Automated Assessment of Artefact Relevance Using Artefact Metadata and Correlated Timeline Events

Suvarna Chaure[1,2] and Vanita Mane[1(✉)]

[1] Department of Computer Engineering, Ramrao Adik Institute of Technology, D Y Patil Deemed to be University, Nerul, Maharashtra, India
vanita.mane@rait.ac.in

[2] Department of Computer Engineering, SIES Graduate School of Technology, Nerul, Maharashtra, India

Abstract. Digital forensic data backlogs that span years are a major concern for law enforcement organizations globally and can hinder court processes. This problem is brought on by the increasing number of cases that call for digital forensic examination as well as the exponential rise in the volume of data that is involved in each case. Leveraging the previously analyzed instances and components of digital forensics can be crucial in resolving this issue. There is a chance to use evidence-based automated artificial intelligence training systems for processing by classifying artefacts according to their relevance. These systems can be very helpful to researchers in organizing and compiling data. A method for assessing file artefact significance is provided by one suggested methodology, the Automated Artefact Classification during Digital Forensic Investigation (AACDFI), which is especially helpful during the examination stage of digital forensic investigations. This method allows the utilization of previously identified relevant files to categorize newly discovered files using machine learning algorithms. These files are identified during the acquisition step through an automated artefact detection model, where trained models play a crucial role. By utilizing filesystem metadata and associated temporal events for each artefact, the method assigns a relevancy score based on file similarity.

Keywords: Digital Forensics · Prioritization of Artefact · Artefact analysis. Machine Learning

1 Introduction

In the digital age, the proliferation of automated devices and the exponential growth of digital information have significantly transformed the landscape of forensic investigations. Law enforcement agencies and digital forensic practitioners are faced with the daunting task of analyzing vast amounts of digital evidence to uncover critical insights for criminal investigations and legal proceedings. However, the sheer volume of data,

P. Pareek et al. (Eds.): IC4S 2024, LNICST 599, pp. 392–407, 2025.
https://doi.org/10.1007/978-3-031-77081-4_31

coupled with resource constraints and time limitations, presents substantial challenges in effectively and efficiently processing digital artefacts. The beginning of the twenty-first century is considered to have been the important age of digital forensics [1]. Since then, the varied individual digital gadgets, the quantity of information kept, and the widespread use of cloud facilities have made digital forensic investigations more challenging [2]. The huge quantity of information that regulation enforcement agencies around the world handle is too big to evaluate in a appropriate way. Large backlogs of digital forensic information accumulating over several years have consequently become the norm [3].

Because of the massive amounts of data that accompany them, it is difficult to address the growing amount of cases needing digital forensic examination using the present investigative methodologies [4]. Reducing the amount of data that specialists must analyse or focusing their attention on the most crucial information first can increase the effectiveness of the study. Quick et al. [5] presented a selective imaging method for digital forensic reduction. In the Select process, files are shown and chosen using filters to generate a subsection. These filters primarily target the operating system (OS), applications, browsing information, user-generated files, emails, documents, photos, music, videos, and other file system artefacts. However, the total volume of electronic materials is sometimes still challenging in many situations, such as child sexual exploitation material (CSEM) cases, even after selectively shooting multimedia files. This could have serious psychological repercussions for the investigator, such as shocking stress disorder. Though, in some circumstances, the information sizes may remain high even after information decrease has been completed [6].

Automated evidence analysis methods are needed to enhance the amount of evidence's categorization. Beebe et al. practice gathering for text search consequences to rank evidence according to its matching relevancy score, which increases retrieval effectiveness [7]. Moreover, document clustering may benefit from the document's content. When a pertinent data was originate in the same cluster, Da et al. devised a technique that let the investigator to examine other files from that cluster first [8]. Le and colleagues converted malicious machine readable data into disk images for the purpose of training deep learning models for malicious data sorting.

During the inquiry phase of a project, a method known as timeline examination is used to determine the chronological sequence in which specific events took place on an object [9]. A timeline for additional research is established by the usage of registries and log archives, which are logs of user activity. Though, it is challenging for investigators to comprehend millions of low-level events in the absence of the whole fact. Automatic top digital incident the generation is one of the solution [10]. Additionally, storage system traces document user behaviour on a device [11]. For instance, the date-time tags on this file indicate the precise moment when it was downloaded and stored to the machine. The file system information holds a plethora of useful information that might be discovered during an investigation [12].

Data science is the skill to gather information and have the knowledge to understand, process, extract value from, interpret, and explain it [13]. Guarino et al. [14] state that data science and digital forensics will work together to ensure the timely analysis of massive volumes of data due to big data difficulty that digital forensics is currently

facing. Machine learning (AI) filtering, safer practitioner representation, and automatic forensic devices are critical due to the huge capacity of digital artefact that needs to be examined. Due to the previously mentioned difficulties, a computerised system for file artefact analysis is needed the big data challenges. Rapidly determining which file evidences are most possible to be important to the examination might help expedite the legal procedure significantly. An approach to file artefact ranking is presented in this study. This work's contribution can be summed up as follows:

1. To put into practice a method for locating possible artefacts that is based on machine learning techniques.
2. To create a method based on data reduction techniques for automatically rating file artefacts according to their potential significance.
3. To create a program that will automatically take in data from timelines that are made.

The remainder of the document is organized as follows: An overview of background information, a review of the literature, and details of current implementations are provided before information on the current level of privacy protection is provided in Sect. 2. On the basis of it, the third section of the model for digital forensic investigation will examine automated artefact classification. The suggested machine learning methodology is shown in Sect. 4. Results are included in Sect. 5. Data generation is discussed in Sect. 6.

2 Literature Survey

Digital forensic procedure models are similarly diverse as the variety of digital evidence sources and technologies [15]. There isn't a single process model that works for every kind of research. Reducing the quantity of information needed for time-consuming, personal evaluation can fasten the whole investigation process and considerably contribute to clearing the backlog of digital forensic cases that is all too familiar in law enforcement bureaus worldwide. Investigators can gain from performance clustering and improved partnership between the different roles in an examination by focusing the processing of digital forensic evidence. Since December 2010 [16, 17], the Netherlands Forensic Institute has been conducting forensic investigations applying HANSKEN, a Digital Forensics as a Service technology.

The minimization of needless manual file scrutiny is made possible by data deduplication based on hash digest comparison. One of the main instruments in digital inquiry is hashing [18]. Many uses exist for hash-based methods, such as similarity hashing, which is the process of discovering things that are like one another and known objects [19]. Most popular operating systems and application packages have a record of common hash values available in the National Software Reference Library (NSRL1). You can remove files that are known to be innocuous by using this list.

In 2016, a deduplicated digital forensic procurement and examination system that could be coupled with a DFaaS system, like HANSKEN, was presented. The approach made it possible to identify illegal or important file artefacts early on in an inquiry and avoided the need to repurchase earlier-viewed and recognised files at the time of acquisition. It has been demonstrated that our method can rebuild forensic sound disk images from the deduplicated storage in 2018 [20].

A broad range of techniques are collectively referred to as machine learning, which allows machines to run methods on the basis of predefined commands and gathered data. This gives the machine the ability to learn without human guidance, adapting its algorithm to the circumstances and reprogramming itself. Examples of this include Google and Siri when they are performing voice-activated tasks. Moreover, aberrant activity is tracked by video surveillance.

By employing a sophisticated information architecture with numerous layers to acquire and classify data at multiple levels of abstraction, deep learning is the next stage of machine learning, where a computer can carry out jobs based on images, text, and music, imitating how the human brain accomplishes it. This is exemplified, for example, by building an official repository structure of educational institution important signs to solve issues like identity identification.

In order to choose the optimal neural network models of the vessel routes and subsequently get control activity, neural networks consist of a machine/deep learning-based pattern identification system. This enables neural networks to learn from observational data and provide their own solutions, like a fuzzy regulator-based auto-steering gear system. Natural language processing computers employ machine learning and natural language processing to evaluate language and speech as it is uttered. Developing swarm intelligence and active systems, as well as implementing intuitive software for human-computer interfaces that may be utilized in e-learning and educational settings, are a few examples of this.

Expert systems are made up of software configurations that help produce responses to specific questions submitted by a user or by another software package. Expert knowledge is dedicated to a specific section of the application that uses reasoning to obtain answers based on contextual data and subsequent decision-making. However, visualization often fails to identify essential events since a disc picture often contains a large number of digital events. There are often too many temporal events for human analysis since every user action has the potential to trigger several digital events at an abstract level. It is challenging for investigators trying to piece together the device's narrative to explain millions of low level events. This drastically decreases the amount of events and makes the information much easier to understand. The digital events from multiple sources are integrated into one timeline. A framework called log2timeline (plaso) makes it easier to create a "super timeline" that incorporates electronic occurrences from the OS registry, records, file systems, and application software logs.

This applies to both the file system and the device access levels. The log2timeline has been considered in great detail in the area and is the origin for several notable additional study. Log2timeline generates *.csv log files, which Timeline2GUI was designed to evaluate. In 2020, an abstraction-based timeline restoration method was planned using the timeline information delivered by log2timeline.

To assist with specific tasks, machine learning creates models based on data attributes. Examples of these models include classification models for spam email recognition and regression models for evaluating the urgency of incoming emails. Two supervised learning methods that have been used to digital forensics problems are regression and classification. Both methods rely on tagged datasets being available.

Leveraging the outcomes from the analysis step is made possible in digital forensics investigations through supervised machine learning. A few research techniques are offered to aid in subsequent investigations by using the results of the first investigation to construct machine learning model. Marturana et al. presented a machine learning-based method for electronic device evaluation. Gadget are categories as illegal or legal based on system factors that represent the consumer's behaviours, like the quantity of office/pdf files, the amount of zipped files, the amount of installed applications, the maximum image size, etc. There was also discussion of cases pertaining to intellectual property violation and CSEM transfers.

A technique for automatically identifying incoming file evidence is described by Du et al. Using the record information, classification models are trained on known benign and malicious files. When examining previously unseen file artefacts, the trained model can determine whether or not they are likely to be significant to the investigation. There is a lot of promise for improving investigative efficiency through the automated use of artificial intelligence in digital forensics. Attributes are the pillars of machine learning, according to Flach. The "experience" that has been stored from processing previous investigations can be used to help label data for automatic classification model training.

When there is a tight timeline, the most useful evidence from digital forensics is gathered early in the investigation. Through the triage process, devices and artefacts are ranked according to priority or importance. Many research in the area of digital forensic triage have been carried out in an attempt to develop the process overall.

Rogers recommended utilising a digital forensic triage process methodology while conducting the inquiry. Depending on the type of case—financial crimes, drug activities, CSEM, etc.—file significance varies. Usually, the triage strategy is grounded in practical experience. The devices from the crime scene are often briefly inspected during the triage phase, which is followed by a further in-depth investigation in the digital forensic lab to uncover further pertinent artefacts.

In multi-device investigations, triage reduces the burden. The "Behavioural Digital Forensics Model" from 2018 describes the sub-phase of prioritising devices for inspection. The more file artefacts that are discovered during an investigation, the longer the examination procedure gets. For many different types of cases, examining image files is essential. Additionally, while searching for certain terms on file artefacts, a huge amount of outcomes are regularly reverted.

Through the inspection of a huge amount of file evidences in a short period of time, triage techniques help prioritize work in an investigation. Beebe et al. developed search hit relevance rating approaches to lower the critical load associated with text string search. The training phase of the Support Vector Machine (SVM) model produced a linear discriminant ranking function. Using historical data, a suggested feature list was created; in the experimentation, the rating score was derived from 18 features.

3 Automated Artefact Classification During Digital Forensic Investigation(AACDFI)

During an investigation, it is typical practice to detect known illicit files by relating artefact hash values to a recognized records. Locating known critical files on the device can yield more data than simply having them there for additional research. Using these

identified files, the research's recommended method builds a framework for classification to identify comparable files that may be more pertinent to the examination (see Fig. 1) shows the procedure, which uses records of recognized files retained from the examination of earlier investigations. The following phase involves using the digital events linked to the appropriate documents that have been discovered to train an algorithm for the examination of the unidentified files. The training machine learning model provides a relevant value to each item; the artefacts are subsequently organised based on this score prior to analysis. File artefacts having digital behaviors associated with them which are more comparable to known unlawful records are the ones that are more pertinent to the inquiry. To determine user behavior for every single evidence, a timeline is categorized to show only the events that are pertinent to the object. The model is built using various attributes that have been taken from each file artefact's past. You can get details on metadata, times of access, and the content's updates by using this file artefact timeline. Modern forensic technologies study many types of artefacts independently, such as database forensics, emails analysis, video or audio forensics, surfing the web analysis, etc.

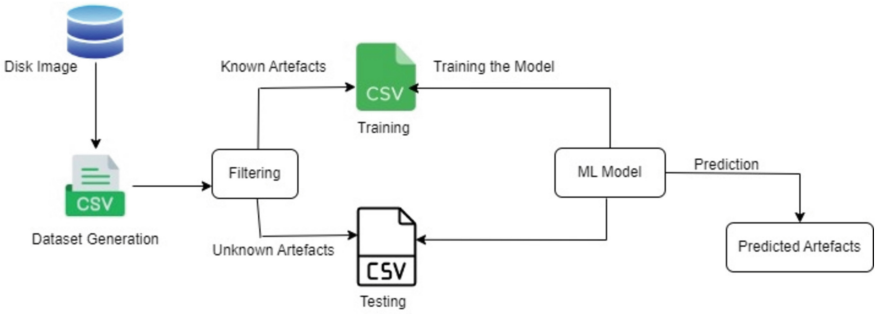

Fig. 1. Automated Artefact Classification Module during Digital Forensic Investigation

4 Relevance Based Machine Learning Approach

This paper describes a method for assisting in the prioritisation of file items that need to be examined by hand.It can be used in an inquiry after the information reduction phase. Data deduplication and hash records evaluation techniques can find files that were previously undiscovered and make known lawful and benign file artefacts visible. Machine learning models can be constructed on existing files to aid in the identification of unknown files. The idea is that files that have "behaviour" comparable to illegal documents are more pertinent to the inquiry and ought to be suggested for additional research.

This method involves of the following steps: 1) reduction of data and deduplication to extract interesting but unknown files in addition to known files. 2) Plaso's building of a "Super Timeline" using an image of a disc. 3) Establishing a timeline for the artefacts. 4) Removing features of the file artefact from the timeline. 5) Train the model with all

of the known file artefacts. 6) Using this methodology, ascertain the relevance score for unidentified formerly undiscovered file artefacts.

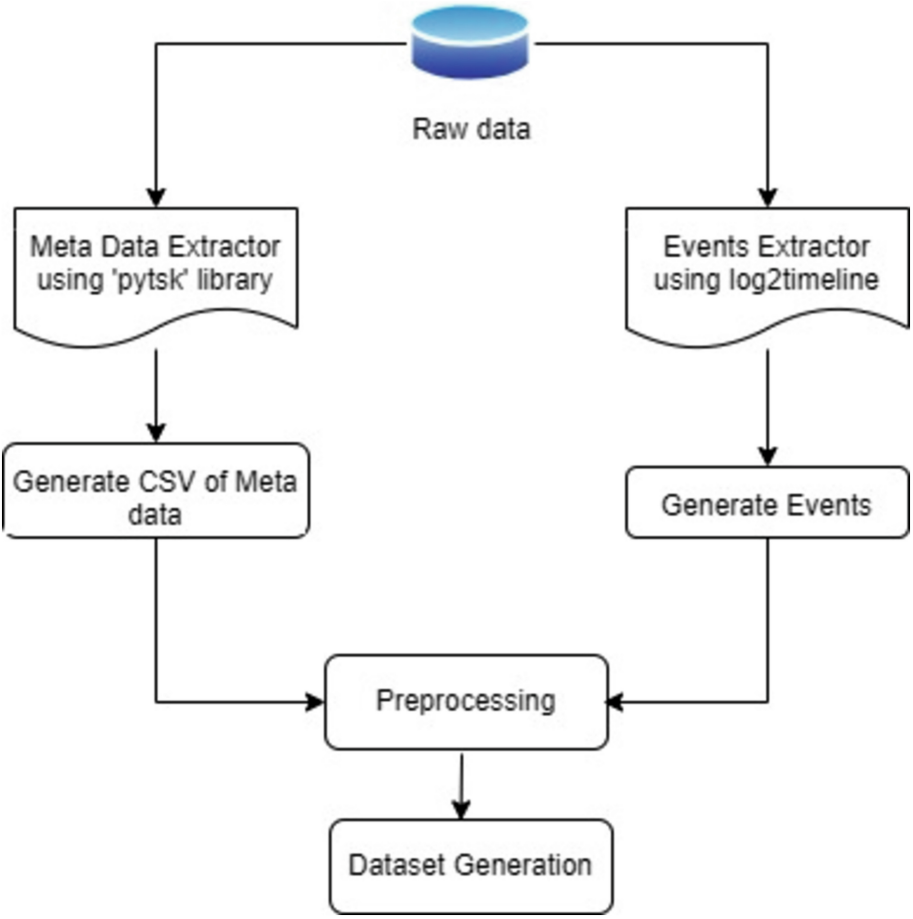

Fig. 2. Data Extraction Process

Filters are constructed using traditional methods of data reduction or triage, which rely on investigative expertise. For instance, searching for a record or image in a financial crime—such as scanned papers—might result in too many results if you only filter by file type. If you search for a specific keyword, you might only get extremely few or no results. When there are too many files to manually comb through, a relevancy score is employed in this study to rank the file artefacts. The machine learning model that produced this relevancy score was trained using well-known file artefacts. As shown in (see Fig. 2), it is a combination that takes into account all of the provided attributes, with more related feature values producing a higher relevancy score (Table 1).

The machine learning library is used in this work are available from Scikit-learn3. Linear modelling techniques, such as Linear SVM and Logistic Regression, can yield

Table 1. Pseudocode for Data Extraction process

Start
Input raw image (disk image)
Dataset Generation in csv
- use '*pytsk3'* for file metadata generator
- use '*log2timeline(plaso)'* command for timeline generator for generating events
Merge two CSV's (metadata and timeline)
Apply filtering using approximate matching algorithms using python library *'pyssdeep'*
Find relevancy score of filtered data using linear modelling i.e. linear SVM
Output categorization of data (relevant, non-relevant)
Stop

coefficients. Additionally, the significance of each characteristic may be determined with flexibility thanks to Random Forest modelling. The script that follows demonstrates how the model's coefficients are obtained. Every file in the FRF is intended to be classified by the machine learning system as either Potentially Conclusive (PC) or Potentially Indecisive (PI) evidence, as was previously mentioned. To put it in formal words, the machine learning approach—a supervised learning technique—is used to solve a two-class classification issue. After the machine learning system has been trained on a few examples of comparable issues that have been addressed, it might start predicting for a new situation. It is imperative that the instructional materials encompass a substantial quantity of instances of every category of case that the investigating agency managed to resolve prior to the solution being used broadly to generate a more comprehensive solution.

Equation (1) can be used to express machine learning algorithm.

$$\mathrm{Y} = \int Xdp(Y/X) \tag{1}$$

where dp(Y/X) represents the conditional probability for y assuming x is true and Y indicates the degree of learning. Any supervised machine learning model typically operates in this manner. We must forecast the value of y for a given value of x.

The supervised learning strategy could be applied using a various learning methods such as Decision Trees or Random Forest, which yield excellent results when the data used for training set is small and the attribute set is sufficiently robust. The learning strategies outlined above, according to the authors, are suitable for the classification job (PC vs. PI) for evolving predictive models for similar scenarios using a comparatively small training data.

An investigating agency could attempt various algorithms such as Support Vector Machine (SVM) and k-Nearest Neighbors (KNN) suppose they possessed a compilation of a significant amount of similar cases, say hundreds or more instances of monetary fraud. Next, the files are arranged in order of relevancy score, from highest to lowest. The framework ensures that the investigator only views the files that are most relevant to her; the others are hidden from her view. The investigator requests the following files, if needed.

The created file timeline includes all of the information that is currently accessible for the creation, modification, and access of each file. A machine learning model that predicts whether undiscovered file artefacts could be pertinent to the inquiry is constructed using the file timelines linked with the known file artefacts. Plaso is utilized to create a timeline of the disk image. The created timeline is kept in a CSV file for later examination. Every row in this csv file represents a single event, complete with source, type, description, and other details. For each specific file, a single CSV file is made to record the digital events connected to it. It is produced by looking up the file name (or previous names) in the timeline. The following chronology features have been selected for this investigation: inode, date, time, MACB, filename, type, source, source type, datetime, and desc. The Plaso timelines is analyzed and file artefact timelines are created using the Pandas2 programme. Timeline-based data can be obtained using the tools. These features can be classified into one of three categories. Digital events: The type of event, including the total amount of digital incidents, the amount of updation in the contents, the amount of updation in metadata, the number of events connected to browsing history, and so on.

Date and time: An event's timing, including the file's creation and final access times, is significant. The timestamp needs to be translated to a category value in order for the model to recognise it. Features like time (early morning, morning, afternoon, night, late night), date (month, workday, weekday), and so on can be classified.

Keywords: The frequency with which a certain research keyword emerges could be another component of the model. For example, records of sales, client data, lists of precursor chemicals, or instructions for manufacturing for drugs might all be considered documentation relevant to a drug-related investigation. These are synonyms that you can use as features. Terms from timelines of files that are known to be illegal can also be added to this list of keywords.

Timelines for file artefacts offer a multitude of easily available features. The amount of features must match the dataset in order to produce the most effective results. Feature selection strategies can be utilised to figure out which features are best suited for a model. The most significant attributes can be identified in order to develop the effectiveness of a machine learning model. Though, it's crucial to find a balance because an overfit model can result from having too few characteristics.

File timelines and file system metadata are the sources of features in this work. The most popular events (together with the related kind, source, etc.) may be used as features; that being said, a timeline may contain a large number of unimportant events, such OS events. Selecting features has the dual purpose of highlighting the most crucial elements that are readily available and preventing any crucial aspects from being overlooked. Which attributes, and how much of them, should there be? Usually speaking, huge datasets can manage more attributes. Therefore, it might be desirable to have fewer features in order to maintain a usable performance when the dataset is constrained.

5 Results

The digital event and metadata extraction tools function on the raw disc image. The input is disc images that are raw. A CSV file is the end result. The utility for extracting metadata uses Pytsk. The Sleuth Kit comes with a Python binding called Pytsk. The

metadata of the file system was retrieved using it as part of this inquiry as shown (see Fig. 3).

```
[ ]  pip install pytsk3
     import pytsk3
     def extract_data(image_path, output_dir):
```

Fig. 3. Data Extraction using pytsk3

The timeline produced by log2timeline mentioned(see Fig. 4) includes digital events from the Windows registry, file system, browsing data, download history, and other sources. Every event has details about its kind, filename, source, and other details.

```
pip install plaso
log2timeline.py --status_view none --hashers all --parsers all -z UTC -f csv output.csv image.E01
```

Fig. 4. Timeline extraction using plaso

A digital event is represented by each row in the created timeline, and the inode can recognize the specific file artefact that the event is connected to. Digital events related with each file artefact can be gathered via the inode. When many partitions are involved, this distinct identity needs to be preserved, which calls for the use of both hash and inode. The combination of time and metadata yields a plethora of information about each file artefact. The code to combine two CSV files is shown (see Fig. 5).

```
[ ]  import pandas as pd

     def merge_csv_files(file1_path, file2_path, output_path):
```

Fig. 5. Merging Metadata and timeline CSV

Not every attribute is useful for teaching models of machine learning. As an example, the prediction job does not profit from numerical features generated at random, inode values, or artefact hash values. Every machine learning model has a distinct aim that requires the use of feature modification. The data provided for model training can be included, altered, or eliminated for every job using features modification.

In reality, the completed design serves as a tool for creating a novel aspect that addresses the issue at hand. Example of hoe features can be extracted from given file is mentioned (see Fig. 6), which will be helpful for finding relevancy of files.

It works by dividing the file artefacts into known and unknown categories by comparing the hash with the database that is previously known. Approximate matching algorithmic methods like ssdeep, sdhash, and others are used for this shown (see Fig. 7). The baseline algorithms' results show that type 2 errors, or False Negatives (FN), are

```
[ ] relevant_columns = ['Owner']
```

```
[ ] for tab in sheet_name:
      df = pd.read_excel(xls, tab)
      print(tab)
      for column in relevant_columns:
        for columns in df.keys():
          if column == columns:
            new_df = df[column]
            print(new_df)
          else:
            pass
```

```
Summary
Operating System Information
0    BillyBob
1    BillyBob
2         NaN
Name: Owner, dtype: object
```

Fig. 6. Feature Extraction Owners info

```
import pyssdeep

def compute_fuzzy_hash(file_path):
    with open(file_path, 'rb') as file:
        data = file.read()

    fuzzy_hash = pyssdeep.fuzzy_hash_filename
    return fuzzy_hash

if __name__ == "__main__":
    file_path = "/content/sample_data/california_housing_test.csv"

    fuzzy_hash = compute_fuzzy_hash(file_path)
    print("Fuzzy Hash:", fuzzy_hash)
```

```
Fuzzy Hash: <function fuzzy_hash_filename at 0x79a0f2f7a5f0>
```

Fig. 7. .Filtering using pyssdeep

difficult for any algorithm to handle. The high FN values are also inappropriate from a digital forensics perspective because they allow legitimate evidence files to elude the investigator and appear to be innocent files. However, as they mark innocent data as potential evidence, False Positives (FP), commonly referred to as type 1 errors in machine learning, can also provide difficulties for digital investigators. However, all the FP would be easily recognized at that moment because the investigator obtains all of the files anticipated by the recommended machine learning approach beforehand for final decision-making.

Comparative analysis is done (see Fig. 8) using various machine learning algorithms such as SVM, multicast logistic regression, decision tree and random forest and it is found that multicast logistic regression gives better accuracy of 98% than other algorithms.

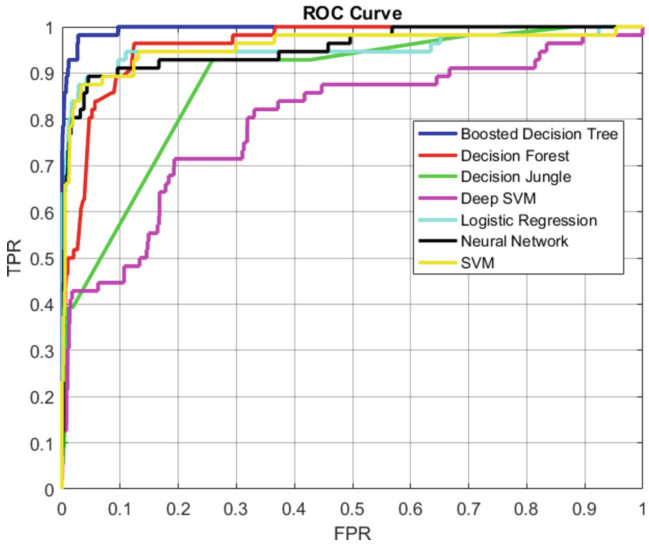

Fig. 8. ROC Curve

The PF, PCF, and NPF data were grouped by the authors using the K-Means and Hierarchal clustering algorithms as shown in (see Fig. 9). Furthermore, because various media files are kept in PF, the authors allotted each one a unique numerical code. Each PCF document file is assigned a code, which is unique.

5.1 Experimental Data Generation

Particularly for the experiments described in this paper, experimental data of Hacking case, data theft and financial fraud were collected from CFReDs portal. It is easy access to documented digital forensic picture datasets using this site. These datasets can help with many activities, including as teaching general practitioners, testing tools, and coming up with additional unexpected uses that the dataset users can come up with. They can also help with building familiarity with tool behaviour for specific tasks. The majority of databases include a description of the various types and locations of important artefacts. To locate datasets by author, year of creation, or dataset properties, descriptions and locating aids are provided. Various file types of Hacking Case is shown (see Fig. 10).

The timeframes and log files from the dataset were then extracted. This process is seen in Fig. 3. Plaso was used to convert the virtual hard drive (VHD) images of the VMs to raw format and extract the corresponding timelines. Every digital event connected to any of the file evidence on the disk image is represented in this timeline. To construct the unique file artefact timelines, this is categorized for every file artefact. Features for training the model are high frequency terms, sources, and kinds of digital events.

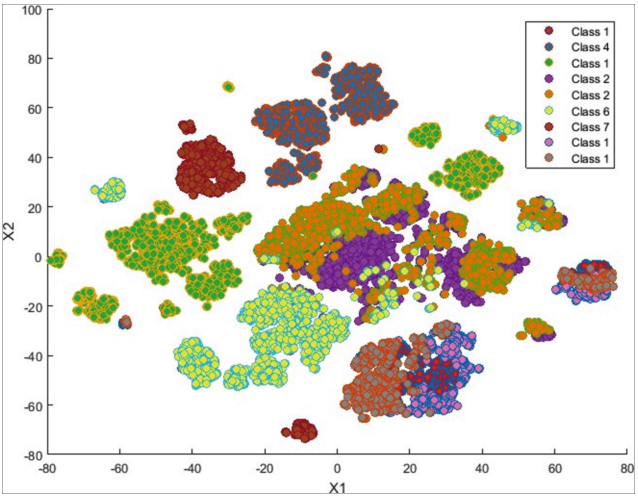

Fig. 9. K Means Clustering

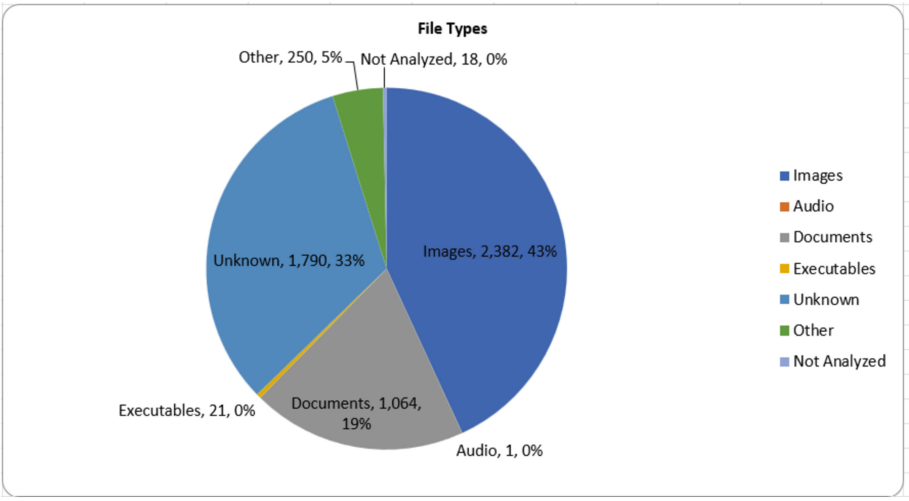

Fig. 10. Files types in Hacking case dataset

5.2 Scenarios

Three scenarios are mentioned below which gives idea about how various types of artefacts are collected and analyzed by the investigators.

Organizational Data Theft Investigation: The suspect used an USB for copying data from victims machine. All types of documents like Presentations, text files, log files Images and videos are investigated. Case investigation involved the sizing of a suspect's PC. Investigators utilize the known hash database to filter out files that are identified to be

illegal; a data filtration tool then obtains a collection of customer files that are most frequently used to locate relevant files. These include email, photo, and chat log files. Some of these image files have been identified by a known hash database as being unlawful. For training, the investigator inserts these files into an SVM model. Ultimately, additional unidentified data were included to the model and arranged according to relevancy score for additional examination.

Hacking Case Investigation: During an investigation into a hacking case, a computer was taken. The culprit has an email address. Several files are found when "username" and "password" are searched for. These text files discuss using scripts for wire-less network hacking and password cracking. Forensic agents provide data into a model to find remaining files that are comparable.

Banking Scam Investigation: To trick victims into providing their mail address, password, and other private information, the suspicious sets up a phishing website. The suspect commits online fraud using their accounts. When investigating a financial fraud case, investigators look for possibly false assets, payments, or other financial data. Certain pertinent files can be found by searching for the term "in-voice(s)" in PDF and Doc files from the raw disk image. Investigators then construct a model to identify comparable files using the analysis results.

This section presents three things: 1) the experimental disc timeline obtained, which clarifies the contents of a full disc image timeline and the source of the file timelines. 2) how the entire disc timeline was used to create these file timelines, including details about user action; and 3) the conclusion of the case study procedure.

5.3 File Artefact Timeline Analysis

This section provides a scenario of making a file artefact timelines. It is the result of file digital incidents gathered from the whole disc image timeline and indicates where the file characteristics were pulled from. Rather than the file on which the event occurs, the field filename indicates the source file of the digital incidents. For instance, the file system metadata on a Windows computer with an NTFS file system comes from the $MFT. The filename of the generated file is located in the description (field desc).

To extract related digital events, use the file names. Use the file name as a keyword to search each column of the timeline's file artefact attribute in order to complete this step. The created file timeline, which is made up of numerous sources of digital events, can be used to locate and confirm action traces of file artefacts.

5.4 Case Investigation and Relevancy Prioritization

The experimentation and research procedure that was carried out on each of the simulated case scenarios is presented in this section. Various elements are employed for every case, taking into account the distinct areas of research. The features that are added to the model depend on the relevant files that have been found and the particular similarities or qualities that are searched for. The following features were taken from each example and used to develop the model:

In the hacking case scenario, some associated text files and Python programmes for hacking user passwords were found. The ZIP file containing the Python project was unzipped. These specifics indicate that the features employed are "hack," "python," "py," "txt," "zip," and "unzip.

The examination is concentrated on pictures, films, etc. for the data theft case scenario. The illicit files that were discovered have digital events linked to them from browsing. File timings also revealed multiple file copying and shifting operations for numerous of the files. The following features are utilized to train the algorithm to find more files with similar usage patterns: "chrome," "child," "png," "jpg," and "MFT."

Upon investigating the financial fraud scenario, phony invoices (PDF files) were attached to emails. The files had been accessed by the user shortly before the machine was confiscated. The following features are used in the model construction for further exploration: "pdf," "invoice," "email," "fraud," "last access time," and "creation time."

6 Conclusion

This paper describes a methodology that gives priority to file artefacts that resemble previously examined relevant files. Machine learning models and feature extraction tools are developed to support the automated procedure. The findings demonstrate the approach's benefits and point to the possibility of accelerated research. Because of this, this strategy would be most effective early in the examination to steer the research in positive directions. This research presents an automated analysis approach that takes into account several information sources. This method's usefulness and accuracy could be increased by adding more features from additional sources.

References

1. Garfinkel, S.L.: Digital forensics research the next 10 year. Digit. Investig. **7**, S64–S73 (2010)
2. Scanlon, M.: Battling the digital forensic backlog through data deduplication. In: Proceedings of the Sixth International Conference on Innovative Computing Technology (INTECH), pp. 10–14 . IEEE, Dublin (2016)
3. Lillis, D., Becker, B., O'Sullivan, T., Scanlon, M.: Current challenges and future research areas for digital forensic investigation: arXiv preprint arXiv:1604.03850 (2016)
4. Mohammed, H., Clarke, N., Li, F.: An automated approach for digital forensic analysis of heterogeneous big data. J. Digit. Forensics Secur. Law **11**(2) (2016)
5. Quick, D., Choo, K.-K.R.: Big forensic data reduction: digital forensic images and electronic evidence. Clust. Comput. **19**(2), 723–740 (2016)
6. Sanchez, L., Grajeda, C., Baggili, I., Hall, C.: A practitioner survey exploring the value of forensic tools, AI, filtering, & safer presentation for investigating child sexual abuse material (CSAM). Digit. Investig. **29**, S124–S142 (2019)
7. Beebe, N.L., Clark, J.G.: Digital forensic text string searching: improving information retrieval effectiveness by thematically clustering search results. Digit. Investig. **4**, 49–54 (2007)
8. da Cruz Nassif, L.F., Hruschka, E.R.: Document clustering for forensic analysis: an approach for improving computer inspection. IEEE Trans. Inf. Forensics Secur. **8**(1), 46–54 (2012)
9. Le, Q., Boydell, O., Mac Namee, B., Scanlon, M.: Deep learning at the shallow end: malware classification for non-domain experts. Digit. Investig. **26**, S118–S126 (2018)

10. Hargreaves, C., Patterson, J.: An automated timeline reconstruction approach for digital forensic investigations. Digit. Investig. **9**, S69–S79 (2012)
11. Casey, E.: Digital Evidence and Computer Crime Forensic Science, Computers, and the Internet. Academic Press (2011)
12. Rowe, N.C., Garfinkel, S.L.: Finding anomalous and suspicious files from directory metadata on a large corpus. In: Gladyshev, P., Rogers, M.K. (eds.) Digital Forensics and Cyber Crime. ICDF2C 2011. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 88, pp. 115–130. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35515-8_10
13. Loukides, M.: What Is Data Science?. O'Reilly Media, Inc. (2011)
14. Guarino, A.: Digital forensics as a big data challenge. In: Reimer, H., Pohlmann, N., Schneider, W. (eds.) ISSE 2013 Securing Electronic Business Processes, pp. 197–203. Springer Vieweg, Wiesbaden (2013). https://doi.org/10.1007/978-3-658-03371-2_17
15. Du, X., Le-Khac, N.-A., Scanlon, M.: Evaluation of digital forensic process models with respect to digital forensics as a service. In: Proceedings of the 16th European Conference on Cyber Warfare and Security, ACPI, vol. 6, pp. 573–581. ECCWS, Dublin, Ireland (2017)
16. Van Baar, R., Van Beek, H., Van Eijk, E.: Digital forensics as a service: a game changer. Digit. Investig. **11**, S54–S62 (2014)
17. Van Beek, H., van Eijk, E., van Baar, R., Ugen, M., Bodde, J., Siemelink, A.J.: Digital forensics as a service: game on. Digit. Investig. **15**, 20–38 (2015)
18. Roussev, V.: Hashing and data fingerprinting in digital forensics. IEEE Secur. Priv. **7**(2), 49–55 (2009)
19. Lillis, D., Breitinger, F., Scanlon, M.: Expediting MRSH-v2 approximate matching with hierarchical bloom filter trees. In: Matoušek, P., Schmiedecker, M. (eds.) Digital Forensics and Cyber Crime. ICDF2C 2017. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 216, pp. 144–157. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73697-6_11
20. Du, X., Ledwith, P., Scanlon, M.: Deduplicated disk image evidence acquisition and forensically-sound reconstruction. In: Proceedings of the 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering, pp. 1674–1679. IEEE, New York (2018)