APPLICATIONS OF MACHINE LEARNING, DEEP LEARNING AND NEURAL NETWORKS IN CLOUD FORENSICS

ABSTRACT:

Cloud computing has recently experienced rapid expansion, making it a major target for cybercrimes. Though cloud computing has many advantages, there are also substantial security issues related to confidentiality, integrity, privacy, and availability. Being a relatively new discipline, cloud forensics faces numerous challenges and issues with interpreting and analyzing data. Forensic investigators and law enforcement confront numerous difficulties in data collection, data protection, and evidence access. Several sophisticated models have been proposed in recent years aiming to accelerate the entire investigation process or address several issues that arise frequently in forensic investigations. This review paper seeks to comprehend the significance of various machine learning models which would help cloud forensics advance significantly.

1. INTRODUCTION:

Cloud computing has seen substantial change in the recent period. It is expected to be the most breakthrough technology ever developed [1]. Cloud computing has drastically transformed and changed the way IT resources are managed. Businesses are abandoning traditional methods of employing IT resources, in pursuit of cloud computing. Simply put, cloud computing is the delivery of computing services over the internet, i.e., the cloud, which offers quicker and more flexible resources. Cloud computing is cost-effective since customers only pay for what they use. Under the pay-as-you-go concept, cloud computing assists organizations in lowering their operating expenses. Cloud computing is gaining popularity among businesses because it improves speed, efficiency, outcomes, reliability, and security [4]. Cloud computing is flexible and adaptive enough to enable remote data access and scalability as business requirements change [3].

Cloud computing protects huge volumes of privately owned data, which unavoidably contributes to the operation of cybercrime. Digital forensics on the cloud is gaining popularity as a result of numerous cybercrimes that are being committed there. By the official definition offered by NIST [1,5], digital forensics describes the application of technology to the identification, assessment, gathering, and analysis of information while safeguarding the data and keeping a precise chain of custody for such data. Cloud crime, as described by Ruan et al., is a crime that takes place in a cloud-based computing environment and uses the cloud as the object, subject, as well as tool of the crime [11].

Cloud forensics refers to the use of digital forensics within cloud technology and is a subdivision of network forensics. Traditional approaches to forensic inquiry are less effective and successful due to decentralized data processing. Cloud forensics aids in overcoming the drawbacks of conventional methods [1]. Numerous problems and challenges are brought on by the emergence of cloud computing, notably in cloud forensics. Advancements in Big Data Analytics, Artificial Intelligence, and machine learning made cloud forensics more advanced. The overall execution of cloud-based forensics seems complex, with several problems and difficulties involved at each level of cloud forensics, according to several studies undertaken by various academics [6].

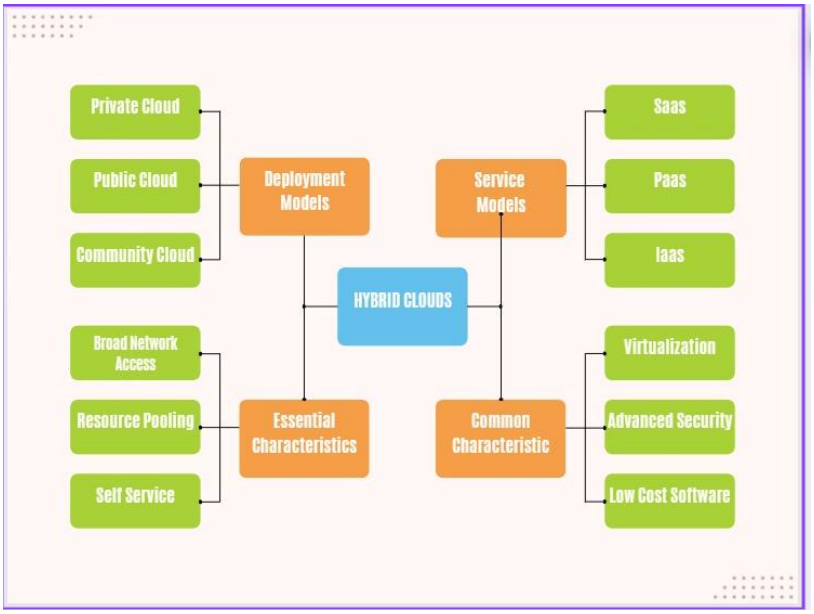
In digital forensics, machine learning (ML) approaches can locate evidence more quickly than manual evaluation of enormous volumes of data produced from numerous sources. Investigators will therefore need to focus more on analyzing criminal dynamics and disclosure. Additionally, a range of digital criminal scenarios can benefit from the use of pattern-matching algorithms, anomaly detection tools, as well as other supervised and unsupervised ML models to enhance cyber forensics outcomes. Furthermore, deep learning algorithms may aid in locating intended evidence in unorganized data by creating connections and

identifying other hidden patterns. Forensic investigators are using big data and Deep Learning (DL) approaches to solve this problem. Deep Neural Networks (DNN) have lately become successful at classification and recognition tasks. As a result, DNN systems with human-like accuracy on underlying data have gained wide acceptance and deployment.

In this review paper, we demonstrate various machine learning, deep learning, and neural network models applied in cloud forensics and show how they assist in simplifying and improving cloud forensics. The scope of this review article will be on understanding how various machine learning, deep learning, and neural network models are employed in forensic investigations in the cloud. The rest of the paper is structured as follows. Section 2 provides an overview of cloud computing, Section 3 discusses cloud security issues, section 4 discusses cloud forensics, and challenges faced in cloud forensics, while Section 5 discusses how machine learning, deep learning, and neural networks models are used in cloud forensics to overcome challenges and Section 6 discusses the existing models in cloud forensics, drawbacks of existing models, and Section 7 discusses the future scope of cloud forensics with machine learning, deep learning, and neural networks models, and Section 8 concludes the paper with discussion and conclusion.

2. CLOUD COMPUTING

Cloud computing is the latest advancement in the IT sector, which presents many promising technological and economic opportunities. Cloud computing is a novel paradigm, also described as on-demand computing [8], that isolates applications and information services from the underlying architecture and delivery methods [12]. Companies rely heavily on third-party vendors to deploy their apps rather than owning a substantial Information technology infrastructure to host them. Furthermore, cloud computing could also save IT expenses by eliminating unnecessary computer energy and storage, cutting maintenance requirements, and minimizing capital spending constraints [9]. Cloud computing has become a sort of technology that delivers internet-based services to users. Cloud computing does have the advantage of allowing users to maintain their information on remote computers that are easily accessible over the internet. It's also described as offering a computer model as a service through the internet. Cloud computing allows users and organizations to use hardware and software which are maintained by cloud service providers (CSP). Users can access any information from any remote location from any device if they are connected to internet utilising the cloud computing [10]. The National Institute of Standards and Technology (NIST) described cloud computing as a concept for providing accessible, on-demand access to a collective pool of computer resources such as servers, services, storage, applications, etc... Cloud computing has five main features, three delivery types, and four deployment models [10]. It is shown in the below figure.



CLOUD DEPLOYMENT MODELS

Many cloud computing models have emerged, as the cloud has increased in popularity. Three of the six deployment model types—Private, Public, and Hybrid models —are the most common. Other deployment models are community, Virtual Private, and Inter-Cloud, which contain two different types of clouds: Federated Clouds and Multi-clouds [45].

Private Cloud: This is a technology that is solely used by one company. It could be stored locally at the company or in the cloud service provider's data centre [10]. It may also be referred to as a corporate or internal model. The organisation that controls a private cloud centrally controls and manages the system. Despite the fact that third parties can host private cloud servers (such as a service provider). Most companies opt to retain the infrastructure in the data centre closest to them. The organization's internal employees may then monitor and manage everything [45].

Public Cloud: The public or significant infrastructure firms are provided with these [12]. The customer obtains services online while the company's whole infrastructure is located on a public cloud. Customers do not need to manage their own infrastructure, and it is simple for them to add more users or processing power as needed. The consumers receive services through the internet and the whole IT infrastructure is housed in a public cloud. Customers may quickly add additional users and computer processing capabilities as needed and aren't required to worry about maintaining one's own IT up to date. Several tenants share the IT infrastructure of the cloud provider in this arrangement [10].

Hybrid Cloud: Public and private clouds are integrated into hybrid clouds. They are constructed in such a way that applications and data flow seamlessly together, so the two platforms communicate smoothly. The public or significant infrastructure firms are provided with these [12]. The customer obtains services online while the company's whole infrastructure is located on a public cloud. Customers do not need to manage their own infrastructure. It is the ideal answer for a company or organization that could use both cloud models, which are often company and capacity, dependent. In general, a hybrid cloud initiates only as a private cloud and subsequently expands the assimilation to include several cloud computing services. When a company has confidential material that cannot be hosted on the web or legal obligations that need data security, storage, and other services, this deployment strategy makes sense [45].

3.Cloud Service Model:

Cloud computing is typically divided into three levels of service offering such as Software as a Service, Platform as a Service, and Infrastructure as a Service.

PaaS (Platform as a Service): A certain percentage of application software, including certain integration with a standard set of applications, can be provided together with the hardware (as it does in IaaS). You may construct your applications on a foundation of databases or programming functions. Platform as a Service (PaaS) has always been a platform for developing and deploying applications that are made available to programmers as a service through the Internet. It provides all the facilities needed to support the full life cycle of creating and providing web apps and services that are altogether available from the Web, making it easier to build and run applications without the expense and complexity of purchasing and trying to manage the underlying core infrastructure. The infrastructure software that makes up this platform generally comprises the database, middleware, and developer tools [45].

SaaS (Software as a Service): This is the notion that someone may provide you with a hosted set of software (operating on a platform as well as infrastructure) which users do not really own but are expected to pay for some part of usage – by the client, or some other type of consumption basis. You don't need to perform any coding or development work here, but you may be required to come in and set up the (very flexible, adaptable, and occasionally customised) software. You are not required to make any purchases. Just what you use is charged. An application is often hosted and managed by a

SaaS provider within their own data centre and made accessible to several tenants and users through the Internet [45].

IaaS (Infrastructure as a Service): (It is based on the service being introduced upon the virtual server, and the service is paid to the customer in a short period of time. It makes cloud infrastructure, which consists primarily of servers, storage, CPU, memory, and other tools, accessible via any Internet technology [10]. The provision of hardware (server and storage including network) and related services is known as IaaS. IaaS is simply software as a service, including file systems and virtualized operating systems. It is a development of conventional hosting which does not call for a commitment over an extended period of time and enables users to provide resources as needed. Users must deploy and administer the software services themself, just as they would do in their own datacenter, in contrast to PaaS services, where the IaaS provider manages the data centre primarily only to keep it operating [10].

4. CLOUD SECURITY

Since its start, the cloud has seen a significant improvement. As the cloud gained prominence in the IT sector, hackers also began to pay attention to it. At first, there were trust difficulties with cloud computing. Businesses were apprehensive about moving their data onto the cloud. However, the cloud's benefits—mostly financial—outweigh its drawbacks. Organizations may benefit from the cloud in many ways, but there are also security risks and worries that come with it. Traditional security technologies and methods are sometimes ineffective in adequately securing cloud-based infrastructure because it differs greatly from the on-premises data centers.

SECURITY RISKS:

Any company that attempts to keep information on either host systems or on a public cloud loses the ability to physically access the servers that are housing its own data. This makes potentially sensitive information susceptible to insider assaults. Recent analysis according to the Cloud Security Alliance places insider attacks as the sixth greatest danger to cloud computing. People who have physical access to data centre servers must pass extensive background checks in order to prevent dangers of this nature. Data centres must be regularly checked for any unusual activity [13]. The biggest security issue with the cloud is discovered to be trust, secrecy, and encryption.

1. Trust Problem: The SLA contract binds the servicing ends since the person at the receiving end may never be certain that the supplying end is giving reliable data. The SLA agreement outlines the service provider's obligations and upcoming actions. A trust management paradigm regarding security within cloud environment is the SLA framework [13].

2. Confidentiality Problem: It forbids the leaking of any private data. Data privacy and confidentiality may be attained if we construct a safe storage service in a public cloud architecture and use cryptographic analysis. Today, a new P2P reputational system is utilized to protect privacy. It provides virtualized defences and discusses attribute-based cryptography to enable flexible and responsive data sharing by users [13].

3. Authenticity Problem: Any user who's using the cloud must be an authenticated user. The use of digital signatures often solves the authentication problem. A reliable and decentralized control system for access control has been presented, in which the cloud maintains the user's data and validates the identity of the user without knowing any personal information about them. Legitimate users can decode the information that has been stored [13].

4. Encryption Problem: Encryption is the most used method of data security in cloud computing. But encryption's fundamental flaw is that it requires a lot of computer power. Because decryption is needed each time a query is done, it also lowers overall database speed. Combining cryptographic algorithms, as opposed to using a single method, effectively increases performance. The cloud tables usually are kept in such a way that these techniques may be used to encrypt data [13].

5. Key Management Problem: A significant security issue in the cloud is how to handle keys, encryption, as well as decryption in an effective manner. The encrypted key is stored in the cloud, which is a pretty sophisticated method. To store keys safely, a tiny database should be kept up to date. Because of the extra hardware and data that are necessary to achieve this, access control mechanisms are needed, which raises the cost of implementation. This issue is resolved by two-level encryption [13].

6. Data Splitting Problem: It functions more quickly than encryption and is a substitute for encryption. Data splitting would be a method for distributing data to many hosts at once when they are unable to interact separately. In order to get the original data, customers must thus visit all service allocators when they want their data back. Additionally, it has a few security issues. The Multi-Cloud Databases Model may be employed to ensure integrity following the splitting process if many clouds are being used. Security is given more importance as a result of the need to replicate data across many clouds. because the potential for the attacker to access several clouds at once will be reduced [13].

7. Multitenancy problem: Resources being dispersed across many cloud environments' geographies may cause confidentiality issues. Applications and systems should be isolated for adequate confidentiality; failing to do so might result in security problems. When data is kept virtually, a virtual machine running a malicious application might have an impact on the entire system [13].

5. CURRENT THREATS AND ATTACKS IN CLOUD:

The security of cloud computing is open to several threats. Based on security requirements (confidentiality, integrity, and availability), this table provides an overview of the risks and attackers to the cloud service model [22–24].

Table 1: Cloud Threats and Attacks [9]

Cloud Threats	Attacks on Cloud
Insider threat scenarios	Malware Injection Attack: ● SQL injection attack ● Cross-sites scripting attack
threats and external attackers	Denial-of-Service attack
Leakage of data	Side Channel attack
Segregation of data	Authentication attack: ● Brute Force Attacks ● Dictionary Attack ● Shoulder Surfing ● Replay Attacks ● Phishing Attacks ● Key Loggers
Client access a physical disturbance Utilizing poor recovery process	Man-In-The-Middle Attacks: ● DNS Spoofing ● Session Hijacking

6.Cloud Forensics:

Cloud computing forensics is described as "the use of scientific principles, technological procedures as well as derived and proved methods to recreate previous cloud computing incidents through recognition, gathering, retention, investigation, explanation, and disclosure of digital evidence" [2] by the NIST. These days, as digital technology develops quickly, it takes a lot of processing capacity to interpret the data these gadgets produce. The idea of a "Forensic Cloud" is put up with the intention of enabling an investigator to concentrate entirely on the investigative procedures [15].

The definition of cloud forensics is "A method of cloud forensics is described as an examination of a cybercrime which requires proof or evidence obtained from any of the cloud-based platforms or cloud-based services" [2]. Since systems crime scene investigation controls scientific investigations in any type of system, whether it be private or public, cloud forensics may also be seen as a component of system legal sciences. As a result, cloud computing relies on extensive and widespread system access and, with very few techniques specifically designed for cloud computing environments, adopts the core standards established in the system measurement approach [1]. The amount of crimes involving computers and web significantly increased over the past ten years, which has led to an equal rise in businesses that aim to help law enforcement by utilising digital evidence to identify the offenders, techniques, casualties, and chronology of cyber-crimes. Digital forensics improved as a result, ensuring accurate presentation of data used as proof of criminality in court. However, as storage capacity outpaces network performance and latency advancements, forensic data is beginning to expand rapidly, making it more difficult to analyse them quickly [16]. We must use the forensics process provides the cloud examination since, as was previously said, cloud forensics would be a cross-disciplinary field of study including both cloud computing & digital forensics. The cloud is a type of data container in which digital information is stored in virtual pools, physical storage would be spread over several servers, and these systems hold a significant quantity of differential data that must be carefully protected. In order to execute forensics acquisitions of the cloud-based environment in the event of an intrusion as well as fake access, one needs a precise and quick-tracking instrument or method. Because the storage space is so large and might be spread over several servers in various places, it is not viable to undertake forensics on the entire cloud area. The amount of time it takes to acquire the entire cloud is unlimited. This paper only discusses the forensics of a small region where some fraudulent activity has been committed or an area that is linked to that area. Here, "forensics" refers to the process of analyzing cloud data to discover incriminating information about the intrusive party. The forensics effort will entail a wide range of tools & software that, to a certain extent, can assist us in doing cloud data analysis [17]. Lack of physical access to computers presents new and disruptive issues for forensic experts in cloud forensics. Traditional methods of evidence gathering, and recovery no longer work due to the decentralized nature of data handling on the cloud. The technological features of digital forensics in a dispersed cloud setting are the main emphasis of this research [18].

PROCESS OF TRADITIONAL FORENSIC INVESTIGATION:

The traditional computer forensics method consists of several processes, however it may be roughly divided into the four essential phases that are described here.

Collection: The initial step in the process would be to locate, classify, record, and collect data from any potential relevant data sources while adhering to policies and procedures that protect the data's integrity.

Examination: In order to evaluate and extract data of specific relevance while maintaining the data integrity, examinations require forensically processing huge volumes of acquired data using a combination of automated and human approaches.

Analysis: The procedure then moves on to the analysis of the examination's findings using methodologies and methods that are legally acceptable in order to provide information that is helpful in answering the queries that served as the basis for the gathering and analysing.

Reporting: Reporting the analysis' findings involves summarising the steps taken, elaborating on the choices made regarding tools and procedures, identifying any additional steps that must be taken, and making

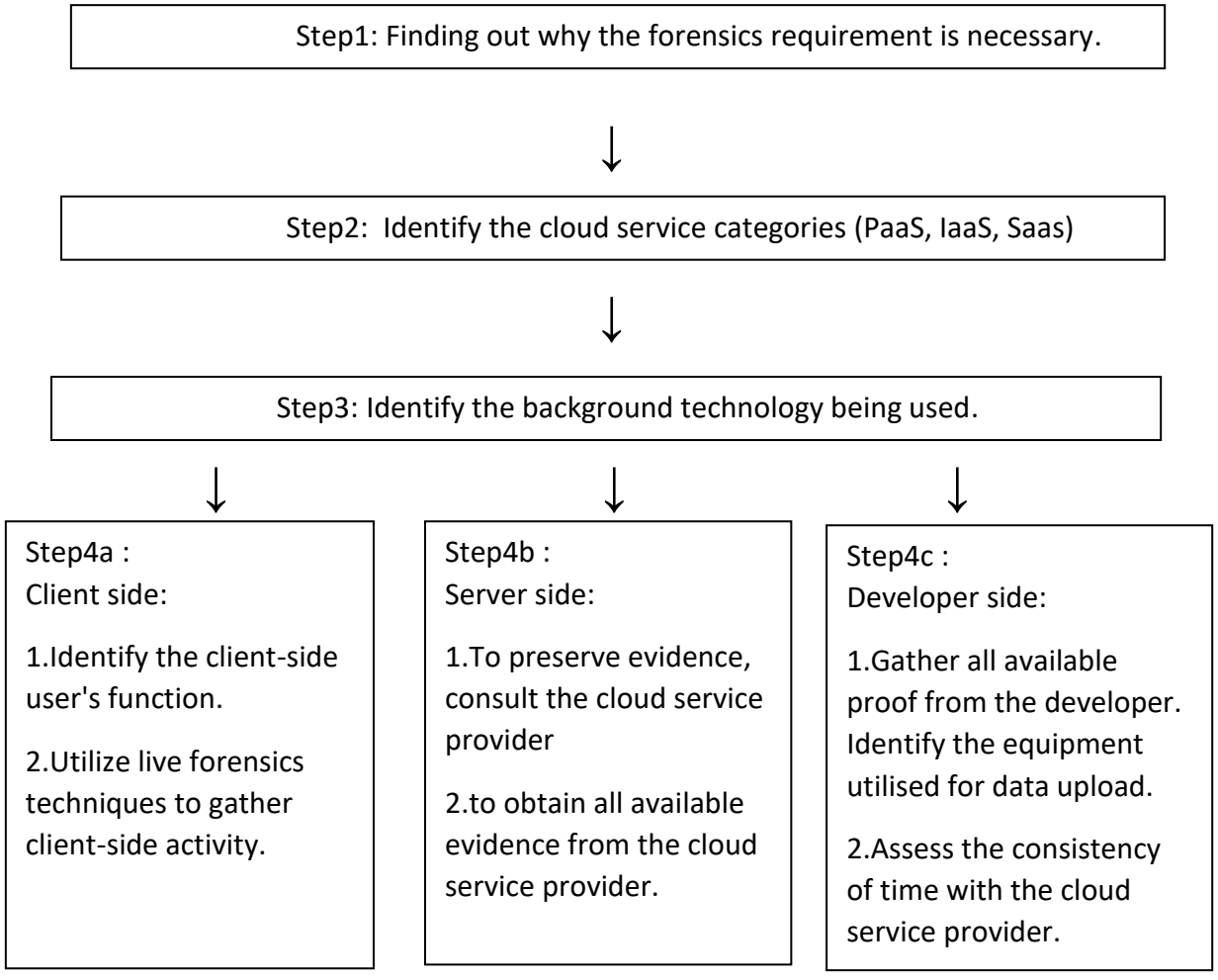
recommendations for changes to be made to policy initiatives, standards, procedures, instruments, and other facets of the forensic examination [19].

PROCESS OF FORENSICS INVESTIGATIONS IN CLOUD:

The development of cloud computing has altered the forensic process's workflow today. In these virtual worlds where discs, storage, and network connections are shared and conventional ownership borders are blurred, we hardly have the power to physically acquire items. Very little study has been done to date on the condition of the tools, procedures, and approaches for obtaining legally acceptable forensic evidence in the cloud. The Forensics and Cloud Security Alliance experts concur that further study is required to create a framework of techniques and procedures that can withstand scrutiny in a court of law.

They advise "having the ability to restore systems to past states, or even a necessity to go beyond 6-12 months for a recognised config." In order to enable the forensic monitoring of event logs while keeping in mind legal options and duties, corrective action might also be needed [20].Basic forensic concepts and procedures apply when conducting forensic investigations in cloud settings, whether for retention, presentation in court, or the independent inquiry of employee misconduct.

The forensic process is broken into four distinct steps: fig 1



The same forensic procedure is used in cloud forensics, but it presents a problem because it must combine different logical and physical locations. These include the following areas [21]:

1. **Client-side:** technical checks or controls put in place on computers and networks that are owned or controlled by clients (IDS, firewalls, access, chat logs locally etc.)

2. **Server-side:** technical checks or monitoring done on systems and networks used by cloud clients (access, transaction, and usage logs, etc.)
3. **Developer-side:** Technical checks and balances made on the networks and systems that make up or enable the cloud service (firewalls, admin access logs, IDS, etc.)

7.CHALLENGES IN CLOUD FORENSICS:

To finish the investigations' findings, it is necessary to overcome various difficulties with the cloud forensics methodologies and investigation models. This section is a list of some of these difficulties.

Recreation of crime scene:

The investigative methods may call for reenacting the entire illegal conduct in order to locate the evidence. All the actions taken at the crime site are repeated and simulated during this process. This is made more difficult by cloud computing since clouds are kept as virtual instances that may be erased after a crime to leave no evidence and provide space for recreation.

Multi-national laws:

The investigations are carried out in accordance with the laws of the nations where the illegal crimes are committed. Due to the fact that cloud data centres are situated all over the world, certain rules must be adhered to in order to carry out the investigations in the case of cloud forensics.

Presenting of evidence in court:

For the case investigations, the presentation of the discovered evidence in court is necessary. Members of the jury might not be familiar with the intricate design of underlying cloud computing models. Evidence presentation thus becomes a challenging job in cloud forensics.

Crime Scene Reconstruction:

It is sometimes necessary for the detectives to recreate the scene of the crime in order to look into the harmful behaviour. In conventional computer forensics, the investigators can quickly determine the number of devices utilised in the crime or even the participants. To rebuild the scene of a crime and determine the extent of the damage, however, as a result of the extremely dynamic structure of the communications, the cloud context necessitates real-time and autonomous interaction between multiple nodes.

Evidence Segregation:

Utilizing virtualization, many instances operating on a same physical computer are separated from one another in the cloud. Despite the fact that their data instances are kept on a single system, many users behave as though they are each operating on a separate host. Therefore, it is very difficult for CSPs & law enforcement to segregate them throughout investigations without jeopardising the privacy of other users of the infrastructure.

Lack of User Information:

Usually cloud service providers enforce user-friendly standards and only ask for the bare minimum of customer data. As a result, it is difficult for the investigators to find the culprit with the little information offered by the scant user data.

Data Origin:

In contrast to typical digital forensics investigations, there is less assurance regarding the origin of the data in clouds because it might originate from hundreds of different user geographic areas and it is extremely challenging to determine who or what generated and/or amended the data item there.

Lack of Transparency and Data Protection in Cloud Services:

The majority of the time, IaaS providers like Google Cloud, Amazon web services, Microsoft Azure and numerous others handle storage and data protection. However, for storage encrypted data and archiving, several services employ standard keys.

Less Control in Clouds:

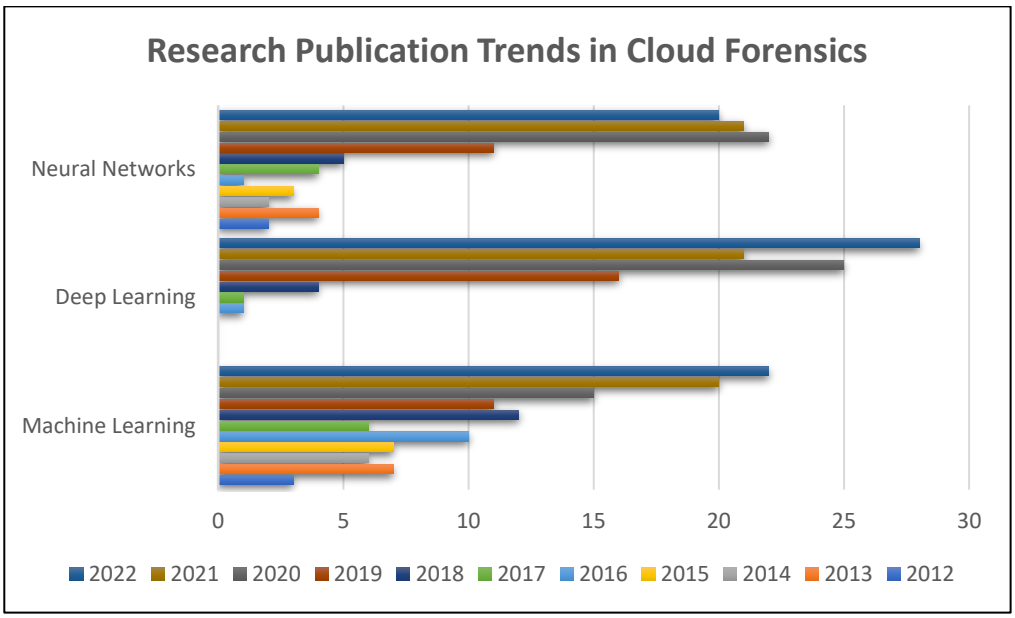
In contrast to digital forensics, where the investigator has complete access to the devices, cloud forensics bases the investigator's access level to cloud devices on the various service models.

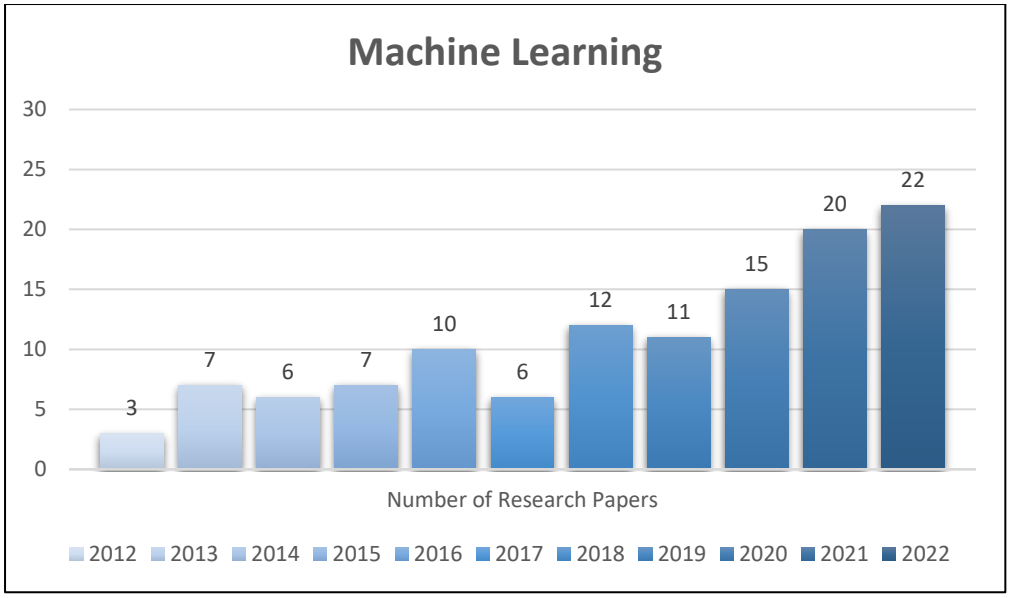
8. ML, DL, NN HELPS IN CLOUD FORENSICS:

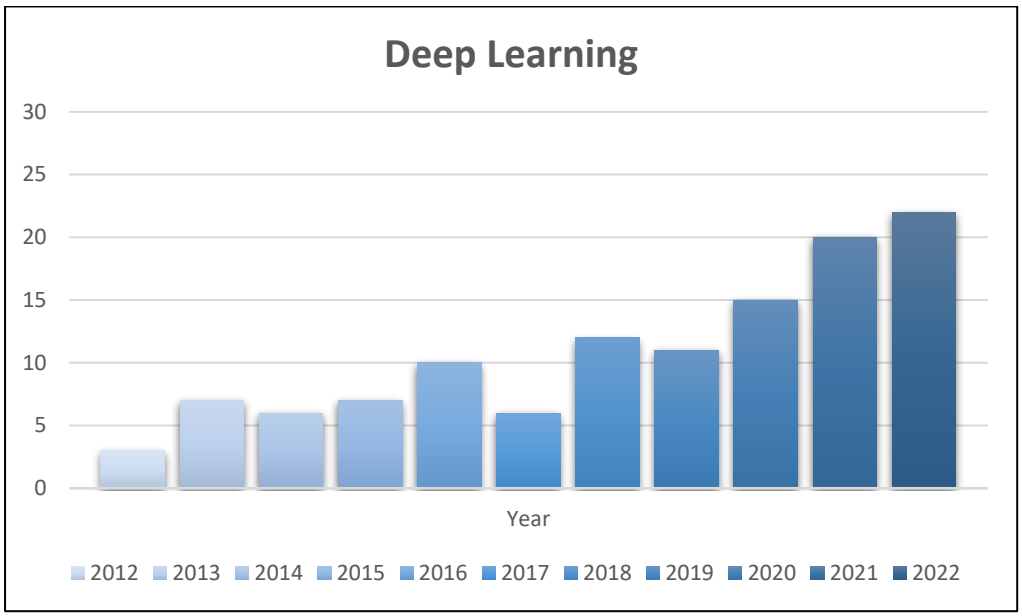
The priorities of modern man have fundamentally shifted as a result of digitization. The demand for digital forensics is growing because of the rapid increase in cybercrimes [25]. But as of right now, we must cope with big data when doing forensic analysis [26]. Large amounts of content are present in big data. Additionally, it is virtually hard to manually do error-free analysis with such a vast amount of data. AI and machine learning (ML) are the hottest technologies in use today. Everything is moving in the direction of automation. To analyze the data more effectively, DF must likewise evolve alongside the digital age [27].

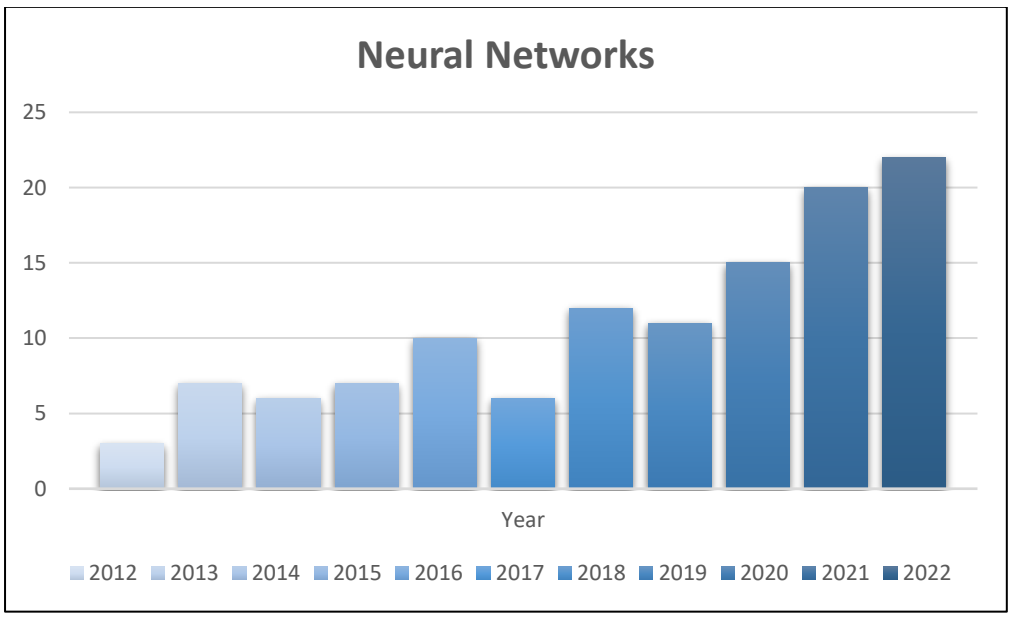
ML has gotten a lot of attention recently. It is a skill that is learned via practice and experience rather than through any sort of programming [28]. These algorithms must first be applied to a trained dataset. Online datasets are widely available for the use of ML in a variety of scientific domains. The selection of the algorithm is another crucial component of ML. Due to these decisions, ML can be divided into two categories. Unsupervised learning is the second and supervised learning is the first. We must train our well-labeled dataset for supervised learning. The training dataset for unsupervised learning must be unlabeled. Deep Learning is a sort of unsupervised learning that employs artificial neural networks at several degrees of hierarchy (ANN). The accuracy of machine learning (ML) relies upon the dataset and the algorithm. Many academics have looked into how well ML systems apply to DF. Malware analysis, network forensics, and mobile/memory forensics are the four key knowledge areas where it has primarily been used [27]. ML algorithms are typically run-on libraries and software tools. These software tools are fed datasets and the data gathered during the collecting phase of DF.

9.Research Trends in Cloud Forensics:




Machine Learning
30
25
20
15
10
5
0
3
7
6
7
10
6
12
11
15
20
22
Number of Research Papers
2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022


Deep Learning
30
25
20
15
10
5
0
Year
2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022


Neural Networks
25
20
15
10
5
0
Year
2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022

MACHINE LEARNING ALGORITHMS:

Machine learning (ML) tactics are often flexible statistical methods for inferring conclusions or categorising data. These methods are often explained by the algorithms which give specifics, although the predictions are made using the data and may produce a wider variety of predictors, also known as high-dimensional information [46]. Machine learning is to program computers for optimizing a performance criterion past experience [47]. There are three types of learning. Supervised learning describes a scenario in which the experience contains significant information that is missing in the unseen to which the learned expertise is to be applied[48]. Unsupervised learning Data used as input or for training are not labelled. By inferring pre-existing patterns or clusters from the datasets, a classifier is created. Both labelled and unlabelled data are present in the semi-supervised learning testing dataset. The classifiers undergo training to understand the patterns needed to classify, identify, and predict the data[49].

Table of most common ML algorithms [50]:

ALGORITHM	**STRENGTH**	**WEAKNESS**
KNN	efficiency, competitive classification performance, and simplicity across numerous domains.	poor run-time performances after a large training set.
DECISION TREE	Simple to grasp and might be used in conjunction with other decision-making processes	Unstable
NAÏVE BAYES	simplicity, ease of implementation, and dataset scalability.	NB models are frequently defeated by models that have been appropriately trained and optimised utilising the aforementioned approaches.
SVM	able to simulate nonlinear decision limits and resistant to overfitting	They require a lot of memory and perform poorly with bigger datasets.
RANDOM FOREST	use huge datasets effectively while maintaining accuracy	Random forests are biased in favour of qualities with more levels when it comes to variables with varied number of levels.
LINEAR REGRESSION	Simple to comprehend and explain, with the ability to be regularised to prevent	when there are non-linear connections, performs badly. not adaptable enough to recognise

	overfitting.	intricate patterns
ANN	Processing in parallel and fault tolerance	The network's lifetime is unclear, and hardware reliance
DNN	simultaneous calculations and automated deduction features	big data volume, high training costs owing to complicated data models
K-MEAN	quick, easy, and adaptable	The number of clusters must be specified; however, it is challenging to do so.
C4.5	Handle properties that are both continuous and discrete.	develops bare branches and is sensitive to noise

DATASETS USED IN DF:

The accuracy of machine learning (ML) relies mostly on dataset as well as the algorithm. The ML Algorithms must first be applied to a training dataset. Online datasets are widely available for the use of ML in a variety of scientific domains. ML algorithms are often run on library resources and software tools. These software tools get fed datasets and the data gathered during the collecting phase of DF. It has been discovered from the examined publications that the dataset's correctness is highly crucial. However, it has been observed that various logical issues, including the choice of ML method and dataset type, have troubled researchers. Making the most efficient use of datasets is said to be the key difficulty in ML-based DF. Datasets are essential for precise outcomes [27].

Table: Below is the table which shows some of the datasets that are used in various fields [27].

SOURCE	**DF TYPE**	**SAMPLES**
Virus Share	Malware Analysis	Present it has 37,309,072 samples of malware
VX Heaven	Malware Analysis	271092 samples
Comodo Cloud Security Center	Malware Analysis	37,930 samples of malware
Pascal VOC 2012	Image Forensics	20 classes, roughly 6,929 segmentations in 11, 530 pictures, and 27, 450 objects
MS-COCO	Image Forensics	2,500,000 occurrences over 330 thousand photos, 80 different item types
IMDB-WIKI	Image Forensics	523,051 instances
Karina	Video Forensics	16 videos
Image Net	Image Forensics	14,197,122 instances
YFCC100M	Video Forensics	100 million

CAIDA	Network Forensics	33 datasets
Bot-IoT	Network Forensics	9543 benign + 73360900 instances of network attacks
Real Data Corpus	Memory Forensics	6748 GB Corpus
2007 INEX Wikipedia	Files/ Memory Forensics	75047 files

Table: open-source ML tools[27]

Tool	**Description**
WEKA	an open-source tool with a wide selection of ML algorithms.
Python WEKA Wrapper	a programme that connects Python and WEKA libraries
LIBSVM	Open-source C++ library that supports SVM for classification and regression analysis.
RapidMiner	a machine learning and data mining tool
Dlib	ML toolkit in C++ that supports several algorithms

10.RELATED WORK:

Sachdeva S, Ali A in et al. [29] Designed a new hybrid model which uses a genetic algorithm for frequent pattern analysis and k-nearest algorithm is applied to keep in track the selection process of KNN, MLP. Their plan incorporates KNN and genetic algorithms to enhance the precision of classification of attacks. The accuracy of the proposed machine learning (KNN + MLP) algorithm came out to be 99.93 in compared to the existing ones. In this research paper a prototype called trust surveillance system was used on the provisioned server and a range of freeware cloud tools were examined with conservative forensic system on the client side for evidence, and it is used on existing ICMP attack, TCP Sync attack, UDP Attack, log analysis.

Saini, P. et al. [30] They used WEKA, a machine learning tool, to detect attacks in the data sources and applied different classifiers such as Naïve Bayes, Random Forest, MLP, and J48 (C4.5). The dataset used includes four types of attacks: UDP, HTTP floods, Smurf, and SIDDoS. They used a confusion matrix to evaluate the performance of the classifiers. J48, MLP, Random Forest, and Naive Bayes accuracy were calculated to be 98.64%, 98.63%, 98.10%, and 96.93%, respectively. The results show that J48 outperformed all 4 classifiers for calculating different classes, while Nave Bayes produced the worst results.

Patrascu A, Velciu M, Patriciu V in et al. [31] They presented a novel solution that allows digital forensics investigators to supervise user behaviour across a cloud platform and detect malicious actions in a reliable and secure manner. They utilised K-nearest neighbour (KNN), SVM classifier, and C4.5 decision tree. The results show that decision trees outperform SVM, KNN, and decision trees in terms of overall performance.

Toshev et al. created a method centered upon Deep Neural Networks to classify images in [32], which was then used to object detection. They developed a multi-scale inference process that a few network applications can use to quickly and affordably detect objects at high resolution. The test set from the Pascal 2007 Visual Object Challenge (VOC) is the dataset that was used. The softmax classifier calculates the detection score.

In [33] they have developed an innovative hybrid technique based on data forensics, machine learning (ML), dynamic malware analysis, and cyber threat intelligence. Big data forensics is used to anticipate IP reputation at the pre-acceptance stage and classify related zero-day attacks using behavioural analysis using Decision Tree (DT) technique.

In [34] they sought to determine whether various machine learning (ML) methods might be used to track the activations of a historical file system to find increasing evidence. The training datasets have been gathered using VMware. The gathered dataset was subsequently fed to 7 different machine learning (ML) algorithms, including Feed Forward Neural Network (FFNN), Support Vector Machine (SVM), Random Forest (RF), Classification and Regression Trees (CART), Naive Bayes (NB). Based on various evaluation metrics, the performance of these algorithms was compared. According to the experimental findings, NN and RF typically generated the best outcomes.

In [35] they presented a two-tier architecture utilizing data mining and neural networks to identify a network-based intrusion. Through the use of two classifiers, they examined network behaviour in their article that may be divided into misuse detection and anomaly detection. Utilizing hierarchical agglomerative clustering and an autonomous model on the training set, the input data was first categorized. Using KNN classification, the second phase categorized the input data as either regular traffic patterns or intrusions. For abuse detection, they employed the MLP algorithm, while for anomaly detection they used the reinforcement algorithm. Their experiments yielded a TP rate of 99% and a false positive rate of 1%.

In [36] They concentrated on the method of analysing network security threats using machine learning algorithms and proposed Cloud-based Intelligent Security Technology (CIST) for tailored security service pro-visioning (unsupervised learning). The new-Kyoto 2006+ training dataset was utilized. SVM, Decision Tree, Neural Network, and Random Forest all performed less well than Random Forest.

In [37,38] They proposed a framework for collective anomaly detection utilizing a partitioned clustering technique to find anomalies. By comparing their findings to those from other techniques using benchmark data sets, they validated their methodology. Additionally, they developed the issue of identifying DoS attacks. They experimentally analysed the 1998 DARPA, 1999 KDD Cup, and Kyoto datasets to validate their methodology and showed that it outperformed existing clustering-based anomaly detection algorithms.

In [39] This study focuses on DDoS attack detection and prevention. They used the Nave Bayes as well as Random Forest algorithms. The false percentage of pockets and the true percentage of packets were detected more efficiently by Nave Bayes than by random forest. They selected one cloud-based site to attack using the Parrot Sec Operating System. The analysed data has been trained in the widely used but powerful tool 'WEKA'.

In [40] a powerful integrated Weighted Fuzzy K-means clustering, and Auto Associative Neural Network (WFCM-AAN) malware detection method is suggested for locating the system's malware. When compared to the existing technique, the suggested system detects malware with the highest precision of 92.45%, the highest recall of 75.48%, and the highest F measure of 58.47%, yielding competent results based on validating and analysing the performance metrics using graphical results. In this study, they discussed the two problems of anomaly and normal detection. This approach will be used in future studies to detect numerous forms of attacks and boost the efficiency of malware detection.

Using VM snapshots, the authors Linda Joseph et al. [41] have suggested a method to identify malware from virtual machines. To categorize the VM snaps as attacked and non-attacked snapshots, machine learning techniques are used. Naive Bayes classifier as well as Random Forest were employed by the authors Amjad HB et al. [42] to regulate traffic on the network among cloud VMs.

In [43] A three-stage system for detecting cloud anti-forensic attacks is proposed known as the suspicious packets identification framework (SPIF). NSL-KDD is the dataset which is utilized. In this suggested approach, both signature analysis as well as anomaly detection across cloud levels are performed to classify the sort of attack which affected the packet. For performance evaluation, a variety of algorithms such as k-means, SVM, KNN, and Naive Bayes are utilized.

In [44] they proposed a generic digital forensic framework with a fusion algorithm for the cloud. The dataset that is used is NSL-KDD with ICMP Attacks, TCP Sync Attacks, and UDP Attacks. Various classifiers are utilized like MLP, Random Forest, and Naive Bayes. The total accuracy of MLP, Random Forest, and Naive Bayes were 98.6%, 98.02%,

and 96.91%, respectively. The best retrieval and precision scores were obtained by MLP, with Naive Bayes performing the worst of the bunch.

11.Comparitive analysis:

Literature	objective	type	Algorithms	Performance metrics	Dataset/ attacks used
[29]	Used to observe frequent pattern analysis in forensics	Supervised machine learning, Artificial neural network	k-nearest algorithm, KNN, MLP	99.93% accuracy	ICMP attack, TCP Sync attack, UDP Attack, log analysis
[30]	detect attacks in the data sources	Machine learning	Naïve Bayes, Random Forest, MLP, and J48 (C4.5)	J48 outperformed all 4 classifiers, while Nave Bayes produced the worst results.	UDP, HTTP floods, Smurf, and DDoS
[31]	supervise user behaviour across a cloud platform and detect malicious actions	Machine learning	utilised K-nearest neighbour (KNN), SVM classifier, and C4.5 decision tree	decision trees outperform SVM, KNN, and decision trees in terms of overall performance.	----------
[32]	Classify images/object detection	Deep Neural Networks	softmax classifier	----------	Pascal 2007 Visual Object Challenge (VOC)
[33]	classify related zero-day attacks using behavioural analysis	Machine learning	Decision Tree (DT) technique	---------	----------
[34]	To track the activations of a historical file system to find increasing evidence	Machine learning	Forward Neural Network (FFNN), Support Vector Machine (SVM), Random Forest (RF), Classification and Regression Trees (CART), Naive Bayes (NB)	NN and RF typically generated the best outcomes	VMware
[35]	to identify a network-based intrusion	neural networks	KNN classification, MLP algorithm, reinforcement	a TP rate of 99% and a false positive rate of 1%.	

			algorithm		
[36]	proposed Cloud-based Intelligent Security Technology (CIST) for tailored security service pro-visioning	Unsupervised and supervised machine learning	SVM, Decision Tree, Neural Network, and Random Forest	Random Forest outperformed all others.	new-Kyoto 2006+ training dataset
[39]	DDoS attack detection and prevention	Machine learning	Naive Bayes, Random Forest	Naïve bayes was more efficient than RF.	WEKA tool was used.
[40]	Malware detection method	ML, ANN	Weighted Fuzzy K-means clustering, and Auto Associative Neural Network (WFCM-AAN)	precision of 92.45%, the highest recall of 75.48%, and the highest F measure of 58.47%	
[43]	A three-stage system for detecting cloud anti-forensic attacks	ML algorithms	k-means, SVM, KNN, and Naive Bayes are utilized.	Proposed RBNN+ k-means + Correlation shows high accuracy, while KNN+ k-means+ correlation shows least accuracy.	NSL-KDD is the dataset that is used.

12.CONCLUSION AND FUTURE WORK:

An integrated review for machine learning-based cloud forensics was built in this review study. Cloud forensics uses a variety of machine learning techniques, including deep learning and artificial neural networks. Different datasets that are used are explained, and as data sizes grow, it is becoming increasingly challenging to conduct forensics on them. In this context, ML technology has demonstrated excellent outcomes. We have highlighted several research articles that support ML-based cloud forensics. The most popular algorithm among researchers, deep learning is demonstrated in reviewed literature to play a significant role in cloud forensics. The most effective use of datasets is the key difficulty in the entire ML-based cloud forensics process. Sets of data are crucial for precise outcomes. In general, future improvements to the cloud forensics process will probably concentrate on enhancing the effectiveness of the investigation process and more effectively integrating new technologies and approaches into the models.

REFERENCES:

[1] Naaz, S., & Ahmad, F. (2016). Comparitive Study of Cloud Forensics Tools. *Communications on Applied Electronics*, *5*(3), 24–30. https://doi.org/10.5120/cae2016652258

[2] Purnaye, P., & Kulkarni, V. (2022). A Comprehensive Study of Cloud Forensics. *Archives of Computational Methods in Engineering*, *29*(1), 33–46. https://doi.org/10.1007/s11831-021-09575-w

[3] Alenezi, A., Zulkipli, N. H. N., Atlam, H. F., Walters, R. J., & Wills, G. B. (2017). The impact of cloud forensic readiness on security. *CLOSER 2017 - Proceedings of the 7th International Conference on Cloud Computing and Services Science*, 511–517. https://doi.org/10.5220/0006332705390545

[4] Zawoad, S. (n.d.). *SECURING THE CLOUD Digital Forensics in the Cloud*.

[5] Le-Khac, N.-A., & Scanlon, M. (2017). *Evaluation of Digital Forensic Process Models with Respect to Digital Forensics as a Service*. https://www.researchgate.net/publication/318981575

[6] Shahzad, F., Javed, A.R., Jalil, Z., Iqbal, F. (2022). Cyber Forensics with Machine Learning. In: Phung, D., Webb, G.I., Sammut, C. (eds) Encyclopedia of Machine Learning and Data Science. Springer, New York, NY. https://doi.org/10.1007/978-1-4899-7502-7_987-1

[7] Aditya, K., Grzonkowski, S., & Lekhac, N. (2018). Enabling Trust in Deep Learning Models: A Digital Forensics Case Study. *Proceedings - 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications and 12th IEEE International Conference on Big Data Science and Engineering, Trustcom/BigDataSE 2018*, 1250–1255. https://doi.org/10.1109/TrustCom/BigDataSE.2018.00172

[8] Sri Shakthi Institute of Engineering and Technology, Institute of Electrical and Electronics Engineers. Madras Section, All-India Council for Technical Education, & Institute of Electrical and Electronics Engineers. (n.d.). *2020 International Conference on Computer Communication and Informatics : January 22-24, 2020, Coimbatore, India*.

[9] Shah, J. J., & Malik, L. G. (2013). Cloud forensics: Issues and challenges. *International Conference on Emerging Trends in Engineering and Technology, ICETET*, 138–139. https://doi.org/10.1109/ICETET.2013.44

[10] Aldawibi, O. O., Sharf, M. A., & Obaid, M. M. (2022). Cloud Computing Privacy: Concept , Issues And Solutions. *2022 IEEE Symposium on Industrial Electronics & Applications (ISIEA)*, 1–4. https://doi.org/10.1109/ISIEA54517.2022.9873688

[11] Zawoad, S. (n.d.). *SECURING THE CLOUD Digital Forensics in the Cloud*.

[12] Guo, H., Jin, B., & Shang, T. (2012). Forensic investigations in Cloud environments. *Proceedings - 2012 International Conference on Computer Science and Information Processing, CSIP 2012*, 248–251. https://doi.org/10.1109/CSIP.2012.6308841

[13] Mondal, A., Paul, S., Goswami, R. T., & Nath, S. (2020, January). Cloud computing security issues & challenges: A review. In 2020 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-5). IEEE.

[14] Agarwal, A., & Agarwal, A. (2011). The security risks associated with cloud computing. International Journal of Computer Applications in Engineering Sciences, 1(Special Issue on), 257-259.

[15] D. Svantesson, R. Clarke, "Privacy and Consumer Risks in Cloud Computing," Computer Law & Security Review, pp. 391-397, 2010.

[16] Vladimir Dobrosavljević, Mladen Veinović, Ivan Barać, "Standard Implementation in Cloud Forensics"Singidunum University, Danijelova 32, Belgrade, Serbia 2015.

[17] Dorey P.G., Leite A,"Commentary: Cloud computing –A security problem or solution?"Information Security Technical Report, 16 (3–4), pp. 89–96, Elsevier (2011)

[18] S. D. Wolthusen, "Overcast: Forensic Discovery in Cloud Environments," Fifth International Conference on IT Security Incident Management and ITForensics,pp. 3-9, 2009.

[19] National Institute of Standards and Technology, "Guide to Interating Forensic Techniques into Incident Response", 2006

[20] Cloud Security Alliance, "Security Guidance for Critical Areas of Focus in Cloud Computing", December 2009, [1] http://en.wikipedia.org/wikilElectronic _identity_card, 2010.

[21] Scott Zimmerman and Dominick Glavach, "Cyber Forensics in the Cloud", IAnewsletter Vol 14 No 1, 2011.

[22] Sen, J., 2013. Security and Privacy Issues in Cloud Computing. Architectures and Protocols for Secure Information Technology, (iv), p.42.

[23] Chou, T., 2013. Security Threats on Cloud Computing Vulnerabilities. International Journal of Computer Science and Information Technology, 5(3), pp.79–88.

[24] Chouhan, P. & Singh, R., 2016. Security Attacks on Cloud Computing With Possible Solution. International Journal of Advanced Research in Computer Science and Software Engineering, 6(1), pp.92–96.

[25] S. Iqbal and S. A. Alharbi, "Advancing Automation in Digital Forensic Investigations Using Machine Learning Forensics". Digital Forensic Science, 2020

[26] G. S. Chhabra, V. P. Singh and M. Singh, "Cyber forensics framework for big data analytics in IoT environment using machine learning",Multimedia Tools Appl., pp. 1-20, Jul. 2018,

[27] Qadir, S., & Noor, B. (2021, May 20). Applications of Machine Learning in Digital Forensics. *2021 International Conference on Digital Futures and Transformative Technologies, ICoDT2 2021*. https://doi.org/10.1109/ICoDT252288.2021.9441543

[28] S. Iqbal and S. A. Alharbi, "Advancing Automation in Digital Forensic Investigations Using Machine Learning Forensics". Digital Forensic Science, 2020

[29] Sachdeva, S., & Ali, A. (2021). A Hybrid approach using digital forensics for attack detection in a cloud network environment. In *International Journal of Future Generation Communication and Networking* (Vol. 14, Issue 1).

[30] Institute of Electrical and Electronics Engineers, Institute of Electrical and Electronics Engineers. Delhi Section, & INDIAcom (Conference) (14th: 2020 : New Delhi, I. (n.d.). *2020 7th International Conference on Computing for Sustainable Global Development (INDIACom).*

[31] Patrascu, A., Velciu, M. A., & Patriciu, V. V. (2015). Cloud computing digital forensics framework for automated anomalies detection. *SACI 2015 - 10th Jubilee IEEE International Symposium on Applied Computational Intelligence and Informatics, Proceedings*, 505–510. https://doi.org/10.1109/SACI.2015.7208257

[32] C. Szegedy, A. Toshev and D. Erhan, "Deep neural networks for object detection",Proc. Adv. Neural Inf. Process. Syst., 2013.

[33] N. Usman, S. Usman, F. Khan, M.A. Jan, A. Sajid, M. Alazab, P. Watters, "Intelligent Dynamic Malware Detection using Machine Learning in IP Reputation for Forensics Data Analytics", Future Generation Computer Systems, Volume 118, 2021, Pp 124-141

[34] R. M. Mohammad and M. Alqahtani, "A comparison of machine learning techniques for file system forensics analysis",Journal of Information Security and Applications, vol. 46, no. 1, pp. 53-61, 2019.

[35] Divyatmika, Sreekesh, Manasa: Two-tier network anomaly detection model: a machine learning approach. In: International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 42–47 (2016)

[36] Kim, H., Kim, J., Kim, Y., Kim, I., & Kim, K. J. (2019). Design of network threat detection and classification based on machine learning on cloud computing. *Cluster Computing*, *22*, 2341–2350. https://doi.org/10.1007/s10586-018-1841-8

[37] Ahmed, M., Mahmood, A.N., Hu, J.: A survey of network anomaly detection techniques. J. Netw. Comput. Appl. 60, 19–31 (2016)

[38] Ahmed, M., Mahmood, A.N.: Novel Approach for Network Traffic Pattern Analysis using Clustering based Collective Anomaly Detection, pp. 111–130. Springer, Berlin (2015). https://doi.org/10.1007/s40745-015-0035-y

[39] Amjad, A., Alyas, T., Farooq, U., & Tariq, M. A. (2019). Detection and Mitigation of DDoS Attack in Cloud Computing Using Machine Learning Algorithm. *EAI Endorsed Transactions on Scalable Information Systems*, *6*(23), 1–8. https://doi.org/10.4108/eai.29-7-2019.159834

[40] Yadav, R. M. (2019). Effective analysis of malware detection in cloud computing. *Computers and Security*, *83*, 14–21. https://doi.org/10.1016/j.cose.2018.12.005

[41] Joseph L, Mukesh R (Sep 2018) Detection of malware attacks on virtual Machines for a Self-Heal Approach in cloud computing using VM snapshots. J Commun Software Syst 14(3):249–257

[42] Amjad HB, Sabyasachi P, Debasish J (2013) Machine learning approach for intrusion detection on cloud virtual machines. Int JAppl Innov Eng Manag 2(6):57–66

[43] Radha Rani, D., & Geethakumari, G. (2083). *A framework for the identification of suspicious packets to detect anti-forensic attacks in the cloud environment*. https://doi.org/10.1007/s12083-020-00975-6/Published

[44] Sachdeva, S., & Ali, A. (2022). Machine learning with digital forensics for attack classification in cloud network environment. *International Journal of System Assurance Engineering and Management*, *13*, 156–165. https://doi.org/10.1007/s13198-021-01323-4

[45] B. Patel, Prof. H., & Kansara, Prof. N. (2021). Cloud Computing Deployment Models: A Comparative Study. *International Journal of Innovative Research in Computer Science & Technology*, *9*(2), 45–50. https://doi.org/10.21276/ijircst.2021.9.2.8

[46] D. Buskirk, A. Kirchner, A. Eck, and C. S. Signorino, "An Introduction to Machine Learning Methods for Survey Researchers what are machine learning methods ?," pp. 0–3, 2018, doi: 10.29115/SP-2018-0004.

[47] E. Alpaydın, Introduction to Machine Learning Second Edition. .

[48] .Ben-david, Understanding Machine Learning : From Theory to Algorithms. 2014.

[49] Das and R. N. Behera, "A Survey on Machine Learning : Concept ," pp. 1301–1309, 2017, doi: 10.15680/IJIRCCE.2017.

[50] Makkawi, A. M., & Yousif, A. (2021, February 26). Machine Learning for Cloud DDoS Attack Detection: A Systematic Review. *Proceedings of: 2020 International Conference on Computer, Control, Electrical, and Electronics Engineering, ICCCEEE 2020*. https://doi.org/10.1109/ICCCEEE49695.2021.9429678