IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 January 2024, accepted 24 January 2024, date of publication 26 January 2024, date of current version 2 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3359030

RESEARCH ARTICLE

CSE-ARS: Deep Learning-Based Late Fusion of Multimodal Information for Chat-Based Social Engineering Attack Recognition

NIKOLAOS TSINGANOS[1], PANAGIOTIS FOULIRAS[1], IOANNIS MAVRIDIS[1], AND DIMITRIS GRITZALIS[2]

[1]Department of Applied Informatics, University of Macedonia (UoM), 546 36 Thessaloniki, Greece
[2]Department of Informatics, Athens University of Economics and Business (AUEB), 104 34 Athens, Greece

Corresponding author: Dimitris Gritzalis (dgrit@aueb.gr)

This work was supported in part by the Hellenic Ministry of Digital Governance to the Research Center of Athens University of Economics and Business, from 2022 to 2024.

ABSTRACT With the increasing prevalence of chat-based social engineering (CSE) attacks targeting unsuspecting users, the need for robust defenses has never been more critical. In this paper, we introduce Chat-based Social Engineering Attack Recognition System (CSE-ARS), an innovative and effective CSE defense system. CSE-ARS employs a late fusion strategy that integrates the findings of five specialized deep learning models, each focused on detecting distinct CSE attack enablers: critical information leakage recognizer (CRINL-R), personality traits recognizer (PERST-R), dialogue acts recognizer (DIACT-R), persuasion recognizer (PERSU-R), persistence recognizer (PERSI-R). The system harnesses weighted linear aggregation and employs simulated annealing with 10-fold cross-validation, ensuring optimal model performance. CSE-ARS is trained on the CSE-ARS Corpus, a carefully curated dataset tailored to the intricacies of CSE attacks. Extensive evaluation reveals that CSE-ARS achieves satisfactory results in identifying and neutralizing CSE threats, enhancing user security in online interactions.

INDEX TERMS Corpus, cybersecurity, deep learning, natural language processing, social engineering.

I. INTRODUCTION

Social engineering is a multifaceted technique that manifests both in real-life and digital environments. It is related to a manipulative form of communication that exploits human personality traits either in a mass personal or interpersonal way [1], [2]. From ''fake news'' to psychographic advertisements and from cognitive hacking to spear-phishing, there is a plethora of different types of social engineering attacks [3], [4], [5]. In the current year, system intrusion, basic web application attacks, and social engineering have emerged as the predominant attack patterns, collectively responsible for 90% of reported breaches as presented in 2023 Data Breach Investigations Report [6]. Notably, Business Email Compromise (BEC) attacks, which essentially fall under the pretexting category, have experienced a significant surge in 2023, now constituting over 50% of incidents within the social engineering pattern. Furthermore, a substantial 74% of all breaches involve a human element, with individuals being implicated through errors, privilege misuse, stolen credentials, or involvement in other social engineering tactics. The increase in social engineering incidents compared to the previous year can be attributed primarily to the widespread adoption of pretexting, with occurrences almost doubling since the preceding year. Moreover, the median monetary loss resulting from these incidents has also risen over the past couple of years, reaching a substantial $50,000. Consequently, social engineering remains a prominent threat, ranking among the top three attack patterns, accounting for 17% of reported breaches and 10% of incidents. It's important to clarify the distinction between phishing and the more intricate forms of social engineering. If you received an email with a suspicious attachment or a malicious link, urging you to update your password, this is a classic example of

The associate editor coordinating the review of this manuscript and approving it for publication was Dominik Strzalka.

phishing, constituting 44% of all social engineering incidents. Now, if you have received an email or a direct message on social media from a friend or family member urgently requesting financial assistance then this exemplifies social engineering, specifically pretexting, which demands a higher level of skill. Proficient social engineers can manipulate your thoughts, making you believe that a loved one is in distress. Leveraging information, they've gathered about you and your close associates, they create a convincing scenario that preys on your emotions and induces a sense of urgency. Pretexting has now surpassed phishing in terms of prevalence within social engineering incidents. Contemporary practices of attackers have been studied thoroughly [4], [5], [7], [8] and numerous solutions have been suggested within the academic domain as well as within the public sector [9], [10]. Recently, CSE attacks increased due to the wide-spread use of electronic medium communication tools that is boosted due to the COVID-19 pandemic [11] and are now considered mainstream. We define as CSE attack, the attack that involves manipulation of individuals through various forms of online communication such as instant messaging, or social media, to deceive them into revealing sensitive information, taking harmful actions, or compromising their security in some way. These attacks exploit human psychology and personality traits to achieve the attacker's goals, often by impersonating a trusted entity or creating a sense of urgency or fear. CSE attacks are tightly related to pretexting, in which a storyline is methodically planned out in advance and the attacker builds a persona with specific characteristics to approach the human target. The most known pretexts were created by Kevin Mitnick, a legendary social engineer, whose stories met wide media coverage and can be found in [12]. In CSE attacks, pretexts often exploit a simple fact: human personality has vulnerabilities that can be exploited broader using cultural dynamics, social stereotypes, or gender roles. The complexity of the phenomenon imposes an interdisciplinary approach to deal with the many different factors that are engaged. Although several cyber-defense mechanisms have been proposed [5], [13], [14], [15], [16] to defend from social engineering attacks, most of the solutions focus on technical countermeasures to improve users' protection. Being technical these mechanisms do not account for known CSE attack enablers as identified and illuminated in [17]. Currently, the majority of CSE attacks include phishing attacks, pre-texting, social media scams, covid-19 scams, and whaling. The attackers' favorite method of approach is impersonation where they pretend to be someone else to deceive and gain unauthorized access to sensitive information. Furthermore, the attackers recently use artificial technology techniques to create realistic videos and audio of individuals known as deepfakes [18], [19], [20] which can be used to impersonate them in CSE attacks. The intricate interplay between human psychology and CSE attacks underscores the importance of an interdisciplinary approach to cybersecurity. By integrating insights from Cialdini's principles of Persuasion [21], [22], [23] and the Big Five Theory of Personality [24], [25], [26] we can develop more robust defense mechanisms and ultimately reduce the success rate of CSE attacks. CSE attacks are a growing threat that can lead to various negative consequences, including financial loss, damage to reputation, loss of productivity, or legal consequences. Financial loss can occur due to fraudulent transactions or data breaches, which can be costly for organizations. A successful CSE attack can damage the reputation of an organization, leading to a decline in business and loss of customer trust. Loss of productivity can occur when important data is lost, leading to the need for system rebuilding, and decreased efficiency. Legal and regulatory penalties can be incurred if an organization is held liable for data breaches. Given the significant impact of CSE attacks, it is crucial to develop effective detection and prevention mechanisms. These systems should be cost-effective, easy to use, and enable individuals and organizations to better protect themselves against imminent risks. In this paper:

- We propose CSE-ARS, a chat-based social engineering recognition system that employs an interdisciplinary approach, considering personality, linguistic, behavioral, and information technology characteristics [27].
- The ensemble includes different deep learning models, such as RNNs, CNNs, and transformers, for recognizing various CSE attack enablers. Recognizers like CRINL-R [28], PERST-R, DIACT-R [29], PERSU-R [30], and PERSI-R [31] are utilized.
- We introduce an augmented CSE Corpus for training and evaluating CSE-ARS [28].
- We utilize a late fusion approach that combines predictions from individual recognizers, leveraging their strengths.
- We utilize an optimization approach that involves weighted linear aggregation with weights optimized using simulated annealing and k-fold cross-validation.

The rest of the paper is organized as follows: Section II gives the required background information around the context of this work. Section III presents the related work of fellow researchers trying to detect social engineering attacks. In section IV CSE-ARS' design considerations and architecture decisions are described. In Section V, we briefly present the system's individual recognisers each of which is designed to recognize a different CSE attack enabler. Section VI details the training of CSE-ARS. Section VII presents the evaluation results obtained from testing the proposed system. In Section VIII, we discuss the findings and limitations of this study. Finally, Section IX concludes the paper by summarizing the contributions of this work and highlighting future directions for research.

II. BACKGROUND

A. SOCIAL ENGINEERING

Social engineering [5], [32] is a tactic used by attackers to manipulate individuals into divulging sensitive information or performing actions that may compromise cyber security. It is a common tactic used in cyber-attacks, and it relies

on exploiting human psychology and behavior. Social engineering attacks can take various forms, including phishing, pretexting, baiting, and quid pro quo [33], [34]. These attacks can be conducted through various channels such as chat, email, phone, or in-person interactions. Recent years have seen an increase in social engineering attacks, and they are becoming more sophisticated and harder to detect. Traditional technical security measures such as firewalls and antivirus software are not effective in preventing social engineering attacks, and there is a growing need for new techniques to detect and prevent such attacks. Machine learning and deep learning (DL) techniques have been utilized as a solution for detecting social engineering attacks. These techniques have been shown to be effective in identifying patterns and anomalies in text, which can be used to detect CSE attacks. Detection methods also involve natural language processing (NLP) to identify specific queries, commands, or predefined blacklisted topics. Attackers often use social networking sites to gather information and manipulate their targets. Automated data retrieval from semi-structured web pages is a common strategy used to approach targets with useful information. To achieve high detection accuracy and low false negative rates, it is crucial to include as many influencing factors as possible, such as psychological profiles of interlocutors, persuasion techniques etc. However, these techniques alone are not sufficient to detect sophisticated CSE attacks, and there is a need for an interdisciplinary approach that combines the results of multiple recognizers. A very common and dangerous type of CSE attack is pretexting, which targets specific victims and compromises their confidential data to get access into a sensitive system. Pretexting attacks [35] are usually unleashed by sending malware or sending a URL link to the targets using fake identities to manipulate the victim [36], [37].

B. PERSUASION

Robert Cialdini in his seminal work [21], [22], [23] in the field of social psychology, sheds light on the psychological mechanisms that social engineers often exploit. He identifies the principles of persuasion, namely Reciprocity, Commitment and Consistency, Social Proof, Authority, Liking, and Scarcity, which serve as powerful tools in the arsenal of social engineers. Social engineers employ the principle of reciprocity by offering small favors or gifts, creating a sense of indebtedness. Victims, influenced by their innate inclination to reciprocate, may inadvertently provide access or information. Exploiting the desire for commitment and consistency, attackers manipulate individuals into taking small initial actions that align with the attackers' ultimate goals. Once committed, individuals tend to stay on the course, even if it leads to further compromises. Social engineers, also, frequently use social proof to convince targets that their actions are in alignment with those of a larger group. The fear of missing out or the desire to conform to perceived social norms can lead individuals to make ill-advised decisions. Furthermore, people tend to defer to authority figures. Attackers posing as trusted individuals or experts can easily gain victims' trust and coerce them into revealing confidential information or engaging in harmful actions. Building rapport and forging a personal connection are paramount for social engineers. By cultivating a sense of liking or familiarity, attackers can manipulate victims into lowering their guard. Finally, creating a perception of limited availability or urgency is a potent tactic. Individuals, driven by their fear of missing out on opportunities, are more susceptible to manipulation under such circumstances.

C. BIG-5 THEORY

Personality, in psychology, refers to an individual's unique and enduring pattern of thoughts, feelings, and behaviors. The Big Five Personality Theory [25], [38], also known as the Five-Factor Model, identifies five fundamental dimensions of personality, namely Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (often abbreviated as OCEAN). The degree of susceptibility to CSE attacks has also been connected to the five personality traits. Responsible behavior concerning security best practices has been positively correlated with low openness [39], meaning that high levels of openness could potentially facilitate risky security behavior. Lower levels of Conscientiousness have been found to predict deviant workplace behavior such as irresponsible conduct or rule-breaking [26]. High levels of Extraversion have been shown to be predictive of increased vulnerability to phishing attacks [40]. Agreeableness has consistently been associated with phishing in multiple studies [41], [42]. Agreeable individuals may be more susceptible to manipulation due to their tendency to establish trust with the target, which is a characteristic of agreeableness. On the other hand, lower levels of Neuroticism are associated with higher susceptibility [43].

D. DEEP LEARNING & EVALUATION METRICS

Deep learning is a subset of machine learning and it is a cutting-edge approach that has revolutionized artificial intelligence by enabling computers to automatically learn and extract intricate patterns from large datasets. This technology is inspired by the structure and function of the human brain's neural networks, consisting of multiple layers of interconnected nodes. Deep learning models, known as neural networks, excel in tasks such as image and speech recognition, and natural language processing. To rigorously assess the performance of deep learning models and ensure their effectiveness, evaluation metrics play a vital role. These metrics, including accuracy, precision, recall, F1 score, and others, provide quantitative measures that allow researchers to gauge how well their models are performing and make informed decisions about refining and optimizing them. We used several performance metrics to evaluate the quality of the proposed deep learning models, and to assess how well they can make accurate predictions on new, unseen data.

More specifically, throughout our experiments, the following performance metrics were used:

- Receiver Operating Characteristic (ROC) is a graphical representation of the performance of a binary classifier model at different classification thresholds. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings, where the TPR is the proportion of actual positive instances that are correctly identified by the model, and the FPR is the proportion of negative instances that are incorrectly identified as positive.
- Accuracy is a measure of the proportion of correct predictions made by the model, expressed as a percentage. It is calculated as the ratio of the number of correct predictions to the total number of predictions made by the model.
- Precision is a measure of the proportion of true positive predictions among all positive predictions made by the model. It is calculated as the ratio of the number of true positive predictions to the total number of positive predictions made by the model.
- Recall, also known as sensitivity, is a measure of the proportion of true positive predictions among all actual positive instances. It is calculated as the ratio of the number of true positive predictions to the total number of actual positive instances
- F1, is an evaluation measure that combines precision and recall to assess the performance of a binary classification model. It provides a single numerical value that represents the harmonic mean of precision and recall, giving equal importance to both measures. The F1 metric is particularly useful when there is an imbalance between the positive and negative classes in the dataset.
- Active Intent Accuracy: The fraction of user turns for which the active intent has been correctly predicted.
- Requested Slot F1: The macro-averaged F1 score for requested slots overall eligible turns. Turns with no requested slots in ground truth and predictions are skipped.
- Average Goal Accuracy: For each turn, we predict a single value for each slot present in the dialogue state. This is the average accuracy of predicting the value of a slot correctly.
- Joint Goal Accuracy: This is the average accuracy of predicting all slot assignments for a given service in a turn correctly. Also, Harmonic mean between seen and unseen classes.

III. RELATED WORK

Several efforts have been made to detect social engineering attacks and mitigate the risk of being victimized. We can dissect these research works based on the approach taken. Until 2017 there were a lot of works based on statistical learning and machine learning. These works were mainly focused on semantic characteristics of the natural language and were utilizing traditional machine learning algorithms. In 2005, In [44] and [45] the authors introduced the Social Engineering Defense Architecture (SEDA) which is designed to detect social engineering attacks during real-time phone conversations. Although the model was successful in detecting the attacks, it did not incorporate previous activity history or personality traits recognition of both the attacker and the victim. In their 2010 paper, researchers [46] introduced an architecture named the Social Engineering Attack Detection Model (SEADM), which assists users in making decisions through the use of a simple binary decision tree model. However, the researchers rely on several unrealistic assumptions to justify the logic of their proposed system. SEADM was revisited in a subsequent paper in [47] where the system was adapted to account for social engineering attacks that include unidirectional, and bidirectional communication between the attacker and the victim. The work of [42] proposed a taxonomy of social engineering attacks based on Cialdini's Influence principles. They investigated the relationship between the Big-5 Theory, which pertains to personality traits, and Cialdini's influence principles. They proposed a Social Engineering Personality Framework (SEPF) and outlined a complete research roadmap for future work on SEPF. The findings of [48], suggest that the most successful social engineering attacks involve a conversation between the attacker and the victim. Their methodology involves utilizing a predefined Topic Blacklist (TBL) to check dialogue sentences. The authors report achieving a precision rate of a recall rate of 88.9% with their approach. The study of [49] expanded upon the aforementioned work by implementing advanced language processing techniques that balance syntactic and semantic analysis. The reported results demonstrate 100% precision and 60% recall. However, they used a small dataset with only three conversations which limits the precision and recall's success as a measure. Moreover, the algorithm did not consider any contextual information during the classification process, making it unaware of the specific environment in which it operates.

Recent developments in the field of social engineering attack recognition prominently employ DL and NLP tools and methodologies, significantly enhancing their effectiveness. In 2018, in their work [50], propose a methodology identifying malicious statements through the application of NLP techniques. The authors conducted an analysis of these statements with the specific aim of discerning those that may indicate a phishing attack. The detection of malicious intent within these statements is achieved through a comprehensive analysis. To evaluate the effectiveness of their proposed algorithm, the researchers employed a benchmark phishing email dataset as the basis for their assessment. In the works conducted by [51], [52], and [53] NLP techniques and neural networks were employed to identify instances of social engineering attacks. The researchers outlined an approach in which both offline and online textual materials were subjected to NLP processing, followed by analysis using artificial neural networks to distinguish between genuine

content and potential social engineering attacks. The initial phase involved parsing the text and applying NLP techniques to assess syntactical and grammatical aspects. Subsequently, an artificial neural network was utilized to classify potential instances of social engineering attacks. Their proposed methodology achieved high levels of accuracy during the evaluation phase, which involved the utilization of both real-world and semi-synthetic datasets for model training. Additionally, the study explored various classification models to provide a comparative analysis of the two datasets. In another study [54] introduced a two-stage DL model that relies on NLP techniques for the detection of social engineering attacks. This model specifically focuses on identifying the principles of persuasion, drawing from the work of Cialdini. To evaluate its effectiveness, the authors employed a semi-synthetic dataset and demonstrated that their approach achieved highly accurate results in detecting social engineering attacks. In the research conducted by [55], the focus was on examining conversational agents, specifically chatbots, designed to influence and change people's opinions. To accomplish this, the study utilized an annotated dataset derived from human dialogues. The authors made predictions regarding ten distinct persuasion strategies, and they integrated these predictions with the demographic and psychological profiles of the conversational partners. Their approach employed a hybrid region-based convolutional network model, incorporating three types of features: sentence embedding, context embedding, and sentence-level features. This model yielded favorable results in the prediction of persuasion strategies within the context of opinion change. However, it's important to note that the proposed solution is not suitable for addressing CSE attacks. This limitation arises because the persuasion methods employed in CSE attacks typically revolve around tactics related to authority and commitment, which are not fully covered by the strategies investigated in this study. In the work by [56], the authors introduced a neural network architecture designed to quantify persuasiveness and identify persuasive strategies. Their proposed model outperformed several baseline learners and offered improved interpretability. The architecture of this model included a semi-supervised neural network comprising a sentence encoder and a document encoder. It was trained using a custom dataset, ultimately leading to the identification of five distinct persuasion strategies. It is noteworthy that while their approach yielded comparable performance results to the one discussed in our study, it also exhibited a growing complexity in its architectural design. In their research documented in [57], researchers propose a method for capturing both the syntactic and semantic attributes of natural language by harnessing a pre-trained BERT model. The proposed model exhibits a robust resilience to adversarial attacks when attackers intentionally substitute keywords with synonyms. In a recent study [58], the authors present a system designed to safeguard against CSE attacks by implementing a series of NLP components within a pipeline. These NLP components encompass NER, dialogue engineering, stylometry, and the utilization of ask and framing questions. The system employs an active defense strategy to identify the social engineer's intent, subsequently diverting her attention and resources, thus mitigating the potential harm. In [59], a URL classifier is introduced that leverages Random Forest models and gradient boosting classifiers. This URL classifier was designed with the objective of enhancing the algorithm's effectiveness in identifying malicious websites. To achieve this goal, the classifier incorporated features associated with both the host and linguistic attributes of the URL. Employing machine learning algorithms, the researchers were able to significantly reduce the time required to detect malicious URLs. As a result, their approach provides real-time protection for safe web browsing while simultaneously conserving computational resources. The work of [60], proposes a social engineering detection model, relying on deep neural networks. This model demonstrates the capability to identify instances of deception and phishing attempts through the analysis of textual content. In its initial phase, the chat history is subjected to processing and analysis utilizing NLP techniques, while the contextual semantics are extracted and explored through a bi-directional Long Short-Term Memory Model (bi-LSTM). Additionally, the incorporation of user characteristics and chat content attributes as features for classification is achieved using ResNet. The findings of [5], investigated the performance of nine distinct machine learning models, which were trained using three separate datasets. They extracted features related to threats and evaluated them against a set of twenty-seven threat detectors. The primary objective was to identify general social engineering attacks that do not target specific techniques. While the results of their research show promise, it's worth noting that the study's broad approach does not allow for the detection of persuasion tactics employed within chat-based conversations. In their work [61],the authors introduce a system that goes beyond identifying persuasion cues. It also addresses aspects such as framing, objectivity/subjectivity, guilt/blame, and the use of emphasis in text analysis. To train the learner, the authors employed a custom dataset and integrated traditional NLP tools, including linguistic inquiry and word count (LIWC), topic modeling, and sentiment analysis. These tools contributed to feeding a random forest classifier. The performance of the system was found to be satisfactory when compared to other methods such as Labeled-LDA and Long Short-Term Memory (LSTM). However, it's important to note that the presented system lacks the capacity to leverage modern word representation techniques and the flexibility offered by CNNs or other more recent neural network architectures.

IV. CSE-ARS ARCHITECTURE

A chat may begin with the user or the potential social engineer, who initiates the conversation by entering utterances into the chat software. CSE-ARS deep-learning based system, acts as an intermediary between the two interlocutors which can detect and recognize CSE attacks. As shown in

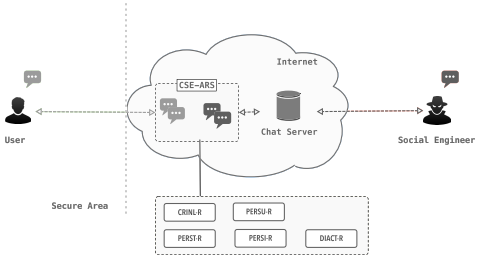

FIGURE 1. Magnetization as a function of applied field. It is good practice to explain the significance of the figure in the caption.

Figure 1 CSE-ARS [27] system captures and analyzes the content of the utterances utilizing deep learning algorithms and techniques and makes inference regarding the existence enablers that can lead to successful social engineering attacks.

There are several enablers that can be triggered in order to achieve a successful CSE attack. By definition [62], [63], an enabler is something that enables or facilitates an action. In the context of CSE attacks, enablers refer to the various tactics and techniques used by attackers to trick their targets into disclosing sensitive information or performing certain actions. These enablers include persuasion techniques, deception techniques, critical information leakage tricks, paraphrasing methods, and specific dialogue acts, which are often used in combination to increase the chances for a successful CSE attack. Recognition of these enablers is important and prevents CSE attacks. In [17] the following critical enablers were identified:

- Critical information leakage: Social engineers may attempt to extract sensitive or confidential information from their targets.
- Personality traits: Social engineers may attempt to exploit certain personality traits of their targets in order to gain their trust.
- Dialogue acts: Social engineers may use certain dialogue acts, such as questions or commands, to elicit information or manipulate the conversation. In this attempt they may utilize previous conversations (with the same interlocutors) in order to deceive them.
- Persuasion attempts: Social engineering attacks often involve attempts to persuade the target to take a certain action.
- Persistent behavior: Social engineers may attempt to paraphrase information they've already obtained in order to confirm its accuracy or to obtain additional information.

We designed, and implemented a resilient and efficient cyber-defense system that predicts whether a chat-based dialogue is evolving into a CSE attack. The proposed system, CSE-ARS, utilizes the recent trends in DL and NLP to examine and classify utterances taking into account the aforementioned CSE attack enablers. CSE-ARS is a synthesis of individual recognizers that utilize a different DL model for each CSE attack enabler. The DL models were evaluated using the same context to draw safe conclusions. This means that we paid special attention to the training data, the hyperparameters, the experimental setup, and the number of experiments conducted for the comparison to be trustworthy. As a principle during our evaluations, we preferred to consider false positives as less important and focused on not having too many false negatives. This trade-off was a basic design decision for all DL models that were tested and evaluated. CSE-ARS brings a multimodal approach to solving the problem of recognizing CSE attacks. In the context of our study, multimodal refers to the use of multiple modalities or sources of information to make predictions [64], [65], [66]. The proposed system is utilizing a late fusion approach, which takes each recognizer's output and produces a final prediction through a weighted linear aggregation. The weights are determined using k-fold cross validation and are optimized to maximize the performance of the system. This allows us to exploit the strengths of each different DL model and to account for the fact that different types of enablers may be utilized to successfully conduct a CSE attack. More analytically, CSE-ARS incorporates the following recognizers:

- Critical Information Leakage Recognizer (CRINL-R): a bi-LSTM NER model [67], recognizing critical data information leakage.
- Personality Recognizer (PERST-R): a BERT [68] model that predicts personality traits of the interlocutors based on the Big-5 theory.
- Dialogue-act Recognizer (DIACT-R): a BERT model that recognizes dialogue acts that can lead to deception taking into account the full chat history.
- Persuasion Recognizer (PERSU-R): a convolutional neural network (CNN) [69] that predicts persuasion attempts.
- Persistence Recognizer (PERSI-R): a BERT model that predicts persistent behavior by identifying paraphrasing in chat dialogue.

After the individual recognizers generate their outputs, the late fusion model combines them to determine whether the chat text constitutes a CSE attack. Each recognizer has its feature extraction network customized for its specific objective and state. The outputs of the recognizers are fused using a weighted linear aggregation method. All the models, including the CSE-ARS but excluding PERST-R, are trained on different variants of CSE corpus [28]. The multimodal fusion architecture of deep learning based CSE-ARS is depicted in **Figure 2**. The system's flexible design allows for the addition or removal of individual recognizers as needed. This, in turn, allows the system to be adapted to new types of social engineering enablers as they emerge or identified. By using a combination of different recognizers, the system captures a wide range of enablers and provides a comprehensive assessment of the likelihood of a particular input being a CSE attack.

Moreover, this architecture allows CSE-ARS to effectively use the strengths of the different recognizers and make a more accurate prediction by combining their results. We used

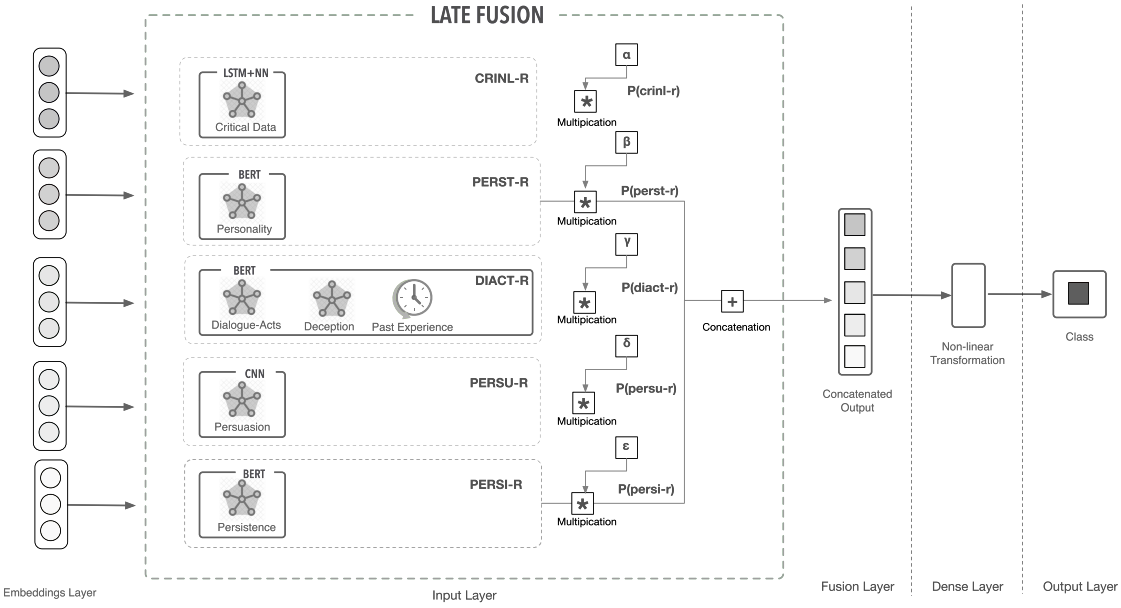

FIGURE 2. CSE-ARS architecture.

concatenation for the fusion layer to combine the outputs of multiple binary recognizers denoting the independence of the different enablers and keeping the system simple. The concatenation layer takes the output of each recognizer and concatenates them into a single vector, which can then be fed into a fully connected layer for the final prediction. The outputs of the recognizers are treated as separate features and the information from each recognizer is retained in the final combined vector. This is useful when the recognizers are designed to capture different aspects of the input and the outputs are independent of each other. Thus, the final combined vector provides a multimodal representation of the input and can be used to make a prediction.

V. CSE-ARS RECOGNIZERS

In this section, we provide a concise summary of each individual recognizer within CSE-ARS, all of which have been previously introduced in our prior works, with the exception of PERST-R, which we introduce in this paper.

A. CRINL-R

CRINL-R, employs NER [70] to identify sensitive information disclosure in chat conversations using a bi-LSTM to extract contextual information from unstructured chat text. CRINL-R is a pre-trained model that is fine-tuned using an appropriately annotated version of CSE corpus, which identifies personal information, information technology (IT) terms, and enterprise information. The terms identified came from the CSE ontology [28], which connects CSE and cybersecurity concepts, focusing on information disclosure during chat interactions. Each sentence was annotated in IOB format, and tokens were represented using a pre-trained English language model from spaCy [71]. A bi-LSTM model is constructed, comprising a bi-directional LSTM layer and a classification layer to capture context from preceding and succeeding tokens. The model is trained using cross-entropy as the loss function and batch-wise gradient descent algorithm. Performance assessment involved applying the trained model to validation data, generating predicted tags for each token, and comparing them to true tags to evaluate the model's accuracy in identifying named entities within chat dialogues. The training of the bi-LSTM neural network involved several sequential steps: token-tag pairs from dialogue sentences were transformed into word embeddings using spaCy. The bi-LSTM model considered prior information at each step, generating an output sequence. A backward pass was performed, and forward and backward outputs were combined and used in a classifier to predict tags (Personal, IT, Enterprise) for input words. This process enabled the bi-LSTM model to learn and predict tags for new data. After training the model for 10 epochs, we reached the scores presented in **Table 1** for each term contained in one of the three categories:

TABLE 1. CRINL-R performance results.

tag	F1	Precision	Recall	Support
PERSONAL	0.67	0.63	0.51	65
IT	0.86	0.93	0.64	122
ENTERPRISE	0.71	0.84	0.56	89

B. PERST-R

In psychology, the term "human personality" denotes the unique variations in patterns of cognition, affect, and

behavior that distinguish one individual from another. The Five-Factor Model (FFM) [26], also known as the Big-5 Theory, represents a widely recognized framework for the classification of personality traits. These factors are typically assessed on a continuum ranging from 0 to 1 [72] and play a significant role in CSE attacks, as they are exploited by attackers to manipulate vulnerable individuals. Several efforts have already been made by researchers to recognize personality traits using DL models [73], [74], [75] The five personality traits have also been linked with high or low susceptibility to CSE attacks. Although earlier work showed contradictory results regarding Openness [76], newer research shows that high Openness has been positively linked to unresponsible behavior regarding security best practices [77], [78], [79], and thus would potentially facilitate dangerous security behavior. Consciousness at low levels predicted deviant workplace behavior in the form of irresponsible conduct or rule-breaking [26]. Extraversion at high level is predictive of increased vulnerability to phishing attacks [40]. Agreeableness is most associated with phishing and multiple studies [41], [42] have reached similar conclusions. Agreeable people may be manipulated by establishing trust with the target, as this represents a facet of agreeableness. Neuroticism at low level leads to higher susceptibility [43]. **Table 2** depicts when a personality trait can lead to increased susceptibility of CSE attacks.

TABLE 2. Personality traits and susceptibility of CSE attacks.

Personality trait	Value	Susceptibility of CSE attacks
Openness	High	High
Consciousness	Low	High
Extraversion	High	High
Agreeableness	High	High
Neuroticism	Low	High

The main objective of the PERST-R model is to identify the personality traits, as defined in the Big-5 theory utilizing the chat dialogue. PERST-R utilizes a pre-trained BERT model which is fine-tuned on a large corpus of text data that has been labeled for each of the five personality traits. The model was trained to predict the likelihood of each of the five personality traits for a given input text. We employed BERT (bert-base-uncased) with a hidden size of 768 and 12 attention heads, encapsulating 12 hidden layers. The model utilizes a GELU [80] activation function, a batch size of 32 with learning rate $3 * 10^{-}5$, and a maximum sequence length of 512 tokens with attention and hidden dropout probability of 0.1. The performance of the trained model was evaluated by measuring its accuracy on a held-out test dataset. The aforementioned workflow is depicted in **Figure 3**.

The training corpus used is the FriendsPersona [81] which is a large-scale conversational dataset that was constructed using scripts from the popular American TV show ‘‘Friends’’. It contains 1,175 dialogues between pairs of characters, totaling 105,784 utterances. The dataset is annotated with the Big-5 personality traits, with each dialogue annotated by three human raters. The FriendsPersona dataset is unique in that it provides both conversational data and personality trait annotations, which enables researchers to explore the relationship between personality traits and conversational behavior. In terms of inter-annotator agreement, the creators achieved an average pair-wise kappa of 54.92% among 2 annotators and Fleiss’ kappa of 20.54% among 3 annotators across five personality traits. **Table 3** presents the corpus details.

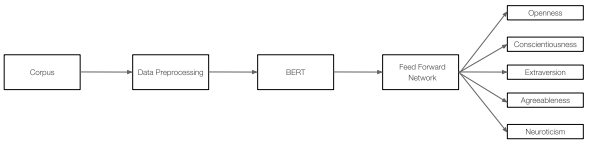

FIGURE 3. PERST-R pipeline.

TABLE 3. FriendsPersona corpus details.

Characteristic	Value
Total dialogues	1,175
Total utterances	105,784
Annotated personality traits	Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism
Annotation method	Three human raters per dialogue
Demographic information	Age, gender, occupation
Relationship labels	Friends, family, romantic partners
Source	Scripts from the TV show "Friends"
Language	English
Release year	2021
License	Creative Commons Attribution 4.0 International (CC BY 4.0)

PERST-R model achieved satisfactory accuracy in recognizing Big-5 personality traits, with an overall accuracy of 71,12%. **Table 4** and **Figure 4** present the performance results and ROC graph. The area under the ROC curve (AUC) for each of the Big-5 personality traits was also satisfactory, with values of 0.83 for Openness, 0.62 for Conscientiousness, 0.79 for Extraversion, 0.57 for Agreeableness, and 0,80 for Neuroticism. These results demonstrate the effectiveness of the PERST-R model in efficiently recognizing Big-5 personality traits and suggest that this approach could be useful in CSE attack recognition.

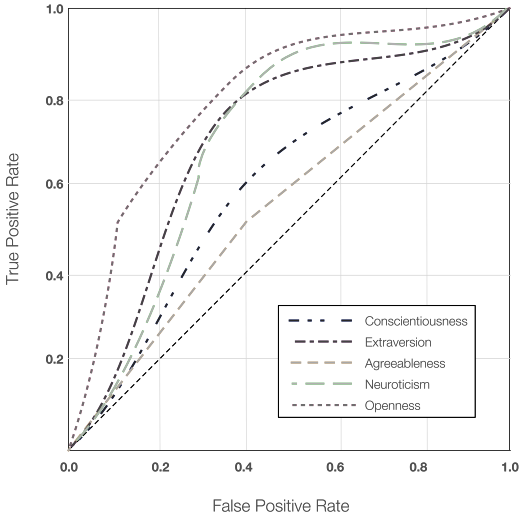

FIGURE 4. PERST-R ROC curves.

TABLE 4. PERST-R accuracy results.

Model	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism	Average
BERT	80,12	66,79	71,09	65,37	72,27	71,12

C. DIACT-R

Dialogue acts (DAs) represent the communicative functions or intentions conveyed during conversations and are exploited by attackers in CSE attacks to manipulate targets. Recognizing dialogue acts in chat text offers advantages in safeguarding against CSE attacks. DIACT-R is built around a BERT model, called SG-CSE BERT, which is fine-tuned on the custom SG-CSE Corpus [29]. It employs a schema-guided paradigm for CSE attack state tracking, utilizing previous interactions to establish a baseline of normal behavior and detect anomalies indicative of CSE attempts. For example, inconsistencies in impersonation attempts can be revealed through monitoring interaction history. For the identification of CSE DAs as shown in **Figure 5**, a Schema-Guided Chat-based Social Engineering Ontology (SG-CSE Ontology) is created using the CSE conceptual model and ontology. Four CSE attack types are extracted and represented as services, each mapped to a specific schema. DIACT-R operates within a schema-guided framework, allowing for conditioning on the CSE attack service schema using intent and slot descriptions. The schema includes multiple intents mapped to corresponding DAs. This schema-guided paradigm leverages domain ontologies to define service schemas, and in this research, the domain is CSE attacks. The system predicts the service, user intention, and requested slot-value pairs, enhancing the ability to detect and thwart CSE attacks in chat-based interactions.

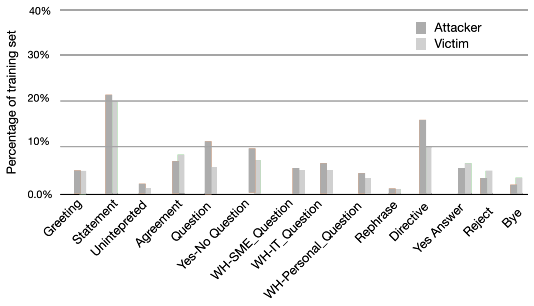

FIGURE 5. Distribution of CSE DAs in the SG-CSE corpus.

DIACT-R uses the schema-guided SG-CSE BERT model to represent unseen intents and slots by encoding them into embedded representations. The model, as shown in **Figure 6**, accommodates varying schema structures in CSE attacks. Input sequences are paired and processed by the SG-CSE BERT encoder, with schema embedded as U_{CLS} and token-level representations as t.

The Hugging Face [82] library and the BERT uncased model with 12 layers, 768 hidden dimensions, and 12 self-attention heads were utilized to form the SG-CSE BERT model. To train the model, we set the batch size at 32 and used a dropout rate of 0.2 for all classification heads. A linear

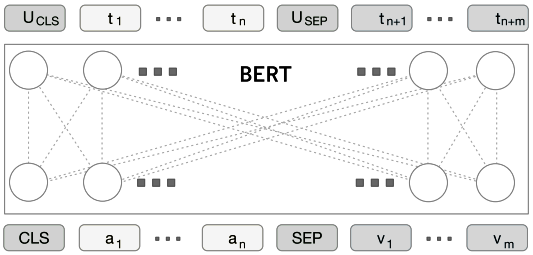

FIGURE 6. The SG-CSE BERT model.

warmup strategy with a duration of 10% of the training steps was employed, as well as the AdamW optimizer [83] with a learning rate of 2e-5. **Table 5** depicts DIACT-R's performance, with Active Intent Accuracy and Requested Slots F1 displaying high efficiency, while Average Goal Accuracy and Joint Goal Accuracy exhibit lower efficiency.

TABLE 5. DIACT-R performance results.

System	Model	Parameters	Active Intent Accuracy	Req Slot F1	Acg Goal Accuracy	Joint Goal Accuracy
DIACT-R	BERT $_{BASE}$	110M	85.2	89.6	74.1	56.7
Seen	BERT $_{BASE}$	110M	53.8	48.3	31.4	24.9
Unseen	BERT $_{BASE}$	110M	69.5	68.9	52.7	40.8

D. PERSU-R

PERSU-R recognizer, identifies persuasive content in utterances. Cialdini's principles of persuasion serve as a foundation for understanding persuasive techniques. These principles are widely recognized in interdisciplinary scientific research on persuasion. PERSU-R combines a CNN and a Multi-Layer Perceptron (MLP) for this purpose. The CNN functions as a feature extractor, capturing local patterns in sentences that may indicate persuasive cues. It applies a nonlinear function to token windows and uses a pooling layer to create a one-dimensional vector, which is then integrated into the overall network architecture. The parameters and values of the filters applied by the CNN are optimized through back-propagation during the training phase, ensuring effective identification of persuasive content in sentences. PERSU-R was trained on the CSE-PUC Corpus [30], which is an appropriately annotated version of the CSE corpus. Before the training process, we specified several hyperparameters, such as the number of filters and filter size, as shown in **Table 6**.

The performance results f the PERSU-R recognizer are presented in **Table 7**.

E. PERSI-R

Paraphrasing involves restating information in different ways and is used by malicious actors to avoid detection while persisting in their efforts to extract sensitive data. To address

TABLE 6. PERSU-R training details.

Description	Values
Batch size	16
Word Embeddings	fastText
Word Embeddings size	300
Filter sizes	$2, 3, 4, 5$
Number of filters	$100, 100, 100, 100$
Stride length	1
Zero padding	Yes
Activation function	ReLU
Pooling method	$1-\max$
Dropout rate	0.1
Training epochs	10

TABLE 7. PERSU-R performance results.

Model	Accuracy	Macro F1
PERSU-R	71.6%	58.2%

this issue, PERSI-R recognizer employs a specialized neural network architecture developed for paraphrase recognition in the context of CSE attacks. PERSI-R is a fine-tuned version of BERT model, specifically tailored for the task of recognizing paraphrasing in the context of CSE attacks. Following the SNLI paradigm [84], each instance of the CSE-Persistence corpus [31] is composed of two sentences and is manually labeled as being a member of one of the following three categories:

- Identical (I): the two sentences are semantically close and share a common term targeting the same leaf entity in the CSE ontology.
- Similar (S): the two sentences are semantically related and share a common intent, which translates into a higher-level entity in the CSE ontology, targeting a different leaf entity
- Different (D): the two sentences are not semantically related, and they do not share a common higher-level or leaf entity.

The PERSI-R model takes pairs of sentences as input and produces a binary classification result, determining whether the sentences are paraphrases or not. To prepare the input sentences for the model, they are tokenized, padded to a fixed length, and passed through the BERT model to generate contextualized word embeddings. The sentence representations are then derived by calculating the mean of the contextualized word embeddings across all tokens within each sentence. This approach enhances the ability to detect and thwart such malicious activities during chat-based interactions. After fine-tuning, PERSI-R compared to the simple BERT-base model achieved the accuracy values presented in **Table 8**.

TABLE 8. PERSI-R performance results.

Model	Training Accuracy (%)	Training Loss (%)	Validation Accuracy (%)	Validation Loss (%)
BERT-base	84.01	0.24	76.79	0.37
PERSI-R	84.96	0.21	78.03	0.36

VI. CSE-ARS TRAINING

To comprehensively assess the effectiveness of our proposed CSE-ARS, a ten-fold cross-validation experiment was conducted. An augmented version of CSE Corpus, called CSE-ARS Corpus, underwent a randomization process, resulting in its division into ten distinct subsets. Throughout each iteration of the experiment, nine subsets were further subdivided into training (80%) and validation (20%) sets, while the remaining subset was designated as the testing set. By employing this methodology, prediction scores for each testing subset were obtained after ten rounds, which were subsequently amalgamated to derive an overall prediction score. To enhance the performance of our approach, several strategies were implemented. Initially, we explored diverse configurations and trained each recognizer using a single domain training set, with the validation set utilized to mitigate the risk of overfitting. Subsequently, the optimal configuration parameters were selected based on the evaluation criterion of the area under the receiver operating characteristic curve (AUC) value. Finally, after the training of the specific recognizers, an exhaustive examination of various coefficient combinations was conducted until the classification performance, as assessed by the AUC value, reached its maximum on the validation set.

A. CSE-ARS CORPUS

CSE-ARS was trained on the CSE-ARS Corpus [27], which is an augmented version of the CSE corpus [28]. The CSE-ARS raw dialogues have been released on DRYAD [85] and is composed of realized and fictional CSE attack dialogues. 9 presents the summary of the CSE-ARS corpus.

TABLE 9. CSE-ARS corpus details.

Characteristic	Value
Corpus Name	CSE-ARS corpus
Collection Method	Web-scraping, pattern-matching text extraction
Corpus size	(N) 180 text dialogues
Vocabulary size	(V) 7106 terms
Total no. of turns	3565
Avg. tokens per turn	9,34
Content	Chat-based dialogues
Collection date	Jun 2018 - Dec 2022
Language	English
Release year	Sep 2023

B. LATE-FUSION

Ensemble techniques offer a valuable means of enhancing system performance by combining multiple machine learning models. Such techniques prove particularly beneficial when individual models exhibit high accuracy but differ in the types of errors they make, as often encountered in multimodal approaches. Within the context of this study, we consider as ''modality'' [64] each distinct acquisition framework that captures information regarding the same phenomenon from various types of recognizers under different conditions, across multiple experiments or subjects. The late fusion

approach, akin to decision-level fusion, constitutes an ensemble technique that combines the output multiple models to generate a final prediction. In late fusion, the unimodal decision values are acquired and merged to derive the ultimate decision. This approach facilitates flexible training and straightforward predictions, even when one or more modalities are absent, albeit at the expense of disregarding certain low-level interactions between modalities.

The output of CSE-ARS recognizers is concatenated and subsequently fed into a final classifier to formulate the conclusive prediction. Leveraging late fusion methodology leads to enhanced performance when compared to utilizing a solitary model, rendering it a commonly adopted technique across various machine learning domains including image classification, NLP, and speech recognition. The key advantage of late fusion lies in its capacity to allow individual recognizers to specialize in their respective areas of expertise, thereby contributing to a more accurate final prediction. If a recognizer m_i is used on modality i using input k_i where $i = 1, 2, \ldots, M$ then the final prediction of a late fusion system is given by $p = f(m_1(k_1), m_2(k_2), \ldots, m_M(k_M))$. The strengths of late fusion systems are relatively simple to implement compared to other models, as they simply combine the outputs of different models into a single prediction.

C. SIMULATED ANNEALING

The novelty of our approach resides in the architectural design and the fusion of multi-dimensional data. The fundamental idea behind our multimodal fusion approach is to integrate the distinct output to enhance CSE attack recognition. The outputs of the individual recognizers are represented by probability distributions, obtained through the application of the SoftMax [86] classifier. The SoftMax classifier formula is depicted in **Formula 1**:

$$softmax(y)_i = \frac{e^{y_i}}{\sum_{j=1}^{n} e^{y_i}} \tag{1}$$

where y_i represent the data in j_{th} column of the output vector and n represents the output vector dimension. The SoftMax layer's output constitutes a probability distribution across the two possible classes, whether an utterance is considered a CSE attack or not. We conducted multimodal fusion based on weighted linear aggregation; the specific construction steps are as follows:

STEP1: The trained CSE-ARS recognizers (CRINL-R, DIACT-R, PERSI-R, PERST-R, and PERSU-R) are applied to the particular validation sets. The output of each recognizer

$$(i.e., output_{CRINL-R}, output_{DIACT-R}, output_{PERSI-R}, output_{PERST-R}, and output_{PERSU-R})$$

has the form of a matrix with dimensions equal to the number of samples in the corresponding validation set and the number of classes.

STEP2: To fuse the results of the individual recognizers, we define the fusion methods shown in **Formulas 2, 3, 4**:

$$\begin{aligned} &output_{CSE-ARS} \\ &= \alpha * output_{CRINL-R} + \beta * output_{DIACT-R} \\ &\quad + \gamma * output_{PERSI-R} + \delta * output_{PERST-R} \\ &\quad + \epsilon * output_{PERSU-R} \end{aligned} \tag{2}$$

$$\alpha + \beta + \gamma + \delta + \varepsilon = 1 \tag{3}$$

$$0 \le \alpha \le 1,\ 0 \le \beta \le 1,\ 0 \le \beta \le 1,\ 0 \le \beta \le 1,\ 0 \le \varepsilon \le 1 \tag{4}$$

where $output_{CSE-ARS}$ represent the result of feature fusion, α represents the weight of the CRINL-R output, β represents the weight of the DIACT-R output, γ represents the weight of the PERSI-R output, δ represents the weight of the PERST-R output and ϵ represents the weight of the PERSU-R output.

STEP3: We find the best combination of $(\alpha,\ \beta,\ \gamma,\ \delta,\ \varepsilon)$ on the validation set so that the cross entropy between $output_{CSE-ARS}$ and label (one-hot encoding) is close to the theoretical minimum $(\alpha,\ \beta,\ \gamma,\ \delta,\ \varepsilon)$. The step is equivalent to a new round of feature learning.

STEP4: In order to find the optimal solution of $(\alpha,\ \beta,\ \gamma,\ \delta,\ \varepsilon)$ on the validation set, we define the optimization problem as shown in **Formulas 5, 6, 7** :

$$min(Loss(output_{CSE-ARS}, Label) \tag{5}$$

$$\alpha + \beta + \gamma + \delta + \varepsilon = 1 \tag{6}$$

$$0 \le \alpha \le 1,\ 0 \le \beta \le 1,\ 0 \le \beta \le 1,\ 0 \le \beta \le 1,\ 0 \le \varepsilon \le 1 \tag{7}$$

To obtain the global optimal solution, we utilize the Simulated Annealing (SA) algorithm [87]. SA is an optimization algorithm used for discovering the global optimum of a complex objective function. The probability of a particular state of x is determined by **Formula 8**:

$$p(x) = e^{\frac{-\Delta f(x)}{kT}} \tag{8}$$

where f(x) is the configuration of energy, k is Boltzmann's constant, and T is temperature. The algorithm that describes the optimization procedure is shown in **Algorithm 1**

To summarize, the Simulated Annealing (SA) algorithm is an optimization algorithm used to find the global optimal solution of a complex objective function. It starts with an initial solution and evaluates it using the objective function. Then, it perturbs the solution and evaluates the new solution. If the new solution is better, it becomes the new current solution. If it's worse, it may still be accepted with a probability based on the temperature parameter. The objective function in this work is defined as shown in **Formula 5** and the SA algorithm is used to determine the optimal values of $(\alpha, \beta, \gamma, \delta, \epsilon)$ on the validation set. Finally, the tuple $(\alpha, \beta, \gamma, \delta, \epsilon)$ is transferred to the test set to predict CSE attacks using the fused features, and the results are compared to those obtained through single modality feature learning.

Algorithm 1 Simulated Annealing

```
procedure Optimize
    Generate a random initial solution x0
    Calculate objective function
    Parameter Initialization (T, k, c)
    while control condition not true do
        for number of new states do
            Pick new solution x0+Δx in neighborhood
            #Evaluate new state
            if f(x0 + Δx) > f(x0) then
                f = f(x0+); x = x + Δx
            else
                Δf = f(x0 + Δx) − f(x0)
                if r>exp(−Δf(x)/(kt)) then
                    f = f(f +Δx), x = x Δx
                else
                    fnew = f(x0
                end if
            end if
            f = fnew
            Decrease the temperature periodically: T = Txc
        end for
    end while
end procedure
```

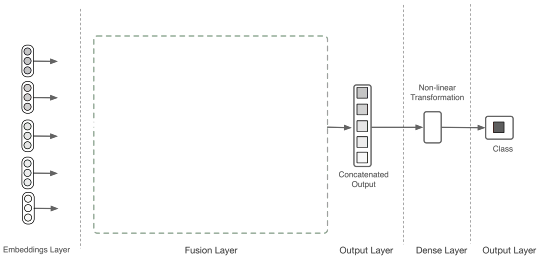

FIGURE 7. Workflow of the CSE-ARS multimodal fusion.

The workflow of the CSE-ARS multimodal fusion is shown in **Figure 7**.

The layers depicted in the figure above are as follows:

- Embeddings Layer: This layer receives input from the individual recognizers.
- Fusion Layer: This layer combines the outputs from the individual recognizers into a single representation. The fusion layer is implemented as a weighted linear aggregation.
- Output Layer: The output of the recognizers is concatenated.
- Dense Layer: This layer is used to apply non-linear transformations to the fused representation to produce the final prediction. The dense layer is implemented as a fully connected layer.
- Inference Layer: This layer produces the final prediction, which could be the probability of a given text being a CSE attack or not.

These steps are repeated until the system reached satisfactory performance. For this specific problem of CSE attack recognition, a binary cross-entropy loss function is used. For the optimizer, our method of choice was the Adam optimizer [88], which adaptively adjusts the learning rate for each parameter based on the historical gradient information. Furthermore, it tends to converge faster and with a better convergence minimum compared to Stochastic Gradient Descent (SGD) [89]. The choice of a binary cross-entropy loss function and the Adam optimizer is a reasonable one for the late fusion system for CSE attack recognition, as it provides a robust and efficient way to measure the error and update the system parameters.

VII. CSE-ARS EVALUATION

During our research, every experiment was tracked as it relates to the progress and the results. Furthermore, the level of detail logged concerning each experiment was at such a level that we could be able to recreate each experiment or compare it at a later time with another experiment. For experiment tracking (e.g., loss curve, model performance metrics, hyperparameters, etc.) and experiment versioning, we utilized the Weight & Biases platform [90]. Different combinations of hyper-parameter values, as expected, gave different performance results. We exhaustively searched all possible combinations to find the optimal hyper-parameter values. The performance of each hyper-parameters set was evaluated against a dedicated validation set. Our test split was never used for hyper-parameter tuning, to avoid overfitting and the best model was selected based on its performance on the validation set. Afterward, this model's performance was measured against the test split. CSE-ARS system has been extensively evaluated on the CSE-ARS corpus.

Due to the lack of similar CSE recognition systems, we utilized a baseline system based on the majority voting [91] method, in which the predictions of all the classifiers are combined and the class label with the highest frequency is selected as the final prediction. By comparing the results of the two models, we aim to show the benefits of using a more sophisticated late fusion model over a simple majority voting ensemble model for CSE attack recognition. The results show that CSE-ARS is outperforming the majority voting ensemble model. For each different recognizer, we suppose that if recognition is true (e.g., persuasion is recognized) then a CSE attack is underway. Thus, there are two different classes we need to predict: CSE-attack and Neutral. The predicted class label $\widehat{y}$, if there are m different classifiers, is given by **Formula 9**

$$\widehat{y} = mode\,\{C_1(x),\ C_2(x),\ \ldots,\ C_m(x)\} \tag{9}$$

We evaluated the model performance of each recognizer via 10-fold cross-validation on the training dataset before we combine them into an ensemble recognizer: e.g., the global optimization outcome for $(\alpha, \beta, \gamma, \delta, \epsilon)$ on the validation set yielded (0.134, 0.294, 0.201, 0.169, 0.202). Upon acquiring the key parameters $(\alpha, \beta, \gamma, \delta, \epsilon)$ for the multimodal fusion model, we tested both the individual recognizers and the fusion model on the test set. Their respective prediction accuracies were 56,70%, 71,20%, 68,70%, 64,90%, 58,90%, and 79,96%. The multimodal fusion model achieved a prediction accuracy of 8.76% greater than the optimal single modality model. The loss value of each model was calculated

TABLE 10. Average accuracy of individual recognizers and ensemble models.

Model	Avg. Accuracy	Avg. Precision	Avg. Recall	Avg. F1-score
PERST-R	0.567	0.512	0.501	0.560
DIACT-R	0.712	0.648	0.633	0.604
PERSU-R	0.687	0.601	0.600	0.653
PERSI-R	0.749	0.667	0.640	0.630
CRINL-R	0.589	0.554	0.550	0.601
Majority Voting	0.753	0.700	0.689	0.701
CSE-ARS	**0, 799**	**0.702**	**0.702**	**0.701**

TABLE 11. AUC values of the CSE-ARS and individual recognizers.

Model	AUC
PERST-R	$0.5479(+/-0.14)$
DIACT-R	$0.6783(+/-0.12)$
PERSU-R	$0.7010(+/-0.11)$
PERSI-R	$0.6971(+/-0.14)$
CRINL-R	$0.5783(+/-0.9)$
Majority Voting	$0.6490(+/-0.12)$
CSE-ARS	$0.7432(+/-0.10)$

as shown in **Formula 10**

$$Loss\left(L_t,\ L_t^*\right) = -\frac{1}{n}\sum\left[L_t(i) * logL_t^*(i)\right] + \lambda R(w) \quad (10)$$

where L_t is the correct label of the sample, L_t^* is the network output, and λ is the weight of the regularization term.

The loss value for each recognizer was 0.304, 0.250, 0.378, 0.401, 0.290 and the multimodal fusion model was only 0.179. We used 10 times 10-fold cross-validation on the individual recognizers and multimodal fusion model and the results are shown in **Figure 8**.

Each set represents a 10-fold cross-validation. All individual recognizers are represented with gray-scaled colors in each set. In the ten times of validations, the average prediction accuracy of the models of CSE-ARS multimodal fusion, and particular recognizers is shown in **Table 10**.

To evaluate the performance of CSE-ARS in predicting CSE attacks, we computed ROC curves for each class and determined the AUC values for each model. **Figure 9** shows the ROC curves of the models used to evaluate the prediction performance of CSE attacks, and **Table 11** displays the calculated AUC values for each model.

VIII. DISCUSSION

Deep learning is a powerful approach for building recognizers as it allows for the learning of high-level abstractions from data. This is particularly useful in the case of CSE attacks, where the patterns of behavior and language used by attackers can be complex and difficult to identify. Deep learning models such as recurrent neural networks and transformer-based networks can learn these complex patterns and make accurate predictions about the likelihood of a CSE attack. Furthermore, the use of pre-trained deep learning models is beneficial as these models are trained on large amounts of data, which enables them to generalize well to new data and new types of CSE attacks. This is important as CSE attacks are constantly evolving, and being able to adapt to new types of attacks is crucial for the effectiveness of a CSE attack recognition system. For the implementation of the individual deep learning recognizers, we mainly used the popular deep learning frameworks PyTorch [92] and HuggingFace [93]. CSE-ARS outperformed the individual recognizers in terms of accuracy, precision, recall, F1 score, and AUC value. PERSI-R had the highest performance among the individual models, but the multimodal fusion model achieved a higher accuracy in 10-fold cross-validation. However, there was a large gap between the AUC value of the multimodal fusion model and PERSI-R. It is suggested that the superior performance of the CSE-ARS model is due to the strength of the late fusion approach to utilize multiple modalities, which may capture more comprehensive information about CSE attacks. Overall, the results suggest that the multimodal fusion approach has the potential for improving the accuracy of CSE attack prediction. However, further research is needed to validate the findings on larger and more diverse datasets and to explore the generalizability and robustness of the approach. Additionally, it may be beneficial to investigate the interpretability of the model and the specific features that contribute to the prediction. Many trade-offs had to be considered during model selection such as computing requirements, and performance. As already mentioned, the design decision was to prefer models that had fewer false positives neglecting the false negatives metric. Furthermore, it was also a design decision to choose the models based solely on their performance and not on computing requirements. Neural networks demand more powerful machines (e.g., GPU over CPU) to deliver high accuracy with an acceptable inference latency. It was crucial to maintain a comprehensive record of definitions required to replicate an experiment alongside its pertinent artifacts, which refer to the files created during an experiment such as those displaying loss curves, evaluation loss graphs, logs, or intermediate results of a model throughout a training process. This practice facilitates the comparison of distinct experiments and aids in the selection of the optimal experiment tailored to one's specific requirements. The proposed CSE-ARS has shown promising results in recognizing the various enablers such as personality traits, dialogue acts, persuasion attempts, persistent behavior, and critical information leakage. The system's performance has been evaluated using CSE-ARS Corpus comprised of real-world and fictional chat conversations and the results indicate that it can accurately recognize these enablers with high precision and recall. Another contribution of this study is the use of the late fusion method to combine the separate outputs of the individual recognizer. This approach allows for the strengths of each technique to be leveraged and results in a more robust and accurate recognition system. The individual recognizers used in the proposed system also deserve further discussion. The use of a convolutional network for recognizing persuasion attempts has been shown to be effective in identifying patterns in the language used

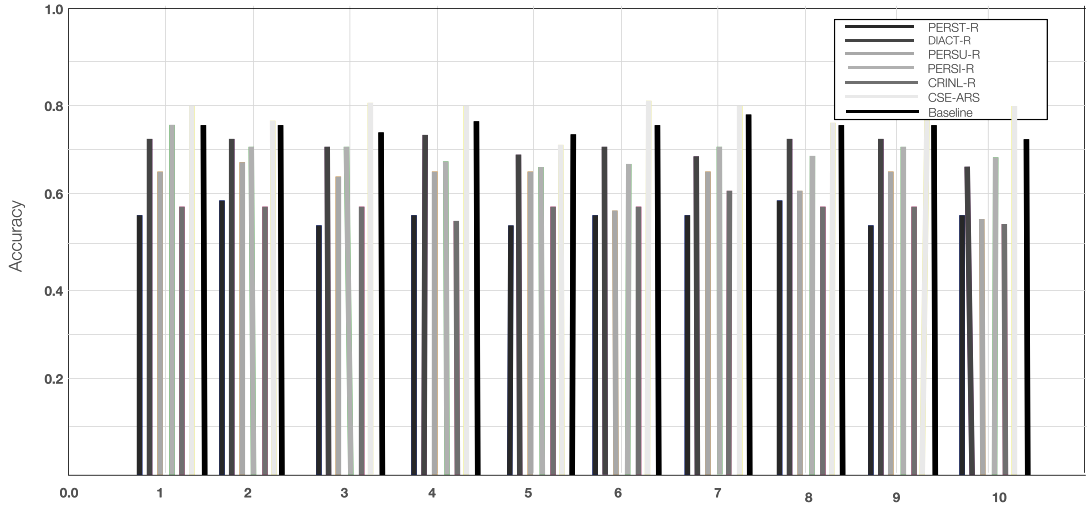

FIGURE 8. Sequence of 10-fold cross-validation.

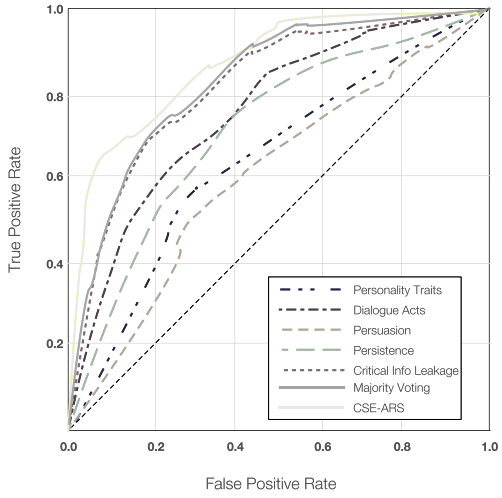

FIGURE 9. ROC curves of CSE-ARS and individual recognizers.

in persuasion attempts. The use of BERT for recognizing personality traits, dialogue acts, and persistent behavior is also well-suited for these tasks as BERT is a pre-trained language model that has been shown to have strong performance in natural language understanding tasks. It is worth noting that BERT models used in this study were fine-tuned on specific and appropriate corpora, and their performance may be improved by further fine-tuning on larger and more diverse corpora. Moreover, the use of such models allows for the proposed system to be easily updated and improved as new data and techniques become available, making it more adaptive to the constantly evolving social engineering attacks. Overall, the individual recognizers used in CSE-ARS have been shown to be effective in recognizing the various enablers of social engineering attacks.

However, SE attacks nowadays utilize a diverse set of communication channels such as audio, video etc. CSE-ARS could significantly benefit from incorporating image, video, and audio inputs, enabling a more comprehensive analysis of social engineering attacks. The integration of diverse modalities would enhance the system's ability to detect nuanced deceptive cues, such as visual anomalies, audio manipulation, or behavioral patterns, thereby improving its overall accuracy and resilience against multifaceted social engineering strategies. Sophisticated image, video and audio processing techniques may enhance CSE-ARS's ability to analyze and interpret visual and audio content in diverse contexts, potentially improving its overall performance. For example advancements in nighttime image enhancement presented in [94], nighttime haze removal method [95] leveraging the innovative gray haze-line or NightHazeFormer's transformer-based framework [96] that demonstrates superior nighttime haze removal by addressing multiple adverse effects. Such techniques may contribute to improved visual content analysis in challenging nighttime conditions, aligning with CSE-ARS's future goal of recognizing social engineering attacks across diverse contexts.

IX. CONCLUSION

In this work, we introduced CSE-ARS, demonstrating its effectiveness in detecting and mitigating CSE attacks. CSE-ARS adopts an interdisciplinary approach, considering factors such as personality traits, linguistic aspects, behavioral characteristics, and information technology attributes. Through multimodal late fusion, it integrates predictions from various deep learning models, leveraging their strengths.

We optimized performance by aggregating recognizer outputs through a weighted linear combination, determined

using simulated annealing and k-fold cross-validation. This ensured peak performance while maintaining a balanced integration of individual recognizers.

Comprehensive evaluation substantiates CSE-ARS's effectiveness in recognizing CSE attacks, making it a valuable tool for safeguarding individuals and organizations. Future research will expand capabilities by incorporating additional factors influencing CSE attacks. The emergence of Foundational Models [97] opens opportunities for identifying deepfake content, though challenges exist.

REFERENCES

[1] I. Del Pozo, M. Iturralde, and F. Restrepo, "Social engineering: Application of psychology to information security," in *Proc. 6th Int. Conf. Future Internet Things Cloud Workshops (FiCloudW)*, Aug. 2018, pp. 108–114.

[2] R. W. Gehl and S. T. Lawson, *Social Engineering: How Crowdmasters, Phreaks, Hackers, and Trolls Created a New Form of Manipulativ e Communication*. Cambridge, MA, USA: MIT Press, 2022.

[3] R. Heartfield and G. Loukas, "A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks," *ACM Comput. Surv.*, vol. 48, no. 3, pp. 1–39, Feb. 2016.

[4] F. Salahdine and N. Kaabouch, "Social engineering attacks: A survey," *Future Internet*, vol. 11, no. 4, p. 89, Apr. 2019.

[5] Z. Wang, H. Zhu, and L. Sun, "Social engineering in cybersecurity: Effect mechanisms, human vulnerabilities and attack methods," *IEEE Access*, vol. 9, pp. 11895–11910, 2021.

[6] *DBIR Report 2023—Master's Guide*, Business Verizon, Ashburn, VA, USA, 2023.

[7] K. Chetioui, B. Bah, A. O. Alami, and A. Bahnasse, "Overview of social engineering attacks on social networks," *Proc. Comput. Sci.*, vol. 198, pp. 656–661, Jan. 2022.

[8] V. Y. Sokolov and O. Y. Korzhenko, "Analysis of recent attacks based on social engineering techniques," 2019, *arXiv:1902.07965*.

[9] C. Catal, G. Giray, B. Tekinerdogan, S. Kumar, and S. Shukla, "Applications of deep learning for phishing detection: A systematic literature review," *Knowl. Inf. Syst.*, vol. 64, no. 6, pp. 1457–1500, Jun. 2022.

[10] W. Syafitri, Z. Shukur, U. A. Mokhtar, R. Sulaiman, and M. A. Ibrahim, "Social engineering attacks prevention: A systematic literature review," *IEEE Access*, vol. 10, pp. 39325–39343, 2022.

[11] S. Venkatesha, K. R. Reddy, and B. R. Chandavarkar, "Social engineering attacks during the COVID-19 pandemic," *Social Netw. Comput. Sci.*, vol. 2, no. 2, pp. 1–9, Apr. 2021.

[12] K. Mitnick, *Ghost in the Wires: My Adventures as the World's Most Wanted Hacker*. Hachette, 2011.

[13] S. M. Albladi and G. R. S. Weir, "Predicting individuals' vulnerability to social engineering in social networks," *Cybersecurity*, vol. 3, no. 1, pp. 1–19, Dec. 2020.

[14] T. Li, K. Wang, and J. Horkoff, "Towards effective assessment for social engineering attacks," in *Proc. IEEE 27th Int. Requirements Eng. Conf. (RE)*, Sep. 2019, pp. 392–397.

[15] M. Mattera and M. M. Chowdhury, "Social engineering: The looming threat," in *Proc. IEEE Int. Conf. Electro Inf. Technol. (EIT)*, May 2021, pp. 056–061.

[16] A. Vishwanath, *The Weakest Link: How to Diagnose, Detect, and Defend Users From Phishing*. Cambridge, MA, USA: MIT Press, 2022.

[17] N. Tsinganos, G. Sakellariou, P. Fouliras, and I. Mavridis, "Towards an automated recognition system for chat-based social engineering attacks in enterprise environments," in *Proc. 13th Int. Conf. Availability, Rel. Secur.*, Aug. 2018, pp. 1–10.

[18] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton Ferrer, "The DeepFake detection challenge (DFDC) dataset," 2020, *arXiv:2006.07397*.

[19] S. Lyu, "Deepfake detection: Current challenges and next steps," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2020, pp. 1–6.

[20] M. Westerlund, "The emergence of deepfake technology: A review," *Technol. Innov. Manag. Rev.*, vol. 9, no. 11, pp. 39–52, Jan. 2019.

[21] B. C. Robert, *Influence, New and Expanded: The Psychology of Persuasion*. HarperCollins Publishers, 2021.

[22] R. B. Cialdini and N. J. Goldstein, "The science and practice of persuasion," *Cornell Hotel Restaurant Admin. Quart.*, vol. 43, no. 2, pp. 40–50, 2002.

[23] R. B. Cialdini, *Influence: Science and Practice*, vol. 4, Boston, MA, USA: Pearson, 2009.

[24] G. W. Allport, "Personality: A psychological interpretation," Tech. Rep., 1937.

[25] L. R. Goldberg, "The structure of phenotypic personality traits," *Amer. Psychologist*, vol. 48, no. 1, pp. 26–34, 1993.

[26] J. F. Salgado, "The big five personality dimensions and counterproductive behaviors," *Int. J. Selection Assessment*, vol. 10, nos. 1–2, pp. 117–125, Mar. 2002.

[27] N. Tsinganos, "Utilizing deep learning and natural language processing to recognise chat-based social engineering attacks for cyber security situational awareness," Ph.D. thesis, School Inf. Sci., Dept. Appl. Inform., Univ. Macedonia, Thessaloniki, Greece, Oct. 2023.

[28] N. Tsinganos and I. Mavridis, "Building and evaluating an annotated corpus for automated recognition of chat-based social engineering attacks," *Appl. Sci.*, vol. 11, no. 22, p. 10871, Nov. 2021.

[29] N. Tsinganos, P. Fouliras, and I. Mavridis, "Leveraging dialogue state tracking for zero-shot chat-based social engineering attack recognition," *Appl. Sci.*, vol. 13, no. 8, p. 5110, Apr. 2023.

[30] N. Tsinganos, I. Mavridis, and D. Gritzalis, "Utilizing convolutional neural networks and word embeddings for early-stage recognition of persuasion in chat-based social engineering attacks," *IEEE Access*, vol. 10, pp. 108517–108529, 2022.

[31] N. Tsinganos, P. Fouliras, and I. Mavridis, "Applying BERT for early-stage recognition of persistence in chat-based social engineering attacks," *Appl. Sci.*, vol. 12, no. 23, p. 12353, Dec. 2022.

[32] Z. Wang, Y. Ren, H. Zhu, and L. Sun, "Threat detection for general social engineering attack using machine learning techniques," 2022, *arXiv:2203.07933*.

[33] J. M. Hatfield, "Social engineering in cybersecurity: The evolution of a concept," *Comput. Secur.*, vol. 73, pp. 102–113, Mar. 2018.

[34] J. H. Bullée, L. Montoya, W. Pieters, M. Junger, and P. Hartel, "On the anatomy of social engineering attacks—A literature-based dissection of successful attacks," *J. Investigative Psychol. Offender Profiling*, vol. 15, no. 1, pp. 20–45, Jan. 2018.

[35] K. Krombholz, H. Hobel, M. Huber, and E. Weippl, "Advanced social engineering attacks," *J. Inf. Secur. Appl.*, vol. 22, pp. 113–122, Jun. 2015.

[36] Y. P. Atmojo, I. M. D. Susila, M. R. Hilmi, E. S. Rini, L. Yuningsih, and D. P. Hostiadi, "A new approach for spear phishing detection," in *Proc. 3rd East Indonesia Conf. Comput. Inf. Technol. (EIConCIT)*, Apr. 2021, pp. 49–54.

[37] V. Shakela and H. Jazri, "Assessment of spear phishing user experience and awareness: An evaluation framework model of spear phishing exposure level (SPEL) in the Namibian financial industry," in *Proc. Int. Conf. Adv. Big Data, Comput. Data Commun. Syst. (icABCD)*, Aug. 2019, pp. 1–5.

[38] J. S. Wiggins, *The Five-Factor Model of Personality: Theoretical Perspectives*. Guilford Press, 1996.

[39] F. Sudzina and A. Pavlicek, "Propensity to click on suspicious links: Impact of gender, of age, and of personality traits," Tech. Rep., 2017.

[40] S. M. Albladi and G. R. S. Weir, "Personality traits and cyber-attack victimisation: Multiple mediation analysis," in *Proc. Internet Things Bus. Models, Users, Netw.*, Nov. 2017, pp. 1–6.

[41] S. Anawar, D. L. Kunasegaran, M. Z. Mas'ud, and N. A. Zakaria, "Analysis of phishing susceptibility in a workplace: A big-five personality perspectives," *J. Eng. Sci. Technol.*, vol. 14, no. 5, pp. 2865–2882, 2019.

[42] S. Uebelacker and S. Quiel, "The social engineering personality framework," in *Proc. Workshop Socio-Technical Aspects Secur. Trust*, Jul. 2014, pp. 24–30.

[43] M. McBride, L. Carter, and M. Warkentin, "Exploring the role of individual employee characteristics and personality on employee compliance with cybersecurity policies," *RTI Int.-Inst. Homeland Secur. Solutions*, vol. 5, no. 1, pp. 45–83, 2012.

[44] M. Hoeschele and M. Rogers, "Detecting social engineering," in *Advances in Digital Forensics*. Berlin, Germany: Springer, 2005, pp. 67–77.

[45] M. Hoeschele, "CERIAS tech report 2006–15 detecting social engineering," Tech. Rep., 2006.

[46] M. Bezuidenhout, F. Mouton, and H. S. Venter, "Social engineering attack detection model: SEADM," in *Proc. Inf. Secur. South Afr.*, Aug. 2010, pp. 1–8.

[47] F. Mouton, L. Leenen, and H. S. Venter, "Social engineering attack detection model: SEADMv2," in *Proc. Int. Conf. Cyberworlds (CW)*, Oct. 2015, pp. 216–223.

[48] R. Bhakta and I. G. Harris, "Semantic analysis of dialogs to detect social engineering attacks," in *Proc. IEEE 9th Int. Conf. Semantic Comput. (IEEE ICSC)*, Feb. 2015, pp. 424–427.

[49] Y. Sawa, R. Bhakta, I. G. Harris, and C. Hadnagy, "Detection of social engineering attacks through natural language processing of conversations," in *Proc. IEEE 10th Int. Conf. Semantic Comput. (ICSC)*, Feb. 2016, pp. 262–265.

[50] T. Peng, I. Harris, and Y. Sawa, "Detecting phishing attacks using natural language processing and machine learning," in *Proc. IEEE 12th Int. Conf. Semantic Comput. (ICSC)*, Jan. 2018, pp. 300–301.

[51] M. Lansley, N. Polatidis, and S. Kapetanakis, "SEADer: A social engineering attack detection method based on natural language processing and artificial neural networks," in *Computational Collective Intelligence*. Hendaye, France: Springer, 2019, pp. 686–696.

[52] M. Lansley, S. Kapetanakis, and N. Polatidis, "SEADer++v2: Detecting social engineering attacks using natural language processing and machine learning," in *Proc. Int. Conf. Innov. Intell. Syst. Appl. (INISTA)*, Aug. 2020, pp. 1–6.

[53] M. Lansley, F. Mouton, S. Kapetanakis, and N. Polatidis, "SEADer++: Social engineering attack detection in online environments using machine learning," *J. Inf. Telecommun.*, vol. 4, no. 3, pp. 346–362, Jul. 2020.

[54] M. Lansley, N. Polatidis, S. Kapetanakis, K. Amin, G. Samakovitis, and M. Petridis, "Seen the villains: Detecting social engineering attacks using case-based reasoning and deep learning," Tech. Rep., 2019.

[55] X. Wang, W. Shi, R. Kim, Y. Oh, S. Yang, J. Zhang, and Z. Yu, "Persuasion for good: Towards a personalized persuasive dialogue system for social good," 2019, *arXiv:1906.06725*.

[56] D. Yang, J. Chen, Z. Yang, D. Jurafsky, and E. Hovy, "Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms," in *Proc. Conf. North*, 2019, pp. 3620–3630.

[57] Y. Lee, J. Saxe, and R. Harang, "CATBERT: Context-aware tiny BERT for detecting social engineering emails," 2020, *arXiv:2010.03484*.

[58] A. Dalton, E. Aghaei, E. Al-Shaer, A. Bhatia, E. Castillo, Z. Cheng, S. Dhaduvai, Q. Duan, B. Hebenstreit, and M. M. Islam, "Active defense against social engineering: The case for human language technology," in *Proc. 1st Int. Workshop Social Threats Online Conversations, Understand. Manag.*, 2020, pp. 1–8.

[59] F. O. Catak, K. Sahinbas, and V. Dörtkardeş, "Malicious URL detection using machine learning," in *Artificial Intelligence Paradigms for Smart Cyber-Physical Systems*. IGI global, 2021, pp. 160–180.

[60] Y. Lan, "Chat-oriented social engineering attack detection using attention-based bi-LSTM and CNN," in *Proc. 2nd Int. Conf. Comput. Data Sci. (CDS)*, Jan. 2021, pp. 483–487.

[61] H. Shi, M. Silva, L. Giovanini, D. Capecci, L. Czech, J. Fernandes, and D. Oliveira, "Lumen: A machine learning framework to expose influence cues in texts," *Frontiers Comput. Sci.*, vol. 4, Aug. 2022, Art. no. 929515.

[62] ISACA. (2023). *In Pursuit if Digital Trust | ISACA*. Accessed: Sep. 27, 2023. [Online]. Available: https://www.isaca.org/

[63] P. Bernard, *COBIT 5-A Management Guide*. Van Haren, 2012.

[64] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.

[65] M. Pawlowski, A. Wróblewska, and S. Sysko-Romanczuk, "Effective techniques for multimodal data fusion: A comparative analysis," *Sensors*, vol. 23, no. 5, p. 2381, Feb. 2023.

[66] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.

[67] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[68] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[69] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[70] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, Jan. 2011.

[71] X. Schmitt, S. Kubler, J. Robert, M. Papadakis, and Y. LeTraon, "A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, gate," in *Proc. 6th Int. Conf. Social Netw. Anal., Manag. Secur. (SNAMS)*, Oct. 2019, pp. 338–343.

[72] C. Spielberger, *Encyclopedia of Applied Psychology*. Cambridge, MA, USA: Academic Press, 2004.

[73] H. Ahmad, M. U. Asghar, M. Z. Asghar, A. Khan, and A. H. Mosavi, "A hybrid deep learning technique for personality trait classification from text," *IEEE Access*, vol. 9, pp. 146214–146232, 2021.

[74] J. E. Arijanto, S. Geraldy, C. Tania, and D. Suhartono, "Personality prediction based on text analytics using bidirectional encoder representations from transformers from English Twitter dataset," *Int. J. FUZZY Log. Intell. Syst.*, vol. 21, no. 3, pp. 310–316, Sep. 2021.

[75] N. H. Jeremy, G. Christian, M. F. Kamal, D. Suhartono, and K. M. Suryaningrum, "Automatic personality prediction using deep learning based on social media profile picture and posts," in *Proc. 4th Int. Seminar Res. Inf. Technol. Intell. Syst. (ISRITI)*, Dec. 2021, pp. 166–172.

[76] M. Pattinson, C. Jerram, K. Parsons, A. McCormac, and M. Butavicius, "Why do some people manage phishing e-mails better than others?" *Inf. Manag. Comput. Secur.*, vol. 20, no. 1, pp. 18–28, 2012.

[77] I. Alseadoon, M. Othman, and T. Chan, "What is the influence of users' characteristics on their ability to detect phishing emails?" in *Proc. 1st Int. Conf. Commun. Comput. Eng.* Cham, Switzerland: Springer, 2015, pp. 949–962.

[78] T. Halevi, J. Lewis, and N. Memon, "A pilot study of cyber security and privacy related behavior and personality traits," in *Proc. 22nd Int. Conf. World Wide Web*, May 2013, pp. 737–744.

[79] A. C. Johnston, M. Warkentin, M. McBride, and L. Carter, "Dispositional and situational factors: Influences on information security policy violations," *Eur. J. Inf. Syst.*, vol. 25, no. 3, pp. 231–251, May 2016.

[80] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[81] H. Jiang, X. Zhang, and J. D. Choi, "Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings (student abstract)," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 13821–13822.

[82] *Hugging Face—The AI Community Building The Future*, May 2023. [Online]. Available: https://huggingface.co/

[83] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[84] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," 2015, *arXiv:1508.05326*.

[85] Nikolaos Tsinganos, *CSE-ARS Corpus*, Univ. Macedonia, Thessaloniki, Greece, 2023.

[86] J. Bridle, "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2, 1989, pp. 1–7.

[87] L. M. R. Rere, M. I. Fanany, and A. M. Arymurthy, "Simulated annealing algorithm for deep learning," *Proc. Comput. Sci.*, vol. 72, pp. 137–144, Jan. 2015.

[88] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[89] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*.

[90] L. Biewald, "Experiment tracking with weights and biases," *Softw. Available From Wandb. Com*, vol. 2, p. 233, Jan. 2020.

[91] A. Jain, A. Kumar, and S. Susan, "Evaluating deep neural network ensembles by majority voting cum meta-learning scheme," in *Soft Computing and Signal Processing*, vol. 2. Berlin, Germany: Springer, 2022, pp. 29–37.

[92] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[93] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, and M. Funtowicz, "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Syst. Demonstrations*, 2020, pp. 38–45.

[94] Y. Liu, Z. Yan, J. Tan, and Y. Li, "Multi-purpose oriented single nighttime image haze removal based on unified variational retinex model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1643–1657, Apr. 2023.

[95] W. Wang, A. Wang, and C. Liu, "Variational single nighttime image haze removal with a gray haze-line prior," *IEEE Trans. Image Process.*, vol. 31, pp. 1349–1363, 2022.

[96] Y. Liu, Z. Yan, S. Chen, T. Ye, W. Ren, and E. Chen, "NightHazeFormer: Single nighttime haze removal using prior query transformer," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 4119–4128.
[97] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, and E. Brunskill, "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.

NIKOLAOS TSINGANOS received the B.Sc. degree in computer science and the M.Sc. degree in pervasive and mobile computing systems from Hellenic Open University, Patras, Greece, and the Ph.D. degree in information systems security from the University of Macedonia. He is currently a member of the InfoSec Research Group, Multimedia, Security, and Networking Laboratory (MSNLab). He has participated in several international and nationally-funded research and development (R&D) projects. His current research interests include the design and performance evaluation of cyber defense mechanisms utilizing machine learning methods, particularly in the context of social engineering attack recognition.

PANAGIOTIS FOULIRAS received the B.Sc. degree in physics from the Aristotle University of Thessaloniki, Greece, and the M.Sc. and Ph.D. degrees in computer science from the University of London, U.K. (QMW). He is currently a permanent Assistant Professor with the University of Macedonia, Thessaloniki, Greece. He has participated in several national and European-funded (H2020) research projects and published articles in many international journals. His research interests include computer networks and network security, blockchain, and system evaluation methods.

IOANNIS MAVRIDIS received the Diploma degree in computer engineering and informatics from the University of Patras, Greece, and the Ph.D. degree in information systems security from the Aristotle University of Thessaloniki, Greece. He is currently a Professor of information security with the Department of Applied Informatics, University of Macedonia (UoM), Greece. He is also the Director of the Multimedia, Security and Networking Laboratory (MSN Lab). His research interests include AI-based attack detection, cybersecurity education, risk management, access control, cyber threat intelligence, digital forensics, and security economics. He serves as an Area Editor for the *Array* journal (Elsevier).

DIMITRIS GRITZALIS received the B.Sc. degree in mathematics from the University of Patras, Greece, the M.Sc. degree in computer science from the City University of New York, New York, NY, USA, and the Ph.D. degree in information systems security from the University of the Aegean, Greece. He has served as an Associate Rector for Research, the President of the Greek Computer Society, and an Associate Data Protection Commissioner of Greece. He is currently a Professor of cybersecurity with the Department of Informatics, Athens University of Economics and Business, Greece, where he serves as the Director of the M.Sc. Programme in information systems security. His research interests include risk assessment, cybersecurity education, malware, and cyber conflicts. He is an Academic Editor of the *Computers and Security* journal.