International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-5, June 2019

Digital Forensics Using Supervised Ml

Rajyashree.R, Nimisha Praveen, Dheeraj.J, Sudharshan.R

Abstract*: Digital forensics is a branch of forensic science where it focused on recovery and investigation of artefacts found on the digital devices. Its ability to totally reduce or even eliminate sample risk – This is the biggest advantage of forensic science over the external auditors. The forensic science helps in gathering of evidence of physical investigations, so care must be exercised in digital forensic collection to ensure that the data being collected for analysis must be as pure and undisturbed as possible. Digital forensics using supervised ml, uses network forensics and forensic data analysis. In network forensic when the data passed over the network like (LAN WAN MAN), data packets may be lost during the transmission, the lost data packet can be traced by the linear regression algorithm.In data forensics it helps to check the integrity of the data with the existing data by the support vector machine algorithm.These two algorithms of machine learning will help us to find the best trends and predict were the missing packets have gone and the data provided is valid or not.

Keywords: Digital Forensics, data collection, data forensics, evidence, integrity, linear regression, network forensics, support vector machine, transmission.

I. INTRODUCTION

The term for forensics science involves forensics (in Latin forensic), means a public discussion or a debate. In a more modern context, however, forensic applies to courts or the judicial system. Combine that with science, and forensic science means applying scientific methods and processes to solving crimes. Digital forensics is a branch of forensic science that includes the identification, recovery, investigation, validation, and presentation of facts regarding digital evidence found on computers or similar digital storage media devices. Digital forensics is a constantly evolving scientific field with many sub-disciplines. Some of these sub-disciplines are:

- Computer Forensics: The identification, preservation, collection, analysis and reporting on evidence which is found on e-hardware devices like computers, laptops and storage media for the support of investigations and legal proceedings.
- Network Forensics: The monitoring, capture, storing and analysis through network activities or events in order to discover the source of security breaches, intrusions or other problem incidents, i.e. worms, virus or malware attacks, abnormal network traffic activities and security breaches
- Mobile Devices Forensics: The recovery of electronic evidence from electronic devices like mobile phones, smartphones, SIM cards, PDAs, GPS devices, tablets and game consoles etc.
- Digital Image Forensics: The extraction and analysis of digitally validate photographic images where their

  authenticity by recovering the metadata of the image file to ascertain its history.
- Digital Video/Audio Forensics: The collection, analysis and evaluation of sound and video recordings from extracting it from videos and sound. The science is the establishment of authenticity as to whether a recording is original or valid and whether it has been tampered with, either maliciously or accidentally on purposely.
- Memory forensics: The recovery of evidence from the RAM of a running computer, from its memory also called live acquisition.

II. RELATED WORKS:

In the early 1980s, the use of personal computers began, and cybercrime was one of the consequences of the problem. Data forensics has been developed to solve this problem and to recover the stored information as a means of recovery. Information Forensic tends to have information about the crime. Today, researchers use forensic information for online fraud, personal system hacking, data theft, and crime with online violent crime. Computerized proof of manufacture is literally produced in the court as the same standard as a physical proof. As a result, information forensic proofs must be authentic, encrypted, accurate and reliable.

This paper is related to the project called ForNet. It is a distribution network Forensic Framework and integrated Wide Area Network (WAN) logging process. ForNet has been developed for Digital Forensic Management and Objectives. The structure consists of two components: SynApps and Forensic Server. SynApp integrates with network devices, such as switches and router. These SynApps can be organized into a peer-to-peer architecture. In the absence of central control, architecture they finally works with each other. All SynApps form a network within a domain in hierarchical architecture. They are associated with the Forensic Server of that domain. Actually, central administrative controls for Forensic Server domains that run a group of domain synopses. The Forensic Server receives searches from outside its domain and processes it using Synopses and returns verification to the sender and the results after the certificate. The SynApp network constitutes the first level domination of the ForNet Hierarchical Architecture.

Manuscript published on 30 June 2019.
***** Correspondence Author (s)
Ms. R. Rajyashree, Assistant Professor, Computer Science and Engineering, Srm Institute of Science and Technology, Tamil Nadu, India
Ms. Nimisha Praveen, Student, BTECH Computer Science and Engineering Srm Institute of Science and Technology Tamil Nadu, India.
Mr. Dheeraj. J, Student, BTECH, Computer Science and Engineering Srm Institute of Science and Technology Tamil Nadu, India
Mr. Sudharshan. R, Student, BTECH, Computer Science and Engineering Srm Institute of Science and Technology Tamil Nadu, India.

Forensic Servers can also be networked for inter-domain collaboration, which form the second stage of the sequence. Through the appropriate Forensic Server to cross the Domain Bound. The only gateway to the queries sent to the domain from outside a forensic server domain boundary In other words, a query that is sent to a domain goes to the forensic server of that domain, is approved by the server and submit it to the appropriate domain listings. Similarly, the SynApps results are sent to certified and return domain charges forensic servers. In fact, the question of page node starts at the upper level in the chain of Dyne forensic server and ends in the leaf nodes in the other branch. Queries usually travel in opposite directions to attack.

III. PROPOSED SYSTEM:

From the administrative point of view, the main challenge of information forensics is acceptance of appropriate standards and regulation of forensic practice data. Although there is a wide variety of acceptable standards for information forensics, there is a lack of criteria. In terms of information forensic rule, no algorithm is specified, and results are used in prediction. Furthermore, the lack of this prediction is due to the lack of algorithms and logarithms. There are technical, legal and administrative challenges facing forensic data. Technical factors affecting data forensic effects include encryption problems, device space consumption and anti-forensic methods. Anti-forensics refers to attempts to bypass data forensic tools, either through process or software. Registry challenges can also arise in data forensics and can confuse or mislead an investigation. An example of this would be description issues due to a malicious program like a Trojan. Trojans are malware that avoid themselves as a harmless file or application. Since trojans and other malware are capable of performing malicious activities without the user's knowledge, it can be difficult to determine if cybercrime was deliberately committed by a user or if they were executed by malware.Our project works with algorithm guidelines and follows strict prediction process. This task expands the calculation of each step of the prediction process. It is possible to maintain a way of progress and maintain lost data with the algorithm. Two algorithms are projected out throughout our project. The classification and regression algorithms are the main objectives given and used this project. Regression and Classification are categorized under the same umbrella of supervised machine learning. To make both predictions, share the same concept of using known data sets known as training data sets. The main difference between them is that the output of regression is variable numerical or continuous, when it is distinct or isolated for classification.

Support Vector Machine (SVM) algorithm is used for classification and Linear Regression (LR) is used for regression. SVM is a discriminating classifier that is formally defined by a dividing plan. As a classification method, SVM is a global classification model that generates non-overlapping partitions and usually uses all attributes. It can solve linear and non-linear problems and works well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyper_plane that separates the data into classes.LR Survey Learning is a machine learning algorithm. It carries out crime law. To predict regression models worth a target based on an independent variable. The regression technique x (input) and y (output) of a linear relationship is between. LR equation forms the equation y = a + bx where Y is the dependent variable, X the independent variable, B line slope and a Y-intercept.

IV. SYSTEM ARCHITECTURE:

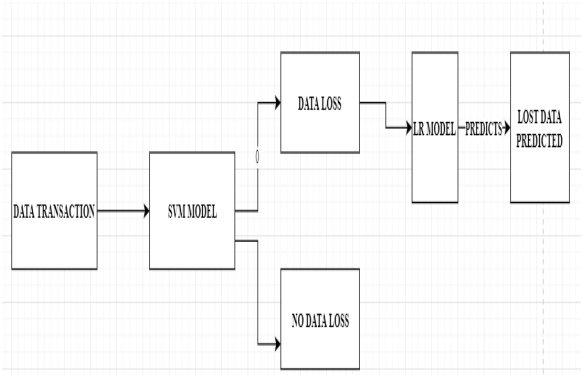

FIG 1. System architecture

The dataset would be used to train the "SVM Model" . Once the model is trained, a new data transaction would be given to this module and the output would give whether a data packet was lost or not and what was the lost data if there was any. In the proposed system the data to be transmitted is first split in test and training data using the train _test_split function(). Once the data is split into train and test, the training data is trained to determine if the data transmitted reaches the destination or not. Ten fold cross validation preferred to produce better accuracies. Sklearn consists of a function .svm(), which helps in classification of models. The training data is the fit using the Sklearn function .fit(). Sklearn library consists of the .fit() function which helps to fit the svm model with the training data along with its class labels. The output of the svm model determines if the data is lost or not. The data that is lost is labelled 1 and the data that is not lost is labelled 0.Thus the data packets which is transmitted through the network can be classified. The reason SVM is more preferred over other classification model is that is produces better accuracies.

- Determining the Data Lost:

The data lost has to be determined. To determine the data that is lost we use linear regression. We can use the data transaction without the data packet loss and remove some portions of them to make a virtual data transaction with packet loss and use the removed data as the predicted data packet lost. Here the data packets which are not lost are taken as training data. The labels column from the training data is removed and is trained using linear regression. The system thus learns all the information about the data which is not lost. This metric is then used by the system to determine the data which is lost. Linear Regression can be imported directly using the sklearn library. The accuracy of this model is low. But the use of linear regression provides a partial solution to the everlasting problem of finding attenuated data. The data packet may be in the form of a binary sequence, but the sequence varies completely from one data packet to another thus this is a regression model. The output of the model thus produces a sequence of bytes which can be translated into appropriate forms to retrieve the data which is lost.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

The regression model is an experimental model use to determine the data lost. Further fine tuning of the algorithm is necessary to produce absolute results.

V. BENEFICIARY OF THE SYSTEM:

1. The usage of the support vector machine algorithm which is a popular strategy for supervised ML where it helps for classifying the algorithms. SVM is more efficient than any other classification algorithm where its optimization rate is very low.
2. When the SVM classifies the data set it uses decision boundary where the classification is much easier rather than using a logical egression where its very difficult to classify.
3. During the transaction of the data packet is trained by by the machine learning process to find the lost data packet.
4. The support vector machine will check whether the data packet is lost or not and it will classify in different classes and then it is further checked by linear regression
5. The linear regression will find out where the lost packet have gone, it will identify where actually the data packet have gone and it will try to retrace it efficiently.

VI. FUTURE WORK:

The machine learning algorithm predominantly is used for predicting analysis, finding the nearest trend between the algorithms. Combining both machine learning and Internet of Things (IOT) could make a powerful tool to finding the best trends. The IoT is a network of interconnection of objects in day to day life called things that have been augmented with a small measure of computing capabilities to ease the human labor. In recent times, IoT which has been affected by a variety of botnet activities. A botnet is a interconnected device which has a malicious or a harmful software where it can affect the electronic devices and it has been damaged over the past few years, existing Network forensic techniques where it can identify where the data packet has been lost but it cannot exactly track the data packet. Using IoT combined with ML we can identify it. And we cause also use the upcoming technology blockchain to increase the security with data passed over any network. Block chain is immutable where the data sent the integrity would be same no one can alter it.

VII. CONCLUSION:

There is an apparently scarcity of scientific research and technology in the area of database forensics. This despite uses wide range of usage and importance of databases in today's Information systems. Using supervised ML in digital forensic where the network forensic will help us to classify the data packet into different classes and it is done by Support Vector Machine algorithm which is the best algorithm for classification. After classifying the data packet it will be further identified by whether the data packet is lost or not if the data packet is not lost the classification stops there but if the data packet is lost it will help us to identify where the data packet have been lost and it will ease our work helping us to trace the lost data packet by using the slope from linear regression.

REFERENCES:

1. https://scikitlearn.org/stable/modules/generated/sklearn.svm.SVC.html
2. https://scikitlearn.org/stable/auto_examples/linear_model/plot_ols.html
3. M. Fasan and M. S. Olivier, "Reconstruction in Database Forensics," in Advances in Digital Forensics VIII, G. Peterson and S. Shenoi, Eds. Heidelberg, Germany: Springer, 2012, pp. 273–287.
4. H. Pieterse and M. S. Olivier, "Data Hiding Techniques for Database Environments," in Advances in Digital Forensics VIII, G. Peterson and S. Shenoi, Eds. Heidelberg, Germany: Springer, 2012, pp. 289–301.
5. W. K. Hauger and M. S. Olivier, "The role of triggers in Database Forensics," in Proceedings of the 2014 Information Security for South Africa Conference, Johannesburg, South Africa, Aug. 13–14, 2014.
6. "The Impact of Triggers on Forensic Acquisition and Analysis of Databases," Africa Research Journal, vol. 106, pp. 64–73, Jun. 2015.
7. K. Fowler, SQL Server Forensic Analysis. London, Great Britain: Pearson Education, 2009. D. Litchfield, The Oracle Hacker's Handbook: Hacking and Defending Oracle. Indianapolis, IN: John Wiley & Sons, 2007.
8. M. T. Pereira, "Forensic analysis of the Firefox 3 Internet history and recovery of deleted SQLite records," Digital Investigation, vol. 5, pp. 93–103, Mar. 2009.
9. H. Chivers and C. Hargreaves, "Forensic data recovery from the Windows Search Database," Digital Investigation, vol. 7, pp. 114–126, Apr. 2011.
10. M. S. Olivier, "On metadata context in Database Forensics," Digital Investigation, vol. 5, pp. 115–123, Mar. 2009.
11. M. Atkinson et al., "The Object-Oriented Database System Manifesto," 1989.
12. J. Beall. (2014, Nov.) Google Scholar is Filled with Junk Science. [Online]. Available: http://scholarlyoa.com/2014/11/04/google-scholaris-filled-with-junk-science/
13. C. Grimes. (2010, Jun.0
14. Casey, Eoghan. "Tool review—WinHex." Digital Investigation 1.2 (2004): 114-128.
15. Wazid, Mohammad, et al. "Hacktivism trends, digital forensic tools and challenges: A survey." Information & Communication Technologies (ICT), 2013 IEEE Conference on. IEEE, 2013.
16. Yasin, Muhammad, and Muhammad Abulaish. "DigLA–A Digsby log analysis tool to identify forensic artifacts." Digital Investigation 9.3 (2013): 222-234.
17. Marrington, Andrew, et al. "CAT Detect (computer activity timeline detection): a Tool for Detecting Inconsistency in Computer Activity Timelines." digital investigation 8 (2011): S52- S61.
18. Olsson, Jens, and Martin Boldt. "Computer forensic timeline visualization tool." digital investigation 6 (2009): S78-S87.
19. Casey, Eoghan, and Aaron Stanley. "Tool review–remote forensic preservation and examination tools." Digital Investigation 1.4 (2004): 284-297.
20. Jansen, Wayne, and Rick Ayers. "An overview and analysis of PDA forensic tools." Digital Investigation 2.2 (2005): 120-132.
21. Inoue, Hajime, Frank Adelstein, and Robert A. Joyce. "Visualization in testing a volatile memory forensic tool." Digital Investigation 8 (2011): S42-S51

AUTHORS PROFILE

Ms. R. RAJYASHREE
Assistant professor Computer science and engineering
Srm institute of science and technology
Tamil Nadu, India

Ms. NIMISHA PRAVEEN
Student, BTECH Computer science and engineering
Srm institute of science and technology Tamil Nadu, India.

Mr. DHEERAJ.J, Student, BTECH, Computer science and engineering Srm institute of science and Technology Tamil Nadu, India

Mr. SUDHARSHAN.R
Student, BTECH, Computer science and engineering Srm institute of science and technology Tamil Nadu, India.