IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A REVIEW PAPER ON "DIGITAL FORENSIC TO DIAGNOSE CYBER CRIME USING MACHINE LEARNING"

Mamta Sakpal[#1], Dr. Surendra kumar yadav[*2]

Research Scholar, Vivekananda Global University, Jaipur Rajasthan

*Professor, Vivekananda Global University, Jaipur Rajasthan

Abstract: One of the many types of crimes that exist in the technological age is cybercrime. Internet usage is rising quickly as more and more things become digital, and attackers are also exploiting it to perpetrate cybercrimes. Examples of invasions include hacking, banking fraud, email spamming, and others. In order to look into these fraudulent acts, the investigating authorities should use technology, which is a crucial component. In the domain of cyber forensics known as "digital forensic scrutiny," digital information that can be used as evidence in court is preserved and examined using scientific techniques and innovations. In order to reconstruct events, forensic investigators struggle with data collecting and analysis. Because people interact with one other quite a bit on a daily basis, machine learning enables investigators to use various algorithms to carry out investigations that are more successful and efficient.. This field of study focuses on developing computer models and algorithms that enable Every machine learning algorithm is domain-specific, capable of doing specialized tasks without the need for programming, like training and testing datasets, and can facilitate research. This study explores several machine learning approaches that look at and evaluate digital evidence while conducting an investigation. It also provides a complete overview of current digital forensics techniques and their applications. Furthermore, based on the features, this study assesses machine learning algorithms according to recognized digital forensics criteria.

Keywords: Digital Forensics, Cyber-attacks, Forensic Science, Security, Machine Learning algorithm, Digital Evidence.

I. INTRODUCTION

Due to the increase of the Internet and technology's quick development, enormous volumes of data and information must be preserved. Everybody possesses gadgets, including computers and smart phones, which are susceptible to attacks by con artists and have led to a significant rise in digital crimes. Investigations into crimes including hacking, banking fraud, and email spamming are conducted using the broad fields of digital forensics and cyber forensics.

Digital forensics is the field of study that integrates all the investigations and research needed to solve these types of digital crimes [1]. Cyber and digital forensics share a same theme. Its main goal is to identify digital data storage technologies.

Digital forensics (DF) is a process that analyzes and presents data from computers, databases, and digital images [1]. Most of the time, data and evidence gathered from a gadget may be removed following the commission of a crime. Because it enables them to recognize the victims and ascertain the precise nature of the crime, this technique is crucial for investigators [3]. Unfortunately, without sufficient human resources, conducting a thorough investigation can take a long time. Although numerous methods, including Hadoop ,

can be used to manage the enormous amount of data gathered by a digital forensics investigator, they are not as effective as the human brain. Instead, for the analysis and gathering of data efficiently investigators employ machine learning (ML) [4]. This system can pick decisions based on facts and learn from many examples and experiences [5]. Support Vector Machine, Decision Tree , K-Means, K-Nearest Neighbor, Nave Bayes, Principal Component Analysis, Logistic Regression, Singular Value Decomposition, and Apriori are just a few of the various techniques it contains. Each algorithm is responsible of a certain duty, such as extracting features, categorizing network attacks, finding modified photos, etc. [6].
This paper's structure is outlined as follows: Digital forensics and machine learning are covered in Section 2. Section 3 discusses the suggested machine learning methods used in digital forensics. Section 4 discusses the limitations of machine learning techniques in digital forensics.

II. MACHINE LEARNING & DIGITAL FORENSIC

A scientific field called "digital forensics" is dedicated to the analysis and preservation of data that has been gathered and saved on various types of media. Although the field's origins may be found in the 1980s, the development of wide-area, multi-user, and multi-tasking networks in the 1990s expedited the field's evolution [7]. Because of the increase in security risks and attacks, it has emerged as one of the most important areas of security. The goal of machine learning, a branch of artificial intelligence, is to build systems that are able to learn from facts. The domains of behavior prediction, analysis, and data mining commonly make use of this technology [8]. The issues, models, and phases of the digital forensics inquiry and different machine learning algorithms are described in this section.

2.1 DIGITAL FORENSICS

A subfield of criminalities called "digital forensics" focuses on the legal processes involved in examining and safeguarding digital data. It entails locating and obtaining data from multiple sources. The information can then be used to assess the evidence in a civil or criminal trial [9]. This technique entails analyzing the data that various digital objects have produced using scientific and technological methodologies [10]. The goal of digital forensics is to gather information that can be utilized to ascertain the details of an incident. Investigations frequently ask the 5WH questions, including who was involved, where the incident happened, how it happened, and when it happened. The solution to these queries helps the investigators confirm the incident [11].

A. DIGITAL FORENSICS INVESTIGATION PROCESS.

The four techniques and methodologies employed in the digital forensics process, according to the National Institute of Standards and Technology, are intended to aid businesses in comprehending the significance of their investigations, as depicted in Fig. 2. Depending on the intricacy of the investigation, they can be carried out in a variety of ways [12]. The number of data sources that can be gathered has increased as a result of the development of digital technologies. The steps involved in a digital forensics investigation are shown in Figure 1.

Fig 1. Process for Investigating Digital Forensics

Below is an explanation of each step.

a) **Data collection:** Finding prospective sources of this data is the initial step in conducting an investigation. Typically, servers, desktop computers, and laptops are where the data is gathered. When examining an organization's operations, analysts should take into account additional data sources in addition to more conventional ones. For instance, they can examine the Internet service provider's logs to learn more about an organization's operations. [13].

b) **Examination**: The acquired data will be examined in the second stage. The required pieces of knowledge are obtained from the data using digital forensics techniques and tools. Define the data files that include information of interest, including information hidden by access restriction, encryption, and file compression [13, 14].

c) **Analysis:** An analysis is a process that comprises using scientific methods in a scientific setting to produce items such identifying people, places, and events and determining the relationships between these aspects [15]. As part of this process, the data gathered from various sources is analyzed. Correlating and gathering data can be made easy with the help of solutions like security event management software; for instance, an audit log may contain information about a particular host, while an IDS log might contain knowledge about an individual user [14].

d) **Reporting**: The investigation's last stage, reporting involves examining the information gathered during the analysis stage and presenting the results to the analyst in a formal report. Finding the reason behind an event or giving a precise explanation can be difficult, but by using the data, an analyst can gain a better knowledge of the incident and help to avoid a repeat of it in the future [15].

B. DIGITAL FORENSICS MODELS.

Investigation models for digital forensics include the End-to-End Digital Investigation Process Model (EEDIP), the Integrated Digital Investigation Process Model (IDIP), the Abstract Digital Forensics Model (ADFM), and the Digital Forensics Research Workshops Model (DFRWS) [16]. Each model has been created for a certain stage or activity. The digital forensics models and associated operations are shown in Figure 2.

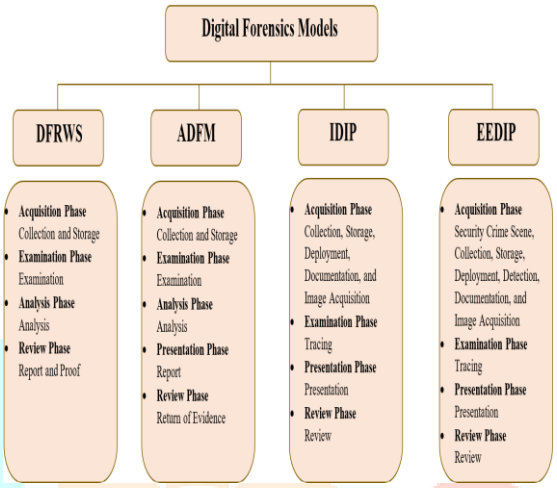

Fig. 2. Models of the Digital Forensics

C. DIGITAL FORENSICS THREADS

The growth of digital forensics has been hampered by the growing quantity of digital gadgets. The intricacy of encryption-using cell phones, software platforms, and hardware makes gathering digital evidence extremely difficult. As a result, the sector now needs new approaches and techniques to deal with its problems. The growing diversity of operating systems and file formats prevents the International Journal of Organizational and Collective Intelligence from creating standardized procedures and tools for digital forensics, claim Montanari et al. [17]. The amount of data that can be gathered and evaluated has grown more difficult as digital technology has become more complicated. Large-scale data collection and analysis is now feasible due to new data formats like low binary. Another issue is complexity, according to Horsman et al. It gets harder to design systems that can interpret data acquired quickly as it is more collected. Additionally, a significant problem is the absence of standards in the formatting and storing of digital evidence. The sharing of digital proof is difficult. By establishing a uniform set of protocols, this problem may impact the effectiveness of investigations and improve the efficiency with which law enforcement shares information [19].

When creating digital analysis tools, correlation and consistency are said to be the major obstacles by Quick et al. Since the evidence is gathered from many sources, accurate data analysis and correlation are required. This can take a lot of time and money out of an inquiry [20].
According to Pandey's research, digital forensics experts encounter the time-lining challenge when divergent sources offer contradictory interpretations of the evidence. This matter may impact the effectiveness of an inquiry. Another important problem that jeopardizes the growth of the sector and the authenticity of the data is the lack of awareness regarding the most recent digital forensics techniques. Because forensic science is developing so quickly, professionals in the field need to be able to apply new technologies with efficiency.

2.2 MACHINE LEARNING

Machine learning, which enables systems to learn and evaluate without the need for extra training, is one of the most popular methods to artificial intelligence (AI) [22]. It has the ability to forecast and classify inputs automatically [23]. This technology can detect fraud and provide security by using the right algorithms in different areas. Supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning are the four primary subcategories of machine learning. An input to an output linking is an example of a supervised learning process. It makes use of training data that has different training examples labeled on it. The most widely utilized methods in this procedure are regression and classification [24]. The clustering technique, sometimes referred to as unsupervised learning, can assist in locating hidden patterns and structures within the datasets [25].

On the other hand, semi-supervised learning is a branch of machine learning that works with both labeled and unlabeled data to carry out different tasks. Between supervised and unsupervised learning [26] is where it lies. Reinforcement learning, which can resolve issues of regulate specific circumstances, places a focus on rewarding behaviors and penalizing those who do not meet standards [27]. The following is a description of these algorithms:

A. **Support Vector Machine** Encourage Both classification and regression issues can be handled by vector machines. SVM classifies objects using examples from the training data set. Depending on the kernel function, it may process complex functions on both structured and semi-structured data. This approach finds a hyperplane that divides each data item into two classes after taking the amount of characteristics into account. It reduces mistakes while increasing the marginal distance.

B. **Decision Tree Algorithm** One learning technique that can be applied to both task categorization and regression is the decision tree. It is simple to understand and can link test results to the categorization of data elements. A decision tree model models the several decision logics into a structure like a tree. the root node, which is the topmost node in a DT tree. A decision tree's internal nodes represent tests pertaining to the input variables or attributes. The classification algorithm branches to the relevant child node after the test is finished. Until the leaf node is prepared to make a decision, this process is continued [29].

C. **The K-Nearest Neighbor Algorithm** The K-Nearest Neighbors algorithm is a learning technique that is non-generalizing and does not prioritize building a general model. It maintains all training data instances in an n-dimensional space. The K-Nearest Neighbors algorithm is capable of handling a variety of tasks, including regression and classification, which handle data training to deliver accurate data based on the quality of data [30]. It uses data to categorize new data points.

D. **The Naïve Bayes algorithm.** An unsupervised learning algorithm used for classification or grouping problems is called Naïve Bayes. It can be used as a clustering approach and does not require the specification of a result [31]. For the algorithm to estimate the required parameters, a minimal amount of training data is needed. Nave Bayes is a supervised learning method because it depends on both the input and the goal variables. When used as a classifier, it creates a tree made up of outcome-probability-based Bayesian networks [32].

E. **K-Means Algorithm** The K-Means algorithm is a straightforward and effective way to divide datasets into K centers. Because it works better when the variables are significant, it is comparable to hierarchical clustering. Thus, the implementation and data interpretation are the effective components of this method [30].

F. **Principal Component Analysis** Algorithm The process of principal component analysis involves taking into account the observations of several potential correlated variables and converting them into values that are linearly uncorrelated. By using the Orthogonal Transformation algorithm, it may be completed fast and easily. As a result, the model can no longer be computed using previous knowledge. PCA offers several more characteristics, including data feature categorization and estimate, in addition to data clustering and classification [33].

G. **Logistic Regression** In machine learning, classification issues are resolved through the employment of a logistic regression model. It facilitates determining the class linked to a certain instance. The model's outcome falls between zero and one because it is a probability. Thus, it can be applied as a binary classifier [34].

H. **Singular value Decomposition** The idea behind the SVD factoring technique is frequently applied to matrices. By taking into account the dominating patterns, the SVD method produces a low-dimensional representation of a high-dimensional data set. The primary foundation of this strategy is data that is

gathered without the need for expertise or intuition. Singular values can be utilized to derive invariance information from an image or a signal using the decomposition approach [35]. made to function well in a database with several transactions. However, a number of factors, including the need for "n" numbers of frequent item sets in the database searches, could cause its performance to deteriorate [36].

I. **Apriori Algorithm**. The Apriori method is a popular tool in data mining because it can identify connections between different data sets. It often uses the candidate generation method to mine item sets. Additionally, it is made to function well in a database with several transactions. However, a number of factors, including the need for "n" numbers of frequent item sets in the database searches, could cause its performance to deteriorate [36].

III DIGITAL FORENSICS USING MACHINE LEARNING

Digital forensics and cyber security both use machine learning. Large-scale data sets kept in different cloud computing environments and networks are analyzed by digital forensics investigators using machine learning methods [42]. The behavior of the users can then be predicted using these data sets. These algorithms also have the ability to recognize patterns. Investigators utilize a set of criteria and approaches that can be used to locate intriguing data patterns to identify possible criminal activities through the use of machine learning techniques. This section outlines many algorithms that have been suggested to find digital evidence and enhance the investigative procedure

3.1 DIGITAL FORENSICS USING SUPPORT VECTOR MACHINE ALGORITHM

Islam et al. presented a methodology that uses discrete cosine transformation (DCT) and local binary pattern (LBP) operations to identify copy-move and splice assaults in color images. The SVM kernel was used to assess the suggested system. In order to identify micro-patterns, the DCT and LBP operators record variations in the local frequency distribution. The suggested approach organizes the LBP blocks' inter-cell values as feature vectors. Next, the generated photos are categorized using the SVM and radial basis function (RBF) into authentic and tampered ones. The findings of the investigation demonstrate the suitability of the suggested strategy for accuracy metrics and image forgery detection [43].

A method for identifying contrast enhancement in JPEG compression by use of an adaptive histogram was presented by Barni et al. This technique is built on an SVM detector's color SPAM characteristics. Once taught, it can recognize JPEG-compressed photos with improved contrast. The system's performance was evaluated by the researchers through training it on a collection of JPEG-compressed photos with varying quality factors (QFs). It is only effective when the QFs are longer than 80 and the one used corresponds with the test QF [44].

3.2 DIGITAL FORENSICS USING DECISION TREE ALGORITHM

An architectural framework combining the MapReduce framework, the Hadoop Distributed File System, and the decision tree algorithm was proposed by Chhabra et al. The enormous volume of data that can be gathered and stored was handled by the suggested architecture. The process is broken down into four steps: gathering network traffic, transforming it into a format that can be read by humans, filtering packets, looking for malicious activity in the data, and finally displaying a threat analysis and visualization. A decision tree increases accuracy and time efficiency in each phase by classifying threats as malicious or benign. The algorithm could identify 99% of all harmful and non-malicious traffic, according to the study's findings [47].

Usman et al. (2021) presented a hybrid strategy that combined the capabilities of multiple data forensics approaches, including machine learning, dynamic malware analysis, and cyber threat intelligence, to address concerns pertaining to the IP reputation system. It can anticipate the chance of a specific assault occurring before it happens using big data forensics, and it can then categorize them based on their behavioral traits. The system was assessed using different machine learning approaches, including DT, SVM, and NB, against a variety of current reputation systems. Recall, F-measure, and precision scores were all high for the DT [48].

3.3 DIGITAL FORENSICS USING NAÏVE BAYES ALGORITHM

In order to determine the accuracy of the distributed denial of service attack, Yudhana et al. examined the information gathered from the network traffic log. They gathered network traffic datasets and extracted network features using the Wireshark program in order to find patterns in the data. The Nave Bayes algorithm was then used to carry out a network package categorization process, and a neural network technique was used to train the system utilizing multiple neurons. The neural network had an accuracy of 95.2381%, whereas the naive Bayes had an accuracy of 99.999%, according to the study and testing. The Nave Bayes method and artificial neural networks, according to the researchers, can be used in network forensics to increase the accuracy of the findings during investigations [51].

3.4 DIGITAL FORENSICS USING PRINCIPLE COMPONENT ANALYSIS ALGORITHM

Roy created a framework for digital forensics to examine the origin and source of an image. Using random forest, the framework was able to classify it. This feature's primary benefit is that it enables investigators to pinpoint the many camera sources that result in various JPEG compression issues. By using the PCA technique, the framework also increased the accuracy of its classification. The dimensionality of the characteristics was significantly reduced using this technique [54].

3.5 DIGITAL FORENSICS USING LOGISTIC REGRESSION ALGORITHM

The kinds of malware that frequently target the registry in Windows operating systems have been discovered by Ali et al. Malware might result in the loss of critical time while conducting an investigation. They gave insightful information about how certain kinds of viruses communicate with the registry. The researchers experimented with a variety of classifiers, including decision trees and neural networks. Their study's findings demonstrated that modified timestamps and machine learning algorithms can be used to carry out digital forensics investigation. The 47 registry locations that malware frequently targets have been identified by the authors. Through their investigation, the researchers found that the Boosted tree accurately classified more than 72% of the malware. This technique enables investigators to quickly distinguish between malware types that are present and those that aren't.[55].

3.6 DIGITAL FORENSICS USING SINGULAR VALUE DECOMPOSITION ALGORITHM

Ahmed et al. presented a novel approach based on the Kolmogorov-Smirnov test and singular value decomposition to identify copy-move forgeries. A steerable pyramid is used to extract image features from different blocks, and feature vectors—which correlate to the pixel's associated features—are then saved with the indices of the original blocks. In digital picture forensics, four processing methods are investigated: color reduction, brightness adjustment, contrast adjustment, and image blurring. The suggested approach yielded good results in terms of F1 score, recall, and precision. It received a 95% for brightness adjustment and 77.5%, 82.7%, and 75% for picture blurring [57].

3.7 APRIORI ALGORITHM IN DIGITAL FORENSICS

Huan et al. used the K-means and Apriori algorithms to create a mobile forensics system. The Apriori method generates frequent item sets and extracts the rules that satisfy the minimum confidence criteria in order to increase mining efficiency utilizing mining rules. Additionally, by representing the data in a vertical structure, it improves the database's inherent qualities. The relationships between the various individuals are taken into consideration while classifying the clustering findings. The data was analyzed by the researchers using the association rules. They discovered that the high confidence criteria show that the user's regular behaviors are in line with the data's features [60] (Table 1).

Table1. **Summary of Machine Learning Algorithms in Digital Forensics Investigation**

Focused Area	ML Algorithm	Forensic Type	DF Phase	Advantage	Disadvantage
Copy-move and splice attacks [43]	SVM	Image forensics	Examination	High accuracy and trained both semi-structured and structured dataset	Less perormance on overlapping images
Contrast enhancement and identify JPEG-Compressed image [44]	SVM	Image forensics	Examination	Fast data analysis	The detector (QF) work well in a specific QF only
Detect manipulated videos and photos [45, 46]	SVM	Image and video forensics	Analysis	High accuracy	Required more processing time
Labelled malicious and non-malicious traffic [47]	DT	Network forensics	Analysis	Accurate data and time efficiency	Complex calculation
Classify attack behavioural [48]	**DT**, SVM and Na¨ıve Bayes	Network forensics	Analysis	High performance on unknown samples and reduces security issues	Long time to train
DDoS attack [49]	**KNN**, Na¨ıve Bayes	Network forensics	Examination and analysis	Flexible classification	Lazy learner and not working with another attack rather than DDoS Attack
Gender classification [50]	**RF**, KNN, RF, AB, SVM	Video forensics	Examination and analysis	High accuracy in the dark videos	Low performance on the prediction stage
DDoS attack [51]	Na¨ıve Bayes	Network forensics	Analysis	Simplicity	Zero-frequency problem
Features classification [52]	K-Means	Network forensics	Examination and analysis	High accuracy rate	Set K value in advance
Identify and recover digital evidence [53]	K-Means	File system/ memory forensic	Analysis	Discover hidden evidence	Low performance on the noisy dataset
Determine image source [54]	**PCA**, RF	Image forensics	Examination	Improve accuracy and reduce the dimensionality of the features	Loss of data if the components are not set correctly
Determine malware location in Windows Registry [55]	**LR**, DT	Malware forensics	Analysis	Possible to build it into existing forensic tools without requiring frequent updates	Used for prediction feature
Email classification [56]	**LR**, SVM, RF, DT	Email forensic	Analysis	High accuracy with bi-gram features	Each variable requires a minimum of 10 data points
Image falsification [57]	SVD	Image forensics	Examination and analysis	High precision	Reduce the block size on the low-quality image
Extract features for copy-move forgery in images [58]	SVD	Image forensics	Analysis	High performance and less computational	-

Anti-malware framework for forensics analysis [59]	**SVD**, PCA	Malware forensics	Examination and analysis	High accuracy	Not understanding data transformation
Mobile forensics application [60]	**Apriori**, K-Means	Mobile forensics/ database forensic	Analysis	High confidence and improve data mining efficiency	Required further resources

IV MACHINE LEARNING LIMITATIONS IN DIGITAL FORENSICS

The creation and application of machine learning models are significantly impacted by the absence of testing procedures and model openness. New models developed in research labs are frequently rapidly applied in real-world settings, yet they can also go wrong in these situations. Unfortunately, many machine learning models must be created to provide forensic practitioners with the required transparency and testing methods. Having the tools and resources to reproduce models can help professionals in a variety of industries solve problems more quickly and avoid problems like bias. This problem may hinder their ability to effectively explain various outputs through their systems [61].

The interpretability of deep learning algorithms is one of the main problems because, despite their potential appeal, machine learning models are useless if they cannot be properly understood. As a result, it is crucial that they be applicable in real-world situations [62].

Numerous methods, including support vector machines, decision trees, and clustering, are employed in big data analysis and prediction to examine and forecast user behavior that is anonymous. Neural networks require the right training data in order to function correctly because of their intricacy. Reusing the data won't yield the desired results as their design evolves and so does their data requirement. The inability of current reputation systems to detect zero-day abnormalities and their reliance on third parties make them problematic. One major issue that has to be addressed is the dearth of reliable data sources. Providing low-quality data might affect a model's accuracy, even while having enough information can sometimes be the same as not having any at all [63].

There are several benefits to believing computer algorithms. The capacity of humans to automate procedures and analyze enormous volumes of data has been extremely beneficial to them. Unfortunately, bias can also affect them. Since humans create and train algorithms, bias is difficult to eradicate. Nevertheless, who should be held responsible if something goes wrong? Machine learning has many benefits, but it is far from flawless, so in the future we will need to create a framework that would allow people to believe in it [64].

Table 2. Machine Learning Algorithm Limitation

ML Algorithm	**Limitation**
Support Vector machine	Unsuitable for large dataset.
Decision Tree	insufficient to address regression problems.
KNN	Less efficient when dealing with big date sets and lots of dimensions.
K-Means	Indicate the K value right away.

V CONCLUSION

The field of digital forensics has expanded in a number of ways. The challenges forensic analysts encounter in analyzing large amounts of data—pictures, videos, and other types of media—that could potentially shed light on past events have been demonstrated. In the field of digital forensics, several new difficulties are appearing with time. This resulted in the application of automation and clever strategies that make investigators' jobs easier. This study has validated a number of machine learning methods, including SVM, KNN, DT, PCA, SVD, K-Means, NB, ANN, LR, and RF, to address digital forensic problems. Algorithms separate real data from fraudulent data for use as legal evidence. Ultimately, the study provided an overview of the optimal procedures for every digital forensics method based on its attributes, benefits, and drawbacks.K-Means focuses on retrieving deleted digital evidence from memory locations based on the suggested research articles. The greatest techniques to use in an image forensics inquiry are the SVM, PCA, and SVD, whilst network forensics is supported by KNN and NB. In the last few years, machine learning researchers have made significant strides toward training these computers to think like people. They now carry out intricate jobs and make judgments after thorough investigation. Even if there has been progress, machine learning still faces several obstacles, including moral dilemmas, interpretability issues, a lack of data for machine learning, and repeatability issues.

REFERENCES

[1] Joakim Kävrestad. *Fundamentals of Digital Forensics*. Springer, 2020.
[2] Konstantinos Karampidis, Ergina Kavallieratou, and Giorgos Papadourakis. A review of image steganalysis techniques for digital forensics. *Journal of information security and applications*, 40:217–235, 2018.
[3] Graeme Horsman. Tool testing and reliability issues in the field of digital forensics. *Digital Investigation*, 28:163–175, 2019.
[4] Godson Kalipe, Vikas Gautham, and Rajat Kumar Behera. Predicting malarial outbreak using machine learning and deep learning approach: a review and analysis. In *2018 International Conference on Information Technology (ICIT)*, pages 33–38. IEEE, 2018.
[5] Anand Handa, Ashu Sharma, and Sandeep K Shukla. Machine learning in cyber-security: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1306, 2019.
[6] R Saravanan and Pothula Sujatha. A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 945–949. IEEE, 2018.
[7] Athanasios Dimitriadis, Nenad Ivezic, Boonserm Kulvatunyou, and Ioannis Mavridis. D4i-digital forensics framework for reviewing and investigating cyber attacks.*Array*, 5:100015, 2020.
[8] Sana Qadir and Basirah Noor. Applications of machine learning in digital forensics. In *2021 International Conference on Digital Futures and Transformative Technolo- gies (ICoDT2)*, pages 1–8. IEEE, 2021.
[9] Stefania Costantini, Giovanni De Gasperis, and Raffaele Olivieri. Digital forensics and investigations meet artificial intelligence. *Annals of Mathematics and Artificial Intelligence*, 86(1):193–229, 2019.
[10] Eoghan Casey. *Handbook of digital forensics and investigation*. Academic Press, 2009.
[11] Owen Defries Brady. *Exploiting digital evidence artefacts: finding and joining dig- ital dots*. PhD thesis, King's College London, 2018.
[12] Karen Kent, Suzanne Chevalier, and Tim Grance. Guide to integrating forensic techniques into incident. *Tech. Rep. 800-86*, 2006.
[13] Flora Amato, Aniello Castiglione, Giovanni Cozzolino, and Fabio Narducci. A semantic-based methodology for digital forensics analysis. *Journal of Parallel and Distributed Computing*, 138:172–177, 2020.
[14] Karen Kent, Suzanne Chevalier, and Tim Grance. Guide to integrating forensic techniques into incident. *Tech. Rep. 800-86*, 2006.
[15] Stefania Costantini, Giovanni De Gasperis, and Raffaele Olivieri. Digital forensics and investigations meet artificial intelligence. *Annals of Mathematics and Artificial Intelligence*, 86(1):193–229, 2019.

[16] Gurpal Singh Chhabra, Varinder Pal Singh, and Maninder Singh. Cyber foren-sics framework for big data analytics in iot environment using machine learning. *Multimedia Tools and Applications*, 79(23):15881–15900, 2020.

[17] Reza Montasari, Richard Hill, SimoParkinson, Pekka Peltola, Amin Hosseinian-Far, and Alireza Daneshkhah. Digital forensics: challenges and opportunities for future studies. *International Journal of Organizational and Collective Intelligence (IJOCI)*, 10(2):37–53, 2020.

[18] Darren Quick and Kim-Kwang Raymond Choo. Impacts of increasing volume of digital forensic data: A survey and future research challenges. *Digital Investigation*, 11(4):273–294, 2014.

[19] Graeme Horsman and James R Lyle. Dataset construction challenges for digital forensics. Forensic Science International: Digital Investigation, 38:301264, 2021.

[20] Quick D& Choo K-KR. Impacts of increasing volume of digital forensic data. Digit. Investig, 11:273–294, 2014.

[21] Abhishek Kumar Pandey, Ashutosh Kumar Tripathi, Gayatri Kapil, Virendra Singh, Mohd Waris Khan, Alka Agrawal, Rajeev Kumar, and Raees Ahmad Khan. Current challenges of digital forensics in cyber security. Critical Concepts, Stan- dards, and Techniques in Cyber Forensics, pages 31–46, 2020.

[22] Nasrin Sultana, Naveen Chilamkurti, Wei Peng, and Rabei Alhadad. Survey on sdn based network intrusion detection system using machine learning approaches.Peer-to-Peer Networking and Applications, 12(2):493–501, 2019.

[23] Devanshi Dhall, Ravinder Kaur, and Mamta Juneja. Machine learning: a review of the algorithms and its applications. Proceedings of ICRIC 2019, pages 47–63, 2020.

[24] Iqbal H Sarker, ASM Kayes, and Paul Watters. Effectiveness analysis of machine learning classification models for predicting personalized context-aware smart- phone usage. Journal of Big Data, 6(1):1–28, 2019.

[25] Ansam Khraisat, Iqbal Gondal, Peter Vamplew, and Joarder Kamruzzaman. Survey of intrusion detection systems: techniques, datasets and challenges. Cyberse- curity, 2(1):1–22, 2019.

[26] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning.Machine Learning, 109(2):373–440, 2020.

[27] Zhe Wang and Tianzhen Hong. Reinforcement learning for building controls: The opportunities and challenges. Applied Energy, 269:115036, 2020.

[28] Shahadat Uddin, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making, 19(1):1–16, 2019.

[29] Iqbal H Sarker. Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2(3):1–21, 2021.

[30] Susmita Ray. A quick review of machine learning algorithms. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMIT- Con), pages 35–39. IEEE, 2019.

[31] Mei Sze Tan, Siow-Wee Chang, Phaik Leng Cheah, and Hwa Jen Yap. Integrative machine learning analysis of multiple gene expression profiles in cervical cancer. PeerJ, 6:e5285, 2018.

[32] Joshua P Parreco, Antonio E Hidalgo, Alejandro D Badilla, Omar Ilyas, and Rishi Rattan. Predicting central line-associated bloodstream infections and mortality using supervised machine learning. Journal of critical care, 45:156–162, 2018.

[33] Loong Chuen Lee and Abdul Aziz Jemain. On overview of pca application strategy in processing high dimensionality forensic data. Microchemical Journal, 169:106608, 2021.

[34] Lian Niu. A review of the application of logistic regression in educational research: Common issues, implications, and suggestions. Educational Review, 72(1):41–67, 2020. A review of the application of logistic regression in educational research: Common issues, implications, and suggestions. Educational Review, 72(1):41–67, 2020.

[35] Steven L Brunton and J Nathan Kutz. Data-driven science and engineering: Machine learning, dynamical systems, and control. Cambridge University Press, 2022.

[36] M Sornalakshmi, S Balamurali, M Venkatesulu, M Navaneetha Krishnan, Laksh- mana Kumar Ramasamy, Seifedine Kadry, Gunasekaran Manogaran, Ching-Hsien Hsu, and Bala Anand Muthu. Hybrid method for mining rules based on enhanced apriori algorithm with sequential minimal optimization in healthcare industry. Neural Computing and Applications, pages 1–14, 2020.

[37] Dijana Jovanovic, Milos Antonijevic, Milos Stankovic, Miodrag Zivkovic, Marko Tanaskovic, and Nebojsa Bacanin. Tuning machine learning models using a group search firefly

algorithm for credit card fraud detection. Mathematics, 10(13):2272, 2022.
[38] Nebojsa Bacanin, Catalin Stoean, Miodrag Zivkovic, Dijana Jovanovic, Milos Antonijevic, and Djordje Mladenovic. Multi-swarm algorithm for extreme learn- ing machine optimization. Sensors, 22(11):4204, 2022.
[39] Nebojsa Bacanin, Miodrag Zivkovic, Fadi Al-Turjman, K Venkatachalam, Pavel Trojovsky`,Ivana Strumberger, and Timea Bezdan. Hybridized sine cosine algorithm with convolutional neural networks dropout regularization application. *Sci- entific Reports*, 12(1):1–20, 2022.
[40] Mohamed Salb, Luka Jovanovic, Miodrag Zivkovic, Eva Tuba, Ali Elsadai, and Nebojsa Bacanin. Training logistic regression model by enhanced moth flame opti- mizer for spam email classification. In Computer Networks and Inventive Commu- nication Technologies, pages 753–768. Springer, 2023. Nebojsa Bacani
[41] n, Miodrag Zivkovic, Marko Sarac, Aleksandar Petrovic, Ivana Strumberger, Milos Antonijevic, Andrija Petrovic, and K Venkatachalam. A novel multiswarm firefly algorithm: An application for plant classification. In Interna- tional Conference on Intelligent and Fuzzy Systems, pages 1007–1016. Springer, 2022.
[42] Ehsan Nowroozi, Ali Dehghantanha, Reza M Parizi, and Kim-Kwang Raymond Choo. A survey of machine learning techniques in adversarial image forensics. Computers & Security, 100:102092, 2021.
[43] Mohammad Manzurul Islam, Gour Karmakar,Joarder Kamruzzaman, Manzur Murshed, Gayan Kahandawa, and Nahida Parvin. Detecting splicing and copy- move attacks in color images. In 2018 Digital Image Computing: Techniques and Applications (DICTA), pages 1–7. IEEE, 2018.
[44] Mauro Barni, Ehsan Nowroozi, and Benedetta Tondi. Detection of adaptive his- togram equalization robust against jpeg compression. In 2018 International Work- shop on Biometrics and Forensics (IWBF), pages 1–8. IEEE, 2018.
[45] Sara Ferreira, M´ario Antunes, and Manuel E Correia. Exposing manipulated pho- tos and videos in digital forensics analysis. Journal of Imaging, 7(7):102, 2021.
[46] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. Unmask- ing deepfakes with simple features. arXiv preprint arXiv:1911.00686 2019.
[47] Gurpal Singh Chhabra, Varinderpal Singh, and Maninder Singh. Hadoop-based analytic framework for cyber forensics. *International Journal of Communication Systems*, 31(15):e3772, 2018.
[48] Nighat Usman, Saeeda Usman, Fazlullah Khan, Mian Ahmad Jan, Ahthasham Sajid, Mamoun Alazab, and Paul Watters. Intelligent dynamic malware detection using machine learning in ip reputation for forensics data analytics. *Future Gen- eration Computer Systems*, 118:124–141, 2021.
[49] Amit V Kachavimath, Shubhangeni Vijay Nazare, and Sheetal S Akki. Distributed denial of service attack detection using naïve bayes and k-nearest neighbor for network forensics. In *2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA)*, pages 711–717. IEEE, 2020.
[50] Paola Barra, Carmen Bisogni, Michele Nappi, David Freire-Obreg´on, and Modesto Castrill´on-Santana. Gait analysis for gender classification in forensics. In *Interna- tional Conference on Dependability in Sensor, Cloud, and Big Data Systems and Applications*, pages 180–190. Springer, 2019.
[51] Anton Yudhana, Imam Riadi, and Faizin Ridho. Ddos classification using neural network and naïve bayes methods for network forensics. *International Journal of Advanced Computer Science and Applications*, 9(11), 2018
[52] T Satya Sudha and Ch Rupa. Analysis and evaluation of integrated cyber crime offences. In *2019 Innovations in Power and Advanced Computing Technologies (i- PACT)*, volume 1, pages 1–6. IEEE, 2019.
[53] Muhammad Faris Ruriawan, Bintaran Anggono, Isaac Anugerah Siahaan, and Yudha Purwanto. Development of digital evidence collector and file classification system with k- means algorithm. In *2019 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, pages 64–68. IEEE, 2019.
[54] Dixit Roy. Naskar, & chakraborty.(2020). digital image forensics theory and imple- mentation. *Studies in Computational Intelligence*, 755.
[55] Muhammad Ali, Stavros Shiaeles, Nathan Clarke, and Dimitrios Kontogeorgis. A proactive malicious software identification approach for digital forensic examiners. *Journal of Information*

Security and Applications, 47:139–155, 2019.
[56] Maryam Hina, Mohsan Ali, Abdul Rehman Javed, Gautam Srivastava, Thippa Reddy Gadekallu, and Zunera Jalil. Email classification and foren- sics analysis using machine learning. In *2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Comput- ing & Communications, Internet of People and Smart City Innovation (Smart- World/SCALCOM/UIC/ATC/IOP/SCI)*, pages 630–635. IEEE, 2021.
[57] Jobin Varghese and C Sathish Kumar. Robust copy-move forgery detection algo- rithm using singular value decomposition and discrete orthonormal stockwell trans- form. *Australian Journal of Forensic Sciences*, 52(6):711–727, 2020.
[58] Turker Tuncer, Fatih Ertam, and Sengul Dogan. Automated malware identifica- tion method using image descriptors and singular value decomposition. *Multimedia Tools and Applications*, 80(7):10881–10900, 2021.
[59] Huan Li, Bin Xi, Shunxiang Wu, Jingchun Jiang, and Yu Rao. The application of association analysis in mobile phone forensics system. In *International Conference on Intelligence Science*, pages 126–133. Springer, 2018.
[60] Timothy Boll´e, Eoghan Casey, and Ma¨elig Jacquet. The role of evaluations in reaching decisions using automated systems supporting forensic analysis. *Forensic Science International: Digital Investigation*, 34:301016, 2020.
[61] Abiodun A Solanke. Explainable digital forensics ai: Towards mitigating distrust in ai-based digital forensics analysis using interpretable models. *Forensic Science International: Digital Investigation*, 42:301403, 2022.
[62] Nighat Usman, Saeeda Usman, Fazlullah Khan, Mian Ahmad Jan, Ahthasham Sajid, Mamoun Alazab, and Paul Watters. Intelligent dynamic malware detection using machine learning in ip reputation for forensics data analytics. *Future Gen- eration Computer Systems*, 118:124–141, 2021.
[63] Felix Anda, David Lillis, Nhien-An Le-Khac, and Mark Scanlon. Evaluating auto- mated facial age estimation techniques for digital forensics. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 129–139. IEEE, 2018
[64] .Felix Anda, David Lillis, Nhien-An Le-Khac, and Mark Scanlon. Evaluating auto- mated facial age estimation techniques for digital forensics. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 129–139. IEEE, 2018.