information

Review

Cybercrime Intention Recognition: A Systematic Literature Review

Yidnekachew Worku Kassa [1,*], Joshua Isaac James [2] and Elefelious Getachew Belay [1]

[1] School of Information Technology and Engineering, Addis Ababa Institute of Technology (AAiT), Addis Ababa University, Addis Ababa P.O. Box 1176, Ethiopia; elefelious.getachew@aau.edu.et
[2] DFIR Science LLC, Bangkok 10110, Thailand; joshua@dfirscience.org
* Correspondence: yidnekachew.worku@aau.edu.et or yidnekacheworku@gmail.com

Abstract: In this systematic literature review, we delve into the realm of intention recognition within the context of digital forensics and cybercrime. The rise of cybercrime has become a major concern for individuals, organizations, and governments worldwide. Digital forensics is a field that deals with the investigation and analysis of digital evidence in order to identify, preserve, and analyze information that can be used as evidence in a court of law. Intention recognition is a subfield of artificial intelligence that deals with the identification of agents' intentions based on their actions and change of states. In the context of cybercrime, intention recognition can be used to identify the intentions of cybercriminals and even to predict their future actions. Employing a PRISMA systematic review approach, we curated research articles from reputable journals and categorized them into three distinct modeling approaches: logic-based, classical machine learning-based, and deep learning-based. Notably, intention recognition has transcended its historical confinement to network security, now addressing critical challenges across various subdomains, including social engineering attacks, artificial intelligence black box vulnerabilities, and physical security. While deep learning emerges as the dominant paradigm, its inherent lack of transparency poses a challenge in the digital forensics landscape. However, it is imperative that models developed for digital forensics possess intrinsic attributes of explainability and logical coherence, thereby fostering judicial confidence, mitigating biases, and upholding accountability for their determinations. To this end, we advocate for hybrid solutions that blend explainability, reasonableness, efficiency, and accuracy. Furthermore, we propose the creation of a taxonomy to precisely define intention recognition, paving the way for future advancements in this pivotal field.

Keywords: digital forensics; digital investigation; intention recognition; goal recognition; plan recognition; cyberattack; cybercrime; model

check for updates

Citation: Kassa, Y.W.; James, J.I.; Belay, E.G. Cybercrime Intention Recognition: A Systematic Literature Review. *Information* **2024**, *15*, 263. https://doi.org/10.3390/info15050263

Academic Editors: Cihan Varol, Amar Rasheed, Umit Karabiyik, Narasimha Shashidhar and Rui Zhang

Received: 18 March 2024
Revised: 22 April 2024
Accepted: 23 April 2024
Published: 5 May 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Cyberspace, which incorporates hardware infrastructures and devices, the connectivity among those devices, the software that runs on those devices, and the information maintained within those infrastructures, is growing exponentially in all aspects [1]. This growth is redefining the world as traditional world activities such as communication, industrialization, social interaction, military, education, transportation and more are being enabled by cybertechnologies. This positions cyberspace as a crucial new element and central focus, to the point where it is regarded as the fifth domain of security and warfare. [2]. Currently, the widespread deployment of IoT devices has enhanced network connectivity across the globe, leading to the development of smart villages, smart cities, smart transportation systems, and smart homes. The software industry is also growing at an unprecedented rate, with artificial intelligence (AI) replacing human intelligence in most expert systems. Social media platforms are also experiencing tremendous growth, with almost everyone connected to one or more platforms [1]. The world is enjoying the benefits of cyberspace, and globalization is more prevalent than ever before.

However, the growth of cyberspace has also redefined security perspectives for individuals, organizations, and countries alike [2]. The attack surface has widened so much that cyberattacks are growing from year to year, becoming a major risk to the world. Moreover, the Global Risk Report has labeled cyberattacks as the sixth most high-impact risk [3]. As a result, digital crime investigation has become a major focus for countries and organizations, as cases that are facilitated or fully committed by computers increase exponentially.

1.1. Digital Forensics

Digital forensics (DF) applies computer science for the investigation of digital crime by following proper investigation procedures such as chain of custody, validation, search authority, and reporting [4]. The Digital Forensics Investigation (DFI) process passes through four major stages as defined by the National Institute of Standards and Technology (NIST). The first stage is the collection stage, which involves identifying and collecting digital pieces of evidence related to the crime. The second stage is examination, which involves filtering the relevant information from the collected evidence. The third stage is analysis, which involves analyzing the evidence and connecting the dots to reconstruct the crime scene. The final stage is reporting, which involves presenting the case to the court [4]. International Standards Organization (ISO) has its own version of the DF investigation process [5].

Many technical and governance challenges make winning a crime case very difficult. There are awareness challenges, from the victim to the police officers, the investigators, the judges, and within the overall society such that cases are lost because of the necessary precautions.On the other hand, in relation to the volume of data collected for a crime case being too much, currently, there is a huge backlog almost in every DF laboratory that continues to grow from time to time. Raghavan [6] categorizes these challenges into five groups.

- First, the complexity problem which arises from the huge and heterogeneous data volume problem that requires complex data mining solutions;
- Second, the diversity problem that arises from the solutions for different evidence sources being different and this created too many tools and techniques;
- Third, the consistency and correlation problem which is the result of the diversity of the evidence sources, and their solutions are expected to correlate, but there are challenges in correlating the solutions;
- Fourth, the volume problem that arises from the dynamic growth of cyberspace results in a dramatic increase in the number of crime cases as well as the number and size of storage devices;
- Fifth, the unified timeline problem that arises from the different timezone and timestamp interpretations makes investigation difficult.

Quick et al. [7] summarize research on the challenges of DF in relation to the big volume data and the solutions proposed. The large volume of data is a result of the increase in cases with DF involved, the increase in the number of devices seized per case, and the big size of each individual device. These big data create a huge backlog such that even the big DF laboratories are challenged. They show how the data volume is continuously increasing with a high slope linear scale by collecting data from 1997 to 2014. The backlog in turn creates problems, including suspect suicide, the suspect being denied access to family, the suspect being denied work, and access to personal data. They also review the solutions proposed by researchers for the backlog problem including data mining which enables the extraction of useful information from big data, data reduction and subset which helps to reduce the data to be analyzed, DF triage which prioritizes the evidence according to their relevance, and using intelligence analysis such as profiling to filter data. Other techniques such as distributed and parallel processing, visualization, DF as a service (DFaaS), and the use of AI techniques are also discussed.

1.2. Intention Recognition

The huge data generated by humans as well as devices is a huge problem for analysis by human beings and consequently, it is creating a delay not only in DF but also in every sector that utilizes digital technologies. The paramount importance of owning machines that are capable of reasoning through given data becomes clearer than ever. Data mining is the process of discovering interesting patterns and knowledge from large amounts of data [8]. However, it is very crucial to have a mechanism to add domain knowledge while mining data in order to obtain meaningful extraction [9].

Intention recognition (IR) is a process by which an agent (the observer/recognizing agent) becomes aware of the intentions of others (the observed/intending agent). In simple terms, the process of IR can be explained as follows: having an intention, the observed agent executes the action or sequences of actions to achieve their intent. Those actions will impact (change states) the environment. Based on those actions, or the state changes, an observer agent can predict the intent of the observed agent. Heinze [10] modeled IR in three levels of intentional behavior of the intending agent and IR of the observer agent: intentional level, activity level, and state level. However, in reality, the process becomes complicated, as there are many other factors: there may be multiple cooperating observed agents, multiple intentions, multiple plans to achieve the intention, multiple hypotheses, different observability levels, different contexts, and many more.

The conceptualization of IR was subject to misinterpretation within the research community until the work of Van-Horenbeke et al. [11] delineated the tripartite structure of the recognition process: activity, intent, and plan. Activity/behavior recognition is characterized as the foundational processing of input streams, wherein patterns are discerned, correlated, and subsequently labeled. Intent/goal recognition, often predicated on the outputs of activity recognition, delves deeper, processing these preliminary findings to extrapolate the agent's underlying objectives. Plan recognition represents the most holistic approach, encapsulating a broader analysis that not only identifies individual intents but also elucidates the interconnections between them, offering an in-depth understanding of the agent's strategies.

IR methodologies are instrumental across a multitude of disciplines, necessitating a nuanced understanding of fundamental constructs such as intent, agents, the environment, and the recognition processes. These constructs must be contextualized to align with the specific domain's characteristics. In the realm of cybercrime, 'intent' refers to the objectives pursued by attackers as delineated by Chen et al. [12]. The 'agents' encompass two distinct roles: cybercriminals, who are the actors with malicious intent, and the juxtaposed observers, comprising investigators and cybersecurity professionals. The 'environment' is represented by the expansive digital landscape of cyberspace, while the 'states' are manifested through the data footprint left by cybercriminals' transactions and activities during their illicit endeavors. The 'recognition process' is represented by models engineered to infer intent from the observable activities and resultant state alterations.

In the multifaceted landscape of cyberoperations, 'intent' manifests across deliberate malevolent attacks to inadvertent actions spurred by negligence or lack of awareness. The 'agents' span a diverse array from malefactors with harmful objectives to cybersecurity experts, as well as autonomous malicious software and even unwitting employees. The 'environment' (cyberspace) comprises technological infrastructures, systems, platforms, and devices—all interwoven to form the operational theater. Moreover, the scholarly pursuit to comprehend these dynamics has given rise to numerous models. This paper conducts a review of selected models that satisfy the established filtering criteria, thereby contributing to the discourse on IR within the DF and cybersecurity domains.

Identifying criminals' intent plays a crucial role in defining and understanding the context of the crime, and it can serve as a clue to investigate the crime from different perspectives [13–15]. It leads the investigation in the right direction by helping in discovering interesting patterns and knowledge: data mining, which may clarify which exhibit is more relevant and who else is involved in the crime. It also has paramount importance

in reducing and subsetting the data generated for the case by applying different filters to facilitate the effective extraction and mining of court-admissible evidence. Proficient evidence gathering is imperative, particularly for active offenses, to mitigate subsequent harm through the application of generated intelligence. Moreover, IR is important to the triage process, assigning precedence to exhibits for in-depth examination. Broadly, IR addresses the voluminous data challenges inherent in DF by integrating it in the four solution types as delineated by [7].

1.3. IR in DF Model Categories

Different researchers have employed different models for IR in DF and related domains. This paper provides a systematic review of these works classified into three major categories—logic-based, classical machine learning, and deep learning-based approaches—as discussed in [11]. Logic-based approaches use logic-based formalisms to represent and reason about the actions, plans, and goals of the observed agent [12,16–20]. They usually rely on predefined domain knowledge and rules to infer the most likely explanation for the observed behavior. Statistical methods and machine learning techniques are employed by classic machine learning approaches to learn patterns and models from data [13,21–24]. These models can be used to recognize the actions and goals of the intending agent. Although these approaches do not require much domain knowledge or human intervention, they need a large amount of labeled data to train the models. Deep learning approaches are the current state of the art in the AI industry [14,25–34]. They utilize deep neural networks to learn high-level features and representations from data, which can be used to recognize intents. However, these approaches lack explainability, as they are considered a black box approach.

1.4. Paper Organization

This paper aims to systematically review the current status of IR related to the DF and cybercrime domain. The remainder of this paper is structured as follows: Section 2 provides an overview of previous reviews as related work. Section 3 outlines the systematic literature review approach adopted in this study. In Section 4, we analyze each study according to its methodological approach to IR, and the challenges are discussed in Section 5. Section 6 presents the findings and trends observed during the analysis. Finally, Section 7 concludes the work and suggests possible areas for future research.

2. Related Works

To the best of our knowledge, there is no systematic review study that has focused on the contribution of IR related to DF or cybercrime. There are a few related studies, and this section discusses them. Table 1 summarizes their contributions and limitations.

In 2017, Ahmed et al. [35] conducted a review of research papers related to the different approaches to recognizing attack intentions. The authors categorized the approaches into four main categories, namely causal network, path analysis, graphical attacks, and dynamic Bayesian network. They discussed each approach in detail and pointed out the advantages and the limitations of the approaches. They also concluded that the causal network approach is more efficient in detecting attacks with similar intentions. However, their review included a small set of studies and they do not discuss how they selected the papers. Additionally, the review's scope is somewhat dated, as it includes papers published prior to 2017.

In 2021, Van-Horenbeke et al. [11] conducted a review of the problem of recognizing human actions, plans, and goals. The paper provides a general view of the problem, both from the human perspective and from the computational perspective, and proposes a classification of the main types of approaches that have been proposed to address it (logic based, classical machine learning, deep learning, and brain inspired), together with a description and comparison of the classes. They included papers from multiple disciplines and application areas. However, since their review is a general review of papers up to 2020

which tries to include all application areas, the attention given to DF or cybercrime is much smaller. It is not known how they selected the papers they reviewed, and the number of DF or cybersecurity-related papers is very few.

This study is meant to be a systematic review that focuses on IR related to DF and cybercrime. The study includes papers from 2018 to 2023. We adopted the approach categorization proposed by Van-Horenbeke et al. [11]; however, we eliminated the brain-inspired category, as we did not find a single paper that could be categorized under this category.

Table 1. Summary of research reviews related to IR and DF.

Article	Contributions	Limitations
Ahmed et al., 2017 [35], Attack Intention Recognition: A Review	• They reviewed some papers related to attack intention recognition. They classified the approaches into four categories (causal networks, path analysis, graphical attacks, and dynamic Bayesian network) and discussed the papers under these approaches.	• It is not known how the papers were selected, as the inclusion and exclusion criteria are not explained. • Many related studies are not included. • The study is a bit old, as it includes papers before 2017.
Van-Horenbeke et al., 2021 [11], Activity, Plan, and Goal Recognition: A Review	• It is a comprehensive study in the sense that it includes papers from all disciplines that utilize activity, goal, and plan recognition. • They categorized the papers into four according to the higher-level approaches (logic based, classical machine learning based, deep learning based, and brain inspired).	• The study inclusion and exclusion criteria are not documented, and it is not known how the reviewed papers were selected. • The study includes papers up to 2020 and is a bit older.

3. Method

There are many approaches to performing systematic literature reviews [36–38]. This research followed the six-phase systematic review approach prepared by Jesson et al. [36], which is adapted from PRISMA. We also employed the latest PRISMA flow diagram for searching and filtering approach [39]. The phases can be summarized as, first, planning the systematic review by preparing the protocol, the questions, the keywords, the criteria, and data extraction sheet. Second is exhaustive search with the keywords, tuning the keywords when necessary, screening using the title and the abstract, and documenting the results. Third is quality assessment by reading the paper and deciding whether to include it or not, documenting the reason for exclusion, and maintaining the result. Fourth is data extraction, extracting the relevant information by using the data extraction sheet. Fifth is synthesizing the data from each article. Six is writing up a balanced, impartial, and comprehensive report.

1. The review is focused on the solutions or models provided to solve the IR problem in DF and cybercrime. The review includes research from 2018 to 2023 which are written in English or have English versions. The data extraction sheet is prepared and attached in Appendix A. The search keywords include the following:
   (a) Search keyword “Intention recognition”:
      - Search string for “intention” is “intent” OR “goal” OR “plan” OR “activity” OR “pattern”;
      - Search string for “recognition” is “recognition” OR “detection”;
      - The above two word categories will be combined with AND yielding strings such as “plan” AND “recognition”. This will be part of the complete search string.

(b) Search keyword "Digital forensics" related words are "attack" OR "cyber-attack" OR "cybercrime" OR "digital forensics investigation" OR "digital investigation".
(c) Search keyword "model": related words are "framework" OR "tool" OR "Algorithm".
(d) Search keywords 1, 2, and 3 are combined with AND to form the complete search string.

2. Exhaustive searches are performed with the known search engines. Google Scholar, IEEE Xplore, Association for Computing Machinery (ACM) Digital Library, ScienceDirect, MDPI, and Springer are searched with the keyword combinations. We include journal articles and conference proceedings from reputable journals with a high impact factor. Higher-level filtering is performed using the titles and to some extent by using the abstracts, and 80 papers are selected.
3. Further filtering is performed by reading the abstracts and to some extent by reading/skimming the full paper, and 22 papers are selected for detailed systematic review. These papers are focused on providing solutions and contributing to the IR problem in DF. They provide different solutions to enhance the efficiency and effectiveness of DFI. Figure 1 shows the process followed in searching and filtering the reviewed articles.
4. Data are extracted using the data extraction sheet. The data extraction includes the following:
   - Subdomains as depicted by Al-Dhaqm [40] (Network Forensics, Computer, Database, Small devices, Memory, Mobile, Multimedia, Software) and Social Media forensics as explained in [41];
   - Modeling approaches as categorized in [11] (such as Logic-based, Classical Machine Learning, Deep Learning);
   - Specific algorithms adopted, as extracted from the papers reviewed (such as similarity based, attack graph-based, signal gaming model, attack tree, hidden Markov, fuzzy min-max, transfer learning, LSTM, CNN, and YOLO);
   - Content type (text, image, audio, video);
   - Intention level as defined in [11] (activity, intent/goal, and plan).
5. The synthesis is discussed in each modeling approach, and a combined synthesis together with the findings is documented in the discussion section.

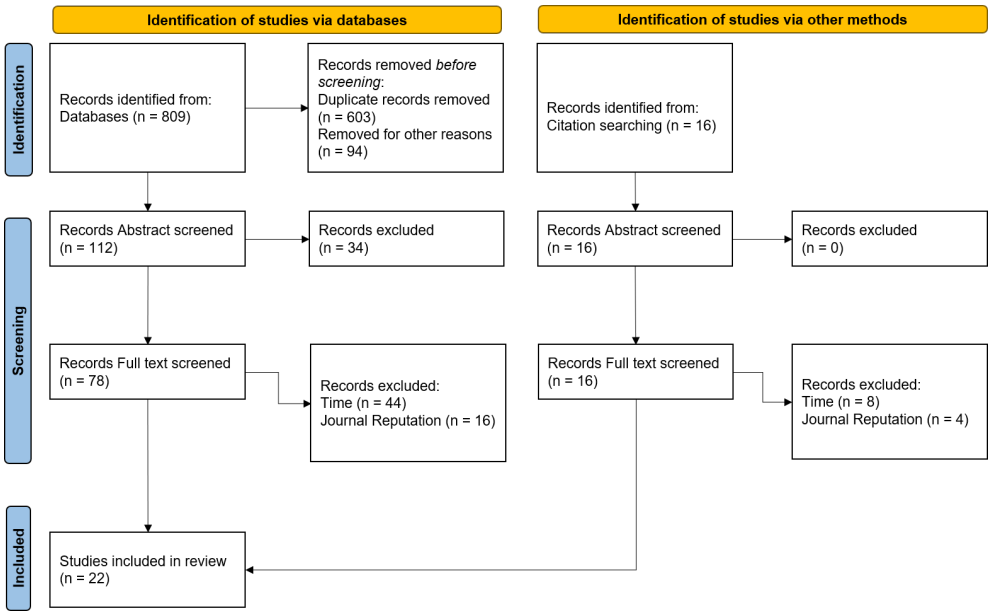

Figure 1. PRISMA flow diagram for searching papers from different databases and filtering papers for the review.

Analysis

The data collected comprise a modest volume of qualitative data, which inherently simplifies the analytical process, obviating the need for advanced analytical methodologies. Nevertheless, the different data retrieved from the respective publications have been organized and analyzed utilizing the functionalities of Microsoft Excel 2021.

4. Review of IR in DF and Cybercrime

DF encompasses a range of activities designed to address cybersecurity incidents, including cyberattacks and computer-facilitated offenses such as cyberbullying, as well as traditional crimes with a digital component, like the theft of mobile devices. The expansion of technology has led to a convergence of conventional criminal and cybercrime investigations. In the context of cyberattacks, DF is intricately linked with incident response and cybersecurity measures. For instance, in the aftermath of an attack, the incident response team is tasked with the reactive measures of recovery, the DF team undertakes the legal investigative process, and the cybersecurity team focuses on system hardening to prevent future breaches. Recognizing this interdependency, NIST has developed a guideline that harmonizes the DF process with incident response protocols [4]. Both domains necessitate the careful collection, handling, and analysis of digital data or evidence, employing a shared arsenal of tools and methodologies. However, while incident response prioritizes the immediate containment of threats and the restoration of systems to safeguard against ongoing attacks, DF delves into the legal aspects of the investigation, aiming to elucidate the factual narrative of the incident.

In this paper, we reviewed diverse criminal activities that have incorporated IR methodologies. While a substantial portion of the literature pertains to cyberincidents [16,18,20,31], there exist significant studies related to computer-facilitated offenses such as harassment [14,21,23,32,34], as well as crimes involving computers such as shoplifting [25,28,30,33]. In these scholarly works, IR techniques were instrumental during the collection [16,20,21], examination [16–19], and analysis [12,16,19,23] phases of DFI. We classified the selected studies into three predominant modeling paradigms: logic-based, classical machine learning-based, and deep learning-based approaches.

4.1. Logic Based

A logic-based, also symbolic AI, approach in AI is a methodology that uses formal languages like logic to represent knowledge and reasoning about problems and domains. They encode human knowledge in a compact and usable manner and can manipulate symbols to make deductions and inferences based on predefined rules. They can also learn new knowledge from examples and existing domain knowledge [42].

Abduction, hybrid logic–probabilistic, and causal reasoning approaches are some examples of logic-based approaches which use formal languages like logic to represent knowledge and reasoning about problems and domains. Abduction is a form of logical reasoning that starts with single or multiple observations and then seeks to find the most likely explanation or conclusion for the observation. Abductive reasoning is useful for commonsense reasoning, diagnosis, planning, and natural language. Hybrid logic–probabilistic approaches are methods that combine logic and probability to handle uncertainty and complexity. Causal reasoning is an approach that involves the use of causal relationships to infer the effects of actions, events, or interventions. It can also be used to explain why something happened or to predict what will happen under different scenarios. Causal reasoning is based on the assumption that there are causal mechanisms that govern the behavior of systems and that these mechanisms can be represented by causal models, such as causal graphs, causal networks, or structural causal models [43]. In this section, we review the following papers that employ logic-based approach for modeling IR in DF and related domains.

Cheng et al. [17] address the problem of cyber situation comprehension for Internet of Things (IoT) systems, which are vulnerable to Advanced Persistent Threat (APT) attacks,

by utilizing the concepts of IR. They argue that existing methods for cyber situation awareness are not suitable for IoT systems, as they do not consider the semantic and logical relationships among different types of data. Therefore, they propose a similarity-based method for the comprehension of APT attacks in IoT environments. In order to do this, they build a framework called APTALCM, which consists of an ontology of the APT potential attacks and two modules for alert and log correlation. The ontology models the concepts and properties to formalize APT attack activities in IoT systems. It depicts the attacks using the classes (alerts and logs), attributes, domain, relationships among instances, and similarity of instances. They use an alert class with seven attributes and six log classes with 19 attributes to calculate the similarity within each class. The alert and log correlation modules use a similarity-based method based on SimRank to recognize the APT attack intentions and scenarios. SimRank is a general similarity measure that exploits the object-to-object relationships in graphs, based on the idea that "two nodes are similar if they are pointed to (have incoming edges) from similar nodes". The alert correlation module uses SimRank to reconstruct APT attack scenarios by measuring the similarity between alert instances. In contrast, the log correlation module uses SimRank to detect log instance communities by measuring the similarity between log instances. As a result, APTALCM can accomplish the cyber situation comprehension effectively by recognizing the APT attack intentions in the IoT systems. The experimental results demonstrate that the two kernel modules, i.e., Alert Instance Correlation Module (AICM) and Log Instance Correlation Module (LICM) in APTALCM achieve a low false positive rate of 4.2% and a high true positive rate of 83.7%.

Mirsky et al. [16] propose two new metric-based algorithms for goal recognition in network security by adapting previously proposed planner-based algorithms. The first algorithm is Plan Edit Distance (PED), which calculates the distance metric between the optimal plan and the observation sequence without requiring online planner execution. The second algorithm is Alternative Plan Cost (APC), which finds the minimal mapping from the states visited by the attacker to the states in the optimal plan. They experiment on a network of 60 hosts and compare five algorithms, including PED, APC, two planner-based algorithms proposed by previous researchers, and one planner-based algorithm which is modified to run offline. The experiments confirm that PED and APC outperform the planner-based algorithms in terms of prediction quality, noisy observations, and running times. However, in terms of missing observations, the planner-based algorithms are shown to be more robust.

Chen et al. [12] propose an attack graph-based method to recognize the intention of attackers in network security, especially for complex and multi-step attacks. In the first step of their method, they identify the key assets in the network by calculating the confidentiality, integrity, and availability (CIA) triads for each asset and ranking them according to their security importance. Then, they generate hypothetical attack intents based on the security requirements of the key asset and the network topology. An attack intent is defined as a specific goal that an attacker wants to achieve by exploiting the vulnerabilities in the network. Next, they adopt an attack path graph generation algorithm based on vulnerability attributes, network accessibility, and causality model. An attack path graph is a directed graph that represents the possible attack paths from the attacker's entry point to the target asset. Finally, they identify the network attack intent by employing qualitative and quantitative attack intent analysis. The qualitative analysis matches the attack path information to a corresponding attack intent, while the quantitative analysis quantifies the degree of concealment of vulnerabilities, the probability of successful utilization, and the similarity between the attack path and the hypothetical attack intent. They also conduct an experiment involving three network domains and eight hosts and show that their method can successfully identify the intents of attackers.

Shinde et al. [19] propose a model for cyberattack IR using the Interactive Partially Observable Markov Decision Process (I-POMDP), a framework for modeling strategic interactions under uncertainty. They apply their model to a cyberdeception domain, where

the defender and the attacker interact on a single honeypot host system. They consider three types of attackers with different objectives and preferences: the data exfil attacker, who aims to steal sensitive data; the data manipulator, who aims to modify critical data; and the persistent threat, who aims to maintain a strong presence for future attacks. Their model actively deceives the attacker by providing fake data and observes the attacker's reactions to infer their behavior and intent. Their model also estimates the attacker's beliefs, capabilities, and preferences, and uses them to calculate how the deception affects the attacker's mental state. They conduct simulation-based and agent-based experiments to compare their model with other strategies for IR. They show that their model can effectively recognize the attacker's type and intent, and provide appropriate deception strategies. They claim that their model achieves significantly higher accuracy and robustness in predicting the attacker's actions and goals than the other commonly known strategies.

Kim et al. [18] propose an attack detection application for the Android OS to protect users' personal information from theft. The application uses an attack tree approach to detect the intention of the attacks. The algorithm has two phases: pre-phase and post-phase. The pre-phase consists of four steps: collect, normalize, create a tree, and apply levels. In phase one, the attack intents are categorized into three: interception, modification, and system damage. Interception attacks aim to steal personal information from the user's device, such as passwords, credit card details, or other sensitive data. Modification attacks aim to alter the user's data or settings, such as changing the user's password or modifying the user's contacts. System damage attacks aim to damage the user's device or the system, such as deleting files or rendering the device unusable. The post-phase also consists of four steps: log collect; compare and analyze; visualize; and warn or block. The system is tested using two attacks, smishing (which is SMS phishing) and backdoor, and it successfully detects them.

The work by Zhang et al. [20] introduces an innovative approach for recognizing attack intentions in network security. Their research centers around the premise that the dynamics of attack–defense interactions resemble a strategic game, characterized by opposition, non-cooperation, and strategy-dependent decision-making. To unravel the true intents behind network attacks, the authors propose a framework grounded in signaling game theory. They identify key assets and categorize the possible attacks on each key asset. They also map attackers' intent to security requirements (CIA) and generate possible hypotheses of attack intentions. In their methodology, they generate attack intention hypotheses, leveraging the signaling game model. They then compute the probabilities associated with each attack intention by solving game equilibria. To validate their approach, they employ NetLogo simulations, providing empirical evidence of its effectiveness. The authors claim that the method effectively improves the accuracy of attack IR.

Summary

The logic-based approach remains the prevailing method in addressing the challenge of IR within DF and related domains. This preference may stem from the domain's inherent need for explainability, as DF investigators are tasked with elucidating the rationale behind a suspect's culpability, and this approach provides a structured framework for explaining both why and how conclusions are derived [44]. Over the past years, this approach has consistently dominated the field as highlighted by Van-Horenbeke et al. [11].

An analysis of the available literature, as listed in Table 2, reveals that the majority of research efforts on IR related to DF center around the subdomain of network security. These studies primarily delve into the analysis of various alerts and network traffic data. Notably, the work by Cheng et al. [17] focuses on IR in the APT on IoT subdomain, while Kim et al. [18] contribute to the role of intent in mobile security. These show that there exist notable gaps in the application of IR technology across different DF categories. Furthermore, most works focus on the IR level, while the work by Mirsky et al. [16] operates at a higher level of plan recognition. In contrast, there is no study that focuses on malicious activity detection, operating at a granular level.

The logic-based approach, while valuable for IR, faces several challenges and limitations. First, scalability remains an issue; these models can be computationally expensive and struggle to handle large and complex domains, especially when dealing with uncertainty, inconsistency, or incomplete information. Second, integration poses difficulties; logic-based methods may not seamlessly combine with other AI techniques, such as subsymbolic approaches (e.g., neural networks) or hybrid models that leverage the strengths of both paradigms. Third, while logic-based systems are generally more interpretable than subsymbolic counterparts, they can still be too abstract or complex for human understanding. Unfamiliar symbols, technical jargon, or lengthy proofs may hinder trust in their results. Fourth, the inherent rigidity of rule-based systems demands that cases neatly fit predefined rules for accurate identification. Finally, the manual introduction of new knowledge by experts is a necessity. However, in extensive and intricate domains, this reliance on human expertise introduces the risk of errors and limitations in keeping up with evolving scenarios.

Table 2. Summary of research on IR in DF and cybercrime: using logic-based method.

Article	Subdomain	Approach	Intent Level	Accuracy
Mirsky et al., 2019 [16], New goal recognition algorithms using attack graphs	Network security	Attack graph, metric based algorithm	Plan	Online test in seconds: • R&G + SC: 0.6578, • PED: 0.0002, • AED: 0.3246
Cheng et al., 2019 [17], Cyber situation comprehension for IoT systems based on APT alerts and logs correlation	APT on IoT	Similarity based	Intent	• True positive: 83.7 • False negative: 4.2
Kim et al., 2019 [18], Attach detection application with attack tree for mobile phone using log analysis.	Mobile forensics	Attack tree	Intent	-
Chen et al., 2020 [12], Attack intent analysis method based on attack path graph	Network security	Attack path graph	Intent	-
Shinde et al., 2021 [19], Cyber-attack intent recognition and active deception using factored interactive POMDPs	Network security	Partially observable Markov decision process	Intent	-
Zhang et al., 2021 [20], Network attack intention recognition based on signaling game model and Netlogo simulation	Network security	Signal gaming model	Intent	-

4.2. Classical Machine Learning

Classic machine learning approaches use statistical methods and machine learning techniques to learn patterns and models from data that can be used to recognize the actions, and intents of the observed agent. They usually do not require much domain knowledge or human intervention, but they need a large amount of labeled data to train the models. They can handle uncertainty and noise in the data, but they may not capture the underlying

structure and semantics of the problem domain. They also may not generalize well to new or unseen situations. These algorithms can be further divided into two categories: supervised learning and unsupervised learning.

In supervised learning, the algorithm is trained on labeled data, where the correct answer is provided to the algorithm. Some widely used supervised learning algorithms include k-Nearest Neighbor (KNN), Support Vector Machines (SVMs), decision tree, and logistic regression. The first three algorithms are used for both classification and regression tasks, while logistic regression is used for regression only. KNN works by finding the k-nearest data points to the input data point and then classifying the input data point based on the majority class of the k-nearest neighbors. SVMs work by finding the hyperplane that best separates the data points into different classes. The hyperplane is chosen such that the margin between the hyperplane and the closest data points from each class is maximized. Decision tree works by recursively splitting the data into subsets based on the values of the input features until a stopping criterion is met. The stopping criterion can be a maximum depth, a minimum number of samples per leaf, or a minimum reduction in impurity. Logistic regression works by modeling the probability of the input data point belonging to a certain class using a logistic function that maps any real-valued input to a value between 0 and 1, which can be interpreted as a probability.

On the other hand, unsupervised learning algorithms are used to find patterns in data without any prior knowledge of the data's structure. Some widely used supervised learning algorithms include the following. K-Means clustering works by partitioning the data into k clusters based on the similarity of the data points. The algorithm starts by randomly selecting k centroids and then iteratively assigns each data point to the nearest centroid. The centroids are then updated based on the mean of the data points assigned to them, and the process is repeated until convergence. Hierarchical clustering works by creating a hierarchy of clusters by recursively merging the most similar clusters. The algorithm starts by treating each data point as a separate cluster and then iteratively merges the two closest clusters until all the data points belong to a single cluster. These two algorithms are used for clustering tasks. Principal Component Analysis (PCA) works by finding the principal components of the data, which are the directions in which the data vary the most. The algorithm then projects the data onto these principal components, reducing the dimensionality of the data while retaining most of the information. t-Distributed Stochastic Neighbor Embedding (t-SNE) works by mapping high-dimensional data to a low-dimensional space while preserving the pairwise distances between the data points. The algorithm is particularly useful for visualizing complex, nonlinear structures in the data. We review studies that utilize the classical machine learning approach in this section.

Ahmed et al. [13] propose a method for recognizing the intentions of cyberattackers based on similarity analysis. They define two types of attack intentions: general and specific. The general intentions correspond to the security objectives of availability, confidentiality, and integrity, while the specific intentions refer to the actual attacks or violations such as DDoS. The main contribution of their paper is the creation of attack patterns, which are the key to IR. The attack patterns are constructed by extracting the features of the main attributes of the known attacks and formulating them as evidence. The second contribution is the improvement in the process of investigating the similarity between the created patterns and the new attacks, which is the core of their method. They devise a similarity metric-based algorithm using the fuzzy min-max (FMM) neural network technique. The algorithm compares a new attack with the existing attack patterns and evaluates the level of similarity between them to identify the attacker's intentions. Their method is able to create a new class of signature or pattern if the new attack is not similar to any of the existing patterns. The authors claim that their method provides useful information and increases the possibility of recognizing attack intentions in advance by eliminating similar cases using the FMM neural network model. They test their method on a subset of the page block dataset and demonstrate its high accuracy and efficiency.

Considering the fact that criminals often use slang expressions to communicate, plan, and execute their illicit activities online, to capture the hidden meanings and intention behind these expressions, Mendonça et al. [21] propose a framework to detect and classify criminal intentions in social media texts ciphered with slang. The framework, called Ontology-Based Framework for Criminal Intention Classification (OFCIC), combines semantic web, semiotics, speech act theory, and machine learning techniques to select, decipher, and classify posts with criminal slang expressions according to their illocutionary classes, which are the types of speech acts that convey the speaker's intention. The framework consists of four main steps: (1) data collection and preprocessing, (2) ontology-based post-selection, (3) ontology-based post deciphering, and (4) intention classification. The framework utilizes machine learning models such as SVM, neural networks, and random fields to classify the texts according to their criminal intent. They show that their framework can effectively identify posts with criminal slang expressions, translate them into standard language, and classify them into eight illocutionary classes: proposal, inducement, forecast, wish, assertion, valuation, palinode, or contrition. The authors evaluate the framework on a dataset of 8.8 million tweets and demonstrate its effectiveness in automatically classifying criminal intentions from social media texts with slang. The paper contributes to the field of cybercrime prevention by providing interdisciplinary approach to analyze social media slang-ciphered texts in Portuguese.

The article by Abarna et al. [23] presents an algorithm for detecting cyberharassment and intention from text on social media platforms, using Instagram comments as a case study. The paper utilizes a conventional scheme that analyzes the lexical meaning of the text using natural language processing (NLP) techniques, and a fast text model that captures the word order of the text. The authors perform various preprocessing steps to normalize and contextualize the text, and then employ a Bag of Words (BOW) model and a Word2Vec technique to transform the words into vectors. To identify the intention of the comments, such as bullying, threatening, or trolling, they use a probabilistic similarity technique that compares the vector representations of the words. The authors also devise a score for intention detection that incorporates the frequency of words and the bully–victim participation score, which quantifies the degree of engagement of the users in the cyberharassment scenario. They evaluate the effectiveness of their algorithm using various metrics and benchmark it against seven existing methods, including random fields, SVM, and Bidirectional Long Short-Term Memory (Bi-LSTM). They demonstrate that their algorithm outperforms all the other methods in terms of precision, recall, and F1 score. The authors conclude that their algorithm achieves superior accuracy and a lower error rate than the state-of-the-art methods and that it can robustly detect cyberharassment and its intention on social media platforms.

Li et al. [22] propose an approach to recognize multi-step attacks by employing a hidden Markov model (HMM) with probabilistic reasoning. As multi-step attacks have interrelated attack steps, to accurately obtain the internal relationship between different attacks, they employ the concept of temporal relationship. Considering the dynamic characteristics of the network, they employ runtime rule updating. Furthermore, rather than analyzing the intents of each attack, they consider higher-level intrusion IR and apply probabilistic reasoning. They build three algorithms: the parameter estimation algorithm to estimate the parameters of the HMM model for alerts correlation; the attack intent inference algorithm to infer the attack intent based on the observation sequence for possible attack IR; and the attack prediction algorithm to analyze the possible attack sequence for possible attack prediction. They build three models based on the HMM, HMM with Probabilistic Inference (HMM-PI), and HMM-PI with an updated Conditional Probability Table (CPT) model (HMM-PI-UCM), experiment with the LLDOS1.0 dataset from MIT, compare the three models, and find that the HMM-PI-UCM model performs better.

Recognizing the criticality of discerning intent for law enforcement and crime deterrence, Bokolo et al. [24] implement a comparative analysis utilizing five conventional machine learning algorithms—logistic regression, ridge regression, SVM, Stochastic Gradient

Descent (SGD), and random forests—to ascertain intent from social media communications, with a particular focus on Twitter. The Sentiment140 dataset, comprising 400,000 tweets, serves as the foundation for their study, partitioned into an 80% training subset and a 20% testing subset. Their methodological approach commences with data preprocessing to eliminate extraneous noise and irrelevant content. Subsequently, feature extraction is conducted to determine the relative significance of each term. To reveal underlying patterns, tokenization is employed, followed by the training of the models using the refined dataset. The evaluation of the models is based on metrics such as accuracy, precision, recall, and F1 score. Logistic regression outperforms its counterparts, achieving an accuracy rate of 92.87%, while SVM, random forest, ridge regression, and SGD yield accuracy rates of 92.56%, 92.39%, 90.88%, and 89.51%, respectively. This comparison underscores the efficacy of logistic regression in this context and sets a benchmark for future research in the domain of intent detection within social media landscapes.

Summary

The classical machine learning-based approach is employed by researchers to address the limitations of logic-based methods, particularly those related to rigidity and manual knowledge encoding. Additionally, this approach is well suited for handling uncertainties, as it leverages probability. The introduction of probability also proves valuable in managing partial observability and handling various data noises.

As shown on Table 3, the landscape within the subdomain has undergone a significant shift, transitioning from a focus primarily on network security (in the case of logic-based approaches) to encompassing a broader range of cases [13,22]. Additionally, researchers have delved into identifying intents related to social media utilization as explored by [21,23,24]. Notably, the work by Li et al. [22] stands out, as it operates at a higher level of plan recognition, while the remaining studies primarily address intent or goal recognition.

Table 3. Summary of research on IR in DF and cybercrime: using classical machine learning method.

Article	Subdomain	Approach	Intent Level	Accuracy
Ahmed et al., 2018 [13], SAIRF: A similarity approach for attack intention recognition using fuzzy min-max neural network	General attack	Fuzzy min-max neural network	Intent	94.74%
Mendonça et al., 2020 [21], A framework for detecting intentions of criminal acts in social media: A case study on Twitter	Social media	Similarity based	Intent	• True positive: 83.7 • False negative: 4.2
Li et al., 2020 [22], Attack plan recognition using hidden Markov and probabilistic inference	General attack	Hidden Markov	Plan	-
Abarna et al., 2022 [23], Identification of cyberharassment and intention of target users on social media platforms	Social media	Similarity based	Intent	Precision: 91.45%
Bokolo et al., 2023 [24], Leveraging machine learning for crime intent detection in social media posts	Social media	Logistic regression	Intent	92.87%

However, this method also faces several limitations. Some of these are akin to logic-based approaches, including scalability issues due to the challenges posed by scaling probabilities. Additionally, as the number of parameters increases, manual input becomes necessary. Furthermore, the approach has specific limitations, notably a lack of applicability as understanding how conclusions are inferred can be challenging. This becomes particularly critical in applications related to DF, where explainability is a mandatory requirement.

4.3. Deep Learning

Deep learning approaches use deep neural networks (DNNs) to learn high-level features and representations from data that can be used to recognize the actions, plans, and goals of the observed agent. They usually do not require any domain knowledge or feature engineering, but they need a huge amount of labeled data to train the networks. They can handle complex and multimodal data, but they may not be interpretable or explainable. They also may overfit the data or suffer from catastrophic forgetting.

Some widely used deep learning algorithms include Convolutional Neural Networks (CNNs): these are deep learning networks that are commonly used for image recognition tasks. They work by applying convolutional filters to the input image to extract features and then passing these features through a series of fully connected layers to make a prediction. Recurrent Neural Networks (RNNs): these are deep learning networks that are commonly used for sequence prediction tasks such as speech recognition and NLP. They work by processing the input sequence one element at a time and maintaining an internal state that captures the context of the sequence. Generative Adversarial Networks (GANs): these are deep learning networks that are used for generating new data that is similar to the training data. They work by training two networks: a generator network that generates new data and a discriminator network that tries to distinguish between the generated data and the real data. The two networks are trained together in a process called adversarial training. Long Short-Term Memory Networks (LSTMs): These are deep learning networks that are commonly used for sequence prediction tasks such as speech recognition and NLP. They work by maintaining an internal state that captures the context of the sequence and using this state to make predictions. Different researchers have applied these algorithms to solve IR challenges related to the DF domain, and we dedicate this section to reviewing them.

Navalgund et al. [25] propose a deep learning-based system that can detect criminal intentions in real-time videos and images captured by closed-circuit television (CCTV) cameras in various locations. The system aims to enhance the crime control and prevention capabilities of the existing surveillance infrastructure. The system employs and evaluates different pre-trained models, such as VGGNet-19 and GoogleNet InceptionV3, to identify and localize objects of violence, such as guns and knives, in the input data. The experimental results show that VGGNet-19 outperforms GoogleNet InceptionV3 in terms of accuracy and efficiency in detecting crime objects and inferring criminal intents. They also use Faster RCNN to draw bounding boxes over the detected guns and knives. Furthermore, the system incorporates a text message alert mechanism that notifies the relevant authorities when potential crimes are detected.

Martinez-Mascorro et al. [28] propose a deep learning model leveraging 3D Convolutional Neural Networks (3D CNNs), a state-of-the-art methodology for spatio-temporal analysis, to preemptively identify shoplifting intentions from surveillance video. The focus of the study is the proactive identification of criminal intent, specifically targeting the behavioral precursors to shoplifting—referred to as precrime behavior (PCB)—before the individual exhibits overtly suspicious actions. The research utilizes the UCF-Crime dataset, which comprises 1900 real-world surveillance videos totaling 129 h, to train and test the proposed model. They conduct four preliminary experiments aimed at optimizing the model's configuration. This phase is critical in determining the most effective parameters for the 3D CNN architecture. Following this, two confirmatory experiments are carried out to validate the model's performance with the optimized configuration across a larger dataset. The results of these confirmatory test demonstrate that the model achieves an

accuracy rate of 75%, showcasing its potential as a tool for early detection of criminal intent, thereby offering valuable insights for security and law enforcement agencies.

Pandey et al. [14] propose a distributional semantic approach to detect malicious intent in Twitter conversations related to sexual assault. The authors aim to detect the intention by building a typology for malicious intent using social construction theory. The typology includes three categories of intent: accusational, validational, and sensational. The accusational category refers to messages that accuse someone of sexual assault or harassment. The validational category refers to messages that validate the experience of sexual assault or harassment. The sensational category refers to messages that focus more on politics or provocation than on the issue of rape or sexual assault. The authors adopt a CNN to model the system and test their model using Twitter messages collected over four months. They compare their model against several baseline models and find that their system performs better.

In order to detect query-based adversarial black-box attacks on DNN at an early stage, Pang et al. [26] introduce a model called AdviMind. The model has three variants: Naive Intent Estimator, which only serves as a passive observer of the adversaries' queries, provides a baseline understanding of intent but lacks robustness and proactive features. Robust Intent Estimator, which is built upon the naive model, is capable of identifying fake queries even in the presence of adversarial noise. It maintains reliability while estimating intent. Proactive Intent Solicitation, which is the most advanced model, not only estimates intent robustly but also actively prompts adversaries to reveal their true intent. By synthesizing query results, it deters successful attacks and achieves early-stage detection. Empirical evaluation of the models on different datasets demonstrates that these models can detect attack intents with an accuracy of over 75% after observing fewer than three query batches. Additionally, they increase the query cost of adaptive attacks by more than 60%.

The paper by Zhao et al. [27] aims to demystify cyberattack intent by analyzing the preference of intruders using a novel framework called HinAp. The framework uses attributed heterogeneous attention networks and transductive learning to analyze the attack preferences of intruders. They first build an Attributed Heterogeneous Information Network (AHIN) of attack events to model attackers, vulnerabilities, exploited scripts, compromised devices, and 20 types of meta-paths describing interdependent relationships among them, in which attribute information of vulnerabilities and exploited scripts are embedded. Then, they propose the attack preference prediction model based on the attention mechanism and transductive learning. They collect social data to train and test their model. Finally, an automated model for predicting cyberattack preferences is constructed by stacking these two basic prediction models, which are capable of integrating more complex semantic information from meta-paths and meta-graphs to characterize the attack preference of intruders. They compare their model with six other models and their model outperforms all.

Hsu et al. [30] propose an approach to detect malicious activity in physical environments. The proposed method is aimed at reducing the risk of malicious activities by combining three fundamental defense systems, namely access control, surveillance, and host defense systems. Firstly, they employ a Multilayer Perceptron (MLP) model to identify anomalies in access control systems. By analyzing login attempts and the duration of successful logins, the MLP effectively pinpoints suspicious behavior. Secondly, the researchers harness the power of NLP, specifically leveraging techniques like Word2Vec and deep learning, to detect anomalies arising from executed commands. This linguistic analysis provides valuable insights into potentially harmful actions. Thirdly, the team utilizes the YOLOv5 object detection model to identify unauthorized entry points. By monitoring physical spaces, they can swiftly detect any breaches. To assess the proximity of individuals to restricted areas, they employ distance measurement methods such as Intersection Over Union (IOU) and Intersection Over Area (IOA). These metrics help determine whether people are accessing unauthorized zones. Finally, the researchers integrate the results from all three anomaly detection components, aggregating threat scores to generate a malicious

activity alarm. The authors execute experiments on their model and claim that their method successfully detects malicious activity.

Kang et al. [31] propose a framework called ActDetector that detects attack activities automatically from the raw Network Intrusion Detection System (NIDS) alerts, which will greatly reduce the workload of security analysts. The framework consists of three components: an extractor, an embedder, and a classifier. The extractor extracts attack phase descriptions by using a knowledge base of adversary tactics and techniques. The embedder uses doc2vec embedding to obtain the numerical representation of the attack phase descriptions. Finally, the classifier employs a temporal-sequence-based LSTM model to detect the attack activity type from the attack activity description. The authors evaluate ActDetector with three datasets, and their experimental results demonstrate that ActDetector can detect attack activities from the raw NIDS alerts with an average of 94.8% precision, 95.0% recall, and 94.6% F1-score.

The paper by Tsinganos et al. [32] proposes CSE-PersistenceBERT, a transfer learning-based model that can detect the persistence of chat-based social engineering (CSE) attacks, which are malicious attempts to manipulate the behavior of online users by exploiting their psychological vulnerabilities. The paper argues that persistent CSE attackers use different chat texts to achieve the same malicious goal, such as phishing, fraud, or malware installation, and that recognizing the persistence of CSE attacks is an important step to prevent them from succeeding. The paper adapts BERT-base, a pre-trained language model that has shown impressive results in various NLP tasks, and fine-tunes it on a small size corpus that they create, called CSE-Persistence, which contains more than 16 thousand pairs of chat texts, annotated as similar, identical, or different in terms of their intentions. The paper evaluates CSE-PersistenceBERT on a test set of CSE-Persistence and compares it with BERT-base. The paper reports that CSE-PersistenceBERT outperforms BERT-base in terms of accuracy, precision, recall, and F1-score, demonstrating its effectiveness and robustness in detecting the persistence of CSE attacks. The CSE-PersistenceBERT model can be used as a specific part of a general CSE attack detection system, which can alert the users or the administrators of potential threats and prevent them from falling victim to the CSE attacks.

To add more to the chat-based social engineering (CSE) attack detection system, Tsinganos et al. [34] propose a deep learning-based model for recognizing the intentions of CSE attacks using dialogue state tracking. They create an ontology and a small corpus called SG-CSE, and, adopted from BERT-base, they build a model called SG-CSE BERT.They test their model by using the dataset to evaluate their approach and achieve promising results.

Tang et al. [29] present a method for detecting the attack intentions of malicious actors in power systems using graph convolutional networks (GCNs). Their proposed model, called Attack Intention Detection for Power System Using Graph Convolutional Networks (AIGCN), consists of two main steps. First, they identify the abnormal IPs based on their log execution behaviors, using four tuples: destination IP, destination port, event time, and protocol. This step aims to filter out the normal IPs and reduce the noise in the data. Second, they model a graph from the interactive relationship among abnormal IPs, construct an attack graph, and apply a GCN model to learn the patterns and classify the attack intentions. This step leverages the graph structure and the node features to capture the complex and dynamic behaviors of the attackers. They evaluate their model on two datasets that they prepared from real-world network logs and compare it with five baseline methods, such as LSTM and BERT. The results show that AIGCN achieves a high precision of 97.34% and 98.25% for both datasets, outperforming the baseline methods, which demonstrates the effectiveness and robustness of the AIGCN model for detecting attack intentions in power systems.

Bhugul et al. [33] propose a deep learning model for detecting suspicious activities in private settings, such as bank robberies. While security cameras are already commonplace, real-time reaction and 24/7 monitoring are essential for automated detection techniques. This study addresses the critical need for preventive measures against gunshots and terrorist

attacks in public areas with heavy foot traffic. The focus of their study is on identifying suspicious human activity related to weapons. Specifically, they consider two parameters, a person with a weapon (gun), and a person wearing a helmet with a weapon. They introduce an algorithm for multiple gun detection using a modified dense deep learning neural network (CNN) model to detect guns from video frames. The temporal complexity of the model across various hardware platforms is also explored, and the proposed system is able to detect all types of guns with an impressive 99.3% accuracy, outperforming existing methods, such as YOLO v3, v4, v5 and SVM.

Summary

The deep learning approach overcomes some of the limitations of the logic-based and classical machine learning approaches. One of the main advantages of the approach is that it can automatically learn features from the data, which means that it does not require the features to be hand-engineered. Because of that, they can learn different patterns and uncover non-linear relationships in data that would be difficult to detect through traditional methods. This makes it a useful tool for extracting insights from big data. The approach has paramount importance particularly for tasks where the features are difficult to define, such as image recognition. Deep learning algorithms can handle large and complex datasets that would be difficult for classical machine learning- and/or logic-based algorithms to process. They are also good at dealing with uncertainty, partial observability, and noise, which makes them a useful tool for IR.

The literature reviewed on deep learning for IR, as shown on Table 4, reveals that the subdomains have shifted from network security to social media (4 out of 10 articles) and physical security (3 out of 10 articles), while only 2 articles focus on network security. This shift in focus from network security to social media and physical security suggests that IR is becoming more relevant in these domains. Additionally, a new subdomain related to AI security has emerged. The emergence of this new subdomain highlights the need for IR-based models in the context of securing AI itself. Transfer learning is employed in many cases to improve the performance of deep learning models. This also indicates that deep learning models can benefit from pre-trained models to improve their performance.

Table 4. Summary of research on IR in DF and cybercrime: using deep learning method.

Article	Subdomain	Approach	Intent Level	Accuracy
Navalgund et al., 2018 [25], Crime intention detection system using deep learning	CCTV	Transfer learning	Intent	92%
Pandey et al., 2018 [14], Distributional semantics approach to detect intent in Twitter conversations on sexual assaults	Social media	Distributional semantic and CNN	Intent	-
Pang et al., 2020 [26], AdvMind: Inferring adversary intent of black-box attacks	Black-box attack	DL	Intent	75%
Zhao et al., 2021 [27], Automatically predicting cyberattack preference with attributed heterogeneous attention networks and transductive learning	Social attack	Attention mechanism and transductive learning	Intent	-
Martinez-Mascorro et al., 2021 [28], Criminal Intention Detection at Early Stages of Shoplifting Cases by Using 3D Convolutional Neural Networks	CCTV	3D CNN	Intent	75%
Q. Tang et al., 2022 [29] AIGCN: Attack intention detection for power system using graph convolutional networks	Network security	Graph convolutional networks	Intent	97.34 %
Hsu et al., 2022 [30], Detection of malicious activities using machine learning in physical environment	Access control, surveillance, and host defense systems	YOLOv5 object detection model	Intent	-
Kang et al., 2022 [31], ActDetector: A Sequence-based framework for network attack activity detection	Network security	Temporal-sequence-based LSTM	Activity	Precision: 94.8%

Table 4. *Cont.*

Article	Subdomain	Approach	Intent Level	Accuracy
Tsinganos et al., 2022 [32], Applying BERT for early-stage recognition of persistence in chat-based social engineering attacks	Social engineering attack	Transfer learning	Intent	78.03%
Bhugul et al., 2023 [33], Novel deep neural network for suspicious activity detection and classification	CCTV	CNN	Intent	99.3%
Tsinganos et al., 2023 [34], Leveraging dialogue state tracking for zero-shot chat-based social engineering attack recognition	Social engineering attack	Transfer learning	Intent	78.03%

However, deep learning approaches also have several disadvantages. Firstly, they require a large amount of training data to achieve high accuracy, similar to classical machine learning approaches. Secondly, they are not explicable, to the extent that even the designers do not know how the conclusions are inferred from the input evidence. This lack of transparency can also make it difficult to debug and improve the model. Thirdly, most deep learning models cannot learn new classes from live/online data. This means that if the model encounters a new class of data that it has not seen before, it will not be able to recognize it. Finally, deep learning models require high computational power to train and run, which can be a significant barrier to entry for many researchers and organizations. These limitations can make it challenging to use deep learning approaches for IR in practice.

5. Open Challenges

IR models face various challenges as identified by Van-Horenbeke et al. [11]. While most of these challenges are inherited in the context of DF (DF) and related domains, the magnitude and characteristics of the challenges differ. Our review identified the following challenges, though this is not an exhaustive list:

- Contextualization: This is the need to recognize the context in which an action is performed, as it can affect the interpretation of the action. In DF, the cybercrime context can provide valuable information about the intent of the actor [12,17,18,21,23,25,29].
- Missed activities, partial observability, or handling noise: This is the difficulty of recognizing an activity, and consequently intent, when only a part of it is observed or when some of the actions are totally missed. It is common to encounter incomplete or partial data about cybercrime. This can make it difficult to recognize an attack or to determine the intent behind it [16,19,29].
- Predictive capabilities: This is the ability to predict future actions based on current observations. Such capabilities can be useful in DF to reconstruct the crime scene and identify potential future attacks in order to predict the behavior of an attacker [16,19,22].
- Handling uncertainty: This is the need to handle uncertainty in the observations and the predictions. Logic-based systems generally lack such capability, while classical machine learning and deep learning handle it by using probabilistic and statistical approaches. In DF, there is often a high degree of uncertainty, as the observed agents try to hide themselves and usually take high precautions. This can make it difficult to determine the intent behind an action or to make accurate predictions [12,26].
- Managing multiple hypotheses: Different IR models generate multiple hypotheses about the intent of the agent based on the observed activities. It is important to manage these hypotheses and select the most plausible one by evaluating them based on the available evidence [12,29].
- Multi-step attack: Nowadays, attackers, instead of attacking immediately, build their attack through time by advancing step by step. Identifying such a situation from network traffic is difficult, and consequently, recognizing the intent is challenging [12,22].

- Scalability: DF generates a huge volume of heterogeneous data that IR models are expected to scale in the sense of size and type of data as well as the complexity of the environment [16].
- Cooperating agents: This is the need to recognize the intentions of multiple agents that are cooperating to achieve a common goal. Identifying cooperating agents can help identify the scope and nature of an attack [13].
- Adversarial agents: Such agents can intentionally try to mislead the recognition system by performing actions that are inconsistent with their true intentions [12–14,18,19,22,26].
- Understanding the attacker's ability and belief: This can help in predicting their future actions and identifying potential vulnerabilities [19,27,29].
- Lack of standard intent classification: This makes it difficult to compare and evaluate different IR systems [12–14,18,19].
- Interpretability: This is the ability to explain the reasoning behind the recognition of an action, intent, or plan so as to develop trust in a court of law. It is not only important but also a mandatory requirement in DF, as courts require an explanation as well as supporting evidence to criminalize or free a suspect.

6. Discussion

In this systematic review, we delve into the intricate world of IR within the realms of DF and cybercrime. Our investigation involves a careful review of existing literature, drawing from reputable journals. Our primary objective is to understand how IR models operate and to identify gaps in the current landscape. Following the categorization proposed by Van-Horenbeke et al. [11], we categorized the collected papers based on their modeling approaches. This systematic grouping allowed us to discern patterns and variations across different studies.

In our analysis, we see that researchers often resort to creating their own intent categories or borrowing from prior studies because there are no predefined and specific intention classes. Table 5 shows the intent classes created by the reviewed studies. This lack of standardized classes underscores the need for a more comprehensive taxonomy. Chen et al. [12] take a commendable step by attempting to generalize technical intents into security triads (CIA). This broader perspective opens up new path for understanding intentions beyond traditional boundaries. Intentions can even be categorized at higher levels such as the management level, and for DF, the legal definitions of intentions can also be considered.

In the synthesis of findings across the three tables (Tables 2–4), a discernible thematic concentration emerges: intent recognition. The majority of scholarly endeavors within this domain are dedicated to unraveling the intricacies of attackers' intentions. Notably, two studies delve into the realm of plan recognition. The first, conducted by Mirsky et al. [16], adopts a logic-based approach, while the second, authored by Li et al. [22], harnesses classical machine learning paradigms. However, there is only one study for activity recognition: in their work, Kang et al. [31] employ a sequence-based LSTM framework to detect network attack activities. In practical terms, the pursuit of IR necessitates not only the identification of discrete actions (commonly referred to as activity detection) but also the consideration of state transitions. Consequently, activity recognition becomes an indispensable facet woven into the fabric of each IR study.

As shown in Figure 2, in recent years, the field of modeling IR has undergone a significant transformation. Traditionally, logic-based methods held sway, but now we observe a transition towards machine learning, particularly deep learning. This evolution reflects the dynamic landscape of research and its practical applications. Van-Horenbeke et al. [11] highlight the historical dominance of logic-based approaches in recognizing activity, intent, and planning. However, contemporary trends reveal a surge in the adoption of deep learning techniques. Most of the current research, as shown in the chart (Figure 2), endeavors to identify malicious intents to prevent cyberattacks. Deep learning models, with their ability to discern intricate patterns from vast data, excel in this domain. However,

when dealing with already committed cybercrimes, the explainability requirement becomes paramount. Logic-based approaches, rooted in formal reasoning and rule-based systems, offer a structured framework for post-incident analysis. Their deterministic nature ensures rigorous adherence to predefined rules, making them indispensable in solving complex cases. While classical machine learning and deep learning models thrive in proactive cybersecurity, they fall short in fulfilling the stringent explainability criteria. So in the context of DFI, logic-based reasoning and data-driven techniques become essential.

Table 5. Intent classes used in studies reviewed.

Article	Intent Classes
Kim et al., 2019 [18], Attach detection application with attack tree for mobile phone using log analysis.	• Interception: steal personal information for the users' device • Modification: alter users' data or setting • System damage: damages the users' devices or system
Zhang et al., 2021 [20], Network attack intention recognition based on signaling game model and Netlogo simulation (They map attackers' intent to security requirements)	• Confidentiality • Integrity • Availability
Ahmed et al., 2018 [13], SAIRF: A similarity approach for attack intention recognition using fuzzy min-max neural network (They divided the intent into general (CIA) and specific (attack target)	• Confidentiality • Integrity • Availability
Mendonça et al., 2020 [21], A framework for detecting intentions of criminal acts in social media: A case study on Twitter	• Proposal • Inducement • Forecast • Wish • Assertion • Valuation • Palinode • Contrition
Pandey et al., 2018 [14], Distributional semantics approach to detect intent in Twitter conversations on sexual assaults	• Accusational • Validational • Sensational

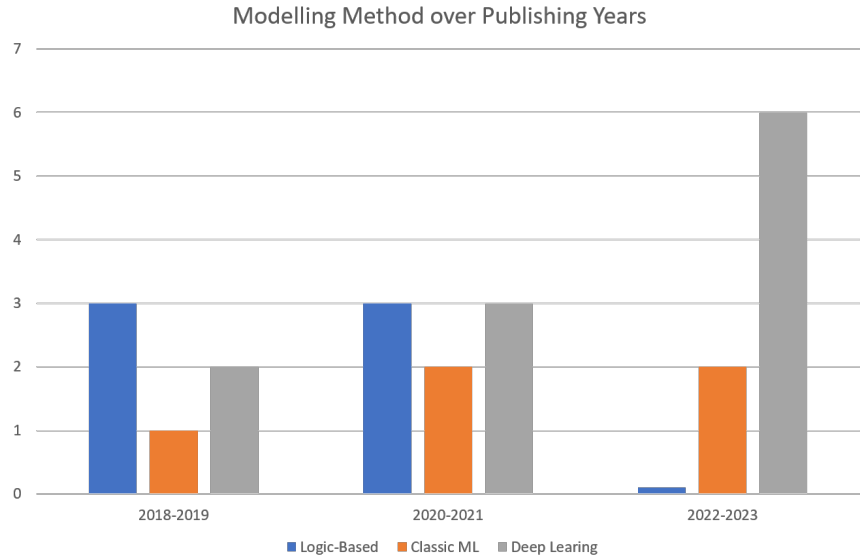

Figure 2. Transitions of IR modeling approaches in DF over the years.

The landscape of IR has expanded beyond its traditional stronghold in network security. While the protection of network infrastructure once dominated this field, recent developments reveal intriguing diversification. Historically, IR found its primary application in safeguarding network boundaries [45]. However, as shown in Figure 3a, and with the perspective of the data type they utilize as shown in Figure 3b, a notable shift has occurred towards social engineering attacks, which, driven by human psychology and manipulation, now prominently employ IR. Classical machine learning and deep learning

techniques are utilized in deciphering the intricate motives of cybercriminals in social engineering attacks. In addition, IR models that utilize the data from IoT devices, specially CCTV cameras, have emerged, employing a deep learning approach. Beyond external threats, researchers explore the application of IR in securing AI systems themselves. As AI algorithms proliferate across domains, safeguarding their integrity and decision-making processes becomes paramount. IR aids in identifying anomalies, unauthorized access attempts, and adversarial inputs. By scrutinizing AI behavior, we fortify the very systems that drive technological advancements.

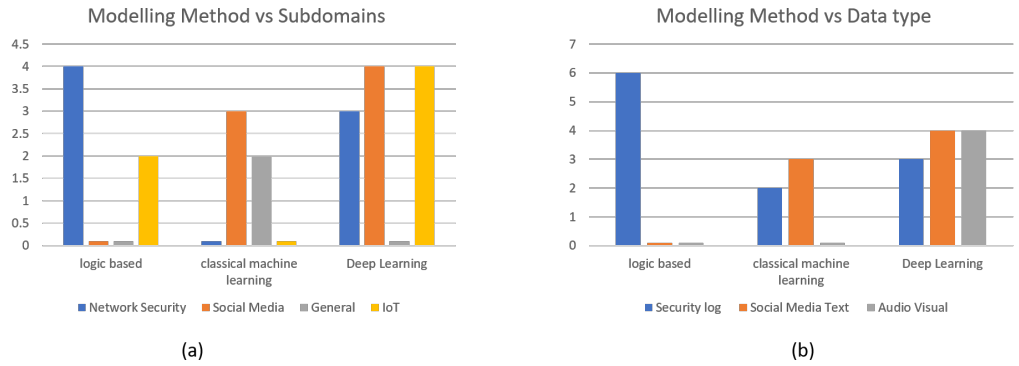

Figure 3. IR modeling approaches in DF per subdomains (**a**) and per data types (**b**).

This multifaceted journey—spanning network security, social engineering, physical security, database security, AI self-defense, and general security—represents a divergence. It serves as a springboard for interdisciplinary exploration. Its fusion with cognitive science, linguistics, and behavioral analysis opens doors to innovative solutions. Whether combating cyberthreats or enhancing AI resilience, IR remains a powerful tool.

Finally, in the adjudication of criminal cases, the judiciary mandates complete elucidations concerning the methodologies employed in crime investigations and the inferential processes applied to derive conclusions from the available evidence. Thus, it is imperative that models developed for DF possess intrinsic attributes of explainability and logical coherence, thereby fostering judicial confidence, mitigating biases, and upholding accountability for their determinations [46]. While deep learning paradigms offer robust pattern recognition capabilities, they inherently lack the faculty to rationalize their outcomes or elucidate their decision-making processes. Explainable Artificial Intelligence (XAI) emerges as a viable alternative, proffering transparency in investigative procedures [47]; nonetheless, such explanations do not inherently ensure the logical soundness of the resultant inferences. Conversely, as shown in Table 6, logic-based methodologies, rooted firmly in logical reasoning, systematically arrive at conclusions through inference, although they suffer from inefficiencies and inaccuracies due to their susceptibility to human error [42].

To address these challenges, future research endeavors should explore hybrid solutions that synergistically combine the strengths of both deep learning and logic-based methodologies. By leveraging the interpretability of logic-based reasoning alongside the predictive capabilities of deep learning, such hybrid models can mitigate their respective weaknesses. This interdisciplinary approach holds promise for advancing the field of DF, by employing models that amalgamate reasonableness with interpretability, ensuring that the mechanisms underlying their conclusions are both comprehensible and justifiable.

Table 6. Advantages and disadvantages of the modeling approaches.

Modeling Approaches	Advantages	Disadvantages
Logic-based approaches	• Compact knowledge representation • Ability to deduct and infer based on logical rules • Able to learn new knowledge from examples • Explainable and understandable by human • Prediction capability	• Not scalable, computationally expensive • Difficult to integrate with other models • Rigid as they use rules • Manual introduction of new rules • Limitation in handling uncertainty
Classical machine learning approach	• Learn from data • Handle uncertainty, as it leverages probability • Minimal human intervention • Handle partial observability • Handle data noises	• Not scalable, though better than logic based • Needs manual encoding of parameters, features • Not explainable
Deep Learning Approach	• No need for feature engineering • Learn from data, and extract insights from big data • Handle complex and multimodal data • Handle uncertainty • Handle partial observability • Handle noises	• Require huge data for training • Not explainable • Overfitting • Catastrophic forgetting • Difficulty to learn new classes • Require high computational power

7. Ethical Considerations

The integration of AI in IR within DF engenders a complex array of ethical considerations. When IR paradigms are deployed for DFI, they inherently gain access to the personal data of individuals under scrutiny. This access harbors the potential for substantial privacy violations, necessitating a judicious equilibrium between the efficacy of the investigative process and the imperative to protect individual privacy rights. Moreover, adherence to ethical and legal standards during the evidence acquisition phase is paramount, as it unequivocally influences the integrity and outcome of the investigation [48]. Furthermore, the automation of the investigative process via AI must be executed with impartiality and without bias. Consequently, AI-driven IR frameworks for DF must embody principles of transparency and accountability to ensure equitable and ethical application [49,50].

8. Conclusions and Future Work

In this systematic literature review, we examined the contributions of IR within the DF and cybercrime domains. Our study followed a rigorous six-step approach, including protocol development and adherence. The following can be considered the key takeaways:

- First, categorization of research: We curated research articles from reputable journals. These studies were classified into three distinct modeling approaches: logic-based, classical machine learning based, and deep learning based.
- Second, expanding beyond network security: Our analysis revealed a significant shift in the application of IR. While it was predominantly utilized for network security in the past, we now observe its adoption across various cybersecurity subdomains. Notably, IR plays a pivotal role in addressing challenges related to social engineering attacks, AI black box vulnerabilities, and physical security.
- Third, deep learning dominance: Among the modeling approaches, deep learning (DL) has emerged as the de facto choice. Its ability to overcome limitations associated with other methods has positioned DL at the forefront. However, a critical consideration arises in the context of DF: the need for explainability.
- Fourth, the explainability conundrum: In the DF domain, explainability is not merely desirable; it is mandatory. DL models, while powerful, often lack transparency. Therefore, researchers should seize this opportunity to explore hybrid solutions, that are transparent and reasonable. Combining the strengths of DL with interpretable tech-

niques such as XAI and logic-based approach could yield more robust and accountable IR systems.
- Fifth, taxonomy development: We emphasize the importance of defining IR more precisely, especially within the context of DF. To address this, we propose the creation of a taxonomy and formal definition. Such a taxonomy would provide clarity, and standardization, and facilitate further advancements in this critical field.

Author Contributions: Conceptualization, methodology, investigation, writing—original draft preparation, Y.W.K.; and supervision, writing—review and editing, J.I.J.; and supervision, writing—review and editing, E.G.B.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The author Joshua Isaac James was employed by the DFIR Science LLC company. All authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DF	Digital Forensics
DFI	Digital Forensics Investigation
IR	Intention Recognition
DL	Deep Learning

Appendix A

The following data extraction sheet is used to extract the data from each article reviewed.

Table A1. Data Extraction Sheet.

Article	DF Category	Content Type	Modeling Approach	Subdomain	Level of IR	Targeted Problem
-	-	-	-	-	-	-

References

1. Malik, J.K.; Choudhury, S. Cyber Space—Evolution and Growth. *East Afr. Sch. J. Educ. Humanit. Lit.* **2019**, *2*, 170–190.
2. Mbanaso, U.M.; Dandaura, E.S. The Cyberspace: Redefining A New World. *IOSR J. Comput. Eng.* **2015**, *17*, 17–24.
3. Granados Franco, E. Global Risks 2020: An Unsettled World. In *The Global Risks Report*; World Economic Forum LLC: New York, NY, USA, 2020; pp. 8–17.
4. Kent, K.; Chevalier, S.; Grance, T.; Dang, H. *Guide to Integrating Forensic Techniques into Incident Response*; The National Institute of Standards and Technology: Gaithersburg, MD, USA, 2006.
5. *ISO/IEC 27037:2012*; Information Technology—Security Techniques—Guidelines for Identification, Collection, Acquisition and Preservation of Digital Evidence. ISO: Geneva, Switzerland, 2012; p. 38.
6. Raghavan, S. Digital forensic research: Current state of the art. *CSI Trans. ICT* **2013**, *1*, 91–114. [CrossRef]
7. Quick, D.; Choo, K.K.R. Impacts of increasing volume of digital forensic data: A survey and future research challenges. *Digit. Investig.* **2014**, *11*, 273–294. [CrossRef]
8. Agarwal, S. Data mining: Data mining concepts and techniques. In Proceedings of the 2013 International Conference on Machine Intelligence and Research Advancement, Katra, India, 21–23 December 2013; pp. 203–207. [CrossRef]
9. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From data mining to knowledge discovery in databases. *AI Mag.* **1996**, *17*, 37–53.
10. Heinze, C. Modelling Intention Recognition for Intelligent Agent Systems. DSTO Systems Sciences Laboratory. 2004. Available online: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA430005 (accessed on 1 April 2024).
11. Van-Horenbeke, F.A.; Peer, A. Activity, Plan, and Goal Recognition: A Review. *Front. Robot. AI* **2021**, *8*, 643010. [CrossRef]
12. Chen, B.; Liu, Y.; Li, S.; Gao, X. Attack Intent Analysis Method Based on Attack Path Graph. In Proceedings of the 2019 9th International Conference on Communication and Network Security, New York, NY, USA, 13 January 2020; ICCNS'19, pp. 97–102. [CrossRef]

13. Ahmed, A.A.; Mohammed, M.F. SAIRF: A similarity approach for attack intention recognition using fuzzy min-max neural network. *J. Comput. Sci.* **2018**, *25*, 467–473. [CrossRef]
14. Pandey, R.; Purohit, H.; Stabile, B.; Grant, A. Distributional Semantics Approach to Detect Intent in Twitter Conversations on Sexual Assaults. In Proceedings of the 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Santiago, Chile, 3–6 December 2018; pp. 270–277. [CrossRef]
15. Cai, Z.; Zhang, Q.; Zhang, R.; Gan, Y. Intrusion intention recognition and response based on weighed plan knowledge graph. *Comput. Model. New Technol.* **2014**, *18*, 151–157.
16. Mirsky, R.; Shalom, Y.; Majadly, A.; Gal, K.; Puzis, R.; Felner, A. New Goal Recognition Algorithms Using Attack Graphs. In *Cyber Security Cryptography and Machine Learning, Proceedings of the Beer-Sheva, Israel, 27–28 June 2019*; Dolev, S., Hendler, D., Lodha, S., Yung, M., Eds.; Springer: Cham, Switzerland, 2019; pp. 260–278.
17. Cheng, X.; Zhang, J.; Chen, B. Cyber Situation Comprehension for IoT Systems based on APT Alerts and Logs Correlation. *Sensors* **2019**, *19*, 4045. [CrossRef]
18. Kim, D.; Shin, D.; Shin, D.; Kim, Y.H. Attack Detection Application with Attack Tree for Mobile System using Log Analysis. *Mob. Netw. Appl.* **2019**, *24*, 184–192. [CrossRef]
19. Shinde, A.; Doshi, P.; Setayeshfar, O. Cyber Attack Intent Recognition and Active Deception Using Factored Interactive POMDPs. In Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS'21, Richland, WA, USA, 3–7 May 2021; pp. 1200–1208.
20. Zhang, X.; Zhang, H.; Li, C.; Sun, P.; Liu, Z.; Wang, J. Network Attack Intention Recognition Based on Signaling Game Model and Netlogo Simulation. In Proceedings of the 2021 International Conference on Digital Society and Intelligent Systems (DSInS), Chengdu, China, 3–4 December 2021; pp. 162–166. [CrossRef]
21. de Mendonça, R.R.; de Brito, D.F.; de Franco Rosa, F.; dos Reis, J.C.; Bonacin, R. A framework for detecting intentions of criminal acts in social media: A case study on twitter. *Information* **2020**, *11*, 154. [CrossRef]
22. Li, T.; Liu, Y.; Liu, Y.; Xiao, Y.; Nguyen, N.A. Attack plan recognition using hidden Markov and probabilistic inference. *Comput. Secur.* **2020**, *97*, 101974. [CrossRef]
23. Abarna, S.; Sheeba, J.I.; Jayasrilakshmi, S.; Devaneyan, S.P. Identification of cyber harassment and intention of target users on social media platforms. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105283. [CrossRef] [PubMed]
24. Bokolo, B.G.; Onyehanere, P.; Ogegbene-Ise, E.; Olufemi, I.; Tettey, J.N.A. Leveraging Machine Learning for Crime Intent Detection in Social Media Posts. In *International Conference on AI-generated Content*; Zhao, F., Miao, D., Eds.; Springer Nature: Singapore, 2023; pp. 224–236.
25. Navalgund, U.V.; Priyadharshini, K. Crime Intention Detection System Using Deep Learning. In Proceedings of the 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), Kottayam, India, 21–22 December 2018; pp. 1–6. [CrossRef]
26. Pang, R.; Zhang, X.; Ji, S.; Luo, X.; Wang, T. AdvMind: Inferring Adversary Intent of Black-Box Attacks. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD'20, New York, NY, USA, 6–10 July 2020; pp. 1899–1907. [CrossRef]
27. Zhao, J.; Liu, X.; Yan, Q.; Li, B.; Shao, M.; Peng, H.; Sun, L. Automatically predicting cyber attack preference with attributed heterogeneous attention networks and transductive learning. *Comput. Secur.* **2021**, *102*, 102152. [CrossRef]
28. Martínez-Mascorro, G.A.; Abreu-Pederzini, J.R.; Ortiz-Bayliss, J.C.; Garcia-Collantes, A.; Terashima-Marín, H. Criminal Intention Detection at Early Stages of Shoplifting Cases by Using 3D Convolutional Neural Networks. *Computation* **2021**, *9*, 24. [CrossRef]
29. Tang, Q.; Chen, H.; Ge, B.; Wang, H. AIGCN: Attack Intention Detection for Power System Using Graph Convolutional Networks. *J. Signal Process. Syst.* **2022**, *94*, 1119–1127. [CrossRef]
30. Hsu, T.; Tang, C. Detection of Malicious Activities Using Machine Learning in Physical Environments. In Proceedings of the 2022 International Conference on Computational Science and Computational Intelligence (CSCI), Los Alamitos, CA, USA, 14–16 December 2022; pp. 1047–1052. [CrossRef]
31. Kang, J.; Yang, H.; Zhang, Y.; Dai, Y.; Zhan, M.; Wang, W. ActDetector: A Sequence-based Framework for Network Attack Activity Detection. In Proceedings of the 2022 IEEE Symposium on Computers and Communications (ISCC), Rhodes, Greece, 30 June–3 July 2022; pp. 1–7. [CrossRef]
32. Tsinganos, N.; Fouliras, P.; Mavridis, I. Applying BERT for Early-Stage Recognition of Persistence in Chat-Based Social Engineering Attacks. *Appl. Sci.* **2022**, *12*, 12353. [CrossRef]
33. Bhugul, A.M.; Gulhane, V.S. Novel Deep Neural Network for Suspicious Activity Detection and Classification. In Proceedings of the 2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 18–19 February 2023; pp. 1–7. [CrossRef]
34. Tsinganos, N.; Fouliras, P. Leveraging Dialogue State Tracking for Zero-Shot Chat-Based Social Engineering Attack Recognition. *Appl. Sci.* **2023**, *13*, 5110. [CrossRef]
35. Ahmed, A.A.; Ahlami, N.; Zaman, K. Attack Intention Recognition: A Review. *Int. J. Netw. Secur.* **2017**, *19*, 244–250. [CrossRef]
36. Jesson, J.; Matheson, L.; Lacey, F.M. *Doing Your Literature Review: Traditional and Systematic Techniques*; SAGE Publications Ltd.: London, UK, 2011.
37. Okoli, C.; Schabram, K. A guide to conducting a systematic literature review of information systems research. *Sprouts Work. Pap. Inf. Syst.* **2010**, 10. [CrossRef]

38. Caulley, D.N. Conducting research literature reviews: From the internet to paper. *Qual. Res. J.* **2007**, *7*, 103–105. [CrossRef]
39. 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. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
40. Al-Dhaqm, A.; Ikuesan, R.A.; Kebande, V.R.; Razak, S.A.; Grispos, G.; Choo, K.K.R.; Al-Rimy, B.A.S.; Alsewari, A.A. Digital Forensics Subdomains: The State of the Art and Future Directions. *IEEE Access* **2021**, *9*, 152476–152502. [CrossRef]
41. Arshad, H.; Jantan, A.; Omolara, E. Evidence collection and forensics on social networks: Research challenges and directions. *Digit. Investig.* **2019**, *28*, 126–138. [CrossRef]
42. Calegari, R.; Ciatto, G.; Denti, E.; Omicini, A. Logic-based technologies for intelligent systems: State of the art and perspectives. *Information* **2020**, *11*, 167. [CrossRef]
43. Kraft, D.; Moloney, C. *Introduction to Artificial Intelligence*; Springer International Publishing AG: Geneva, Switzerland, 2016; pp. 1–6. [CrossRef]
44. Marques-Silva, J. Logic-Based Explainability in Machine Learning. In *Reasoning Web. Causality, Explanations and Declarative Knowledge* ; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2023; Volume 13759, pp. 24–104. [CrossRef]
45. Geib, C.W.; Goldman, R.P. Plan recognition in intrusion detection systems. In Proceedings of the DARPA Information Survivability Conference and Exposition II. DISCEX'01, Anaheim, CA, USA, 12–14 June 2001; Volume 1, pp. 46–55. [CrossRef]
46. Dodge, J.; Liao, Q.V.; Zhang, Y.; Bellamy, R.K.E.; Dugan, C. Explaining models: An empirical study of how explanations impact fairness judgment. In Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI'19, New York, NY, USA, 17–20 March 2019; pp. 275–285. [CrossRef]
47. Xu, F.; Uszkoreit, H.; Du, Y.; Fan, W.; Zhao, D.; Zhu, J. Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In *Natural Language Processing and Chinese Computing: Proceedings of the 8th CCF International Conference, NLPCC 2019, Dunhuang, China, 9–14 October 2019*; Tang, J., Kan, M.Y., Zhao, D., Li, S., Zan, H., Eds.; Springer: Cham, Switzerland, 2019; pp. 563–574.
48. Maratsi, M.I.; Popov, O.; Alexopoulos, C.; Charalabidis, Y. Ethical and Legal Aspects of Digital Forensics Algorithms: The Case of Digital Evidence Acquisition. In Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance, ICEGOV'22, New York, NY, USA, 4–7 October 2022; pp. 32–40. [CrossRef]
49. Jinad, R.; Gupta, K.; Simsek, E.; Zhou, B. Bias and fairness in software and automation tools in digital forensics. *J. Surveill. Secur. Saf.* **2024**, *5*, 19–35. [CrossRef]
50. Felzmann, H.; Fosch-Villaronga, E.; Lutz, C.; Tamò-Larrieux, A. Towards Transparency by Design for Artificial Intelligence. *Sci. Eng. Ethics* **2020**, *26*, 3333–3361. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.