Abstract—Over the past few years, the number of crimes related to the worldwide diffusion of digital devices with large storage and broadband network connections has increased dramatically. In order to better address the problem, law enforcement specialists have developed new ideas and methods for retrieving evidence more effectively. In accordance with this trend, our research aims to add new pieces of information to the automated analysis of evidence according to Machine Learning-based “post mortem” triage. The scope consists of some copyright infringement court cases coming from the Italian Cybercrime Police Unit database. We draw our inspiration from this “low level” crime which is normally sat at the bottom of the forensic analyst’s queue, behind higher priority cases and dealt with the lowest priority. The present work aims to bring order back in the analyst’s queue by providing a method to rank each queued item, e.g. a seized device, before being analyzed in detail. The paper draws the guidelines for drive-under-triage classification (e.g. hard disk drive, thumb drive, solid state drive etc.), according to a list of crime-dependent features such as installed software, file statistics and browser history. The model, inspired by the theory of Data Mining and Machine Learning, is able to classify each exhibit by predicting the problem dependent variable (i.e. the class) according to the aforementioned crime-dependent features. In our research context the “class” variable identifies with the likelihood that a drive image may contain evidence concerning the crime and, thus, the associated item must receive an high (or low) ranking in the list.

Keywords: computer forensics, automated analysis of evidence, data mining, machine learning, “post mortem” triage.

I. INTRODUCTION

Computer Forensics is the application of digital investigative techniques aiming to analyze seized computers, maintaining a documented chain of evidence, for presentation in courts. Over the past few years, the spread of computer crimes caused by the large availability of low-cost, sophisticated and heterogeneous devices with large storage, resulted in an increased complexity and a proportional expansion of digital forensics examinations. A serious problem is that, despite the effort to analyze hundreds of terabytes of data, the outcome usually is not proportional to the amount of crime-related evidence retrievable from seized drives. The reason is twofold: on one side, the growing amount of information to process and, on the other, the forensic analyst’s habit to look for cues and evidence by means of traditional, manually intensive and time-consuming procedures. As a consequence, new works in the field, addressed to narrow the search, could reverse this negative trend.

Drawing inspiration from Marturana F. et al. [1,2] in the field of Mobile Forensics, this paper deals with an implementation of the “post mortem” triage model described by Berté R. et al. [3] and concerning the crime of copyright infringement. In particular, the proposed classification method aims to rank each seized device according to a predefined set of crime-related features extracted from the corresponding forensic image. As a result, this activity aims to identify as soon as the most relevant computers to focus on.

Berté R. et al. [3] describe “post mortem” triage as workflow consisting of the following four-phases: forensic acquisition, feature extraction and normalization, context and priority definition, data classification and triaging. The first one is the classical forensic hard disk image creation. The second is devoted to extract a set of features (e.g. configuration files, registry settings, installed software, file statistics and browser history) from the disk image and normalize them creating a two-dimensional matrix, called complete matrix, representing the data-set of a generic Machine Learning scheme. The third is in charge of (a) isolating the crime-specific features, (b) introducing in the model the timeline of interest and (c) creating the so called reduced matrix. The fourth, finally, is assigned the task of analyzing the reduced
matrix’s features and calculating the *class* variable by means of the aforementioned machine learning scheme. Assuming that the criminal conduct concerning copyright infringement is a time-independent variable i.e. it is unnecessary to know when the offense occurred to be able to prosecute it, we propose a three-phases workflow excluding from our implementation the timeline of interest and, thus, skipping the aforementioned third phase.

II. RELATED WORK

Over the past few years, Computer Forensics has been supported by several theories and tools to retrieve evidence on seized computers more effectively. A new research field, called triage, which allow to rank groups of seized devices and quickly identify the most relevant ones from a crime perspective has recently emerged. Triage has two main applications, called “live” and “post mortem”, which differ in the way they are implemented. Dealing with specific crimes such as murder, child abductions, missing persons, death threats, for example, the need for a timely identification, analysis and interpretation of digital evidence is crucial since it could be the difference between life and death for the victim. In those cases the inspection of each powered-on computer and digital device found on-scene with a “live” triage tool provides, indeed, investigators with cues to proceed with the search. On the other hand, when the aim is to search for evidence in lots of seized devices or to get through a consistent backlog of data to analyze, the “post mortem” triage tools could provide a viable way to prioritize and rank each device in order to simplify the subsequent forensic analysis.

Rogers M. K. et al. [4] proposed a new “live” forensics methodology called Cyber Forensic Field Triage Process Model (CFFTPM), which deals with “live” on-site activities for providing the identification, analysis and interpretation of digital evidence in a short time frame, without the requirement of taking the system back to the lab for an in-depth examination or acquiring a complete forensic image. The proposed methodology, although entailing a real risk of exhibit pollution, is justified by the need to provide investigative leads quickly in time critical situations.

Recently a new trend which combines computer forensic principles, Data Mining and Machine Learning statistical approach is taking hold in the research community (Veena H. B. et al. [5]).

As far as Mobile Forensics is concerned, moreover, Marturana F. et al. [1,2] have recently proposed two possible applications of Data Mining based classifications to “post mortem” triage. The first [1] deals with a procedure to identify the most relevant seized mobiles from an investigative point of view by predicting the device owner’s usage profile. Adopting the proposed methodology, investigators are able, indeed, to split up relevant and less important aspects of the case under investigation by ranking each involved device, person and crime. After a quick memory search of the whole set of seized phones, it is possible to create a ranking of items, ordered by probative value, and to identify the ones requiring additional lab processing. The latter [2], also related to Mobile Forensics and based on a research conducted with some Italian law enforcement cybercrime specialists, deals with self-knowledge algorithms for classification of mobile phone allegedly used to commit child pornography.

Further, Decherchi S. et al. [6,7] conducted an important study on the application of Data Mining theory and clustering text mining techniques for text analysis purposes during digital investigations. The work addressed forensic text clustering based on an adaptive model which arranges unstructured documents into content-based homogeneous groups.

We finally mention a valuable research conducted by Garfinkel S. et al. [8] in the field of digital forensic experiments reproducibility. The authors make an important effort to explain the need to test new forensic tools and techniques against standardized and well-known data corpora. They also contributed to create 4 large-scale standardized forensic corpora, available for research and educational purposes and downloadable at [9,10].

III. COPYRIGHT INFRINGEMENT: CRIMINAL CONDUCT AND RELATED FEATURES

This paragraph defines the crime of copyright infringement by identifying features and associated traces retrievable from a seized digital device. Copyright infringement is the criminal offense associated with the unauthorized or prohibited use of works under copyright, infringing the copyright holder's exclusive rights. It occurs when a copyrighted work is reproduced, distributed, performed, publicly displayed, or made into a derivative work without the permission of the copyright owner.

This criminal offence can be equated with both piracy and theft since the first is the practice of predating statutory copyright law, intentionally committed for financial gain while the latter is considered a misuse of the exclusive rights of the copyright holder for personal gain and without authorization.

The following is a list of platform independent crime-related features retrievable from a storage drive owned by an alleged copyright infringer:

- **audio/video files.** Files are classified by extension (.wav, .mp3, .mp4, .avi, .dvx, .mpeg, .mpg etc.) and recognized by means of the relative file header. A different weight is assigned to compressed audio and video files (e.g. MP3 and MPEG) since these are likely illegal copies.
- **ISO files or compressed archives.** An illegal reproduction of a copyright protected software, indeed, is usually stored within an ISO image or a compressed file (.zip, .rar, .tar, .gz) with an associated key generator to unlock the installation process.
- **Hacking tools** (key generators, password crackers etc.).
- **Peer-to-peer clients** (e.g. Emule, Kazaa, uTorrent etc.). P2P networks indeed are the most common way to share copyrighted material.
- **Specific multimedia players** such as Winamp or VLC commonly used by infringers to play copyrighted music or movies. Copyright infringers, indeed, are used to dislike multimedia players like Windows Media Player which doesn’t include divx audio/video codec.
– Web URL history. The URL history may uncover illegal traces since some P2P clients, such as uTorrent, to work properly, need first to collect the illegal file URL from a search engine such as www.isohunt.com for example.

The aforementioned features fall into the following three groups: installed software, file statistics and browser history and represent our model’s independent variables.

IV. METHODOLOGY ALGORITHMIC FOUNDATIONS

Our model’s dependent variable, i.e. the class, is a binary random variable with the following possible values: Copyright Infringer or Non Infringer. When a relationship exists between the device-under-triage and a copyright infringement court case the class is given the value Copyright Infringer otherwise its value is Non Infringer. The research goal is, thus, to train a machine learning scheme (i.e. the classifier) to classify each device-under-triage, according to the aforementioned crime’s dependent features. The preliminary step to train the learning scheme is thus to create a training-set, i.e. a collection of crime-dependent representative samples each with known classification (either Copyright Infringer or Non Infringer). Once trained, the classifier is assigned the task of elaborating unclassified real patterns (i.e. the test-set) and predict their classes. It is important to mention that the training phase is crucial for the whole classification process. It is possible to evaluate classifier’s learning effectiveness with the iterative and predictive method called 10 folds cross-validation, described by Witten I. H. et al. [11]. Such procedure splits the training-set into ten approximately equal partitions, each in turn used for testing and the remainder for training and it is repeated ten times so that every sample has been used exactly once for testing. Classifier’s learning effectiveness is evaluated according to the following performance indicators: Precision, Recall and F-measure [11], defined respectively as:

- Precision = TP / (TP + FP),
- Recall = TP / (TP + FN),
- F-measure = 2*Recall*Precision / (Recall+Precision).

Where TP= True Positive, FP=False Positive and FN=False Negative.

V. IMPLEMENTED PROCESS MODEL

This paragraph describes the implementation steps of the copyright infringement “post mortem” triage model. A scheme of the followed three-phases workflow is shown in Fig.1:

![Fig.1 – “post mortem” triaging model](image)

The process starts by creating a forensically sound image of the seized drive (i.e. forensic acquisition). The next stage, called feature extraction and normalization is in charge of extracting the crime’s dependent features e.g. the video files average size or the number of installed hacking tools etc., from each disk image and creating the data-set, resulting from the combination of training-set (i.e. a set of samples with known classification) and test-set (i.e. a set of samples to classify).

The following is a list of aggregated features of the model:

- number and category of installed programs (chat, P2P, crypto, browsers, entertainment, game, disk mount utilities)
- number and category of visited URLs (hacking, warez, illegal download)
- number and type of picture files (produced or downloaded), max and average file size
- number of video/music files, max and average file size
- number of ISO files, max and average file size
- number of doc/pdf files, max and average file size
- number of compressed files, max and average file size
- number of crypto files and max file size.

In this case the data-set set consists of 9 hard drives, both related and unrelated to copyright infringement investigations, and 4 hard drives of Garfinkel S. M57-Patents corpus [10]. It is represented by the matrix in Fig.2, where feature name is indicated in the leftmost column while a set of sample drive images are listed in the others:

![Fig.2 – Dataset structure](image)

Excluding the header row and line, the matrix is a rectangular N x M (where N=45 is the number of features and M=13 is the number of available samples). The data classification and triaging stage provides then a classification of the data-set by processing it with one or more Machine Learning schemes. We made a benchmark comparison of the four most popular ones (i.e. Bayesian Networks, Decision Trees, Locally Weighted Learning and Support Vector Machines), assessing their compared performance. The following is a brief definition of the aforementioned classifiers:

**Bayesian Networks (BN)** estimate the conditional probability distribution of the values of the class attribute (i.e. the usage class), given the values of the other attributes. BN are drawn as a network of nodes, one for each feature, connected by directed edges in a directed acyclic graph.

**Decision Trees (DT)** apply a “divide-and-conquer” approach to the problem of learning from a set of independent samples where the binary tree representation consists of nodes implying a test on one or more attributes and leafs giving a classification to all the samples that reach it. For example, to be classified, an unknown sample is routed down the tree.
according to the values of the attributes tested in successive nodes and, when a leaf is reached, the sample is assigned the same class.

**Locally Weighted Learning (LWL)** is a general algorithm associated with any learning technique that can handle weighted samples. It assigns weights using an sample-based method and builds a classifier from the weighted samples, assuming independence within a neighborhood, not globally in the whole sample space.

**Support Vector Machines (SVM)** are algorithms used for classification, regression, or other tasks which constructs a hyper-plane or set of hyper-planes in a high-dimensional space. A SVM model is a representation of the sample data as points in space, mapped so that samples of separate categories are divided by a clear gap. New samples are mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

To perform the benchmark experiment we used WEKA [12], a powerful open-source Java-based machine learning workbench. WEKA brings together many machine learning algorithms and tools under a common framework with an intuitive graphical user interface. WEKA Experimenter allows large scale experiments to be run with results stored for further retrieval and analysis.

**A. Complete data-set analysis**

In this scenario we adopted the iterative and predictive method called 10 folds cross-validation, described in section IV, to assess the compared performance of Bayesian Networks, Decision Trees, Locally Weighted Learning and Support Vector Machines. The analysis takes into account the complete data-set (i.e. the N x M matrix, where N=45 is the number of features and M=13 is the number of available samples).

We considered the following performance indicators: percentage of correctly classified samples, mean absolute error, root mean square error, weighted average precision, weighted average recall, weighted average F_measure and the experimental results are summarized in Table 1:

<table>
<thead>
<tr>
<th>Performance parameter</th>
<th>Machine Learning schemes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BN</td>
</tr>
<tr>
<td>Percentage correct (%)</td>
<td>99</td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>0.03</td>
</tr>
<tr>
<td>Root mean square error</td>
<td>0.04</td>
</tr>
<tr>
<td>Weighted avg Precision</td>
<td>0.99</td>
</tr>
<tr>
<td>Weighted avg Recall</td>
<td>0.99</td>
</tr>
<tr>
<td>Weighted avg F_Measure</td>
<td>0.99</td>
</tr>
</tbody>
</table>

As we can see, in this scenario Bayesian Networks is the best performing classifier with a 99% of correctly classified samples and weighted average precision of 0.99 while Support Vector Machines ranked second with 93.5% of Percentage correct and weighted average precision of 0.93.

**B. Classifiers comparative analysis**

We chose to follow an incremental approach with 3 different data-set calculated on the same samples of the previous scenario (i.e. 13) but with a different and a gradually increasing number of features, respectively 15, 30 and 45. The goal was to assess classifiers compared performance with regards to the increase of available features in the model. Experimental results are summarized in the following tables:

<table>
<thead>
<tr>
<th>Table 2: Percentage correct comparative results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
</tr>
<tr>
<td>--------------------------------</td>
</tr>
<tr>
<td>Copyright_Infringement_15</td>
</tr>
<tr>
<td>Copyright_Infringement_30</td>
</tr>
<tr>
<td>Copyright_Infringement_45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3: Weighted_avg_Precision comparative results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
</tr>
<tr>
<td>-----------------------------------</td>
</tr>
<tr>
<td>Copyright_Infringement_15</td>
</tr>
<tr>
<td>Copyright_Infringement_30</td>
</tr>
<tr>
<td>Copyright_Infringement_45</td>
</tr>
</tbody>
</table>

Comparing results, it is possible to note that half of the classifiers (i.e. Bayes Networks and Support Vector Machines) have an increasing precision with regards to the number of available features while the others (i.e. Decision Trees and Locally Weighted Learning) behave contrary to expectations. Hughes G.F. [13] described this behavior, concerning the mean accuracy of statistical pattern recognizers, showing that, an abnormal features increase may turn into a significant performance degradation with an increasing error rate (i.e. the number of misclassified samples out of the total). In particular, with a fixed number of training samples (M=13), the overall accuracy (OA) initially grows up to a max OA corresponding to N* features and then decreases with the further increase of N. This is due to the limited number of training samples which is inadequate to provide reliable figures about the classifier.
parameters. Hall M.A. and Holmes G. [14] and Wang B. et al. [15] proposed and compared different feature reduction techniques to reduce the feature-space dimension and showed how it is possible to generally improve classifiers performance.

VI. Conclusion

The paper deals with Computer Forensics “post mortem” Triage related to court cases of copyright infringement. The research questions that we addressed were the following: "which are the potential benefits of applying Machine Learning schemes to automated analysis of evidence? Is it possible to implement a “post mortem” triage tool which could (a) reduce backlogs, (b) bring order back in the lab queue and (c) increase investigation efficiency?”. In this context we decided to adapt the “post mortem” machine learning based triage model proposed by Berté R. et al. [3] to the crime of copyright infringement. The proposed crime-dependent categorization model consists of the following three phases: forensic acquisition, feature extraction and normalization and data classification and triaging. The methodology aims to predict associations among devices-under-triage and copyright infringement investigations. Once applied to a number of seized devices, thus, the method is able to identify the relevant ones requiring further lab analysis. We associate the aforementioned likelihood with the model’s dependent variable i.e. the class. This variable is calculated upon a set of features which belong to one of the following categories: installed software, file statistics and browser history and represent our model’s independent variables. Classification of the dataset is performed with four Machine Learning schemes (i.e. Bayesian Networks, Decision Trees, Locally Weighted Learning and Support Vector Machines) whose performance were compared with a benchmark analysis performed by WEKA Experimenter, an open-source Java-based machine learning workbench.

VII. Future Work

The model described in this paper can be virtually extended to (a) the management of INFORMATION SECURITY (INFOSEC), simplifying the identification of possible threats towards corporate copyrighted material and the search for countermeasures [16,17] and (b) a number of possible implementations concerning crimes such as child pornography, hacking, murder, terrorism. In this regard, interested readers who want to try their own implementations should spend some time to identify the crime-related features (i.e. the independent variables) and collect a consistent set of classified samples related to one of the aforementioned crimes. It is important to mention that classification success depends on the accuracy adopted when creating the training-set which is the core of the whole process. The higher the number of analyzed hard disk images (i.e. the training samples) is, indeed, the better the model classifies new unclassified samples. It is also important to make a benchmark comparison among the available classifiers to find the ones that best fit to the problem.

REFERENCES