





Overview of Social Engineering Protection and Prevention Methods

Konstantinos Kontogeorgopoulos^(✉) and Kyriakos Kritikos

Department of Information and Communication Systems Engineering,
University of Aegean, Mytilene, Greece
{kontogeorgopoulosk,kkritikos}@aegean.gr

Abstract. Recently, with the increasing use of social networks, services, and computers in general plus the enhanced capabilities of remote working, especially during quarantine periods due to Covid-19, social engineering attacks are a growing phenomenon. These attacks are, nowadays, the most common, since no matter how protected an information system is from security attacks, the weakest link is the human factor. As such, it is imperative to address and prevent such attacks. This paper reviews the most common social engineering attack prevention and protection methods and classifies them based on various criteria. Based on the analysis, it identifies the most effective methods in their protection degree, while it supplies some challenges to maximise such degree.

Keywords: Social engineering · attacks · protection methods · review

1 Introduction

Social engineering is the manipulation of individuals to extract information, especially confidential and sensitive data. These attacks are so widespread as no matter how strong the security of an information system and the strength of its protection mechanisms are, the system can be penetrated due to external factors, such as people [1]. The social engineering attack methods used do not require as much time and effort as other types of attacks that exploit system vulnerabilities, since humans are dominated and operate based on emotions. This makes these attacks among the most dangerous [1] since they cannot be yet addressed with a complete and definitive security solution while their confrontation also requires the proper training of the people who access and operate the systems in question.

As such, social engineering protection and prevention methods have witnessed significant advancement. Organizations are increasingly investing in security awareness and training programs, which aim to educate employees about the risks of social engineering attacks and how to identify and respond to them. Further, new technologies, such as machine/deep learning and natural language processing (NLP) are being developed to address social engineering attacks in

Supported by organization x.

real-time. The increasing adoption of these protection methods has significantly reduced the success rate of social engineering attacks and improved the overall security posture of organizations [2].

Systemizing protection methods allows organizations to streamline processes and standardize practices. Categorizing and evaluating different protection methods helps identify effective approaches and prioritize their implementation. This promotes efficiency, reduces effort duplication, and ensures consistent application of social engineering protection measures. By understanding the methods' effectiveness, organizations can allocate resources and prioritize measures based on risk levels, aligning strategies with their objectives. Documenting and categorizing effective protection methods facilitates sharing best practices and lessons learned, fostering collaboration and enhancing defence against social engineering attacks. Having a comprehensive understanding of social engineering protection and prevention methods is crucial so as to achieve this vision.

Our paper introduces significant advancement and value by exploring machine, deep and hybrid learning plus scenario-based attack detection methods. By extending from [3] and [4], our study goes beyond the existing scope to delve into the realm of deep learning algorithms, harnessing their potential to enhance attack detection accuracy and robustness. Moreover, by incorporating hybrid learning techniques the detection system's effectiveness is further strengthened. Our paper also introduces the concept of scenario-based attack detection, which considers real-world scenarios and contextual information to improve the system's ability to identify and mitigate emerging threats.

Our paper adds extra value by introducing original criteria for comparing protection methods. By applying such criteria, our research goes beyond the existing literature and provides a comprehensive and objective framework to assess the protection approaches effectiveness. This novel contribution allows for a more systematic understanding of the strengths and weaknesses of various methods, enabling researchers and practitioners to make informed decisions when selecting and implementing cybersecurity measures. Further, creating such criteria opens up new avenues for future research, as they can serve as a benchmark to evaluate and refine protection techniques in an evolving threat landscape.

The remainder of this paper is structured as follows. Section 2 explains the methodology used to select the defence and prevention methods. Section 3 introduces the evaluation criteria, evaluates the methods based on them and analyzes the evaluation results. Finally, Sect. 4 concludes the paper.

2 Method Selection Methodology

The selection of suitable social engineering protection and prevention methods is critical to ensure the security of an organization's sensitive information and systems. In this paper, a systematic literature review was conducted to evaluate existing measures and methods (policies and tools) to address social engineering attacks. The review's main research questions to answer were the following:

- What are the main categories of methods utilized to protect against social engineering attacks?
- What are the main pros and cons of each protection method in each category?
- Which is the best protection method based on which criteria and aspects?

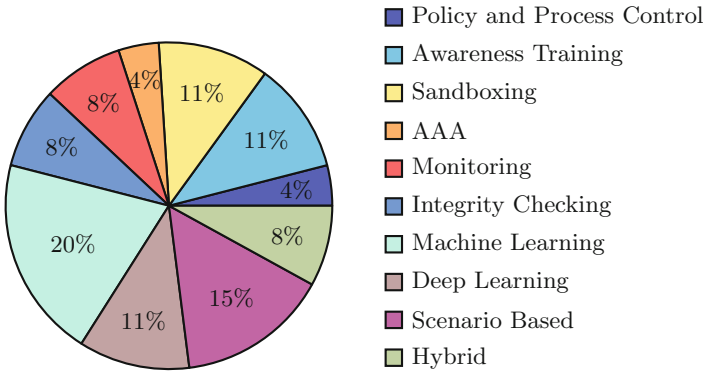


Fig. 1. The distribution of different protection and prevention methods.

The search strategy employed used keyword patterns to search relevant literature on Google Scholar, Science Direct, and Web of Science. The keywords used were “Social Engineering || Phishing || Impersonation” && “Attack” && “Detection || Prevention || Protection.” || means logical OR, while && logical AND.

Several articles and studies appearing with the above keywords were examined. To separate and filter the studies, eligibility criteria (exclusion, inclusion and quality) were applied to determine whether to include or exclude each identified article from the subsequent analysis.

The inclusion criteria were as follows:

- Literature publications which include research articles from scientific journals, conferences and workshops and doctoral theses.
- Publications proposing techniques, tools, methodologies, strategies and solutions focusing on social engineering attacks prevention and addressing.

The exclusion criteria were as follows:

- Exclusion of publications published before 2008. By focusing on more recent literature, we capture the latest advancements, trends, and insights in the field.
- Exclusion of publications written in a language other than English.
- Exclusion of publication with charged access to their content

The quality criteria were as follows:

- Exclusion of publications supplying an unmeaningful solution.

- Exclusion of publications with no kind of assessment of their contribution.
- Exclusion of publications with unverifiable assessment results.

The articles originally identified were 367 while the ones retained after the application of the eligibility criteria were 66. Of these 66 articles, 50 mapped to Protection and Prevention Methods that we chose to examine while the rest were literature review ones, kept for respective knowledge extraction and utilization by our paper. The 50 methods are approximately distributed based on their protection method category in Fig. 1 while their distribution based on their publication kind is shown in Fig. 2.

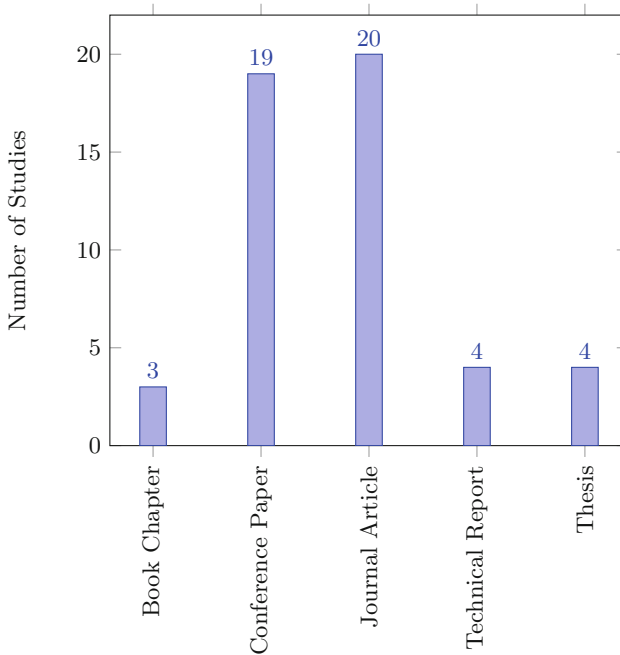


Fig. 2. The publication distribution of the different studies.

3 Analysis

This section overviews the main method categories. Based on this categorization, it then classifies the protection methods selected according to the methodology in Sect. 2. It also evaluates these methods based on some key evaluation criteria.

3.1 Overview of Protection Method Categories

This sub-section analyzes the main categories of social engineering protection methods by extending the method categorisation in [3] with the categories of Deep Learning, Hybrid Learning and Scenario-based Attack Detection.

Policy and Process Control. Policy and process control provide hierarchical control via some management and process frameworks, as opposed to technical systems, such as security software. They are essential in an organisation as they provide a comprehensive protection approach. Most importantly, they are procedures to prevent and detect potential attacks, but also steps and procedures to react to such attacks. They are designed to reduce exposure to social engineering attacks. Well-maintained policy and organizational procedures help mitigate the occurrence of an exploit without relying on the technical capabilities of the system users. Policy and Procedure Control are the backbone of the organizations' security, while the overall security countermeasures and tools to protect against such attacks are decided based on them.

Awareness Training. Since attacks target the system's users, attacks can be greatly reduced by proper user training and awareness. Education is a key defence model element. In particular, in social engineering attacks, it concerns introducing and applying training programs to compensate for and mitigate the technical security mechanisms inadequacy.

Empirical studies and research have shown that awareness training programs can enhance individuals' ability to recognize and identify social engineering attacks. Training participants become more alert to common tactics, such as phishing emails, impersonation attempts, and phone scams. Further, awareness training encourages individuals to report suspicious activities or social engineering attempts. This can lead to quicker incident response and mitigation of potential security breaches. When employees are educated about the risks and consequences of social engineering attacks, they become more proactive to safeguard information and are more likely to adopt secure behaviors [5].

Social engineering tactics evolve rapidly such that individuals require regular and updated training to stay informed. One-time training sessions may not be sufficient to combat the ever-changing landscape of social engineering attacks.

Technical. "Technical" are those protection and prevention methods in which the human factor is either irrelevant or has little significance in addressing social engineering attacks. These methods create mechanisms that either prevent and protect against social engineering attacks entirely on their own or create an infrastructure helping the user to identify and defend against such attacks.

We categorize Technical Protection and Prevention Methods into:

Sandboxing Mechanisms. In sandboxing, an isolated computer environment is created, usually via virtualization, to test unreliable functions. It has been effectively applied in various IT domains, from specific code platforms to browsers, plus in the field of smartphone security to improve defence against malware. Sandboxing can be used to help in protecting against social engineering attacks by isolating potentially malicious programs or actions from the rest of a system. For example, if a user clicks on a link or opens an email attachment that contains malware, sandboxing can prevent that malware from spreading beyond the sandboxed environment. This can help prevent the attacker from gaining access to sensitive data or causing damage to the system [6].

Authorisation, Authentication, and Accounting (AAA). AAA is a framework for intelligent computer access, enforcing authentication and authorization-related policies, controlling usage, and providing the information needed for service billing. It is typically applied in controlled environments, especially where there is a diverse user landscape that compromises data control and protection. The framework provides controls for accessing resources (Authentication), enforcing organizational policies (Authorization), and controlling resource usage (e.g., devices accessed). Its use is intended to ensure that organisations have a detailed assurance and control level over who has access to a system, based on data about names, roles, skill sets, etc. By using strong authentication methods, such as multi-factor authentication, organizations ensure that only authorized users access sensitive data or systems. Further, by implementing strong accounting policies, logging and monitoring of user activities are enabled, providing an audit trail to be used for post-incident analysis and forensic investigations. This can help detect and mitigate the effects of social engineering attacks, as organizations can identify respective suspicious behaviour.

Monitoring. Monitoring concerns observing a computer system's behaviour, generated by user/programmer actions, programs, services and processes, via collection, aggregation and analysis mechanisms. Monitoring is a key security mechanism for social engineering attacks, as new attacks can be identified by logging and analysis of network traffic and effective security control where they can be detected by juxtaposition to normal/legal user actions in the system.

Effective monitoring enables timely alerts and notifications when potential social engineering attacks are detected. These alerts can trigger incident response procedures, allowing security teams to investigate and mitigate the attacks before causing significant harm [7].

Social engineering attacks evolve over time; thus, monitoring should be an ongoing process. Regularly reviewing and updating monitoring systems, staying informed about new attack vectors, and adapting monitoring strategies are critical to maintaining an effective defence against social engineering attacks.

Integrity Checking. The integrity of applications and data is difficult to ensure without proof or analysis. Integrity checking provides the user with a visual response and technical assurance as to whether the file, site, or data should be trusted through various tools like Intrusion Detection Systems (IDS).

Integrity checking can effectively identify instances where malware has been injected into files or system components. By regularly verifying file integrity and detecting unexpected modifications, organizations can detect and mitigate the impact of social engineering attacks involving malware injection. Further, it can also help ensure the integrity and authenticity of data by using cryptographic hashing or digital signatures such that organizations can verify data integrity at rest or during transmission. This helps protect against social engineering attacks that involve tampering with sensitive information or data manipulation [8].

Machine Learning. Research has demonstrated that malware detection via machine learning (ML) can be dynamic, where appropriate algorithms, such as support vector machines and neural networks can be applied to profile files

against known and potential exploits and distinguish between legitimate and illegitimate data. ML algorithms have been successfully applied to detect malicious emails using anomaly classification techniques, thus demonstrating their potential for further application to other areas of social engineering attacks.

The effectiveness of ML algorithms in detecting social engineering attacks relies heavily on the availability of high-quality and diverse training data. Collecting data sets that encompass a wide range of social engineering attack scenarios can be challenging. Further, maintaining up-to-date data sets to keep pace with evolving attack techniques is crucial for ML model accuracy [9].

Deep Learning. Similar to ML, Deep Learning (DL) algorithms can prevent social engineering attacks by analyzing patterns in user and system behaviour and detecting anomalies indicative of an attack. DL algorithms can be used to analyze the language used in emails, social media messages, and other communication channels to detect phishing and other social engineering attacks. Natural language processing (NLP) can identify suspicious language patterns or unusual word usage indicative of an attack. These algorithms can also be used to analyze images and videos for signs of social engineering attacks, such as phishing sites, fake login pages, or malware. Recent developments in approaches have suggested that the classification of phishing websites using neural networks should outperform traditional ML algorithms.

DL techniques, e.g., convolutional neural networks (CNNs), have been used to analyze visual content, such as images or video frames, to detect social engineering-related cues or visual elements. For example, DL models can identify spoofed websites or altered images used in social engineering attacks [10].

Hybrid Learning. Hybrid learning (HL) is a training approach that combines different types of learning algorithms or architectures to improve a DL model's performance. By utilizing the advantages of various learning architectures or algorithms, HL seeks to improve upon each one's shortcomings. An example HL approach is to combine supervised and unsupervised learning methods or different types of deep learning architectures, such as CNNs and recurrent neural networks (RNNs). HL can improve detection model accuracy and reduce false positives, whereas multi-modal data analysis models can combine multiple data types, such as images, and speech, to analyze the different aspects of a social engineering attack. The models can also adapt and learn from new attack types and update their detection algorithms accordingly [11].

Scenario-Based Attack Detection. By simulating various attack scenarios and examining user behaviour for signals of an attack, scenario-based attack detection is a technique used to identify and stop social engineering attacks. It involves creating hypothetical situations and common attack patterns that closely resemble strategies and procedures employed by attackers and then keeping an eye on user behaviour to spot potential risks. Red team assessments, involving simulated attacks performed by specialized teams, are often used to evaluate an organization's resilience against social engineering attacks. Such assessments provide empirical evidence by demonstrating how effective existing

security measures are in detecting and mitigating real-world social engineering threats [12].

Scenario-based attack detection should be an ongoing process that evolves alongside emerging social engineering techniques. Regularly updating and refining attack scenarios based on new threats and attack vectors is crucial to maintain the effectiveness of this protection method.

3.2 Criteria for Method Evaluation

This section defines newly devised criteria for evaluating the protection and prevention methods selected. Some of these criteria focus on the applicability of the protection methods to be evaluated.

Method of Protection - MoP. Refers to a specific approach, technique, or countermeasure implemented to safeguard individuals, organizations, or systems against social engineering attacks.

Method of Treatment - MoT. Indicates whether the method of treatment targets *Prevention P*, *Reaction R*, *Detection D* or a mix of these.

Degree of Protection - DoP. Evaluates the effectiveness of a protection method in addressing the attacks it specializes in. The evaluation can lead to assessing that the provided degree of protection is either *Small*, *Medium*, or *Great*. “Small” means that extra measures or improvements are necessary to enhance the protection level and strengthen the security posture, “Medium” indicates that the method offers a satisfactory protection level under typical circumstances while “Great” suggests that the method surpasses the average protection level and is considered highly reliable and secure.

Ease of Implementation - EoI. Assesses how easily the response suggested by a protection method can be implemented in an information system or incorporated into an organization’s plan. EoI can be categorised as *Small*, *Medium*, or *Great*, based on the level of effort, resources, and complexity required to deploy and integrate the suggested response. “Small” indicates that the implementation process is straightforward, requiring minimal changes or adjustments, “Medium” implies a balanced effort level without posing overwhelming obstacles while “Great” suggests that the response can be quickly adopted without causing disruptions or significant changes to existing systems or processes.

Application Part - AppP. Identifies the areas or components to which the proposed method applies. These areas are: *The architecture, Policies and Procedures, Security Mechanisms, People, and Systems*. This categorization clarifies the scope and context in which the method can be effectively implemented.

Implementation Time - ImplT. Refers to the estimated duration, measured in *Hours, Days, Months, or Years*, required to fully implement the proposed method. It represents the time investment needed to deploy and integrate the method into an organization’s existing infrastructure, processes, and security framework. While time estimation can be quite challenging, several factors can

be considered to facilitate it: the method complexity, organization size, resource availability, complexity and integration, plus the training and familiarization. By breaking down the implementation tasks, identifying dependencies, and considering the above factors, a reasonable implementation time estimation can be derived. However, it must be noted that unforeseen challenges or unexpected circumstances may impact the actual implementation time, and regular monitoring and adjustment of the implementation plan may be necessary.

Application Effort - AppE. Evaluates the amount of effort required to fully implement the proposed method. AppE can be assessed as *small*, *medium*, or *great*, based on the resources, time, and complexity involved in the implementation process. “Small” indicates that the method can be readily implemented efficiently without significant disruptions or resource-intensive activities with the available resources and within a reasonable time frame. “Medium” indicates that it can be accomplished within a manageable time frame and with a reasonable resource allocation. Thus, the method is implementable with the organization’s existing capabilities and may require a moderate coordination and planning level. A “Great” value suggests that the implementation may require significant changes to the existing infrastructure, processes, or systems. So, the method is resource-intensive, complex, and may require extra expertise or support for successful implementation. Thus, implementing such a method may involve extensive planning, coordination, and resource allocation.

Implementation Cost - ImplC. Refers to the expenses associated with implementing and maintaining a protection method. It quantifies the financial resources required for the method’s setup, deployment, and ongoing management. A method’s ImplC can be categorized into three general categories: *Small*, *Medium*, and *Great*, representing different cost ranges. “Small” indicates that the method can be implemented without significant financial burden or extra investments. Thus, the cost is manageable and aligns with the organization’s budgetary constraints. “Medium” indicates that the cost is within a balanced range, considering the value provided by the method and the organization’s financial capabilities. Thus, the implementation cost is justifiable and can be accommodated with appropriate budget planning. A “Great” assessment value suggests that the cost may exceed the average budget allocation and might require extra financial resources or long-term commitments. As such, the method may involve expensive infrastructure, specialized tools, or ongoing licensing fees.

3.3 Evaluation Results

The evaluation results are presented via a set of tables. Each table showcases how well each method within a specific category satisfies the above criteria.

3.4 Analysis of Evaluation Results

This section analyses the methods’ effectiveness in terms of their performance against the devised criteria. The analysis is performed per method category.

Table 1. Policy and Process Control

MoP	MoT	DoP	EoI	AppP	ImplT	AppE	ImplC
[13]	D, P, R	Great	Medium	Policies and Procedures	Months	Medium	Medium
[14]	P	Medium	Medium	Policies and Procedures	Months	Medium	Medium

Table 2. Awareness Training

MoP	MoT	DoP	EoI	AppP	ImplT	AppE	ImplC
[15, 16]	P	Medium	Great	People	Days	Great	Great
[17]	P, R	Small	Great	Systems, People	Hours	Great	Great
[18]	P	Medium	Great	Systems, People	Hours	Great	Great
[19]	P	Medium	Medium	Systems, People	Hours	Medium	Great
[20]	D	Great	Small	Systems, People	Days	Medium	Great

Policy and Process Control deals with all security levels (Technical Attacks - Social Engineering Attacks, etc.) but every category method is quite time-consuming to implement. It is a general method of security that stands out in overall organisational security approaches. It is defined around the business and the user environment. However, the security frameworks introduced to address attacks have been added as extra elements to the broader security architecture, rather than to the strategic policy and process control development (i.e., by-design) [62]. More importantly, policy must be inherently structured with people management and embedded at the core of all information systems.

As can be seen in Table 1, EoI is moderate and implementation time corresponds to months since it takes some time to implement such security approaches as a whole. There is a fairly good protection degree but mainly only general guidelines for security procedures are provided for the whole information system and the people participating in it. The application part is Policies and procedures and the implementation costs are moderate.

More effective policies may be developed identifying gaps in current policies and introducing new policies better tailored to social engineering threats. There is also a need to measure more accurately policies and process control effectiveness in preventing and mitigating social engineering attacks with new metrics and evaluation methods. Such data can also be used to train ML and DL models.

Awareness Training is probably the most basic response to social engineering attacks since the weak system link is the human user. As can be seen in Table 2, this method category is mainly concerned with prevention, while its main application targets are humans and systems. The protection degree, ease of implementation, implementation time and cost can vary depending on the program and training type each organization-organization will follow.

There are various training modes, where beyond a simple presentation or seminar, they can take the form of interactive games or training systems, which make the process more interesting and reward the user in the learning process. Training has shown good results as a way of protection as it reduces social engineering attacks to a fairly satisfactory degree, but it must be done thoroughly

Table 3. Technical - Sandboxing

MoP	MoT	DoP	EoI	AppP	ImplT	AppE	ImplC
[21]	D, P, R	Great	Medium	Systems	Months	Medium	Medium
[6]	D	Medium	Medium	Systems	Days	Medium	Medium
[22, 23]	P	Small	Medium	Systems	Weeks	Medium	Medium
[24]	R	Great	Great	Systems	Weeks	Great	Medium
[25]	R	Great	Great	Systems	Days	Great	Great

Table 4. Technical - AAA

MoP	MoT	DoP	EoI	AppP	ImplT	AppE	ImplC
[26]	D	Medium	Medium	Systems, Policies and Procedures	Days	Medium	Medium
[27]	R	Medium	Medium	Systems, Policies and Procedures	Days	Medium	Great

and properly implemented in an organisation to attain such results. It must be also continuously applied to cover new attacks and exploitation modes [63].

Sandboxing mechanisms represent a good protection way at a low cost compared to what they offer. As can be seen in Table 3 the protection provided is *Medium* to *High*, except for some methods in a more experimental stage. The EoI is *Moderate* to *Great* since the system supports the user in making correct decisions as to how to run applications of dubious origin in such an environment via a UI without being obscure and difficult for the ordinary user. Implementation time ranges from days to months depending on the approach taken and the environment choice (widespread and quickly accessible, or experimental-research in development) while implementation effort ranges from *Medium* to *Great* in the examined methods, as they did not use an already existing infrastructure with some exceptions. The application target is systems while cost again varies depending on the environment choice, but ranges towards *Moderate*.

The attack range addressed is wide as many key attack features are covered [64]. The Sandboxing mechanisms already in use are widespread as a security solution, with a fairly good defence rate against large-scale attacks but there is room for improvement in enhancing their detection capabilities and performance.

AAA: It is a moderate protection mode and specific towards large organisations as it provides a centralized management framework for access control, making it easier to manage a large number of users, resources and permissions, and can easily be scaled depending on the use case. As can be seen in Table 4, the coping mode is mainly in detection and reaction with medium protection and EoI as there are many established infrastructures and implementations that are accessible (FreeRADIUS, Globberry, etc.). EoI is also moderate since once the AAA infrastructure is built, it is very easy to cover many people in the organisation. Implementation time is usually within hours and the implementation effort is moderate. The application targets are systems, policies and procedures while the implementation cost is moderate to low if open-source solutions are used. However, the cost of implementation is Medium to Great in the solutions

Table 5. Technical - Monitoring

MoP	MoT	DoP	EoI	AppP	ImplT	AppE	ImplC
[28]	D	Medium	Medium	Systems	Weeks	Medium	Medium
[29]	D	Great	Great	Systems	Hours	Small	Great
[30]	D	Medium	Great	Systems, Policies and Procedures	Hours	Medium	Great
[31]	D	Great	Great	Systems	Hours	Small	Great

Table 6. Technical - Integrity Checking

MoP	MoT	DoP	EoI	AppP	ImplT	AppE	ImplC
[32]	D	Medium	Medium	Security Mechanisms	Months	Small	Medium
[33]	P	Small	Medium	Security Mechanisms	Months	Medium	Medium
[34]	P	Medium	Great	Security Mechanisms	Weeks	Medium	Medium
[35]	D	Medium	Medium	Security Mechanisms	Weeks	Medium	Small

we examined as they require extra resources, such as storage, bandwidth, and processing power, to collect and analyze data.

The authentication methods can be improved to better protect against social engineering attacks relying on credential theft or account takeover. While AAA solutions typically include some authentication form, they may not be sufficient to protect against the latest social engineering techniques. Research in this area could focus on developing new authentication methods more resistant to social engineering attacks, such as biometric authentication or behavioural analysis. While AAA solutions typically include authorization controls, these controls may not be sufficient to prevent social engineering attacks that exploit weaknesses in the authorization process. Research in this area could focus on developing new authorization methods more resistant to social engineering attacks, such as adaptive authorization that considers user behaviour and context.

Monitoring: As can be seen in Table 5, response mode is *Detection* while EoI is medium to great, as there are many open-source implementations (e.g., Wireshark, OSSEC). It is mainly applied to systems. The implementation time depends on the solution and tool choice (Hours - Days - Months) but the monitoring data, to be meaningful, must be collected over a long time period. The implementation cost is from moderate to none as there are several monitoring programs even for free (e.g., SolarWinds IP Monitor). Monitoring delivers good protection results [29] It is one of the most effective protection ways, if there is a cyber security specialist or a mechanism (Software, Model, etc.) in the organization that manages the network traffic, system logs, user activity, etc.

Monitoring may not be always able to detect the latest social engineering attacks. As such, research in this area could focus on developing new, more effective monitoring techniques, such as user activity monitoring, network monitoring, and endpoint monitoring. Further research is needed to compare the effectiveness of different monitoring approaches and identify best practices.

Integrity Checking. As can be seen in Table 6, response mode is prevention and detection, since users are warned of any malicious actions. The protection degree is moderate since the final decision beyond warning is at the user’s discretion. EoI

Table 7. Technical - Machine Learning

MoP	MoT	DoP	EoI	AppP	ImplT	AppE	ImplC
[36]	D	Medium	Great	Security Mechanisms	Months	Great	Medium
[37]	D	Medium	Medium	Systems	Weeks	Medium	Medium
[38]	D	Great	Medium	Security Mechanisms	Months	Medium	Medium
[39]	D	Small	Medium	Security Mechanisms	Months	Medium	Great
[40]	D	Small	Medium	Architecture	Months	Medium	Medium
[41]	D	Medium	Great	Systems	Hours	Medium	Great
[42]	D	Great	Great	Systems	Hours	Medium	Great
[43]	D	Great	Medium	Systems	Months	Medium	Small

Table 8. Technical - Deep Learning

MoP	MoT	DoP	EoI	AppP	ImplT	AppE	ImplC
[44]	P	Small	Small	Architecture	Months	Medium	Small
[45]	D	Small	Medium	Security Mechanisms	Months	Medium	Medium
[46]	D	Great	Small	Systems	Months	Medium	Small
[47]	P	Small	Medium	Security Mechanisms	Weeks	Medium	Medium
[48]	P	Small	Medium	Security Mechanisms	Weeks	Medium	Medium
[49]	D	Great	Medium	Security Mechanisms	Days	Medium	Medium

is medium as there are existing tools (e.g., Tripwire, AIDE); but it may vary in some methods depending on what stage they are and how experimental is their approach. The application target is Security mechanisms. Implementation effort is usually moderate as the existing tools are easily integrated and there is sufficient documentation for such an integration, with some exceptions depending on the research and the algorithms under consideration. Implementation time is moderate - mostly months as there was no existing infrastructure in the examined methods. The cost of implementation is *Medium*.

Integrity checking can be improved by enhancing the methods' scalability to protect against large-scale attacks. The existing methods should be also extended, e.g., by using ML algorithms to identify anomalous behaviour, so as to address attacks that involve manipulation of data or systems.

Machine Learning becomes increasingly common with great research interest since the use and invention of ML algorithms have been quite widespread recently. As shown in Table 7, the coping mode is *Detection*. The protection degree, implementation time and implementation cost vary depending on the approach followed and the volume of data selected each time for learning. The application target varies (Security Mechanisms, Systems, Architecture) in the examined methods. EoI is moderate to high, as once the algorithm has acquired the necessary 'knowledge', it can be included in systems with relative ease.

Using feature variables with behavioural input data sets (usually collected through monitoring), accurate predictions and indicator measurements can be achieved in terms of the significance of a file or user behaviour effect on a system.

Table 9. Technical - Scenario Based

MoP	MoT	DoP	EoI	AppP	ImplT	AppE	ImplC
[50]	D	Small	Medium	Security Mechanisms	Weeks	Medium	Medium
[51, 52]	P	Small	Medium	Security Mechanisms	Months	Great	Medium
[53]	R	Medium	Medium	Security Mechanisms	Weeks	Small	Medium
[54]	P	Small	Medium	Security Mechanisms	Weeks	Medium	Medium
[55]	D	Medium	Great	Security Mechanisms	Days	Small	Great
[56]	D	Small	Small	Security Mechanisms	Weeks	Medium	Great
[57]	D	Small	Medium	Security Mechanisms	Weeks	Small	Medium

Table 10. Technical - Hybrid

MoP	MoT	DoP	EoI	AppP	ImplT	AppE	ImplC
[58]	P	Medium	Medium	Security Mechanisms, Policies and Procedures	Weeks	Medium	Great
[59]	R	Great	Medium	Security Mechanisms, Policies and Procedures	Weeks	Medium	Medium
[60]	D	Medium	Medium	Security Mechanisms, Policies and Procedures	Days	Medium	Medium
[61]	D	Great	Medium	Security Mechanisms, Policies and Procedures, People	Days	Medium	Medium

While ML tools have been built, tested and evaluated in research, their application has largely focused on countering phishing attacks [65, 66].

There is room to further optimize ML methods. Developing more effective feature engineering methods is a potential research gap so as to more accurately extract relevant features from social engineering attack data. The interpretability of ML models could be also improved by utilising, e.g., explainable artificial intelligence (XAI) methods [67].

Deep Learning. With neural networks being increasingly used and slowly replacing traditional ML algorithms, DL approaches are becoming more and more common. As can be seen in Table 8, DL methods supply detection and prevention abilities since they are similar to ML methods. The protection degree can vary from small to great as it depends on the model created and the data used to create it. EoI is moderate as it depends on the way the problem is approached and the amount of data to be learned. The application target can vary (Architecture, security mechanisms, and systems). Implementation time varies from weeks to months, as it is quite time-consuming to train a neural network with a large data volume. Implementation effort is medium and the cost is small to medium in the studied methods, as suitable, advanced tools to train the models already exist (e.g., Tensorflow, Keras). These values may vary depending on the integration requirements of each implementation effort.

The DL research results are modest at present but may improve further due to the intensity of research being conducted. DL solutions can be improved by more diverse and relevant data collection, incorporating contextual information into DL models and developing DL models that can detect and respond to social engineering attacks in real-time. It is also worth addressing the potential for bias in the DL algorithms so as to increase prediction accuracy.

Scenario-Based. It can be seen from Table 9 that the response mode varies while the application target is security mechanisms. The protection degree, EoI, implementation time and effort can vary as they depend on the kind and complexity of the scenario chosen by the researchers. As such, such methods can be used for more specific situations and organisation needs; however, we can use combinations of scenario-based attack detection methods to get better results and attain a larger coverage.

Research must be conducted for the more accurate measurement of the methods' effectiveness via the use of objective and standardized metrics. Further, there is a need to develop personalized training scenarios that are tailored to the specific needs and vulnerabilities of individual employees or employee groups.

Hybrid. As shown in Table 10, the response mode varies. The implementation effort is medium in the examined methods as hybrid proposals with specific use cases can have great cost-benefit analysis and a well-defined incident-response plan. The application target is Security mechanisms, policy and processes plus people, as hybrid methods use a union of technical and non-technical approaches. The implementation time ranges from days to weeks. The cost can vary from little to great depending on the method complexity and integration. Hybrid protection methods are quite effective but with a slightly higher implementation cost due to the expertise and infrastructure needed.

Hybrid approaches often involve integrating multiple solutions and technologies, which can be challenging. Thus, there is a need to develop standardized frameworks and protocols to enable seamless integration of different protection methods. There is also the need to develop optimization algorithms that can take into account the strengths and weaknesses of the different protection methods to achieve optimal hybrid approach performance.

Overall Analysis: Determining a single method category as universally better than the others based on all criteria is challenging, as the methods' effectiveness against social engineering attacks depends on various factors. Different criteria hold different weights of importance for different organizations or systems, making it difficult to establish a definitive superiority across all categories. However, there are categories that may be deemed better than others based on specific criteria. For instance, ML methods promise to detect and mitigate social engineering attacks by utilizing advanced algorithms and pattern recognition. They can adapt and evolve to new attack vectors, making them highly effective in certain scenarios. Similarly, categories like Policy and Process Control, focusing on establishing robust security procedures and guidelines, can provide comprehensive protection and help organizations maintain a strong defence against attacks. In terms of EoI, some categories may require significant effort and time to fully implement, such as Policy and Process Control. These methods typically involve developing comprehensive security procedures and guidelines for the entire information system, which can be time-consuming and resource-intensive. On the other hand, categories like Monitoring and Integrity Checking may have a relatively easier implementation process, as there are existing tools and open-source solutions available. Similarly, the protection degree can vary

across categories. While some methods may offer great protection against social engineering attacks, such as Sandboxing Mechanisms and ML, others may provide only moderate or small protection levels. It depends on the specific features and capabilities of each method in addressing the attacks they specialize in.

It is also challenging to pinpoint a specific category as generally worse than others. Each category has its own strengths and weaknesses, and their effectiveness can vary depending on the context and specific criteria. For example, categories like Awareness Training, aiming to educate and empower users to recognize and resist social engineering attempts, can be highly effective only when implemented correctly. However, if not properly executed or lacking regular updates, their impact may be limited.

In real-world scenarios, the implementation of social engineering prevention methods often requires a hybrid approach that combines multiple strategies to address the multifaceted nature of social engineering attacks. These hybrid solutions leverage a combination of technological, procedural, and educational measures to create a robust defence against ever-evolving threats. For instance, an organization might employ advanced email filtering systems to detect and block phishing attempts, complemented by periodic security awareness training for employees to identify and report suspicious messages. Additionally, access controls and multi-factor authentication mechanisms can be integrated to prevent unauthorized access to critical systems, mitigating the risk of social engineering attacks that exploit human error. The adoption of hybrid solutions allows organizations to create a layered defence, where each protective measure reinforces the effectiveness of others, thereby significantly reducing the likelihood of successful social engineering attacks.

To determine the most suitable category or combination of methods, organizations should carefully evaluate their needs, assess the potential risks they face, and consider those criteria that hold the highest priority for their operation. By conducting thorough assessments and understanding the strengths and limitations of each category and method especially against the devised criteria, organizations can make informed decisions on which methods are most suitable for their needs to establish a multi-layered defence against social engineering attacks.

4 Conclusion

This paper has supplied an analysis of protection and prevention methods against social engineering attacks, facilitating the selection of such methods based on the user/organisation needs as well as the development of countermeasures and conduction of further research in this area. It has classified the methods based on some key dimensions by extending the work in [3] and assessed them based on specific evaluation criteria. The evaluation results obtained were then analysed to infer some interesting conclusions, such as how effective these methods are.

References

1. Klimburg-Witjes, N., Wentland, A.: Hacking humans? Social engineering and the construction of the “deficient user” in cybersecurity discourses. *Sci. Technol. Hum. Values* **46**, 1316–1339 (2021)
2. Khalid, A., Nazir, M., Hussain, S., Asim, M.: A comprehensive review of social engineering attacks and defense mechanisms. *J. Inf. Secur.* (2016)
3. Heartfield, R., Loukas, G.: A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks. *ACM Comput. Surv.* **48**(3), 1–39 (2016). <https://doi.org/10.1145/2835375>
4. Odeh, A.E.N.A., Eleyan, D.: A survey of social engineering attacks: detection and prevention tools (2021)
5. Aldawood, H., Skinner, G.: Reviewing cyber security social engineering training and awareness programs—pitfalls and ongoing issues. *Fut. Internet* **11**(3), 73 (2019). <https://doi.org/10.3390/fi11030073>
6. Greamo, C., Ghosh, A.: Sandboxing and virtualization: modern tools for combating malware. *IEEE Secur. Priv.* **9**(2), 79–82 (2011)
7. Ghafir, I., Prenosil, V., Svoboda, J., Hammoudeh, M.: A survey on network security monitoring systems, pp. 77–82, August 2016
8. Subha, T., Jayashri, S.: Efficient privacy preserving integrity checking model for cloud data storage security. In: 2016 Eighth International Conference on Advanced Computing (ICoAC), pp. 55–60 (2017)
9. Xue, M., Yuan, C., Wu, H., Zhang, Y., Liu, W.: Machine learning security: threats, countermeasures, and evaluations. *IEEE Access* **8**, 74720–74742 (2020)
10. Samakovitis, G., Petridis, M., Lansley, M., Polatidis, N., Kapetanakis, S., Amin, K.: Seen the villains: detecting social engineering attacks using case-based reasoning and deep learning, July 2019
11. Sedjelmaci, H., Senouci, S.-M., Ansari, N., Boualouache, A.: A trusted hybrid learning approach to secure edge computing. *IEEE Consum. Electron. Mag.* **11**(3), 30–37 (2022)
12. Krombholz, K., Hobel, H., Donko-Huber, M., Weippl, E.: Advanced social engineering attacks. *J. Inf. Secur. Appl.* **22**, 10 (2014)
13. Peltier, T.R.: Information Security Policies, Procedures, and Standards: Guidelines for Effective Information Security Management (2001)
14. Frauenstein, E.D., von Solms, R.: An enterprise anti-phishing framework, March 2011
15. Kumaraguru, P.: PhishGuru: a system for educating users about semantic attacks, p. 199, April 2009
16. Arachchilage, N.A.G., Love, S., Scott, M.: Designing a mobile game to teach conceptual knowledge of avoiding ‘phishing attacks’. *Int. J. e-Learn. Secur.* **2**(1), 127–132 (2012). <https://doi.org/10.20533/ijels.2046.4568.2012.0016>
17. Lin, E., Greenberg, S., Trotter, E., Ma, D., Aycock, J.: Does domain highlighting help people identify phishing sites?, pp. 2075–2084, May 2011
18. Lee, J., Bauer, L., Mazurek, M.: Studying the effectiveness of security images in internet banking. *IEEE Internet Comput.* **13** (2015)
19. Kritzinger, E., von Solms, S.H.: Cyber security for home users: a new way of protection through awareness enforcement. *Comput. Secur.* **29**(8), 840–847 (2010)
20. Anderson, B., Kirwan, B., Jenkins, J., Eargle, D., Howard, S., Vance, A.: How polymorphic warnings reduce habituation in the brain: insights from an fMRI Study, pp. 2883–2892, April 2015

21. Barth, A., Reis, C.: The security architecture of the chromium browser (2009)
22. Mozilla Wiki-Security/Sandbox (2015)
23. The chromium projects-sandbox (2015)
24. Lu, L., Yegneswaran, V., Porras, P., Lee, W.: BLADE: an attack-agnostic approach for preventing drive-by malware infections, pp. 440–450, October 2010
25. Bianchi, A., Corbetta, J., Invernizzi, L., Fratantonio, Y., Kruegel, C., Vigna, G.: What the app is that? Deception and countermeasures in the android user interface, pp. 931–948, July 2015
26. Desmond, R.A.B., Richards, J., Lowe-Norris, A.G.: Active Directory, 5th edn. (2013)
27. Motiee, S., Hawkey, K., Beznosov, K.: Do windows users follow the principle of least privilege? Investigating user account control practices, July 2010
28. Salem, M.B., Stolfo, S.J.: Modeling user search behavior for masquerade detection. In: Sommer, R., Balzarotti, D., Maier, G. (eds.) RAID 2011. LNCS, vol. 6961, pp. 181–200. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23644-0_10
29. Lu, L., Perdisci, R., Lee, W.: SURF: detecting and measuring search poisoning, pp. 467–476, October 2011
30. Li, Z., Alrwais, S., Xie, Y., Yu, F., Wang, X.: Finding the linchpins of the dark web: a study on topologically dedicated hosts on malicious web infrastructures, pp. 112–126, May 2013
31. Lee, S., Kim, J.: WARNINGBIRD: detecting suspicious URLs in Twitter stream, January 2012
32. Udzir, N., Samsudin, K.: Towards a dynamic file integrity monitor through a security classification. *Int. J. New Comput. Archit. Appl. (IJNCAA)* **3**, 789–802 (2011)
33. Dhanalakshmi, R., Chellappan, C.: Detection and recognition of file masquerading for e-mail and data security. In: Meghanathan, N., Boumerdassi, S., Chaki, N., Nagamalai, D. (eds.) CNSA 2010. CCIS, vol. 89, pp. 253–262. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14478-3_26
34. Hara, M., Yamada, A., Miyake, Y.: Visual similarity-based phishing detection without victim site information, pp. 30–36, May 2009
35. Bhardwaj, T., Sharma, T.K., Pandit, M.R.: Social engineering prevention by detecting malicious URLs using artificial bee colony algorithm. In: Pant, M., Deep, K., Nagar, A., Bansal, J.C. (eds.) Proceedings of the Third International Conference on Soft Computing for Problem Solving. AISC, vol. 258, pp. 355–363. Springer, New Delhi (2014). https://doi.org/10.1007/978-81-322-1771-8_31
36. Singhal, P., Raul, N.: Malware detection module using machine learning algorithms to assist in centralized security in enterprise networks. *Int. J. Netw. Secur. Appl.* **4**, 61–67 (2012)
37. Sandouka, H., Cullen, A., Mann, I.: Social engineering detection using neural networks, pp. 273–278, January 2009
38. Basnet, R., Mukkamala, S., Sung, A.H.: Detection of phishing attacks: a machine learning approach. In: Prasad, B. (eds.) Soft Computing Applications in Industry. Studies in Fuzziness and Soft Computing, vol. 226, pp. 373–383. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-77465-5_19
39. Raskin, V., Rayz, J., Hempelmann, C.: Ontological semantic technology for detecting insider threat and social engineering. In: Proceedings New Security Paradigms Workshop, September 2010
40. Xiang, G., Hong, J., Rose, C.P., Cranor, L.: CANTINA+: a feature-rich machine learning framework for detecting phishing web sites. *ACM Trans. Inf. Syst. Secur.* **14**(2), 1–28 (2011)

41. Cova, M., Krügel, C., Vigna, G.: Detection and analysis of drive-by-download attacks and malicious JavaScript code, pp. 281–290, April 2010
42. Aggarwal, A., Rajadesingan, A., Kumaraguru, P.: PhishAri: automatic realtime phishing detection on Twitter. In: eCrime Researchers Summit, eCrime, January 2013
43. Stringhini, G., Thonnard, O.: That ain't you: blocking spearphishing through behavioral modelling. In: Almgren, M., Gulisano, V., Maggi, F. (eds.) DIMVA 2015. LNCS, vol. 9148, pp. 78–97. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20550-2_5
44. Basit, A., Zafar, M., Liu, X., Javed, A.R., Jalil, Z., Kifayat, K.: A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommun. Syst.* **76**(1), 139–154 (2020). <https://doi.org/10.1007/s11235-020-00733-2>
45. Maurya, S., Jain, A.: Deep learning to combat phishing. *J. Stat. Manag. Syst.* **23**, 07 (2020)
46. Subasi, A., Molah, E., Almkallawi, F., Chaudhery, T.J.: Intelligent phishing website detection using random forest classifier, pp. 1–5, November 2017
47. Abdelhamid, N., Thabtah, F., Abdel-jaber, H.: Phishing detection: a recent intelligent machine learning comparison based on models content and features, pp. 72–77, July 2017
48. Mao, J., et al.: Detecting phishing websites via aggregation analysis of page layouts. *Procedia Comput. Sci.* **129**, 224–230 (2018)
49. Lansley, M., Polatidis, N., Kapetanakis, S.: SEADer: a social engineering attack detection method based on natural language processing and artificial neural networks. In: Nguyen, N.T., Chbeir, R., Exposito, E., Anioré, P., Trawiński, B. (eds.) ICCCI 2019. LNCS (LNAI), vol. 11683, pp. 686–696. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28377-3_57
50. Begum, A., Badugu, S.: A study of malicious URL detection using machine learning and heuristic approaches. In: Satapathy, S.C., Raju, K.S., Shyamala, K., Krishna, D.R., Favorskaya, M.N. (eds.) *Advances in Decision Sciences, Image Processing, Security and Computer Vision. LAIS*, vol. 4, pp. 587–597. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-24318-0_68
51. Chouhan, A.Y., Fatima, R., Liu, L., Yasin, A., Wang, J.: Contemplating social engineering studies and attack scenarios: a review study. *Secur. Priv.* **2**, e73 (2019)
52. Al-Hamar, Y., Kolivand, H., Tajdini, M., Saba, T., Ramachandran, V.: Enterprise credential spear-phishing attack detection. *Comput. Electr. Eng.* **94**, 107363 (2021)
53. Fatima, R., Chouhan, A.Y., Liu, L., Wang, J.: How persuasive is a phishing email? A phishing game for phishing awareness. *J. Comput. Secur.* **27**, 1–32 (2019)
54. Chiew, K.L., Yong, K., Tan, C.C.L.: A survey of phishing attacks: their types, vectors and technical approaches. *Exp. Syst. Appl.* **106**, 1–20 (2018)
55. Yao, W., Ding, Y., Li, X.: LogoPhish: a new two-dimensional code phishing attack detection method, pp. 231–236, December 2018
56. Mao, J., et al.: Phishing page detection via learning classifiers from page layout feature. *EURASIP J. Wirel. Commun. Netw.* **2019**, 43 (2019). <https://doi.org/10.1186/s13638-019-1361-0>
57. Sahingoz, O., Buber, E., Demir, O., Diri, B.: Machine learning based phishing detection from URLs. *Exp. Syst. Appl.* **117**, 345–357 (2019)
58. Adebowale, M., Lwin, K., Sanchez, E., Hossain, A.: Intelligent web-phishing detection and protection scheme using integrated features of images, frames and text. *Exp. Syst. Appl.* **115**, 300–313 (2018)

59. Pandey, A., Gill, N., Sai Prasad Nadendla, K., Thaseen, I.S.: Identification of phishing attack in websites using random forest-SVM hybrid model. In: Abraham, A., Cherukuri, A.K., Melin, P., Gandhi, N. (eds.) ISDA 2018 2018. AISC, vol. 941, pp. 120–128. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-16660-1_12
60. Niranjana, A., Haripriya, D.K., Pooja, R., Sarah, S., Deepa Shenoy, P., Venugopal, K.R.: EKRV: ensemble of kNN and random committee using voting for efficient classification of phishing. In: Pati, B., Panigrahi, C.R., Misra, S., Pujari, A.K., Bakshi, S. (eds.) Progress in Advanced Computing and Intelligent Engineering. AISC, vol. 713, pp. 403–414. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-1708-8_37
61. Patil, V., Thakkar, P., Shah, C., Bhat, T., Godse, S.P.: Detection and prevention of phishing websites using machine learning approach, pp. 1–5, August 2018
62. Flowerday, S.: Information security policy development and implementation: a content analysis approach, July 2014
63. Lee, J., Bauer, L., Mazurek, M.L.: The effectiveness of security images in internet banking. *IEEE Internet Comput.* **19**(1), 54–62 (2015)
64. Heartfield, R., Loukas, G.: A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks. *ACM Comput. Surv.* **48**, 02 (2016)
65. Rifat, N., Ahsan, M., Chowdhury, M., Gomes, R.: BERT against social engineering attack: phishing text detection, pp. 1–6, May 2022
66. Wang, Z., Ren, Y., Zhu, H., Sun, L.: Threat detection for general social engineering attack using machine learning techniques, March 2022
67. Arrieta, A.B., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2019)