INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

# A Framework of Forensic Analysis and Visualization: Using WhatsApp Chat Data as a Case Study

Shahnaz Pirzada [a,*], Nurul Hidayah Ab Rahman [a], Niken Dwi Wahyu Cahyani [b], Muhammad Fakri Othman [c]

[a] *Centre of Information Security Research, Universiti Tun Hussein Onn Malaysia, Parit Raja, Johor, Malaysia*
[b] *School of Informatics, Telkom University, Bandung, Malaysia*
[c] *Application and Research on Multimedia, Universiti Tun Hussein Onn Malaysia, Parit Raja, Johor, Malaysia*
*Corresponding author: shahnaz.essa@gmail.com*

***Abstract*—Digital forensic analysis involves studying and analyzing acquired evidence artifacts using methodical approaches. However, unstructured data could be time-consuming and difficult in the forensic examination phase. Automation in digital forensic processes has recently been seen as a potential solution to improve analysis processes. Therefore, we propose a forensic analysis and visualization framework via exploratory data analysis (EDA) using WhatsApp chat datasets as a case study. Univariate and multivariate EDA visualization models were applied to the datasets. The framework's utility was demonstrated through forensic analysis simulation scenarios: linkage (interaction) and attribution (who was responsible). origination (evaluation of source), and sequencing (timeline). It was conducted in a controlled experiment environment using Python scripting. The aim is to test the extent to which EDA visualization models can visualize complete and accurate artifacts based on the scenarios. Our evidence-based findings demonstrated the suitability of specific univariate and multivariate in visualizing complete and accurate data. The framework was able to visualize key metadata such as incoming and outgoing chats, sender identification, communication timeline, and shared media. The findings suggested that the EDA approach aligns with forensic analysis, as it helps describe investigative clues by analyzing data patterns. Additionally, an expert review was conducted, in which the experts confirmed the adequacy of the simulation scenarios and the usefulness of the forensic visualization. Furthermore, the results of this study could aid in presenting evidence in a court of law.**

***Keywords*—Forensic analysis; forensic visualization; instant messaging apps; mobile forensics; mobile communication apps.**

*Manuscript received 3 Jul. 2024; revised 4 Aug. 2024; accepted 3 Oct. 2024. Date of publication 30 Nov. 2024.*


## I. INTRODUCTION

WhatsApp is a widely used communication platform that enables users to share text, images, videos, and audio files across Android, iOS, and Windows devices. As such communication applications are adopted, they increasingly attract cybercriminals who exploit these platforms for illegal activities. The emergence of digital device exploitation by criminals has resulted in more seizures of digital evidence to assist investigations, including forensic analysis activities. Forensic analysis is one of the phases in digital forensics, which involves activities to study and analyze acquired evidence artifacts using methodical approaches for drawing conclusions [1]. However, the growth in the volume of data seized has led to digital forensic backlogs, which have become a notable challenge to the efficiency of forensic analysis activities.

WhatsApp chat data primarily comprises unstructured raw text messages, multimedia files, and various other forms of communication exchanged between users. This data includes various elements such as text conversations, timestamps, sender and recipient information, media files like images and videos, and associated metadata for each message [2]. Users generate a continuous flow of unstructured data as they engage with the platform, complicating analysis and interpretation. The informal nature of WhatsApp messaging further complicates this process, as chats often include slang, acronyms, emoticons, and other non-standard language elements. However, the form of unstructured data or massive data could lead to difficulty in the forensic examination phase as well as time-consuming. Although investigators usually conduct cross-analysis between different forensic software tools, they need advanced visualization techniques to facilitate evidence analysis [3]. Consequently, extracting valuable insights from this unstructured data requires advanced methods to handle such variability. Therefore, the challenge lies in analyzing and interpreting unstructured forms of evidence data. This necessitates the development of

techniques to assist investigators and the judicial system in interpreting evidence metadata.

Automation in digital forensics processes has recently been a potential solution to improve analysis processes. Although automating the investigation activities is not feasible, automation could facilitate analysts in drawing investigation hypotheses. As reviewed by [4], integrating intelligent automation could facilitate quickly identifying trends and anomalies, thus minimizing investigation time and resources. Similarly, [5] highlighted that automated tools, data science, and artificial intelligence approaches could expedite investigations and maximize forensic analysis's accuracy, efficiency, and cost-reduction.

With a focus on big data analytics, data science encompasses many tasks in processing and analyzing data. A branch in data science, Exploratory Data Analysis (EDA), is a first analysis of data to determine the relationships between data measurements. With the use of statistics and visualization tools, EDA can be applied to collect insights on the trends and patterns of various variables within a dataset, as well as their relationships [6]. As classified in [7], EDA methods can be non-graphical or graphical, with each further categorized as univariate (i.e., one variable) or multivariate (i.e., several variables). Various visualization approaches have been developed, including histograms, pie charts, bar charts, scatter plots, and box plots, to deliver the most efficient approach to visualize a particular set of data. Furthermore, combining specific types of statistics with appropriate methods of visualization is significant to supply data scientists readily with needed data [8].

With the emergence of data science, the amount of research conducted on WhatsApp chat analysis has considerably risen over the years, mainly due to the app's popularity as an instant messaging platform. A study by [9] demonstrated a tool that extracts chat database data into a visual form for forensic analysis purposes. This visualization can show the connection of metadata, text frequency, and word count and display a report of analysis activities. However, the study lacks a discussion of the theoretical aspects, focusing only on the application of information visualization. Another limitation is that the timestamp format was presented in epoch format rather than a more user-friendly, human-readable datetime format. This makes it challenging for users to understand and interpret the timing of chat data effectively.

Demonstrating WhatsApp Chat Analyzer tool [10], the study used data preprocessing, statistical analysis, and visualization processes. The tool showed statistical data such as total messages and temporal trends indicating peak activity periods. It also created visualized activity maps. However, the study lacked a detailed discussion of the method. For instance, it did not clearly explain the visualization model used, and there was no discussion of the evaluation approach.

In a study by [11], WhatsApp analytics data was used to describe and visualize students' collaborative trends. The visualization involved a three-pronged approach, which included network analysis, process-oriented analysis, and content-oriented analysis. However, the authors should have provided detailed information about data preprocessing, apart from mentioning raw data processing and converting text files to data frames. Additionally, the study utilized box plots, scatter plots, and word clouds as visualization models.

Focusing on sentiment analysis technique, a study by [12] used the technique to determine whether a conversation in a WhatsApp group is positive or negative. The study utilized the "dplyr" package in the R programming language to gather a set of verbs. The sentiment analysis results were visualized using bar charts representing anger, joy, surprise, and positivity. However, the explanation of the evaluation approach is ambiguous. Another similar study by [13] integrated text classification and sentiment analysis to determine the relevance of views, opinions, or emotions expressed within the WhatsApp group. This involved data collection, transformation, exploration, and visualization. Ranjan et al. [14] utilized Python libraries to present an analysis of WhatsApp chat trends, such as the most active users, the most popular terms, and the busiest chat time. Similarly, the study involves data preprocessing, analysis, and visualizing the data.

Previous studies highlighted several similarities, such as using typical phases like data preprocessing, analysis, and visualization. However, there needs to be more discussion on the theoretical aspects of visualization. From our perspective, the typical phases and visualization align with exploratory data analysis (EDA). This is evidenced by applying visualization models and the statistical analysis the studies integrate. Our observation showed that most of these studies are unrelated to digital forensics investigations. However, it is worth noting that the approach can be relevant in forensic analysis, such as finding clues of criminal activities from chat data. Therefore, this is the research gap that we sought to fill by examining exploratory data analysis to analyze and visualize WhatsApp chat data for digital forensic analysis.

Furthermore, in a recent work [15], we conducted a literature survey highlighting forensic visualization's role in understanding data and interpreting evidence. The survey showed that forensic visualization has been applied in various computing case studies, such as the Internet of Things, malware, and mobile applications. This indicates the broad applicability of forensic visualization in investigations of different digital infrastructures. We observed that the most common evaluation method includes performing technical simulations to assess tool usability, then conducting user testing to evaluate the tool and using a focus group to solve forensic challenges. While many studies have made progress in this area, there has been limited discussion on the selection criteria of forensic visualization techniques for forensic analysis. This is another gap that we aim to address in this study.

Therefore, we proposed a forensic analysis and visualization framework using exploratory data analysis (EDA). EDA entails examining data in a variety of ways to gain insights and applying data-specific domain knowledge in the analysis [16]. In this study, WhatsApp chat datasets were applied as a case study.

The objectives of this study are three-fold:

a. To propose an evidence-based forensic analysis and visualization framework using exploratory data analysis for WhatsApp chat data.
b. To design forensic visualization models suitable for different types of WhatsApp chat data and forensic analysis scenarios

c. To evaluate the utility of the proposed framework using completeness and accuracy metrics, as well as expert evaluation feedback.

Theoretical contributions include presenting a generic forensic analysis framework integrated with visualization models, which will be useful for academics and researchers. Practically, this study will examine automated forensic analysis tools with the specific requirements of forensic investigators.

## II. Materials and Methods

This section outlines the study's detailed approach, describing the Exploratory Data Analysis (EDA) and the proposed framework. Furthermore, it describes the experimental setup and the technical simulations conducted, highlighting how they were designed to test the framework's utility and validate the results.

### A. Exploratory Data Analysis

EDA was adopted as a critical step in describing and visualizing evidence of interest to assist forensic analysts in generating hypotheses. As classified in [7]EDA methods can be non-graphical or graphical, each further categorized as univariate (i.e., one variable) or multivariate (i.e., several variables). Non-graphical EDA methods provide insights into the characteristics and distribution of variables of interest, while graphical EDA methods visualize variables of interest. Table 1 summarizes potential EDA techniques that can be applied to EDA methods.

Exploratory data analysis includes three components: (1) understanding the variables in a dataset, (2) cleaning the dataset, and (3) analyzing the relationships between the variables [17]. Understanding variables involves performing preliminary analyses after importing all the programming libraries required for analysis. Cleaning the dataset includes removing replicate variables, incorrect formats, and null values to ensure data quality before analysis activities. One of the fastest ways to analyze the relationships between variables is to visualize them, for example, using a correlation matrix and a heatmap.

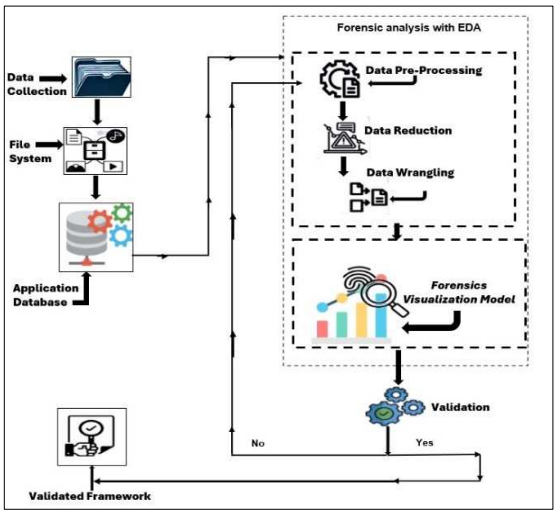



Fig. 1 Proposed framework for forensic analysis visualization

TABLE I
EDA techniques

| Method | | Potential EDA techniques |
|---|---|---|
| Non-graphical | Univariate | Tabulation of categorical data, characteristics of quantitative data: central tendency, spread, and shape of distribution |
| | Multivariate | Cross-tabulation, covariance, and correlation |
| Graphical | Univariate | Histograms, stem plots, boxplots, 2D line plots, probability plots |
| | Multivariate | Side-by-side boxplots, scatterplots, curve fitting, heatmaps, 3D surface plots |

### B. The Proposed Evidence-based Forensic Visualization Framework

This study aimed to demonstrate an evidence-based forensic visualization framework that uses the EDA approach. The proposed framework is presented in Fig. 1. The data collection involves gathering relevant WhatsApp chat data from public repositories. After collecting the data, it undergoes data cleaning, reduction, and wrangling. The data is then transformed into a suitable format for analysis. The forensic visualization model is a crucial step that includes the selection of EDA techniques that are based on the WhatsApp chat data type and/or analysis objective [7]. The selection involves simulating a forensic analysis scenario using a proof-of-concept prototype. The scenarios comprised linkage, attribution, origination, and sequencing [18]The evaluation metrics for visualizing each scenario are completeness and accuracy. The simulation results are compiled into a test plan, which was then validated by an expert review.

*1) Data collection*: This study used two publicly available Android disk images as datasets. The first Android disk image was downloaded from Digital Corpora, a widely used research corpus of actual data in computer forensic education research [19]. The second disk image was downloaded from Dataset for Cyber Forensics, a work by [20] comprising a collection of available datasets for digital forensic researchers [21].

The following steps were conducted to obtain raw data of WhatsApp databases from the Android disk images and after extracting related databases and tables:

- Unzip each Android disk image to extract WhatsApp databases. The databases can be extracted from the following directory path: Android system/data /data/com. WhatsApp/database/…. Fig. 2 presents the screenshot of some of the extracted WhatsApp databases.
- Use DB Browser for SQLite to examine all databases and tables.
- Navigate the directory.

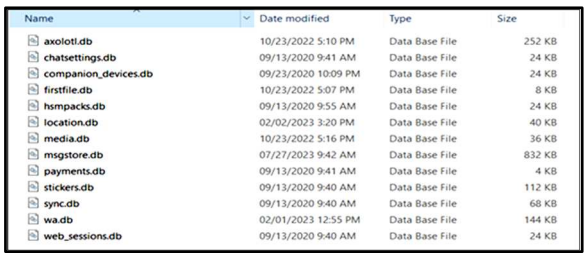



Fig. 2 Examples of extracted WhatsApp database files

Our observation found that the key metadata of the chat history was stored in the msgstore.db file for both datasets, specifically in a massage table with the number of columns and rows shown in Table 2.

TABLE II
SUMMARY OF DATASETS

| Dataset Source | | Smartphone OS | Number of Rows | Number of Columns | Attributes |
|---|---|---|---|---|---|
| Dataset 1 | Digital Corpora (https://digitalcorpora.org/corpora/cell-phones/) | Android 10 | 19 | 40 | Null and non-null attributes |
| Dataset 2 | Datasets for Cyber Forensics (https://datasets.fbreitinger.de/datasets/) | Android 11 | 11 | 40 | Null and non-null attributes |

Dataset 1 (D1) had 40 columns and 19 rows, while Dataset 2 (D2) had 40 columns and 11 rows. The number of rows referred to the number of chats, while the number of columns denoted the number of attributes.

*2) Data pre-processing*: We converted the message tables into a CSV file format by using MySQLite. Next, using DB Browser for SQLite, we examined the WhatsApp databases and named the CSV file 'DCmessageCSV.csv.'

The next step was reading the data entries to reference the dataset summary quickly. We applied Python Panda's pd.read_csv function to read the datasets (Fig. 3).

```
BEGIN
// Import the panda's library for data manipulation IMPORT pandas AS pd
// Import the pyplot module from matplotlib for plotting IMPORT pyplot AS plt
// Load data from a CSV file into a DataFrame READ CSV file 'C:\python\Lib\site-packages\pandas\io\DCmessages10.csv' INTO df
// Print descriptive statistics of the DataFrame to the console PRINT descriptive statistics OF df
// Define a function named csv that currently does nothing FUNCTION csv RETURN None END
```

Fig. 3 Pseudocode of reading the dataset files

The parameters obtained from the results were the range index, data column, non-null count, and data type (see Fig. 4a and Fig. 4b). The range index parameter represents the number of rows, which indicates the total number of chat interactions. As observed, we can quickly interpret that 19 chat interactions occurred in D1, while 11 chat interactions occurred in D2.

The data column parameter refers to the number of attributes. This helps analysts identify the WhatsApp version and the number of attributes. Based on the results of this study, D1 and D2 used the same WhatsApp version, as given by the same number of attributes (i.e., 40) and the same attribute names.

*3) Data reduction:* In this study, data cleaning involved filtering the WhatsApp attributes by identifying null and non-null values in the datasets. The non-null count parameter indicates the number of rows with values for a specific attribute. Zero non-null indicates an attribute with no value, meaning no interaction occurs for the attribute. From the results, for example, both datasets showed '0 non-null' for 'payment_transaction_id', which indicated that the users did not use or activate the WhatsApp payment feature. Therefore, a data summary helps analysts understand attributes with many, average, few, and no interactions.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18 entries, 0 to 17
Data columns (total 40 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   _id                         18 non-null     int64
 1   key_remote_jid              18 non-null     int64
 2   key_from_me                 18 non-null     int64
 3   key_id                      18 non-null     object
 4   status                      18 non-null     int64
 5   needs_push                  18 non-null     int64
 6   data                        3 non-null      object
 7   timestamp                   18 non-null     float64
 8   media_url                   8 non-null      object
 9   media_mime_type             7 non-null      object
 10  media_wa_type               18 non-null     int64
 11  media_size                  18 non-null     float64
 12  media_name                  2 non-null      object
 13  media_caption               2 non-null      object
 14  media_hash                  8 non-null      object
 15  media_duration              18 non-null     int64
 16  origin                      18 non-null     int64
 17  latitude                    18 non-null     float64
 18  longitude                   18 non-null     float64
 19  thumb_image                 11 non-null     object
 20  remote_resource             6 non-null      object
 21  received_timestamp          18 non-null     float64
 22  send_timestamp              18 non-null     int64
 23  receipt_server_timestamp    18 non-null     float64
 24  receipt_device_timestamp    18 non-null     float64
 25  read_device_timestamp       6 non-null      float64
 26  played_device_timestamp     0 non-null      float64
 27  raw_data                    0 non-null      float64
 28  recipient_count             18 non-null     int64
 29  participant_hash            0 non-null      float64
 30  starred                     0 non-null      float64
 31  quoted_row_id               18 non-null     int64
 32  mentioned_jids              0 non-null      float64
 33  multicast_id                0 non-null      float64
 34  edit_version                18 non-null     int64
 35  media_enc_hash              2 non-null      object
 36  payment_transaction_id      0 non-null      float64
 37  forwarded                   18 non-null     int64
 38  preview_type                18 non-null     int64
 39  send_count                  6 non-null      float64
types: float64(16), int64(14), object(10)
emory usage: 5.8+ KB
```

(a)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11 entries, 0 to 10
Data columns (total 40 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   _id                         11 non-null     int64
 1   key_remote_jid              11 non-null     object
 2   key_from_me                 11 non-null     int64
 3   key_id                      11 non-null     object
 4   status                      11 non-null     int64
 5   needs_push                  11 non-null     int64
 6   data                        2 non-null      object
 7   timestamp                   11 non-null     float64
 8   media_url                   2 non-null      object
 9   media_mime_type             1 non-null      object
 10  media_wa_type               11 non-null     int64
 11  media_size                  11 non-null     float64
 12  media_name                  2 non-null      object
 13  media_caption               2 non-null      object
 14  media_hash                  2 non-null      object
 15  media_duration              11 non-null     int64
 16  origin                      11 non-null     int64
 17  latitude                    11 non-null     float64
 18  longitude                   11 non-null     float64
 19  thumb_image                 7 non-null      object
 20  remote_resource             0 non-null      float64
 21  received_timestamp          11 non-null     float64
 22  send_timestamp              11 non-null     int64
 23  receipt_server_timestamp    11 non-null     float64
 24  receipt_device_timestamp    11 non-null     int64
 25  read_device_timestamp       6 non-null      float64
 26  played_device_timestamp     1 non-null      float64
 27  raw_data                    0 non-null      float64
 28  recipient_count             10 non-null     float64
 29  participant_hash            0 non-null      float64
 30  starred                     0 non-null      float64
 31  quoted_row_id               10 non-null     float64
 32  mentioned_jids              0 non-null      float64
 33  multicast_id                0 non-null      float64
 34  edit_version                10 non-null     float64
 35  media_enc_hash              2 non-null      object
 36  payment_transaction_id      0 non-null      float64
 37  forwarded                   10 non-null     float64
 38  preview_type                10 non-null     float64
 39  send_count                  5 non-null      float64
dtypes: float64(21), int64(9), object(10)
memory usage: 3.6+ KB
```

(b)

Fig. 4 Results of read data from (a) D1 and (b) D2

The data type parameter refers to the data type of the attribute. There are three types of data: int64, float64, and object. Int64 and float64 represent attributes with numeric values, whereas other data types (e.g., text) are represented by objects. For non-graphical EDA, numeric values are examined using statistical analysis.

Some attributes may be removed if they do not provide a value or are insignificant for analysis. From the result, the following attributes were removed from D1 and D2: 'played_device_timestamp', 'raw_data', 'participant_hash', 'starred', 'mentioned_jids', 'multicast_id', and 'payment_transaction_id'. In addition, 'remote_resource' was removed from D2. Since there was no data available for these attributes, there was no need for forensic analysis.

*4) Data wrangling*: While some attributes are excluded, others may contain challenging data to understand. Some data may have data entry errors, and some may need data type conversion. In this study, we modified two attributes in the original data: 'timestamp' and 'key_remote_jid'.

We adjusted the 'timestamp' attribute to convert from epoch time to human-readable date and time formats. Improving the timestamp representation is essential to enhance the tool's usability, providing a more accessible and user-friendly experience for forensic investigators and analysts.

```
➢ START
STEP 1 → Convert timestamp from milliseconds to seconds
  • IF timestamp is in milliseconds:
        Go to → Convert timestamp by dividing it by 1000
STEP 2 → Convert timestamp to datetime
  • IF timestamp is not already in datetime format:
        Go to → Convert timestamp to datetime format
STEP 3 → Format datetime to "YYYY-MM-DD HH:MM"
  • IF datetime conversion is successful:
        Go to → Return formatted datetime
  • ELSE:
        Go to → Show error message "Datetime conversion failed"
STEP 4 → Read the CSV file
    IF file is read successfully:
    Go to → Proceed to next step
  • ELSE:
        Go to → CATCH Empty Data Error and show error message "The selected file is
        empty or not a valid CSV file."
        Go to → CATCH Parser Error and show error message "Error parsing the CSV file.
        Please check if it's a valid CSV format."
STEP 5 → Convert the 'timestamp' column to datetime format
  • IF timestamp column exists:
Go to → Apply convert timestamp function to each value in timestamp column
STEP 6 → Split timestamp into separate date and time columns
  • IF timestamp conversion is successful:
        Go to → Convert timestamp to datetime and split into Date and Time columns
  • ELSE:
        Go to → Show error message "Timestamp conversion failed"
STEP 7 → Return the updated Data Frame with Date and Time columns
  • IF all previous steps are successful:
        Go to → Return updated Data Frame
  • ELSE:
        Go to → Show error message "Data Frame update failed"
END
```

Fig. 5 Pseudocode of 'timestamp' attribute conversion

The 'key_remote_jid' attribute contained both text data and contact numbers. It was cleaned by extracting the original data and removing unnecessary text to isolate the contact numbers (see Fig. 6).

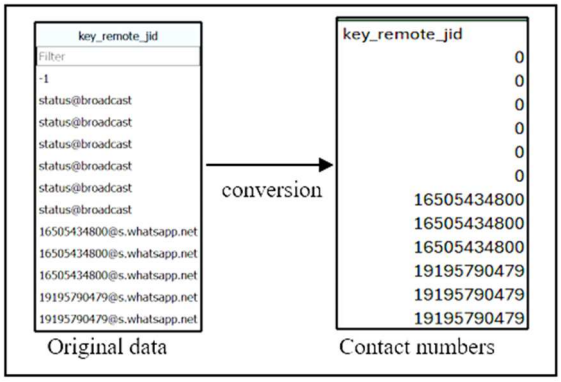



Fig. 6 Conversion of 'key_remote_jid' attribute

*5) Forensic visualization:* WhatsApp attributes are classified into four information visualization data types: relational, text, geospatial, and sequential (see Table 3). We classified them based on the results from data preprocessing and our observation of the message tables.

TABLE III
CLASSIFICATION OF WHATSAPP ATTRIBUTES IN INFORMATION VISUALIZATION

| Data type | Relational | Text | Geospatial | Sequential |
|---|---|---|---|---|
| List of attributes from messages table | status, broadcast, forwarded, send_count, key_ from_me, needs_ push, origin, media_ duration, media_ size, media_wa_type, key_id, Edit_ version, quoted_ raw_id, recipient_ count, message_ type | data media_name media_hash media_enc_hash remote_source media_url media_mime_type media_caption key_remote_ jid text_data | Latitude, longitude | received_ time stamp, send_ time stamp, receipt_ server_ time stamp, receipt_ device_ time stamp, read_ device_ time stamp, time stamp |

Four forensic analysis simulation scenarios, each with sub-scenarios, were designed. The forensic analysis scenarios comprised four scenarios: sequencing (timeline), linkage (interaction), origination (evaluation of source), and attribution (who was responsible). The simulations aimed to test whether the visualization models utilized could visualize the chat data based on the forensic analysis scenario. The simulations were conducted in a controlled experiment environment. Fig. 7 presents the flowchart of the forensic analysis simulation.

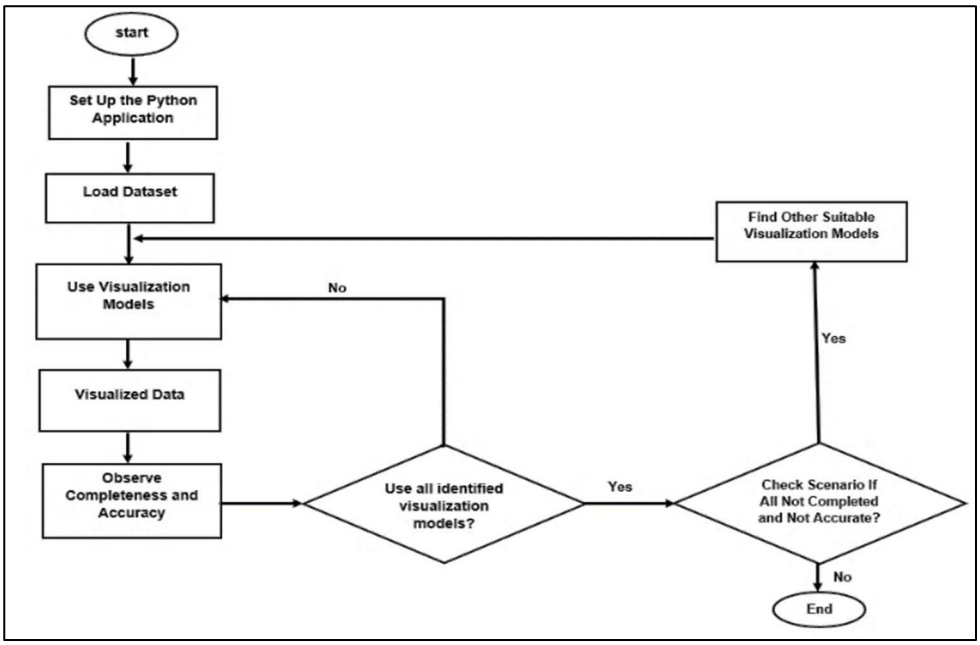



Fig. 7 Flowchart of forensic analysis simulations

The simulation started with preparing a proof-of-concept prototype developed on the Python platform. Using the cleaned datasets, we tested each sub-scenario with the identified visualization model. We chose visualization models based on graphical EDA consisting of univariate and multivariate techniques [7]. We also adopted other visualization models, such as word clouds, geospatial maps, and data tables, from related previous works.

If none of the cases in a particular sub-scenario yielded complete and accurate visualization, we needed to explore additional visualization models and incorporate any new findings into our list. Therefore, it was essential for at least one visualization model to generate a complete and accurate output for each sub-scenario to meet the requirements for forensic analysis and visualization.

*6) Validation:* We conducted expert reviews with the aim of reviewing the test plan and getting feedback on the proposed framework. Semi-structured interview questions were adopted from a study [22] that demonstrated a knowledge-assisted visualization system for malware analysis. The questions were designed to focus on the framework's effectiveness and usability.

Four industry practitioners of digital forensics from government agencies in Malaysia were invited to participate. Three of them have 5 to 10 years of experience in the forensic analysis field, while one has 3 to 5 years of experience (see Table 5).

TABLE V
SUMMARY OF EXPERT REVIEW PARTICIPANTS

| Participant | Position | Years of experience |
|---|---|---|
| P1 | Senior forensic analyst | 5 to 10 years |
| P2 | Senior forensic analyst | 5 to 10 years |
| P3 | Senior forensic analyst | 5 to 10 years |
| P4 | Junior forensic analyst | 3 to 5 years |

Initially, we provided an overview of the prototypes, such as their features and workflow. Next, each participant walked through every simulation scenario result and was interviewed to provide feedback. The interview sessions lasted 90 minutes. The researchers documented the feedback on the paper.

### C. Evaluation metrics

Two metrics were used in this study, which were completeness and accuracy. Completeness refers to the proportion of evidence artifacts present in the forensic image that were identified and reported using the proposed framework [23]. For the completeness metric, three qualitative parameter values were used as follows:

- The value 'full' was considered when all data from the attributes can be visualized,
- 'Partial' referred to partially visualized data, and,
- 'Not available' referred to no visualized data.

Accuracy refers to the proportion of the analysis result from the proposed framework that was correct [23]. For the accuracy metric, two qualitative parameter values were used, namely: (1) as expected (i.e., accurate) and (2) not as expected (i.e., not accurate).

In this study, we presented the results in qualitative, as consistent with widely recognized guidelines, such as Computer Forensic Tool Testing from the National Institute of Standards and Technology (NIST) [24] and the Scientific Working Group on Digital Evidence (SWGDE) [25], as well as guidelines from previous studies in forensic analysis [26], [27].

## III. Results and Discussion

This section presents the results of the statistical summary of WhatsApp chat data, scenario simulations, and expert review. Subsequently, we discuss this study's contributions from theoretical, application, and cyber law perspectives.

### A. Forensic analysis scenario simulations of WhatsApp chats

Samples from each forensic scenario were included to present the results. The results are presented based on the forensic analysis scenarios of linkage, origination, attribution, and sequencing.

*1) Linkage:*

Linkage in WhatsApp forensic analysis involves the interaction between a WhatsApp user or owner and other contacts. We attempted to visualize the interaction of incoming and outgoing chats, text messages, shared media, and shared location. A histogram, a bar chart, and a pie chart were applied to visualize the case scenarios for univariate attributes. Table 6 and Table 7 present our findings.

WhatsApp phone numbers are the key to linking chat activities with the owner. The 'key_remote_jid' attribute was used to identify WhatsApp phone numbers and the total count of interactions. Identifying incoming and outgoing chats can be undertaken by visualizing the 'key_from_me' attribute. Color-coded visualization was applied to distinguish between incoming and outgoing chat interactions. Our observation showed that the pie chart, bar chart, and histogram can entirely and accurately visualize both attributes. Using pie charts, bar charts, and histograms to highlight the data distribution can help forensic analysts simplify and understand the frequency of events.

A word cloud model was applied to visualize and explore the text data type. Three attributes, namely 'data', 'media_name', and 'media_url', were used to test the model. The 'data' attribute provided a word cloud that showed the most frequently used words in the chat. The 'media_name' attribute showed a word cloud of the names of the media that were sent or received in the chat. The 'media_url' attribute displayed a word cloud of the URL links of the shared media. Our observation showed that all attributes were complete and accurate for both datasets.

Media sharing is a crucial feature frequently used by WhatsApp users. Examining shared media is another scenario for forensic analysts to explore to understand the pattern or identify the presence of investigated press as evidence of interest. Up to the time of this research, WhatsApp can support media sharing in various formats, such as audio, video, location, documents, and contact data. We examined the 'media_wa_type' attribute to visualize the shared media types. The result showed that the pie chart and bar chart met the accuracy and completeness evaluations. However, the histogram was incomplete and inaccurate due to the absence of data points; therefore, a histogram does not seem suitable.

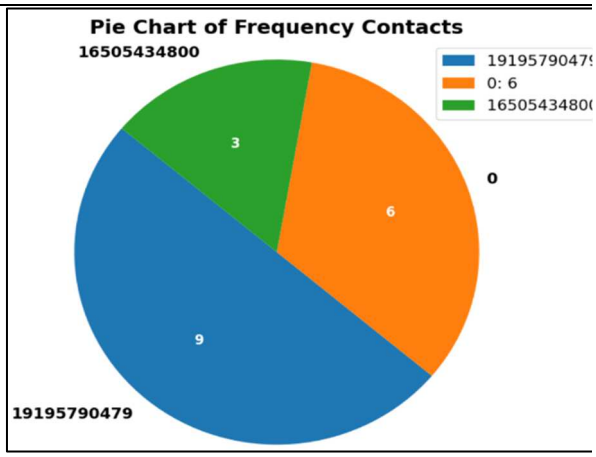
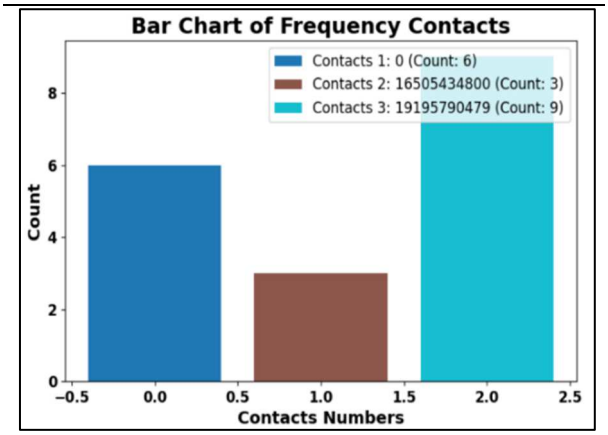
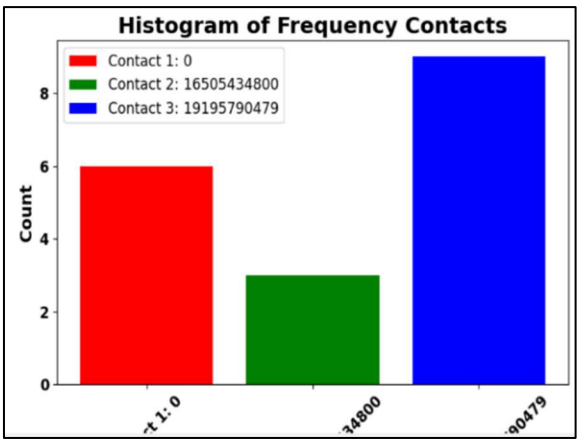
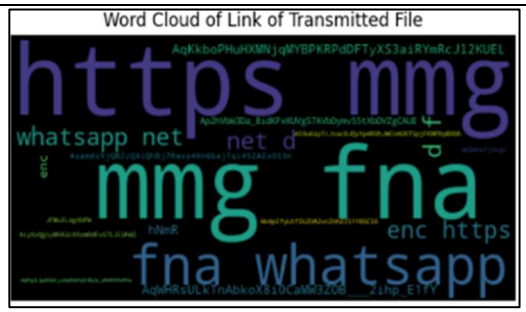
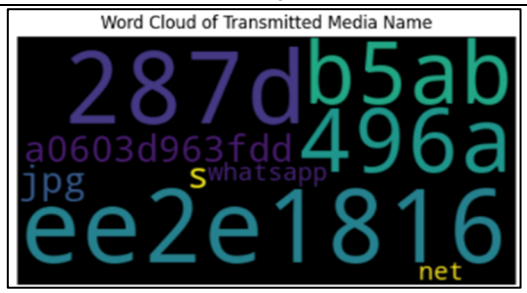
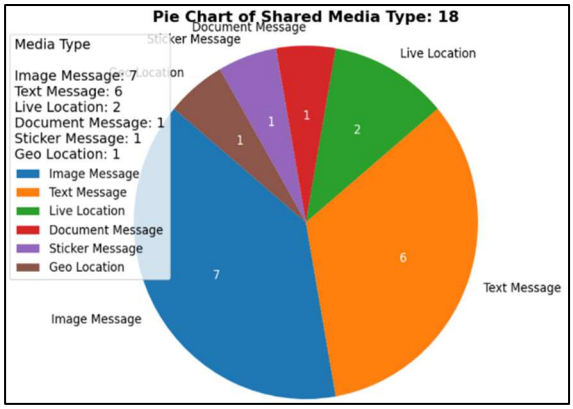
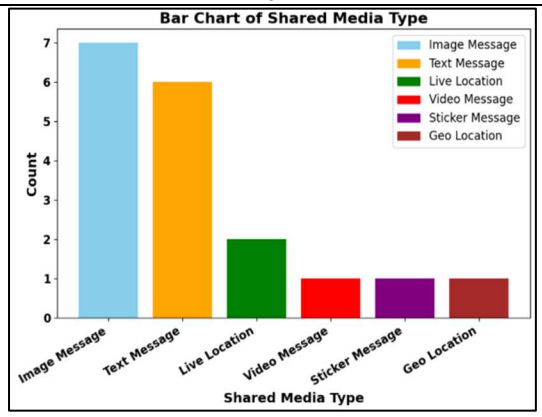
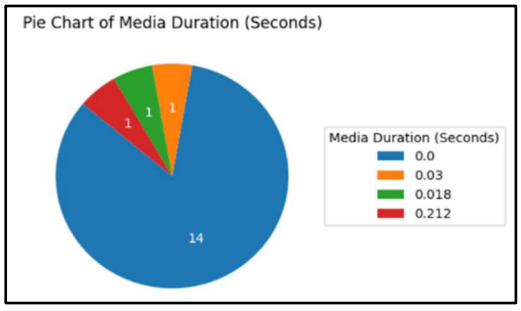
The 'media_duration' attribute can be used to examine the duration of shared media in audio and video formats. Its data are distributed based on the duration range (in seconds). This study's pie chart accurately and completely visualized the shared media durations for D1 and D2. However, the bar chart for both datasets was incomplete and inaccurate. On the other hand, while the histogram was also incomplete and inaccurate for D1, it was complete and accurate for D2. Therefore, the pie chart was the only model with consistent results for both datasets.

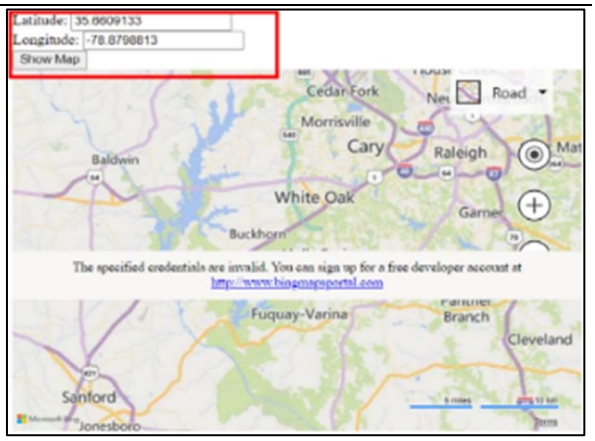
Geospatial data is another interesting piece of evidence found in WhatsApp chat data. It includes shared and live location data. We used the geospatial map to visualize the location map using 'latitude' and 'longitude' attributes. Accurate and complete geospatial maps were observed for both D1 and D2.

TABLE VI
RESULT OF LINKAGE

| **Type of data / Attribute** | **A test case of forensic visualization** | **Case no.** | **Visualization model** | **Completeness (Full, partial, N/A)** | **Accuracy (As expected, not as expected)** |
|---|---|---|---|---|---|
| Relational / key_remote_jid | To display incoming and outgoing contact numbers | A.1.1 | Pie chart | D1, D2: Full | D1, D2: As expected |
| | To display incoming and outgoing contact numbers | A.1.2 | Bar chart | D1, D2: Full | D1, D2: As expected |
| | To display incoming and outgoing contact numbers | A.1.3 | Histogram | D2, D1: full | D1, D2: As expected |
| Relational / key_from_me | To display total number of incoming and outgoing chats | A.1.4 | Pie chart | D1, D2: Full | D1, D2: As expected |
| | To display total number of incoming and outgoing chats | A.1.5 | Bar chart | D1, D2: Full | D1, D2: As expected |
| | To display total number of incoming and outgoing chats | A.1.6 | Histogram | D1, D2: Full | D1, D2: As expected |
| Text / data | To display frequent words in chats | A.2.2 | Word cloud | D1, D2: Full | D1, D2: As expected |
| Text / media_name | To display media name sent or received by user | A.2.9 | Word cloud | D1, D2: Full | D1, D2: As expected |
| Text / media_url | To display links of text communication | A.2.5 | Word cloud | D1, D2: Full | D1, D2: As expected |
| Relational / media_wa_type | To display shred media type | A.3.1 | Pie chart | D1, D2: Full | D1, D2: As expected |
| | To display shred media type | A.3.2 | Bar chart | D1, D2: Full | D1, D2: As expected |
| | To display shred media type | A.3.3 | Histogram | D1, D2: N/A | D1, D2: Not as expected |
| Relational / media_duration | To display duration of media | A.4.4 | Pie chart | D1, D2: Full | D1, D2: As expected |
| | To display duration of media | A.4.5 | Bar chart | D1, D2: N/A | D1, D2: Not as expected |
| | to display duration of media | A.4.6 | Histogram | D1: N/A D2: Full | D1: Not as expected; D2: As expected |
| Geo-spatial / latitude and longitude | to display Google map of user's location | A.5.1 | Geospatial map | D1, D2: Full | D1, D2: As expected |

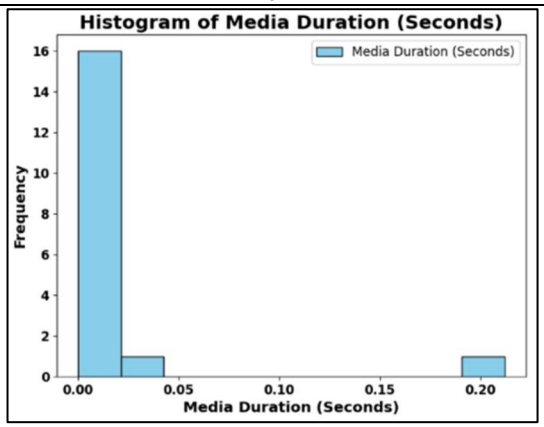
TABLE VII
VISUALIZATION RESULTS OF LINKAGE

| | |
|---|---|
| Full & Accurate |  A.1.1 |
| |  A.1.2 |
| |  A.1.3 |
| |  A.2.2 |
| |  A.2.5 |
| |  A.2.9 |
| |  A.3.1 |
| |  A.3.2 |
| |  A.4.4 |

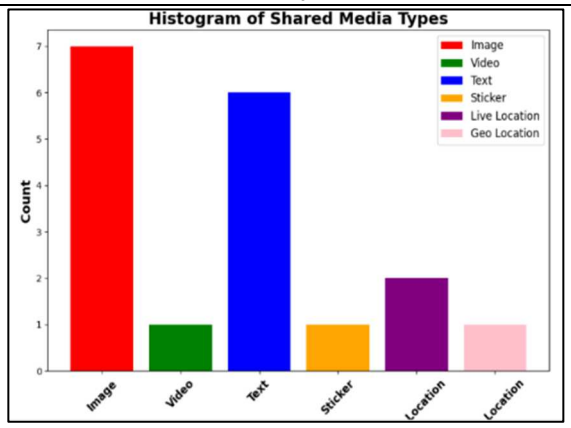



A.5.1

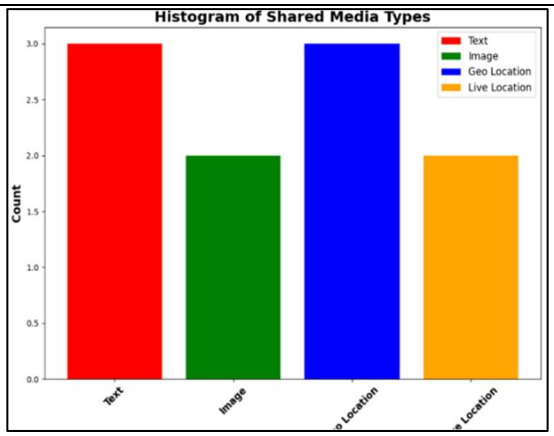
N/A & Not as expected



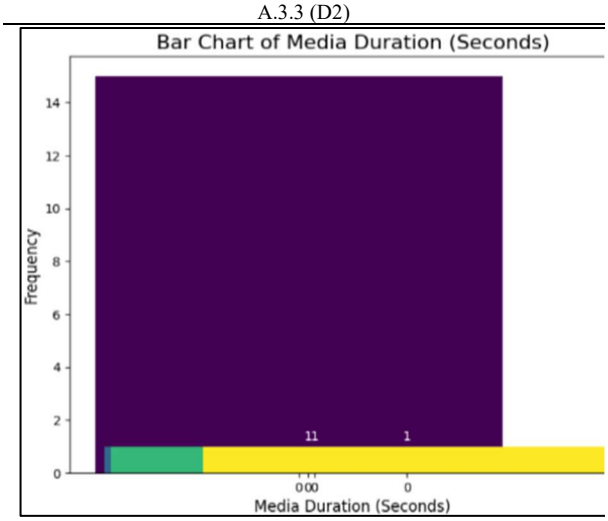

A.4.6




A.3.3 (D1)




A.3.3 (D2)




A.4.6

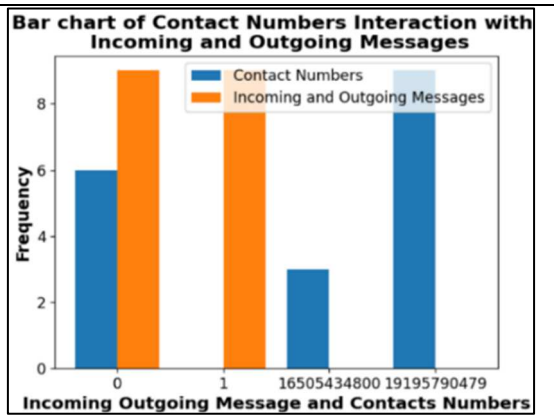
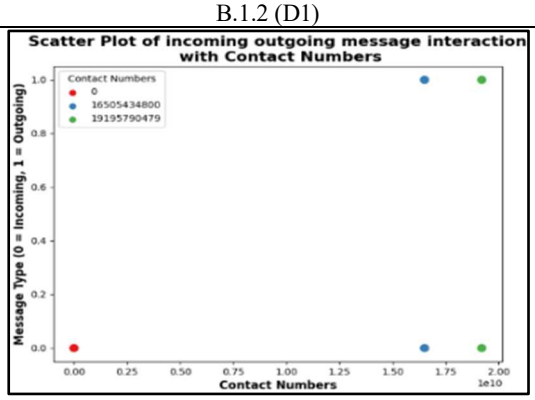
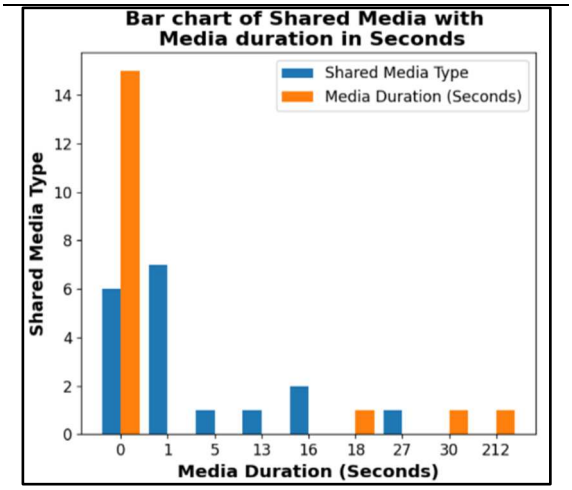
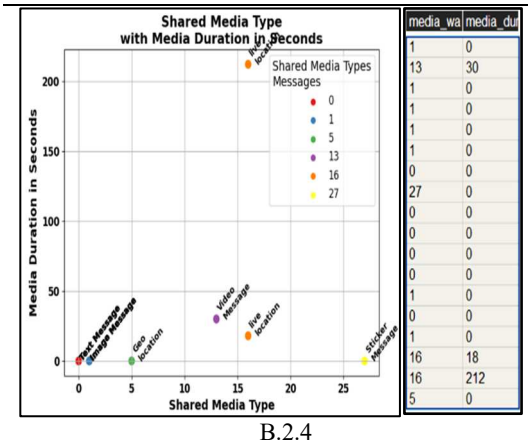
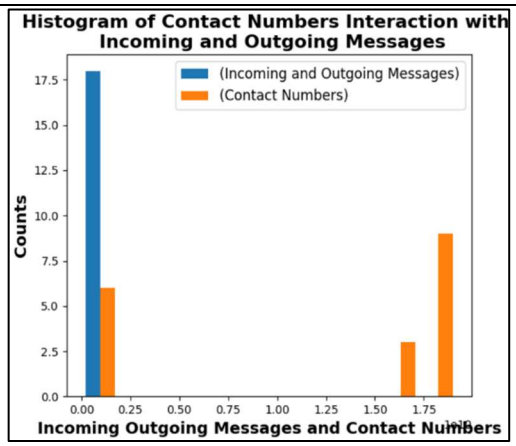
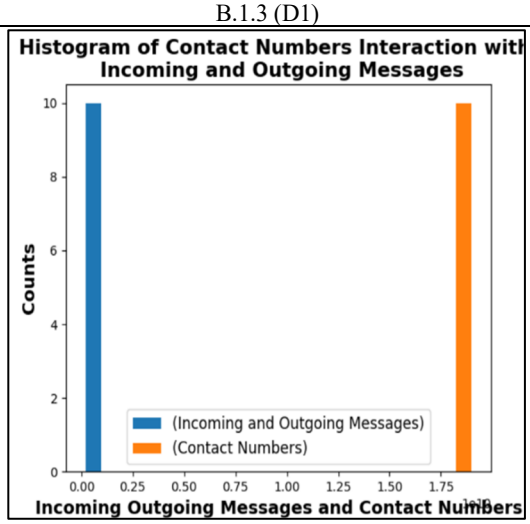
*2) Origination and attribution:*

This study combined the forensic scenarios of origination and attribution since both are similar. These scenarios involved understanding the source of the interaction and who handled the communication actions. Table 8 and Table 9 summarize the results for both scenarios. Multivariate analysis was applied for the scenarios, allowing interactions between two or more variables. In addition, a scatter plot was employed due to its capability to explore more than two variables.

TABLE VIII
RESULT OF ORIGINATION AND ATTRIBUTION SCENARIOS

| Type of data / Attribute | A test case of forensic visualization | Case no. | Visualization model | Complete-ness (Full, partial, N/A) | Accu-racy (As expected, not as expected) |
|---|---|---|---|---|---|
| Relational / key_from_me and key_remote_jid | To display specific contact numbers and related chats | B.1.2 | Bar chart | D1: Full<br>D2: Partial | D1: As expected;<br>D2: Not as expected |
| | To display specific contact numbers and related chats | B.1.3 | Histogram | D1, D2: N/A | D1, D2: Not as expected |
| | To display specific contact numbers and related chats | B.1.4 | Scatter plot | D1, D2: Full | D1, D2: As expected |
| Relational / media_wa_type and media_duration | To display shared media pattern of ‘media_duration’ and ‘media_wa_type’ | B.2.2 | Bar chart | D1, D2: Full | D1, D2: Not as expected |
| | To display shared media pattern of ‘media_duration’ and ‘media_wa_type’ | B.2.4 | Scatter | D1, D2: Full | D1, D2: As expected |

TABLE IX
VISUALIZATION RESULTS OF ORIGINATION AND ATTRIBUTION

| | | |
|---|---|---|
| **Full and As expected** |  <br>B.1.2 (D1) | |
| |  <br>B.1.4 | |
| |  <br>B.2.2 | |
| |  <br>B.2.4 | |
| **N/A and Not as expected** |  <br>B.1.3 (D1) | |
| |  <br>B.1.3 (D2) | |

The attributes 'key_from_me' and 'key_remote_jid' were employed to visualize WhatsApp contact numbers and their incoming or outgoing category. 'key_from_me' was designated as '0' for incoming chats and '1' for outgoing chats. These attributes provide valuable insights into user communication patterns. Our analysis revealed that the scatter plot (Case B.1.4) effectively presented the contact numbers and the chat categories as '0' or '1'. This visualization is valuable for analysts' observing patterns in incoming and outgoing chats associated with specific contact numbers.

Another test was the visualization of shared media and media duration. Similarly, the scatter plot was able to provide meaningful results. In Case B.2.4, we can quickly identify three shared media with durations: two live locations and one video. A data table with the scatter plot was included to provide further information, with color codes and labels distinguishing each media type. On the other hand, the bar chart and histogram could not provide comprehensive visualizations of the two attributes. This scenario demonstrates that while the models can display complete data, they may need to be more accurate for analysis. It highlights the need to evaluate visualization models carefully to ensure meaningful data analysis.

*3) Sequencing:*

Sequencing presents the visualization of timeline interactions. This study focused on analyzing the timeline patterns of chat communications and shared media. The

sequencing involved multivariate analysis, including the 'timestamp' attribute with other attributes to display patterns. Table 10 and Table 11 present the results. We tested the visualization timelines of 'key_remote_jid' and 'key_from_me' attributes to identify the patterns in contact numbers and incoming and outgoing chats in D1 and D2. The findings indicated that neither attribute in the datasets was accurate or complete.

Cases C.2.3–C.2.8 examined the timeline patterns of shared media, which involved 'timestamp' and 'media_wa_type' attributes. Each case presented a specific media type (e.g., text, image, geolocation, live location, sticker message, and document). All cases were accurately depicted in the timeline graphs and met the expected outcome, except for the null value of documents and sticker messages in D2. We used the '_id' attribute (i.e., the sequence number of the record) to label each interaction. Additionally, a table was included to provide key metadata for the chat artifacts.

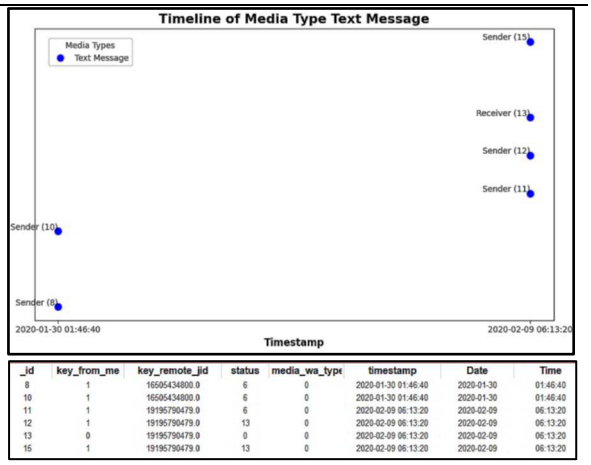
Our findings showed that the scatter plot helped examine an event distribution over time, identifying important activity patterns by referring to the label '_id' in the message table. Additionally, other metadata were displayed in the table. These include 'key_from_me', which indicated whether the communication was incoming or outgoing, and 'key_remote_jid', which displayed the WhatsApp phone number, message status (e.g., '0' for received, '13' for messages opened by the recipient), date, and time. Another important piece of metadata for cyber investigations is the 'status' attribute because it is important to determine if the recipient received and viewed a media.

TABLE X
RESULT OF SEQUENCING SCENARIO

| Type of data / Attribute | A test case of forensic visualization | Case no. | Visualization model | Completeness (Full, partial, N/A) | Accuracy (As expected, Not as expected) |
|---|---|---|---|---|---|
| Sequential / Timestamp (key_remote_jid) | To display timeline of contact numbers from local device | C.1.1 | Timeline graph | D1, D2: N/A | D1, D2: Not as expected |
| Sequential / Timestamp (key_from_me) | To display timeline of incoming and outgoing chats from local device | C.1.2 | Timeline graph | D1, D2: N/A | D1, D2: Not as expected |
| Specific shared media (Text) | To display text messages sent from local device | C.2.3 | Timeline graph with scatter plot | D1, D2: Full | D1, D2: As expected |
| Specific shared media - Image | To display image messages sent from local device | C.2.4 | Timeline graph with scatter plot | D1, D2: Full | D1, D2: As expected |
| Specific shared media (Geolocation) | To display geo-locations sent from local device | C.2.5 | Timeline graph with scatter plot | D1, D2: Full | D1, D2: As expected |
| Specific shared media (Live location) | To display live locations sent from local device | C.2.6 | Timeline graph with scatter plot | D1, D2: Full | D1, D2: As expected |
| Specific shared media (Document) | To display documents sent from local device | C.2.7 | Timeline graph with scatter plot | D1: Full | D1: As expected |
| Specific shared media (Sticker message) | To display sticker messages sent from local device | C.2.8 | Timeline graph with scatter plot | D1: Full | D1: As expected |

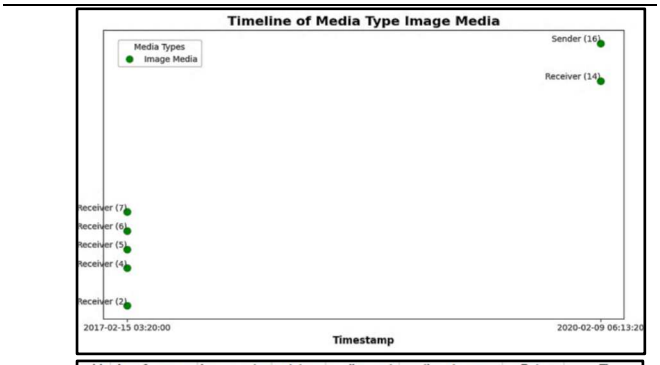
TABLE XI
VISUALIZATION RESULTS OF SEQUENCING

**Full and as expected**



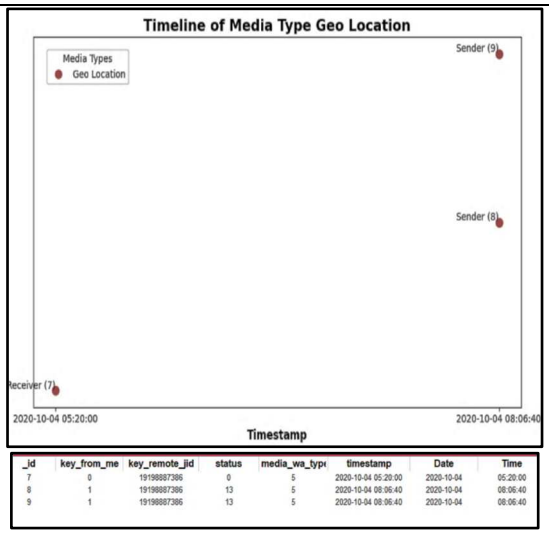

| _id | key_from_me | key_remote_jid | status | media_wa_type | timestamp | Date | Time |
|---|---|---|---|---|---|---|---|
| 8 | 1 | 16505434800.0 | 6 | 0 | 2020-01-30 01:46:40 | 2020-01-30 | 01:46:40 |
| 10 | 1 | 16505434800.0 | 6 | 0 | 2020-01-30 01:46:40 | 2020-01-30 | 01:46:40 |
| 11 | 1 | 19195790479.0 | 6 | 0 | 2020-02-09 06:13:20 | 2020-02-09 | 06:13:20 |
| 12 | 1 | 19195790479.0 | 13 | 0 | 2020-02-09 06:13:20 | 2020-02-09 | 06:13:20 |
| 13 | 0 | 19195790479.0 | 0 | 0 | 2020-02-09 06:13:20 | 2020-02-09 | 06:13:20 |
| 15 | 1 | 19195790479.0 | 13 | 0 | 2020-02-09 06:13:20 | 2020-02-09 | 06:13:20 |

C.2.3



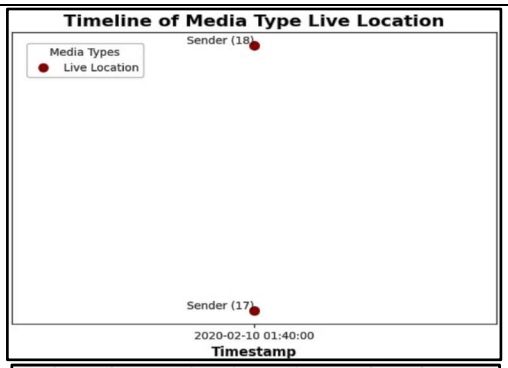

| _id | key_from_me | key_remote_ | status | media_wa_ty | timestamp | Date | Time |
|---|---|---|---|---|---|---|---|
| 2 | 0 | nan | 0 | 1 | 2017-02-15 03:20:00 | 2017-02-15 | 03:20:00 |
| 4 | 0 | nan | 0 | 1 | 2017-02-15 03:20:00 | 2017-02-15 | 03:20:00 |
| 5 | 0 | nan | 0 | 1 | 2017-02-15 03:20:00 | 2017-02-15 | 03:20:00 |
| 6 | 0 | nan | 0 | 1 | 2017-02-15 03:20:00 | 2017-02-15 | 03:20:00 |
| 7 | 0 | nan | 0 | 1 | 2017-02-15 03:20:00 | 2017-02-15 | 03:20:00 |
| 14 | 0 | 19195790479.0 | 13 | 1 | 2020-02-09 06:13:20 | 2020-02-09 | 06:13:20 |
| 16 | 1 | 19195790479.0 | 13 | 1 | 2020-02-09 06:13:20 | 2020-02-09 | 06:13:20 |

C.2.4



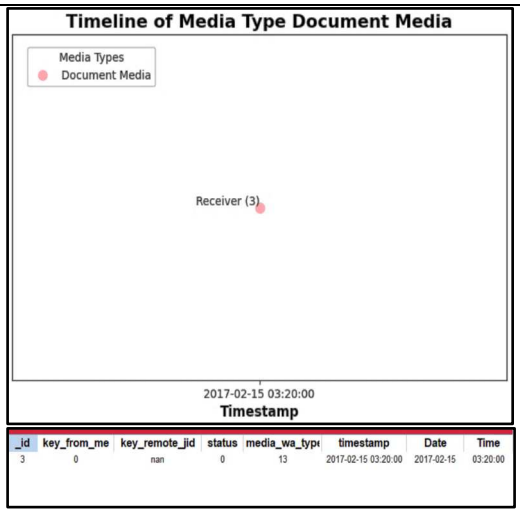

| _id | key_from_me | key_remote_jid | status | media_wa_type | timestamp | Date | Time |
|---|---|---|---|---|---|---|---|
| 7 | 0 | 19198887386 | 0 | 5 | 2020-10-04 05:20:00 | 2020-10-04 | 05:20:00 |
| 8 | 1 | 19198887386 | 13 | 5 | 2020-10-04 08:06:40 | 2020-10-04 | 08:06:40 |
| 9 | 1 | 19198887386 | 13 | 5 | 2020-10-04 08:06:40 | 2020-10-04 | 08:06:40 |

C.2.5



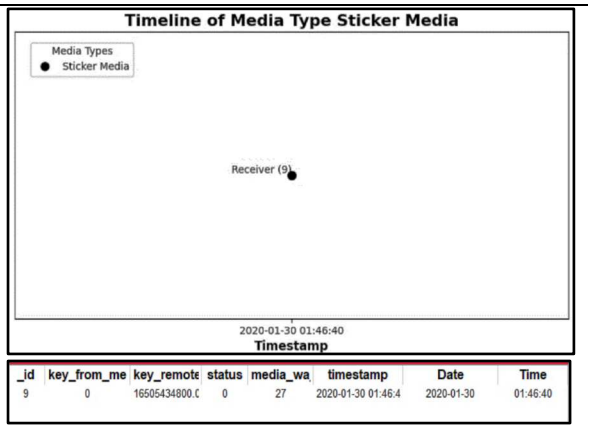

| _id | key_from_n | key_remote_jid | status | media_wa_ty | timestamp | Date | Time |
|---|---|---|---|---|---|---|---|
| 17 | 1 | 19195790479.0 | 13 | 16 | 2020-02-10 01:40:00 | 2020-02-10 | 01:40:00 |
| 18 | 1 | 19195790479.0 | 13 | 16 | 2020-02-10 01:40:00 | 2020-02-10 | 01:40:00 |

C.2.6




| _id | key_from_me | key_remote_jid | status | media_wa_type | timestamp | Date | Time |
|---|---|---|---|---|---|---|---|
| 3 | 0 | nan | 0 | 13 | 2017-02-15 03:20:00 | 2017-02-15 | 03:20:00 |

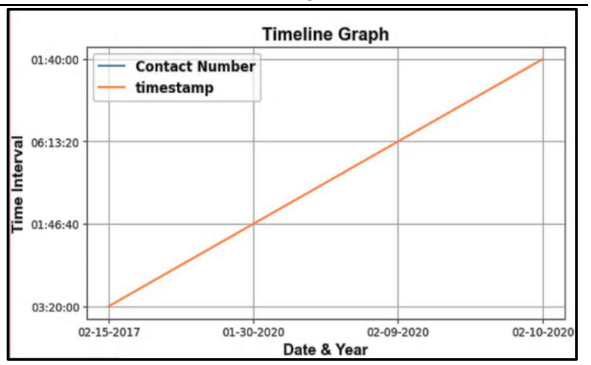
C.2.7



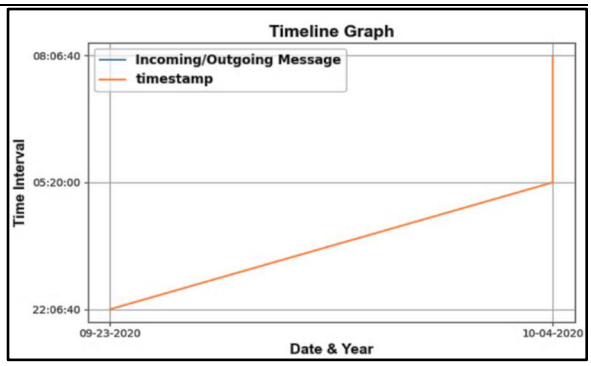

| _id | key_from_me | key_remote | status | media_wa | timestamp | Date | Time |
|---|---|---|---|---|---|---|---|
| 9 | 0 | 16505434800.0 | 0 | 27 | 2020-01-30 01:46:4 | 2020-01-30 | 01:46:40 |

C.2.8

N/A and Not as expected




C.1.1




C.1.2

Our proposed framework supported the forensic analysis scenarios by including the following features:

- Incoming and outgoing chats: Analyzing both sent and received messages.
- Sender identification: Displaying the contact numbers of senders to trace the origin of the communication.
- Communication timeline: Visualizing the timeline of communications to understand the sequence and timing of events.
- Media state: Displaying the current state of the media (e.g., size, duration, received and read status).

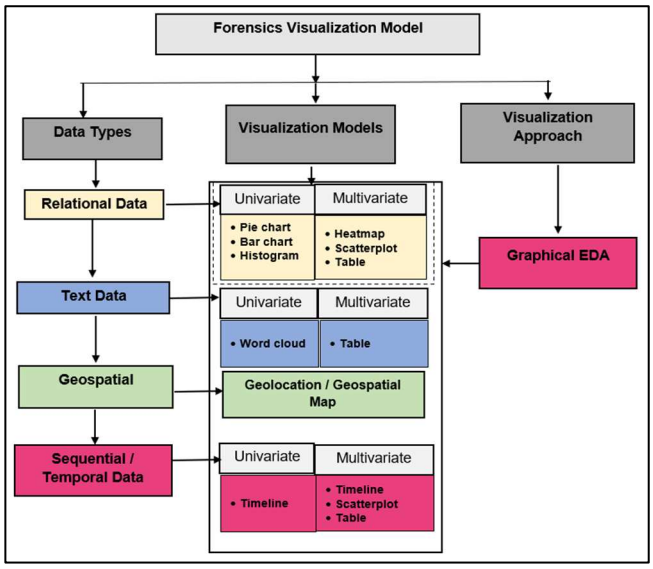



Fig 11 Forensic visualization model

With the exploration of data and the visualization of the evidence of interest, we finalized the following forensic visualization models of instant messaging mobile applications for the proposed framework, using WhatsApp as a case study (see Fig. 11). The finalized visualization model consistently produced complete and accurate outputs for each sub-scenario. It enables the visualization needs for forensic analysis to be met, ensuring that at least one visualization model for each sub-scenario provides the required complete and accurate output. By fulfilling this criterion, the visualization model could be reliably used in forensic contexts involving similar data types where accuracy and completeness are needed.

## B. Expert Review

All participants' statements presented in this section were translated from Malay to English by the authors. The interview result is as follows:

- *Are forensic analysis scenarios (i.e., linkage, origination, attribution, and sequencing) understandable?* All participants unanimously agreed that the forensic analysis scenarios are comprehensive.
- *Are the scenarios of forensic analysis relevant in real-world investigations?* P1–P4 acknowledged that the four forensic analysis scenarios are related to WhatsApp forensic analysis. P1 and P2 added that timeline analysis is the core method in forensic analysis. Usually, investigation officers (IOs) provide a timeline of the occurrence of cybercrime for further analysis of digital evidence. Using a WhatsApp information leak case, P2 explained that the forensic analyst compared the timestamps between when the media was supposed to be legally viewed and when the media was leaked. The difference in the time was used as an evidence clue.
- *Are the scenarios of forensic analysis sufficient for real-world investigations?* P1–P4 agreed that the scenarios will suffice for real-world investigations.
- *Are the evaluation metrics (completeness and accuracy) relevant to industry practitioners (to evaluate tools effectiveness)?* All four participants stated that completeness and accuracy metrics are relevant in industry practices for evaluating digital forensic tools. Furthermore, according to P2 and P3, accuracy is important because the presented data must be correct.
- *Are visualization techniques helpful for forensic analysts?* P1–P4 agreed on the helpfulness of the visualizations. P1 and P2 raised the point that not all attributes require visualization; for example, hash values visualized using a word cloud can be perplexing. Additionally, P3 and P4 recommended that visualizations should be simple, straightforward, and uncomplicated.
- *Do you have suggestions to improve the test plan?* P3 and P4 suggested that including legends in visualization can improve interpretation precision. P1–P4 recommended user testing to identify the extent to which the proposed framework can support investigations.

## C. Research Contributions

This work examined the application of graphical and non-graphical exploratory data analyses in interdisciplinary research areas. The involved areas were data science, information visualization, and digital forensics. We discussed the findings of this study from the viewpoints of data science theory, the application of the framework in forensic visualization, and cyber laws.

### 1) Theoretical:

This study proposed a forensic analysis and visualization framework using the EDA approach. Our findings showed that the EDA approach aligns with forensic analysis, which needs to describe investigation clues by analyzing trends and data relationships. From data pre-processing, investigators can gain an initial review of evidence artifacts through statistical summary and data profiling. The data reduction process is in line with the feedback from the experts, indicating that not all attributes need to be examined in cybercrime investigations involving communication apps. Data reduction is another key point to be highlighted in digital forensic investigations, as many studies, such as those on smart city forensics [28] and those reviewing IoT forensics [29], pointed out the importance of reducing irrelevant data from those to be examined and stored. Furthermore, this can help to address forensic backlog issues.

A key challenge in analyzing mobile communication application data is that it involves several types of data. Attributes from WhatsApp chat data can be categorized into data types to be adopted for information visualization. This study included forensic visualization models based on the data type. Each data type was mapped into potential visualization models adopted from graphical univariate and multivariate EDA. We found that the classification of data types via univariate and multivariate techniques was useful to identify visualization models quickly. This is similar to studies in different fields that applied data visualization to define visualization models using univariate and multivariate techniques, for example, visualization using healthcare data [30] and network data [31]. Graphically EDA is a more natural method than classic descriptive statistics when dealing with non-statistical tables, such as WhatsApp chat data. Furthermore, by leveraging the power of visualization, intricate relationships within the data become readily apparent,

facilitating more informed decision-making and enhancing the efficiency of investigative processes. As reported by studies on big data forensics, visualization is a powerful tool to enhance security and forensic analysis [32], [33].

*2) Application of framework in forensic analysis process:*

In this study, we simulated case scenarios using various data visualization models to examine the models' completeness and accuracy. We found that some WhatsApp data attributes were accurate and aligned with our expectations, confirming the utility of the proposed framework. Our findings presented that the proposed visualization model can present complete and accurate visuals by at least one visualization model for each simulation case scenario. The results validate that the proposed framework can achieve accuracy and completeness in representing artifacts successfully. Completeness and accuracy have been applied in forensic visualization studies to show forensically sound validation, such as those on malware dataset analysis [34], log files analysis [34], and automated forensic analysis (e.g., Android app analysis) [35].

Correlation of metadata is a critical activity that needs further examination from analysts. This study attempted to examine the use of statistical correlation. We found that media size and media duration were positively correlated. Previous research studies have established that media duration, size, and type are key metadata in multimedia forensics. For example, these metadata help detect deepfake media [36]Related forensic analysis scenarios include detecting malware incidents, where media type and size can help determine a malware file. Similarly, in multimedia forensics, media duration and size can be indicators of file authenticity.

When developing automated forensic analysis tools, it is crucial to consider the specific requirements of forensic investigators. Our test plan was designed based on forensic analysis scenarios: linkage, attribution, origination, and sequencing. Findings from the expert review confirmed that the scenarios are sufficient for forensic analysis.

*3) Cyber laws and evidence presentation:*

Forensic investigations should adhere to cyber laws and other related crime laws. In Malaysia, for example, the Communications and Multimedia Act (CMA) 1998[37] regulates the use of communications and multimedia, including the Internet and mobile applications. Section 233 addresses explicitly offensive content, stating that anyone who makes, creates, or initiates the transmission of any offensive content using any application service is considered to have committed an offense. When investigating such incidents, it is vital to consider the following:

- Identification: Tracing the origin of the offensive content to the specific sender.
- Timeline: Establishing when the content was sent and received to understand the sequence of events and communication.
- Read status: Verifying whether the recipient read the content, which can provide insights into the intent and impact of the communication.

Our findings demonstrated the visualization of incoming and outgoing chats and displayed the senders' contact numbers, the communication timeline, and the media's current state. The framework, therefore, can facilitate investigations to examine clues for cyber law enforcement.

Visual representation (e.g., multimedia presentation) is one way to communicate evidence and crime scene information in court proceedings. To be admissible in a court of law, digital evidence must be appropriately presented and validated using forensically sound methods. As highlighted in [38], it is essential to ensure digital evidence is authentic, accurate, complete, and convincing during the gathering and analysis processes to ensure admissibility. In this study, we designed the test plan based on widely recognized guidelines (i.e., SWGDE and NIST). The validation was undertaken in a controlled simulation set, and further validation was conducted via digital forensic expert review. This further indicates that the forensic visualization generated from the framework can support evidence presentation in a court of law.

## IV. Conclusion

In this study, we proposed a forensic analysis and visualization framework using exploratory data analysis. The framework leverages the advantages of data science and forensic analysis to explore digital evidence and extract artifacts, which can facilitate cyber investigations. The proposed framework's evaluation was demonstrated using WhatsApp chat data as a case study through forensic analysis simulations. The simulation test plan involved linkage, origination, attribution, and sequencing scenarios to examine the visualization models' completeness and accuracy in facilitating forensic analysis activities. Subsequently, interviews were conducted with digital forensic experts to validate the proposed framework further.

We identified limitations to this study, which are the dataset variety and the limited interactive feature of the tool. The following are the future research directions based on the present study:

- Validation and assessment of the framework with various communication apps (e.g., WeChat, Telegram) datasets.
- Conduct user acceptance testing with samples of real-world forensic investigative questions involving communication apps.
- Obtain more reviews with experts from other fields, such as data science and law enforcement, to guarantee compliance with regulatory requirements and industry standards.
- Develop the proposed tool with a more advanced and interactive visualization graphical user interface.

## Acknowledgment

This research was supported by the Ministry of Higher Education (MOHE) through the Fundamental Research Grant Scheme (FRGS/1/2020/ICT07/UTHM/03/1). The first author is a PhD student at Universiti Tun Hussein Onn Malaysia. The authors thank the anonymous reviewers for their constructive and generous feedback.

## References

[1] K. Kent, S. Chevalier, T. Grance, and H. Dang, "Guide to Integrating Forensic Techniques into Incident Response," *Natl. Inst. Stand. Technol.*, 2006.

[2] H. Baumeister and C. Montag, *Digital Phenotyping and Mobile*

*Sensing*. 2019.
[3] C. Tassone, B. Martini, and K. K. R. Choo, "Forensic Visualization: Survey and Future Research Directions," in *Contemporary Digital Forensic Investigations of Cloud and Mobile Applications*, Syngress, 2017, pp. 163–184.
[4] A. Jarrett and K.-K. R. Choo, "The impact of automation and artificial intelligence on digital forensics," *WIREs Forensic Sci.*, vol. 3, no. 6, pp. 1–17, 2021.
[5] X. Du *et al.*, "SoK: Exploring the state of the art and the future potential of artificial intelligence in digital forensic investigation," in *ACM International Conference Proceeding Series*, 2020, no. 46, pp. 1–10.
[6] "Univariate Analysis | Exploratory Bivariate and Multivariate Analysis." [Online]. Available: https://www.analyticsvidhya.com/blog/2021/04/exploratory-analysis-using-univariate-bivariate-and-multivariate-analysis-techniques/. [Accessed: 23-Jun-2023].
[7] M. Komorowski, D. C. Marshall, J. D. Salciccioli, and Y. Crutain, *Secondary Analysis of Electronic Health Records*, 1st edn. Cham: Springer Nature, 2016.
[8] C. Vischioni, F. Bove, F. Mandreoli, R. Martoglia, V. Pisi, and C. Taccioli, "Visual Exploratory Data Analysis for Copy Number Variation Studies in Biomedical Research," *Big Data Res.*, vol. 27, p. 100298, 2022.
[9] W. S. Ong and N. H. Ab Rahman, "A Forensic Analysis Visualization Tool for Mobile Instant Messaging Apps," *Int. J. Inf. Commun. Technol.*, vol. 6, no. 2, pp. 78–87, 2020.
[10] S. Yaqub, S. Gochhait, H. A. H. Khalid, S. N. Bukhari, A. Yaqub, and M. Abubakr, "WhatsApp Chat Analysis: Unveiling Insights through Data Processing and Visualization Techniques," *2024 ASU Int. Conf. Emerg. Technol. Sustain. Intell. Syst. ICETSIS 2024*, pp. 862–865, 2024.
[11] F. Duzhin and J. S. Tan, "Analytics for WhatsApp chats: tracking and visualising students' collaboration in project teams," *Int. J. Mob. Learn. Organ.*, vol. 17, no. 1–2, pp. 149–179, 2023.
[12] S. J. Rani, T. N. Prabhu, and J. A. Ida Chellam, "Whatsapp Sentiment Analysis Using R," *2022 3rd Int. Conf. Emerg. Technol. INCET 2022*, pp. 1–4, 2022.
[13] A. Ahmad and M. Abubakar, "Sentiment Analysis and Classification of Asuu Whatsapp Group Post using Data Mining," *J. Confl. Resolut. Soc. Issues*, vol. 1, no. 2, pp. 17–26, 2022.
[14] Ranjan, B. Gupta, V. Kapoor, and D. Bansal, "Analyzing WhatsApp Chat Using Python Libraries," *Proc. 2023 Int. Conf. Intell. Syst. Commun. IoT Secur. ICISCoIS 2023*, pp. 181–184, 2023.
[15] S. Pirzada, N. H. Ab Rahman, N. D. W. Cahyani, and M. F. Othman, "A Survey of Forensic Analysis and Information Visualization Approach for Instant Messaging Applications," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 2, pp. 237–246, 2023.
[16] A. Unwin, "Exploratory Data Analysis," *Int. Encycl. Educ. Third Ed.*, pp. 156–161, 2009.
[17] "An Extensive Step by Step Guide to Exploratory Data Analysis | by Terence Shin | Towards Data Science." [Online]. Available: https://towardsdatascience.com/an-extensive-guide-to-exploratory-data-analysis-ddd99a03199e. [Accessed: 07-Aug-2023].
[18] J. Kizza and F. Migga Kizza, *Digital Evidence and Computer Crime*. 2011.
[19] "Digital Corpora – Producing the Digital Body." [Online]. Available: https://digitalcorpora.org/. [Accessed: 04-Feb-2023].
[20] C. Grajeda, F. Breitinger, and I. Baggili, "Availability of datasets for digital forensics – And what is missing," *Digit. Investig.*, vol. 22, pp. S94–S105, 2017.
[21] "Datasets – Datasets for Cyber Forensics." [Online]. Available: https://datasets.fbreitinger.de/datasets/. [Accessed: 12-Feb-2023].
[22] M. Wagner, A. Rind, N. Thür, and W. Aigner, "A knowledge-assisted visual malware analysis system: Design, validation, and reflection of KAMAS," *Comput. Secur.*, vol. 67, pp. 1–15, 2017.
[23] D. Ayers, "A second generation computer forensic analysis system," *Digit. Investig.*, vol. 6, pp. S34–S42, 2009.
[24] "Computer Forensics Tool Testing Program (CFTT)," 2019. [Online]. Available: https://www.nist.gov/itl/ssd/software-quality-group/computer-forensics-tool-testing-program-cftt.
[25] T. Wu, F. Breitinger, and S. O'Shaughnessy, "Digital forensic tools: Recent advances and enhancing the status quo," *Forensic Sci. Int. Digit. Investig.*, vol. 34, p. 300999, 2020.
[26] M. T. A. Razak, N. H. Ab Rahman, N. D. W. Cahyani, T. X. Hui, and S. K. Taylor, "M-health digital evidence taxonomy system (MDETS): Enabling digital forensics readiness with knowledge sharing approach," *AIP Conf. Proc.*, vol. 2508, no. 1, pp. 020016–1–020016–12, 2023.
[27] C. Anglano, M. Canonico, and M. Guazzone, "Forensic analysis of Telegram Messenger on Android smartphones," *Digit. Investig.*, vol. 23, pp. 31–49, 2017.
[28] Y. C. Tok and S. Chattopadhyay, "Identifying threats, cybercrime and digital forensic opportunities in Smart City Infrastructure via threat modeling," *Forensic Science International: Digital Investigation*, vol. 45. 2023.
[29] M. N. Alam and M. S. Kabir, "Forensics in the Internet of Things: Application Specific Investigation Model, Challenges and Future Directions," *2023 4th Int. Conf. Emerg. Technol. INCET 2023*, pp. 1–6, 2023.
[30] Y. Tong, Y. Cui, L. Jiang, Y. Zeng, and D. Zhao, "Construction, Validation, and Visualization of Two Web-Based Nomograms for Predicting Overall Survival and Cancer-Specific Survival in Elderly Patients with Primary Osseous Spinal Neoplasms," *J. Oncol.*, vol. 2022, p. 21, 2022.
[31] Y. Tong, Y. Cui, L. Jiang, Y. Zeng, and D. Zhao, "Construction, Validation, and Visualization of Two Web-Based Nomograms for Predicting Overall Survival and Cancer-Specific Survival in Elderly Patients with Primary Osseous Spinal Neoplasms," *J. Oncol.*, vol. 2022, 2022.
[32] A. O. Aljahdali; G. Alluhaib; R. Alqarni; M. Alsharef; A. Alsaqqaf, "Big data analysis and forensics." International Journal of Electronic Security and Digital Forensics," *Int. J. Electron. Secur. Digit. forensics(IJESDF)*, vol. 14, no. 6, pp. 579–593, 2022.
[33] J. Najar, M. Tsantekidis, A. Sotiropoulos, and V. Prevelakis, "Enhancing Cyber Threat Hunting: A Visual Approach with the Forensic Visualization Toolkit," *Proc. - 2023 IEEE Int. Conf. Big Data, BigData 2023*, pp. 3035–3042, 2023.
[34] I. Ahmad, M. A. Shah, H. A. Khattak, Z. Ameer, M. Khan, and K. Han, "FIViz: Forensics investigation through visualization for malware in internet of things," *Sustain.*, vol. 12, no. 18, pp. 1–23, 2020.
[35] C. Anglano, M. Canonico, and M. Guazzone, "The Android Forensics Automator (AnForA): A tool for the Automated Forensic Analysis of Android Applications," *Comput. Secur.*, vol. 88, pp. 1–15, 2020.
[36] S. Ferreira, M. Antunes, and M. E. Correia, "Exposing manipulated photos and videos in digital forensics analysis," *J. Imaging*, vol. 7, no. 7, 2021.
[37] "Malaysian communications and multimedia commission act 1998," 2000. [Online]. Available: https://www.mcmc.gov.my/en/legal/acts/communications-and-multimedia-act-1998-reprint-200?nid=2311.
[38] Y.-O. A and B. AD, "Digital Forensics Investigation Jurisprudence: Issues of Admissibility of Digital Evidence," *J. Forensic, Leg. Investig. Sci.*, vol. 6, no. 1, pp. 1–8, 2020.