Asia Pacific Fraud Journal
E-ISSN: 2502-695X, ISSN: 2502-8731
Volume 9, Issue 1 (January-June) 2024
*Asia Pacific Fraud Journal now has been accredited "**SINTA 3**" by Ministry of Research and Technology of The Republic of Indonesia (No SK. 225/E/KPT/2022).*
Available online at: http://apfjournal.or.id/index.php/apf

Advancing Digital Forensic through Machine Learning: An Integrated Framework for Fraud Investigation

✉[1,2]Wishnu Agung Baroto
[1]Doctoral Student, Department of Social and Human Science, Tokyo Institute of Technology Tokyo, Japan
[2]Directorate General of Taxes, Ministry of Finance of the Republic of Indonesia Jakarta, Indonesia

ARTICLE INFORMATION
Article History:
Received July 28, 2023
Revised January 23, 2024
Accepted June 1, 2024

DOI:
10.21532/apfjournal.v9i1.346

ABSTRACT
The rise of cybercrime and cyber-related crime encourages efficient digital forensic investigations more crucial than ever before. Traditional investigation methods can be time-consuming, costly, and resource-intensive, while machine learning algorithms have the potential to reduce the complexity by promoting automation and investigation capabilities. This study begins with an analysis of digital forensics framework using a document analysis methodology. Moreover, exploring current practice and potential implementation of machine learning in digital forensics for fraud investigation is demonstrated through the features of Autopsy 4.15.0, a widely known digital forensics tool. The findings suggest the implementation of a comprehensive digital forensic framework that prioritizes the interpretation phase, with the support of machine learning capabilities. At present, machine learning mainly supports the analysis phase, which happens to be the most time-intensive process of digital forensic investigations. Furthermore, as fraud investigation has a role of fraud detection and prevention, current digital forensics procedures do not support the fraud detection and prevention process, despite the potential for machine learning to support this through pattern recognition.These discoveries are particularly significant in the fight against fraudulent activities, such as tax fraud, data fraud, financial fraud, and asset misappropriation, in the digital age.

Keyword: Digital Forensic, Machine Learning, Fraud Investigation

How to Cite:
Baroto, W. A., (2023). Advancing Digital Forensic Through Machine Learning: An Integrated Framework for Fraud Investigation. *Asia Pacific Fraud Journal*, *9*(1), 1-16. http://doi.org/10.21532/apfjournal.v9i1.346.

✉Corresponding author :
Email: wishnu.ab@gmail.com

1. INTRODUCTION

The massive expansion of the internet and increased use of information and communication technology brought momentous changes in human life. Digital technology has provided us with numerous economic opportunities, such as e-commerce, cryptocurrency, and other digital economy sectors. Unfortunately, these advancements are also in line with the rise of cybercrime and cyber-related crime, as well as other fraudulent activities, such as financial fraud, asset misappropriation, tax fraud, and data fraud. In order to combat fraud, fraud examiners and investigators must thoroughly collect, examine, and analyze all evidence, including digital and electronic data, in addition to traditional physical evidence. This is where digital forensics becomes more critical to combat crime and fraud as a process of investigating and collecting evidence from digital media to be used in a court of law.

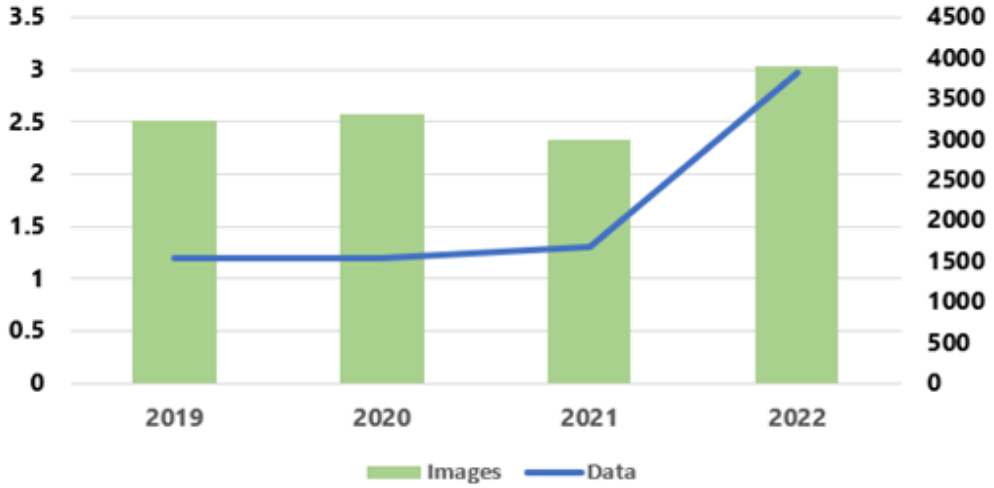
The intersection of traditional crime investigation and digital forensics was highlighted in serial murder cases between 1974 and 1991 in the United States. The BTK (Bind, Torture, Kill) case was solved after the police analyzed metadata of digital evidence in a floppy disk sent by the killer to the police in 2005. In the fraud examination, one famous case was between Ceglia and Zuckerberg, the owner of Facebook. In 2012, Paul Ceglia filed a lawsuit against Mark Zuckerberg, alleging that Zuckerberg had violated an agreement that the two had signed in 2003. Ceglia claims that the agreement stated that he was entitled to 50% of Facebook shares, a claim that Zuckerberg strongly denied. However, the digital forensics team was able to discover evidence that contradicted Ceglia's claims. Specifically, they found that the alleged agreement was forged by Ceglia, who had manipulated the digital footprints of email exchanges between him and Zuckerberg. As a result, the court dismissed Ceglia's lawsuit, and Zuckerberg maintained control over Facebook. This case highlights the importance of digital forensics in resolving legal disputes in the digital era. As more transactions and communications are conducted digitally, it is increasingly important to use digital forensics to investigate the authenticity of evidence presented in court. In this case, the digital forensics team uncovered the truth and prevented a fraudulent claim from being awarded.

In term of data generation, the rapid growth of e-commerce and digital currency has resulted in an exponential increase in electronic data production. This phenomena is commonly referred to as "Big Data." Big Data is characterized by its volume, variety, velocity, and veracity, and more recently, the concept of "Value" has been added as the fifth V to encompass its significance. A significant portion of Big Data consists of unstructured data, lacking predefined data models or schemas, which necessitates specialized procedures for handling and analysis. Consequently, digital investigators not only should possess the expertise to acquire and analyze digital data as compelling evidence in the fight against fraudulent activities but also have to encounter immense volumes of data that necessitate meticulous examination (Platzer et al., 2014).

One of the most well-known institutions that focused on investigating and combating criminal activities related to tax evasion, fraud, and other financial crimes is Internal Revenue Service-Criminal Investigation (IRS-CI). Annually, IRS-CI conducts extensive inquiries into tax-related fraud cases, with a significant reliance on digital forensics techniques for evidence handling. However, the accumulation of forensic data can reach a substantial volume, which requires specialized tools and the application of specialized techniques for extensive data management.

Since data acquired in digital forensics investigation consists of various types and in a large volume, the process of analyzing digital forensic data involves specific tools and the application of specialized techniques for extensive data management

Figure 1. **Forensics Images and Data Acquired of IRS-CI in 2019-2022**



Source: Author Based on IRS-CI Annual Report

(Guarino, 2013). One prominent approach in this domain is machine learning, a branch of artificial intelligence that underlies behavior and pattern analysis. Machine learning enables the development of pattern recognition software capable of handling vast datasets and preemptively preventing criminal activities (Figure 1).

In response to the increasing threat of digital fraud, there is a growing demand for effective investigative methodologies. This research paper searches into the digital forensic process utilized in fraud investigation and investigates the necessity of integrating a machine learning approach. Therefore, the primary objectives of this study are to analyze digital forensic framework for fraud investigation and to conduct a comprehensive analysis of the potential of employing machine learning techniques in digital forensics.

2. LITERATURE REVIEW AND HYPOTHESIS

This study exlpores three different fields: fraud investigation, digital forensics, and machine learning.

Fraud examination

The practice of fraud examination is important to the success of organizations, as it serves various essential objectives. These objectives include identifying improper conduct, determining the responsible individuals, halting fraud activities, delivering a strong message against fraud, quantifying potential losses, aiding recovery efforts, preventing future losses, mitigating consequences, and enhancing internal controls (Wells, 2018). To achieve these goals, fraud examiners engage in four key activities during investigations: gathering evidence, submitting reports, providing testimony, and contributing to the overall effort to detect and prevent fraud.

The effectiveness of a fraud investigation relies on the credibility of the evidence collected. Fraud examiners must have the knowledge and skills required to gather documentary evidence and witness testimony legally and effectively, as these are often crucial in revealing fraudulent activities. Once the evidence has been obtained, analyzed, and interpreted, those conducting the investigation are responsible for communicating their findings to relevant parties, such as management, the board of directors, or the audit committee. The fraud investigation report is a detailed record of the examiner's activities, findings, and any recommendations they may have.

Fraud examiners often find themselves in situations where they are required to testify and present their findings in legal proceedings such as trials. It is essential that they maintain honesty and clarity when providing testimony to ensure the integrity and impact of their statements. In addition, while fraud examiners are not directly responsible for fraud prevention—management or relevant authorities assume such duties—they actively engage in recommending appropriate policies and procedures to deter fraud effectively.

Digital forensics

Digital forensics is a critical branch of forensic science that applies scientific principles, methodologies, and techniques to investigations involving digital evidence. It involves the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence sourced from digital devices Sachowski (2016). Its primary objective is to facilitate the reconstruction of criminal events or anticipate unauthorized actions that may disrupt planned operations (Pearson, 2001).

Numerous scholars have contributed to developing proposed frameworks for digital forensic procedures resulting in at least 21 frameworks being issued between 1995 and 2011 (Oettinger, 2020). Although specific steps may vary, the most common and generalized sequence of digital forensic procedures includes preparation, identification, collection, preservation, examination, analysis, and presentation (Sachowski, 2016).

The first process is the preparation that involves organizing equipment and personnel to ensure a well-prepared and efficient investigative process. The identification process centers around detecting and acknowledging an incident or potential digital evidence. Furthermore, digital forensics examiners should acquire evidence using standardized and approved techniques to ensure its integrity, followed by a preservation process emphasizing the significance of maintaining proper evidence collection and the chain of custody. Preservation aims to safeguard the integrity and admissibility of the evidence.

The subsequent digital forensics process is conducting an examination, which entails the comprehensive evaluation of digital evidence volumes, examination of protected files, and the analysis of registry data. This process is often connected with the analysis process, which focuses on analyzing the content and context of digital evidence, determining its relevance, establishing links, and investigating the root cause of the incident. Lastly, the presentation comprehends detailed reports to document all processes undertaken during the investigation thoroughly.

In 2012, the International Organization for Standardization (ISO) introduced a comprehensive standard for digital evidence handling, to establish a consistent process for digital evidence handling, regardless of the nature of the investigation or the type of evidence involved. The standard, known as ISO/IEC 27037, provides guidelines for the proper handling and management of digital evidence in a manner that is accurate, reliable, and legally admissible. It is widely adopted by law enforcement agencies, forensic investigators, and other organizations involved in digital investigations, as it helps to ensure the integrity of the evidence and the credibility of the investigation. Four distinct phases of the standard are:

a. Identification. This initial phase involves the thorough search, recognition, and documentation of physical devices present at the scene that may contain valuable digital evidence.
b. Collection. Devices identified during the previous phase are collected and either transferred to an analysis facility or acquired on-site to preserve the integrity of potential evidence.

c. Acquisition. The process of acquiring an image of the potential evidence source is undertaken, aiming to create an ideally identical replica of the original data.
d. Preservation. Continuous assurance of evidence integrity, both physical and logical, is of utmost importance throughout the investigation process.

In addition, the examination, analysis, and interpretation of digital evidence are carried out following a standardized framework set by the ISO/IEC 27042:2015. This standard ensures that the digital evidence handling process is systematic, transparent, and accurate, thus maintaining the integrity of the evidence throughout the analysis process. Once the digital evidence has been collected, it undergoes a thorough examination. The digital forensic analysis is conducted to assess the relevance of the evidence to the case, considering the context and objectives of the investigation. Various branches of digital forensics, such as network forensics, image forensics, computer forensics, and mobile forensics, may be involved in this process. The interpretation of digital evidence involves transforming raw data into meaningful information that can be used as evidence in the case. Higher-level analysis techniques, such as content analysis or sentiment analysis, may be applied to assist in this process. The interpretation of digital evidence is crucial as it plays a significant role in determining the outcome of the investigation. Finally, the results of the digital forensic analysis are effectively communicated and disseminated to the relevant parties involved in the investigation. The reporting stage is essential as it allows the investigator to share their findings with the stakeholders involved and helps to build a comprehensive picture of the incident or crime. The ISO/IEC standard provides a comprehensive framework for digital evidence handling, ensuring that the digital forensics process is carried out systematically, accurately, and transparently. The standard helps to maintain the integrity of the digital evidence throughout the investigation process and ensures that the results of the analysis are effectively communicated to all relevant parties.

The Association of Chief Police Officers's (ACPO) Good Practice Guides for Digital Evidence are considered to be one of the most authoritative sources of best practice advice for digital forensics. (Horsman, 2020). One of the key highlights of these guides is the four governing principles for evidence handling, which are integrity, competency, preservation, and responsibility (Williams, 2012).

Preservation of data integrity

Law enforcement agencies, their personnel, and agents must refrain from taking any action that could alter data that might serve as crucial evidence in court proceedings.

Competent handling of original data

If circumstances necessitate accessing original data, the individual conducting such actions must possess the necessary competence and be capable of providing evidence explaining the significance and implications of their actions.

Creation and preservation of audit trail

A comprehensive audit trail or record of all processes applied to digital evidence should be established and preserved. Furthermore, these processes should be subject to examination by an independent third party, capable of achieving identical results.

Responsibility of the investigator

The individual overseeing the investigation holds ultimate accountability for ensuring strict adherence to the law and these guiding principles. These principles represent fundamental pillars of best practices for handling digital evidence, emphasizing the critical importance of maintaining data integrity, ensuring competence in handling original data, documenting all procedures through audit trails, and enforcing responsible oversight by the investigator (Williams, 2012).

Machine learning

Machine learning algorithms are increasingly being used in various applications, including digital forensics. It is essential for digital forensic investigators to understand the underlying process and algorithms used in digital forensic tools to improve their efficiency. Ayyadevara (2018) emphasizes the need for digital forensic investigators to have a comprehensive understanding of the algorithms used in digital forensic tools.

One of the most significant machine learning algorithms used in digital forensics is the Support Vector Machine (SVM). SVMs are supervised learning models that are used for pattern recognition, data analysis, regression analysis, and classification (Jordan et al., 2008). Decision Tree is another commonly used algorithm that generates a flow chart for the results, such as a tree structure, to classify data (Han et al., 2011). The algorithm classifies data in a dataset through a query structure by going bottom-up from the root to the leaf representing one class. The k-Nearest Neighbors algorithm is a nonparametric method used for regression and classification. Naïve Bayes Classification is another algorithm that results from the application of the Bayes theorem. Different types of Bayes classifiers are Gaussian Naïve Bayes and Multinominal Naïve Bayes. Artificial Neural Networks (ANN) are another type of algorithm used in machine learning which are derived from the model or system in the human brain or the human neuron. The layers in an ANN are the input layer, hidden layer, and output layer.

A recent meta-analysis conducted by Nayerifard et al. (2023) systematically reviewed the application of machine learning within the domain of digital forensics. The study encompassed an examination of 608 research papers related to digital forensics and machine learning, with data collected until December 2021. The findings reveal that several machines learning algorithms, including Convolutional Neural Networks (CNNs), Decision Trees, Support Vector Machines, and Deep Learning Neural Networks, are prevalent in the field of digital forensics.

Furthermore, notable research efforts have delved into various methodologies within digital forensics. Sachdeva & Ali (2021) conducted a study focusing on data classification in digital forensics using deep learning techniques, achieving enhanced detection capabilities. Qadir and Varol (2020) explored explanatory approaches to elucidate the workings of machine learning algorithms in this context. Additionally, Goni et al. (2020) contributed to the body of knowledge with a literature review centered on cyber-attacks and digital forensics.

3. METHODS

The aim of the study is to identify the gap in the current use of machine learning in digital forensics for investigating fraud. This focus on fraud investigation is important, as digital forensics has traditionally been used more for criminal and cyber crimes than for cyber-related crimes. To conduct this study, a document analysis methodology was utilized to gather valuable insights and enhance comprehension. Document analysis is a qualitative research that enables the researcher to identify gaps, weaknesses, or areas for improvement in current practices, contributing to a more informed process using documents, including standards, regulations, and research papers. The primary advantage of document analysis is it provides an overview of existing standards and frameworks governing the field. Moreover, it also more efficient, cost-effectiveness, lack of complexity, stability, and availability online (Bowen, 2009). However, document analysis has certain limitations: the selection of documents employed in the research and dependency on interpretative skills of the author.

The fact that document analysis has some limitations does not make it any less valuable of an approach to research (Morgan, 2022). To overcome the limitations associated with document selection and

interpretative skills, this study adopts Flick's (2018) four-factor requirements for selecting documents to be analyzed. These factors include authenticity, credibility, representativeness, and meaning. To meet these requirements, this study only uses internationally recognized guidelines and books that meet the criteria of authenticity, credibility, representativeness, and meaning.

Furthermore, exploring machine learning in digital forensics requires not only literature but also current practice analysis. Since machine learning implementation relies on technology and software and digital forensics also relies upon software application and tools for data acquisition, analysis, and interpretation of digital data (Horsman, 2019), this study explores one of digital forensics tool that used in the practice of digital forensics, the Autopsy 4.15.0.

4. RESULTS AND DISCUSSION

Digital Forensics in Fraud Investigation

Digital forensics process begins with preparation, where the digital forensic examiner gathers relevant case information, prepares necessary software and hardware, and assesses the potential data available. This phase occurs immediately after the examiner accepts the assignment before the incident response process. Based on ISO/IEC 27037: and ISO/IEC 27042:2015 the procedures after identification, then divided into two processes, as illustrates in Figure 2:

a. Acquisition. In this phase, digital evidence is gathered, and image forensics is created. If the number of digital objects is small, acquisition can be carried out promptly.
b. Collection - Preservation – Acquisition. If the data volume is large or the examiner needs to determine which evidence to acquire, the collection process is initiated. This step involves selecting potential digital evidence and securely storing the data in a safe environment, ensuring protection from external interference. The collection process requires analysis of the business process, information system, and data requirements. Temporary preservation is carried out during this phase before proceeding to the acquisition process.

However, due to the nature of the cases, digital forensics in a fraud investigation is commonly conducted with a considerable time lag from the incident. For example, in a tax investigation, the

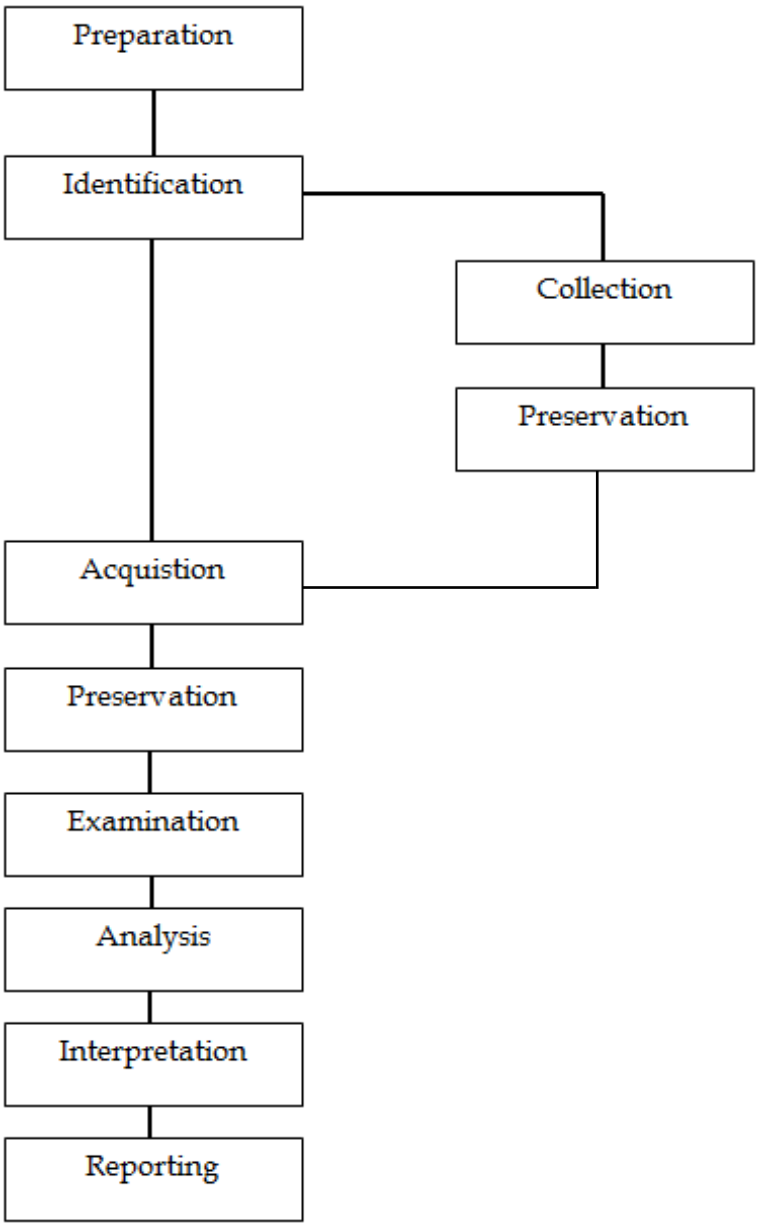
Table 1. **Selected Guidelines/Books**

No.	Guidelines/Books	Description
1.	ISO/IEC 27037:2012 Guidelines for identification, collection, acquisition, and preservation of digital evidence	An international standard for handling digital evidence, which was developed and maintained by the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC).
2.	ISO/IEC 27042:2015 Guidelines for analysis and interpretation of digital evidence	As a widely known guidelines for the analysis and interpretation of digital evidence that was developed and maintained by ISO and IEC.
3.	The International Fraud Handbook by Joseph T. Wells (2018)	An international standard of fraud investigation and guidance for developing an anti-fraud program, written by founder and chairman of the Association of Certified Fraud Examiners (ACFE).

investigator starts after taxpayers pay and report their previous tax return. Based on the tax return and economic transaction, a tax investigation is conducted after preliminary evidence of suspected tax crime is discovered.

One of the critical issues regarding this time lag and data acquisition is the difference in handling procedures for volatile and nonvolatile data. Volatile data is temporary data stored in a device's memory and lost when the device is turned off or restarted. On the other hand, nonvolatile data is permanent data stored on a device's hard drive or other storage media. Prioritizing evidence collection is also crucial because of the nature of volatile evidence. Based on the order of volatility (Sammons, 2015), the descending list work from the most volatile to the least volatile is: computer's cache, routing table, memory, temporary file system, data on hard disk, remotely logged data, and data contained on archival media.

Figure 2. **Digital Forensic Framework in Fraud Investigation**



Source: Author Based on ISO/IEC 27037:2012 and 27042:2015

In many cases, a computer's memory also becomes the only way to recover the password needed to remove the encryption on a hard drive. In a running computer, the encrypted portion of the data storage would be accessible, but only until the computer is turned off, making it essential that the hard drive is copied while the computer is still turned on. Tools are available to make copies of RAM and hard drives on running computers and line-of-business servers that cannot be shut down and still ensure that those copies are forensically sound (Daniel & Daniel, 2012). The volatile nature of electronic evidence, coupled with the ever-increasing likelihood of malicious actors attempting to tamper with or destroy such evidence, necessitates a simultaneous and integrated approach to preservation and collection during investigations (Casey et al., 2010). This concept differs from another digital forensics standard explicitly defining "preservation" and "collection" as distinct processes.

Therefore, the process of digital forensics in the context of fraud investigation is a critical component that warrants careful consideration of the types of data involved. Specifically, data should be categorized into volatile and nonvolatile types, and an assessment of the relevant information system should be conducted to identify potential evidence and relevant business processes. This assessment should include the identification of the devices and systems that engaged in the incident, determination of the scope of the investigation, and collection of relevant information about the devices and systems. Prioritizing data collection and preservation based on the order of volatility is a key step in ensuring that the collected evidence is forensically sound (Figure 2).

Subsequently, digital forensics examiner proceeds to acquire, collect, and preserve digital evidence based on the current framework. The preservation process involves creating a forensic copy of the original data, which is an exact duplicate of the data on the original device. This copy is then stored in a secure location to ensure that it is not tampered with or destroyed. Further processes in digital forensics are conducted in a streamlined manner. The examination, analysis, and interpretation phases involve data extraction, classification, summarization, and interpretation. The most time-consuming phase in digital forensics is the analysis process. During this phase, the examiner analyzes the collected digital evidence to identify any relevant information that can be used to support or refute the allegations of fraud. This analysis involves using specialized software tools and techniques to examine the data and identify any patterns or anomalies that may indicate fraudulent activity.To accelerate the digital forensics process, integration of machine learning algorithms is recommended.

Machine learning algorithms in digital forensics

The integration of machine learning algorithms in digital forensics has significantly transformed the way digital evidence is analyzed. Machine learning algorithms empower the recognition of patterns through classification techniques. These patterns are derived from the training data provided to the machine and subsequently utilized for predictions in various cases. Machine learning techniques such as cluster analysis, network analysis, and predictive analysis have revolutionized digital forensics.

Cluster analysis is one of the most popular techniques used in digital forensics. Clustering involves categorizing the population or data points into groups based on their similarities. The primary objective is to segregate data points with shared traits and assign them to clusters. Prominent algorithms used in cluster analysis include the connectivity model, density model, centroid model, and distribution model.

a. The connectivity model is based on the concept that data points that are close

in data space exhibit greater similarity. The hierarchical clustering algorithm is an example of the connectivity model.

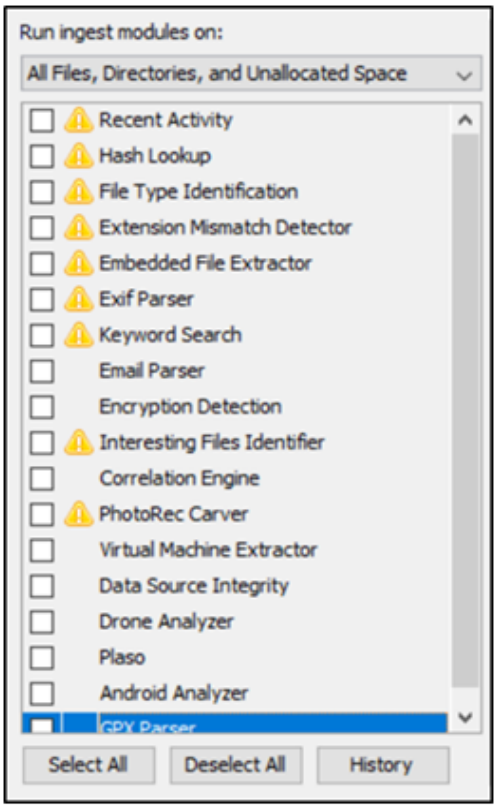
b. The density model, on the other hand, examines the density of data space to assign data points to the same cluster. The density-based spatial clustering of applications with noise (DBSCAN) is an example of the density model.
c. The centroid model, on the other hand, is based on the proximity of a data point to the centroid of clusters. The k-Means clustering algorithm is an example of the centroid model.
d. Lastly, the distribution model is based on the notion that all data points belong to the same distribution. The Naïve Bayes Classification algorithm is an example of the distribution model.

Clustering software can form groupings based on human behavior, while in machine learning, clusters are autonomously created by the program from the data. Cluster analysis is useful in digital forensics as it helps to identify patterns and similarities between data points.

Another important technique used in digital forensics is network analysis, also known as link analysis. Network analysis is employed in law enforcement to create charts representing relationships between suspects and evidence. In machine learning, the process can be automated by examining the information and establishing connections between objects. This analysis provides insights into relationships and reveals the frequency of contacts (closeness and betweenness). Network analysis is useful in digital forensics as it helps to identify relationships and reveals the frequency of contacts between suspects and evidence.

Predictive analysis is a valuable technique used in digital forensics. While cluster and network analysis are valuable in the current analytical process, they also serve as a foundation for predictive

Figure 3. **Ingest Modules in Autopsy**



Source: Author Based on the Autopsy 4.15.0

Figure 4. **Analysis Moduls in Autopsy**

Source : Author from Autopsy 4.15.0

analysis. Utilizing time series models, random forests, and other algorithms, predictive analysis improves analytical predictions. This involves understanding human or data behavior, training the data, and executing the model. Predictive analysis is useful in digital forensics as it helps to identify trends and predict future behavior.

Machine learning in digital forensics tools

Once digital evidence is acquired, the next crucial step is the analysis and interpretation of such evidence. The standard framework for analysis and interpretation of digital evidence is defined in ISO/IEC 27042:2015, which outlines the process of examination, analysis, interpretation, and reporting. To illustrate this process, Autopsy 4.15.0, a widely utilized digital forensics software, is examined in this context. Developed by Basis Technology, Autopsy provides a graphical user interface that enables the deployment of various open-source programs and plugins from The Sleuth Kit (TSK), a collection of utilities for extracting data from digital storage to facilitate forensic analysis of computer systems. Autopsy is commonly used by law enforcement, military, and corporate examiners to investigate digital evidence once it has been collected. The software's user-friendly interface and extensive range of capabilities make it a popular choice in the field of digital forensics.

Examination

The process of digital forensics examination involves the extraction of digital evidence, identification of relevant data, and its classification. In recent years, machine learning principles have been leveraged for the development of digital forensic software. These tools utilize metadata associated with computer files to categorize them into distinct groups, thereby simplifying the data classification and summarization process for investigators.

In Autopsy 4.15.0, once the digital forensics examiner loads the evidence into the software, the first step involves extracting digital evidence from the data. This crucial step lays the foundation for subsequent analysis and interpretation. Several common modules are utilized for data classification using metadata associated with each file. These include modules of File Type Identification, Extension Mismatch Detector, and Exif Parser. Through these modules, data can be effectively categorized based on file metadata. Moreover, Autopsy offers additional functionalities such as file restoration for deleted files, data integrity verification, and file encryption detection. hese features enhance its capabilities as a comprehensive digital forensic software. The data extraction process plays an imperative role in simplifying data classification and summarization for investigators, providing the extracted information in a readable and accessible format. This facilitates a streamlined and efficient analysis of digital evidence during the investigative process (Figure 3).

Analysis

The data analysis stage is a multi-step process and requires machine learning techniques to simplify the process. Autopsy software provides several techniques and methods in data clustering and visualization.

Data clustering

The process of data classification in digital forensics involves grouping data based on file signatures. This initial step serves as a simple and straightforward method for categorizing data. However, as the process advances, a more complex approach

is necessary to accurately classify data using fuzzy hash technique. The fuzzy hash algorithm is a powerful tool for calculating the similarity between digital files. It introduces a level of tolerance for changes, allowing for more accurate analysis of differences between two files. This is achieved by comparing the similarity of each output, which enables a more comprehensive classification of data during digital forensic analysis.

Network analysis

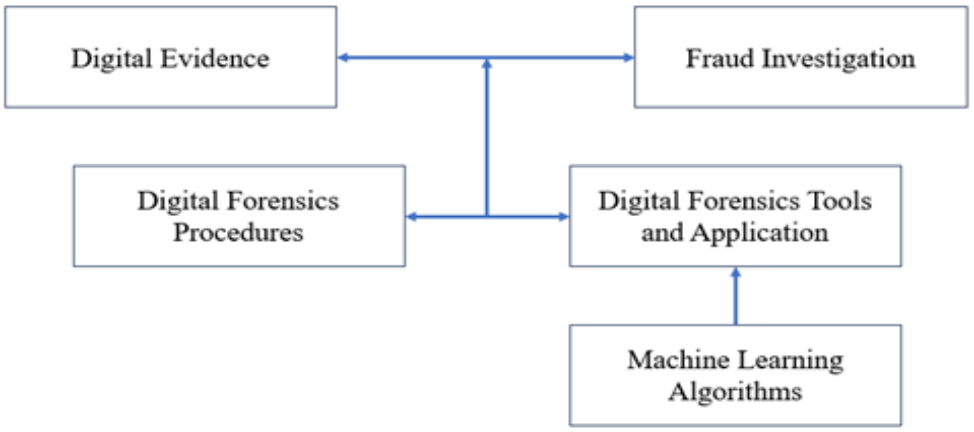
The Autopsy software is a comprehensive and sophisticated tool that offers a broad range of features to facilitate clear and accessible data visualization. Its Communications menu provides an insightful and detailed link analysis of communication activities on various devices or social media backups. By creating edges and nodes, this menu illustrates the connections between entities, allowing users to understand the relationships within the data comprehensively. Moreover, the Geolocation menu conducts an in-depth location analysis for each file's origin, providing users with valuable insights into the geographic distribution of data. This feature offers a visual representation of file locations that enhances data interpretation in a meaningful way. Thus, users can easily identify patterns and trends in data distribution, aiding in a better understanding of the evidence.

In addition, the Timeline menu presents a well-organized and chronological depiction of file creation within the evidence. This timeline visualization allows users to track the sequence of events and better comprehend the temporal aspect of the digital evidence. The timeline is an extremely helpful tool, especially in cases where time is a crucial aspect of the investigation. By providing a clear and detailed overview of the file creation, users can reconstruct the events leading up to the incident and gain valuable insights into the case.

Interpretation

Interpretation is a pivotal step in the digital forensics process as it plays a vital role in comprehending and effectively presenting the significance and implications of the digital evidence uncovered. Without interpretation, the findings from the analysis stage may not be fully understood, and the digital evidence may not be used effectively in legal proceedings. During this phase, digital forensic investigators leverage their expertise and specialized knowledge to clarify the processes, predictions, reconstructions, and potential implications derived from the analytical findings. They may also seek to confirm their findings with other evidence or information to ensure that their interpretation is accurate, relevant, and reliable.

Figure 5. **Relationship among Machine Learning Algorithms, Digital Forensics Tools and Procedures, and Fraud Investigations**



Machine Learning

Since digital forensics requires extensive data analysis to draw meaningful insights and make informed decisions eventhough the process is can be time-consuming, resource-intensive, and prone to human errors, machine learning algorithms have the potential to revolutionize investigative practices by providing predictive analysis. By employing behavioral models trained on extensive data, these algorithms can learn from patterns and trends in data, identify relationships between different variables, and make predictions based on those relationships. This can help investigators focus their efforts on the areas that are most likely to yield results, saving time and resources.

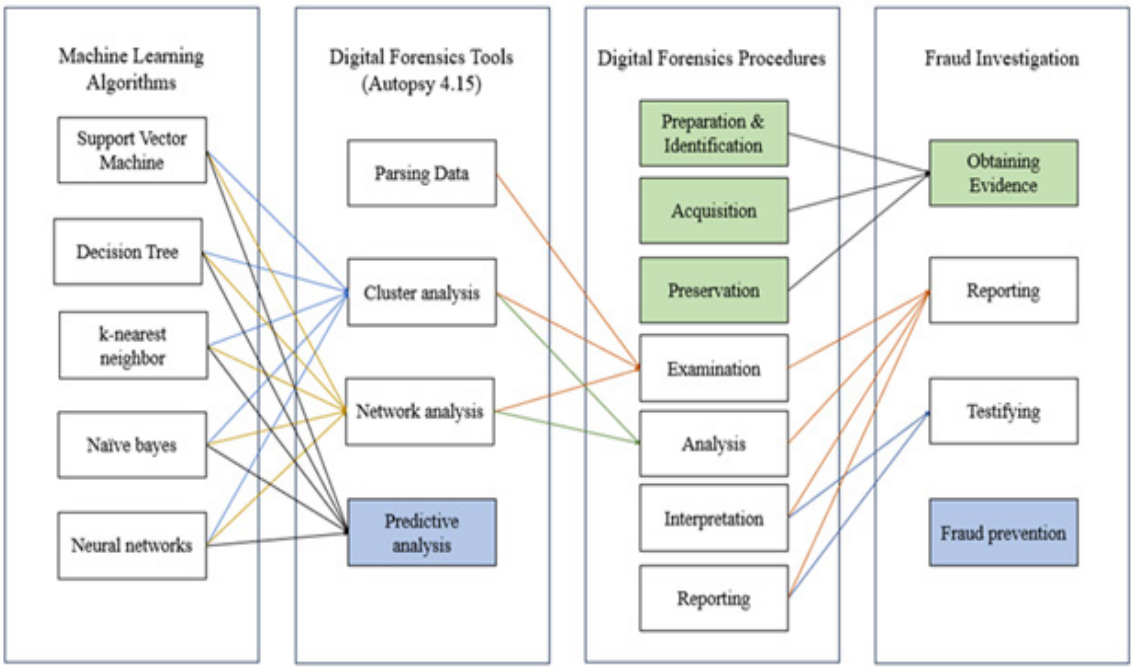
Relationship among Machine Learning, Digital Forensics Tools and Procedures, and Fraud Investigation

In a rapidly-evolving technological landscape, digital evidence has become an requisite element in fraud investigations. The digital footprint left behind by individuals in their online activities is often the key to identifying fraudulent activities and obtaining evidence necessary to prosecute the perpetrators. However, handling digital data can be a very complex and intricate process, requiring specialized expertise, procedures, and tools to ensure its integrity and accuracy. These tools must be designed to handle the complex and varied nature of digital data, which can involve anything from emails and social media posts to financial transactions and chat logs. Additionally, with the vast amounts of data involved in modern investigations, machine learning algorithms can be used to enhance the efficiency of the investigation, as depicted in Figure 5.

Machine Learning and Digital Forensics Tool

Machine learning supports automation in the analysis process especially in large volumes of digital evidence. The process reduce the time and resources required in manual examination. As suggested by Nayerifard et al. (2023), machine learning algorithms such as support vector, decision tree, k-nearest neighbor, naïve Bayes, and neural network are widely embraced in the field of digital forensics.

Figure 6. **Relationship among Machine Learning Algorithms, Digital Forensics Tools and Procedures, and Fraud Investigations**



Autopsy 4.15.0, a software tool for digital forensics, incorporates modules that employ cluster and network analysis concepts such as Timeline and Communication, which use machine learning algorithms for data classification. The algorithms utilized in Autopsy can significantly enhance the analysis of digital evidence, enabling investigators to detect hidden patterns and relationships that may otherwise be undetectable. Furthermore, machine learning algorithms have the potential to be explored in predictive analysis, which may reveal new insights that could lead to improved forensic investigations. The Modules of Autopsy reflect predictive analysis, but this area has not yet been fully explored.

Digital Forensics Tools and Digital Forensics Procedure

As a tool for examination and analysis of digital evidence, Autopsy 4.15.0 has been specifically designed to extract critical data from digital evidence, and then present it in a human-readable format. This tool is essential for analyzing and examining digital evidence while maintaining its original authenticity. Autopsy 4.15.0 is capable of reading digital forensics images in various formats, including raw and e01 which widely used in digital forensics acquisition. Once the image has been loaded, it undergoes a thorough parsing process that involves metadata analysis to transform it into a readable format. During the analysis phase of digital forensics procedures, it is implements the use of cluster and network analysis to improve overall efficiency. However, it is important to note that predictive analysis is not covered in digital forensics procedures due to the unique nature of each case. Therefore, each case requires individual attention and in-depth analysis to uncover all relevant evidence, which can then be used to build a solid case.

Digital Forensics Procedure and Fraud Investigation

Fraud investigation is a complex field due to the digitalization of data. Digital forensics procedures have emerged and provide a structured approach to the collection and preservation of electronic evidence. The process in digital forensics ensures that evidence is handled with accuracy by maintaining data integrity and establishing a chain of custody. Additionally, digital forensics also offers an analytical advantage to investigators, for example in the process of reconstructing events, establishing timelines, and analyzing the patterns of fraudulent behavior.

In fraud investigations, since the investigators are required to establish a fraud report and testimony, digital forensics framework is become more vital to provides valuable assistance in the examination, analysis, interpretation, and reporting process. These sequence procedures ensure that the evidence is properly collected, analyzed, and presented in a way that is admissible in court to build a strong case against the alleged perpetrator. While digital forensics is valuable for post-incident analysis and support investigation in legal proceedings, several challenges in the role of supporting fraud prevention and early warning detection exists. First, the nature of digital forensics id reactive, which limits its effectiveness in providing early warning of fraud. Additionally, the digital forensics process is time-consuming and primarily focuses on case attributes, which detracts from its efficacy in preventing and detecting fraud. The adherence to strict legal and privacy standards when collecting and analyzing digital evidence presents another challenge. The implementation of proactive monitoring systems for fraud detection may be hindered by privacy concerns, which may limit the extent of data usage, despite the potential of machine learning to leverage digital forensics tools for fraud investigations.

5. CONCLUSION

This paper begins by presenting the digital forensics framework in the context of fraud investigation. Then, it explores the use of digital forensics tools to analyze

the potential implementation of machine learning algorithms in digital forensics for fraud investigations. Machine learning plays a significant role in digital forensics, especially in the process of data examination and data analysis. The algorithms are able to support digital investigations more effective. On the other hand, in the phase of data interpretation, the algorithms are less developed. Digital forensics can support fraud investigation by obtaining evidence, reporting, and testifying. However, since there is no predictive process in digital forensics, which is one of the requirements in fraud investigation, the fraud detection and prevention process remains under-explored. Machine learning algorithms have the potential to be implemented for fraud detection and prevention because of their ability to recognize patterns and predict behavior. This represents an area that needs further advancements. Reactive action, focus on case attributes, legal standard, and privacy regulation are factors that undermine digital forensics to fraud detection and fraud prevention.

ACKNOWLEDGEMENTS

The author is studying under the supervision of Professor Kaneko Hironao at the Tokyo Institute of Technology. The Indonesia Endowment Fund for Education Agency (LPDP) provided the author with funding for his study.

REFERENCE

Ayyadevara, K. V. (2018). *A Hands-on Approachto Implementing Agorithms in Python and R: Pro Machine Learning Algorithms.* Apress Berkeley, California.

Bowen, Glenn A. (2009). Document Analysis as a Qualitative Research Method. *Qualitative Research Journal*, *9*(2), 27-40.

Casey, E., Daywalt, C., Johnston, A., and Maguire, T. (2010). *Network Investigations.* Habdbook of Digital Forensics and Investigation.

Daniel L. E. & Daniel, L. E. (2012). *The Foundations of Digital Forensics: Best Practices.* Digital Forensics for Legal Professional.

Flick, U. (2009). *An Introduction to qualitative research (4th Edition).* Sage Publications.

Guarino, A. (2013). *Digital Forensics As A Big Data Challenge. ISSE 2013 Securing Electronic Business Processes:* Springer.

Goni, I., Gumpy, J. M. Maigari, T. U., Muhammad, M., and Saidu, A. (2020). Cybersecurity and Cyber Forensics: Machine Learning Approach. *Machine Learning Research, 5*(4), 46-50.

Han, J., Pei, J., Kamber, M. (2011). *Data mining: Concepts and Techniques.* Elsevier.

Horsman, G. (2019). Tool Testing and Reliability Issues in the Field of Digital Forensics. *Digital Investigation, 28*, 163-175.

Horsman, G. (2020). *ACPO Principles for Digital Evidence: Time for an Update?*. Forensic Science International: Reports.

Internal Revenue Service – Criminal Investigation (2023). Internal Revenue Service Criminal Investigation 2022 Annual Report.

Internal Revenue Service – Criminal Investigation (2022). Internal Revenue Service Criminal Investigation 2021 Annual Report.

Internal Revenue Service – Criminal Investigation (2021). Internal Revenue Service Criminal Investigation 2020 Annual Report.

Internal Revenue Service – Criminal Investigation (2020). Internal Revenue Service Criminal Investigation 2019 Annual Report.

International Organization for Standardization. (2012). Guidelines for identification, collection, acquisition, and preservation of digital evidence. ISO/IEC 27037:2012.

International Organization for Standardization. (2015). Guidelines for the analysis and interpretation of digital evidence. ISO/IEC 27042:2015.

Jordan, M., Kleinberg, J., and Scholkopf, B., (2008). *Support Vector Machines, Information Science and Statistics*. New York: Springer.

Morgan, H. (2022). Conducting a Qualitative Document Analysis. *The Qualitative Report, 27*(1), 64-77

Nayerifard, T., Amintoosi, H., Bafghi, A. G., & Dehghantanha, A. (2023). *Machine Learning in Digital Forensics: A Systematic Literature Review*.

Oettinger, W. (2020). *Learn Computer Forensics, Packt.*

Pearson, G. (2001). *A road Map for Digital Forensic Research*. Digital Forensic and Research Workshop 2001.

Platzer, C., Stuetz, M., and Lindorfer, M. (2014). Skin Sheriff: A Machine Learning Solution for Detecting Explicit Images. *Proceedings of the 2nd International Workshop on Security and Forensics in Communication Systems*, 45-56.

Qadir, A. M. and Varol, A. (2020). The Role of Machine Learning in Digital Forensics. *Proceedings of the 2020 8th International Symposium on Digital Forensics and Security (ISDFS)*.

Sachdeva, S. & Ali, A. (2021). Machine Learning with Digital Forensics for Attack Classification in Cloud Network Envireonment. *International Journal of Systems Assurance Engineering and Management, 13*(1), 1-10.

Sachowski, J. (2016). I*mplementing Digital Forensic Readiness from Reactive to Proactive Process*. Elsevier.

Sammons, J. (2015). *Collecting Evidence. The Basics of Digital Forensics (2nd Edition)*. Syngress Publishing.

Wells, Joseph T. (2018). *The International Fraud Handbook*. John Wiley & Sons, Inc.

Williams, J. (2012). *ACPO Good Practice Guide for Digital Evidence.* Association of Chief Police Officers of England, Wales & Northern Ireland (ACPO).