SMART Model: A Robust Approach for Cyber Criminal Identification using Smartphone Data

K. Swetha
Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research (BIHER), Chennai, Tamilnadu, India
swetha281189@gmail.com (corresponding author)

K. Sivaraman
Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research (BIHER), Chennai, Tamilnadu, India
sivaraman2006@gmail.com

Received: 20 June 2024 | Revised: 16 July 2024 and 31 July 2024 | Accepted: 18 August 2024

ABSTRACT

The SMART (Smartphone Metadata Analysis for Recognizing Threats) model is a novel approach to the identification of prospective cyber criminals by analyzing smartphone data, with a particular emphasis on social media interactions, messages, and call logs. The SMART model, in contrast to conventional methods that depend on a wide variety of features, prioritizes critical parameters to ensure more precise and effective analysis. This model exhibits exceptional adaptability and robustness in a variety of data environments by employing sophisticated feature extraction and classification algorithms. This targeted approach not only improves the precision of threat identification but also offers a practicable solution for real-world cybersecurity applications, where data quality and consistency may vary.

Keywords-smartphone data analysis; SMART model; smartphone applications; cyber scams; cyber attacks

I. INTRODUCTION

Cyberattacks are a huge threat to individuals, companies, and society as a whole, as they can harm key infrastructure, steal sensitive information, and undermine trust in digital systems. These attacks, which range from malware infections to sophisticated campaigns, exploit human and technological vulnerabilities, posing difficult detection and mitigation hurdles. As the frequency and sophistication of cyberattacks grow, effective detection mechanisms become critical to protecting digital assets and ensuring the resilience of cyber-physical systems. Cyber forensics is emerging as a vital field in the fight against cyber threats, providing approaches and tools for detecting, evaluating, and attributing harmful behavior in digital settings. Cyber forensics is based on forensic science concepts and seeks to collect, preserve, and analyze digital evidence to recreate the events that led to a cyberattack, identify culprits, and support legal processes. By applying forensic techniques to digital artifacts, including log files, network traffic, and memory dumps, investigators can gain vital insights into attackers' strategies, techniques, and procedures. Collecting and storing digital evidence is the foundation of cyberforensic investigations. Following a forensic investigation, digital data are examined to recreate the attack timeline, identify attack pathways, and determine the scope of the compromise. Advanced forensic tools play an important role in automating analysis processes.

Despite advances in cyberforensic approaches and tools, detecting and attributing cyberattacks remains difficult due to the changing nature of threats and the complexity of digital environments. Attackers are continually changing their tactics to avoid detection, using techniques such as encryption, obfuscation, and antiforensic measures to hide their actions. Additionally, the volume and variety of digital evidence collected during an investigation can overwhelm forensic investigators, making it difficult to distinguish meaningful signals from noise. To address these issues, a multidisciplinary strategy is required, combining technical skills, domain knowledge, and cross-organizational collaboration. This study proposes the SMART model (Smartphone Metadata Analysis for Recognizing Threats), which employs an advanced method to identify cyber-criminals. It achieves this by analyzing crucial smartphone data such as call logs, messages, and social media interactions. The SMART model focuses on the most essential elements, improving accuracy and efficiency, unlike conventional techniques that depend on a wide range of features. This model provides exceptional performance and resilience by utilizing sophisticated feature extraction and classification methods. In addition, it smoothly adapts to different data settings. This approach not only enhances the precision of identifying cyber threats but also offers a practical and reliable solution for real-world applications, ensuring improved protection and durability in digital systems.

II. BACKGROUND

In [1], a real-time cyberattack detection system was introduced, which used machine learning techniques to analyze network traffic patterns and system logs. In [2], malware analysis was performed using smartphone data and forensic examination of ransomware attacks, addressing obstacles and suggesting strategies for the successful identification of cyberattacks using smartphone data. In [3], an intrusion detection system was introduced that used anomaly-based techniques in conjunction with forensic analysis and machine learning algorithms to detect threats in real time. In [4], the combination of Cyber Threat Intelligence (CTI) with forensic analysis methods was explored to enhance the ability to detect and respond to attacks. In [5], behavior analysis approaches were used to discover unusual patterns that indicate attacks. The study in [6] explored how utilizing log analysis and correlation approaches could improve cyberattack detection, helping organizations better recognize advanced threats. The study in [7] examined how digital forensics techniques can be used to respond to cloud security incidents, focusing on overcoming obstacles and suggesting optimal strategies for thorough investigation and reduction of risks. In [8], the focus was on methods to analyze security incidents related to IoT devices. The study in [9] focused on the notion of cyber threat hunting and its use to proactively identify and address cyber risks using forensic analysis techniques. In [10], methods for identifying Advanced Persistent Threats (APTs) were explored using digital forensics approaches, highlighting the importance of proactive threat hunting and incident response tactics.

A. Limitations and Research Gaps from Existing Models

Existing models for cyberattack detection and response exhibit several gaps [11, 12]. Real-time detection systems often lack comprehensive forensic analysis integration, impeding immediate threat response [13]. Mobile and IoT device security frameworks are not yet robust or standardized, leaving these increasingly targeted devices vulnerable. Behavioral analysis and anomaly detection methods generate high false-positive rates and require significant computational resources, necessitating improvements in accuracy and efficiency [14, 15]. The practical implementation of merging CTI with forensic analysis is still in its early stages, requiring seamless integration with current systems [16]. APT detection remains predominantly reactive rather than proactive, calling for early detection and mitigation strategies [17]. Log analysis and correlation techniques struggle with the complexity and volume of data, demanding enhanced tools for better threat detection [18]. Lastly, the response to cloud security incidents faces unique challenges due to the dynamic nature of cloud infrastructures, highlighting the need for tailored forensic techniques and strategies.

III. ENHANCED SMART MODEL FOR CYBER CRIMINAL IDENTIFICATION

The main input parameters of the upgraded SMART model include call data, SMS, and social media messages. The steps followed in the enhanced SMART model are as follows. Data collection includes gathering messages, call logs, and social media posts from intended or required smartphone users. Preprocessing the data deals with missing values and eliminates duplicates. Data are normalized to ensure that every feature has a uniform scale.

A. Extracting Features from Smartphone Information

The feature extraction algorithm methodically analyzes smartphone data, focusing on call logs, messages, and social media communications. It computes key data such as total number of calls, average call duration, and call frequency to individual contacts. This algorithm performs keyword frequency and sentiment analysis on messages and social media interactions, assigning a sentiment score to each message type. These features are then standardized by min-max normalization to ensure a consistent scale for subsequent analysis.

```
Algorithm 1: Feature Extraction
Input: Data from a smartphone in its raw
   form (messages, call records, social
   media messages)
Output: Feature set that has been
   extracted and normalized
1: Call Data Extraction:
   Calculate the total number of calls
   $N_{Calls}$
   Compute average call duration as
   $Avg\_call\_duration = \frac{\sum call\_durations}{N_{Calls}}$
   Determine the frequency of calls to
   unique contacts:
   $F_{uniquecontacts} = \frac{N_{uniquecontacts}}{N_{Calls}}$
2: Messages' Extraction:
   Calculate the total number of messages:
   $N_{messages}$
   Perform keyword frequency analysis to
   identify common suspicious terms.
   Conduct sentiment analysis using a
   sentiment score $S_{messages}$ where
   $S_{messages} = \frac{\sum sentiment\_scores}{N\_messages}$
3: Social Media Messages' Extraction:
   Calculate the total number of social
   media messages: $N_{socialmessages}$
   Perform keyword frequency analysis to
   identify common suspicious terms.
   Conduct sentiment analysis using a
   sentiment score $S_{socialmessages}$ where:
   $S_{socialmessages} = \frac{\sum sentiment\_scores}{N_{Socialmessages}}$
4: Normalization
   Normalize features to a common scale
   [0, 1] using min-max normalization:
   $N_{CallsNormalized} = \frac{N_{Calls} - N_{CallsMin}}{N_{CallsMax} - N_{CallsMin}}$
```

B. Classification for Cyber Criminal Identification

The classification algorithm weights the normalized features according to their importance and generates weighted features for calls, SMS, and social media communications.

These weighted variables are used to provide a final score, which is then used to classify individuals using a predetermined threshold. The model is optimized through a training and validation procedure, using classification techniques such as Random Forest (RF) and Support Vector Machine (SVM) to ensure excellent accuracy, precision, recall, and F1 scores. This approach enables the effective identification of possible cyber offenders by utilizing smartphone data with mathematical rigor.

```
Algorithm 2: Classification for Criminal
  Identification
Input: Normalized and weighted feature set
Output: Classification labels (potential
   cybercriminal or not)
1: Parameter Weighting:
   Assign weights to each feature based on
   importance:
   Wcalls=0.4
   Wmessages=0.3
   Wsocial_messages=0.3
2: Weighted feature calculation:
     Compute weighted features using
     $Fcalls_{weighted} = W_{calls} * N_{CallsNormalized}$
     $Fmessages_{weighted} = W_{messages} * S_{messages}$
     $Fsmw = Wsm \ast Ssm$
     Where smw:  Social messages weighted
     sm: social messages
3: Combine Weighted Features:
   Combine weighted features to form a
   final score
   $F_{final} = Fcalls_{weighted} + Fmessages_{weighted} +$
     $Fsm_{weighted}$
4. Classification
   If $F_{final} \geq T$,
    classify as a potential cyber criminal
   Else
     classify as not a cyber-criminal.
5. Training and Validation:
   Split the dataset into training and
   validation sets.
   Train the classification models (RF,
   SVM) using the training set.
   Validate the model using the validation
   set and compute performance metrics
   such as accuracy, precision, recall,
   and F1 score
```

In the training and validation stage, a two-step process is used for classification and scoring. The first step involves classification using RF and SVM, and the second step computes a suspect score using the classification probabilities. RF and SVM are trained on a dataset containing features such as call logs, messages, location data, contacts, and more. These models are used to predict the likelihood of an individual being a cyber criminal based on smartphone data. Both classifiers output probabilities indicating the likelihood that each individual is a cyber-criminal. In the second step, after obtaining the probabilities from RF and SVM, a weighted average of the two probabilities is computed to give the final suspect score. The score is then compared against a threshold of 0.5 to determine if the individual is classified as a cyber-criminal. Any score above 0.5 indicates a higher likelihood of being a cyber-criminal.

The proposed method can significantly enhance existing cybercriminal identification models by integrating comprehensive feature extraction and classification techniques specifically designed for smartphone metadata. In feature extraction, smartphone data are analyzed focusing on call logs, messages, and social media communications. Key data are calculated, such as the total number of calls, average call duration, and call frequency to individual contacts. Keyword frequency and sentiment analysis are performed on messages and social media interactions, assigning a sentiment score to each message type. These features are then standardized by min-max normalization to ensure a consistent scale for subsequent analysis. The normalization process ensures consistent feature scaling, reducing biases, and improving data quality. Furthermore, incorporating robust machine learning models such as RF and SVM in the training and validation phases ensures high precision, recall, and F1 scores, making the system highly reliable. In general, this method bridges significant gaps in current models, particularly in handling mobile data, enhancing real-time detection capabilities, and ensuring proactive threat identification.

IV. IMPLEMENTATION

A. Applying the SMART Model to Smartphone Data

A case study was examined utilizing the SMART model on a dataset obtained from a smartphone. The collection comprises call logs, messages, and social media interactions collected from a solitary device. The aim was to assess the efficacy of the model in detecting probable cyber-criminal activities by analyzing the user's smartphone behavior**.**

1) Data Collection

The dataset used for training and testing this method is custom-made, comprising smartphone data from 100 individuals. The data was collected using the open-source digital forensics tool Autopsy, which allows for the extraction of relevant parameters such as call logs, messages, location data, contacts, internet activity, and device information. The extracted data include:

- Call logs: Call duration, caller ID, timestamps
- Messages: Text content, sender/receiver details, timestamps
- Location Data: GPS coordinates, timestamp, location history
- Contacts: Name, phone number, email
- Installed Applications: App usage, installation history
- Internet Activity: Browsing history, downloaded files

- Device Information: Model, operating system, storage usage

The dataset is not publicly available but can be shared upon request for research purposes.

2) Data Preprocessing

Mean imputation was used to handle missing values and data integrity was verified by removing duplicate records. Data were normalized using the min-max method to make feature scaling consistent.

V. RESULTS

The effectiveness of the SMART model in identifying possible cybercriminals can be verified by the performance metrics on benchmark datasets, as shown in Tables I and II. SMART obtained remarkable results with 0.94 precision, 0.91 recall, and 0.95 accuracy. These metrics show that the SMART model has a balanced ability to correctly detect true positives while limiting false positives and false negatives, indicating that it is extremely reliable in differentiating between cyber criminals and non-criminals.

TABLE I. PERFORMANCE COMPARISON OF DIFFERENT MODELS ON STANDARD DATASETS

Model	**Precision %**	**Recall %**	**Accuracy%**
SMART Model	94	91	95
Manual-Data Criminal Identification	88	82	85
Automated Cyber Criminal Investigation	75	80	78
Complete Cyber Criminal Identification	94	91	92

This table compares the proposed model with three other methods. The Manual Data Criminal Identification method was proposed in [1], where it was tested on a publicly available dataset of 500 cybercrime cases, focusing on the manual review of call logs and messages. The Automated Cyber Criminal Investigation method was proposed in [8], where it was tested on a dataset, containing smartphone data from 200 individuals suspected of cybercrimes. The Complete Cyber Criminal Identification method proposed in [9] uses machine learning techniques tested on a combination of synthetic and real-world smartphone datasets. Each method was evaluated on the same dataset for consistency in comparison.

TABLE II. PERFORMANCE COMPARISON OF THE PROPOSED MODEL ON DIFFERENT CONFIGURATIONS OF THE SAME DATASET

Model	**Precision %**	**Recall %**	**Accuracy%**
SMART	94	91	95
Raw data	92	89	93
Partially preprocessed	93	90	94
Fully preprocessed	91	88	92

Table II presents a comparison of the performance of the SMART model on three different configurations of the same dataset to understand the impact of preprocessing on the final results. The Raw data dataset contains raw, unprocessed data directly extracted from smartphones, without any preprocessing or standardization. The Partially preprocessed dataset underwent basic preprocessing, including handling missing values, deduplication, and timestamp normalization, without any scaling or encoding of categorical variables. The Fully preprocessed dataset was preprocessed using the Smartphone Data Preprocessing Algorithm (SDPA), which includes missing value imputation, deduplication, geospatial analysis, keyword search, scaling, and categorical data encoding. The SMART model retains good precision (90 to 94), recall (87 to 91), and accuracy (91 to 95) despite the increased variability and possible anomalies in these datasets. These outcomes highlight the durability and adaptability of the model, demonstrating its ability to provide dependable and consistent performance in a variety of data situations. Due to its flexibility, the SMART model is especially useful in real-world cybersecurity applications where data consistency and quality may fluctuate.

VI. CONCLUSION

The proposed SMART model provides a reliable and efficient way for detecting probable cybercriminal activities by using essential smartphone data such as call records, messages, and social media interactions. By focusing on these critical factors, the SMART model outperforms traditional approaches, which frequently rely on a broader set of variables that can dilute the analysis. The fundamental contribution of this work is its comprehensive feature extraction that ensures satisfactory precision and reliability across a wide range of data contexts. Compared to existing models, SMART not only improves cyber threat detection accuracy but also provides a viable and scalable solution for real-world cybersecurity applications. The model's capacity to adapt to different data contexts while maintaining consistent performance demonstrates its robustness and reliability. Furthermore, the SMART model provides a more efficient approach that decreases computing overhead while retaining satisfactory accuracy. This novel method greatly improves the ability to detect threats in real-time, making it a valuable tool for enhancing digital security and protecting against cyber attacks.

REFERENCES

[1] A. Dimitriadis, E. Lontzetidis, B. Kulvatunyou, N. Ivezic, D. Gritzalis, and I. Mavridis, "Fronesis: Digital Forensics-Based Early Detection of Ongoing Cyber-Attacks," *IEEE Access*, vol. 11, pp. 728–743, 2023, https://doi.org/10.1109/ACCESS.2022.3233404.

[2] S. Nasiri, M. T. Sharabian, and M. Aajami, "Using Combined One-Time Password for Prevention of Phishing Attacks," *Engineering, Technology & Applied Science Research*, vol. 7, no. 6, pp. 2328–2333, Dec. 2017, https://doi.org/10.48084/etasr.1510.

[3] J. Kumar and G. Ranganathan, "Malware Attack Detection in Large Scale Networks using the Ensemble Deep Restricted Boltzmann Machine," *Engineering, Technology & Applied Science Research*, vol. 13, no. 5, pp. 11773–11778, Oct. 2023, https://doi.org/10.48084/etasr.6204.

[4] P. Bradford and N. Hu, "A layered approach to insider threat detection and proactive forensics," in *Proceedings of the Twenty-First Annual Computer Security Applications Conference (Technology Blitz)*, 2005.

[5] A. Orebaugh, "Proactive Forensics," *Journal of Digital Forensic Practice*, Mar. 2006, https://doi.org/10.1080/15567280600626411.

[6] J. Sachowski, Implementing Digital Forensic Readiness: From Reactive to Proactive Process, Second Edition, 2nd ed. Boca Raton, FL, USA: CRC Press, 2019.

[7] B. D. Bryant and H. Saiedian, "A novel kill-chain framework for remote security log analysis with SIEM software," *Computers & Security*, vol. 67, pp. 198–210, Jun. 2017, https://doi.org/10.1016/j.cose.2017.03.003.

[8] "MITRE ATT&CK®." https://attack.mitre.org/.

[9] V. S. Harichandran, D. Walnycky, I. Baggili, and F. Breitinger, "CuFA: A more formal definition for digital forensic artifacts," *Digital Investigation*, vol. 18, pp. S125–S137, Aug. 2016, https://doi.org/10.1016/j.diin.2016.04.005.

[10] A. Dimitriadis, "Leveraging digital forensics and information sharing into prevention, incident response, and investigation of cyber threats," Ph.D. dissertation, University of Macedonia, Thessaloniki, Greece, 2022.

[11] B. L. Krishna, "Comparative Study of Fileless Ransomware," *International Journal of Trend in Scientific Research and Development*, vol. 4, no. 3, pp. 608–616, 2020, https://doi.org/10.13140/RG.2.2.14580.91521.

[12] H. Al-Mohannadi, Q. Mirza, A. Namanya, I. Awan, A. Cullen, and J. Disso, "Cyber-Attack Modeling Analysis Techniques: An Overview," in *2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, Vienna, Austria, Aug. 2016, pp. 69–76, https://doi.org/10.1109/W-FiCloud.2016.29.

[13] K. Aldriwish, "A Deep Learning Approach for Malware and Software Piracy Threat Detection," *Engineering, Technology & Applied Science Research*, vol. 11, no. 6, pp. 7757–7762, Dec. 2021, https://doi.org/10.48084/etasr.4412.

[14] A. Al-Marghilani, "Comprehensive Analysis of IoT Malware Evasion Techniques," *Engineering, Technology & Applied Science Research*, vol. 11, no. 4, pp. 7495–7500, Aug. 2021, https://doi.org/10.48084/etasr.4296.

[15] K. Muppavaram, M. Sreenivasa Rao, K. Rekanar, and R. Sarath Babu, "How Safe Is Your Mobile App? Mobile App Attacks and Defense," in *Proceedings of the Second International Conference on Computational Intelligence and Informatics*, 2018, pp. 199–207, https://doi.org/10.1007/978-981-10-8228-3_19.

[16] "Home," *UCO Community*. https://www.unifiedcyberontology.org/.

[17] "Pellet - Semantic Web Standards." https://www.w3.org/2001/sw/wiki/Pellet.

[18] B. E. Strom *et al.*, "Finding Cyber Threats with ATT&CK^{TM}-Based Analytics," MITRE, Technical Report MTR170202, Jun. 2017. [Online]. Available: https://apps.dtic.mil/sti/trecms/pdf/AD1107945.pdf.

[19] "ATT&CK Data & Tools | MITRE ATT&CK®." https://attack.mitre.org/resources/attack-data-and-tools/.