# Towards Performing Classification in Cloud Forensics with Self Organizing Maps

**Sugandh Bhatia[1], Jyoteesh Malhotra[2]**

[1]Assistant Professor in Computer Science, Punjab School of Economics, Guru Nanak Dev University, Amritsar – 143005, India
[2]Professor and Head, Department of Computer Science and Engineering, GNDU Regional Campus, Jalandhar, India

**Summary**
In the recent years, cloud computing applications and services have witnessed the exponential growth in every sphere of Government, non-government organizations and corporate houses. Security, privacy, compliance, SLA, integrity and availability are the key issues in cloud computing. This research article concentrates on the investigation and analysis by implementing self organizing maps. The investigation is conducted on a set of file systems such as internet operations, email and exif data across two forensic cases named as SEL01 and SEL02. Outcome of the analysis divulged that self organizing maps can be executed as a significant unsupervised clustering tool in digital forensics. Hence, an effort has been made to apply various tools and techniques to extract proof and evidence from different sets of data and on the basis of output; it is feasible to find out the percentage of significant and insignificant files.

***Key words:***
*cloud forensics, self organizing maps, machine learning, cluster, significant.*

## 1. Introduction

Machine learning techniques are ubiquitous. Daily life transactions like financial, automobile, health and research is being controlled and managed by machine learning techniques. Another important subset of machine learning is deep learning [1]. Many companies like Facebook, Amazon, Azure, Google, Hitachi and IBM are deploying it. Deep learning is very useful in large datasets. However, it requires large number of neural networks. As, it can be applied in case of large amount of data. Obviously, for the training purpose, it needs huge amount of memory. In other words, we can say that deep learning models require large amount of data, huge amount of memory, more power for execution and finally, more overheads for the successful implementation and execution of deep learning. One of the most popular, important and useful computing paradigm is cloud computing and towards achieving privacy and security in cloud computing [2], machine learning algorithms can perform a pivot role. Therefore, the objective of this research paper is to explore and investigate cloud forensic cases with the help of self organizing maps [3] in the network size of 3×3, 5×5 and 7×7.

Manuscript received August 5, 2020
Manuscript revised August 20, 2020

## 2. Literature Review

In this segment, we included momentous literature on network forensics, digital forensics and cloud forensics [4]. The principal objective of this research domain is to accentuate the machine learning applications of cloud security and forensics.

Liao et al. [5] suggested a model which is designed on the basis of fuzzy theory. This model is capable to perform inference with the inclusion of network traffic. However, the success rate of this model is 91%, but the major limitation of this model is that, it is implemented only on the DARPA [6] dataset.

Rahman et al. [7] explained a framework for the analysis of big data in cloud. This framework is based upon map reduce [8] technique. The overall working of this model revolves around the healthcare analytics, which is important application of machine learning and healthcare big data is a hot topic of research these days.

Khan et al. [9] proposed a mechanism which performs a crucial role in the design of time line of the associated events. The mechanism is dependent on the neural networks for execution. With the combination of four information sources, a timeline is created. These sources are operations in the file system, log files, free & blank chunks of memory and entries in the registry for windows based system.

Salloum et al. [10] suggested in an article, the criteria to perform analysis on the Apache Spark platform. In case of big data analytics, Spark is one of the important and popular platforms. Graph analysis, structured data processing and fault tolerance are major advantages of Spark. Unexplored areas and benefits of Apache Spark discussed in the article along with machine learning technique such as Mahout, H2O and SAMOA.

Ismail et al. [11] proposed and implemented a framework, which is designed specifically for machines running on Windows based operating system. Digital forensics investigations in real time can be performed with the help of XLIVE framework that is entirely dependent upon XML.

It is an useful tool for analysing huge amount of data and overall functioning of XLIVE is based upon three phases.

Fiore et al. [12] described a mechanism of cloud infrastructure for the analytics of big data. This mechanism deals with environmental change and biodiversity. Therefore, this model depicts the multifarious applications of cloud computing and machine learning. Climate, financial services, data analysis and health are the major areas of cloud computing.

Teing et al. [13] discussed in their research paper regarding the investigation of big data. Cloud and big data are the inter-related techniques. CloudMe forensics framework is explored by authors who is originally owned by Xcerion [14]. Impressive combinations of client and web applications are revealed in the paper. The pivot outcome of the article is that the cache database, log files, web caches and configuration files are equally important for the digital forensic process.

Pichan et al. [15] explained a digital forensic framework and classified the forensic process in to following phases: (i) identification (ii) preservation (iii) acquisition (iv) investigation and analysis (v) presentation and reporting. Moreover, all the phases are further classified in to sub phases along with explanation, challenges and solutions at each level.

Selvaraj et al. [16] suggested a mechanism for the estimation of trust in cloud which is based upon fuzzy logic. It is a dynamic model that is deployed on the basis of evidence and services available in the system. Real time performance is achieved with the help of fuzzy logic. To maintain trust and control the uncertainty, weight averaging operator is used to collect trust values. High effectualness and competency is achieved and proved through simulation. The major limitations of this model are: less safety, low confidentiality and lesser integrity.

Chahal et al. [17] have proposed an expert system to determine the trustworthiness of service provider in the cloud with the help of fuzzy logic. On the basis of important parameters such as security, usability, performance, reliability and trust, a comparison is accomplished on five major cloud service providers. To perform the fuzzy logic in MATLAB, fuzzy logic toolbox is used and positions of cloud service providers displayed as per the ranks on the scale of 1 to 5. This expert system provides security, scalability and reliability. Major drawbacks of this system are less dependability, lesser safety and low confidentiality.

## 3. Cloud Forensic Process

This section focuses on the traditional cloud forensic model. Figure 1 demonstrates the mechanism, which aimed to furnish a thorough chronological process that comprises of following phases:

- Identification: It is the first step in the process and it revolves around the detection of malicious activity like hijacking of accounts, data breach, and malware injection, denial of service and loss of data. The forensic procedure starts with identification of digital evidence. The evidence could be a snapshot of log, virtual machine, file and data stored on the server. The phase of identification can be further classified into two sub phases: (a) identification of incident (b) identification of evidence. The first sub phase reports the occurrence of malicious activity in the system. It can be from any side either customer or service provider. Various resources such as file systems, memory, machines and log tables are identified to collect the required evidence. The second sub phase performs the job of collection and identification of evidence from the resources like selected memory, files and logs. The focus of this sub phase is on the collection of digital artifacts that are permissible in the court of law.
- Acquisition: The acquisition phase is considered as one of the important phase of the process. An error can't be afforded during the collection of data as it can affect the whole procedure of digital forensic in the cloud in an adverse manner. As the nature of operations in the cloud are evanescent and not regular. The evidence artifacts are not easily accessible. Therefore, the acquisition and collection of evidence is a tedious task to perform in cloud environment. At this level, along with acquisition of evidence, one more function is performed and that is preservation. The main goal of preservation phase is to ensure the safety and integrity of acquired evidence. The acquired evidence is to be presented in the court of law. The overall forensic process is dependent upon the seized evidences. Furthermore, the originality and integrity of evidence should be ensured during the entire investigation process.
- Authentication: This phase consists of all the techniques and methods applied for the verification and corroboration of the authenticity of acquired evidence or data. Hashing performs an impressive role at this level. After the completion of acquisition phase, it is required to verify that the data and evidence has taken from the authentic source. MD5 and SHA1 are the popular and useful hashing techniques that can be applied for the purpose of authentication.
- Investigation and Analysis: The process starts after the acquisition and collection of digital evidence and facts. In the field of cloud forensics, investigation can be discussed as a forensic method and tool suitable to the huge variety of

data that was acquired and processed to identify and obtain the required information from the acquired evidence and data. In case, the acquired data or evidence from the acquisition phase is insignificant, then, the whole procedure should be repeated from the first level which is identification phase.

- Results and Summary: The results and summary phase consists of conclusions, findings of the forensic study and presentations. Digital evidence, investigation reports and final summary are submitted in the court. NIST explained reporting as a process that "include describing the actions performed and recommending improvements to policies, guidelines, procedures, tools and other aspects of forensic process."
- Preservation: The last phase in the life cycle of cloud forensic is preservation phase. Once the evidence and documents submitted in the court and judgement given, there is a possibility that in future, the evidences and documents will be required. All the evidences, documents, logs and files must be preserved under the safe custody and it should be ensured that integrity, truthfulness and originality of all the documents is maintained.

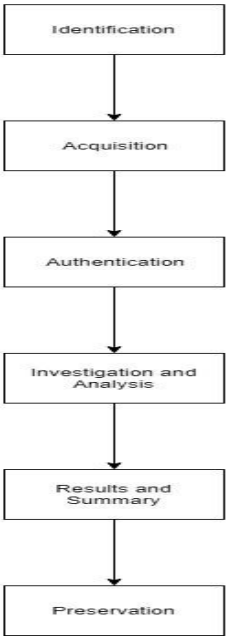


Fig. 1 Cloud Forensic Process

## 4. Gaps in Classical Cloud Forensic Model

It is true that cloud computing has transformed the method of performing computations and data storage in todays' digital world. In earlier days, computing was considered as a product, but with the advent of cloud computing paradigm, computing is considered as a service and not as a product. The success and popularity of any technique depends upon safety and security measures taken in that technique. The same is applied in the case of cloud computing. In the past couple of years, many scientists and researchers have contributed in the development of cloud forensic tools [18], finding the challenges, designing the architecture and evidence acquisition methods in cloud forensic. Although, all the recent developments point out the technical, legal and social challenges in implementing the digital forensics to the cloud computing environment which is acceptable to all the stakeholders like service provider, customer and forensic expert. In the field of cloud forensic and security, there are so many hidden patterns or areas, which require to explore. Therefore, it is need of the modern computing paradigm to endeavour digital forensic research in cloud computing environment. The major research gaps of traditional cloud forensic process and frameworks are expounded below:

a) Unavailability of uniform tools and techniques in cloud forensic.
b) Evidence acquisition and investigation is a challenging and tedious task in virtualized cloud environment.
c) Lack of co-operation between cloud service provider and customer.
d) Dearth of information regarding the operations and location of data centres to cloud forensic experts.
e) Data and details of a virtual machine is of fickle nature and after the termination of a virtual machine, it is not possible to get back its data.
f) Involvement of various sub phases in the cloud forensic life cycle.
g) Due to multi tenancy and multi jurisdiction, it is a complex job to conform the legal concerns associated to cloud forensics.
h) There is no well - defined SLA framework to conduct cloud forensic.
i) Dependency on cloud service provider is a major hindrance in performing data acquisition and evidence collection.
j) Forensic experts have less control over the data. Even customers are totally dependent on the mercy of service provider as there is no clear guideline in SLA regarding the ownership and control of data.

## 5. Machine Learning in Cloud Forensic

Machine learning is a subset of applications of artificial intelligence that furnishes the competency to computer system to perform without specifically programmed. The

system based on machine learning concentrates on the development of programs that can fetch data and train it to use automatically. On the basis of previous research and methodology which focused on the cloud forensic tools, techniques and related challenges, it is conspicuous that available cloud forensic tools and techniques are not up to the mark and there is a huge scope of improvement. There is dire need to develop effective and efficient mechanism that can resolve all the issues related to the humdrum phases of cloud forensic [19]. The primary objective of this section is to apply important and popular unsupervised machine learning technique – self organizing map with in cloud forensic field to explore the range to which SOM has the proficiency to perform automatic clustering with in the cloud forensic. In this research work an effort has been made to predict whether SOM can collaborate with investigator while performing analytical procedures more effectively over pattern searching in the datasets and generating clusters of association between data. This section also scrutinizes and assesses the whole clustering process and explains the mechanism. Moreover, various SOM applications for digital and cloud forensic investigation are discussed. Now, question arises that why self organizing map is chosen among various techniques available? Basically, cluster analysis is a process of grouping a collection of objects in such a manner that objects in the same group are called cluster. The aim is to group objects with major similarity in one cluster whereas keeping the different one in another cluster. The rationale behind the classification of objects is the value of attributes that is used for elucidating the objects. Clustering analysis [20] is a prominent and convenient unsupervised learning technique which can be applied on a given dataset to find hidden patterns. There are numerous clustering algorithms are available such as K-means clustering, mean-shift clustering, hierarchical clustering and self organizing map. When all the techniques are compared then it will be clearly narrated that SOM is the most effective and popular technique in digital and cloud forensic domain. Therefore, SOM is selected as the technique for investigating the automation on the analysis part of the cloud forensic process. Hence, the authors give a bird's eye view on the SOM is given in the next section.

## 6. Self Organizing Map

The self organizing map algorithm was developed by Konohen more than three decades ago, yet its success rate and usefulness in many fields of science and technology during these 35 years surpasses various other neural networks available till date. The adequacy of SOM can be expressed with the following features: (i) visualization (ii) clustering (iii) processing of data (iv) classification. In other words, it can be said that SOM is a technique that is dependent upon unsupervised learning which means that whole mechanism of learning is data driven and there is a healthy competition between the neurons in the output layer. SOM is broadly applied in domains like biomedical analysis, statistical analysis and various areas of computing such as computational intelligence, software security, intrusion & anomaly detection, denial of service attacks, traffic modeling, classification of spam emails, monitoring of SMTP traffic, hardware security and digital & cloud forensic.

## 7. Working and Experiment Methodology

The objective of this experiment is to find out the scope to which SOM can be applied to design the cluster of artifacts. In practice, the category and nature of cloud forensic cases would vary and categorize into various forms due to medley of cloud crime instances. The important parameters from the investigation point of view are nature & size of snapshot, number of files & artifacts and skills & experience of forensic investigator. A study has been conducted to observe and analyze two cases of digital forensics. It is significant to test and compare a proposed mechanism with pre-defined group of inputs with the objective to verify if it could be executed in real life. A comparative statement is to be made after performing the comparison between similar set of outputs that had found earlier during forensic analysis which performed manually by the investigators. The size of a snapshot can be from MBs to GBs. The basic procedure of implementing self organizing map in digital and cloud forensic depends upon EXIF data [21] and metadata related with the artifacts. Important elements of the metadata are path and location of datacenter & virtual machine, size & nature of file and acquired timestamps. The selected case is labeled as SEL01 and deals with financial fraud. The size of image is 12 GB and it holds 96000 artifacts. Although, the relative number of files with timestamps is 18 and these are termed as significant. The result of every cluster will be evaluated in the form of significant and insignificant. By the term significant we mean that information is associated with the case and it is useful while performing investigation. On the other side, insignificant is considered as absurd to the case. The second case is regarding the hacking and labeled as SEL02. The image size of case is 1.5 GB and 5120 files are available as artifacts, whereas, 1082 files are available with metadata such as time & date of creation, access time & date and timestamp. Moreover, 173 files are significant which means that these are available with timestamp. Especially, the accurate number of files which can be associated with the case are 2840. The major reason of difference between the number of artifacts and selected significant is the availability of timestamps. It is essential to discuss that SOM identifies few shortlisted features for

the processing purpose. Many times, features such as path of file, data center & virtual machine, file extension, deleted, curved or encrypted files were chosen for SOM investigation. In the Tables 1, 2 and 3, a detailed list of characteristics of each and every category of metadata is presented.

Table 1: Shortlisted Characteristics on the Basis of File

| Shortlisted Characteristics | Explanation |
|---|---|
| Date and time of creation of file | It displays the actual date and time of the file creation. |
| Accession and modification time | It expresses the accurate date along with time when the file was accessed and if any modification was made to the file. |
| Actual Path | It presents the actual location of the file on the data centre. |
| Deleted File | Files that are deleted from the system permanently and temporarily. |
| Encrypted File | Any file which is encrypted in the system for the purpose of protection and security. |

Table 2: Shortlisted Characteristics on the Basis of E-Mail

| Shortlisted Characteristics | Explanation |
|---|---|
| Sender of Mail | It displays the address and other information of the sender. |
| Receiver of Mail | It holds the email address of the recipient. |
| Subject | It presents the subject matter or nature of the email. |
| CC and BCC | It refers to the carbon copy and blind carbon copy. |
| Submission and delivery date of time | Refers to the time and date when submission of email was performed and delivered. |
| Attachment | Reveals the type and nature of attachment, if any. |

Table 3: Shortlisted Characteristics on the Basis of File

| Shortlisted Characteristics | Explanation |
|---|---|
| Log Files | It displays a record of events, messages, operations and communication between various applications and operating system. |
| Snapshot of VM | Snapshot file of a virtual machine presents the information regarding state of the VM while creating the snapshot. |
| NVRAM of VM | This is non-volatile random access memory that stores the basic input output system details of the VM. |
| Configuration of VM | It keeps the detail of configuration of virtual machine such as processor, memory, operating system, interface and network adaptors. |
| Disk of VM | It is the virtual hard disk of the virtual machine that runs on guest operating system. It can be of two types either fixed or dynamic. |

## 8. Implementation of Self Organizing Maps

Both the cases SEL01 and SEL02 were applied and tested on various network sizes 3×3, 5×5 and 7×7. It is tough to ascertain optimal network size for the solution of this problem. Actually, it is one of the major issue of classification problem [22]. The most effective and impressive way of analysis is to make a comparison between the density of significant and insignificant files available in a cluster. Moreover, significant found in large density is appreciable as it plays an important role in the reformation of data analysis. The experiments were conducted on the basis of network size and the explanation is given below. As discussed that homogeneous classes of related information may not be available in both the cases. Therefore, as per the availability of metadata, analysis has been conducted in each case. The results of self organizing maps are revealed with the help of unified matrices for each network size. As per density of the significant files, only the 5 top most clusters are displayed and all other significant files are exhibited under the remaining clusters category.

Table 4: Clustering Output for Network Size 3×3 of Case SEL01

| Category | | Analyzed Files | | | | Internet | | | | Email | | | | EXIF | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Significant | | Insignificant | | Significant | | Insignificant | | Significant | | Insignificant | | Significant | | Insignificant | |
| Case Code | Cluster ID | Number | Percentage | Number | Percentage | | | | | | | | | Number | Percentage | Number | Percentage |
| SEL 01 | 1 | 12 | 100 | 48 | 1.75 | -- | -- | -- | -- | -- | -- | -- | -- | 2 | 33.33 | 3 | 14.28 |
| | 2 | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | 1 | 16.67 | 2 | 9.52 |
| | 3 | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | 2 | 33.33 | 4 | 19.05 |
| | 4 | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | 1 | 16.67 | 4 | 19.05 |
| | 5 | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | | -- | -- |
| | Remaining Clusters | 0 | 0 | 2695 | 98.25 | -- | -- | -- | -- | -- | -- | -- | -- | 0 | 0 | 8 | 38.10 |
| | Total | 12 | 100 | 2743 | 100 | -- | -- | -- | -- | -- | -- | -- | -- | 6 | 100 | 21 | 100 |

Table 5: Clustering Output for Network Size 3×3 of Case SEL02

| Category | | Analyzed Files | | | | Internet | | | | Email | | | | EXIF | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Significant | | Insignificant | | Significant | | Insignificant | | Significant | | Insignificant | | Significant | | Insignificant | |
| Case Code | Cluster ID | Number | Percentage | Number | Percentage | Number | Percentage | Number | Percentage | Number | Percentage | Number | Percentage | | | | |
| SEL 02 | 1 | 42 | 30.65 | 115 | 17.77 | 9 | 100 | 16 | 100 | 10 | 37.04 | 63 | 25.61 | -- | -- | -- | -- |
| | 2 | 39 | 28.47 | 124 | 19.17 | -- | -- | -- | -- | 8 | 29.63 | 46 | 18.70 | -- | -- | -- | -- |
| | 3 | 27 | 19.71 | 49 | 7.57 | -- | -- | -- | -- | 4 | 14.81 | 35 | 14.23 | -- | -- | -- | -- |
| | 4 | 14 | 10.22 | 70 | 10.82 | -- | -- | -- | -- | 2 | 7.41 | 24 | 9.76 | -- | -- | -- | -- |
| | 5 | 8 | 5.84 | 42 | 6.49 | -- | -- | -- | -- | 2 | 7.41 | 16 | 6.50 | -- | -- | -- | -- |
| | Remaining Clusters | 7 | 5.11 | 247 | 38.18 | -- | -- | -- | -- | 1 | 3.70 | 62 | 25.20 | -- | -- | -- | -- |
| | Total | 137 | 100 | 647 | 100 | 9 | 100 | 16 | 100 | 27 | 100 | 246 | 100 | -- | -- | -- | -- |

Table 6: Clustering Output for Network Size 5×5 of Case SEL01

| Category | | Analyzed Files | | | | Internet | | | | Email | | | | EXIF | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Significant | | Insignificant | | Significant | | Insignificant | | Significant | | Insignificant | | Significant | | Insignificant | |
| Case Code | Cluster ID | Number | Percentage | Number | Percentage | | | | | | | | | Number | Percentage | Number | Percentage |
| SEL 01 | 1 | 12 | 100 | 26 | 0.95 | -- | -- | -- | -- | -- | -- | -- | -- | 2 | 33.33 | 2 | 9.52 |
| | 2 | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | 1 | 16.67 | 1 | 4.76 |
| | 3 | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | 2 | 33.33 | 3 | 14.28 |
| | 4 | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | 1 | 16.67 | 2 | 9.52 |

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 5 | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- |
| | **Remaining Clusters** | 0 | 0 | 2717 | 99.05 | -- | -- | -- | -- | -- | -- | -- | -- | 0 | 0 | 13 | 61.92 |
| | **Total** | 12 | 100 | 2743 | 100 | -- | -- | -- | -- | -- | -- | -- | -- | 6 | 100 | 21 | 100 |

Table 7: Clustering Output for Network Size 5×5 of Case SEL02

| **Category** | | **Analyzed Files** | | | | **Internet** | | | | **Email** | | | | **EXIF** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Significant** | | **Insignificant** | | **Significant** | | **Insignificant** | | **Significant** | | **Insignificant** | | **Significant** | | **Insignificant** | |
| **Case Code** | **Cluster ID** | **Number** | **Percentage** | **Number** | **Percentage** | **Number** | **Percentage** | **Number** | **Percentage** | **Number** | **Percentage** | **Number** | **Percentage** | | | | |
| **SEL 02** | 1 | 24 | 17.52 | 69 | 10.67 | 9 | 100 | 16 | 100 | 7 | 25.92 | 18 | 7.32 | -- | -- | -- | -- |
| | 2 | 21 | 15.33 | 91 | 14.06 | -- | -- | -- | -- | 4 | 14.82 | 13 | 5.28 | -- | -- | -- | -- |
| | 3 | 14 | 10.22 | 40 | 6.18 | -- | -- | -- | -- | 3 | 11.11 | 27 | 10.97 | -- | -- | -- | -- |
| | 4 | 8 | 5.84 | 22 | 3.40 | -- | -- | -- | -- | 2 | 7.41 | 9 | 3.66 | -- | -- | -- | -- |
| | 5 | 11 | 8.03 | 55 | 8.50 | -- | -- | -- | -- | 2 | 7.41 | 6 | 2.44 | -- | -- | -- | -- |
| | **Remaining Clusters** | 59 | 43.06 | 370 | 57.19 | -- | -- | -- | -- | 9 | 33.33 | 173 | 70.33 | -- | -- | -- | -- |
| | **Total** | 137 | 100 | 647 | 100 | 9 | 100 | 16 | 100 | 27 | 100 | 246 | 100 | -- | -- | -- | -- |

Table 8: Clustering Output for Network Size 7×7 of Case SEL01

| **Category** | | **Analyzed Files** | | | | **Internet** | | | | **Email** | | | | **EXIF** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Significant** | | **Insignificant** | | **Significant** | | **Insignificant** | | **Significant** | | **Insignificant** | | **Significant** | | **Insignificant** | |
| **Case Code** | **Cluster ID** | **Number** | **Percentage** | **Number** | **Percentage** | | | | | | | | | **Number** | **Percentage** | **Number** | **Percentage** |
| **SEL 01** | 1 | 12 | 100 | 18 | 0.66 | -- | -- | -- | -- | -- | -- | -- | -- | 2 | 33.32 | 3 | 14.29 |
| | 2 | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | 1 | 16.67 | 2 | 9.52 |
| | 3 | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | 1 | 16.67 | 2 | 9.52 |
| | 4 | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | 1 | 16.67 | 3 | 14.29 |
| | 5 | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | 1 | 16.67 | 1 | 4.76 |
| | **Remaining Clusters** | 0 | 0 | 2725 | 99.34 | -- | -- | -- | -- | -- | -- | -- | -- | 0 | 0 | 10 | 47.62 |
| | **Total** | 12 | 100 | 2743 | 100 | -- | -- | -- | -- | -- | -- | -- | -- | 6 | 100 | 21 | 100 |

Table 9: Clustering Output for Network Size 7×7 of Case SEL02

| **Category** | | **Analyzed Files** | | | | **Internet** | | | | **Email** | | | | **EXIF** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Significant** | | **Insignificant** | | **Significant** | | **Insignificant** | | **Significant** | | **Insignificant** | | **Significant** | | **Insignificant** | |
| **Case Code** | **Cluster ID** | **Number** | **Percentage** | **Number** | **Percentage** | **Number** | **Percentage** | **Number** | **Percentage** | **Number** | **Percentage** | **Number** | **Percentage** | | | | |
| **SEL 02** | 1 | 19 | 13.87 | 76 | 11.76 | 9 | 100 | 16 | 100 | 4 | 14.82 | 24 | 9.77 | -- | -- | -- | -- |
| | 2 | 23 | 16.79 | 84 | 12.98 | -- | -- | -- | -- | 3 | 11.11 | 16 | 6.51 | -- | -- | -- | -- |
| | 3 | 17 | 12.41 | 56 | 8.65 | -- | -- | -- | -- | 2 | 7.41 | 18 | 7.33 | -- | -- | -- | -- |
| | 4 | 12 | 8.76 | 33 | 5.10 | -- | -- | -- | -- | 1 | 3.70 | 10 | 4.06 | -- | -- | -- | -- |

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 5 | 18 | 13.14 | 69 | 10.66 | -- | -- | -- | -- | 1 | 3.70 | 4 | 1.63 | -- | -- | -- | -- |
| | **Remaining Clusters** | 48 | 35.03 | 329 | 50.85 | -- | -- | -- | -- | 16 | 59.26 | 174 | 70.7 | -- | -- | -- | -- |
| | **Total** | 137 | 100 | 647 | 100 | 9 | 100 | 16 | 100 | 27 | 100 | 246 | 100 | -- | -- | -- | -- |

## 10. Interpretation of Results

The results of experiments reveal the importance of network size of self organizing map. From the analysis, it is clear that output may be dissimilar for the same case whenever the size of network is different. The objective of this section is to explore the number of input types, accurate size of artifacts and exact number of significant and insignificant files along with percentage. As demonstrated in Table 10, in both cases SEL01 and SEL02, clustering performed with the help of self organizing map proved a great success. Clusters were implemented to procure significant files on three types of networks. In SEL01, it was noticed that 100% significant files were identified just in one cluster in the network size of 3×3. In other case SEL02, 94.89% significant files were identified in 5 clusters and 5.11% of the files were covered under the remaining category. In 5×5 network of case SEL02, the output was 56.94%, which means 43.06% of the significant files were covered under the remaining clusters category.

Table 10: Details of Significant and Insignificant Values

| **Network Size** | | **3×3** | | **5×5** | | **7×7** | |
|---|---|---|---|---|---|---|---|
| **Case ID** | **Cluster ID** | **Sig.** | **Insig.** | **Sig.** | **Insig.** | **Sig.** | **Insig.** |
| **SEL01** | 1 | 100 | 1.75 | 100 | 0.95 | 100 | 0.66 |
| | **Remaining** | 0 | 98.25 | 0 | 99.05 | 0 | 99.34 |
| **SEL02** | 1 | 30.65 | 17.77 | 17.52 | 10.67 | 13.87 | 11.76 |
| | 2 | 28.47 | 19.17 | 15.33 | 14.06 | 16.79 | 12.98 |
| | 3 | 19.71 | 7.57 | 10.22 | 6.18 | 12.41 | 8.65 |
| | 4 | 10.22 | 10.82 | 5.84 | 3.40 | 8.76 | 5.10 |
| | 5 | 5.84 | 6.49 | 8.03 | 8.50 | 13.14 | 10.66 |
| | **Remaining** | 5.11 | 38.18 | 43.06 | 57.19 | 35.03 | 50.85 |

Table 11: Significant Value of Internet Files in SEL02

| **Case ID** | **Total Internet Files** | **Significant** | **Significant (%)** |
|---|---|---|---|
| **SEL02** | 25 | 9 | 36.00 |

The internet category was available in case SEL02 only and total number of significant files was 9, it means that 36% of the files were significant. In case of SEL01 no file either significant or insignificant was procured under the category of internet files. On every network, the same number of significant and insignificant files obtained as shown in above tables. The email category presented in the case SEL02 and output received from the network size 3×3, 5×5 and 7×7 clearly depicts the impact of network size. In size 3×3, just 3.7% of the significant files were covered under the remaining category, in 5×5, 33.33% of the significant files were placed in the remaining category and in 7×7, the files under the remaining category were maximum to the 59.26% level. The results revealed that

self organizing map has the ability to arrange the data on the basis of similarity. In the three network sizes significant results obtained without the involvement of remaining cluster. It is noticed that there was a fluctuation in case of insignificant files as 38.10% obtained with remaining cluster in 3×3 network size, 61.92% files procured by remaining cluster in 5×5 network size and in 7×7 network size the percentage of insignificant files under the remaining cluster were 47.62%.

Table 12: Experimental Results for Email Category

| **Network Size** | | **3×3** | | **5×5** | | **7×7** |
|---|---|---|---|---|---|---|
| **Cluster ID** | **Sig.** | **Insig.** | **Sig.** | **Insig.** | **Sig.** | **Insig.** |
| 1 | 37.04 | 25.61 | 25.92 | 7.32 | 14.82 | 9.77 |
| 2 | 29.63 | 18.70 | 14.82 | 5.28 | 11.11 | 6.51 |
| 3 | 14.81 | 14.23 | 11.11 | 10.97 | 7.41 | 7.33 |
| 4 | 7.41 | 9.76 | 7.41 | 3.66 | 3.70 | 4.06 |
| 5 | 7.41 | 6.50 | 7.41 | 2.44 | 3.70 | 1.63 |
| **Remaining** | 3.70 | 25.20 | 33.33 | 70.33 | 59.26 | 70.70 |

## 11. Conclusion

A complete study on the clustering for different network sizes and with various input characteristics is performed. Initially, the study conducted by taking into consideration the network size. Analysis performed on the basis of other variables such as number of significant files, insignificant files, artifacts and percentage of each parameter. The analysis clearly reveals that output received in every category has significance in the following manner: significant files grouped in a separate cluster and density of significant files is higher in every cluster when compared with the population. In fact, in various cases, clusters hold only the significant value, which is important from the analysis point of view. On the basis of overall results, it is clear that the clustering performance was impressive. In larger network size self organizing maps, it was noticed that there is autonomous phenomena that major number of clusters would keep significant files and density of these significant files in clusters would be on the higher side which points out a high clustering output. Furthermore, it can be divulge that a lower number of significant files denote good clustering accomplishment. No doubt, clustering performance also dependent upon the nature of artifacts taken for analysis purpose. In this research paper, major artifact type is analyzed files. This artifact was available in both the cases. The output received from three different network sizes formed compatibility with the received data and supported by the artifacts. Hence, self organizing map can obviously be used for implementing the cloud forensics.

## References

[1] Theodoridis., "Neural Networks and Deep Learning", Machine Learning. doi:10.1016/b978-0-12-801522-3.00018-5, pp. 875-936, 2015.

[2] Hyun-Min Son, Nak-Keun Joo, Hyun-Taek Choi and Hyun-Cheol Lee, "Analysis of Cloud Security Vulnerabilities and Countermeasures", International Journal of Computer Science and Network Security, vol. 19, no. 2, pp. 200-206, 2019.

[3] Ballabio, D., Vasighi, M., and Filzmoser, P., "Effects of supervised Self Organising Maps parameters on classification performance", Analytica Chimica, pp. 45-53, 2013.

[4] Kebande, V.R., and Venter, H., "On Digital Forensic Readiness in the Cloud Using a Distributed Agent Based Solution : Issues and Challenges", Australian Journal of Forensic Sciences, pp. 209-238, 2016.

[5] Liao, N., Tian, S., and Wang, T, "Network Forensics Based on Fuzzy Logic and Expert System", Computer Coms, pp. 1881-1892, 2009.

[6] Download Data Sets (n.d.). Retrieved from https:// web.cs.dal.ca/~riyad/site/download.htm.

[7] Rahman, F., Slepian, M., and Mitra, A., "A Novel Big-data Processing Framework for Healthcare Applications", IEEE International Conference on Big Data. Doi: 10.1109/big data 7841018, 2016.

[8] Maitrey, S., and Jha, C., "MapReduce: Simplified Data Analysis of Big Data", Procedia Computer Science, pp. 563-571, 2015.

[9] Khan, M., Chatwin, C., and Young, R, "A Framework for Post Event Timeline Reconstruction using Neural Networks", Digital Investigation, pp. 146-157, doi:10.1016/j.diin.2007.11.001, 2007.

[10] Salloum, S., Dautov, R., Chen, X., Peng, P. X., and Huang, J. Z., "Big data analytics on Apache Spark", International Journal of Data Science and Analytics, pp. 145-164, doi:10.1007/s41060-016-0027-9 2016.

[11] Ismail, L., Masud, M. M., & Khan, L., "FSBD: A Framework for Scheduling of Big Data Mining in Cloud Computing", IEEE International Congress on Big Data. doi:10.1109/bigdata.congress.2014.81, 2014.
[12] Fiore, S., Mancini, M., Elia, D., Nassisi, P., Brasileiro, F. V., and Blanquer, I, "Big data analytics for climate change and biodiversity in the EUBrazilCC federated cloud infrastructure", Proceedings of the 12th ACM International Conference on Computing Frontiers - CF '15. doi:10.1145/2742854.2747282, 2015.
[13] Teing, Y., Dehghantanha, A., and Choo, K. R., "CloudMe forensics: A case of big data forensic investigation", Concurrency and Computation: Practice and Experience, 30(5), e4277. doi:10.1002/cpe.4277, 2017.
[14] Xcerion – The Innovation Holding Company. (n.d.) retrieved from http://xcerion.com
[15] Pichan, A., Lazarescu, M., and Soh, S. T., "Cloud forensics: Technical challenges, solutions and comparative analysis", Digital Investigation, pp. 38-57, 2015.
[16] Selvaraj, A., and Sundararajan, S., "Evidence-Based Trust Evaluation System for Cloud Services Using Fuzzy Logic", International Journal of Fuzzy Systems, 19(2), doi:10.1007/s40815-016-0146-4 pp. 329-337, 2016.
[17] Chahal, R. K., and Singh, S., "Trust Calculation Using Fuzzy Logic in Cloud Computing", Handbook of Research on Security Considerations in Cloud Computing, doi:10.4018/978-1-4666-8387-7.ch007 pp. 127-172, 2015.
[18] Choo, K. R., and Dehghantanha, A., "Contemporary Digital Forensic Investigations of Cloud and Mobile Applications", Rockland, MA: Syngress, 2016.
[19] Simou, S., Kalloniatis, C., Gritzalis, S., and Mouratidis, H. "A Survey on Cloud Forensics Challenges and Solutions", Security and Communication Networks, doi:10.1002/sec.sec.1688, pp. 6285-6314, 2016.
[20] Samar H. Ahmed, Khalid T. Wassif and Emad Nabil, "Clustering Based Sentiment Analysis Using Randomized Clustering Cuckoo Search Algorithm", International Journal of Computer Science and Network Security, vol. 20, no. 7, pp. 159-166, 2020.
[21] Smirnov, K.O., "The Analysis of Digital Images through the Exif – Standard", Interactive Science (12), pp. 243-246, 2017.
[22] Han, J., Pei, J., and Kamber, M., "Data Mining: Concepts and Techniques", Amsterdam, Netherlands, Elsevier, 2011.

**Sugandh Bhatia** is currently working as Assistant Professor (Computer Science) in Punjab School of Economics, Guru Nanak Dev University, Amritsar – 143005, India and pursuing his PhD in the Faculty of Engineering and Technology, Guru Nanak Dev University, Amritsar. His field of research interests includes Cloud Computing, Cloud Forensics, Data Security and Privacy.

**Jyoteesh Malhotra** is Associate Dean of Academics, Student Welfare and Professor and Head in the Department of Electronics and Communication Engineering and Department of Computer Science and Engineering in the Regional Campus Jalandhar of Guru Nanak Dev University, Amritsar. He received PhD, M.Tech (Gold Medalist) and has more than 20 years of experience of teaching and research. He has more than 175 publications in journals of International and National repute in his credit. His research interests include Wireless Networks and Optical Communication.