DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4571
Volume 8, No. 8, September-October 2017
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

ARTIFICIAL INTELLIGENCE BASED DIGITAL FORENSICS FRAMEWORK

Parag H. Rughani
Ph. D.,IFS, Gujarat Forensic Sciences University
Gandhinagar, Gujarat - India

Abstract: With increase in number of Internet and smartphone users, cyber crimes are equally increased. Current resources including man power are not sufficient to investigate and solve cyber crimes with the pace they are committed. Present tools and technology require human interaction at large scale, which slows down the process. There is acute need to optimize speed and performance of Digital Forensic Tools to keep pace with the reported cyber crimes. An Artificial Intelligence Based Digital Forensics Framework is proposed in this paper to overcome above issues. The framework proposed in this paper require minimum user interaction and does majority of routine operations by intelligence acquired from training. Outcome of the work is mentioned in the form of proposed framework to optimize digital forensics process.

Keywords: Digital Forensics, Artificial Intelligence, Machine Learning, AI Framework, Cyber Crime Investigation, Digital Forensic Tools, Cyber Crime

1. INTRODUCTION

Evolution of smart phone and Internet has changed the whole world today. It is almost impossible to imagine world without smart phone and Internet. As per the reports, by 2020, 80% of adults on earth will have a smartphone [1], on other hand till date 49.7% of total population is connected to Internet with the growth of 936% from 2000-2017 [2]. These figures are encouraging for the manufacturers and developers as the exponential growth is an indication of more hardware and software requirements from these users. However, the things are not good from the low enforcement agencies' point of view. The figures shown below are the biggest worry for cyber security professionals.

It is predicated that the projected costs of cyber crimes will reach $2 trillion by 2019 [3]. It is also observed that more than half of firms – around 57% – have experienced a cyber attack in the past year [4]. Not only organizations or giant companies are targeted, individuals are also vulnerable to cyber attacks and it is estimated that there will be around 18 victims / second [5]. The biggest challenge for cyber security experts is to protect the data and as per the report from [6] from 2005 till August 02, 2017, reported records breached are around 904,715,669. The crimes are bound to happen and cannot be stopped completely. Cyber security experts can do their best to reduce number of crimes but with increase in number of new devices and technologies the crimes in this virtual field are impossible to control. It has been observed and represented in many surveys that most of the victims do not report officially for legal procedure because of lack of awareness. The worst part is even if the crimes are reported it is very difficult to solve and investigate them. Following chart gives an insight about the conviction rate of cyber crimes in India between 2013 to 2015.

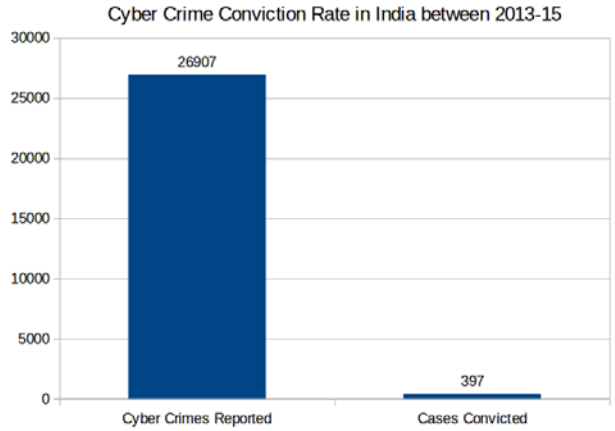

Figure 1.1: Cyber Crimes in India between 2013-15 [7]

Almost similar figures are found in all over the world as reported by [8]. These figures are not socking as they reflect reality and speed by which Cyber Crimes are investigated. Lack of highly automated tools and equally less number of forensic investigators could be the only reasons behind such low conviction rates. Preparing forensic experts for computer crimes is a time taking process and may take months or even years for a country to produce such skilled persons to deal cases with the speed by which they are registered. On other hand the tools available today provide automation of procedure upto certain extent but they require human interaction and intelligence for accurate results. In fact is is very difficult to imagine an expert system which can forensically investigate cyber crimes, however it is not impossible. This paper focuses on possibility of such intelligent framework which can work to reduce burden on digital forensic investigators and hence increase the speed of the investigation.

2. DIGITAL FORENSICS PROCESS AND AVAILABLE TOOLS

Forensics as defined by Oxford is "Scientific tests or techniques used in connection with the detection of crime". If these scientific tests or techniques – mainly computer based – are used in connection with the detection of digital crimes then can be referred as Digital Forensics. In general terms, the methods and tools used in cyber crimes are referred as Digital Forensics. The digital forensics or the forensics consists of 3 steps namely: Acquisition, Analysis and Presentation[9].

Variety of commercial and free tools are available in market to carry out digital forensics. EnCase [10], FTK [11], UFED [12], Nuix [13], IEF [14], Autopsy [15], XRY [16], etc. are some of the famous and widely accepted digital forensics tools. Reports generated from most of the tools mentioned above are accepted in many of the countries all over the world.

Following figure lists few known tools used at various steps of digital forensics.

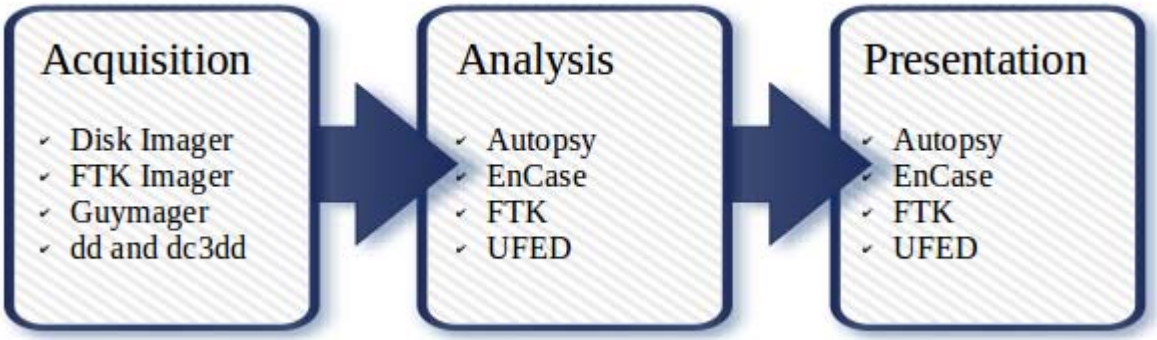

Figure 2.1: Few known tools used in digital forensics process

In last few years these and many other tools have evolved like anything. They provide user friendly GUI and high speed analysis. Custom Filters, Advanced reporting and carving are key features of most of these tools. They are also modified to automate certain processes. However, the automation part is not fully implemented as most of the time user inputs are needed to automate a task.

Looking at current scenario, one can say that the tools available in the market require human interaction and they cannot be operated without skilled resources. Increasing number of skilled resources is one option to speed up investigation process, while other option is to make the tools smart enough to work intelligently for analyzing suspicious data. Many researchers and tool writers are working hard to make their forensic tools automatic and faster. The efforts put by some of the researchers in improving digital forensics tools are discussed in next section.

3. RELATED WORK

Lots of work has been done in improving digital forensic tools and frameworks. May researchers contributed to this process by proposing various frameworks and enhancements. Some of them include digital forensic framework for cloud computing[17], Hierarchical, Objectives-Based Framework [18] and use of GPUs to increase the performance of digital forensics tools [19]. Apart from suggestions and proposals some researchers worked to compare existing model [20] and predict future of these tools [21].

Many authors work on adding intelligence part to existing digital forensic tools. Someone used artificial intelligence for offline intrusion analysis in network forensics [22], while someone developed a multi-agent and case-based reasoning using artificial intelligence [23] and some authors demonstrated the ability to automatically correlate events and objects among a memory image, filesys- tem image, and network capture [24].

However, significant work has been done to improve performance and capability of digital forensic tools for speeding up the investigation process, present tools are still user dependent. There is acute need of a concept for creating intelligent digital forensic tools, which can not only reduce burden of the investigator but can also think by itself. It is also important that the future tools individually or as a component of a kit can address all three steps intelligently. Proposed framework in the next section is aiming to achieve similar tool which can investigate a case with help of initial inputs.

4. PROPOSED FRAMEWORK

The framework proposed in this paper is a conceptual framework which can help manufacturers and developers in designing highly intelligent and automated tools to assist digital forensic investigation process as a whole. The framework is based on Machine Learning concept and uses AI in each step of the process to make sure decisions taken by tools have minimum false positives. However, it is not possible to conceptualize a 100% accurate and intelligent framework, the possibilities cannot be overlooked for mistakes and false positives. The working model may

require rigorous testing before use. Though the user inputs can be minimized at the possible level, a verification at each process is preferable to avoid mistakes, especially the report generated by the system should be verified and cross checked with artifacts before it is presented in the court of law.

An overview of present non-AI framework is shown in following figure.

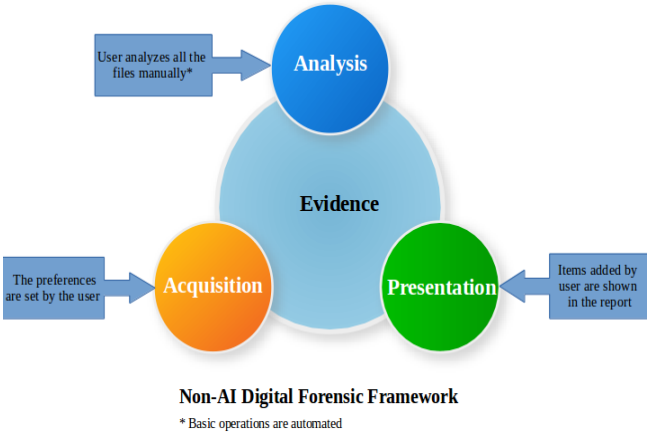

Figure 3.1: Non-AI Digital Forensic Framework

As it can be seen from the above figure, the tools used at various steps are dependent on user inputs. Lack of intelligence in these tools lead to repetitive and tedious tasks to be performed by the investigators even if the evidence and case are identical. Another drawback of this system is user dependency in analysis step. Present tools are very powerful and designed to provide automation like finding broken files, identifying invalid file signatures, retrieving files from unallocated space, retrieving deleted files, categorizing files and artifacts as per common requirements and even grouping items like multimedia files, log history, Internet artifacts, mails, etc. However, the user has to check each and every corner of the evidence to find related files and add them to report. There is an obvious reason that each case is different and criminals use different modus operandi. It is not possible to eliminate role of user in real life, but if the investigator gets 10 files to look instead of 1000 then it is always useful and faster.

An overview of the proposed model which provides intelligence by reducing redundant work of an investigator is given in the following figure.

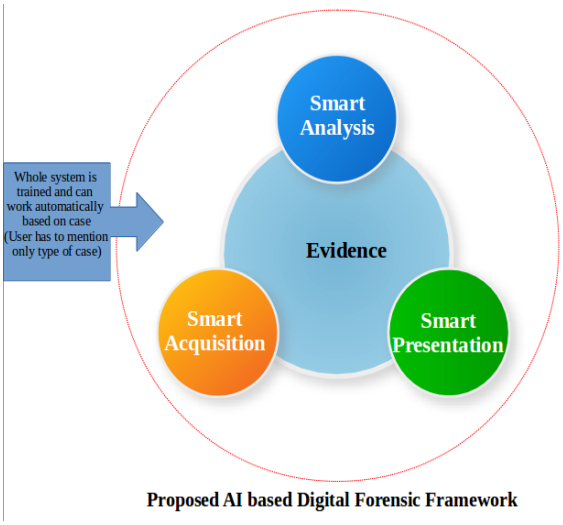

Figure 3.2: Proposed AI based Digital Forensic Framework

Proposed framework is based on cases and considered to be a single kit which can deal with all three steps of digital forensic process. Most of the existing tools provide support for all three steps but they lack intelligence mentioned in the proposed framework. The framework is designed with commonly known artificial intelligence / machine learning process, where the system is trained with existing data sets. These data sets are helpful in making system understand what decision to take in which situation. Following subsections discuss possibility of incorporation of AI at each step. The steps are called smart in this framework as they work based on their experience. After training, each step needs to be trained to see the results. The test data sets can be used in verifying the learning process and if required tool can be trained rigorously with more number of training data sets.

3.1 Smart Acquisition

Acquisition is one of the most crucial step in any forensic process. The paper focuses only on imaging of storage media as other tasks and documentations are not in the scope of this work. Let us consider an example of a threatening mail. This type of case in general is filed by victims after such mails are received. This is one of the common crimes reported and many such instances have been observed over the years. If a non AI forensic tool is used to acquire image from storage device then it will image all the data and will let the investigator or analysis tool to extract files from the whole image. If we consider the capacity of a hard disk then presently almost all computers come with around 1 TB disk. Creating image from 1 TB hard disk is not only time consuming, it is useless as not whole 1 TB need to be analyzed. One should not forget that the image created has to be analyzed by a tool and which in turn may need to process whole disk image.

Here in this proposed model the acquisition tool will be a trained with similar cases. So, if we assume then it is an experienced tool which knows from its past experiences what data is of interest, in our cases we are looking for only emails, we are least concerned about the multimedia files or even software installed on the machine. Following diagram shows smart acquisition tool:

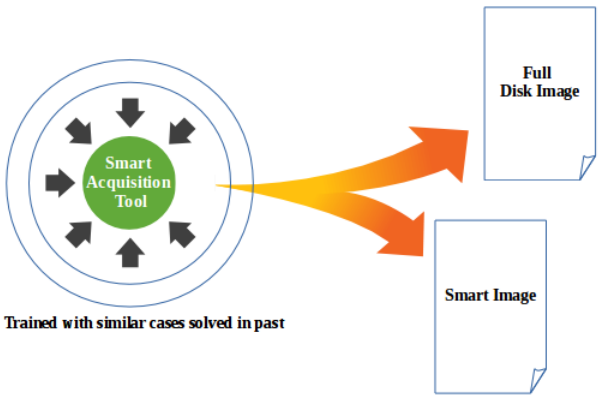

Figure 3.1.1: Smart Acquisition

As shown in above figure, the smart acquisition tool of proposed framework will be trained with similar cases solved in the past. During learning phase the tool will determine types of files and their locations which were used by the investigator in solving the case. The tool will remember all these and during real execution once informed about type of case, it will focus only on the files and locations of the interest. As integrity is one of the major concerns in the evidence handling, it can be argued if only relevant data is acquired. To overcome the integrity issue, the proposed framework suggests to create two copies of evidence: one as the full disk image and the other one consisting of only files of interest for a particular case. Of course, the time will be a concern here and this smart acquisition cannot help in speeding up the acquisition process, however, properly trained tool will help in reducing relatively more time during the next step of analysis.

3.2 Smart Analysis

Tools used in proposed smart analysis are also trained with similar solved cases as mentioned in previous sub section. The smart analysis is considered here as a part of complete AI based framework and also as a separate process. If image to analyze is created using smart acquisition tool then smart analysis reduces large amount of analysis time as it has to process only few files instead of whole disk. But, if the image is created using a non AI based acquisition tool, then also the smart analysis tool can use its experience in filtering files of interest but speed will be reduced in that situation.
While considering case when image is created from a smart acquisition tool, the smart analysis tool which is trained with past cases will use keywords remembered during training with the limited set of files to analyze. It is worth to note that, unbelievable speed can be reduced in identical cases.
When the image is taken from a non AI based acquisition tool, the smart analysis tool will use files, locations and keywords used in solving previous cases similar to the current one. This will be time consuming compared to previous scenario but still will take less time then manual analysis required by a non AI analysis tool.
In either case, the smart analysis tool will identify necessary artifacts and will mark them for final report, which in other case was done by the investigator manually. This also reduces sufficient amount of time.

3.3 Smart Presentation

In general, this step is done as a result of analysis step and in the form of report. The tools used at present, generate reports based on the items marked manually by the investigators. However, in case of smart presentation the report is generated based on the automated and intelligent process done by smart analysis tool. Items added by the smart analysis tools can be further filtered by smart presentation process. The smart reporting tools proposed in this framework are also trained with the reports of similar cases solved in the past.
The smart reporting tool works as a filter and can remove unnecessary items, if any added by smart analysis tool as a false positive. Properly trained smart reporting tool can make the final report highly accurate.
All the tools mentioned in this process can be customized and user can make changes or corrections at any point of time. So, if all the three smart techniques used together then lots of redundant and tedious work can be reduced. Which will make forensic investigators available for other cases including cases with new modus operandi. It will also allow forensic labs to work on parallel cases by keeping more number of AI based toolkits and with less man power to operate them.

4. LIMITATIONS

As mentioned earlier the framework proposed in this paper is a conceptual framework and has many limitations. Some of them are mentioned in this section, there may be other limitations which are difficult to visualize but can be discovered during implementation.
Major limitation of this framework is its dependency on training data sets. If there is a case which is reported first time then this framework may not be able to handle it, however as discussed in previous section all the three smart technologies can work with AI and non AI conditions. Another limitation is size of training data sets, a huge amount of data is required to train this system to make it more accurate and powerful. It may seem difficult from the forensic investigator's point of view to train the system with large amount of data covering majority of common cases, but it is equally easy for the manufacturers. It is already seen in the case of mobile forensic tools, e.g. Cellebrite UFED supports majority of android phones, and it can be assumed that the tool has been trained to understand basics of all supported phones.
Training is crucial step of this proposed framework and can assure reliability of the work done by such system. Another limitation could be uncommon cases, however this framework can be trained from regularly reported crimes like malware infection, document forgery, unauthorized access, etc. but it may not be trained for cases which are committed once in a while. Though such cases are very rare, they are important and can be solved with traditional way of non AI component of the proposed framework.
Final limitation observed at this stage is authenticity of the reports generated by the smart system. However, thorough training will reduce false positives, the investigator is required to verify and cross check the finding of the proposed framework before submitting it in the court of law.

5. CONCLUSION

Framework proposed in this paper can be highly effective if it is trained properly and rigorously. It can reduce time, processing power and even man power in investigating a digital crime. Use of this framework can let the forensic investigators handle more cases with more accuracy and better speed to increase conviction rate of cyber crimes.

6. REFERENCES

1. 2016: Current State of Cybercrime, RSA Whitepaper, 2016
2. World Internet Users and 2017 Population Stats, 10th July, 2017 [http://http://www.internetworldstats.com/stats.htm]
3. Morgan S., Cyber Crime Costs Projected To Reach $2 Trillion by 2019, 2016, Forbes.
4. The Hiscox Cyber Readiness Report - 2017, Hiscox
5. Cyber Crime Statistics and Trends [Infographic], 17 May 2013, Go-gulf [https://www.go-gulf.com/blog/cyber-crime/]
6. Identity Theft Resource Center, 2nd August, 2017 [http://www.idtheftcenter.org/Data-Breaches/data-breaches]
7. http://mha1.nic.in/par2013/par2016-pdfs/ls-190716/23%20E.pdf
8. Grimes R., Why Internet crime goes unpunished, CSO, 2012 [http://www.csoonline.com/article/2618598/cyber-crime/why-internet-crime-goes-unpunished.html]
9. Altheide C. and Carvey H., Digital Forensics with Open Source Tools, 2011
10. https://www.guidancesoftware.com/encase-forensic
11. http://accessdata.com/products-services/forensic-toolkit-ftk
12. https://www.cellebrite.com/en/home/
13. https://www.nuix.com/
14. https://www.magnetforensics.com/magnet-ief/
15. https://www.sleuthkit.org/autopsy/
16. https://www.msab.com/products/xry/
17. Martini, B., & Choo, K. K. R. (2012). An integrated conceptual digital forensic framework for cloud computing. Digital Investigation, 9(2), 71-80.
18. Beebe, N. L., & Clark, J. G. (2005). A hierarchical, objectives-based framework for the digital investigations process. Digital Investigation, 2(2), 147-167.
19. Marziale, L., Richard, G. G., & Roussev, V. (2007). Massive threading: Using GPUs to increase the performance of digital forensics tools. digital investigation, 4, 73-81.
20. Reith, M., Carr, C., & Gunsch, G. (2002). An examination of digital forensic models. International Journal of Digital Evidence, 1(3), 1-12.
21. Garfinkel, S. L. (2010). Digital forensics research: The next 10 years. digital investigation, 7, S64-S73.
22. Mukkamala, S., & Sung, A. H. (2003). Identifying significant features for network forensic analysis using artificial intelligent techniques. International Journal of digital evidence, 1(4), 1-17.
23. Hoelz, B. W., Ralha, C. G., & Geeverghese, R. (2009, March). Artificial intelligence applied to computer forensics. In Proceedings of the 2009 ACM symposium on Applied Computing (pp. 883-888). ACM.
24. Case, A., Cristina, A., Marziale, L., Richard, G. G., & Roussev, V. (2008). FACE: Automated digital evidence discovery and correlation. digital investigation, 5, S65-S75.