The Enhanced Forensic Examination and Analysis for Mobile Cloud Platform by Applying Data Mining Methods

Ibrahim Ali Alnajjar
School of Computing, Universiti Utara Malaysia, Kedah, Malaysia.

Massudi Mahmuddin
School of Computing, Universiti Utara Malaysia, Kedah, Malaysia.
E-mail: ady@uum.edu.my

Received October 09, 2020; Accepted November 18, 2020
ISSN: 1735-188X
DOI: 10.14704/WEB/V18SI01/WEB18006

Abstract

Investigating the mobile cloud environment is a challenging task due to the characteristics of voluminous data, dispersion of data, virtualization, and diverse data. Recent research works focus on applying the latest forensic methodologies to the mobile cloud investigation. This paper proposes an enhanced forensic examination and analysis model for the mobile cloud environment that incorporates timeline analysis, hash filtering, data carving, and data transformation sub-phases to improve the performance of the cloud evidence identification and overall forensic decision-making. It analyzes the timeline of events and filters the case-specific files based on the hash values and metadata using the data mining methods. The proposed forensic model performs the in-place carving on the filtered data to guide the investigation and integrates the heterogeneous file types and distributed pieces of evidence with the assistance of the data mining. Finally, the proposed approach employs LSTM based model that significantly improves the forensic decision making.

Keywords

Mobile Cloud Forensics, Examination and Analysis, Hash Filtering, Data Carving, Data Transformation, and Data Mining.

Introduction

The dramatic development in Information Technology (IT) has introduced the Mobile Cloud Computing (MCC) technology [1] that vividly changes the computing environment from physical to virtual world. Cloud computing has become the dominant technology for operating mobile applications. MCC enables mobile users to store the data and process the
applications in the cloud, which mitigates constraints of the smartphones such as battery life, memory capacity, processing delay, and computational power [2].

Smartphones offer greater capabilities to their users in the form of applications with the support of internet access. In smartphones, most of the mobile applications such as Google Mail and Facebook utilize cloud computing technology due to the advantage of the distributed and scalable cloud environment. The rapid increase of cyber-attacks significantly impacts the MCC that includes Denial-of-Service (DoS), Distributed Denial-of-Service (DDoS), botnets, web-based attacks, physical-based attacks, network-based attacks, or application-based attacks [3]. Nowadays, smartphones are infected by viruses in different ways, such as through Short Messaging Service (SMS) messages, instant messengers, internet Wireless Application Protocol (WAP) application downloads, storage cards, and Bluetooth [4]. Hence, cyber criminals take advantage of the advanced technology such as MCC to commit the crimes in the form of propagating a terrorist ideology, facilitating terrorist communication, or disseminating attacks against the software applications, programs, or digital information of the mobile user. Therefore, conducting a forensic investigation on the mobile cloud environment becomes an emergent need.

Cloud forensics is also in its infancy stage due to the lack of knowing the full impact of the digital forensics community on the cloud model [5]. In addition, applying traditional digital forensic tools and procedures for the investigation of a mobile cloud environment is inappropriate for the volatile mobile environment and virtualization technologies and remote storage-enabled cloud environment [6].

Hence, it is essential to understand the impact of cloud computing on smartphone forensics and analyze whether the existing mobile forensics methodologies, tools, and techniques support mobile cloud scenarios or not [7]. The phases of digital forensics include identification, preservation, collection, examination, analysis, and presentation. Due to the inter-relationship between the mobile and cloud during the execution of the cloud-based mobile applications, the mobile cloud forensics involves the evidence correlation in addition to the traditional forensic procedures in both the mobile and cloud environments. In the context of mobile cloud forensics, the examination and analysis phase plays a significant role in improving the performance of the investigation compared to other investigation phases [8, 9]. It is because the major part of the evidence identification in the cloud scenario heavily relies on the examination and analysis phase in the smartphone, and crime event decision-making relies on the examination and analysis phase in the cloud. Examining evidential artifacts from the mobile cloud services such as
emails and communication applications provide the potential information to the forensic investigator to reconstruct the crime event, which is beneficial for law enforcement to prove the investigation. Thus, the proposed approach targets to enhance the examination and analysis phase and its sub-phases in the mobile cloud environment with the assistance of the data mining methods.

This paper model the mobile cloud forensic methodology, especially enhancing the examination and analysis phase of the mobile and cloud environment. The primary contributions are as follows.

- The proposed forensic examination and analysis model incorporates the hash filtering, data carving, and data transformation sub-phases in both the mobile and cloud environment to improve the investigation performance along with the help of data mining methods.
- Instead of massively analyzing the information acquired from the collection phase, the proposed forensic model performs the timeline analysis and hash filtering to filter the acquired files in the large-scale collaborative environment effectively.
- By performing the in-place carving in the filtered artifacts, the proposed approach extracts the potential features alone and validates the files based on the characteristics of the file types.
- With the target of facilitating the forensic investigation among the heterogeneous type of evidence in the distributed cloud storage, the proposed approach precisely integrates the files and thus, builds the forensic evidence taxonomy for the corresponding crime event.
- The proposed approach improves the investigation performance in the mobile cloud by executing the proposed sub-phases and applying the data mining methods to deliver accurate results.

**Forensic Methodologies**

National Institute of Standards and Technology (NIST) forensic procedure is widely applied to mobile and cloud investigation. This section reviews the conventional research works on mobile forensics, cloud forensics, and cloud-based mobile application forensics.

**Mobile and Cloud Forensic Approaches**

The forensics research [10] automatically monitors the malicious activities performed on Android devices using several forensic components include a tractor beam for an Android device, server, analysis framework, and central database. Even though it leverages the
identification of the malicious android applications through ascertaining monitoring of the
android device, it fails to find more than one deleted applications due to the lack of
accessing the persistent storage location in the android. Droid Watch [11] is an
open-source and automated enterprise monitoring system for an Android device, which
continuously gathers, stores, and transfers the forensically-rich information of the Android
smartphone to a web server after obtaining the user consent without root privileges.
However, there is the possibility of a DoS attack before transferring the data to a Web
server. Mobile Forensic Investigation (MFI) life cycle model [12] resolves the
shortcomings in the traditional digital forensic investigation methods using new methods
and techniques in the lifecycle process. It involves data gathering, preservation, and report
generation with its sub-phases. Harmonized Digital Forensic Investigation Process
(HDFIP) [13] extracts the potential evidence of the mobile device in a forensically sound
manner, which ensures the flexibility, availability, integrity, adaptiveness, confidentiality,
accountability, and comprehensiveness during the evidence acquisition. Smartphone
forensic investigation process model (SPFIPM) [14] performs the smartphone
investigation using a fourteen-stage model for finding potential evidence. The research
work [15] presents the forensic data collection methodology for Android devices. Android
Forensic Data Analyzer (AFDA) [16] targets to reduce the workload for the investigator
by correlating the events of the same or different android applications to improve the
exposure of the hidden forensic data. It assists the investigator in effectively and quickly
analyze the forensic image. Android malware forensic analysis model [17] examines the
malware behaviors on the Android device and reconstructs the malicious events to
detect the suspicious programs quickly. The research work [18] conducts the forensic analysis
on two instant messenger applications that is cloud-based mobile applications such as
WhatsApp and Viber. It extracts the data pertaining to those two instant messengers from
its chat logs, images, chat history, and video files. WhatsApp forensics model [19]
extracts the forensic-rich evidence from the volatile and non-volatile memory of the
android device regarding the WhatsApp application activities and then analyzes the
extracted data in a forensically sound manner. The research work [20] assesses the data
related to the cloud storage applications stored on the client-side android and iOS devices.
In this case, such smartphones act as a proxy for the cloud storage data, which are utilized
for investigating the cloud services with fruitful benefits.

**Examination and Analysis based Approaches**

The research work [21] employs the Visualize Association Inside Emails (VAIE) system
to enable the forensic examiner to acquire the information regarding e-mails. With the
assistance of two layout models such as the spring force model and radial tree model, the
forensic investigator visualizes the E-mail relevant data for better understand-ability. Instead of analyzing the information about Apple iOS mobile devices such as contact details, text messages, and voice messages, the research work [22] examines the evidential artifacts in the third party application that comprises the potential information such as user account, timestamps, geolocation, native files, and additional contact information on the mobile device. Twitter, Facebook, Skype, iBooks, Four Square, Where.com, and Bright Kite are several examples of cloud-based third-party applications. The earlier forensic examination research works have focused on extracting the information regarding username and filename alone of the cloud-based mobile applications such as Skype, Viber, Dropbox, and Facebook. To improve the investigation performance, the research work [23] forensically analyzes the social networking applications on different Smartphones and recovers the events, which helps to extract the case-specific information. By forensically analyzing the Skype application according to work [24], the forensic examiner recovers the potential evidence from the RAM and NAND flash memories of the Android devices. The forensic analysis model [25] extracts the evidence both the logical storage and internal flash memory of WeChat application through MOBIL edit based logical acquisition and Chip-off extraction based physical acquisition in iPhone respectively. Even though it examines the WeChat application folders on the mobile device, it lacks to investigate the application-relevant evidence in the cloud environment.

**System Model**

This section presents a system model for implementing the proposed forensic examination and analysis methodology in the mobile cloud environment. Let the cloud-based mobile application running on the Android device. During the mobile forensics, the investigator accesses the partitions of the android device, including boot system, recovery, data, and cache along with the Secure Digital (SD)card and SD-ext. If the forensic investigator examines the crime event that is launched on the cloud-based mobile application, the investigator acquires the log file entries, file system metadata structures, application logs, registry information, chat logs, network packet, and so on. Similarly, the required metadata information is acquired from the cloud server. These metadata parameters are considered as the input to the forensic tools for the mobile cloud forensic examination and analysis.

Let $X_f^M$ denotes the forensic data belongs to the mobile device, and $Y_f^\text{Ref}$ denotes the forensic data belongs to the reference dataset in the National Software Reference Library (NSRL). Wherein, ‘f’, ‘M’, and ‘Ref’ denotes the features, mobile device, and reference dataset, respectively. Moreover, consider the ‘P’ number of forensic evidences acquired
from the cloud environment, \( FE = \{FE (1), FE (2), \ldots, FE (P)\} \) in which \( FE (i) \) is a particular evidence, and \( FE (i) \) is its neighbor evidence, which is to be matched during co-variance calculation. Each forensic evidence has ‘N’ number of features, \( f = \{1,2,\ldots,N\} \). For instance, each forensic record or data in a set of acquired data from the mobile devices is denoted as \( X_f^M \) and each forensic record or data in a set of good forensic data in the reference dataset is referred to as \( Y_f^\text{Ref} \). Let the forensic data have two features, such as IP address and IMEI number. Hence, \( N = 2 \). For example, the acquired forensic data \( (X_f^M) \) from the suspected mobile device have the IP address \( (X_{f1}^M) \) of 131.175.17.9 and IMEI number \( (X_{f2}^M) \) of 990000862471238. The reference dataset comprises that sample record \( (Y_f^\text{Ref}) \) is 131.175.17.9 as a legitimate IP address \( (Y_{f1}^\text{Ref}) \) and 990000862471854 as the IMEI \( (Y_{f2}^\text{Ref}) \) of the genuine user. In the example above, the IP address and IMEI number are the features of the forensic data or record.

It is essential to ensure that the acquired information is in the form of a structured dataset such as Comma Separated Value (CSV) format to apply the data mining techniques on the acquired forensic artifacts. It is because the proposed forensic model attempts to reduce the burden of forensic investigation and apply the mining techniques. Hence, it transforms the input forensic data into the required table format of rows and columns in which rows represent the unique forensic data or record, and columns denote the fields, features, or attributes. From the application database, the messages are modeled as \( A_M \) with different fields, including receiving message ID, sent message ID, sent and received time, media caption name, file type, file size, and so on. Metadata assists in finding the associated artifacts that are useful to verify the fraud, malicious insider, and other types of cybercrimes in the mobile cloud. It describes the attributes of the applications or files in the digital source of evidence, providing the logical consistency, accuracy, and coherence of the applications or files. The evidence parameters are in different data types such as string (S), integer (I), boolean (B), data and time (DT). If the source of evidence is log file entries, the parameters involve message or event ID, timestamp, Internet Protocol (IP) address, International Mobile Equipment Identity (IMEI), and user ID. Application logs include the application name, version, and a timestamp. The parameters of the File system metadata structures are file name, file size, file type, creation time, modification time, access time, and file status, such as active or hidden. In the network packet, the parameters of the source of evidence include packet length, source IP, and destination IP. Figure 1 illustrates the components involved in the proposed mobile cloud forensic examination and analysis.
An Improved Mobile Cloud Forensic Examination and Analysis Model

In forensic computing, examination and analysis are the essential phases in improving investigation accuracy and efficiency. According to the NIST definition, ‘Examination’ is the process of identifying and extracting the relevant artifacts from the acquired data while protecting the data integrity and ‘Analysis’ is the process of analyzing the examination results to derive the significant artifacts regarding the questions which stimulate the collection and examination process.

Figure 2 shows the main phases of the mobile cloud forensics and designs the necessary sub-phases for the examination and analysis phase in the mobile cloud environment. The sub-phases of the forensic examination and analysis phase include timeline analysis, hash filtering, data carving, data transformation, cross-referencing, and keyword searching. The proposed approach focuses on improving the forensic investigation in the mobile cloud environment.
environment by enhancing the examination and analysis phase, which has depicted in Figure 3.

![Figure 2 The Phases of the Mobile Cloud Forensics with the Processes involved in the Examination and Analysis Phase](image-url)
Figure 3 The Proposed Forensic Examination and Analysis Phase in Mobile Cloud

**Timeline Analysis**

In the context of mobile cloud forensics, timeline analysis plays an essential role in filtering the potential evidence regarding the time of the crime event in both the mobile device and the cloud without analyzing the evidential artifacts which are irrelevant to the
crime event. In order to identify the creation, modification, and deletion time of the file before and after the infection time, the forensic examination and analysis methodology performs the timeline analysis to estimate the activity performed by the attacker after the infection. During the timeline analysis, the four different types of times are considered as the key factors, including the last modification time, the last access time, the last change time of the Master File Table (MFT) entry, and the creation time. To obtain higher investigation accuracy, the forensic examiners need to perform end-to-end temporal analysis in terms of the timeline of events. Temporal characteristics have crucial importance over different aspects such as the interaction of events with people, processes, or objects.

In the cloud environment, forensically validating the logging framework tends to create challenges in the timeline of events. The existing forensic investigation methods examine only for the accessed timestamps, data remnants, and file contents, which consider the logs are not necessary for the forensic investigation. However, the examination of log files provides fruitful benefits to the investigator in connecting the dots during the investigation. To effectively perform the timeline analysis in the mobile cloud environment, the proposed examination and analysis methodology has focused on modeling the knowledge representation. The knowledge representation model comprises a large set of entities and relations between the events, which necessitates the subsequent forensic analysis of other entities. It facilitates the reconstruction of the investigation process to provide credibility to the investigated results. An automated timeline tool forensically analyzes the web servers to determine the previously occurred website links by applying the preprocessing on the raw web log files and path analysis of the URL information. During the timeline analysis, the forensic investigator initially examines the date and day concerning the occurred event.

**Hash Filtering**

The data reduction methods and hash sets address the data size constraint to handle an ever-increasing amount of data during a forensic investigation. Hash filtering is one of the time-saving techniques for the mobile cloud forensics examiner while dealing with large-scale data. Consequently, it reduces the time on investigating the known suitable files and facilitates the examiner in the mobile cloud environment within the scope of the investigation. The proposed model generates the hash files for every file and validates the generated hash files with a set of hashes of the previously calculated known useful files to filter out the matched known and suitable hash files from the forensic database. The files belong to the residual hash files in the forensic database are to be examined further by the
forensic investigator. By applying the deep neural network classifier, the proposed model recognizes the patterns in the data and classifies the relevant files that are to be used for further investigation. In order to enhance the effectiveness of the forensic database that is to be investigated, the proposed model focuses on filtering the files stored in the forensic database based on the relevance to the crime event. Initially, it attempts to determine the ignorable files from the acquired data based on the hash data that reflects the corresponding software. Moreover, it examines the most recent hash values from the hash database to retain only the potential files for further investigation.

![Figure 4 Steps involved in the Hash Filtering Process](image)

In the field of mobile forensics, the current methods have filtered the irrelevant files through reporting the consistent hash files and validating the reported hash data of individual files from its back-to-back acquisitions and subsequent acquisitions, respectively. In the evidence database, the images contain metadata or Exif information
involving the timestamp and several additional attributes that assist in determining the source of the image during the investigation. Metadata is the information about the files that are acquired from the mobile device and the cloud, includes file size, file name, last accessed time, and so on.

\[
\text{Corr}_\text{Score} = \frac{1}{N} \sum_{f=1}^{N} \text{Sim}(X^M_f, Y^\text{Ref}_f)
\]

(1)

Equation (1) computes the correlation score between the data residing in the mobile device \(X^M_f\) and the reference dataset \(Y^\text{Ref}_f\). In Equation (1), ‘f’ refers to the features that vary from ‘1’ to ‘N’, which are based on three main categories such as file metadata, case metadata, and file content-based data. The features include file type, IP address, file name, permissions, file ID, email ID, file content type, URLs, and so on. If the correlation score is high when correlating with the known useful features, the proposed methodology ignores that information from the forensic evidence database during the investigation.

In the field of cloud forensics, segregating the relevant evidence from the acquired evidence is a challenging process due to the cause of the encrypted data. Even though live forensics enhance the forensic capabilities through frequently capturing the images of the running environment, it leads to overhead, cost, and performance issues. Hence, securely maintaining the hash values of the forensic artifacts is essential such as secure log files. Subsequently, the forensic investigator needs to create standard forensic procedures to have access to the decryption key without privacy violations.

\[
\text{Class} = \begin{cases} 
1 & \text{if } \left[ b + \sum_i w_i x_i > 0 \right] \\
0 & \text{if } \left[ b + \sum_i w_i x_i \leq 0 \right]
\end{cases}
\]

(2)

Equation (2) predicts the class of the input data while employing the neural network as a classifier for the forensic good and bad files classification. In Equation (2), ‘\(w_i\)’ represents the weight that shows the relative influence of the input ‘\(x_i\)’, and ‘\(b\)’ denotes the bias in the classification results based on the neurons. By training the neural network classifier with the known good files characteristics or patterns, the proposed methodology filters out
the good information from the acquired evidential artifacts. Initially, the forensic investigator identifies several parameters such as timestamp, IP address, and Media Access Control (MAC) address for forensic analysis to retain the evidence of malicious activity alone. Moreover, search activities of frequent users in terms of verifying their IP addresses, byte transferred, and file access. By analyzing the behavior of the user, the proposed model confirms the IP address of the malicious user.

**Data Carving**

In the field of cyber forensics, the term ‘file or data carving’ has been used, which extracts the data from the raw data in terms of evidential artifacts through identifying and recovering the relevant data based on the forensic analysis. Data or file carving plays a crucial role in forensically determining the deleted or hidden files from the digital media stored in both the mobile and the cloud.

In the smartphone and the cloud, the hidden areas of a file include slack space, lost clusters, and unallocated clusters of the digital media or disk. In order to apply this type of extraction method, there is a need for a standard file signature with the file header and footer, which enforces the search to extract and analyze the file during the file validation.

The existing forensic data carving tools search the header signature of each file in the digital media even when there is the existence of the file system for the corresponding media. In this context, to mitigate the search space, it is adequate to search the header signatures in the slack space, lost clusters, and unallocated clusters on the disk.

Also, to minimize the time of searching the header signature, the data carving methods have focused on searching the header signatures within the few bytes of the cluster and sector, and throughout the sectors. In the data carving method, file structure-based carving identifies the information level in a file format to recognize the file while matching the extracted information on the file format with the raw dataset. The content-based file carving method explores the content of the clusters to identify the relationships between the files regarding a specific file, which assists to reassemble the original file recovery.
Figure 5 shows the proposed data carving process for the mobile cloud forensics. The proposed examination and analysis model applies the file carving in the hidden areas when there is an availability of the file system. Moreover, it performs the file carving from the raw image even when there is no file system. In consequence, the proposed model employs a disk analysis tool to provide the in-place or zero storage carving in the slack space, lost clusters, and unallocated clusters. In essence, the proposed forensic methodology incorporates several key components to perform the file carving process with the time efficiency in the mobile cloud environment. After filtering the irrelevant data from the forensic database, the proposed model attempts to provide the dynamically updated file system to the investigator to mitigate their waiting time when there is unavailable of the file system. In subsequence, it applies the signature, file structure, and content-based carving method to extract the critical data from the hidden areas of the digital media. In the proposed methodology, data carving involves the signature-based carving, file structure based carving, and content-based carving, which extracts the hidden information from the reduced set that acquired evidential artifacts. To extract the data
hidden in the files, the forensic investigator needs to understand the signature, structure, and content of the files due to the availability of the different data hiding process. Figure 6 illustrates the detailed procedure of the data carving process in the proposed mobile cloud forensic examination and analysis methodology.

- **File Signature Based Carving:** With the help of file signatures, the proposed data carving method identifies the file types from the unallocated clusters. The file signature is referred to as a text value, magic number, or numerical value, which assists the identification of file format. File signature-based data carving plays a significant role in defending against the data hiding techniques, which facilitates the forensic investigator in the massive mobile cloud environment. File signature analysis is based on the process of matching the files, headers, and extensions with the existing databases to determine the hidden files. In essence, the forensic investigator focuses on the extraction of the information about the type of the file from the header or footer fields in a file, termed as file signature. The forensic investigator often matches the file signature with the file extensions to identify similar files, in some cases, there are few exceptions, such as mismatches, unknown types, no matches, and anomalous results while matching the files. File signatures are also known as the magic numbers, which have unique values with the replacement of named constants. The file signatures are either in the hexadecimal format or the ISO 8859-1 encoding format for different file formats. For example, the file format of Adobe PDF and JPEG image has the file signature as the 25 50 44 46 and FF D8 FF respectively, which are in the hexadecimal format.

- **File Structure-Based Carving:** In order to determine the fragmented files, extracting the internal structure of the file is essential to facilitate the forensic investigator in recovering a file and identifying the starting and ending point of the data. The forensic technique of file carving focuses on recovering the files based only on the file structure and content without matching the metadata of the file system. It often recovers the files from the unallocated space in the mobile and cloud storage in which unallocated space indicates that the area of holding no files as a long time. Moreover, it recovers the files from entire storage when there is the existence of the missing or damaged file structures in both the mobile and cloud environments. By utilizing the file system structure, the forensic investigator quickly identifies and extracts the undeleted data from the unallocated space. For example, consider a bitmap file comprises the file size in bytes in the footer, JPEG file contains the metadata sequence, and word file comprises the byte strings such as keywords, author, and company. By accessing the file system structure, the forensic investigator can easily recover the deleted file from the presence of the file entry and information linking with the clusters.
• **File Content-Based Carving:** File content based carving relies on the content structure such as XML and HTML and content characteristics such as statistical attributes, character count, text, or language recognition. File content carving is also known as the semantic carving. The file content is mostly encoded in different ways while storing the data in the file for different applications. Hence, the associated applications only access the stored file in terms of reading and extracting the content of the file. Instead of deleting the malicious files or evidence, the malicious individuals or criminals often hide the information inside another file by changing its extensions to misguide the investigator. For example, criminals hide their sensitive information by modifying the file extension, such as changing the.doc file to.jpg file.

During the mobile forensics, the proposed methodology focuses on the dynamic updation of the file system, in-place carving, signature, file structure and content-based carving, and file validation. Whereas, in the context of the cloud forensics, to deal with the massive amount of data storage, the proposed methodology employs the feature extraction techniques such as Principle Component Analysis (PCA) while analyzing the content of the files along with the in-place carving. In both the mobile and cloud, in-place carving enables the investigator to examine the files without copying the content of the files, and it stores the metadata of the files in the forensic database for further forensic examination. The in-place carving process outcomes the details include the file name, starting position of the file in the device or cloud, file length, location of the file, and truncated details. By applying PCA, the proposed forensic examination and analysis methodology selects a set of features based on the principle components of the massive cloud Forensic Evidence (FE). To identify the principle components, the proposed methodology computes the covariance matrix based on the training samples and calculates the eigenvectors and eigenvalues. The feature extraction or feature selection method forms a subset of the features of the raw features, which retains only the potential contents of the overall data. For instance, to identify the file types such as ‘doc’, ‘gif’, ‘pdf’, ‘jpg’, and ‘html’, the proposed model analyzes the features in terms of the frequency of occurring each byte value as well as the content of the file. It assumes that the same file types have the same characteristics even for the different files, which assists in detecting the file type related to the suspected activity. In order to reduce the false positives and speed up the carving time, the proposed methodology performs the file validation based on the file type validator during carving operations. The file validation method conducts the matching process for the retained details of the carving operations in the mobile cloud environment. The file validation of the proposed methodology matches the carved details of the files on the mobile device with the carved details of the files in the cloud to determine the inherently
correlated files, and fine-tune the evidential artifacts towards the potential evidence for the corresponding crime event.

Data Transformation

In the mobile cloud environment, data transformation or integration plays a crucial role mainly to facilitate the hidden patterns and potential information from a large amount of cloud evidential artifacts compared to the mobile evidence. Data integration is the process of combining massive data from the disparate sources and transforms it into a unified structure. In the context of mobile cloud forensics, data integration combines the data
acquired from the various data centers and transforms the heterogeneous types of forensic evidence into a single format. Consequently, it assists in guiding the investigator to make the right decisions at the right time during the investigation. Even though traditional data mining techniques efficiently perform the integration of large volumes of unstructured, semi-structured, or streaming data, it lacks to integrate heterogeneous evidence from different dynamic locations. Most of the existing security systems examine the footprints of the intruders from the log files of the system to identify the potential evidence from the massively acquired evidence regarding the suspicious activities. Even though current forensic systems support the investigation with the trace of network traffic, account management, file system checkers, system monitoring, and system log files, there is a lack of providing adequate pieces of evidence for the event reconstruction due to the independent logging.

Figure 7 Steps involved in the Data Transformation Process
Integrating the Forensic Logs using Long-Short Term Memory (LSTM) Model

Figure 7 illustrates the proposed components incorporated with the data transformation in the mobile cloud forensic examination and analysis phase. In the field of mobile cloud forensics, the proposed forensic examination and analysis model handles the heterogeneous types of evidence acquired from the smartphones and cloud concerning the cloud-based mobile application execution.

If the forensic collection phase acquires the system call and its accompanied activities of the users, the investigator easily reproduces the security breaches or crime events that happened on the smartphone. The acquisition of system calls enables the investigator to accomplish the time efficiency and completeness of logging data through only recording the relevant information. Owing to the existence of the system call logging module in the suspected device, the forensic investigator identifies the modification of logging information from the recorded activities in the backend storage system.

The logging system collects the information about all the activities of the cloud-based mobile application performed by the users and securely stores the information in a forensic server to facilitate the forensic investigation. Thus, in essence, the proposed examination and analysis phase needs to analyze both the volatile and non-volatile information in the aspect of the users and processes, chronological order, and activities of the users. Moreover, with the help of LSTM, the proposed data transformation model integrates the forensic evidence into a unified format from the multiple modalities.

During mobile forensics, the proposed forensic methodology handles the heterogeneous types of the evidence acquired from the activities of the cloud-based mobile application of the suspect’s device and integrates such files with the help of string matching regarding the directory, file name, and metadata.

In cloud forensics, the proposed data transformation model employs the LSTM architecture, which is an extension of the Recurrent Neural Network (RNN). The LSTM architecture is more suitable for the forensic data integration since it contextually memorizes and recalls the forensic images even when the forensic database comprises the snapshots as the raw pixel data of the image in terms of high dimensional inputs. Hence, the proposed forensic methodology combines the LSTM architecture with the multimodal autoencoder to unify the information of various modalities, which assists in realizing the temporal sequence of the multimodal representation for a particular event. In the context
of forensics, the proposed approach considers the heterogeneous type of evidence, such as text log files and snapshot images, as the learning modalities.

Initially, the proposed forensic model extracts a set of feature vectors for each modality using the encoder LSTM and then forms the integrated feature vector of all the modalities with the consideration of temporal information. In subsequence, it employs the multimodal autoencoder to reconstruct the feature vectors for each modality from the integrated feature vector and then exploits the decoder LSTM to decode the input in chronological order from the reconstructed feature vector.

Thus, it facilitates the forensic investigator through heterogeneous data types integration acquired from the distributed data centers regarding a particular cloud-based mobile application execution. In the cloud environment, the forensic data transformation is relatively similar to the forensic logging supporting four services, such as completeness, authenticity, reproducibility, and efficiency.

An LSTM network [26], input gate, forget gate, and cell state are the essential elements. In order to support the forensic logs integration with the knowledge of the crime event and evidence nature, the proposed methodology attempts to model the LSTM with an additional control cell that manipulates the forensic features during the learning process. As a result, the LSTM based forensic methodology decides that the information be retained for further investigation in an integrated form. It categorizes the forensic evidential artifacts based on the relevancy in terms of activities within a particular timestamp or region.

\[
F_t = \left[ \sigma \ W_{\text{event}} w_t + \sum_{l} a_l W_{sf}^l h_{t-1}^l \right] \\
\]

In Equation (3), Wf event and Wsfl are the weight information related to the forensic event and layer-wise similar features, respectively. Wt is the input at time step ‘t’, and a1 is the layer-wise constant. Ft is the forensic gate that is used to learn certain patterns regarding the crime event and forensic characteristics, which is used to update control vector in the LSTM network, i.e. d_t=Ft ∘ d_{t-1}. 

http://www.webology.org
Figure 8 LSTM-based Evidence Integration in the Cloud

Figure 8 depicts the proposed evidence integration process using LSTM in the massive and distributed evidence storage environment of the cloud. The proposed forensic integration and linking among the extracted evidence reveal that the relationships between the same activities and also existing relationships between the targeted victim and the evidence. According to the committed crime in the mobile cloud environment, the proposed forensic examination and analysis methodology captures the possible relevant pieces of evidence that are linked to either attacker or targeted suspect. To reduce the complexity of linking or grouping the evidence related to the same activity or person from the distributed storage, the proposed model employs the LSTM for data integration. The LSTM-based classification model automatically handles this process to reduce manual errors and mitigate time. It finds the relationship between the previously obtained evidence. The primary advantage of the long-term memory in the LSTM enforces the integration of the evidence through its processing of sequence chains. Wherein gates decide which information is relevant to the sequence of evidence to make decisions on
either keeping or forgetting the information during training. In LSTM, sigmoid activation squishes the values between 0 and 1 to update or forget the incoming data. The forget gate decision heavily relies on the outcomes of the sigmoid function. By processing the information from the previous hidden state and current state, the sigmoid function outcome values between 0 and 1. If the value is closer to 0, the LSTM forgets the information; otherwise, it keeps that information. The input gate and output gate are responsible for deciding what information is relevant to add from the current state and decides what information is to be in the next hidden state respectively. Accordingly, at the end of the LSTM, the proposed methodology obtains a set of grouped evidence by applying Equation (3), which facilitates the forensic investigation rather than analyzing the evidence of a particular crime event in a distributed manner.

In the mobile cloud environment, the forensic investigator combines all the forensic logs or evidence into a forensic repository to analyze the acquired files for the corresponding malicious activity performed in the cloud-based mobile application. The log machines or forensic server stores all the log files acquired from both the mobile and cloud environments. In essence, the forensic server contains the forensic data repository. The integration of several mobile cloud forensic logs includes system_log, secure_log, packet_log, and volatile log. Finally, to ease the forensic investigation process over the abundant collection of the cloud forensic evidence, the proposed forensic examination and analysis methodology builds the case-specific forensic evidence taxonomy for the corresponding evidence. In order to model the massive cloud evidence in the form of machine-understandable format, annotation of the evidence with metadata has performed by the proposed forensic methodology. The evidence management process annotates the evidence with the semantic information in a logical manner. The tagged or annotated information assists the investigator to analyze and report the evidence in the cloud easily. In the proposed case-specific evidence taxonomy, the metadata provides the rudimentary details about the corresponding evidence such as folders and files, which comprises the concepts and properties to model the taxonomy with richer description. The proposed forensic methodology improves the forensic decision-making by finding the criminal for the corresponding crime event performed in the cloud-based mobile applications. It quickly responds to the investigation query by retaining and analyzing only the forensically relevant information such as case-specific data in the mobile cloud environment with the support of the data mining techniques. Finally, the mobile cloud forensic framework iteratively correlates the evidence that is analyzed from both the mobile and cloud and provides significant results under the forensically sound conditions.
Experimental Evaluation

This section describes the prototype for evaluating the LSTM-based data transformation in the proposed forensic examination and analysis methodology. The experimental model evaluates the part of the enhanced forensic methodology with a set of logs collected from the Open Nebula cloud environment.

Experimental Setup

The experimental model employs Open Nebula to collect cloud activity logs regarding the execution of mobile cloud applications. Open Nebula is a widely used open-source toolkit for Infrastructure as a Service (IaaS) cloud computing, which is a virtualization tool enabling the computations in a private cloud or public cloud. It comprises three logging systems, such as syslog logging, logging to standard error system, and file-based logging systems. The proposed forensic model utilizes the log files extracted from ‘/var/log/one’ in Open Nebula, which is a file-based logging system. The investigator acquires the evidence from the multiple log files such as oned.log, sched.log, sunstone.log, and VMID.log. In essence, Open Nebula stores the logs for the activities performed in the Virtual Machine as the files with the VM ID similarly, the multiple log files exist in Open Nebula cloud. For example, if a suspect launches the malicious activity during application execution in VM1, the forensic investigator needs to collect the VM1.log file. Moreover, Open Nebula supports features such as multi-tenant computing, data center federation, and virtual data centers on top of vCenter. Multiple OpenNebula data centers or zones are in the form of federation, sharing the information related to the same user accounts, permissions, and groups across the virtual data centers. As a result, the forensic investigator acquires the evidence of a particular suspect or event from the different OpenNebula virtual data centers.

Details for LSTM Implementation

This section demonstrates the final process of the proposed examination and analysis methodology in terms of implementing the LSTM-based data transformation process for the cloud evidence using the Java programming language. The experimental model assumes that the synthetic input dataset of LSTM is the outcome of several proposed examination and analysis processes such as timeline analysis, hash filtering, and data carving. After partially fine-tuning the acquired evidence, the experimental model conducts the experiments on the cloud evidence using LSTM according to the procedures explained in the data transformation process. The synthetic dataset of the cloud evidence consists of several fields such as crime event timestamp, crime type such as identity theft,
stalking, pornography, hacking, software piracy, and illegal electronic surveillance, user_ID, application_ID, event_ID, datacenter or zone_ID, server_ID, log_ID, and timestamp. Also, it comprises several footprints such as file creation time (C Time), file altered time (A Time), file read time (R Time), Master File Table (MFT) changed time (M Time), file permissions, including read-only, archive, offline, hidden, temporary, encrypted, and compressed, file size, file type, access control type such as access allowed, denied, and audit, file name namespace, and file name length. The experimental model applies the LSTM algorithm on such a generated synthetic dataset for the distributed evidence linking based on the sequence of inputs over time. The experimental model splits the generated dataset into a training set and testing set in the ratio of 70:30. From the entire forensic evidence in the generated dataset, 70 percent of the dataset is applied for training the LSTM and the remaining 30 percent for testing. The experimental model employs the crime type field as the target variable which is selected from the training set to integrate or inter-link the forensic evidence from the perspective of distributed evidence and heterogeneous file types of a similar crime event. As a result, the proposed model delivers relevant evidence for the corresponding crime event with the assistance of the LSTM and forensic investigator. According to the federal rules of evidence [27], relevant evidence is defined as “any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence”.

Performance Metrics

**Precision:** It is the ratio between the number of accurately classified or identified files as the relevant evidence and the total number of files classified as the evidence by the system.

**Recall:** It is the ratio between the number of accurately classified or identified files or evidence as the potential evidence and the total number of files or evidence that are relevant to the incident.

Experimental Results

The experimental results illustrate the performance of the Forensic Evidence Integration using LSTM (FEI-LSTM) for the cloud evidential artifacts with the help of the precision and recall performance metrics.

Inter-Linked Evidence Ratio Vs. Precision

Figure 9 illustrates the precision of the FEI-LSTM for the increased amount of inter-linked evidence ratio. Inter-Linked Evidence Ratio (ILER) is the ratio between the
number of evidence that is related to the evidence in different datacenters and the total number of evidence acquired for a particular crime event. Initially, the precision value increases from 86% to 87.5% for the variation of Inter-Linked Evidence Ratio from 0.2 to 0.6.

After reaching the ILER from 0.6, the proposed FEI-LSTM model maintains the precision value in the average of 87.6% by considering the inherent forensic evidence footprints during the integration and classification of the evidence using the LSTM model. By performing the integration of the distributed evidence, the proposed forensic examination and analysis methodology assists the forensic investigator accurately determine a set of relevant evidence regarding a crime event.

![Figure 9 Inter-Linked Evidence Ratio Vs. Precision](image1)

![Figure 10 Inter-Linked Evidence Ratio Vs. Recall](image2)
The performance of the recall or sensitivity or true positive rate is depicted in Figure 10 while varying the Inter-Linked Evidence Ratio for the LSTM based integration and classification model. The FEI-LSTM model assists the forensic investigator to precisely filter the evidence that is relevant to the crime event based on the potential footprints of the criminal activity, primarily, crime type and the crime occurred timestamp. As a result, the recall value gradually increases with the increase of the Inter-Linked Evidence Ratio from 0.2 to 1.0. The FEI-LSTM model yields the average recall to 86%, even when increasing the number of distributed evidence and multi-modal similar evidence. Moreover, the temporal sequence-based deep learning in the proposed forensic examination and analysis methodology ensures that the balanced recall value of the massive collection of heterogeneous file types in the same crime event and distributed evidence in the cloud.

Conclusion

This paper presented an enhanced mobile cloud forensic examination and analysis model with the incorporation of the essential sub-phases along with the adoption of the data mining techniques. The proposed forensic examination and analysis model has improved the sequential process of the forensic investigation on both the mobile and cloud environments. It has ensured that the improved performance of the forensic decision-making by introducing and enhancing the sub-phases of the forensic examination and analysis phase in the mobile cloud environment. By modeling the data mining-assisted enhanced examination and analysis phase, the proposed methodology ensures the improved performance of the investigation in the mobile cloud. The experimental results show that the performance of the LSTM based evidence integration and relevant evidence identification process through precision and recall.

References


