IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Predicting Data Exfiltration using Supervised Machine Learning based on Tactics Mapping from Threat Reports and Event Logs

ARIF RAHMAN HAKIM[1] (Member, IEEE), KALAMULLAH RAMLI[1] (Member, IEEE), MUHAMMAD SALMAN[1] (Member, IEEE), BERNARDI PRANGGONO[2] (Senior Member, IEEE), ESTI RAHMAWATI AGUSTINA[1] (Member, IEEE).

[1]Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia, Depok 16424, Jawa Barat, Indonesia
[2]School of Computing and Information Science, Anglia Ruskin University, CB1 1PT Cambridge, U.K.

Corresponding author: Kalamullah Ramli (email: kalamullah.ramli@ui.ac.id).

Universitas Indonesia supported this work through the Hibah Publikasi Terindeks Internasional (PUTI) Kolaborasi Internasional (Q2) 2023 Scheme under Contract NKB-820/UN2.RST/HKP.05.00/2023. The work of Arif Rahman Hakim was supported by the Lembaga Pengelola Dana Pendidikan (LPDP), Ministry of Finance of the Republic of Indonesia.

ABSTRACT Data breach attacks are unique, especially when attackers exfiltrate data from their target's systems. Furthermore, as data breaches continue to increase in frequency and severity, they pose a growing risk to society and organizations. Unfortunately, no prior research focused on predicting exfiltration occurrence based on a sequence of tactics identified from low-level logs. In addition, integrating low-level logs with a high-level conceptual framework presents a critical challenge. The need for automation in the mapping process and advanced methods to assist defenders in analyzing the occurrence of exfiltration within their systems is urgent. In this paper, we focus on developing a machine learning (ML) model to predict the occurrence of data exfiltration by analyzing the sequence of tactics employed by an attacker. We propose two main contributions, including addressing the gap level between low-level logs and high-level data breach conceptual steps and integrating collected event logs and ML models to predict exfiltration tactics. Our dataset for the ML model is created based on tactics identified in threat reports, cleaned to obtain ten features, and balanced using the SMOTE+ENN technique. The prediction is made using tactics identified from low-level logs that serve as input to the ML model to determine whether the events lead to the occurrence of exfiltration. We benchmarked three resampling methods, five feature selection techniques, and five ML algorithms to achieve optimal ML model performance. A new dataset, comprehensive techniques used to develop the ML model, and the proposed prediction method represent the key contributions of this study compared to existing research. In addition, to demonstrate the effectiveness of our proposed method, we present case studies using event logs from real-world incidents. The investigation shows that our proposed method effectively predicts the occurrence of exfiltration with higher accuracy than existing studies.

INDEX TERMS data exfiltration, event logs, machine learning, tactics mapping, threats report.

I. INTRODUCTION

DATA breaches have become a critical issue globally due to their significant impact on various aspects of society [1], organizations [2], [3], and individuals [4]. These breaches involve unauthorized access to confidential, protected, or sensitive data, resulting in data loss, including financial [5], health [6], and personal information [7]. According to the Identity Theft Resource Center, 2023 marked a notable year for data breaches, with 3,205 incidents disclosed, affecting approximately 353,027,892 people. This represents a 78% increase compared to 2022 [8]. Furthermore, IBM reported that 46% of the breaches in 2024 involved personally identifiable information (PII) of consumers [9] and according to Verizon, 86% of data breaches involved the use of stolen credentials [10].

Moreover, as data breaches continue to increase in frequency and severity, they pose a growing risk to society and organizations [11]. The ongoing threat of data breaches requires continuous efforts to improve cybersecurity measures and mitigate risks [12]. Mitigating risk in cybersecurity

involves the use of specific technologies and processes to prevent adversaries from achieving their tactical goals [13], [14].

In particular, the exfiltration phase is a key reference point for detecting or investigating breaches, providing indicators and patterns that can be monitored and analyzed to justify the existence of data breaches [15]. However, cyber-attacks are becoming more sophisticated, making detection of data exfiltration challenging for organizations [16], [17]. Signature-based detection requires human assistance to create and update signature rules, which can be time-consuming and require adaptation to new threats or signature patterns [18]. To address this issue, we need to develop flexible and efficient mechanisms to respond intelligently and update policies. These mechanisms require analyzing large amounts of cybersecurity data from various sources, identifying insights, and creating automated tools [19]. The data from various sources, including external threat reports and diverse logs from the system's internal environment, are analyzed, and insights are identified through automated processes supported by machine learning techniques.

Threat reports are crucial in strengthening an organization's security posture by providing actionable insights into emerging threats, adversary tactics, and vulnerabilities [20]. These reports can substantially improve an organization's capacity to identify, avert, and address cyber threats [21]. On the other hand, leveraging low-level logs to detect ongoing attacks and investigate incidents is a crucial aspect of modern cybersecurity practices. These logs can provide detailed information on system activities, helping identify anomalies and potential security breaches. Integrating threat reports and low-level logs for incident response and threat hunting offers significant advantages by enhancing incident detection, analysis, and mitigation.

However, integrating threat reports with low-level logs presents several challenges, such as the complexity of the data, the need for accurate parsing, and an automated process. In addition, data breach investigation frameworks are typically high-level concepts that serve as a guide to conduct investigations step by step [22], while the logs collected for analysis are low-level information. Therefore, the investigation process has a level gap between low-level logs and high-level conceptual frameworks.

In this paper, we propose a novel method that integrates threat reports and event logs to assist defenders or investigators in analyzing the occurrence of exfiltration within their systems. To address the challenge of the level gap in this integration, we use the Adversarial Tactics and Techniques and Common Knowledge (ATT&CK) from MITRE [23] as an intermediate level. Furthermore, the main contribution of this study is the automated integration process, where data exfiltration analysis is no longer based on manual examination but is instead enhanced by machine learning-based predictive models.

In developing this machine learning (ML) model, we utilize tactics data extracted from threat reports as features to predict exfiltration as the output variable. Furthermore, event logs collected from the system are mapped to various tactics according to the event IDs recorded in the logs. The sequence of tactics resulting from this mapping serves as the input for the prediction model, which processes the query and produces a prediction of whether exfiltration has occurred in the system. By using our method, organizations can enhance decision-making in detecting or investigating data exfiltration in a more time-efficient manner.

This paper is structured into five sections. Section II discusses previous studies, particularly those relevant to developing advanced methods utilizing the MITRE ATT&CK labeling from different data sources. We also provide a comparative analysis of these advanced methods. Section III explains our comprehensive process in developing our ML-based prediction model, including the approaches we employed to address the imbalance within the dataset, features selection, and evaluation process. Section IV introduces our proposed method, detailing each step, including usage scenarios and case studies. Finally, we offer our conclusions in Section V.

II. RELATED WORK

In this section, we present a summary of selected studies concerning the extraction and mapping of MITRE ATT&CK tactics from diverse information sources. This review highlights the potential to leverage tactics extraction from existing studies to create a new dataset aimed at predicting the occurrence of exfiltration. A detailed discussion of our new dataset creation process is presented in Section III-A. Furthermore, this section also includes a review of selected methods used for predicting adversarial tactics and attack goals in cyber-attacks. These methods were selected based on their alignment with the objectives of this study, which predict exfiltration as an attacker's objective based on the sequence of tactics identified within the system. Additionally, we provide a comparative analysis of the objectives, approaches, and algorithms employed by these methods.

A. MITRE ATT&CK MAPPINGS

In this subsection, we summarize existing models that leverage the mapping of Tactics or Techniques from the MITRE ATT&CK for purposes such as detecting, investigating, or predicting cyber-attacks. Our review focuses on the developed models, the algorithms employed, the datasets used, performance based on the metrics applied, and other key findings.

The paper [24] investigates the implementation of transformer-based models for the automatic mapping of Common Vulnerabilities and Exposures (CVEs) to MITRE ATT&CK tactics, with the objective of enhancing cybersecurity comprehension and defense strategies. The research employs a comprehensive dataset comprising 9985 entries and incorporates security auditing tools to establish connections between CVEs and MITRE ATT&CK tactics. The models that demonstrated superior performance include SecRoBERTa, SecBERT, CyBERT, and TARS, with

SecRoBERTa attaining the highest weighted F1 score of 78.88%. A primary limitation of the study is the inherent difficulty in interpreting certain vulnerabilities due to the conceptual associations between tactics.

The study in [25] automates the labeling of malware threat reports using feature selection techniques and word embeddings to improve the prediction of MITRE ATT&CK tactics. The method uses Mutual Information (MI) and Chi-squared statistics (CHI) for feature selection, improving previous studies. The F.5-score metric showed a 6% increase in performance when utilizing feature selection techniques to predict MITRE ATT&CK tactics. However, the paper acknowledges limitations, such as the lack of metadata connecting sentences to specific tactics and techniques, making it difficult to divide texts into blocks for analysis.

A system that uses a deep learning model and ontology knowledge to extract MITRE techniques from unstructured CTI reports called MITREtrieval is proposed in [26]. The system addresses sparse data and implicit TTPs by leveraging deep learning and ontology. The system demonstrated F2 scores of 58%, 62%, and 69% in multi-label technique identification across 113, 46, and 23 CTI reports, respectively. These results not only exceeded established benchmarks but also enhanced the efficiency of CTI analysis. This innovative approach overcomes the limitations of machine learning-based methods, which often overlook crucial words, impacting technique identification.

In their 2024 study, Gabrys et al. [27] introduced a novel methodology that employs Large Language Models (LLMs) to establish a connection between Intrusion Detection System (IDS) rules and attacker tactics, techniques, and procedures (TTPs) as outlined in the MITRE ATT&CK framework. This research utilizes a dataset containing 972 labeled IDS rules to generate descriptive narratives and predict TTPs. The authors leverage ChatGPT to generate structured textual summaries of these IDS rules and develop a pipeline that transforms existing IDS rulesets into efficient data representations using BERT. Additionally, they train an algorithm to effectively classify observed attacker behaviors into MITRE ATT&CK Enterprise technique IDs. The study evaluates various classification methods, identifying that the Support Vector Machine exhibits superior performance, achieving accuracy rates of approximately 99% for T10 and 90% for T5. Nevertheless, the expansion of this approach is limited by the scarcity of suitably labeled datasets.

The authors in [28] use NLP techniques to map Linux commands to MITRE ATT&CK techniques and sub-techniques. The study uses cosine similarity scoring to extract the top-n ATT&CK Techniques and Sub-Techniques for each command, evaluating performance using recall at n metrics. The aim is to enhance the mapping between Linux commands and MITRE ATT&CK for cybersecurity purposes. The study focuses on one-to-many mapping based on text similarity performance, measured using recall at n metrics. The algorithm used is Cosine similarity scoring, which measures the similarity between Linux commands and MITRE ATT&CK descriptions, aiding in the mapping process. This innovative approach in cybersecurity research aims to improve the mapping between Linux commands and MITRE ATT&CK for cybersecurity purposes.

The paper [29] presents a dataset of 1813 Computer Vision Errors (CVEs) annotated with MITRE ATT&CK techniques and proposes models to link CVEs to techniques using BERT-based language models. The study addresses the imbalance in the training set with data augmentation strategies based on TextAttack, achieving an F1-score of 47.84%. The methodology involves building a labeled corpus of CVEs by manually mapping each CVE to tactics and techniques from the MITRE ATT&CK Enterprise Matrix. The performance of the models is measured using the F1-score metric.

From the review of existing studies, it was found that their methods primarily focus on extracting MITRE ATT&CK tactics or techniques from various sources such as CVE, threat reports, CTI, and IDS rules. However, none of the studies have processed the extracted data from these sources to examine the characteristics of the sequence of tactics leading to a specific attacker goal. Therefore, we processed the extracted data from threat reports into a sequence of tactics that serve as predictors for determining whether an attack leads to exfiltration as the attack's objective. The detailed process of constructing our dataset is explained in Section III-A.

B. SELECTED EXISTING PREDICTION METHODS

In line with the objective of this study, which is to develop a predictive method for identifying the occurrence of exfiltration, we reviewed selected articles that similarly proposed predictive methods for attack tactics, attack techniques, and attack goals. We highlight the objectives, primary methods, key findings, and limitations of each study.

The paper [30] proposes a method to extract ATT&CK techniques from Sysmon logs, improve incident response efficiency, and create a prediction system for lateral movements. It proposes two proposals: one to automatically extract ATT&CK techniques and another to develop an efficient prediction system. The first involves creating a database to record the relationship between ATT&CK techniques and attack commands, while the second uses Quantification Theory Type 3 and device activity logs to visualize attackers' movements and detect lateral movements. However, the method has limitations due to limited ATT&CK techniques from Atomic Red Team log data.

The study in [31] uses the Bayesian ATT&CK Network model to predict attack techniques and attacker goals, using previously detected attacks as evidence. It suggests defense techniques to counter these goals and enhances APT prediction and defense strategies through a data-driven approach. The BAN model effectively predicts advanced persistent threat (APT) attacks and defenses, with a high F1-score. The study highlights four limitations in existing studies on predicting APT attacks, including insufficient use of data-driven methodologies, atypical attack modeling, high system reliance, and unavailability of practical datasets.

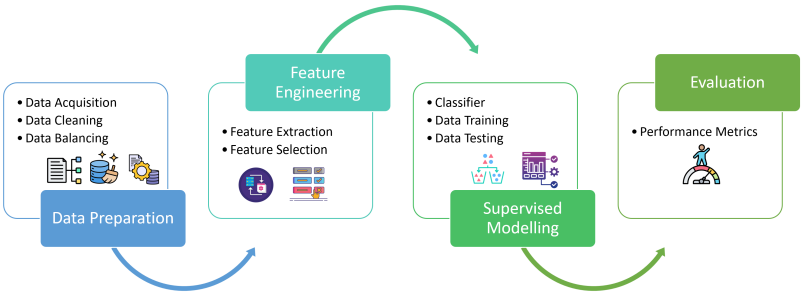

FIGURE 1. Supervised ML process model.

The research conducted by Liu et al. [32] seeks to develop a composite learning approach, referred to as CL-AP 2, to predict cyberattack techniques by analyzing real-time system logs and employing MITRE ATT&CK tactics. This approach generates a Temporal Attack Knowledge Graph (TAKG) and utilizes the Soft Actor-Critic algorithm for predictive modeling. The study undertakes a comparative analysis of CL-AP 2 against established attack prediction methodologies, utilizing anonymized Windows audit logs. The findings emphasize the importance of incorporating advanced NLP models and pertinent threat information to enhance the accuracy of attack predictions. However, the study also notes that existing models may exhibit limitations in generalizing to diverse scenarios due to the specificity of attack patterns and the unstructured nature of real-time logs.

A cooperative system for predicting DDoS attacks using early warning signals from preparation phases called COOPRED DDoS is developed by Neira et al. [33]. The system collects network traffic attributes and uses machine learning techniques like Ada-boost and Multilayer Perceptron to improve prediction accuracy. Four experiments were conducted using two widely used datasets. The paper highlights the limitations of internet data for DDoS attack predictions, the need for knowledge of botnet communication methods, and the challenge of identifying attack preparation signals due to attackers' concealment.

Almomani et al. in [34] evaluates Denial-of-Service (DoS) attack detection in the IoT environment using ML classifiers with the UNSW-NB15 dataset. The research aims to identify the most effective classifier, improve detection accuracy, precision, and sensitivity, and address IDS limitations. The Random Forest classifier outperformed other classifiers, achieving 99.48% accuracy, precision, sensitivity, and F-measure. However, the study may overlook other types of DoS attacks and suggests future research should use modern datasets and advanced techniques.

The novelty of our study resides in the prediction of exfiltration based on a sequence of tactics identified from event logs. This prediction is facilitated by the selection of the most effective machine learning model, determined through a comprehensive comparison of three resampling methods, five feature selection techniques, and five machine learning algorithms. Furthermore, we processed threat report datasets to extract relevant tactics and techniques, thereby creating a new dataset that incorporates ten tactics as features, with exfiltration designated as the target variable. To distinguish our proposed method from existing approaches, we conducted an extensive comparison encompassing objectives, methodologies, algorithms, metrics, and the performance of each method.

III. MACHINE LEARNING MODEL DEVELOPMENT

This section offers a comprehensive description of our ML model development. In our proposed method, we leverage the significant potential of supervised ML to predict attack tactics, especially Exfiltration. Fig.1 illustrates the common process model of supervised ML. Given the ML model's capacity to analyze extensive datasets and recognize patterns, we employ a supervised ML model to provide predictions based on the tactic's patterns identified from event logs, using a dataset of tactics extracted from threat reports. Furthermore, we provide a detailed explanation of creating, addressing imbalanced, and optimizing the feature space of our dataset. Moreover, we offer a comparative analysis of five ML algorithms' performance for our model. The ML model with the best performance will then be employed to predict whether exfiltration has occurred based on the event log data.

A. CREATING DATASET BASED ON IDENTIFIED TACTICS IN THREAT REPORTS

During the data acquisition process, we obtained a dataset from [25], [35] containing descriptions of threat reports from various sources, along with labeled tactics for each row of threat report descriptions. The data cleaning process involved removing rows with incomplete tactic labels. In this study, we excluded two tactics, Reconnaissance, and Resources, because these tactics are primarily conducted on the attackers' side. We utilized the first tactic, Initial Access, as it describes how attackers gain access to the target environment.

Therefore, in this study, we used ten tactics as features: (1) Initial Access, (2) Execution, (3) Persistence, (4) Privilege Escalation, (5) Defense Evasion, (6) Credential Access, (7) Discovery, (8) Lateral Movement, (9) Collection, and (10) Command and Control. Exfiltration was used as the target

TABLE 1. Tactics Occurrence in Cleaned Dataset

Feature	Tactics	#Occurrence
Feature_1	Initial Access	158
Feature_2	Execution	396
Feature_3	Persistence	480
Feature_4	Privilege Escalation	321
Feature_5	Defense Evasion	635
Feature_6	Credential Access	234
Feature_7	Discovery	267
Feature_8	Lateral Movement	269
Feature_9	Collection	165
Feature_10	Command and Control	304
Target	Exfiltration	86

TABLE 2. Comparison of Datasets Size and Ratio

Resampling	#Rows	Ratio (%)
Initial (without resampling)	1284	93:37
Undersampling	301	71:29
Oversampling	2036	59:41
Hybrid sampling (SMOTE+ENN)	1139	85:15

variable. The fundamental idea behind using this dataset is to observe the patterns that emerge from the ten tactics and determine whether the occurrence of these ten tactics leads to exfiltration.

As a result of the data cleaning process, we obtained a dataset consisting of 1284 rows and 11 columns, with 10 columns representing features and 1 column representing the target variable, along with the composition of each tactic, as presented in Table 1. The table shows that the Exfiltration tactic has 86 labels with a value of "1," while 1198 labels have a value of "0." This clearly indicates that the cleaned dataset is imbalanced. Therefore, in Section III-B, we applied resampling methods to address this imbalance.

Next, we performed feature selection in Section III-C to identify the most influential features for prediction outcomes. We proceeded with feature selection based on these calculations using filter and wrapper methods. We applied various techniques for the filter method, including Mutual Information (MI), Pearson Correlation, Feature Importance, and Chi-square. On the other hand, we utilized Recursive Feature Elimination (RFE) for the wrapper method.

B. ADDRESSING IMBALANCED DATASET USING RESAMPLING APPROACHES

The cleaned dataset from the previous stage is clearly imbalanced, with the majority class outnumbering the minority class by a ratio of 93%:7%. The tactics leading to exfiltration are scarce, which challenges the effectiveness of the ML model in detecting and classifying exfiltration. To address this imbalance, we applied three techniques including random undersampling, the Synthetic Minority Over-sampling Technique (SMOTE), and SMOTE with Edited Nearest Neighbors (SMOTE+ENN). We then compared the performance of these methods using a Random Forest classifier, and the best-performing resampling technique was selected for further model development.

Table 2 compares the initial dataset characteristics without resampling and the three datasets generated from each resampling method. The table displays the number of rows for each dataset and the ratio of majority to minority class in percentages. Notably, the oversampled dataset has the largest number of rows and the most balanced ratio. However, this does not guarantee optimal model performance. Therefore, we evaluated the model's performance across the resampled datasets to determine which method yielded the best results. Additionally, the number of identified tactics in each column across all datasets is presented in Table 3.

We employed random undersampling to address the class imbalance in our dataset. Random undersampling is a simple technique that decreases the quantity of samples in the majority class [36]. In terms of our dataset, this involves reducing the number of samples with a value of 0 in the target column, Exfiltration. This process resulted in a new dataset of 301 rows, with a majority-to-minority class ratio of 71% to 29%.

Moreover, we employed the Synthetic Minority Oversampling Technique (SMOTE), a prevalent strategy for addressing imbalanced datasets by creating synthetic samples for the minority class [37]. The oversampling process resulted in a new dataset with 2,036 rows, achieving a majority-to-minority class ratio of 59% to 41%.

Finally, the third resampling method we applied was a hybrid approach that combines SMOTE with Edited Nearest Neighbors (ENN). The SMOTE used in this process is the same as in the oversampling stage but is combined with ENN. This SMOTE+ENN combination effectively addresses dataset imbalance by generating synthetic samples for the minority class and then removing noisy samples. This approach not only improves accuracy but also reduces complexity [38], leading to optimal model performance [39]. As a result, the process yielded a new dataset with 1,139 rows, with a majority-to-minority class ratio of 85% to 15%.

Fig.2 presents the performance measurements of the undersampling, over-sampling, and hybrid-sampling methods using the Random Forest model. The comparison of model performance across the three resampling methods found that hybrid sampling achieved the best performance across all metrics.

TABLE 3. Comparison of identified tactics across feature columns in both the initial (I) dataset and three resampled datasets; Undersampling (U), Oversampling (O), and Hybrid sampling (H)

Tactics	Occurrence in Database			
	I	U	O	H
Initial Access	158	45	241	118
Execution	396	113	810	353
Persistence	480	125	855	419
Privilege Escalation	321	80	560	281
Defense Evasion	635	157	1153	547
Credential Access	234	66	516	226
Discovery	267	90	706	223
Lateral Movement	269	89	639	254
Collection	165	58	496	142
Command and Control	304	96	816	245
Exfiltration	86	86	838	163

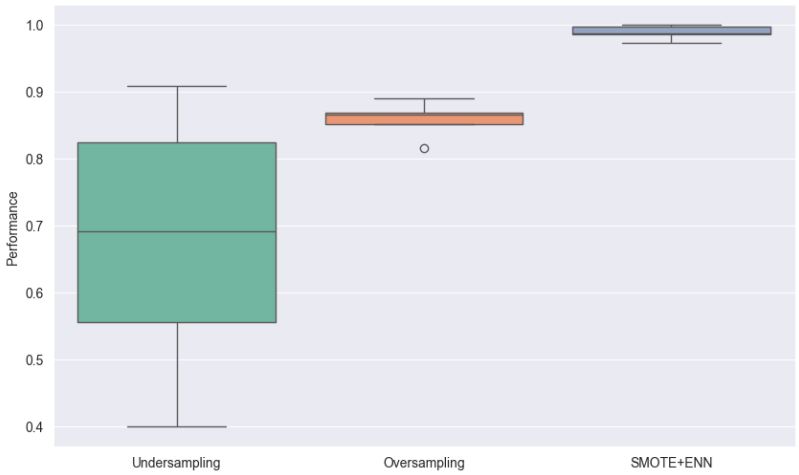

FIGURE 2. Performance comparison between three resampling techniques, Undersampling, Oversampling, and Hybrid (SMOTE+ENN), based on 5 metrics.

The random undersampling technique results in information loss as it randomly removes samples from the majority class without considering the distribution or importance of those samples. This has a negative impact on model performance [40]. On the other hand, the oversampling method using SMOTE already produces better model performance compared to random undersampling, though it is still not optimal. Interestingly, for the precision metric, undersampling outperforms oversampling. SMOTE often struggles with handling noise, class overlap, and small data subsets, which hinders its classification performance and may even degrade it further [41]. Based on this performance comparison, we will continue using the dataset produced by the hybrid sampling method (SMOTE+ENN) in the next stage of model development.

C. OPTIMIZING THE FEATURE SPACE THROUGH FEATURE SELECTION

Feature selection aims to identify significantly informative features while eliminating those less relevant to prediction outcomes. This results in a reduced feature set with lower dimensionality than the initial set, improving efficiency when implementing the chosen model. We explored the characteristics of the dataset's features by calculating Mutual Information, Pearson Correlation, and Feature Importance. Using these three methods complementarily enables the identification and selection of the most relevant and informative features, thereby enhancing the accuracy and efficiency of the predictive model.

1) Mutual Information

We utilized Mutual Information (MI) to measure the dependence or association between two variables in our dataset, capturing continuous and categorical relationships without assuming linearity. This approach allows us to uncover nonlinear and complex interactions between features and the target, enabling the identification of features that provide the most information about the target. Mathematically, the Mutual Information between two variables is defined in (Eq. 1) [42]:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \quad (1)$$

Where:

- $I(X;Y)$ is the mutual information between feature X and target Y.
- $p(x,y)$ is the joint probability distribution of X and Y.
- $p(x)$ and $p(y)$ are the marginal probability distributions of X and Y, respectively.

TABLE 4. Mutual Information of Each Feature from Highest to Lowest

Feature	Tactics	MI
Feature_9	Collection	0.083
Feature_10	Command and Control	0.076
Feature_8	Lateral Movement	0.056
Feature_7	Discovery	0.049
Feature_6	Credential Access	0.039
Feature_2	Execution	0.026
Feature_4	Privilege Escalation	0.025
Feature_3	Persistence	0.005
Feature_5	Defense Evasion	0.004
Feature_1	Initial Access	0.000

The MI value is always non-negative, ranging from 0 to 1. A value of 0 indicates that the two features are completely independent and share no information, while a higher MI value suggests that more information is shared between the two variables. For each feature, we measured the MI to determine whether the feature was highly informative for predicting the target. Consequently, features with a high MI value relative to the target are important for the model's inclusion.

Table 4 shows the MI values for each feature in relation to Exfiltration as the target variable. It can be observed that

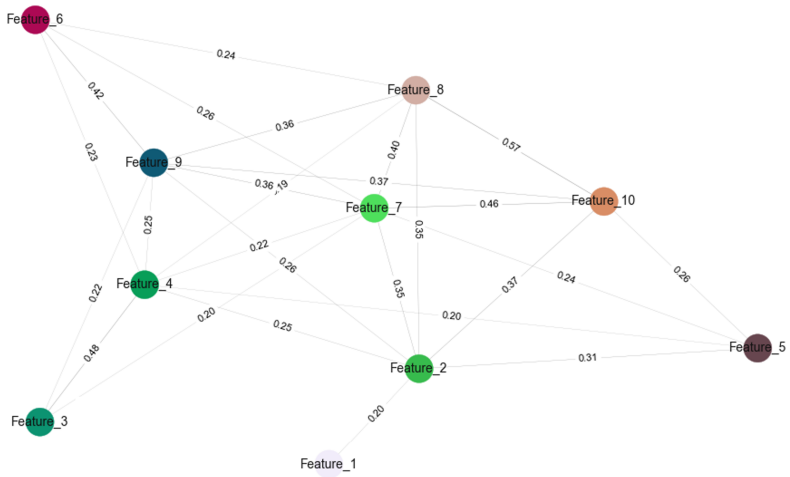

FIGURE 3. Pearson Correlation of the features in our dataset.

Feature_9, Feature_10, and Feature_8 have the highest MI values, indicating that these three features provide the most information and should be prioritized for retention in the model. We also evaluated the consistency of these three features in the subsequent step, Feature Selection.

2) Pearson Correlation

Pearson correlation is a metric that quantifies the intensity and direction of the linear relationship between two variables. [43]. We applied this method to identify feature pairs with strong correlations, aiding in feature selection by considering the relevance between variables. Mathematically, the Pearson correlation coefficient (r) is defined in (Eq.2) [44]:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \qquad (2)$$

Where:

- r is the Pearson correlation coefficient between feature X and target Y.
- X_i and Y_i are the individual sample values of X and Y.
- $\bar{X}$ and $\bar{Y}$ are the means of X and Y, respectively.
- n is the number of samples.

The coefficient value ranges from -1 to 1, with the following interpretation: a value of -1 indicates a perfect negative linear relationship between two features, a value of 0 means there is no linear relationship, and a value of 1 signifies a perfect positive linear relationship. In Fig.3, we present a network graph of the Pearson correlation value, with each node representing each feature and each weighted edge representing the correlation value between features. Based on Fig.3, several features exhibit prominent correlation values, indicating a strong linear relationship, specifically Feature_7, Feature_8, Feature_9, and Feature_10. We will examine the consistency of this characteristic with the results from other data measurements.

3) Feature Importance

Feature importance analysis is crucial for identifying and prioritizing significant features in model performance and computational efficiency by reducing dataset dimensionality [45]. We analyze feature importance by calculating the relative importance score of each feature using Random Forest algorithm [46]. The higher score significantly contributes to the predictive model by minimizing impurity at decision nodes. In contrast, the lower score has less impact due to weaker relationships. Higher scores are crucial for model performance and features with above-average scores should be prioritized for selection, while those with lower scores may be considered for dimensionality reduction. As shown in Fig.4, Feature_9 was identified as the most impactful, while Feature_4 was found to be the least impactful. Moreover, we selected each feature that has a relative importance score above 10 (mediant point). Therefore, Feature_9, Feature_10, Feature_6, Feature_7, and Feature_8 are selected based on their feature importance score.

4) Chi-square

From the previous discussion, we have explored the characteristics of the features within the dataset, including correlations between features, those with the most information, and the most impactful features for the prediction model. Moreover, we conducted filter-based feature selection using Chi-square score calculations. Chi-square is a statistical test used to determine the independence between two events, and in the context of feature selection, it assesses the relevance of features to the target variable while enhancing model performance [47], [48].

The results of the Chi-square calculations for our dataset are presented in Fig. 5 in accordance with Equation 3. The Chi-square scores for the ten features range from 0 to 175. It is clear that the top four features exhibit significantly higher scores compared to the bottom six. We have two options: we

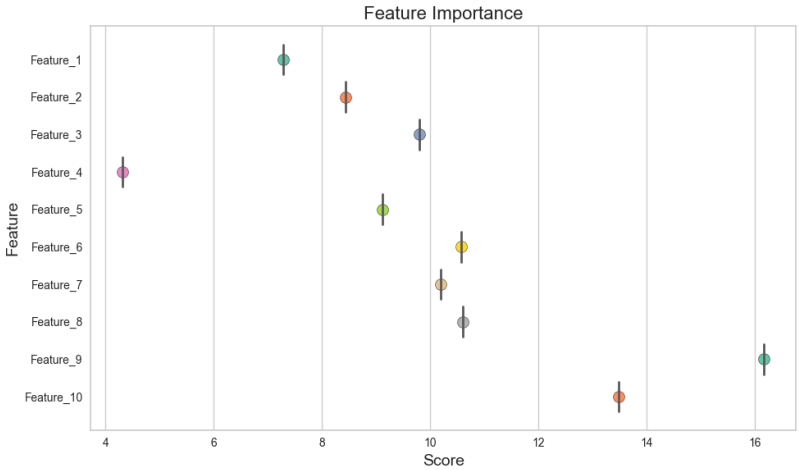

FIGURE 4. **Feature Importance of 10 features in our dataset.**

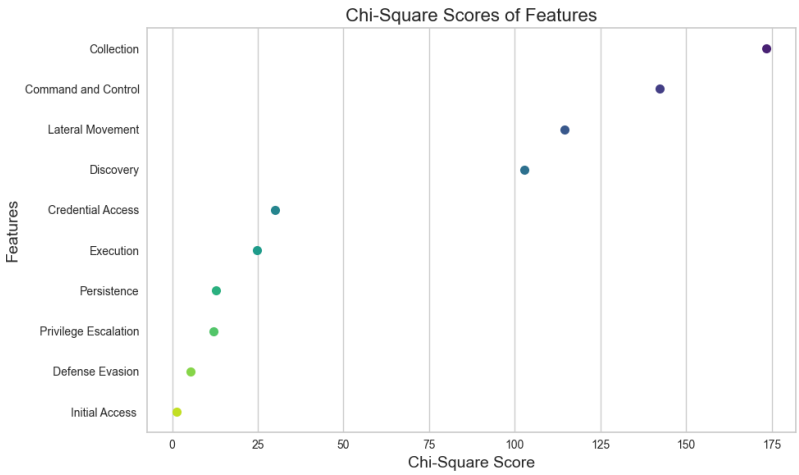

FIGURE 5. **Chi-square score of 10 features in the dataset used.**

can either select the top four features with the highest scores or set a threshold of 86, which represents the median score. In this case, features with Chi-square scores exceeding 100 will be retained for the final model implementation, while the others will be excluded. This method effectively reduces the dimensionality of the dataset, thus improving the computational efficiency of the final model. In the next section, we will compare the performance of the model using the complete set of ten features with that of the selected reduced feature set.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{3}$$

Where:

- χ^2 is the Chi-Square score for a specific feature in the dataset.
- O_i is the observed frequency for the i-th class (e.g., the actual occurrences of a class label given a feature value).
- E_i is the expected frequency for the i-th class (e.g., the expected occurrences under the assumption of independence between the feature and the target variable).
- k is the total number of possible classes.

5) Recursive Feature Elimination (RFE)

Besides the four filter-based methods explained above, feature selection can also be performed using wrapper-based approaches. In contrast to filter approaches, which examine features independently of the model, wrapper methods concentrate on evaluating different subsets of features by training the model on each subset and measuring its performance to identify the most effective features. [42]. This approach enables wrapper methods to identify subsets of features that are more relevant in the context of the model used. One of the techniques in wrapper methods is Recursive Feature

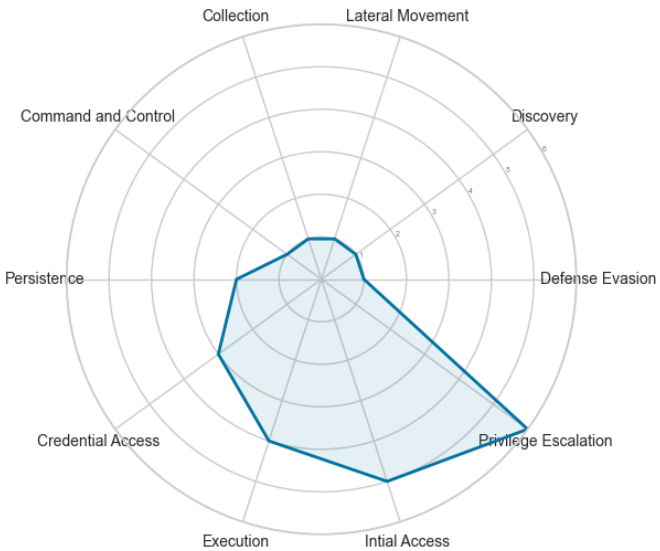

FIGURE 6. **RFE Feature Ranking of 10 features in dataset used.**

Elimination (RFE), which we selected due to its iterative nature of training the model with all features and progressively removing those with the lowest contribution while evaluating model performance at each step. Fig.6 illustrates the ranking of features using RFE, highlighting the top five features: Discovery, Lateral Movement, Defense Evasion, Collection, and Command and Control.

Following the feature selection techniques applied, we identified which features consistently appear across all techniques. Table 5 presents the results of the techniques, where selected features in each technique are marked with an "S" to denote their selection. Notably, four features—Discovery, Lateral Movement, Collection, and Command and Control—consistently have the "S" mark across all techniques. Therefore, these four features were selected for use in the subsequent steps.

TABLE 5. **Summary of Feature Selection techniques used in our dataset**

Features	Tactics	Feature Selection				
		MI	PC	FI	CS	RFE
Feature_1	Initial Access	-	-	-	-	-
Feature_2	Execution	-	-	-	-	-
Feature_3	Persistence	-	-	-	-	-
Feature_4	Privilege Escalation	-	-	-	-	-
Feature_5	Defense Evasion	-	-	-	-	S
Feature_6	Credential Access	-	-	S	-	-
Feature_7	Discovery	S	S	S	S	S
Feature_8	Lateral Movement	S	S	S	S	S
Feature_9	Collection	S	S	S	S	S
Feature_10	Command and Control	S	S	S	S	S

D. DETERMINING OUR FINAL ML MODEL

1) Comparing Ml Algorithms for Optimal Performance

To identify the best-performing model for our proposed method, we compared five commonly used ML algorithms on imbalanced datasets. We considered Random Forest (RF) [49]–[52], Support Vector Machine (SVM) [50], [53], [54], Logistic Regression (LR) [50], [51], [55], XGBoost [55]–[57], and Naïve Bayes (NB) [54], [58], [59], as these algorithms are frequently employed for imbalanced datasets. In this comparative analysis, we utilized a comprehensive set of metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, with the results presented in Fig. 7. It is clear that Random Forest outperforms all other algorithms across all metrics. Logistic Regression had the lowest performance, followed closely by Naïve Bayes, which showed only a marginal improvement. Support Vector Machine (SVM) and XGBoost exhibited similar performance levels and delivered satisfactory results, although both were still inferior to Random Forest. Consequently, the Random Forest algorithm was selected as the final model to be integrated into the complete workflow of our proposed method, which is detailed in the next section.

2) Validating our ML model using Cross-Validation and Repeated Cross-validation

We conducted a thorough validation procedure on our final model to ensure its reliability and confirm that the performance achieved was not coincidental. Additionally, validation plays a crucial role in assessing the stability of performance estimates, especially in studies with limited sample sizes, where the selection of training and testing samples can significantly impact the results [60]. Two essential techniques employed in this validation process are cross-validation (CV) and repeated cross-validation (RCV).

The CV allows for an unbiased estimate of model performance by partitioning the dataset into training and validation sets, thus ensuring that the model's performance is not overly optimistic due to overfitting [61]. K-fold cross-

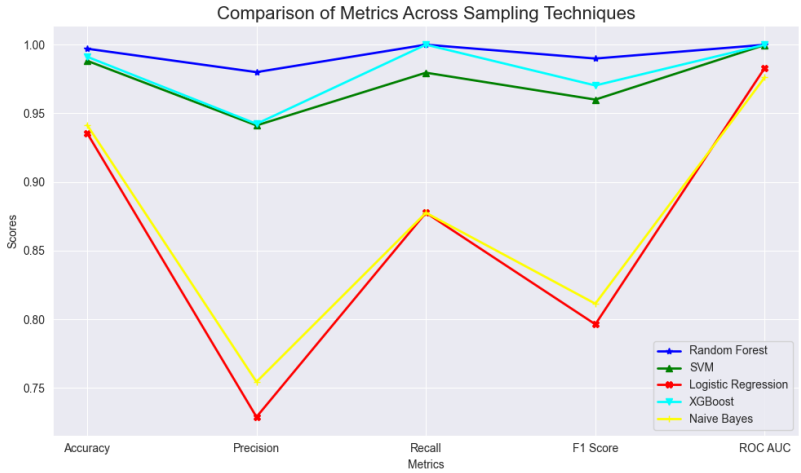

FIGURE 7. **Comparison of model performance using different ML algorithms.**

TABLE 6. **Comparison of objectives, approaches, and performance between selected works.**

Authors	Objective	Approach	Performance
Okada et al. [30]	Detecting targeted attacks	Deep learning analysis	Accuracy 0.90
Kim et al [31]	Predicting attack goal	Bayesian network	F-1 Score 0.565
Gabrys et al [27]	Predicting tactics and techniques	SVM, KNN, Naïve Bayes, XGBoost, RF	Accuracy 0.92-0.95
Our model	Predicting attack goal (exfiltration tactic)	LR, Naïve Bayes, SVM, XGBoost, RF	Accuracy 0.935, 0.941, 0.988, 0.991, 0.997

validation, where the dataset is divided into K subsets, is particularly prevalent in ML applications, as it provides a robust method for assessing model generalization [62]. We consider conducting the CV of our RF model using 5-folds and 10-folds, which obtained mean accuracy of 0.991 and 0.998, respectively.

The RCV repeatedly applies the CV technique to reduce performance estimate variability and stabilize model correctness [63]. Furthermore, the choice of cross-validation technique can significantly impact the results, emphasizing the need for careful selection based on the specific characteristics of the dataset and the model being evaluated [64]. Here, we utilized 5-folds with 5 repetitions with our RF model, resulting in 0.9972 mean accuracy. Moreover, we conducted 10-folds with 10 repetitions and obtained a mean accuracy of 0.9974.

Based on the validation results using four different approaches, including 5-fold cross-validation, 10-fold cross-validation, repeated cross-validation with 5-folds and 5 repetitions, as well as 10-folds with 10 repetitions, it was found that the model's mean accuracy ranged between 0.993 and 0.997. This consistent accuracy range indicates that the model exhibits highly stable and reliable performance, regardless of the variation in validation techniques. These results reinforce the claim that the model not only performs well on training data but also demonstrates resilience against overfitting. Therefore, the model can be considered sufficiently generalizable for use with new data, with a very low potential for bias.

3) Benchmarking the performance between our model and the existing studies

Based on the experimental results obtained, we present a performance benchmark of selected existing studies [27], [30], [31] that share similar objectives with our research. A review of these selected studies is provided in Section II-B. Table 6 summarizes the objectives and approaches of each study, along with their respective performance outcomes, to illustrate the benchmarking between studies. It is evident that our ML model, particularly the one utilizing the RF algorithm, outperforms the results achieved by the existing studies.

IV. PROPOSED METHOD

This section offers a comprehensive description of our proposed methodology, outlining each step involved. As described in previous sections, our method aims to predict the occurrence of exfiltration within a system using a machine learning approach. Benefiting our final ML model in Section III, we leverage an integration of internal data, specifically event logs, with our ML model to predict whether identified tactics observed from event logs lead to exfiltration. We also highlight a challenge encountered when integrating low-level logs and ML predictions for identifying data breach steps at a conceptual level. In this section, we explain the approach taken to overcome this challenge. Therefore, our proposed method has two main contributions, including (1) addressing the gap level between low-level logs and high-level data breach conceptual steps, and (2) integrating collected event logs and ML models to predict exfiltration tactics. Further-

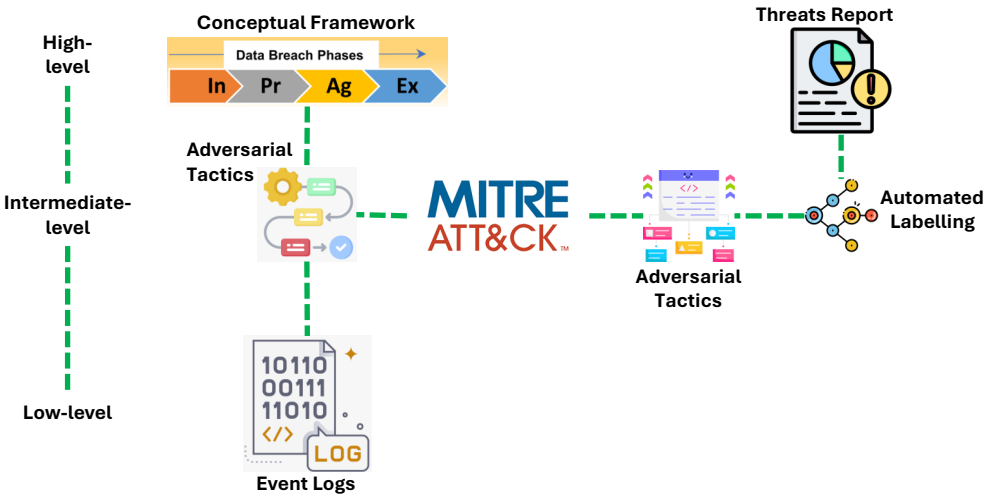

FIGURE 8. **Adversarial Tactics from ATT&CK as intermediate level to address the gap.**

more, we present scenarios to illustrate how our proposed method operates in real-world cases.

A. ADDRESSING THE GAP LEVEL BETWEEN LOW-LEVEL LOGS AND HIGH-LEVEL DATA BREACH CONCEPTUAL STEPS

In our previous study [22], the content analysis process, particularly the sub-process of mapping artifacts to the Data Breach Breakdown (DBB) phase, was conducted manually and relied heavily on the investigator's expertise. The investigator needed to identify an artifact, typically low-level logs, and assess its relevance to the DBB phase, which consists of high-level data breach conceptual steps, such as Infiltration (In), Propagation (Pr), Aggregation (Ag), or Exfiltration (Ex). When a large volume of low-level logs needs to be analyzed and mapped, this manual approach poses the risk of human error and inefficiency in terms of time.

Therefore, this paper focuses on developing an automated method for analyzing low-level logs and mapping them to the DBB phase. We concentrate on the final DBB phase, Exfiltration (Ex), as previously explained, which is a crucial determinant of a data breach. The primary challenge in automating this process is the information relevance gap between logs and the conceptual steps, which prevents a direct connection between them. We utilize ATT&CK, specifically Enterprise Tactics, to address this challenge at an intermediate level.

We employed a similar approach to integrate relevant data from threat reports with event logs. These two data sources are bridged through the perspective of adversarial tactics. The adversarial tactics from threat reports are obtained through an automated labeling process. In contrast, the adversarial tactics from event logs are automatically mapped based on the event IDs found within the logs. An illustration of this concept can be seen in Fig. 8.

B. INTEGRATING COLLECTED EVENT LOGS AND ML MODEL TO PREDICT EXFILTRATION TACTIC

In this section, we provide a detailed explanation of the proposed method. The final model is employed to make predictions based on new queries derived from event logs collected by the system. Furthermore, the intermediate level is integrated into a unified workflow within the proposed method, as illustrated in Fig.9.

The process begins with event ID parsing to capture the event IDs contained within the event logs. These extracted event IDs are then mapped to their corresponding ATT&CK tactics based on a lookup table adopted from the mapping [65]. The lookup table consists of more than 270 EVTX samples mapped to MITRE ATT&CK tactics and techniques. When identical event IDs appeared across different samples, repeated mappings were frequently found in the provided lookup table. To address this, we cleaned up these redundant mappings to streamline the process and improve time efficiency. Additionally, as our study focuses on four selected features, the mapping was also narrowed to include only event IDs related to these four features. Mappings to other tactics were removed from our lookup table to enhance the efficiency of the mapping process further.

Each identified tactic is then encoded into a query of size {1,4}, where the four columns are filled with 1 or 0, respectively, representing the presence or absence of the detected tactics. The query is structured as a single row with four feature columns aligned with the results of the feature selection process that selected the four most significant features for the prediction.

C. CASE STUDY

This section presents a case study analysis illustrating the application of our method in real-world scenarios. For this purpose, we used three incidents. A, B, and C, that were

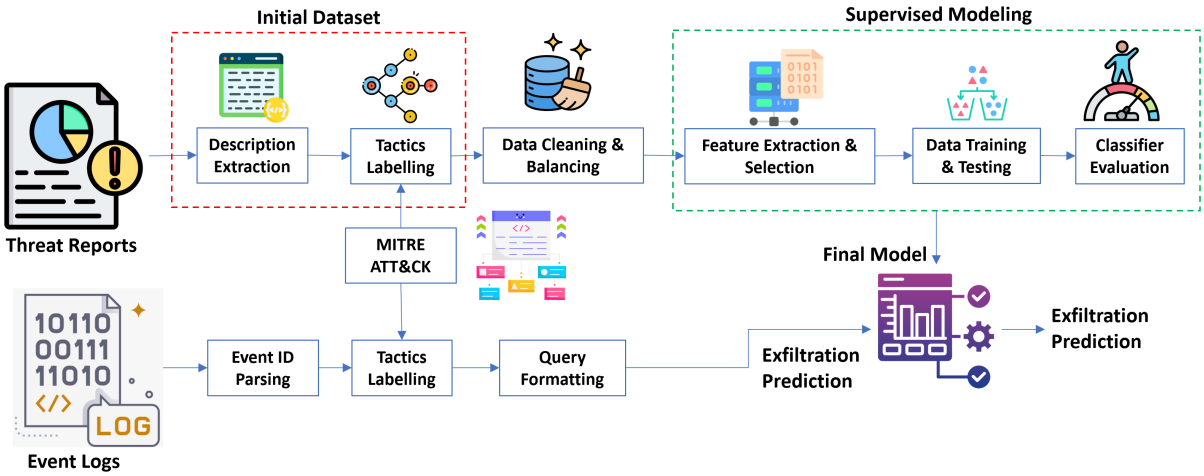

FIGURE 9. **Proposed Method: Exfiltration Prediction using Supervised Machine Learning based on Tactics Mapping from Threat Reports and Event Logs.**

TABLE 7. **Comparison of objectives, approaches, and performance between selected works.**

Incident	Tactics Obtained	Prediction	Analysis
A	Discovery, Lateral movement, collection	Exfiltration	accurate
B	Discovery, Lateral movement	Non-Exfiltration	accurate
C	Discovery, Lateral movement, collection, Command and Control	Exfiltration	accurate

duplicates of actual incidents in a laboratory environment. Furthermore, we used system and security event logs that were preserved from each host involved in the respective incidents. In addition, we considered the composition of two incidents in which exfiltration occurred and one incident where no exfiltration took place. This composition ensures that our method is proven for its ability to accurately predict both exfiltration and non-exfiltration incidents.

First, incident A involved exfiltration. In this case, the hacker exploited a vulnerability in an outdated version of the web application framework, generating suspicious web shell files in the web directory, which were then used to remotely access the web server and dump data from the database. Next, incident C, which also involved exfiltration, occurred when a hacker remotely controlled the user's PC after the user ran malware attached to a downloaded email. The hacker subsequently created a domain administrator account in Active Directory by exploiting the Serologon vulnerability and used the account to steal files from the File Server. In contrast, incident B was a ransomware attack that did not involve exfiltration. This incident occurred when a user's PC was compromised remotely via an open RDP port. The weak password on the PC allowed a brute force attack, enabling the hacker to log in, copy, and execute ransomware, resulting in file encryption on the PC.

Table 7 presents the results of tactics extraction from the automated mapping process and the exfiltration predictions for the three incidents using our method. Both incidents A and C were accurately predicted as involving exfiltration. Similarly, for incident B, our method correctly predicted that exfiltration did not occur. Through this case study analysis, our proposed method demonstrates its effectiveness and precision in predicting the occurrence of exfiltration.

V. CONCLUSION

This study aimed to develop a machine learning (ML) model designed to predict exfiltration as the objective of an attack, based on the sequence of tactics employed by the attacker. These tactics were identified through the automated mapping of collected event logs, which were subsequently analyzed using the ML model developed in this research. We created a new dataset by processing the identified tactics from existing threat reports, aligning with the main idea of this study—predicting exfiltration as the target variable based on a set of ten tactics used as features. Although the dataset was imbalanced, the best-performing ML model was obtained through a comprehensive comparison of three resampling methods, five feature selection techniques (using both filter and wrapper methods), and five commonly used ML algorithms for imbalanced datasets. The results of this investigation indicate that hybrid sampling with SMOTE+ENN was the most effective resampling method, and four features—Discovery, Lateral Movement, Collection, and Command and Control—were consistently selected across all five feature selection techniques. Furthermore, the Random Forest model demonstrated superior performance compared to Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayes, and XGBoost in terms of overall performance. This study further demonstrates the applicability of the proposed method to real-world incidents. Subsequent research

may investigate the development of generative machine learning models utilizing a range of new queries input into the model. Moreover, advancing the model through transfer learning—where knowledge is transferred not only from threat reports but also from vulnerability reports and data breach notification letters—represents a promising direction for future studies.

ACKNOWLEDGMENT

This publication is supported by the Hibah PUTI Q2 2023, Universitas Indonesia, under Contract NKB-820/UN2.RST/HKP.05.00/2023. The work of Arif Rahman Hakim was funded by the Lembaga Pengelola Dana Pendidikan (LPDP), Ministry of Finance of the Republic of Indonesia. The authors also extend their gratitude to the Indonesia Cyber Awareness and Resilience Center, Universitas Indonesia (IdCARE-UI) through the IT Forensics Enablement Workshop supported by the Japan International Cooperation Agency (JICA) for providing the case study data essential to this research. Additionally, AI-based generative tools were employed for text editing and grammar checking, ensuring clarity and precision in the manuscript's language.

REFERENCES

[1] C. D. Raab, "Information privacy, impact assessment, and the place of ethics," *Computer Law & Security Review*, vol. 37, p. 105404, 2020.

[2] A. A. Teoh, N. B. A. Ghani, M. Ahmad, N. Jhanjhi, M. A. Alzain, and M. Masud, "Organizational data breach: Building conscious care behavior in incident," *Organizational data breach: Building conscious care behavior in incident response. Computer Systems Science and Engineering*, vol. 40, no. 2, pp. 505–515, 2022.

[3] K. Masuch, M. Greve, S. Trang, and L. M. Kolbe, "Apologize or justify? examining the impact of data breach response actions on stock value of affected companies?" *Computers & Security*, vol. 112, p. 102502, 2022.

[4] D. Kolevski, K. Michael, R. Abbas, and M. Freeman, "Cloud data breach disclosures: The consumer and their personally identifiable information (pii)?" in *2021 IEEE Conference on norbert wiener in the 21st century (21CW)*. IEEE, 2021, pp. 1–9.

[5] Z. Zhang, J. Hu, L. Ma, R. Pei, and P. Wang, "Bvfb: Training behavior verification mechanism for secure blockchain-based federated learning," *Computing and Informatics*, vol. 41, no. 6, pp. 1401–1424, 2022.

[6] A. H. Almulihi, F. Alassery, A. I. Khan, S. Shukla, B. K. Gupta, and R. Kumar, "Analyzing the implications of healthcare data breaches through computational technique." *Intelligent Automation & Soft Computing*, vol. 32, no. 3, 2022.

[7] H. N. Chua, J. S. Ooi, and A. Herbland, "The effects of different personal data categories on information privacy concern and disclosure," *Computers & Security*, vol. 110, p. 102453, 2021.

[8] I.T.R.C., "2023 data breach report," available:. [Online]. Available: https://www.idtheftcenter.org/publication/2023-data-breach-report/

[9] I.B.M., "Cost of a data breach report 2024," *IBM Security*, pp. 1–73,, available:. [Online]. Available: https://www.ibm.com/security/data-breach

[10] Verizon, "Dbir: Data breach investigations report 2023," online]. Available:. [Online]. Available: https://www.verizon.com/business/resources/T8b/reports/2023-data-breach-investigations-report-dbir.pdf

[11] A. Ibrahim, D. Thiruvady, J.-G. Schneider, and M. Abdelrazek, "The challenges of leveraging threat intelligence to stop data breaches," *Frontiers in Computer Science*, vol. 2, p. 36, 2020.

[12] P. Zanke and D. Sontakke, "Safeguarding patient confidentiality in telemedicine: A systematic review of privacy and security risks, and best practices for data protection," *International Journal of Current Science Research and Review*, vol. 07, no. 06, pp. 3910–3922,.

[13] S. B. Son, S. Park, H. Lee, Y. Kim, D. Kim, and J. Kim, "Introduction to mitre att&ck: concepts and use cases," in *2023 International Conference on Information Networking (ICOIN)*. IEEE, 2023, pp. 158–161.

[14] V. Solange, M. Legoy, M. Thesis, and M. Caselli, "Retrieving attck tactics and techniques in cyber threat reports," no. January, 2019, [Online]. Available:. [Online]. Available: https://otx.alienvault.com/browse/pulses

[15] D. Molitor, A. Saharia, V. Raghupathi, and W. Raghupathi, "Exploring the characteristics of data breaches: A descriptive analytic study," *Journal of Information Security*, vol. 15, no. 2, pp. 168–195, 2024.

[16] J. King, G. Bendiab, N. Savage, and S. Shiaeles, "Data exfiltration: methods and detection countermeasures," in *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*. IEEE, 2021, pp. 442–447.

[17] B. Sabir, F. Ullah, M. A. Babar, and R. Gaire, "Machine learning for detecting data exfiltration: A review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–47, 2021.

[18] W. Sus and P. Nawrocki, "Signature-based adaptive cloud resource usage prediction using machine learning and anomaly detection," *Journal of Grid Computing*, vol. 22, no. 2, p. 46, 2024.

[19] I. H. Sarker, A. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: an overview from machine learning perspective," *Journal of Big data*, vol. 7, pp. 1–29, 2020.

[20] N. M. Karie and L. F. Sikos, "Cybersecurity incident response in the enterprise," in *Next-Generation Enterprise Security and Governance*. CRC Press, 2022, pp. 83–119.

[21] M. R. Rahman and L. Williams, "From threat reports to continuous threat intelligence: a comparison of attack technique extraction methods from textual artifacts," *arXiv preprint arXiv:2210.02601*, 2022.

[22] A. R. Hakim, K. Ramli, T. S. Gunawan, and S. Windarta, "A novel digital forensic framework for data breach investigation," *IEEE Access*, vol. 11, pp. 42 644–42 659, 2023.

[23] M.I.T.R.E., "Enterprise tactics," *Accessed: Aug*, vol. 28, online]. Available:. [Online]. Available: https://attack.mitre.org/tactics/enterprise/

[24] I. Branescu, O. Grigorescu, and M. Dascalu, "Automated mapping of common vulnerabilities and exposures to mitre att&ck tactics," *Information*, vol. 15, no. 4, p. 214, 2024.

[25] E. Domschot, R. Ramyaa, and M. R. Smith, "Improving automated labeling for att&ck tactics in malware threat reports," *Digital Threats: Research and Practice*, vol. 5, no. 1, pp. 1–16, 2024.

[26] Y.-T. Huang, R. Vaitheeshwari, M.-C. Chen, Y.-D. Lin, R.-H. Hwang, P.-C. Lin, Y.-C. Lai, E. H.-K. Wu, C.-H. Chen, Z.-J. Liao *et al.*, "Mitretrieval: Retrieving mitre techniques from unstructured threat reports by fusion of deep learning and ontology," *IEEE Transactions on Network and Service Management*, 2024.

[27] R. Gabrys, M. Bilinski, S. Fugate, and D. Silva, "Using natural language processing tools to infer adversary techniques and tactics under the mitre att&ck framework," in *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2024, pp. 0541–0547.

[28] Y. Andrew, C. Lim, and E. Budiarto, "Mapping linux shell commands to mitre att&ck using nlp-based approach," in *2022 International Conference on Electrical Engineering and Informatics (ICELTICs)*. IEEE, 2022, pp. 37–42.

[29] O. Grigorescu, A. Nica, M. Dascalu, and R. Rughinis, "Cve2att&ck: Bert-based mapping of cves to mitre att&ck techniques," *Algorithms*, vol. 15, no. 9, p. 314, 2022.

[30] S. Okada, Y. Katano, Y. Kozai, and T. Mitsunaga, "Predicting and visualizing lateral movements based on att&ck and quantification theory type 3," *Journal of Cases on Information Technology (JCIT)*, vol. 26, no. 1, pp. 1–14, 2024.

[31] Y. Kim, I. Lee, H. Kwon, K. Lee, and J. Yoon, "Ban: Predicting apt attack based on bayesian network with mitre att&ck framework," *IEEE Access*, 2023.

[32] Y. Liu and Y. Guo, "Cl-ap2: A composite learning approach to attack prediction via attack portraying," *Journal of Network and Computer Applications*, vol. 230, p. 103963, 2024.

[33] A. B. de Neira, A. M. de Araujo, and M. Nogueira, "An intelligent system for ddos attack prediction based on early warning signals," *IEEE Transactions on Network and Service Management*, vol. 20, no. 2, pp. 1254–1266, 2022.

[34] O. Almomani, A. Alsaaidah, A. A. A. Shareha, A. Alzaqebah, and M. Almomani, "Performance Evaluation of Machine Learning Classifiers for Predicting Denial-of-Service Attack in Internet of Things," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 1, pp. 263–271, 2024.

[35] V. legoy, "Reports Classification by Adversarial Tactics and Techniques," 2020. [Online]. Available: https://github.com/vlegoy/rcATT

[36] Z. Sun, W. Ying, W. Zhang, and S. Gong, "Undersampling method based on minority class density for imbalanced data," *Expert Systems with Applications*, vol. 249, p. 123328, 2024.

[37] H. Sug, "An oversampling technique with descriptive statistics," *WSEAS Transactions Information Science Applications*, vol. 21, pp. 318–332,.

[38] I. Riantika, B. Sartono, and K. A. Notodiputro, "Effectiveness of smote-enn to reduce complexity in classification model," *Indonesian Journal of Statistics and Its Applications*, vol. 8, no. 1, pp. 70–82, 2024.

[39] R. Bounab, B. Guelib, and K. Zarour, "A novel machine learning approach for handling imbalanced data: Leveraging smote-enn and xgboost," in *2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*. IEEE, 2024, pp. 1–7.

[40] S. Feng, J. Keung, Y. Xiao, P. Zhang, X. Yu, and X. Cao, "Improving the undersampling technique by optimizing the termination condition for software defect prediction," *Expert Systems with Applications*, vol. 235, p. 121084, 2024.

[41] N. A. Azhar, M. S. M. Pozi, A. M. Din, and A. Jatowt, "An investigation of smote based methods for imbalanced datasets with data complexity analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 7, pp. 6651–6672, 2022.

[42] G. Yuan, L. Lu, and X. Zhou, "Feature selection using a sinusoidal sequence combined with mutual information," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 107168, 2023.

[43] L. Xu, Y. Wang, L. Mo, Y. Tang, F. Wang, and C. Li, "The research progress and prospect of data mining methods on corrosion prediction of oil and gas pipelines," *Engineering Failure Analysis*, vol. 144, p. 106951, 2023.

[44] C. Hongsong, M. Caixia, F. Zhongchuan, and C.-H. Lee, "Novel ldos attack detection by spark-assisted correlation analysis approach in wireless sensor network," *IET Information Security*, vol. 14, no. 4, pp. 452–458, 2020.

[45] A. A. Megantara and T. Ahmad, "Feature importance ranking for increasing performance of intrusion detection system," in *2020 3rd International Conference on Computer and Informatics Engineering (IC2IE)*. IEEE, 2020, pp. 37–42.

[46] I. M. Zubair and B. Kim, "A group feature ranking and selection method based on dimension reduction technique in high-dimensional data," *IEEE Access*, vol. 10, pp. 125 136–125 147, 2022.

[47] J. A. Widians, R. Wardoyo, and S. Hartati, "Feature selection based on chi-square and ant colony optimization for multi-label classification." *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 14, no. 3, 2024.

[48] A. F. Rasheed, M. Zarkoosh, and S. S. Al-Azzawi, "The impact of feature selection on malware classification using chi-square and machine learning," in *2023 9th International Conference on Computer and Communication Engineering (ICCCE)*. IEEE, 2023, pp. 211–216.

[49] A. More and D. P. Rana, "Review of random forest classification techniques to resolve data imbalance," in *2017 1st International conference on intelligent systems and information management (ICISIM)*. IEEE, 2017, pp. 72–78.

[50] R. M. Tischio and G. M. Weiss, "Identifying Classification Algorithms Most Suitable for Imbalanced Data," *Proceedings of the 15th International Conference on Data Science, 106-111, Las Vegas, NV.*, pp. 1–6, 2019.

[51] H. Luo, X. Pan, Q. Wang, S. Ye, and Y. Qian, "Logistic regression and random forest for effective imbalanced classification," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 1. IEEE, 2019, pp. 916–917.

[52] B. Pes, "Learning from high-dimensional and class-imbalanced datasets using random forests," *Information*, vol. 12, no. 8, p. 286, 2021.

[53] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 4, pp. 42–47, 2012. [Online]. Available: http://www.ijetae.com/files/Volume2Issue4/IJETAE_0412_07.pdf

[54] G. Oreski and S. Oreski, "An experimental comparison of classification algorithm performances for highly imbalanced datasets," in *Central European Conference on Information and Intelligent Systems*. Faculty of Organization and Informatics Varazdin, 2014, p. 4.

[55] N. H. N. B. M. Shahri, S. B. S. Lai, M. B. Mohamad, H. A. B. A. Rahman, and A. B. Rambli, "Comparing the performance of adaboost, xgboost, and logistic regression for imbalanced data," *Mathematics and Statistics*, vol. 9, no. 3, pp. 379–385, 2021.

[56] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," *International Journal of Distributed Sensor Networks*, vol. 18, no. 6, 2022.

[57] S. He, B. Li, H. Peng, J. Xin, and E. Zhang, "An Effective Cost-Sensitive XGBoost Method for Malicious URLs Detection in Imbalanced Dataset," *IEEE Access*, vol. 9, pp. 93 089–93 096, 2021.

[58] C. K. Aridas, S. Karlos, V. G. Kanas, N. Fazakis, and S. B. Kotsiantis, "Uncertainty Based Under-Sampling for Learning Naive Bayes Classifiers under Imbalanced Data Sets," *IEEE Access*, vol. 8, pp. 2122–2133, 2020.

[59] M. C. L. Amit Gupta and M. Manchanda, "Financial fraud detection using naive bayes algorithm in highly imbalance data set," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 24, no. 5, pp. 1559–1572, 2021. [Online]. Available: https://doi.org/10.1080/09720529.2021.1969733

[60] V. K. Singh, M. J. Pencina, J. Liang, D. S. Berman, and P. J. Slomka, "Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging," *Scientific Reports*, 2021.

[61] Y. Xu and R. Goodacre, "On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning," *Journal of Analysis and Testing*, 2018.

[62] I. Tougui, A. Jilbab, and J. El Mhamdi, "Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications," *Healthcare informatics research*, vol. 27, no. 3, pp. 189–199, 2021.

[63] G. M. Merola, "Blocked cross-validation: A precise and efficient method for hyperparameter tuning," *arXiv preprint arXiv:2306.06591*, 2023.

[64] A. Nurhopipah and U. Hasanah, "Dataset splitting techniques comparison for face classification on cctv images," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 14, no. 4, pp. 341–352, 2020.

[65] Mdecrevoisier, "EVTX-to-MITRE-Attack," 2024. [Online]. Available: https://github.com/mdecrevoisier/EVTX-to-MITRE-Attack

ARIF RAHMAN HAKIM (Member, IEEE) received his bachelor's degree in cryptography from the National Crypto Institute, Bogor, Indonesia, in 2008 and a master's degree in communication and information systems from the School of Electronics and Information Engineering, Beihang University, Beijing, China, in 2016. He is currently pursuing a Ph.D. in electrical engineering at the Faculty of Engineering, Universitas Indonesia, Depok, Indonesia. Since 2017, he has been a lecturer at National Cyber and Crypto Polytechnic, Indonesia. His research interests include cybersecurity, digital forensics, and cryptography.

KALAMULLAH RAMLI (Member, IEEE) received a master's degree in telecommunication engineering from the University of Wollongong, Wollongong, NSW, Australia, in 1997 and a Ph.D. degree in computer networks from Universitaet Duisburg-Essen (UDE), NRW, Germany, in 2003. He has been a lecturer at Universitas Indonesia (UI) since 1994 and a professor of computer engineering since 2009. He currently teaches advanced communication networks, embedded systems, object-oriented programming, and engineering and entrepreneurship. His research interests include embedded systems, information and data security, computers and communication, and biomedical engineering. He is a prolific author, with more than 125 journals and conference papers and eight books and book chapters published.

MUHAMMAD SALMAN (Member, IEEE) is currently a lecturer and researcher at Computer Engineering, Faculty of Engineering, University of Indonesia (UI), especially in the field of Network and Information Security. He holds a Doctoral degree in Information Network Security from Universitas Indonesia, and Master degree in Information Technology from Monash University, Melbourne, Australia. He is also a Head of Computer Engineering Study Program of Faculty Engineering, UI. Regarding his field of expertise, he was a former board member as a Vice Chairman of ID-SIRTII (Indonesia Security Incident Response Team on Internet Infrastructure) under the Ministry of Information and Communication Technology, Republic of Indonesia. He is also co-founder and Board Member of Id-CARE.UI (Indonesia Cybersecurity and Resilience Center, Universitas Indonesia) for developing capacity building and research in cybersecurity-related fields. He has represented UI and Indonesia in various regional and international information security forums. His involvement in the field extends to delivering keynote addresses and participating in numerous conferences, workshops, and training sessions on topics such as information security, network infrastructure, ICT community development, professional education, and academic-industry partnerships.

BERNARDI PRANGGONO (Senior Member, IEEE) is an Associate Professor in Cyber Security and Computer Networks at the School of Computing and Information Science, Anglia Ruskin University, Cambridge, UK. Dr. Pranggono received his B. Eng. degree in Electronics and Telecommunication Engineering from Waseda University, Japan, an M. DigComms degree in Digital Communications from Monash University, Australia, and a Ph.D. degree in Electronics and Electrical Engineering from the University of Leeds, UK. He has previously held academic and research positions at Sheffield Hallam University, Glasgow Caledonian University, Queen's University Belfast, and the University of Leeds. He has held industrial positions at Accenture, Telstra, Oracle, and PricewaterhouseCoopers. His research interests include cybersecurity, the Internet of Things, AI, and green ICT.

ESTI RAHMAWATI AGUSTINA (Member, IEEE) received her bachelor's degree in cryptography from the National Crypto Institute, Bogor, Indonesia, in 2008; a master's degree in communication and information systems from School of Electronics and Information Engineering, Beihang University, Beijing, China, in 2018 and a master's degree in technology management, Faculty of Engineering, Universitas Krisnadwipayana, Indonesia, in 2022. She is currently pursuing a Ph.D. degree at the Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia. Since 2009, she has worked at National Cyber and Crypto Agency, Indonesia. Her research interests include cryptography and information security-related topics, especially IT security product evaluation, security and privacy preserving, and cryptographic protocols in VANET.