IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 September 2024, accepted 20 September 2024, date of publication 27 September 2024, date of current version 8 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3469552

RESEARCH ARTICLE

Clustering APT Groups Through Cyber Threat Intelligence by Weighted Similarity Measurement

ZHENG-SHAO CHEN[1], R. VAITHEESHWARI[1], ERIC HSIAO-KUANG WU[1],(Member, IEEE), YING-DAR LIN[2], (Fellow, IEEE), REN-HUNG HWANG[3], (Senior Member, IEEE), PO-CHING LIN[4],(Member, IEEE), YUAN-CHENG LAI[5], AND ASAD ALI[6]

[1]Department of Computer Science and Information Engineering, National Central University, Taoyuan 320317, Taiwan
[2]Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu 300093, Taiwan
[3]College of Artificial Intelligence, National Yang Ming Chiao Tung University, Tainan 711, Taiwan
[4]Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi 621301, Taiwan
[5]Department of Information Management, National Taiwan University of Science and Technology, Taipei 106335, Taiwan
[6]National Institute of Cyber Security, Ministry of Digital Affairs, Taipei 100057, Taiwan

Corresponding author: Eric Hsiao-Kuang Wu (hsiao@csie.ncu.edu.tw)

This work was supported in part by the Cybersecurity Center of Excellence (CCoE) Program through the National Science and Technology Council (NSTC), Taiwan.

ABSTRACT Advanced Persistent Threat (APT) groups pose significant cybersecurity threats due to their sophisticated and persistent nature. This study introduces a novel methodology to understand their collaborative patterns and shared objectives, which is crucial for developing robust defense mechanisms. We utilize MITRE ATT&CK Techniques, software, target nations, and industries as our primary features to understand the characteristics of APT groups. Since essential information is often buried within the unstructured data of Cyber Threat Intelligence (CTI) reports, we employ Natural Language Processing (NLP) and Named Entity Recognition (NER) to extract relevant data. To analyze and interpret the complex relationships between APT groups, we compute similarity among the features using weighted cosine similarity metrics and Machine Learning (ML) models, enhanced by feature crosses and feature selection strategies. Subsequently, hierarchical clustering is used to group APTs based on their similarity scores, helping to identify common behaviors and uncover deeper relationships. Our methodology demonstrates notable clustering performance, with a silhouette coefficient of 0.76, indicating strong intra-cluster similarity. The Adjusted Rand Index (ARI) of 0.63, though moderate, effectively measures agreement between our clustering and the ground truth. These metrics provide robust validation, surpassing commonly recognized benchmarks for effective clustering in cybersecurity. Our methodology successfully classifies 23 distinct APT groups into six clusters, highlighting the importance of techniques and industry features in the clustering process. Notably, techniques such as T1059 (Command and Scripting Interpreter) and T1036 (Masquerading) are prevalently deployed, observed in 18 out of 23 APT groups across all six clusters.

INDEX TERMS Advanced persistent threat (APT) groups, cyber threat intelligence (CTI) report, feature engineering, hierarchical clustering, named entity recognition, weighted similarity measurement.

I. INTRODUCTION

A significant concern for enterprises and organizations in the digital age is the surge in cyber threats. With the number and scale of Advanced Persistent Threat (APT) [1] groups escalating, the need for effective defense measures has never been greater. Organizations face an increasing demand to protect their digital assets and sensitive data against hackers, cybercriminals, and state-sponsored entities. To combat these risks, they must be armed with current and accurate Cyber Threat Intelligence (CTI), which provides insights into potential attackers, their tactics, techniques, and procedures,

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

as well as vulnerabilities and attack methods. The adoption of CTI has seen a marked increase, moving from a smaller fraction of organizations to becoming a more widely adopted practice over a recent period [2], [3], with a substantial 80.8% of organizations affirming an enhancement in their security posture, particularly in detection, prevention, and response capabilities. This reliance underscores the evolving complexity of the cyber threat landscape and the growing consensus on CTI as an indispensable element of contemporary cyber defense.

Cybersecurity firms, serving as the primary sources of this intelligence, cater to a diverse clientele with state-of-the-art solutions. They have broad access to threat intelligence data and produce comprehensive reports that delve into emergent cyber threats, offering guidance on best practices and defense strategies.

Recent trends underscore an increasing focus on key sectors by APT, with these adversaries conducting in-depth reconnaissance followed by sophisticated incursions designed to mask their presence and secure long-term access to target networks. The complexity and persistence of these attacks have been escalating, with entities repeatedly victimized by the same or similar APT groups. This trend is visually represented in Fig. 1, which illustrates the similarities between APT Group A and Group B, particularly highlighting shared MITRE ATT&CK technologies, locations, and target countries. Such patterns, as accentuated by data from FireEye's M-trends report [4], showcase a pressing need for in-depth, actionable intelligence that can inform and refine organizational defense strategies.

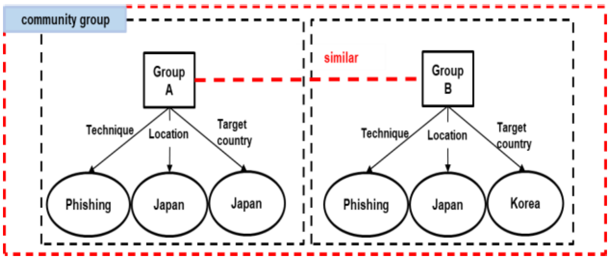

FIGURE 1. Sample illustration of similarities between APT Group A and Group B - highlighting MITRE ATT&CK techniques, locations, and target countries.

Studies such as, Wang et al. [5] made significant progress in analyzing APT behavior by incorporating key features such as TTPs, software, location, and target industries. However, their approach did not involve extensive feature engineering or employ similarity measures and deep learning techniques, which limited the depth and sophistication of their analysis. Similarly, [6] introduced the APT detection framework AULD (Advanced Persistent Threats Unsupervised Learning Detection), which applies a clustering algorithm to detect suspicious domains in APT attacks. While it effectively clusters malicious domains, it is limited by its narrow focus on domain (DNS)-based detection. Meanwhile, [7] proposed a hierarchical clustering framework to identify IoT-based APT attacks using IoT honeypots and TTP (Tactics, Techniques and Procedures) extraction. Reference [8] also focused on identifying attack groups by comparing the similarity of distributed domains, helping to detect recurring cyber-attacks from the same group. Although these approaches are effective within their specific contexts (DNS or IoT), they lack the integration of broader multi-dimensional analysis required for comprehensive APT clustering.

Recent works have attempted to address these limitations by incorporating additional features. For example, [9] presented a machine learning approach using hierarchical clustering to discover significant correlations between MITRE ATT&CK techniques, demonstrating that certain techniques can predict others with high accuracy. Li et al. [10] proposed an innovative method to automatically discover correlations between APT groups using rough set theory, achieving a high correlation precision. Their work focused on quantifying relationships between attack behavior patterns, offering an additional perspective on APT group relevance. However, these methods still lack a comprehensive, multi-dimensional view of APT behavior.

Our research builds on these efforts by integrating weighted feature importance across multiple dimensions, including MITRE ATT&CK techniques, software, target industries, and geographic locations. This allows for a more nuanced understanding of APT group behavior, offering improved clustering accuracy and actionable insights for defense strategies. We also extend previous efforts like [11], which utilized multimodal feature fusion and heterogeneous graph attention networks to capture deeper relationships between IOCs and APT groups. Our approach enhances this by incorporating decision trees and deep feature extraction, refining feature importance and improving clustering accuracy.

APTs are defined by their deliberate sequence of malicious activities—Tactics, Techniques, and Procedures (TTPs)—each chosen to advance the adversary's strategic goals. Security professionals must decipher these granular details to counter the threat effectively. However, much of this essential information is buried within unstructured CTI reports, complicating extraction and analysis. Advances in Natural Language Processing (NLP) and Named Entity Recognition (NER) [12], [13] have enabled the categorization of APT groups by their distinct characteristics, revealing potential inter-group dynamics.

While behavioral analysis in sandboxes [14], [15] and binary analysis [16], [17] offer ways to match malicious samples used by attackers to known or novel APT families, it is insufficient for identifying the groups behind attacks, given the many-to-many relationships between APT groups and techniques. Data-driven approaches have demonstrated strong performance in network security, as they allow for the discovery of broader patterns and relationships within large datasets. Our research builds on these strengths to develop a more accurate clustering of APT groups, considering multiple dimensions of their behavior.

Despite technological advancements, the industry continues to face significant challenges. The nuanced characteristics and covert interrelationships between APT groups are frequently missing from CTI reports, leaving gaps in collective cybersecurity knowledge. For instance, the ''Fox Kitten Campaign'' revealed a hidden alliance between Iran-linked APT groups APT34/OilRig and APT33/Elfin [18]. Identifying such strategic collaborations is crucial for crafting anticipatory defense mechanisms.

While prior studies have expanded our understanding of APT behavior, they often concentrated solely on data-level analysis, overlooking the critical semantic connections that illuminate the strategies and origins of APTs. Our research addresses this gap by delving into both the explicit tactics and subtler semantic ties evident in CTI reports. This dual approach strives for a deeper understanding and more effective counteractions against these threats.

Our primary focus lies in dissecting APT activities, scrutinizing pivotal aspects such as MITRE ATT&CK techniques, software usage, and the geographical locations and industries targeted. This in-depth analysis aims to provide a holistic view of APT operations, underscoring their tactical movements, technological capabilities, and strategic intents. By employing a robust NLP model for NER tasks, our research accurately labels crucial features (MITRE ATT&CK techniques, software, locations, and industries) across a range of open-source CTI datasets. Feature selection and feature crosses further refine our analysis. Subsequently, our approach utilizes weighted similarity metrics, informed by decision trees, and further enhanced by deep feature extraction using Deep Neural Networks (DNNs). This method allows for a more nuanced and accurate analysis of APT group relationships. In summary, our study makes the following contributions to the field of CTI.

- We propose a novel AI-enhanced weighted similarity metric combining Feature Crosses, Feature Selection, and weighted cosine similarity metrics using decision trees and DNNs. This method addresses existing literature gaps by capturing explicit relationships (e.g., common MITRE ATT&CK techniques and software) and subtle relationships (e.g., patterns in geographical and industry targets) among APT groups. It significantly improves clustering accuracy over traditional methods that rely on shallow feature extraction or simplistic techniques.
- We created a comprehensive dataset from 709 CTI reports, annotated with specific APT groups and their associated features. In total, 35 distinct APT groups and their associated features were identified. This dataset was split into 23 labeled APT groups, which were clustered to train and validate our clustering model, and 12 unlabeled APT groups, which were used to uncover hidden patterns and enhance the model's robustness and depth of our findings.
- We provide in-depth insights into APT group operations by analyzing their behavior by considering features such as MITRE ATT&CK Techniques, software, geographic, and industrial targets based on the CTI reports. This detailed focus reveals critical patterns and trends in APT activities, offering actionable intelligence for better defense strategies.

The paper is structured as follows: Section II outlines the research methods, Section III reviews related literature, and Section IV defines the problem statement. Methodology and execution are detailed in Section V, followed by results and evaluation in Section VI. Discussions are presented in Section VII, and the Conclusion in Section VIII summarizes key findings and contributions.

II. BACKGROUND

The background section outlines the essential methodologies for clustering the APT groups, including data extraction and preparation, sophisticated feature engineering, and advanced clustering and similarity measurement techniques.

A. MITRE ATT&CK FRAMEWORK

CTI reports are critical for detailing the activities of threat actors, especially APT groups. These reports offer insights into the groups' goals, the industries and regions they target, and the methods they employ in their attacks [19]. However, the inherently unstructured format of CTI reports presents challenges for extracting clear and actionable insights.

To navigate these complexities, the MITRE ATT&CK framework [20] serves as a cornerstone resource, offering a structured compilation of known adversary behaviors. The work [21] statistically analyzes the MITRE ATT&CK dataset to enhance security strategies for enterprises, ICS, and mobile infrastructures, offering a structured analysis from threat profiles to techniques, and providing actionable insights for future cybersecurity research.

Leveraging the framework, we have advanced our capability to distill pertinent details from CTI reports, categorizing the diverse techniques and software deployed by APTs. By aligning the insights gained from MITRE ATT&CK with the intelligence extracted from CTI reports, our analysis has become more nuanced and comprehensive.

B. NAMED ENTITY RECOGNITION

NER is a subtask of NLP that focuses on identifying and classifying named entities, such as industries, locations, and software, within unstructured text data. In our approach, we leveraged NER to automatically identify and extract these important entities from the CTI reports based on a model combining BERT [22] (Bidirectional Encoder Representations from Transformers) with a Conditional Random Field (CRF) layer. This model, known as BERT-CRF, harnesses BERT's powerful contextual embeddings with the sequence modeling capabilities of a CRF layer to improve the accuracy of entity classification.

Given the sentence from the CTI report, ''BRONZE BUTLER is a cyberespionage group with headquarters in Japan,'' the BERT-CRF model would operate as follows: BERT first processes the entire sentence to understand the context around each word. It identifies "BRONZE BUTLER" as a probable named entity of the APT group and "Japan" as a geographic location. The CRF layer then uses this information, along with the learned transitions between entity labels from the training data, to predict the most likely sequence of labels for the entire sentence.

However, we encountered a challenge when dealing with technique IDs defined by MITRE ATT&CK. These IDs were typically presented in tabular form in some of the reports. To address this, we manually supplemented the technique ID information into our dataset to ensure its completeness and accuracy.

C. FEATURE ENGINEERING METHODS

Effective clustering of APT groups necessitates a meticulous selection and engineering of features that capture the essence of their behavior and impact. In our study, we focus on a curated set of features that are pivotal for understanding and distinguishing between APT groups: MITRE ATT&CK techniques, software utilized, industries targeted, and geographic locations of the attacks.

1) FEATURE SELECTION

Before we perform the clustering of APT groups, it is imperative to identify the most informative features within our dataset. Feature selection is a crucial step in our machine learning pipeline, designed to manage the complexities inherent in the high-dimensional data associated with APT groups.

Our methodology employs the filter method [23] to identify a subset of significant features. This process is articulated as

$$Score\left(F_i\right) = f\left(F_i, Y\right), \tag{1}$$

where $Score\left(F_i\right)$ quantifies the importance of each feature F_i concerning the target variable Y. In our study, Y represents the cluster of APT groups based on their behaviors and characteristics. The filter method relies on statistical measures such as correlation coefficients, mutual information, and chi-squared tests to evaluate the relevance of each feature to Y. High-scoring features are indicative of greater relevance to the clustering outcome and are thus selected for further analysis.

2) FEATURE CROSSES

With the relevant features identified, we utilize Feature Crosses to create new combined features that can model the interactions and relationships between the original features [24]. This technique is vital for capturing complex, non-linear relationships that may exist in the interactions between different APT group behaviors. For example, if we consider the interaction between any two features deployed by APT groups, a new feature might be represented as

$$F_{new} = f\left(F_i, F_j\right), \tag{2}$$

where F_{new} represents the newly formed feature through the interaction of F_i and F_j. Specifically, if we examine the interaction between technique and software features, this relationship can be represented as

$$F_{new} = F_{technique} * F_{software}. \tag{3}$$

By creating these crossed features, we are equipped to uncover hidden patterns and dependencies, enhancing our clustering framework's ability to discern the nuanced behaviors that define each APT group.

D. SIMILARITY MEASUREMENT AND CLUSTERING

In our analysis, we quantify the similarity between APT groups using the weighted cosine similarity metric, which is formulated as

$$Similarity = \frac{\sum_{i=1}^{n} w_i a_i b_i}{\sqrt{\sum_{i=1}^{n} \left(w_i a_i\right)^2} \sqrt{\sum_{i=1}^{n} \left(w_i b_i\right)^2}}. \tag{4}$$

In this expression, a_i and b_i are the feature vectors of the APT groups being compared, w_i is the weight assigned to each feature based on its importance derived from the feature selection method and n is the total number of features.

Our methodology employs both decision tree models and DNNs to derive these critical feature importance scores. The decision tree model, trained on a foundational dataset comprising labeled data points, uses measures such as entropy or Gini impurity to optimize its decision-making process. This results in the generation of feature importance scores that reflect each feature's contribution to the model's predictive accuracy.

Parallelly, DNNs offer a complementary perspective by capturing complex, non-linear relationships between features, further enriching the feature importance assessment. This dual approach enables a comprehensive evaluation of feature significance, which is instrumental in calculating our weighted similarity measure. To refine the weighted similarity scores further, we incorporate the models' accuracy into weighting process and apply normalization techniques to address data imbalances, ensuring equitable representation across all classes. Building upon these refined similarity measures, we employ hierarchical clustering [25] to categorize APT groups.

Despite its computational intensity, hierarchical clustering is invaluable for revealing intricate patterns and relationships that might otherwise remain obscured. By combining weighted cosine similarity with hierarchical clustering, our framework effectively identifies significant structures within the data, facilitating a comprehensive exploration of APT group behaviors and interactions.

III. RELATED WORK

Several studies have contributed significantly to cluster analysis and APT group clustering, utilizing different features and methodologies. The key contributions are compared

TABLE 1. Comparison of various related work.

Category	Paper	Features Used	Similarity Measures	Clustering Algorithm
Behavioral and IoC-Based Clustering	[26]	TTP	Phi coefficient	Hierarchical
	[27]	Indicators of compromise (IoC)	N/A	Spectral
	[28]	IoC	Cosine, Jaccard	Hierarchical, Spectral, DB-scan
Graph and Similarity-Based Clustering	[10]	Attack behavior patterns, TTP	Rough set theory	Rough set theory
	[11]	IOCs, TTPs, Network data	N/A	Heterogeneous Graph Attention Networks
Feature-Specific Clustering	[7]	IoT-based TTP, Honeypots	N/A	Hierarchical
	[6]	DNS logs, Suspicious domains	N/A	K-means
	[5]	TTP, Software, Industry, Country	N/A	Girvan Newman algorithm
Multi-Feature and Multi-Dimensional Clustering	[9]	TTP	N/A	Hierarchical
	This work	TTP, Software, Industry, Country	Weighted Cosine	Hierarchical

in Table 1, focusing on the features used, similarity measures applied, and the clustering algorithms employed.

Ding et al. [26] constructed a knowledge graph to uncover hidden correlations between APT groups and their techniques using hierarchical clustering. While this approach provides insights into TTP relationships, it lacks the integration of broader features like industry and geography, which are critical for understanding APT group behavior in different contexts. In contrast, Fu et al. [27] analyzed attack events using IoC features with spectral clustering, demonstrating the potential of honeypot data to enhance defense mechanisms. Although useful for small-scale attacks, this work does not consider a multi-dimensional approach that includes strategic features such as software and industry targeting. Similarly, Faridi et al. [28] explored malware clustering using behavioral attributes and cosine and Jaccard similarity, but their focus on malware types rather than APT group behaviors limits its applicability to broader APT clustering tasks.

Reference [10] tackled APT clustering through rough set theory to automatically discover correlations between APT groups, offering high precision but lacking a multi-dimensional integration of APT behaviors. Xiao et al. [11] introduced a more advanced technique by utilizing multimodal feature fusion and heterogeneous graph attention networks to capture relationships between IOCs and APT groups. This method provides a deeper understanding of APT behaviors by combining semantic and structural features. However, [11] is heavily focused on graph-based analysis, and while it captures intricate relationships between technical indicators, it does not incorporate broader strategic features, such as targeted industries or geographic locations, limiting its strategic applicability.

Wang et al. [5] took a multi-dimensional approach by clustering APT groups using TTPs, software, industry, and country. However, this work lacks similarity measures and deep learning techniques, reducing its ability to capture complex relationships between features. Al-Shaer et al. [9] improved on this by incorporating cosine similarity and hierarchical clustering to find correlations between MITRE ATT&CK techniques, demonstrating high predictive power. However, [9] still lacks broader feature integration, limiting its ability to handle more complex APT behaviors. In contrast, [7] proposed Group Tracer, using hierarchical clustering to identify IoT-based APT attacks with IoT honeypots and TTP extraction. While effective for IoT-specific attacks, it lacks generalization across industries or locations. Similarly, [6] introduced an unsupervised learning model to cluster the suspicious domains via DNS logs. However, its narrow focus on DNS limits its broader applicability to APT groups, which often use multiple attack vectors.

Our research builds on and extends previous studies by incorporating multi-dimensional feature integration, addressing gaps in earlier works that primarily focused on individual features like TTPs, IoT data, or DNS logs. In contrast to these feature-specific approaches, our work integrates features such as TTPs, software, industries, and geographic locations, providing a comprehensive view of APT group behavior. By emphasizing feature importance and appropriate weighting, we overcome biases in technique usage across approximately 190 techniques and 130 APT groups, as illustrated in Fig. 2 from our dataset. This integration enables us to refine clustering accuracy, offering deeper insights into the strategic behavior of APT groups for more effective threat analysis and defense strategies.

Our approach also enhances traditional methods by incorporating a weighted similarity measurement algorithm, positioning our research as a significant advancement in CTI analysis.

IV. PROBLEM FORMULATION

Our goal is to develop a weighted similarity measure for clustering 'n' APT groups using four features: MITRE ATT&CK Techniques, software, target countries, and industries, aiming to create 'm' clusters. We assess our clustering approach with two metrics:

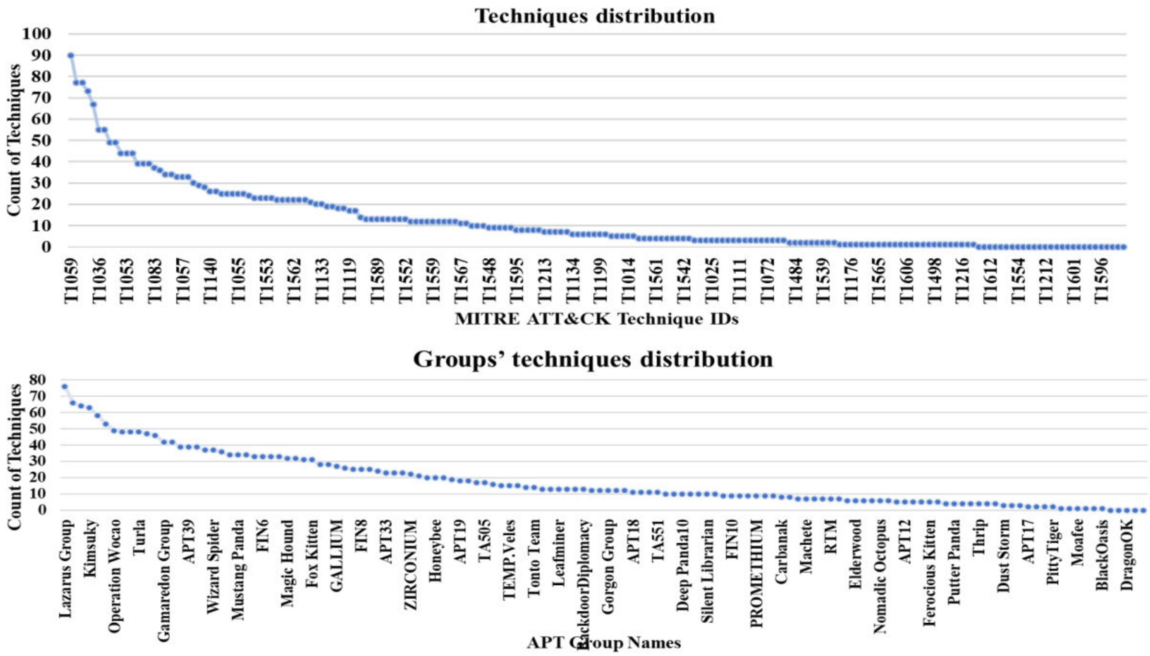

FIGURE 2. Distribution of techniques used by APT groups.

1) The Silhouette Coefficient, which measures how well data points fit within their cluster compared to other clusters, is suitable for unsupervised learning scenarios.
2) The Adjusted Rand Index (ARI), which quantifies the similarity between the clustering outcome and the ground truth labels, is useful for evaluating clustering with known labels.

Due to limited labeled data, we use a mix of decision tree-derived training data and additional labeled data not used in training to calculate ARI, providing a robust validation of our clustering effectiveness. This method, complemented by an evaluation of a separate labeled dataset, enables a comprehensive assessment of our weighted similarity metric and clustering strategy's accuracy and generalizability.

V. METHODOLOGY FOR CLUSTERING APT GROUPS

In this section, we present our framework for clustering APT groups. The overall system architecture is depicted in Fig. 3. The architecture has four fundamental components: Feature Extraction, Data Aggregation, Ground Truth Generation, and Clustering Method. This approach ensures a comprehensive analysis by integrating these critical aspects to understand and classify APT group behaviors accurately.

A. FEATURE EXTRACTION

During the Feature Extraction phase, our primary goal is to meticulously identify and extract relevant features that accurately represent the behaviors and characteristics of APT groups from a comprehensive dataset of over 700 CTI reports. To achieve this, we integrate NER with the advanced capabilities of the BERT-NER model. This model, adapted from the open-source implementation [22], is further augmented with a CRF layer to enhance its precision in identifying and classifying named entities related to APT groups within the text.

The process of feature extraction through the BERT-NER model can be described as follows:

1) Data Preprocessing: Each CTI report is tokenized into words or sub-words. These tokens serve as input for the BERT-NER model.
2) BERT-NER Model with CRF Layer: The tokenized data is passed through the BERT-NER model, where the BERT component generates contextual embeddings for each token. The CRF layer then utilizes these embeddings to predict the most probable sequence of labels (e.g., technique ID, software used, target industry) for the tokens, considering the context and dependencies between labels.

   During the training phase, the model's parameters are optimized by minimizing the objective function, defined as

$$L(\theta) = -\Sigma_{i=1}^{N} \log P(y_i \mid x_i; \theta) + \frac{\lambda}{2} ||\theta||^2, \tag{5}$$

   where, x_i represents the input token sequence, y_i represents the corresponding sequence of labels, θ is the parameters of the model and λ is the regularization parameter.
3) Training Parameters: To fine-tune our BERT-NER model, we adopt a batch size of 8 and train the model for 8 epochs. This setup is chosen to balance between model performance and computational

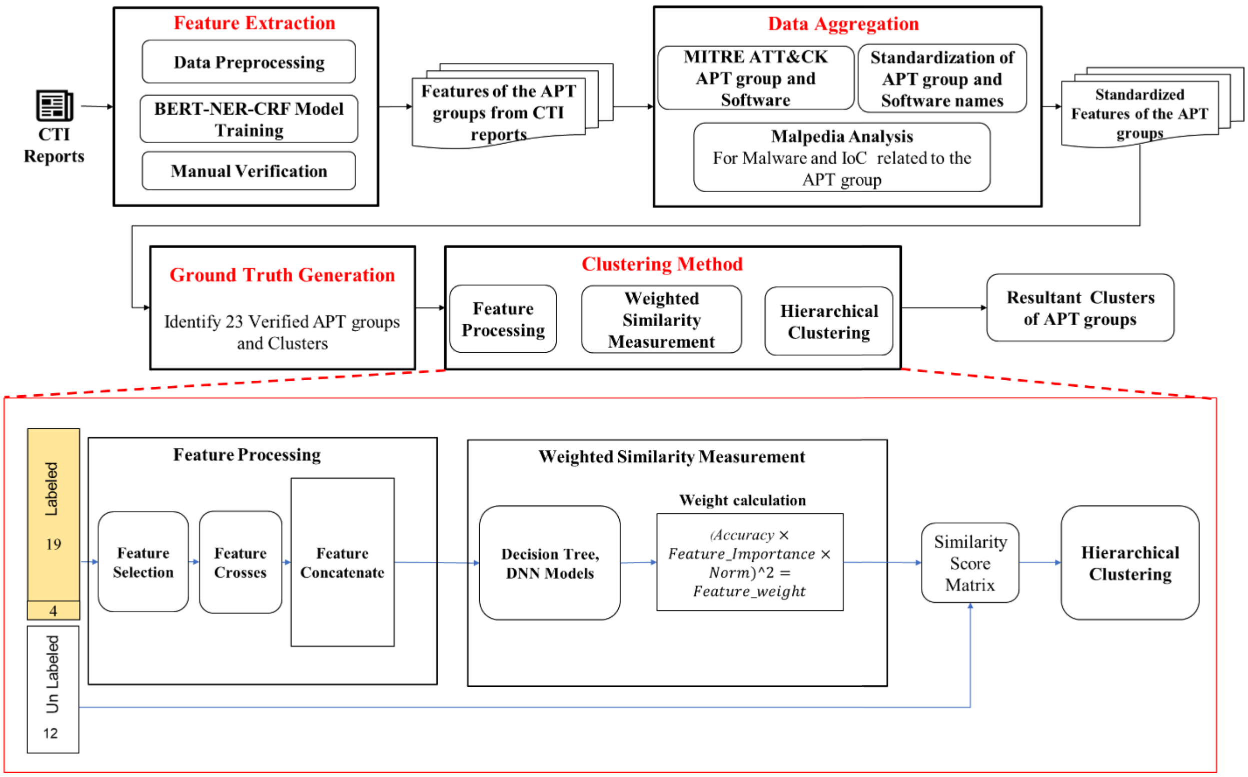
CTI Reports
Feature Extraction
Data Preprocessing
BERT-NER-CRF Model Training
Manual Verification
Features of the APT groups from CTI reports
Data Aggregation
MITRE ATT&CK APT group and Software
Standardization of APT group and Software names
Malpedia Analysis For Malware and IoC related to the APT group
Standardized Features of the APT groups
Ground Truth Generation
Identify 23 Verified APT groups and Clusters
Clustering Method
Feature Processing
Weighted Similarity Measurement
Hierarchical Clustering
Resultant Clusters of APT groups
Labeled
19
4
Un Labeled
12
Feature Processing
Feature Selection
Feature Crosses
Feature Concatenate
Weighted Similarity Measurement
Weight calculation
Decision Tree, DNN Models
(Accuracy × Feature_Importance × Norm)^2 = Feature_weight
Similarity Score Matrix
Hierarchical Clustering

FIGURE 3. System architecture for clustering APT groups.

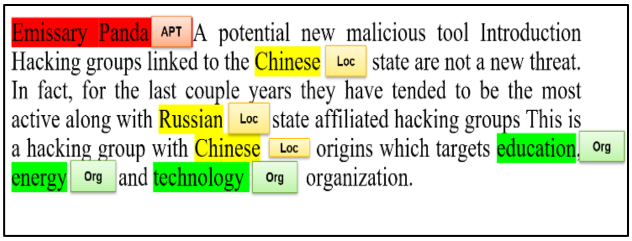
Emissary Panda APT A potential new malicious tool Introduction Hacking groups linked to the Chinese Loc state are not a new threat. In fact, for the last couple years they have tended to be the most active along with Russian Loc state affiliated hacking groups This is a hacking group with Chinese Loc origins which targets education. Org energy Org and technology Org organization.

FIGURE 4. Example of a CTI report processed by the BERT-NER model. [LOC- location, APT- group name, Org – Sector].

efficiency. An example of a CTI report processed by the BERT-NER model is shown in Fig. 4.

4) Manual Verification: Given the variability in how CTI reports present information, techniques detailed in tables require manual checking to complement the automated NER process. The parameters used to manually check the MITRE Techniques include the technique ID, description, and examples provided by the MITRE database. These details help ensure accurate identification and contextualization of each technique. Also, the validation of the manual check on MITRE Techniques is performed through cross-verification with existing CTI reports with labels and the MITRE ATT&CK framework to ensure accuracy and consistency.

By leveraging the BERT-NER model enhanced with a CRF layer, we significantly reduce the time and manual effort required for feature extraction from CTI reports. Thus, the combination of automated and manual methods ensures a thorough and accurate representation of APT group characteristics, laying a solid foundation for further analysis in the subsequent phases of our methodology.

B. DATA AGGREGATION AND GROUND TRUTH GENERATION

In the process of data aggregation for our research, we meticulously synthesized information from a wide array of sources to construct a comprehensive dataset on APT groups. This involved incorporating data from the MITRE ATT&CK framework, which served as a foundational resource due to its detailed compilation of cyber threat tactics and techniques. Additionally, we curated over 700 CTI reports from sources including [29] and [30], and our own collections, providing a broad perspective on APT operations and strategies. A significant challenge in cybersecurity research is the presence of multiple aliases for a single software or APT group, as reported by different cybersecurity firms. To tackle this, we standardized the names based on MITRE ATT&CK's "Associated Groups", ensuring consistency across our dataset. For example, the APT group known variably as "Ajax Security Team," "Operation Woolen-Goldfish," "AjaxTM," "Rocket Kitten," "Flying Kitten," and "Operation Saffron Rose" was uniformly referred to as "Ajax Security Team" in our dataset. Similarly, the software "ZxShell," also known as "Sensocode," was consistently

TABLE 2. Clustering of APT groups based on shared characteristics and documented instances in CTI reports.

Cluster	APT Groups	Instances in CTI Reports
Cluster_1	APT41, Winnti Group, Earth Lusca, Emissary Panda, Ke3chang, Axiom, APT17, APT19, and Deep Panda	• Axiom, APT17, and Ke3chang are closely linked to Winnti Group. • The attack by Emissary Panda is linked to the DRBControl campaign in early 2020 to Chinese APT groups APT27 and Winnti. • Earth Lusca has used malware commonly used by other Chinese threat groups, including APT41 and the Winnti Group cluster.
Cluster_2	APT37, Lazarus Group, Andariel, APT38 and Kimsuky	• Some security researchers report all North Korean state-sponsored cyber activity under the name Lazarus Group instead of tracking clusters or subgroups, such as Andariel, APT37, APT38, and Kimsuky.
Cluster_3	Dragonfly, ALLANITE, and Gamaredon Group	• ALLANITE technical operations combined with activity with a group Symantec calls Dragonfly.
Cluster_4	APT28 and Sandworm Team	• Some of APT28 attacks were conducted with the assistance of GRU Unit 74455, which is also referred to as Sandworm Team.
Cluster_5	Carbanak and FIN7	• Carbanak may be linked to groups tracked separately as Cobalt Group and FIN7 that have also used Carbanak malware.
Cluster_6	APT33 and APT40	• Iran-linked APT34/OilRig and APT33/Elfin have cooperated in the ”Fox Kitten Campaign”.

labeled as “ZxShell.” This approach not only streamlined our dataset but also mitigated the potential confusion arising from the diverse nomenclature used across different cybersecurity reports.

We also consulted Malpedia [31], an esteemed online repository of malware, to enrich our dataset with detailed information on malware families, attack techniques, and Indicators of Compromise (IoCs). This extensive compilation of data allowed us to create a standardized taxonomy for APT group aliases and their associated software, enhancing the dataset’s coherence and utility for analysis.

Our integrated dataset featured 35 distinct APT groups, from which we identified 23 that met our strict criteria for classification as “ground truth”. These criteria included consistent reporting across multiple sources and the presence of verifiable IoCs. The remaining 12 groups were excluded due to insufficient data to meet these standards. Table 2 showcases the clusters, associated APT groups, and instances in CTI reports, validating our clustering approach and illustrating the effectiveness of our method in categorizing APT groups. This table acts as a pivotal element in generating the ground truth of our clustering process.

C. CLUSTERING METHOD

Our clustering methodology is meticulously designed to address the complexities of cyber threat analysis by harnessing a combination of ML models such as decision tree modeling and DNN for enhanced feature importance assessment and similarity measurement. Visualized in Fig. 3, the process unfolds in three stages: Feature Processing, Weighted Similarity Measurement, and Hierarchical Clustering. This part becomes the key contribution of our solution architecture.

1) FEATURE PROCESSING

We initiate our clustering methodology with the collection of 19 labeled data points (from a subset of 23 APT groups identified as “ground truth”), which forms the foundation of our decision tree model. This model is pivotal in uncovering the relationships between different features (such as MITRE ATT&CK techniques, software used, and targeted industries) and their corresponding labels representing unique APT groups.

For the feature selection, the decision tree algorithm uses entropy and Gini impurity measures to calculate the importance of each feature [32] in predicting APT group behaviors. These importance scores are crucial for understanding which features contribute most significantly to distinguishing between APT groups. After identifying the crucial features, we employ Feature Crosses to explore interactions between attributes. For instance, if a decision tree identifies both a specific MITRE ATT&CK technique (e.g., Tech_T1027) and a piece of software (e.g., Soft_ZxShell) as significant, a Feature Cross might combine these into a single feature to capture their interplay. This step is vital for uncovering the complex, non-linear relationships among APT group behaviors.

After feature selection and crossing, we perform Feature Concatenation, merging selected and crossed features into a unified set. For example, if APT28 has been identified using technique T1027 and software ZxShell, and targets the finance industry in Russia, the concatenated feature vector could be [1, 1, 0, 1, 1] combining individual and crossed features. This process ensures a holistic representation of APT group attributes for AI model training, enriching analysis and improving clustering accuracy. Table 3 illustrates a sample of feature processing for two APT groups.

TABLE 3. Sample illustration of feature processing.

Features	APT 28	APT 33
Tech -T1027	1	0
Software -ZxShell	1	1
Industry - Finance	0	1
Location - Russia	1	0
Cross: Tech_Software	1	0
Concatenation	[1,1,0,1,1]	[0,1,1,0,0]

2) WEIGHTED SIMILARITY MEASUREMENT

Building upon the foundation laid in the Feature Processing phase, we advance to the Weighted Similarity Measurement segment. Here, our methodology incorporates both machine learning models—specifically, decision trees—and DNNs to refine the representation of features.

Decision trees help determine the importance of features, while DNNs extract complex patterns, producing a sophisticated set of features. The feature weights are calculated as

$$F_{weight} = \sqrt{\frac{\sum_{i=1}^{n} \left(Accuracy_i * F_{importance} * Norm_i\right)^2}{n}}, \quad (6)$$

where, F_{weight} represents the final weighted importance of each feature. n is the total number of features, $Accuracy_i$ is the accuracy of the prediction model for the i^{th} feature, $Norm_i$ is the normalization parameter applied to i^{th} feature to counteract data imbalance and ensure equitable representation across all classes. The resultant weighted features form the basis of our similarity measurement, facilitating a nuanced comparison between APT groups through weighted cosine similarity.

3) HIERARCHICAL CLUSTERING

The final step in our methodology is Hierarchical Clustering, where we employ the derived similarity scores to categorize APT groups into distinct clusters. This technique leverages dendrograms to visually represent the structure of the data and elucidate the hierarchical relationships within clusters. By calculating similarity scores based on the weighted features, we ensure that our clustering process is grounded in a robust understanding of APT group behaviors.

While hierarchical clustering builds a dendrogram to illustrate how data points are nested within clusters, we determine where to cut the dendrogram to form meaningful clusters based on a similarity threshold. For our clustering tasks, we set a threshold at 0.8 for the weighted similarity measurement. Values greater than this threshold indicate strong similarity, ensuring that only highly similar APT group behaviors and attributes are clustered together. This approach enhances the robustness and accuracy of our clustering methodology. This hierarchical clustering not only reveals potential affiliations and subgroupings among APT entities but also offers a comprehensive view of the cyber threat landscape.

VI. EVALUATION

A. DATASET DESCRIPTION

To evaluate the robustness of our clustering methodology, we leveraged a dataset encapsulated in Table 4. Our dataset comprised 709 CTI reports collected from various sources such as rcATT [29], GitHub [30], and self-collected reports from security firms including FireEye, Google, and Microsoft. These reports provide a rich source of data, covering 193 different MITRE ATT&CK techniques, 636 types of software, operations across 199 countries, and targeting 52 industries. Each CTI report was meticulously annotated to identify specific APT groups and their associated features using BERT-NER. This involved tagging reports with relevant MITRE ATT&CK techniques, software used, target nations, and industries. Thus, we created a comprehensive CSV file containing this information. These CSV files were further utilized for feature engineering and the clustering process. The selection of the strategic features—MITRE ATT&CK techniques, software used, target industries, and target countries—was based on their relevance to understanding the operational behavior of APT groups.

- Techniques: These provide insights into the tactics and methods employed by threat actors, helping to map out their operational strategies.
- Software: The types of software used by APT groups reflect their technical capabilities and preferred tools.
- Industries: Target industries highlight sector-specific risks and vulnerabilities that threat actors exploit.
- Locations: Understanding the geographical focus of attacks allows for the identification of regional threat trends and geopolitical motivations.

This strategic selection ensures that the clustering process is guided by features that directly impact threat analysis and mitigation strategies. Our dataset includes 35 distinct APT groups. Of these, 23 groups were labeled and assigned to specific clusters based on consistent reporting and verifiable IoCs. The remaining 12 groups were unlabeled due to insufficient data. This split allowed us to validate our clustering results effectively. The connection between APT groups and their features was established through a relational mapping

TABLE 4. Dataset overview.

Category	Data type	Quantity
CTI reports	Total reports analyzed	709
MITRE ATT&CK Framework	Distinct APT groups	35
	Distinct Techniques	193
	Distinct Software	636
	Target Countries	199
	Target Industries	52

in the CSV files, where each APT group is linked with its corresponding techniques, software, and targets.

B. EVALUATION METRICS

We used the following evaluation metrics to validate the accuracy of the clustering process.

1) Silhouette Coefficient: The Silhouette Coefficient, first proposed in the work [33], is a robust metric for evaluating clustering quality by measuring cohesion (similarity within clusters) and separation (difference from other clusters). This metric was calculated using the complete set of 35 APT groups to measure the cohesion and separation of data within clusters. The Silhouette Coefficient is essential for our study as it quantifies how well each APT group's behaviors and attributes are grouped versus separated from other groups.
   Generally, a Silhouette value greater than 0.7 indicates strong clustering performance, while values between 0.5 and 0.7 are considered moderate, and values below 0.5 suggest weak clustering. This validation confirms the effectiveness of our clustering methodology in accurately identifying patterns in APT activities.
2) ARI: This metric is frequently used in cluster validation as it measures the agreement between two partitions: one given by the clustering process and the other defined by external criteria [34]. For our task, the ARI is particularly valuable because it allows us to assess the accuracy of our clustering results against known classifications of APT groups. By comparing the clusters generated by our methodology with the ground truth, the ARI helps validate that our clustering process accurately reflects relationships among the APT groups.
   The ARI ranges from -1 to 1, where 1 indicates perfect agreement, 0 indicates random clustering and negative values indicate worse than random clustering. An ARI score greater than 0.65 is typically considered indicative of good clustering performance, providing strong validation for our clustering approach in identifying meaningful patterns in APT activities.
   In our case, ARI calculations were based on a selected group of 19 labeled APT groups (chosen from the 23 APT groups labeled as "ground truth" in Table 2). These labels served as a benchmark to quantify the alignment between our hierarchical clustering results and the existing knowledge base.
3) Observational Analysis: Four APT entities (the remaining four APT groups from the original set of 23 labeled as "ground truth") Earth Lusca, Emissary Panda, Andariel, and APT38 were scrutinized. These groups were not included in the training or testing phase, providing an unbiased ground to validate our clustering approach's generalization capability.

Our evaluations aimed to position the APT groups within their respective clusters accurately. In the case of Earth Lusca and Emissary Panda, a successful clustering result would place them within or adjacent to APT41 in Cluster 1. Similarly, for Andariel and APT38, a corresponding placement would be alongside the Lazarus Group in Cluster 2. The precision of this classification within the dendrogram hierarchy, mirroring the established threat group associations, signifies the effectiveness of our feature processing techniques and underpins the contribution of our approach to Cyber Threat Intelligence analysis.

C. EFFECT OF FEATURE SELECTION

The significance of feature selection in our clustering model is illustrated in Table 5, where 'X' represents the characteristics of APT groups, encompassing technology, industry, software, and location features for various APT groups. By varying 'X', we assessed the impact of different feature set sizes on the model's performance, focusing on the ARI and Silhouette Coefficient metrics.

TABLE 5. Performance evaluation of feature selection.

Feature type	Feature Selection	Silhouette Coefficient	ARI	Validation
Technique	1<X	0.55	0.20	None
	1<X<17	0.66	0.45	Emissary Panda, Earth Lusca
	1<X<20	0.57	0.33	Emissary Panda, Earth Lusca
Software	1<X	0.68	0.34	Earth Lusca, Andariel
	1<X<17	0.65	0.34	Earth Lusca, Andariel
	1<X<20	0.69	0.34	Earth Lusca, Andariel
Target Country	1<X	0.48	0.22	None
	1<X<17	0.65	0.33	None
	1<X<20	0.50	0.31	None
Industry	1<X<17	0.70	0.14	None
	1<X<17	0.72	0.22	None
	1<X<20	0.55	0.25	Earth Lusca

Our analysis prominently features four APT groups: APT37, Andariel, APT38 (Cluster 2), and APT33 (Cluster 6). We observed that certain techniques were frequently used across these groups, as previously highlighted through statistical analysis. By applying mutual information and correlation analysis, we effectively filtered out these commonly used techniques, leading to a notable enhancement in the model's performance. This optimization process resulted in more accurately defined clusters, as reflected in improved ARI scores.

Interestingly, while the software features demonstrated minimal overlap among the APT groups, their exclusion based on usage frequency had a marginal impact on the model's effectiveness. This observation suggests that software features, though not predominant, can play a role in clustering APT groups when shared between them. This is corroborated by our model's ability to accurately associate two of the scrutinized APT entities, highlighting the potential of shared software features in revealing underlying connections among APT groups.

D. EFFECT OF FEATURE CROSSES AND FEATURE CONCATENATION

This section delves into the intricate effects of feature crosses and feature concatenation on the clustering model's accuracy. Experimenting with feature crosses for all attributes (excluding ‘‘software’’) resulted in noticeable variations in performance metrics. The data in Table 6 showcases these differences, indicating a general improvement in silhouette coefficients and ARI with the application of feature crosses. Notably, a depth of 3 for ‘‘Technique’’ feature crosses led to a slight reduction in ARI, suggesting that the complexity introduced by higher dimensionality could potentially diminish the model's discriminative capability.

Subsequent evaluation on feature concatenation illustrated optimal clustering when ‘‘Technique’’ was paired with ‘‘Industry’’, as depicted in Table 7. Contrarily, incorporating the ‘‘Software’’ feature in concatenations seemed to lessen its impact on clustering efficacy. These experimental findings underscore the critical role of feature selection strategies in enhancing the model's performance.

TABLE 6. Performance evaluation of feature crosses.

Feature Type	Feature Crosses	Silhouette Coefficient	ARI	Validation
Technique	1	0.66	0.45	Emissary Panda, Earth Lusca
	2	0.54	0.57	Emissary Panda, Earth Lusca
	3	0.67	0.42	Earth Lusca
Target Country	1	0.48	0.22	None
	2	0.68	0.37	Earth Lusca
	3	0.79	0.38	Earth Lusca
Industry	1	0.72	0.22	None
	2	0.67	0.21	None
	3	0.73	0.24	Earth Lusca

TABLE 7. Performance evaluation of feature concatenation.

Feature Type	Silhouette Coefficient	ARI	Validation
Techniques	0.54	0.57	Emissary Panda, Earth Lusca
Target Country	0.79	0.38	Earth Lusca
Software	0.68	0.34	Earth Lusca, Andariel
Industry	0.73	0.24	Earth Lusca
Techniques, Target Country	0.57	0.64	Emissary Panda, Earth Lusca
Techniques, Industry	0.70	0.55	Emissary Panda, Earth Lusca
Techniques, Software	0.62	0.36	None
Techniques, Target Country, Industry	0.64	0.44	Earth Lusca

TABLE 8. Performance evaluation of various similarity measurement and clustering approaches.

Approaches	Silhouette Coefficient	ARI	Validation
Cosine Similarity	0.05	0.06	None
Cosine Similarity + Decision Tree-Accuracy × *Feature_Importance*	0.68	0.28	Earth Lusca
Cosine Similarity + XGBoost-Accuracy × *Feature_Importance*× *Norm*	0.59	0.51	Earth Lusca
Cosine Similarity + Random Forest-Accuracy × *Feature_Importance*× *Norm*	0.64	0.55	Emissary Panda, Earth Lusca
Cosine Similarity + *Decision Tree-Accuracy ×Feature_Importance× Norm*	0.70	0.55	Emissary Panda, Earth Lusca
Cosine Similarity + *Decision Tree+ DNN - Accuracy ×Feature_Importance× Norm*	0.76	0.63	Emissary Panda, Earth Lusca
K-Means Model	0.49	0.36	None
Meanshift Model	0.57	0.51	None

E. PERFORMANCE OF VARIOUS SIMILARITY MEASURES AND ML METHODS

Our evaluation highlights a diverse range of performances across different clustering methodologies, with machine learning models enhancing the basic cosine similarity approach especially when applied to the ‘Technique’ and ‘Industry’ features which demonstrated robust performance in our feature engineering phase. As observed in Table 8, cosine similarity alone yielded a low Silhouette Coefficient (0.05) and ARI (0.06), indicating a limited capacity for clustering APT groups effectively. This underscores the need for more sophisticated methods in high-dimensional data environments.

Incorporating ML into the similarity measurement significantly improved the outcomes. The integration of Decision Trees with cosine similarity boosted both the Silhouette Coefficient to 0.68 and ARI to 0.28, successfully identifying the Earth Lusca group. The subsequent application of XGBoost, paired with normalization, further enhanced the performance, achieving a Silhouette Coefficient of 0.59 and an ARI of 0.51, again validating the Earth Lusca group.

The introduction of Random Forest techniques, alongside cosine similarity and normalization, resulted in even more pronounced improvements, evidenced by a Silhouette

Coefficient of 0.64 and ARI of 0.55, accurately validating the Emissary Panda and Earth Lusca groups. However, the most significant advancement was observed with the integration of DNN. This combination achieved the highest Silhouette Coefficient (0.76) and ARI (0.63), effectively validating multiple APT groups, including Emissary Panda and Earth Lusca. This progression demonstrates the substantial impact of blending traditional similarity metrics with advanced machine learning and deep learning techniques.

In addition to our primary approach, we used K-means [35] and MeanShift [36] clustering algorithms as baselines to evaluate our features. MeanShift is a centroid-based algorithm that iteratively calculates the expected movement of the center point until convergence. K-means, on the other hand, divides the sample set into a predefined number of clusters, minimizing the within-cluster sum of squares. For our analysis, we set K to 6, reflecting the number of clusters identified in our primary method. The performance of these models is compared in the last two rows of Table 8, illustrating the superior clustering accuracy of our approach.

F. INSIGHTS DERIVED FROM THE CLUSTERING OF APT GROUPS

Our methods effectively revealed complex connections between APT groups. A high silhouette coefficient of 0.76 indicates strong cluster cohesion, reflecting precise and insightful clustering. Fig. 5 illustrates the percentage of techniques (Fig. 5.a) and targeted industrial organizations (Fig. 5.b) utilized by APT groups within three clusters, showcasing the distribution and prevalence of key features across the clusters. The subsequent explanations, however, delve into detailed observations, uncovering insights into commonalities and unique traits within and across these clusters.

In cluster 1, APT 41 and the Winnti Group demonstrate a notable similarity, with approximately 19% (10 sectors) overlapping in their targeted industry sectors. This is emblematic of shared strategic interests and potential collaborative behaviors. Moreover, common usage of specific software tools—namely ShadowPad, PlugX, and others—by approximately 6 groups within this cluster underscore shared operational tactics and toolsets. Within cluster 1, Deep Panda and APT 19 stand out for their high frequency of attacks on organizations, despite their limited presence. Deep Panda, targeting 11.8% of organizations, leads with the highest number of attacks, making up 43% within the cluster. APT 19, though present in only 7% of organizations, follows closely with a significant 17% of the attacks, underscoring its aggressive approach despite targeting fewer entities. This pattern is indicative of a focused approach and possible prioritization of certain organizations by these groups.

Cluster 2 is unified not just by common targets but also by methodology, with 11 specific techniques (constituting 6% of the total) and convergence on three specific industries (also 6%), delineating a cohesive operational profile within this group.

Specifically, APT37's expanded target range to include a diverse set of Western and Asian countries signifies a strategic shift in their operations, pointing towards a broader ambition and reach beyond their historically concentrated regional focus.

Cluster 3 is marked by a geographical concentration on the United States, with defense and electronics sectors being notably more targeted compared to other clusters. Within this cluster, Dragonfly emerges as a dominant entity, engaging with 50% of the organizations that are also targeted by other APT groups within the same cluster, highlighting its significant role and shared operational interests with other members.

In cluster 4, despite the absence of commonly used software, the APT groups share 59 techniques. This cluster is distinguished by its use of unique techniques, notably T1598, which is rarely employed by APT groups in other clusters. Similar to Cluster 1, APT groups in Cluster 4 also predominantly target the energy sector.

Clusters 5 and 6 are characterized by their distinct focus on the chemical industry, alongside government and finance sectors. These clusters exhibit fewer unique techniques compared to others, indicating potential specialized operational expertise or strategic goals.

Beyond the confines of individual clusters, our analysis has unveiled several key patterns common across all 35 APT groups:

- 'Technique' and 'Industry' features are pivotal in the clustering of APT groups. Predominantly targeted industries such as government, finance, and entertainment, which account for approximately 75% of the focus among APT groups (equivalent to 26 out of 35 groups), have emerged as primary targets. This suggests that these sectors are perceived as high-value and vulnerable by the threat actors.
- Among the 35 unique APT groups examined, techniques such as T1036 (Masquerading) and T1059 (Command and Scripting Interpreter) are extensively utilized, employed by over 88% of these groups (amounting to 31 APT groups). The ubiquity of these techniques highlights their importance in the arsenal of APT groups and the need for defenses to prioritize these attack vectors.

VII. DISCUSSION

Our research endeavors to push the boundaries of understanding and analyzing APT groups through sophisticated clustering techniques. By meticulously evaluating feature selection, crosses, and concatenation, alongside the deployment of various similarity measurement approaches enhanced by machine learning models, we've demonstrated marked improvements in clustering performance.

Our methodical approach, underscored by the dataset comprising 709 CTI reports detailing attributes of 35 APT groups, emphasizes the necessity of nuanced feature engineering in CTI analysis. The effect of feature selection highlighted in our evaluation, where specific techniques consistently

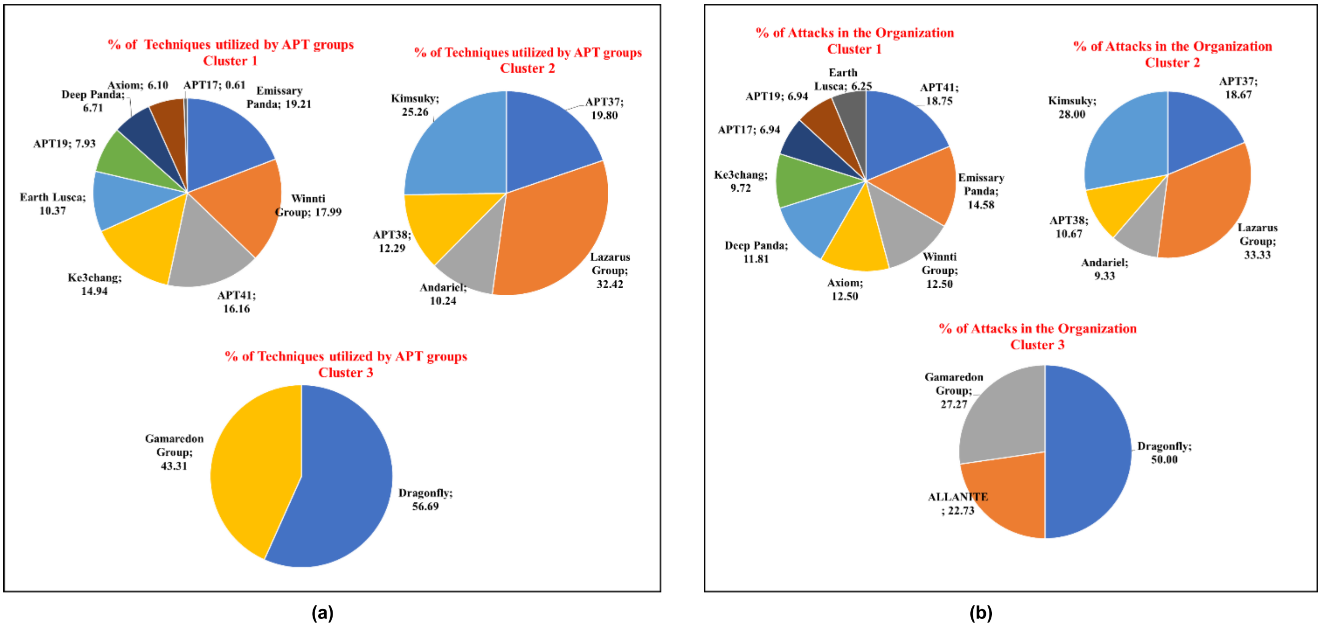

FIGURE 5. Distribution of techniques and industry targets across clusters. (a) Percentage of attack techniques utilized by APT groups in each cluster. (b) Percentage of industry sectors targeted by APT groups in each cluster.

deployed across groups were identified and filtered, aligns with the work of Ding et al. [26] and further validates the importance of strategic feature selection in revealing underlying threat dynamics.

Our clustering methodology, enhanced by machine learning models and similarity measures, marks a significant performance improvement. This is particularly evident when compared to works like Wang et al. [5], which lacked the application of similarity measures and did not engage in extensive feature engineering. While Wang et al. successfully clustered APT groups using multiple features (e.g., TTPs, software, industry), the absence of machine learning techniques and weighted feature importance limited their depth of analysis. Our approach addresses these gaps by incorporating feature weighting and similarity measurement algorithms, refining both the granularity and accuracy of APT group clustering.

We observed a significant performance leap facilitated by the integration of AI models with traditional similarity measures, underscoring the transformative potential of AI in cybersecurity analysis. For example, [9] focused on clustering MITRE ATT&CK techniques using hierarchical clustering with cosine similarity. However, our work extends beyond technique clustering, integrating additional features to provide a more comprehensive understanding of APT behaviors.

The clustering process also revealed key discriminators, such as prominent techniques used by APT groups and their new geographic targets, further illuminating APT group behaviors. For instance, we found that groups such as APT41 and the Winnti Group share technique T1047 and employ tools like Cobalt Strike, Mydoor, and Winnti for Linux, which hints at shared methods or operational goals. This aligns with the findings of Faridi et al. [28], who used behavioral attributes for malware clustering but focused primarily on malware types rather than the broader contextual features that our research incorporates.

Furthermore, studies like [6] and [7], which focused on specific environments like IoT-based attacks and DNS-based detection, respectively, were limited by their scope. Our research extends beyond these feature-specific approaches by employing multi-dimensional integration, allowing us to uncover subtle patterns and possible collaborations among APT entities. For example, the clustering of APT41 and Winnti Group suggests potential collaboration, as they share not only techniques but also similar software and tools. In conclusion, our research significantly advances the current state of APT group clustering by integrating advanced feature engineering, machine learning, and multi-dimensional feature integration.

VIII. CONCLUSION

In our study, we introduced an advanced approach to cluster Advanced Persistent Threat (APT) groups by examining their unique characteristics and affiliations. Utilizing Named Entity Recognition (NER), we pinpointed crucial APT attributes, such as software choices, target countries, and impacted industries. By amalgamating Feature Crosses, Feature Selection, and a weighted cosine similarity metric, we amplified the silhouette coefficient and Adjusted Rand Index (ARI) to 0.76 and 0.63 values respectively. These values not only surpass the industrial norms but also affirm the robustness of integrating technique and industry features in identifying APT clusters. The clear patterns that surfaced — such as the favored attack techniques and the changes in geographic targets — provide crucial insights into APT clusters. This nuanced understanding of APT group behaviors supports the development of targeted defense strategies, contributing to a more secure digital environment.

Looking forward, several potential avenues warrant exploration: enhancing feature extraction through sophisticated NER applications, curating larger labeled datasets, creating predictive frameworks to identify and classify emerging APT entities, and ensuring that our clustering methods scale seamlessly with the increasing influx of APT data.

Fundamentally, our research represents preliminary strides toward a more profound understanding of APT trends and predictions. By intertwining machine learning with iterative methodological enhancements, we aspire to expand our comprehension of the dynamic world of adversarial actors, setting the stage for a more robust cybersecurity landscape.

REFERENCES

[1] B. E. Strom, A. Applebaum, D. P. Miller, K. C. Nickels, A. G. Pennington, and C. B. Thomas, "MITRE ATT&CK: Design and philosophy," MITRE Corporation, McLean, VA, USA, Tech. Rep., MP180360R1, 2018.

[2] R. Brown and R. M. Lee, "The evolution of cyber threat intelligence (CTI): 2019 sans CTI survey," SANS Inst., USA, Tech. Rep. 38790, 2019. [Online] Available: https://www.sans.org/whitepapers/38790

[3] O. Cherqi, Y. Moukafih, M. Ghogho, and H. Benbrahim, "Enhancing cyber threat identification in open-source intelligence feeds through an improved semi-supervised generative adversarial learning approach with contrastive learning," *IEEE Access*, vol. 11, pp. 84440–84452, 2023.

[4] FireEye M-Trends. *Annual Reports*. Accessed: Mar. 20, 2022. [Online]. Available: https://www.fireeye.com/current-threats/annual-threatreport/mtrends.html

[5] W. Wang, B. Tang, C. Zhu, B. Liu, A. Li, and Z. Ding, "Clustering using a similarity measure approach based on semantic analysis of adversary behaviors," in *Proc. IEEE 5th Int. Conf. Data Sci. Cyberspace (DSC)*, Jul. 2020, pp. 1–7.

[6] G. Yan, Q. Li, D. Guo, and B. Li, "AULD: Large scale suspicious DNS activities detection via unsupervised learning in advanced persistent threats," *Sensors*, vol. 19, no. 14, p. 3180, Jul. 2019.

[7] Y. Wu, C. Huang, X. Zhang, and H. Zhou, "GroupTracer: Automatic attacker TTP profile extraction and group cluster in Internet of Things," *Secur. Commun. Netw.*, vol. 2020, pp. 1–14, Dec. 2020.

[8] H. Cho, S. Lee, B. Kim, Y. Shin, and T. Lee, "The study of prediction of same attack group by comparing similarity of domain," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2015, pp. 1220–1222.

[9] R. Al-Shaer, J. M. Spring, and E. Christou, "Learning the associations of MITRE ATT & CK adversarial techniques," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, Jun. 2020, pp. 1–9.

[10] J. Li, J. Liu, and R. Zhang, "Advanced persistent threat group correlation analysis via attack behavior patterns and rough sets," *Electronics*, vol. 13, no. 6, p. 1106, Mar. 2024.

[11] N. Xiao, B. Lang, T. Wang, and Y. Chen, "APT-MMF: An advanced persistent threat actor attribution method based on multimodal and multilevel feature fusion," *Comput. Secur.*, vol. 144, Sep. 2024, Art. no. 103960.

[12] R. Marinho and R. Holanda, "Automated emerging cyber threat identification and profiling based on natural language processing," *IEEE Access*, vol. 11, pp. 58915–58936, 2023.

[13] V. Yadav and S. Bethard, "A survey on recent advances in named entity recognition from deep learning models," 2019, *arXiv:1910.11470*.

[14] M. Alazab, "Profiling and classifying the behavior of malicious codes," *J. Syst. Softw.*, vol. 100, pp. 91–102, Feb. 2015.

[15] S. S. Hansen, T. M. T. Larsen, M. Stevanovic, and J. M. Pedersen, "An approach for detection and family classification of malware based on behavioral analysis," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, Kauai, HI, USA, Feb. 2016, pp. 1–5.

[16] M. Ahmadi, D. Ulyanov, S. Semenov, M. Trofimov, and G. Giacinto, "Novel feature extraction, selection and fusion for effective malware family classification," in *Proc. 6th ACM Conf. Data Appl. Secur. Privacy*, New Orleans, LA, USA, Mar. 2016, pp. 183–194.

[17] A. Makandar and A. Patrot, "Malware analysis and classification using artificial neural network," in *Proc. Int. Conf. Trends Autom., Commun. Comput. Technol. (I-TACT-15)*, Bangalore, India, Dec. 2015, pp. 1–6.

[18] (2020). *Fox Kitten—Widespread Iranian Espionage Offensive Campaign, Clear Sky Team*. Accessed: Oct. 2023. [Online]. Available: https://www.clearskysec.com/fox-kitten/

[19] N. Virvilis and D. Gritzalis, "The big four—What we did wrong in advanced persistent threat detection?" in *Proc. Int. Conf. Availability, Rel. Secur.*, Sep. 2013, pp. 248–254.

[20] MITRE. *Adversarial Tactics, Techniques and Common Knowledge*. Accessed: Oct. 2023. [Online]. Available: https://attack.mitre.org/

[21] B. Al-Sada, A. Sadighian, and G. Oligeri, "Analysis and characterization of cyber threats leveraging the MITRE ATT&CK database," *IEEE Access*, vol. 12, pp. 1217–1234, 2024.

[22] Kyzhouhzau. *'BERT-NER' GitHub Repository*. Accessed: Mar. 2023. [Online]. Available: https://github.com/kyzhouhzau/BERT-NER

[23] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, "Benchmark for filter methods for feature selection in high-dimensional classification data," *Comput. Statist. Data Anal.*, vol. 143, Mar. 2020, Art. no. 106839.

[24] *Feature Crosses*. Accessed: Jun. 2023. [Online]. Available: https://developers.google.com/machine-learning/crash-course/feature-crosses/video-lecture

[25] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: An overview," *WIREs Data Mining Knowl. Discovery*, vol. 2, no. 1, pp. 86–97, Jan. 2012.

[26] Z. Ding, D. Cao, L. Liu, D. Yu, H. Ma, and F. Wang, "A method for discovering hidden patterns of cybersecurity knowledge based on hierarchical clustering," in *Proc. IEEE 6th Int. Conf. Data Sci. Cyberspace (DSC)*, Oct. 2021, pp. 334–338.

[27] C. Fu, Y. Rui, and L. Wen-mao, "Internet of Things attack group identification model combined with spectral clustering," in *Proc. IEEE 21st Int. Conf. Commun. Technol. (ICCT)*, Oct. 2021, pp. 778–782.

[28] H. Faridi, S. Srinivasagopalan, and R. Verma, "Performance evaluation of features and clustering algorithms for malware," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2018, pp. 13–22.

[29] V. Legoy, M. Caselli, C. Seifert, and A. Peter, "Automated retrieval of ATT&CK tactics and techniques for cyber threat reports," 2020, *arXiv:2004.14322*.

[30] CyberMonitor. *Apt & Cybercriminals Campaign Collection*. Accessed: Mar. 2023. [Online]. Available: https://github.com/CyberMonitor/APT_CyberCriminal_Campagin_Collections

[31] Malpedia. *Library of Reports*. Accessed: Apr. 2023. [Online]. Available: https://malpedia.caad.fkie.fraunhofer.de/library/1/?search=Github

[32] M. Gupta. *How Feature Importance is Calculated in Decision Trees? With Example*. Medium. Accessed: Apr. 2023. [Online]. Available: https://medium.com/data-science-in-your-pocket/how-feature-importance-is-calculated-in-decision-trees-with-example-699dc13fc078

[33] P. J. Rousseeuw and L. Kaufman, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: Wiley, 2009.

[34] J. M. Santos and M. Embrechts, "On the use of the adjusted Rand index as a metric for evaluating supervised classification," in *Proc. Int. Conf. Artif. Neural Netw.* Berlin, Germany: Springer, 2009, pp. 175–184.

[35] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm," *Appl. Statist.*, vol. 28, no. 1, pp. 100–108, 1979.

[36] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 8, pp. 790–799, Aug. 1995.

ZHENG-SHAO CHEN received the master's degree in computer science and information engineering from National Central University, Taiwan, in 2023. His research interests include cyber threat intelligence analysis and natural language processing.

R. VAITHEESHWARI received the B.E. and M.E. degrees in electronics and communication engineering from Anna University, Chennai, Tamil Nadu, India, in 2017 and 2019, respectively. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Information Engineering, National Central University. Her research interests include the applications of artificial intelligence, natural language processing, and cybersecurity.

ERIC HSIAO-KUANG WU (Member, IEEE) received the B.S. degree in computer science and information engineering from National Taiwan University, in 1989, and the master's and Ph.D. degrees in computer science from the University of California at Los Angeles (UCLA), in 1993 and 1997, respectively. He is currently a Professor of computer science and information engineering with National Central University, Taiwan. His research interests include wireless networks, mobile computing, and broadband networks. He is a member of the Institute of Information and Computing Machinery (IICM).

YING-DAR LIN (Fellow, IEEE) received the Ph.D. degree in computer science from the University of California at Los Angeles (UCLA), in 1993. He was a Visiting Scholar with Cisco Systems, San Jose, from 2007 to 2008; the CEO with the Telecom Technology Center, Taiwan, from 2010 to 2011; and the Vice President of National Applied Research Labs (NARLabs), Taiwan, from 2017 to 2018. He was the Founder and the Director of the Network Bench Marking Laboratory (NBL), from 2002 to 2018, which reviewed network products with real traffic and automated tools and has been an approved test laboratory of the Open Networking Foundation (ONF). He co-founded L7 Networks Inc., in2002, later acquired by the D-Link Corporation, and O'Prueba Inc., as a pin-off from NBL, in 2018. He is currently a Chair Professor of computer science with National Yang-Ming Chiao Tung University (NYCU), Taiwan. He has published a textbook *Computer Networks: An Open Source Approach*, with Ren-Hung Hwang and Fred Baker (McGraw-Hill, 2011). His research interests include network security, wireless communications, network softwarization, and machine learning for communications. His work on multi-hop cellular was the first along this line and has been cited over 1000 times and standardized into IEEE 802.11s, IEEE 802.15.5, IEEE 802.16j, and 3GPP LTE-Advanced. He is an IEEE Distinguished Lecturer (2014–2017) and the ONF Research Associate (2014–2018). He received the 2017 Research Excellence Award and the K. T. Li Breakthrough Award. He has served or is serving on the editorial boards of several IEEE journals and magazines, including the Editor-in-Chief of IEEE COMMUNICATIONS SURVEYS AND TUTORIALS (COMST,1/2017-12/2020).

REN-HUNG HWANG (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of Massachusetts, Amherst, MA, USA, in 1993. He is currently the Dean of the College of Artificial Intelligence, National Yang Ming Chiao Tung University (NYCU), Taiwan. Before joining NYCU, he was with National Chung Cheng University, Taiwan, from 1993 to 2022. His research interests include deep learning, network security, wireless communications, the Internet of Things, and cloud and edge computing. He received the Best Paper Award from the Sixth International Conference on Internet of Vehicles 2019, IEEE Ubi-Media 2018, IEEE SC2 2017, IEEE IUCC 2014, and the IEEE Outstanding Paper Award from IEEE IC/ATC/ICA3PP 2012. He received the Outstanding Technical Achievement Award of the IEEE Tainan Section, in 2022. He served as the General Chair for the International Computer Symposium (ICS), in 2016; the International Symposium on Pervasive Systems, Algorithms, and Networks (ISPAN), in 2018; the International Symposium on Computer, Consumer and Control (IS3C), in 2018; and the 2019 IEEE DataCom (the Fifth IEEE International Conference on Big Data Intelligence and Computing). He is currently on the editorial boards of IEEE COMMUNICATIONS SURVEYS AND TUTORIALS and *IEICE Transactions on Communications*.

PO-CHING LIN (Member, IEEE) received the Ph.D. degree in computer science from National Chiao Tung University, Hsinchu, Taiwan, in 2008. He joined as a Faculty Member with the Department of Computer Science and Information Engineering, National Chung Cheng University (NCCU), in August 2009, where he is currently a Professor. His research interests include network security, network traffic analysis, and performance evaluation of network systems.

YUAN-CHENG LAI received the Ph.D. degree from the Department of Computer and Information Science, National Yang Ming Chiao Tung University, in 1997. He joined as a Faculty Member with the Department of Information Management, National Taiwan University of Science and Technology, in August 2001, and has been a Distinguished Professor, since June 2012. His research interests include performance analysis, wireless networks, network security, and machine learning.

ASAD ALI is currently a Senior Engineer with the National Institute of Cyber Security (NICS), Ministry of Digital Affairs (MoDA), Taiwan. His research interests include cyber threat intelligence, cellular networks, network security, and optimization.