JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 07 Issue 02
2024

Research Article
https://doi.org/10.56979/702/2024

Exploring Phishing Attacks in the AI Age: A Comprehensive Literature Review

Muhammad Saeed Liaqat[1*], Gohar Mumtaz[1], Nazish Rasheed[1], and Zeeshan Mubeen[2]

[1]Faculty of Computer Science and Information Technology, Superior University, Lahore, 54000, Pakistan.
[2]Riphah International University, Lahore, 54000, Pakistan.
[*]Corresponding Author: Muhammad Saeed Liaqat. Email: saeedliaqat0786@gmail.com

Received: May 21, 2024 **Accepted:** August 19, 2024 **Published:** September 01, 2024

Abstract: Over the years, phishing attacks have also evolved to more difficult types of phishing such as spear-phishing and clone-phishing. By nature, these attacks target not only human but also technical vulnerability resulting in massive financial losses and data exposure. So the more complex these methods become – such as phishing, cyberbullying or Packet Sniffing — offers in need new cybersecurity protocols to get rid of those threats. We have done this study by using Kitchenham Systematic Literature Review (SLR) framework, which consists of three phases: planning; conducting and reporting. These programs were reviewed because of the increased use in AI oriented operations such as financial phishing attacks, with new difficulties for detection systems. Furthermore, the study undertook extensive database searches on computers routines like IEEE Access, Research Gate and Google Scholar rich in recent scientific studies. Methods: A two-phase screening process rigorously identified 20 high quality articles out of an initial pool of 250 studies for further analysis. The results are a good demonstration that AI-driven phishing campaigns continue to evolve in complexity, and in many cases can be more difficult to spot. Moreover, the current AI-based detection systems are mainly not claimed fully secured as they can easily tricked through adversarial attacks which may needs to updated or refine time after one another. The research indicates that combining contextual and behavioral investigation may improve the ability to detect threats as they take place. In addition, it is advisable to deploy a multi-layered security approach that combines traditional AI methodologies with human oversight through machine learning for more effective threat detection and prevention. The research highlights the need for preventative security strategies and continued detection innovations against ever more sophisticated phishing campaigns.

Keywords: Phishing Attacks; Cyber Security; Artificial Intelligence; Threats Detection; Machine Learning.

1. Introduction

Many people and organizations are highly affected by phishing attacks that have become a major and serious problem in the digital world. These cyberattacks leverage human and technical weaknesses, often leading to significant financial losses and exposing sensitive data. The term "phishing" was coined in the mid-1990s after hackers breached America Online (AOL) accounts by using social engineering methods. Over time, phishing has grown into an even more sophisticated threat serving as a vehicle for various communication channels including email, SMSs, social media networks even voice calls purposed on hoodwinking victims into sharing protected information or even downloading malware **[1]**. This paper discusses how phishing attacks have changed over time, the techniques used by cyber criminals and how they can be combated.

Phishing is simply a combination of social engineering and exploiting technology. In order to get sensitive information from users, cybercriminals often pretend to be reliable bodies such as financial institutions, online payment companies or social networking sites. For this purpose, they create counterfeit websites that are virtually indistinguishable from real ones. According to APWG (Anti-Phishing Working

Group), throughout 2014 the number of phishing incidents has risen significantly with around 255,000 new threats being detected per day during the peak [1]. Such increase in phishing cases emphasizes on the importance of having strong mechanisms for detection and prevention.

Adaptability and continuous evolution are some of the key characteristics of phishing attacks. Phishing techniques used by hackers include clone phishing, spear-phishing, and malware-based phishing among others. Spear-phishing is an approach that focuses on a particular group or individual where attackers gather comprehensive information to come up with more convincing phishing emails [2]. Such emails usually pretend to be from valid sources and they are personalized for recipient thus increasing chances of success. Clone-Phishing on the other hand involves making a copy of a legitimate email received previously by the target but replacing its contents with malicious links or attachments [3]. These sophisticated ways show how much it might be difficult in detecting and preventing phishing attacks.

Phishing's impact goes beyond individuals to include organization and also national security. When phishing is done on a large scale, it causes immense financial loss, tarnishes the reputation of corporate brands and leads to unauthorized access which interferes with classified data within the companies. Corporate networks are infiltrated through phishing which is used as a tool resulting in data breach leading to compromise of customer information along with intellectual property. Experts in cybersecurity face an uphill task due to global reach of phishing attacks coupled with AI and machine learning use by hackers [4]. The use of AI in creating better strategies has led to automated creation of highly-convincing phishing emails and websites thereby complicating detection efforts further.

Different techniques have been designed to counter the increasing threat posed by phishing. These encompass technological means such as machine learning algorithms for identifying phishing websites and educational activities that increase user knowledge. Phishing URLs can be effectively detected by making use of machine learning models like Random Forest and XGBoost which have been shown to achieve remarkable accuracy [5]. The features about URLs used in these models consist of the age of the domain, whether HTTPS is being used or not, URL length among others, all of which are taken into account when assessing potential threats. User training remains quite important in combating phishing despite the improvement in detection technologies because it helps users identify suspicious email messages and encourages them to adopt security measures such as two-factor authentication that could prevent this type of attacks from happening [5].

In addition, there are browser-based defenses that have been built to shield users on real-time. For example, browser extensions that use artificial intelligence applications can identify and prevent the phishing attempts while users are browsing. These kinds of extensions function by examining the content of a website and then comparing it with previously recognized phishing indicators. Additionally, some of them also include real-time reporting systems which allows one to flag suspicious sites and play part in the overall fight against phishing [6]. The tools have been successful at preventing users from unknowingly navigating into malicious sites and therefore they are useful in protecting against zero-day phishing attacks where attackers adopt new approaches that are not known to people.

But still, phishers and those that protect against them are locked in an arms race that continues to unfold. New techniques are constantly being developed by attackers in order to overcome security systems. Attackers have started using encryption messaging services as well as social media platforms to distribute phishing links making it difficult for conventional detection systems to detect the threat. In fact, personalized phishing attacks have become successful with the use of AI by attackers especially in spear-phishing campaigns [7]. Such a development highlights the necessity of constantly innovating anti-phishing solutions and approaches.

Combating phishing through AI and machine learning is one of the most promising developments, not just for detection but also predictive analytics. By going through extensive data sets of phishing attempts, AI models can recognize constancies in them and forecast threats that are to be encountered later by the organization (future attacks) thus protecting it against emerging dangers. It is useful in zero-day attacks identification and mitigation because they exploit previous unexplored flaws [8]. AI embeddings into cybersecurity frameworks would probably have the most significant impact on future anti-phishing fight.

The main contribution of this research study is as follows:

- New attack methods, such as cyberbullying and packet sniffing, introduce challenges that require updated cybersecurity protocols.
- The study focuses on the growing role of AI in phishing, which introduces new difficulties for detection systems.
- The study recommends combining traditional AI methods with human oversight for a more effective, multi-layered defense against phishing attacks.

2. Literature Review

Phishing attacks have changed a lot since their birth, and the tactics used by hackers have become more sophisticated, making it difficult to identify them. The literature on phishing contains full descriptions of different types of attack, how cyber criminals adapt their methods and strategies designed to deal with these perils. By discussing the results from recent research this survey shows what is happening now in terms of combating phishing.

Phishing was originally known as a form of internet fraud that deceptively collects sensitive information and is now one of the most widespread security issues in modern times. As shown in Figure 1, at first, early phishing mainly targeted email as a means of delivery whereby attackers would imitate credible persons or organizations to manipulate users into revealing personal data about themselves. Nonetheless, phishing has gone beyond just emails as other channels like SMSs, social media platforms or even voice calls are involved thus it is a complex issue whose prevention must be holistic **[1]**. An upward trend in its occurrence is evidenced by Anti-Phishing Working Group (APWG) reports which indicate an escalating incidence of phishing cases where hundreds of thousands new threats emerge every day **[1]**.

Social engineering is a staple of phishing — threats trick their victims into acting on behalf! Threat actors create highly-convincing emails and sites pretending to be real companies, essentially taking advantage of the trust people have in these organizations. Because it is a targeted kind of phishing, spear-phishing has been very potent. Hackers use specific information about the victim via cookie steal, passphrase dump etc and send phishing message according to it. In fact, a spear phish is one of the hardest forms to catch as it bypasses traditional security methods [2]. Moreover, attackers using clone phishing as shown in Figure 2 — where malicious emails contain the same visual elements as legitimate ones and point only to their own installation file instead of compromising a website [3]— make it even harder for detection techniques.

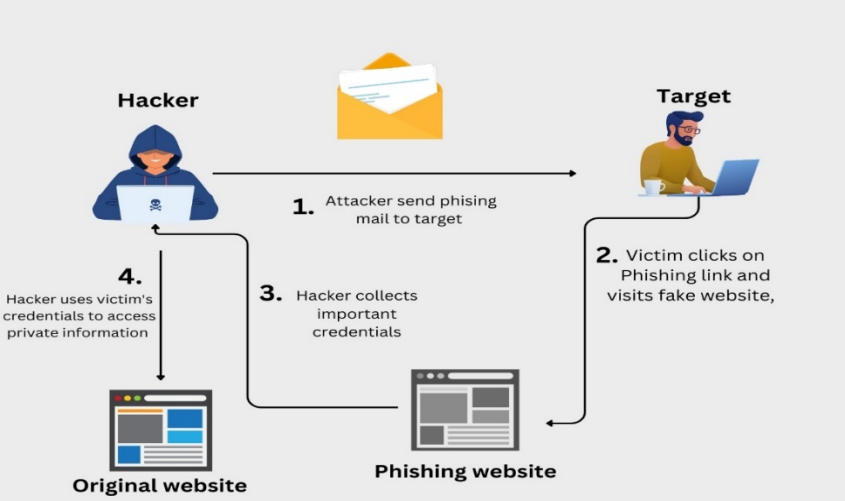

Figure 1. How is a phishing attack done?

Phishing affects not only safe user habits but also as a threat to organizations and even national security. One of the leading reasons for data breaches is phishing, where attackers gain unauthorized access to company networks by tricking staff members into revealing login credentials. When breaches occur, the potential costs are high in multiple dimensions (financial loss and reputational hit among them). On the other hand, phishing attacks are both global and pervasive [4], with adversaries using machine learning tools for at considered phase of an attack lifecycle presenting a challenging task to those in cybersecurity. Phishing attacks have gotten more sophisticated too — with AI, attackers can now build convincing phishing emails and websites by automating the post-delivery process as well.

In the recent past, a number of studies were carried out to analyze effectiveness of different anti-phishing techniques from both technological and educational perspective. Phishing Phases: Machine learning is one the effective means for phishing detection, as it can easily analyze bulk data set to extract patterns which are indicative of being a phishing. Random Forest, XGBoost and other deep learning models are built on it that have high efficiency in identifying phishing URLs. These models take into account features such as domain age, whether the site uses HTTPS and URL length to predict the chance of a website being malicious [5]. But, of course the models are only as good as their data (read here for more on the issues associated with synthetic ML), and they need monitoring to keep them from being hijacked by newer types of phishing.

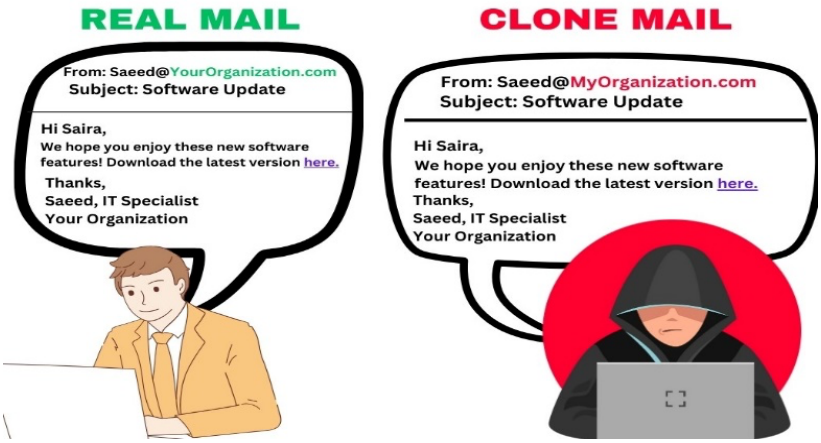

Figure 2. Protect Yourself: Distinguishing Between Real and Fake Emails

Besides technological solutions, user education has been a central theme in combating phishing. According to research, tech-savvy users are harder targets for phishing because they can easily distinguish between a legitimate email and one from the attackers; Phishing awareness training programs, in combination the use of security features like two factor authentication can greatly reduce such a risk [6]. Human factors are a significant weakness despite advances in detection technologies; thus, constant education and awareness efforts should be the highest priority.

Real time protection against phishing continues with browser-based defenses. Such defenses can also encompass ML-backed browser extensions, capable of analyzing web traffic and alerting users to potential phishing attempts. Behind the scenes such extensions are using content analysis of websites and matching that against known indicators for phishing. A number of extensions also have real-time reporting capabilities, so users can flag malicious sites and do their part to keep the entire ecosystem safe from phishing [6]. These are very useful features that can protect users from inadvertently accessing malicious sites, and they serve as a unique tool form of defense against zero-day phishing attacks where attackers use new, undocumented methods.

Phishers are also constantly evolving, and it's fair to say that this is an arms race where there will always be more techniques phishers try out as new defenses come into play. There has been a shift to encrypt messages services and social networks platforms where the phishing links are shared among users which makes detection by common security solutions more difficult. The fact that the attackers make use of AI for sending personalized phishing emails is therefore contributing to an even greater percentage that spear-phishing attacks work [7]. This shows the merits of on-going innovation in anti-phishing methods and technologies.

AI and machine learning holds one of the most potential advancements against phishing not just for detection but also during prediction analytics. The application of artificial intelligence can address these issues, as AI is capable of sorting through massive volumes and varieties of phishing attacks to detect patterns originating with a subset, identifying similar characteristics in future attacks that will help organizations including front line defenders stay ahead. An advance feature of this method come in detecting and preventing zero-day attacks, which uses new vulnerabilities [8]. AI is anticipated to be a major in cybersecurity frameworks and the future of phishing defense.

In addition, many studies proposed frameworks and guidelines to improve the detection rate of phishing systems. These frameworks typically employ a mixture of machine learning, user education and policy enforcement to create an in- depth defense against phishing. For example, studies have proposed to use behavioral analysis in addition to traditional phishing detection methods [3], so that know happening at the impersonated end by identifying a link between certain user behavior and receiving fraudulent messages or using malicious websites instead of utilizing some attributes from the content of an e-mail because cyber criminals always find ways how their classic tools become ineffective (e.g. protection mechanisms are implemented into up-to-date Web browsers). It is a supplementary security-measure to protect you from advanced phishing methods that sometimes get through normal detection systems.

3. Materials and Methods

We performed a systematic literature review (SLR) following the framework established by Kitchham [9]. This methodology comprises planning, implementation, and dissemination phases, each with various tiers. As illustrated in Figure 3, the SLR involves three primary steps. The subsequent sections will outline each of these steps.

3.1. Phase 1: Planning the review:

Meanwhile, we determined the main purpose of the analysis and performed the following tasks, explaining each step in detail.

3.1.1. Identification of the need for a review

Rapid rise observed in the recent years with phishing activities being recorded by multiple statistics points out a clear need to over haul current detection strategies and defense mechanisms. In 2021, saw SlashNext cataloging almost 50,000 new URLs daily as part of its phishing detection efforts in a demonstration of increasingly intricate and elaborate phishing attacks. This number surged to 80,000 URLs per day in 2022 which is an increase of up to 61% leading us at a huge xx phishing attacks detected last year [10]. The detection of 870,555 credential phishing links (compared to 2022—and a rising trend) was possible also using advanced machine-learning and image analysis technologies [12]. These disturbing statistics prove that phishing threats are both dynamic and constant in evolution, thus demanding a systematic review regarding the effectiveness assessment of existing methodologies for detection, identification of gaps in current research, and investigation into new technological innovations for strengthening cybersecurity defenses.

PHASE 1: PLANNING THE REVIEW	PHASE 2: CONDUCTING THE REVIEW	PHASE 3: REPORTING THE REVIEW
• Identification of the need for a review • Specifying the research question(s) • Identifying the relevant bibliographic databases	• Identification of research • Selection of primary studies • Extraction of data • Synthesis of data	• Specifying dissemination mechanism • Formatting the main report

Figure 3. Systematic literature review phases and activities.

3.1.2. Specifying the research question(s)

This huge increase in phishing attacks, especially in regard to AI-driven cyber threats, puts pressure on researchers and professionals working for cybersecurity. As methods of phishing continue to evolve and include more sophisticated AI and machine learning algorithms, it gets way harder to detect and prevent such attacks. While AI technologies improved threat identification and counteracting, their deployment added a host of new vulnerabilities that can be exploited by adversaries. The current situation requires a systemic review of the literature to estimate the effectiveness of AI-based detection methods and explore emerging trends in phishing attacks that pinpoint areas where existing research may be insufficient. This review addresses this gap between the current state of knowledge and the rapidly

evolving threat landscape, setting a foundation for future research so as to enhance defenses against AI-powered phishing attacks.

1. How could AI-powered phishing attacks, in which the artificial intelligence has so thoroughly learned human communications patterns that they are utterly indistinguishable from them, be successfully detected without an unacceptable level of false positives?
2. To what extent do current AI-based phishing detection systems have vulnerabilities that adversarial attacks could take advantage of, and how can this be mitigated?
3. How much can AI and Machine Learning help in not just detecting phishing but actually anticipating and preventing new phishing strategies, even before they might be used in the wild?
4. In what ways does the incorporation of AI into the methods of phishing attacks alter the traditional lifecycle of phishing campaigns, and what has this implications for real-time threat detection?

3.1.3. Identifying the relevant bibliographic databases

According to the research questions, the following digital libraries were searched for the necessary articles: Google Scholar, PubMed, Science Direct, Springer, and Scopus. These digital libraries were selected primarily because they compile studies in the fields of computer science and medical science, indexing articles from a variety of publication outlets, including journals, conferences, books, and workshops. In this study, the search was restricted to journal articles and conference proceedings published between 2010 – 2024.

3.2. Phase 2: conducting the review

For this systematic literature review on "Phishing Attacks in the Age of AI," a rigorous, structured approach will be employed, guided mainly by well-established methodologies like PRISMA. The detailed search will begin in academic databases such as IEEE Access, Google Scholar and ResearchGare, among others. This research will be based on keywords and phrases that best describe artificial intelligence, phishing attacks, machine learning, and cybersecurity. The selection of publications within the last decade will be based on the inclusion criteria restricted to peer-reviewed journals, conference papers, and technical reports in order to retain relevance to existing technological changes. In the final stage, a two-step screening process will be conducted on all identified publications after the literature search. First, they will be checked for relevance based on titles and abstracts, followed in the second stage by a full-text review to determine whether the studies meet the objectives of the research. Data extraction will then be done to capture systematically information from every study concerning the kind of methodologies used, findings obtained, and gaps identified by the authors. The review will synthesize the extracted data for common themes, trends, and areas where further research is needed to provide a deeper understanding of how AI is shaping the phishing attack landscape and defenses.

3.2.1. Study Selection

Selection of studies for this systematic literature review on "Phishing Attacks in the Age of AI" will involve critical steps to ensure that only high-quality and pertinent research is being taken into account. First of all, there will be a broad search across primary academic databases such as IEEE Access, ResearchGate, and Google Scholar with the outlined keywords: "phishing attacks," "AI," "machine learning," "cybersecurity," and "threat detection" [10]. It will, therefore, be limited to only peer-reviewed articles, conference papers, and technical reports published within the past decade or so, representing the newest develop.

The selection process will proceed with a two-phase screening. First, the researchers will screen the titles and abstracts of the retrieved articles, excluding studies that have no direct relation to AI-driven methods for phishing detection or prevention. The second stage will involve a full-text review of the remaining articles, based on their relevance to the objectives of the review, considering only studies reporting empirical data, introducing new methodologies, or making a critical review of existing AI-based methods for the detection of phishing [11].

3.2.2. Quality assessment

In the systematic literature review entitled "Phishing Attacks in the Age of AI," there will be a strict quality assessment ensuring that only studies of high quality are to be included, producing reliable and meaningful insights. Each of the selected studies will be graded in relation to some predefined criteria like clarity and relevance of the question, solidity of methodology, and validity of the findings. In this respect, more focus will be devoted to studies that apply the empirical method. For example, studies that deal with

real-life datasets and implement AI-based models for phishing detection are most likely to provide actionable insights [5].

The quality assessment will also look at the transparency of the data analysis process and at the possibility of replicating the results, which are important in making sure that the findings can actually be taken up with confidence to inform future research and practice. Studies found with clear failures in these criteria, such as using uncritical methods or with an insufficiency of data supporting conclusions, are best excluded from the review. In short, the current review operationalizes the abovementioned stringent process of assessment on the literature to synthesize only the most solid and relevant research in providing a sound understanding of both the current state of AI in phishing detection and its main areas of further exploration.[6].

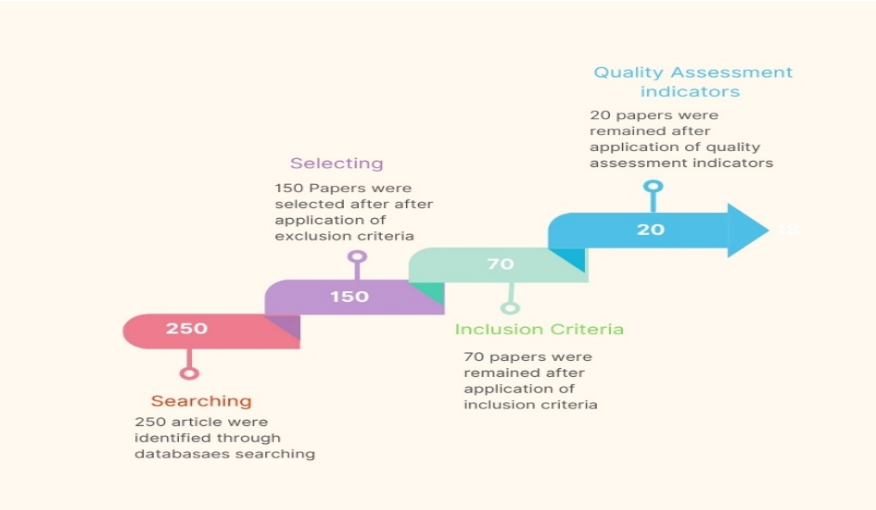

Figure 4. Flow diagram regarding the systematic review, inclusion and exclusion of studies in this review

3.3. Data Extraction and Synthesis

A data extraction form was used to collect relevant information from the selected articles to answer the research questions.

RQ1: To answer this research question how to detect highly sophisticated AI-powered phishing attacks that closely mimic human communication patterns, without generating excessive false positives.

RQ2: To answer this question, the vulnerabilities in current AI-based phishing detection systems to adversarial attacks and how these vulnerabilities can be mitigated.

RQ3: The question explores how AI and Machine Learning can be used to anticipate and prevent new phishing strategies before they are deployed.

RQ4: To answer this question investigates how AI changes the traditional lifecycle of phishing campaigns and its impact on real-time threat detection.

Different strategies were adopted to generate the data obtained to answer the research questions. A descriptive linking approach was generally used to answer the research questions. In addition, visual aids such as tables and diagrams are also used as research questions.

4. Results

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

This section presents the results obtained during this review (Phase 3: reporting the overview). First, we provide an overview of the results of the selection process; then, we show all the results separately for each research question.

4.1. Overview of the selected studies

By searching five databases we identified 250 matching files, as shown in Figure 4. An additional 150 files were analyzed in detail to identify as many relevant data as possible. After reviewing titles, abstracts, and summaries, only articles that met at least one of the criteria were included. Finally, 100 sentences remained. At the end of this stage, the quality of the selected text will be examined as a whole; finally, 20 articles were used to extract data and provide answers to the research questions in this SLR. Figure 5 demonstrates the distribution per years of the 20 selected papers.

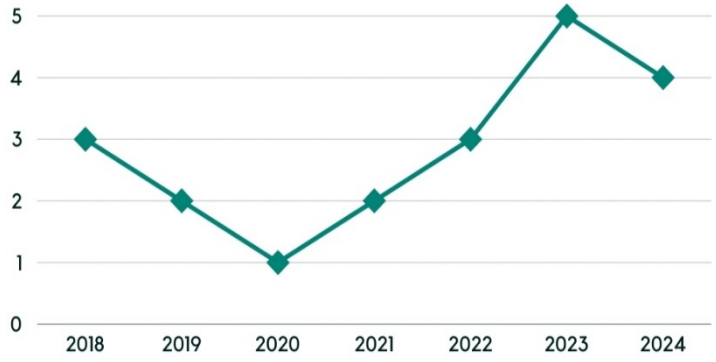

Figure 5. Distribution of articles by publication year

4.2. Phase 3: Reporting the Review

RQ1: How could AI-powered phishing attacks, in which the artificial intelligence has so thoroughly learned human communications patterns that they are utterly indistinguishable from them, be successfully detected without an unacceptable level of false positives?

AI completely imitates communication patterns of human and there are really few possibilities to find this phish among legitimate communications. These sophisticated phishing attacks use machine learning models built on huge amounts of real human reactions to make them seem more genuine and less detectable with simple detection like Static File analysis. But a multi-dimensional approach leveraging deep AI, contextually processing and behavioral analysis on top of continuous learning has the potential to improve detection but in minimizing false positives.

4.2.1. Advanced AI and Machine Learning Models

Anomaly Detection: Current phishing detection approaches are based on looking for what we know and exception handling. These patterns are more hidden in the area of AI-powered phishing. In simple terms, machine learning models (with a focus on deep learning and natural language processing or NLP especially) can be trained to distinguish between minor variances in the patterns of communication that even some state-of-the-art AI might not perfectly mimic. They detect the smallest deviations from normal human conversation; it may be an error in syntax, phrasing or context.

Contextual Awareness: Integrating powerful and faster detection patterns within the AI models by including contextual understanding For example, the system will know that you have sent a similar SMS in the past. If such an SMS comes from your number all of a sudden but isn't what I expect at this time when compared to my history with replying and interacting like me (based on data), then it can flag as being suspect even if done well, maybe better than just by rules.

4.2.2. Behavioral Analytics

User Behavior Analysis: By contrast, one strategy is to observe a user's behavior over time and build inferences about what constitutes standard activity. Phishing attack triggers: Even if the communication looks legitimate, launching a phishing attack may activate unusual behaviors like following unknown links; asking strange transactions and provide private data in nonstandard contexts when these departures from normal behavior occur, if also a deviated pattern of the critical system output is noticed AI will highlight it out that requires human attention.

Historical Data and User Profiling: Historical data analysis to create users' profile can also be used in detecting AI driven phishing attacks. By noticing subtle variances in calls with current interactions given past communications, systems can recognize suspicious behavior where ill-intentions may be at play. Alternately, if a user hasn't requested certain types of requests suddenly makes this type of request, it might not be consistent with their overall behavior.

4.2.3. Multilayered Defense Mechanisms

Combining Multiple Detection Layers: Using a layered approach including AI along with traditional detection approaches, such as heuristic analysis, rule-based filters, and blacklists to increase the accuracy

of overall detections. AI can pick up advanced threats, but traditional approaches will handle basic volume phishing (and help prevent the AI from being overloaded and stop any false positives).

Human-AI Collaboration: Human analysts can be roped in to alleviate false positives wherever the detection systems are not confident. AI has the ability to detect suspicious items and needs a human touch for getting things out of them. The division of responsibilities is important for maintaining the human-AI collaboration in which high-stakes decisions — such as blocking communications or other defensive measures — are not determined purely by automated systems.

4.2.4. Continuous Learning and Adaptation

Adapting to New Threats: As tactics of phishing are now quickly evolving with AI. With continuous learning models that are regularly updated with fresh data, the algorithms can grow and improve exponentially. Integrating feedback loops, in the form of threat data sampling or other mechanisms where identified threats are examined and used to update models so as to anticipate new varieties of attacks before they even materialize.

User Feedback and Reporting: By incentivizing users to report suspicious communications, you can use this data for fine-tuning the alert/detection systems if end-users report phishing signals, information will be collected to feed machine learning models and prevent false positives from occurring in the future.

4.2.5. Reducing False Positives

Threshold Tuning and Confidence Scoring: On the other hand, we can tune detection thresholds and confidence scoring in order to deal with false positives. Detect Threats With Probability Grades: AI systems determine the confidence for each detected threat to decide on an appropriate response.(utils) This will also help in minimizing false positives and alerts for cases when the system is less confident, warnings can be shown or a higher level of verification may be needed instead of just blocking.

Feedback Integration: Having feedback mechanisms which allow users to confirm or dismiss potential phishing detection can help tighten the screws. This feedback over time aids in the reduction of false positives and overall system accuracy.

RQ2: To what extent do current AI-based phishing detection systems have vulnerabilities that adversarial attacks could take advantage of, and how can this be mitigated?

Extent of Vulnerabilities: Vulnerabilities in AI-based phishing detection systems could be exploited by adversarial attacks. Specifically, these are attacks where the input into an AI model is intentionally modified such that it returns incorrect outputs via certain attack vectors — for our case here: allowing phishing emails through. Adversarial attacks are potent because they can modify phishing content in subtle ways—say, by slightly changing a few pixels within an image or adjusting a couple of words throughout the body text of an email—which cause enough confusion to fool AI models but go unnoticed by human users.

Adversarial Examples: Attackers have found this to be one of the main weak points in AI detection systems. These inputs are adversarial examples to the model, crafted with a fine-tuned design to let the model make an error. For example, a phishing URL can end up with ever slight modifications that are still not perceptible to the humans however suffice for an AI model to differentiate. However, such adversarial attacks could considerably lower the accuracy and credibility of phishing detection systems — in particular if an AI model only has experienced a small number of test cases [4].

Overfitting to Training Data: A second type of vulnerability is overfit to its training examples. This type of model is more static in nature one can say, as however effective it may be on detecting the phishing tactics we already know about, when facing a modified or previously unknown technique that has been through our traditional defenders and managed to pass this phase an ML based approach might recognize things differently. Attackers can move in on this by designing phishing attacks slightly different than the ones we know examples of it will not set off detection as spam [12].

Lack of Contextual Understanding: AI models, especially ones based on machine learning simply do not have the understanding of what this content means. Attackers can take advantage of this shortcoming to engineer phishing attacks that are in some cases contextually nuanced yet equally deceiving. A common type of these attacks involves phishing emails designed to closely model the tone, style and content of legitimate communications: human operators tasked with social engineering can write exceptionally convincing messages that are difficult for AI models to spot — especially if they only use superficial cues (e.g., based solely on keyword matching or URL structure) as opposed parsing conduct analysis [11].

Mitigation Strategies: Several strategies can be adapted to strengthen AI-based phishing detection systems against vulnerabilities.

Adversarial Training: Adversarial training is one of the most powerful methodologies to mitigate adversarial attacks. A new angle on data augmentation, this takes the approach of injecting adversarial examples directly into the training set so that it can learn what an adversary looks like (to be able to classify them properly). Training the model with a broader array of possible attack strategies helps teach it to better handle such attacks in real life [6].

Ensemble Learning: Security measures like using an ensemble of different machine learning models can also contribute to reducing the risk of adversarial attacks. In the worst case, one adversarial attack can fool a model in image classification but if we pool predictions from many models — some of which are good at avoiding such attacks and others that are not — they reduce the risk of this type of problem. This technique reduces the possibility of a successful cyber-attack on user against which an particular vulnerability has been exploited in one model [8].

Continuous Learning and Updates: The secret in the sauce is Momentum quickly learns from new data to continuously update and retrain your AI phish-detection model. This includes leveraging data on newly detected phishing campaigns, as well as reports (input from end users re: phish), among other things. Their effectiveness over time can be maintained by ensuring that the model remains current with today's detailed modes of threats [1].

Contextual and Semantic Analysis: One can refine his AI models with respect to contextual and semantic-based analysis so that they can be much more formidable against more sophisticated attacks. The model, having the capability to analyze the context and meaning of such content, will give attackers a much harder time evading detection by simply modifying superficial characteristics. Doing this also helps in reducing false positives since the model will filter better between legitimate and malicious content from deeper analysis.

RQ3: How much can AI and Machine Learning help in not just detecting phishing but actually anticipating and preventing new phishing strategies, even before they might be used in the wild?

AI and machine learning are increasingly promising in detection of phishing attacks and still foretelling new phishing strategies before they are even deployed. These technologies can analyze huge amounts of data, recognizing patterns within it that would forecast the threat ahead and, hence, protect from the continuing fight against phishing. How that can be achieved through AI and machine learning, together with challenges and strategies to improve performance, are discussed below.

Predictive Analytics and Pattern Recognition: AI and machine learning models are very adept at analyzing historical data to recognize patterns and trends that may indicate potential future dangers. By investigating earlier instances of phishing attacks, these models will learn common features and tactics used by the attackers, such as popular keywords, URL structures, and email formatting. Armed with such knowledge, it will be quite easy for such a model to predict new strategies for phishing if it sees similar patterns or discovers that some methods have been derived from earlier ones. It can, for example, learn subtle differences in emails or web pages to recognize a new phishing attack technique—thereby developing countermeasures before the attack spreads [5].

Zero-Day Attack Detection: Identifying zero-day attacks, which leverage newly-discovered vulnerabilities, is a challenging problem in cybersecurity. With that sort of attack getting more and more sophisticated all the time, we see many developments in AI-based systems to detect such attacks by the odd one out (anomalies) or unusual behavior compared with norm. A prime example would be training a machine learning model to recognize when the behavior of an otherwise normal website starts taking on attributes commonly seen in phishing sites — such as selecting confidential information under abnormal circumstances. As AI can recognize these inconsistencies sooner, it can also give security teams a head up signal to detect possible threats before they become full-blown attacks [6].

Adaptive Learning and Continuous Improvement: Phishing detection machine learning models are capable of evolving as they can be trained with new data to improve their performance over time. Especially with adversaries that are always coming up with new evasions, this is key. These models can update their algorithms rapidly by including real-time data feeds along with user feedback in order to identify and respond new malicious attack vectors. The system adaptability in learning, makes the AI systems still work by preventing new types of phishing to have a successful effect [12].

Behavioral Analysis and Contextual Understanding: AI can also look at the content on emails and websites, as well analyze user behavior and context where interactions happen. For instance, an AI system might monitor the typical ways that someone communicates and identify anything out of the ordinary as potentially suspicious or even a phishing attempt; anytime an email comes from someone you know, but the wording or request is strange / odd, even if it passes all other checks that traditional phishing detection and prevention would check for. Lead with the knowledge that this "contextual understanding" creates another layer of security which must add complexity into an attacker's phishing strategy [13].

Proactive Threat Hunting: It can involve AI and machine learning for a proactive threat hunting, where it will intelligently search through the data manipulation to find potential threats before they get to your users. This could include searching the web for similar domains that have been recently registered or lurking around dark web forums looking to see if anyone is talking about new phishy techniques. Early detection of these threats allows security teams to respond proactively, such as adding specific malicious domains / alerting users on some predefined methods-based attacks [4].

Challenges and Limitations: Although there is significant promise in the possibility of AI and machine learning detecting phishing attempts, challenges remain. Ultimately, these technologies are only as accurate as the data upon which they were trained and failed detections or false positives can result from biases in training data that do not accurately reflect reality. Also, advanced attackers could craft adversarial techniques to bypass AI-based detection systems so that a permanent enhancement and evolution of the function model is needed [1].

RQ4: In what ways does the incorporation of AI into the methods of phishing attacks alter the traditional lifecycle of phishing campaigns, and what has these implications for real-time threat detection?

AI has basically transformed the traditional lifecycle of phishing campaigns by engraining the methodologies involved, hence inventing new challenges and opportunities for real-time threat detection. AI improves the effectiveness and efficiency of phishing campaigns through automation and optimization of their respective attack stages, starting from target selection to crafting and delivery of phishing content. The following sections discuss in greater detail how AI impacts a phishing campaign's lifecycle and the implications for real-time threat detection.

1. Automation of Phishing Campaigns:

AI in Phishing campaigns has changed the way what previously were doing manually. For instance, artificial intelligence can be leveraged in scraping data from online profiles and social media activities of probable targets. As a consequence, the data can be employed to make very compelling phishing emails that tailor according to user taste and maximize potential click through rate of victims. This automation of tasks enables attackers to more easily scale their campaigns, impacting a wider range of targets with less effort [5].

2. Enhanced Crafting of Phishing Content:

This is the best example of how attacking tools based on AI algorithms and generating phishing content similar to legitimate communication. The AI would be taught to recognize the patterns, language and design elements in a massive collection of emails (or websites or any other form of communication) which represent legitimate content. This gives access to attackers to generate phishing emails and websites which are almost indistinguishable from the real ones adding more chances of success. Thus, the old techniques to rely solely on differences between legitimate and phishing content for detection are losing their effectiveness [6].

3. Dynamic Adaptation and Learning:

The prior learning and constant improvement capabilities AI systems have over a certain threshold after running their algorithm on new tasks for some time. By leveraging AI-based phishing tools can learn about the responses for past attacks and hence they continuously evolve from their successes as well as failures, which thus help prepare better strategies in the next campaigns. If a certain format of phishing email is successful, the AI can mimic that in future campaigns as well. This aspect of learning makes phishing attacks impermeable to the conventional detection techniques as the AI model has evolved and kept on improving in order to bypass new security blocking [4].

4. Real-Time Phishing Attacks:

Real-time Phishing — automating the process of generating and delivering phishing content at scale with AI. The ability to execute simultaneous, multiple, highly sophisticated phishing attack types is a

significant headache for real time threat detection system authors. AI-driven phishing campaigns are the speed and scale of AI, which also necessitates our detection systems to work faster and with more accuracy than ever before [14].

5. Implications for Real-Time Threat Detection:

There are many challenges in real-time threat detection where AI is integrated into phishing campaigns.

- Increased False Positives and False Negatives:

The complexity of AI-generated phishing or BEC has the potential to raise both false positives (legit email treated as a phish) and negatives (phishing emails outsmart existing detections). This hampers the efficacy of real-time detection systems, costing user irritation and even breaching [7].

- Evasion of Traditional Detection Methods:

Blacklists and heuristics, which are the more traditional phishing detection methods therefore do not work effectively against AI-powered phishing attacks. But these approaches require recognizing established tricks or aberrations, that AI are able to create phishing materials suitably imitating official communications will be tough for such systems to breakup [1].

- Need for Advanced AI-Based Detection Systems:

This demand has promoted the Artificial Intelligence to detect tasks performed by AI like Phishing attacks, etc. The ability to analyze both content and users' behaviors in addition systems > must have attacks of any, giving context with perception. For instance, AI can be employed to track how users normally behave and identify any anomalies that might signal a phishing operation [13].

- Continuous Learning and Adaptation:

In a similar way that attackers are using AI to learn and evolve so should our detection systems. See how the stakes are higher than ever before, and why ongoing education is crucial to keep up with these increasingly sophisticated AI phishing tactics. So it needs an AI models updated on regular basis and real-time data to be integrated, so that detection systems are effective [15].

4.2. Phishing Attack Detections Using Machine Learning

In 2021, the report noted a substantial increase in phishing activities. By the end of 2021, SlashNext detected approximately 50,000 malicious URLs daily. This reflects a significant volume of phishing attacks being detected, but specific annual totals for 2021 are not directly mentioned.

In 2022, SlashNext detected 80,000 malicious URLs daily, representing a 61% increase from 2021. This increase equates to 255 million phishing attacks detected in 2022 using machine learning and other AI-based detection methods [5].

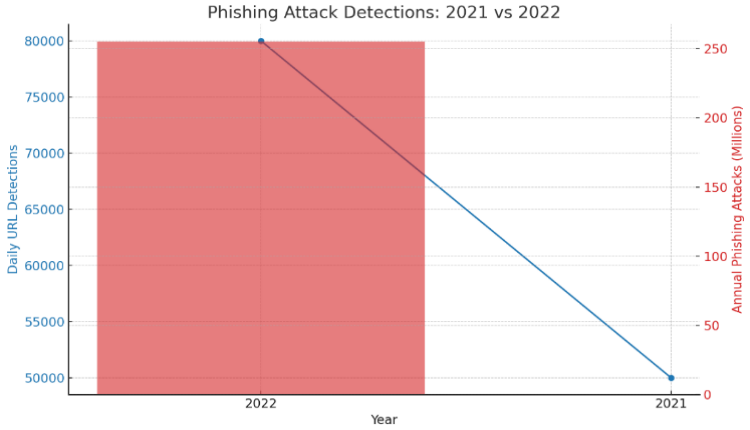

Figure 6. Here is the graphical representation of phishing attack detections for 2021 and 20222

Blue Line (Left Y-Axis): This line shows the daily number of malicious URLs detected. In 2021, approximately 50,000 URLs were detected daily, which increased to 80,000 in 2022. Red Bar (Right Y-Axis): This bar represents the total number of phishing attacks detected annually in millions. In 2022, approximately 255 million phishing attacks were detected.

In 2023, a total of 870,555 credential phishing links were identified using Computer Vision, an advanced image analysis and machine learning (ML) technology designed to detect credential phishing emails by analyzing site content, such as branded elements and login forms. This represents a substantial

increase of 263% compared to 2022, highlighting a significant escalation in the prevalence of such attacks. Shown in Figure 7.

These visualizations offer a clear summary of the performance of machine learning models for phishing detection and the increasing trend of phishing attacks over recent years. No specific experimental results were reported, as the document focused more on discussing existing technologies and their effectiveness in different contexts.

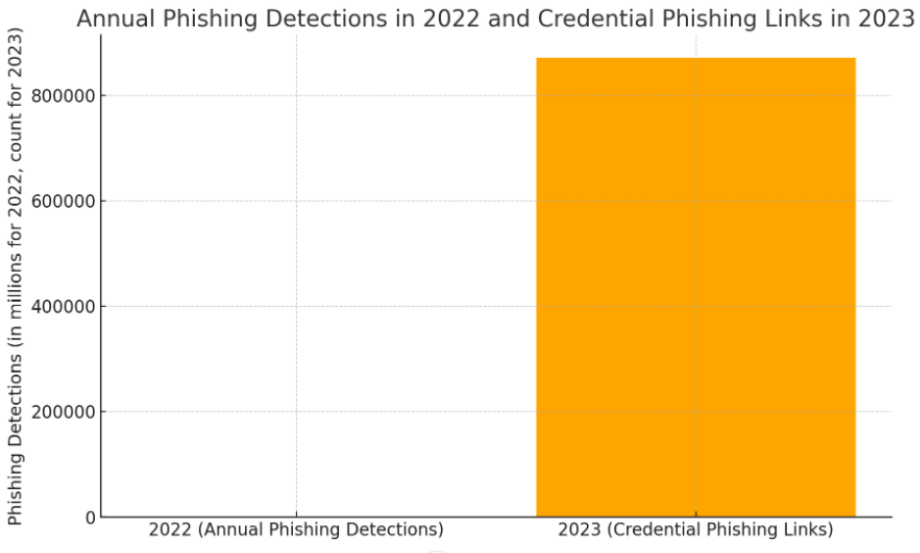

Figure 7. Annual Phishing Detections in 2022 and Credential Phishing Links in 2023

5. Conclusions

In reviewing the current phishing scenario, especially considering AI-backed tricks, we can see that these threats are growing in complexity and escaping common measures. AI-powered phishing campaigns: AI technology changing conventional stages of a Phishing lifecycle; Automated Target Selection (3) Craft deceptive content and rich HTML form creation. This progression has created for unique stumbling blocks in real-time threat discovery, such as AI-generated phishing content looks convincingly like actual communication does that standard discovering systems are incapable of differentiating between what is legitimate and malicious.

As well, existing AI-based methods for phishing detection have been demonstrated vulnerable to adversarial attacks, it indicates the importance of constantly evolving and refining these systems. The risk of adversarial examples is that attackers modify inputs in a way so subtle they are virtually impossible to humans to detect, thus enabling them thwarting detection methods based on pattern recognition or anomaly detection. These developments suggest a defense strategy that incorporates in-depth, adaptive AI techniques layered with human oversight and continuous learning around emerging threats.

Generally, the promise of AI and machine learning solutions to improve phishing detection was highlighted but it remains as an ongoing arms race between attackers and defenders demonstrating that all organizations need a proactive approach with continuous technical security testing. With the types of AI phishing campaigns we have seen, it is crucial to create stronger detection mechanisms that can predict and prevent new strategies for operations before they are put into action.

References

1. Junaid Ahsenali Chaudhry 1 , Shafique Ahmad Chaudhry 2 , Robert G. Rittenhouse, Phishing Attacks and Defenses, 2016 International Journal of Security and Its Applications, V ol. 10, No. 1 (2016), pp.247-256
2. Zainab Alkhalil, Chaminda Hewage, Liqaa Nawaf and Imtiaz Khan, Phishing Attacks: A Recent Comprehensive Study and a New Anatomy , 2021 Frontiers in Computer Science, March 2021 – Volume 3 – Article 563060
3. Pawankumar Sharma1, Bibhu Dash2, Meraj Farheen Ansari, Anti-Phishing Techniques – A Review of Cyber Defense Mechanisms, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 11, Issue 7, July 2022
4. M. Elatoubi, Phishing in Web 3.0: Opportunities for the Attackers, Challenges for the Defenders, ARIS2- Journal, vol. 3, no. 2, pp. 11–25, Dec. 2023.
5. P. Amba Bhavani, Chalamala Madhumitha, Pinnam Sree Likhitha, Chanda Pranav Sai, Phishing Websites Detection Using Machine Learning, Department of Information Technology, Maturi Venkat Subba Rao (MVSR) Engineering College
6. Akshaya Arun, Nasr Abosata, Next Generation of Phishing Attacks using AI powered Browsers, Department of Computer Science and Engineering Northumbria University London, United Kingdom
7. Prashanth Rajivan and Cleotilde Gonzalez, Creative Persuasion: A Study on Adversarial Behaviors and Strategies in Phishing Attacks , Frontiers in Psychology , February 2018 – doi: 10.3389/fpsyg.2018.00135
8. Arshpreet Singh Sohal, Deepakmoney Banga, Kevin Antony, PhishNET: A Phishing Websites Detection Tool, Department of Computer Science and Engineering Dr. B. R. Ambedkar National Institute of Technology Jalandhar -144008, Punjab (India) May 2024
9. Barbara Kitchenham, Stuart M. Charters, Guidelines for performing Systematic Literature Reviews in Software Engineering, Department of Computer Science University of Durham Durham UK, 9 July, 2007
10. The State of Phishing, SlashNext, 2022.
11. Sajjad, R., Khan, M. F., Nawaz, A., Ali, M. T., & Adil, M. (2022). Systematic analysis of ovarian cancer empowered with machine and deep learning: a taxonomy and future challenges. Journal of Computing & Biomedical Informatics, 3(02), 64-87.
12. Valentine Adeyemi Onih, Phishing Detection Using Machine Learning: A Model Development and Integration., International Journal of Scientific and Management Research Volume 07 Issue 04 (April) 2024 ISSN: 2581-6888 Page: 27-63
13. Shah, A. M., Aljubayri, M., Khan, M. F., Alqahtani, J., Sulaiman, A., & Shaikh, A. (2023). ILSM: Incorporated Lightweight Security Model for Improving QOS in WSN. Computer Systems Science & Engineering, 46(2).
14. Riyadh Rahef Nuiaa, Selvakumar Manickam, A Critical Review: Revisiting Phishing Attacks Classification and Analysis of Techniques Employed in Taxonomies, Wasit Journal for Pure Sciences Vol. (2) No. (2), 30 Jun 2023
15. Jerson Francia , Derek Hansen , Ben Schooley , Matthew Taylor, Shydra Murray and Greg Snow, Assessing AI vs Human-Authored Spear Phishing SMS Attacks: An Empirical Study Using the TRAPD Method, arXiv:2406.13049v1 [cs.CY] 18 Jun 2024
16. Rana Alabdan, Phishing Attacks Survey - Types, Vectors, and Technical Approaches, Future Internet 2020, 12, 168; doi:10.3390/fi12100168
17. R. Butler, Investigation of Phishing to Develop Guidelines to Protect the Internet Consumer's Identity Against Attacks by Phishers, South African Journal of Information Management, Vol.7(3) September 2005
18. Hafeez, M. A., Imran, A., Khan, M. I., Khan, A. H., Nawaz, A., & Ahmed, S. (2022). Diagnosis of liver disease induced by hepatitis virus using machine learning methods. In 2022 8th International Conference on Information Technology Trends (ITT) (pp. 154-159). IEEE.