Computers in Digital Forensics using Machine Learning and Big date.

Dr. Jambi Ratna Raja Kumar
Principal, Genba Sopanrao Moze College of Engineering,
Balewadi, Pune-411045
ratnaraj.jambi@gmail.com
Prof. Dr. B.K. Sarkar
Mahatma Education Society's, Pillai Hoc College of Engineering &Technology
Rasayani Taluka Panvel, Dist, Navi Mumbai, Maharashtra 410207, India.
dr.bksarkar20032yahoo.in

Prof. Dr. Reena Singh
Mahatma Education Society's, Pillai Hoc College of Engineering &Technology
Rasayani Taluka Panvel, Dist, Navi Mumbai, Maharashtra 410207, India.
dr.bksarkar20032yahoo.in

Abstract.
Due to the rise in cybercrimes in recent years, digital forensics is becoming a crucial subject to research in order to gather reliable evidence. In order to recreate events, forensic investigators encounter challenges with data collecting and processing. Machine learning enables investigators to conduct more effective and efficient investigations utilising a variety of algorithms because of the significant contact that people have on a regular basis. A subset of artificial intelligence is machine learning. It is a branch of science that focuses on creating algorithms and computer models that can carry out particular activities without programming, such as dataset training and testing, and its potential to support investigations. In the course of an investigation, digital evidence is examined and analysed using a variety of machine learning techniques, which are reviewed in this paper. Every machine learning algorithm operates on a certain Based on the characteristics, each machine learning algorithm targets a particular area of digital forensics and solves challenges like complexity, data volume, timing, correlation, consistency, etc. Additionally, this study contrasts machine learning algorithms based on accepted standards. Forensic Chemistry can be defined as the practice of application of our knowledge in the field of chemistry to solve crimes. A forensic chemist can assist in the identification of unknown materials found at a crime scene There are several methods that a chemist can adopt from chemistry to help solve uncertainties at a crime scene. Forensic chemists use a variety of instruments to identify unknown substances found at a scene. Some examples of applications of forensic chemistry Spectroscopy techniques

are used to check the purity of materials.Detecting illegal drugs and narcotics using identification and separation techniques.

Keywords: Digital Forensics • Machine learning Algorithms •Investigation • Digital Evidence • Swarm Intelligence

Introduction

Data obtained from databases, computers, and digital pictures are analysed and presented using a method called digital forensics (DF) [1]. There are many various types of data, with many distinct categories and features, produced by the growing number of smart gadgets in our everyday lives. To assist investigators in identifying and preventing unauthorised access to the information acquired, the digital forensics investigation process gathers and analyses data [2]. The majority of the time, once a crime has been committed, the data and evidence gathered from a device can be wiped. For detectives, this procedure is crucial because it enables them to identify the victims and pinpoint the precise nature of the crime [3]. Sadly, it might take a long time to conduct a comprehensive inquiry when there aren't enough human resources available.

Although numerous methods, including Hadoop, may be used to manage the enormous quantity of data gathered by a digital forensics investigator, they are not as effective as the human brain. Instead, to efficiently analyse and gather data, researchers employ machine learning (ML) [4]. This system can pick decisions based on facts and learn from many instances and experiences [5]. Support Vector Machine (SVM), Decision Tree (DT), K-Means, K-Nearest Neighbour (KNN), Nave Bayes (NB), Principal Component Analysis (PCA), Logistic Regression (LR), Singular Value Decomposition (SVD), and Apriori are just a few of the several techniques it incorporates. Each algorithm is in charge of a certain task, such as extracting characteristics, categorising network threats, or spotting distorted photographs, among others.

Digital Forensics and Machine Learning

In the field of research known as "digital forensics," data that has been gathered and saved on various media is examined and preserved. 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]. It has become one of the most crucial areas of security as a result of the growth in cyber threats and assaults. A subfield of artificial intelligence called machine learning is concerned with creating machines that can learn from data. This technology is frequently employed in the fields of data mining, analysis, and behaviour prediction [8]. The issues, models, and phases of the digital forensics inquiry are described in this section.

Digital Forensics

A subfield of criminalistics called "digital forensics" focuses on the legal processes involved in assessing and safeguarding digital material. 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 analysing the data that various digital objects have produced using scientific and technical methodologies [10]. The goal of digital forensics is to gather information that may be utilised to ascertain the details of an occurrence. 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 Investigation procedure for digital forensics. The four techniques and methodology employed in the digital forensics process, according to the National Institute of Standards and Technology, are intended to aid organisations in comprehending the relevance of their investigations. These techniques and methodologies are depicted in Fig. 2. Depending on the intricacy of the investigation, they can be carried out in a variety of methods [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. Below is an explanation of each step.

a) Finding probable sources of this data is the first step in gathering it for an inquiry. Typically, servers, desktop computers, and laptops are where the data is gathered. When examining an organization's operations, analysts should take into account other data sources in addition to more conventional ones. For instance, they can learn about the operations of a company through the logs kept by its Internet service provider [13].

b)

Fig.1. Digital Forensics Investigation Process.

b) Examination: The second stage aims to examine the data that has been collected. Through the use of digital forensics techniques and tools, the necessary pieces of information from the data are extracted. Moreover, defining the data files that contains information of interest, including information concealed through file compression, access control, and encryption [13,14].

c) Analysis: An analysis is a process that involves carrying out scientific procedures in a scientific setting to produce elements such as identifying people, places, and events, as well as determining how these elements are related [15]. This process involves analyzing data collected from various sources. For instance, an IDS log may contain information about a specific user, while audit logs may include details about a particular host, with the help of tools such as security event management software, it can be easy to correlate and gather data [14].

d) Reporting: The investigation's last phase entails examining the information gathered during the analysis phase and presenting the results to the analyst in a formal report. Finding the reason for an occurrence or giving a precise explanation might 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 [1][5].

B. Models for digital forensics. There are various investigation models used in digital forensics, including 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) [1][6]. Each model has been created for a certain stage or activity. The digital forensics models and associated operations are shown in Figure 2.

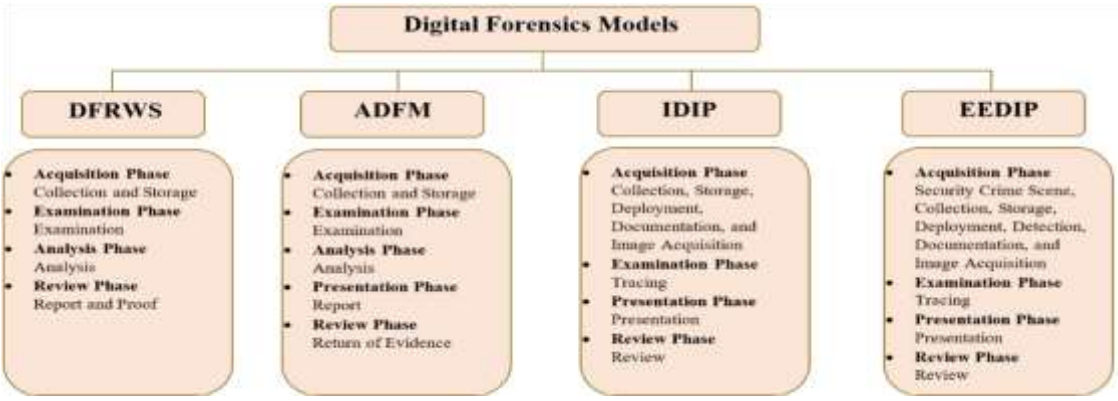

Fig.2. Models of the Digital Forensics

C. Threads in digital forensics. The growing number of digital devices has slowed the development of digital forensics. The complexity of the hardware, software, and mobile devices that use encryption makes obtaining digital evidence

very challenging. Therefore, in order to address the problems the industry is presently experiencing as a result of this, new strategies and methods are needed. According to Montana et al. of the International Journal of Organisational and Collective Intelligence, consistent digital forensics techniques and processes cannot be developed because of the increasing diversity of file formats and operating systems [1][7]. The complexity of digital techno the amount of data that can be collected and analysed has become more challenging. With new data formats, such as the low binary, it is now possible to collect and analyze large volumes of data [1][8].

According to Horsman et al., intricacy is an additional difficulty. The challenge of creating tools that can swiftly analyse the data collected grows as more data is collected. A major problem is also the absence of standardisation in the presentation and preservation of digital evidence. The exchange of digital proof is difficult. By establishing a standardised set of processes, this problem might impact how successfully law enforcement exchanges information during investigations [19]. According to Quick et al., correlation and consistency are the biggest challenges when developing digital analysis tools. Since the evidence is collected from different sources, the data must be analysed and correlated correctly. This can be time-consuming and drain an resources for the inquiry [2][10].

Pandey investigated the time-lining difficulty that experts in digital forensics have when different sources offer opposing interpretations of the same evidence. The effectiveness of the inquiry may be impacted by this problem. Another key problem that endangers the growth of the sector and the reliability of the data is the lack of information about the most recent digital forensics techniques. Because forensic science is developing so quickly, professionals in the area need to be prepared to employ the latest technologies efficiently. This problem may stop the creation of new digital technologies [2][1].

Machine Learning

Machine learning, which enables computers to learn and analyse without the need for extra training, is one of the most popular methods to artificial intelligence (AI) in [2][12]. It is capable of automatically categorising and predicting inputs [2][3]. In a variety of contexts, including security and fraud detection, this technology can employ the proper algorithms. Supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning are the four primary divisions of machine learning. Examples of supervised learning include translating inputs into outputs.

It makes use of training data that has been labelled with different training examples. The most often utilised methods in this procedure are regression and classification [2][4]. Unsupervised learning, sometimes referred to as the clustering approach, may be used to find hidden patterns and structures in datasets [2][5]. While using both unlabelled and labelled data to carry out various tasks is the emphasis of semi-supervised learning, a sort of machine learning. Between supervised and unsupervised learning [2][6] is where it lies.

Reinforcement learning, which may resolve issues or regulate specific circumstances, has a focus on rewarding behaviours and penalising those that do not meet standards [2][7]. The following is a description of these algorithms:

A. Algorithm for support vector machines. Support Vector Machine is capable of handling both classification and regression issues. SVM classifies the objects using examples from the training data set. Depending on the kernel function, it can handle structured and semistructured data and carry out sophisticated operations. This approach finds a hyperplane that divides each data item into two classes after taking the amount of characteristics into account. It maximises the marginal separation between the two classes while minimising mistakes [2][8].

B. Algorithm using Decision Trees. A decision tree is a learning technique that may be applied to problems requiring classification and regression. It is simple to understand and may relate test results to the categorization of data points. A decision tree model takes into account the numerous types of decision logic and represents them in the form of a tree. the root node, which is the topmost node in a DT tree. A decision tree's interior nodes represent tests pertaining to the input variables or characteristics. The categorization algorithm branches to the appropriate child node when the test is finished. Until the leaf node is prepared to make a decision, this process continues [2][9].

C. The algorithm K-Nearest Neighbour. A non-generalizing learning technique, the K-Nearest Neighbours algorithm doesn't concentrate on developing a broad model. It maintains all training data instances in an n-dimensional space. The K-Nearest Neighbours algorithm utilises data to categorise new data points and is capable of performing a number of tasks, including classification and regression, which manage data training and offer correct data based on the quality of the data [3][10].

D. No-nonsense Bayes algorithm. An unsupervised learning method called Naive Bayes is used for classification or grouping tasks. It may be used as a clustering approach and does not need the declaration of a result [3][1]. Only a little quantity of training data is needed for the algorithm to predict the required parameters. Nave Bayes is a supervised learning method since it depends on both the target and the input variables. It creates a tree of Bayesian networks, which are tree models based on outcome probabilities, as a classifier [3][2].

E. Algorithm K-Means. A quick and effective way to divide datasets into K centres is to use K-Means. Because it is more effective when the variables are substantial, it is comparable to hierarchical clustering. The implementation and data interpretation of this method are hence its efficient components [3][10].

F. Algorithm for Principal Component Analysis. The principle component analysis is a method that transforms the observed values of different potential correlated variables into linearly uncorrelated values. By using the Orthogonal Transformation algorithm, it may be completed rapidly and easily. As a result, the model may be computed without the requirement for prior knowledge. PCA offers a number of additional characteristics, such as data feature categorization and estimate, in addition to data clustering and classification [3][3].

G. Algorithm for Logistic Regression. Machine learning uses a logistic regression model to address categorization issues. It aids in determining which class belongs to a certain instance. Given that it is a probability, the model's result falls between 0 and 1. Therefore, it may be utilised as a binary classifier [3][4].

H. Algorithm for Singular Value Decomposition. In matrices, the idea of the SVD factorization technique is frequently employed. By taking into account the dominating patterns, the SVD method produces a low-dimensional representation of a high-dimensional data set. This approach relies mostly on data acquired, not expertise or common sense. Singular values and the decomposition approach may be used to extract invariance characteristics from a picture or a signal, respectively [3][5].

Apriori algorithm, first. The Apriori method is popular in data mining because it uncovers connections across disparate data sets. It typically uses the candidate generation approach to mine item sets. Additionally, it is made to function effectively in a database with numerous transactions. The necessity for "n" numbers of frequent item sets in the database scans is one of the issues that may cause it to perform less well [3][6].

Swarm Intelligence

Swarm intelligence is a term used to describe algorithms that get their inspiration from the behaviours of diverse creatures seen in nature. The optimisation of synthetic bee colonies and particle swarm optimisation are two of the most notable instances of this form of met heuristics [3][7]. The development of several computational models and algorithms created to solve the complexity of real-world issues in mathematics, statistics, and Blockchain has heavily used the met heuristic method. Among them is the optimisation of tasks involving artificial neural networks [3][8].

According to a study of the literature, swarm intelligence approaches should have been employed more to enhance the performance of machine learning models, even if the met heuristic approach has been widely used in the development of many computer models and algorithms. This is unexpected because other study fields have effectively applied these techniques. A pair of algorithms created to increase the effectiveness of activities carried out by Extreme Learning Machines (ELMs) are among the most successful uses of this form of met heuristics [3][9].

Algorithm in Digital Forensics/ Results

Ahmed et al. suggested a brand-new approach based on the Kolmogorov-Smirnov test and singular value decomposition for identifying copy-move forgeries. It entails employing a steerable pyramid to extract picture characteristics from various blocks, after which the indices of the original blocks are saved with feature vectors that match to the relevant features of the pixels. In digital picture forensics, the brightness, contrast, image blurring, and colour reduction processing methods are all investigated. The recall, accuracy, and F1 score of the suggested approach were good. It received a 95% rating for brightness adjustment, 77.5%, 82.7%, and 75% [5][7] for picture blurring.

Varghese et al. suggested a technique to extract characteristics from each block using a mix of discrete orthonormal Stockwell transform and singular value decomposition in order to identify a copy-move forgery in photos. Two threshold settings allow for the lexicographic sorting of the generated characteristics, which may then be separated from other pictures. The simulations show that the suggested approach is more reliable and invariant than previous cutting-edge methods for identifying copy-move forgeries. It also shown strong precision in a variety of tasks, including rotation [5][8].

A strategy for categorising different kinds of malware and creating a powerful anti-malware model was put out by Tuncer et al. The suggested technique extracts features using a local binary pattern (LBP) and SVD, then uses PCA to lessen their complexity. The proposed technique, which was based on the LBP-SVD-LTPNet architecture, had an 88.08% success rate. In terms of accuracy, it outperformed deep learning techniques [5][9].

Conclusion

The field of digital forensics has expanded in various ways. Forensic analysts have demonstrated the numerous challenges they encounter while analysing massive data, such as photographs, video, etc., that may help expose events. In digital forensics, several brand-new problems are starting to surface over time. As a result, automated systems and clever procedures were developed to make it easier for investigators to do their jobs. This study has verified a number of machine learning (ML) methods, including SVM, KNN, DT, PCA, SVD, K-Means, NB, ANN, LR, and RF, to address issues in digital forensics. In order to provide evidence in court, algorithms distinguish between real and fraudulent data. The study concluded by summarising the recommended practises for each algorithm in digital forensics in light of its characteristics, benefits, and drawbacks. K-Means focuses on retrieving deleted digital evidence from memory locations in accordance with the recommended research papers. The greatest techniques to use in an image forensics inquiry are the SVM, PCA, and SVD, while the KNN and NB help network forensics. In the past several years, machine learning researchers have made considerable strides in creating computers that think like humans. They now carry out complicated activities and make choices after careful consideration. Although there has been improvement, machine learning still has several drawbacks, including moral concerns, a lack of interpretability, a lack of repeatability, and a shortage of data to train the computers.

References

1. Joakim K¨avrestad. Fundamentals of Digital Forensics. Springer, 2022.
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, 2022.
3. Graeme Horsman. Tool testing and reliability issues in the field of digital forensics. Digital Investigation, 28:163–175, 2021.
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 cybersecurity: 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 Technologies (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 digital 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.