and M. P. Singh, "Game-Theoretic and Machine Learning-based Approaches for Defensive Deception: A Survey," pp. 1–37, 2021, [Online]. Available: http://arxiv.org/abs/2101.10121
30. S. Mittal, D. Sengupta, and T. Chakraborty, "Hide and Seek : Outwitting Community," 2021.
31. N. Van Huynh, D. N. Nguyen, D. T. Hoang, and T. X. Vu, "Defeating Super-Reactive Jammers With Deception Strategy : Modeling , Signal Detection , and Performance Analysis," 2021.
32. H. Ahmed and F. Zhang, "Deception Detection in Videos using the Facial Action Coding System Deception Detection in Videos using the Facial Action Coding System," pp. 0–2, 2021.
33. V. Kozlov, D. Vovchuk, and P. Ginzburg, "Radar Range Deception with Time-Modulated Scatterers," 2021.
34. A. O. Gara, "Hoodwinked : Deception and Cooperation in a Text-Based Game for Language Models," pp. 1–9, 2023.
35. G. H. Reviews, "MAiDE-up: Multilingual Deception Detection of GPT-generated Hotel Reviews," 2023.
36. I. Processing, A. V. Healthcare, H. Chou, and C. Lee, "Automatic Deception Detection using Multiple Speech and Language Communicative," vol. 10, 2021, doi: 10.1017/ATSIP.2021.6.
37. D. Mouratidis and M. N. Nikiforos, "Deep Learning for Fake News Detection in a Pairwise Textual Input Schema," 2021.
38. W. Steingartner and D. Galinec, "Cyber Threats and Cyber Deception in Hybrid Warfare," vol. 18, no. 3, 2021.
39. T. Brennen, "The Science of Lie Detection by Verbal Cues : What Are the Prospects for Its Practical Applicability ?," vol. 13, no. April, 2022, doi: 10.3389/fpsyg.2022.835285.

40. L. E. M. A. R. L. Göritz, *Detecting Greenwashing ! The Influence of Product Colour and Product Price on Consumers ' Detection Accuracy of Faked Bio - fashion*, vol. 46, no. 2. Springer US, 2023. doi: 10.1007/s10603-023-09537-8.
41. A. Mambreyan and E. Punskaya, "Dataset Bias in Deception Detection," 2021.