**APPLICATIONS OF MACHINE LEARNING IN CLOUD FORENSICS**

Malluri Reddy Theertani  
Student, Computer Science and Engineering , Vellore Institute of Technology , Vellore.Vellore , India  
[theerthanimalluri@gmail.com](mailto:theerthanimalluri@gmail.com)

Kakelli Anil Kumar  
Associate Professor Sr. School of Computer Science EngineeringVellore Institute of Technology, VelloreVellore, India  
[anilkumar.k@vit.ac.in](mailto:anilkumar.k@vit.ac.in)

Neha Valeti   
Student, Computer Science and Engineering, Vellore Institue of Tehnology,VelloreVellore, India  
[nehavaleti01@gmail.com](mailto:nehavaleti01@gmail.com)Darshini Solanki   
Student, Computer Science and Engineering, Vellore Institute of Technology,VelloreVellore, India  
[darhsinisingh@gmail.com](mailto:darhsinisingh@gmail.com)

*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. It also provides a detail information on cloud forensics as well as the research trends that have been going on in this field over the past few years. A detailed comparative analysis of different approaches of machine learning on development of cloud forensics has been presented in this research study.

Keywords—Cloud Computing, Cloud Forensics, Machine Learning.

# 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. Cloud computing, which offers quicker and more flexible resources, is simply the transmission of computing services through the internet, i.e., the cloud. 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].The term "cloud forensics" describes the application of digital forensics within cloud computing. 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. Cloud forensics has advanced due to advancements in Big Data Analytics, Artificial Intelligence, and machine learning. 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 analysing 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 Machine Learning approaches to solve this problem.

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 article will be on understanding how various machine learning models are employed in forensic investigations in the cloud. The rest of the paper is structured as follows. Section II discusses the existing machine learning models in cloud forensics, Section III gives an overview of cloud computing, Section IV discusses cloud forensics followed with Section V with tradition process of forensic investigation and Section VI with challenges in cloud forensics. Section VII explains the cases solved using digital forensics. Section VIII explains the role of machine learning in cloud forensics followed with machine learning algorithms in Section IX. Section X presents the different datasets used in digital forensics. Section XII shows the comparative analysis of different machine learning models that were used in development of cloud forensics and Section XII concludes with the findings of this review paper.

# RELATED WORK

S. Sachdeva, A. Ali, and others in [29] A new hybrid model was developed, using a genetic algorithm for the study of frequent patterns and the k-nearest neighbor method to maintain track of the KNN and MLP selection process. KNN and genetic algorithms are used in their strategy to improve the accuracy of attack classification. In comparison to the current ones, the accuracy of the suggested machine learning (KNN + MLP) algorithm was 99.93%. In this study, a prototype known as the trust surveillance system was used on the provided server. A variety of freeware cloud tools were also evaluated, along with a conservative forensic system on the client side, for possible evidence.

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. To assess how well the classifiers were performing, they employed a confusion matrix. The computed accuracy for J48, MLP, Random Forest, and Naive Bayes were 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, and others in [31] They presented a novel solution that allows digital forensics investigators to supervise user behavior across a cloud platform and detect malicious actions in a reliable and secure manner. They utilized K-nearest neighbor (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 detect objects quickly and affordably 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. They have created a ground-breaking hybrid method based on data forensics, machine learning (ML), dynamic malware analysis, and cyber threat intelligence in [33]. Big data forensics is used to categorize linked zero-day attacks using behavioral analysis utilizing Decision Tree (DT) technique and estimate IP reputation at the pre-acceptance stage.

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. Following data collection, the dataset was put into 7 different machine learning (ML) algorithms, including Feed Forward Neural Network (FFNN), Support Vector Machine (SVM), Random Forest (RF), Classification and Regression Trees (CART), and 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 behavior 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. The second step classified the input data using KNN as either regular traffic patterns or intrusions. They used the MLP algorithm for misuse detection and the reinforcement algorithm for anomaly identification. In their tests, the TP rate was 99% while the false positive rate was 1%.

In [36] They concentrated on the method of analyzing 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.

Using a partitioned clustering technique, they provided a framework for collective anomaly detection in [37,38]. They validated their methods by comparing their results to those from other methodologies using benchmark data sets. They also ran into the issue of recognizing DoS assaults. They used experimental analysis to demonstrate that their methodology beat currently used clustering-based anomaly detection algorithms on the 1998 DARPA, 1999 KDD Cup, and Kyoto datasets.

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 analyzed data has been trained in the widely used but powerful tool 'WEKA’.

For locating the system's infection, [40] suggests using a potent combined Weighted Fuzzy K-means clustering and Auto Associative Neural Network (WFCM-AAN) malware detection approach. The proposed approach, which is based on evaluating and analyzing the performance metrics using graphical results, identifies malware with the highest precision of 92.45%, the highest recall of 75.48%, and the highest F measure of 58.47% when compared to the existing technique. They covered the two issues of anomaly and

regular detection in this work. 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.

# III.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. The ability to retain user data on remote machines that are readily accessible over the internet is an advantage of cloud computing. It is also explained as delivering a computer model as a service online. Cloud computing allows users and organizations to use hardware and software which are maintained by cloud service providers (CSP). Users can use any device to retrieve any information from any remote location, if they are connected to internet utilizing the cloud computing [10]. Cloud computing is a concept that allows accessible, on-demand access to a collective pool of computer resources, comprising servers, services, storage, applications, etc.., according to the National Institute of Standards and Technology(NIST).Cloud computing has three delivery types that are Software as a Service(Saas),Platform as a Service(Paas),Infrastructure as a service(Iaas). It also has three deployment models that are Private, Public and Community cloud.

IV. CLOUD FORENSICS

Cloud computing forensics is described as “the use of scientific principles, technological procedure as well as derived and proven methods to recreate previous cloud computing incidents through recognition, gathering, retention, investigation, explanation, and disclosure of digital evidence” [2] by the NIST. Due to the rapid advancement of digital technology, it requires a lot of computer power nowadays to understand the data that these devices produce. The idea of a "Forensic Cloud" has been taken into consideration to allow an investigator to fully concentrate on the investigative processes [15]. "A method of cloud forensics is described as an analysis of a cybercrime that requires proof or evidence gathered from any of the cloud-based platforms or cloud-bases services," according to the definition of the term [2]. Cloud forensics may also be viewed as a part of system legal sciences since systems crime scene investigation regulates scientific investigations in any form of system, whether it be private or public. Due to the lack of approaches specially created for cloud computing environments, cloud computing relies on large and ubiquitous system access and embraces the fundamental standards defined by the system measurement methodology [1]. Over the past ten years, there has been a significant increase in the number of crimes involving computers and the internet, which has resulted in a corresponding increase in businesses that aim to assist law enforcement by using digital evidence to pinpoint the perpetrators, methods, victims, and timing of cyber-crimes. As a result, digital forensics advanced, ensuring the accuracy of data presented in court as evidence of criminal activity. Yet, forensic data is starting to grow swiftly, making it more challenging to analyze them quickly as storage capacity outpaces network performance and latency developments [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 . And due to its convenience in data management and storage, cloud computing has a variety of impact on forensics Due to the enormous volume of logs that are produced, and the time and resources needed to filter and evaluate them all, processing logs even becomes challenging. 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 of 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 time required to obtain the entire cloud is limitless. In this context, "forensics" refers to the procedure of reviewing cloud data to find evidence against the invasive party. A wide variety of technologies and software will be used in the forensics effort, some of which can help us analyze cloud data [17]. Lack of physical access to computers presents new and disruptive issues for forensics experts in cloud forensics. Traditional methods of evidence gathering, recovery no longer work due to the decentralized nature of data handling on the cloud. The 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].

# V. addressing the Differences: Traditional Forensics vs. Cloud Forensics

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**: Locating, categorizing, recording, and collecting data from any potentially relevant information sources would be the first phase in the procedure, with regulations and processes that ensure the integrity of the data having implemented at all periods.

**Examination**: Examinations necessitate forensically processing vast quantities of acquired data using a combination of automated and manual procedures to analyze and extract data of special significance while maintaining the data security.

**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 analyzing.

**Reporting**: Reporting the analysis' findings involves summarizing 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].

Application Forensics, yet another field of forensics that examines applications and associated data to find any potential security flaws or concerns. Lack of application access, poor logging, data encryption, dynamic data, and anti-forensic methods are a few of the frequent issues encountered in application forensics.

The development of cloud computing has altered the forensic process's workflow today in several ways, including storage, because the data is stored on remote servers rather than the local hard drives. Which means the forensic investigators may need to access the data stored on the cloud rather than the physical device. And the data is encrypted, which makes it difficult for forensics investigators to access data without decryption keys [9]. 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 recognized 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. As a result, cloud computing has had a significant impact on forensics in terms of data collecting, data storage, data privacy, data encryption, and jurisdictional issues [9].

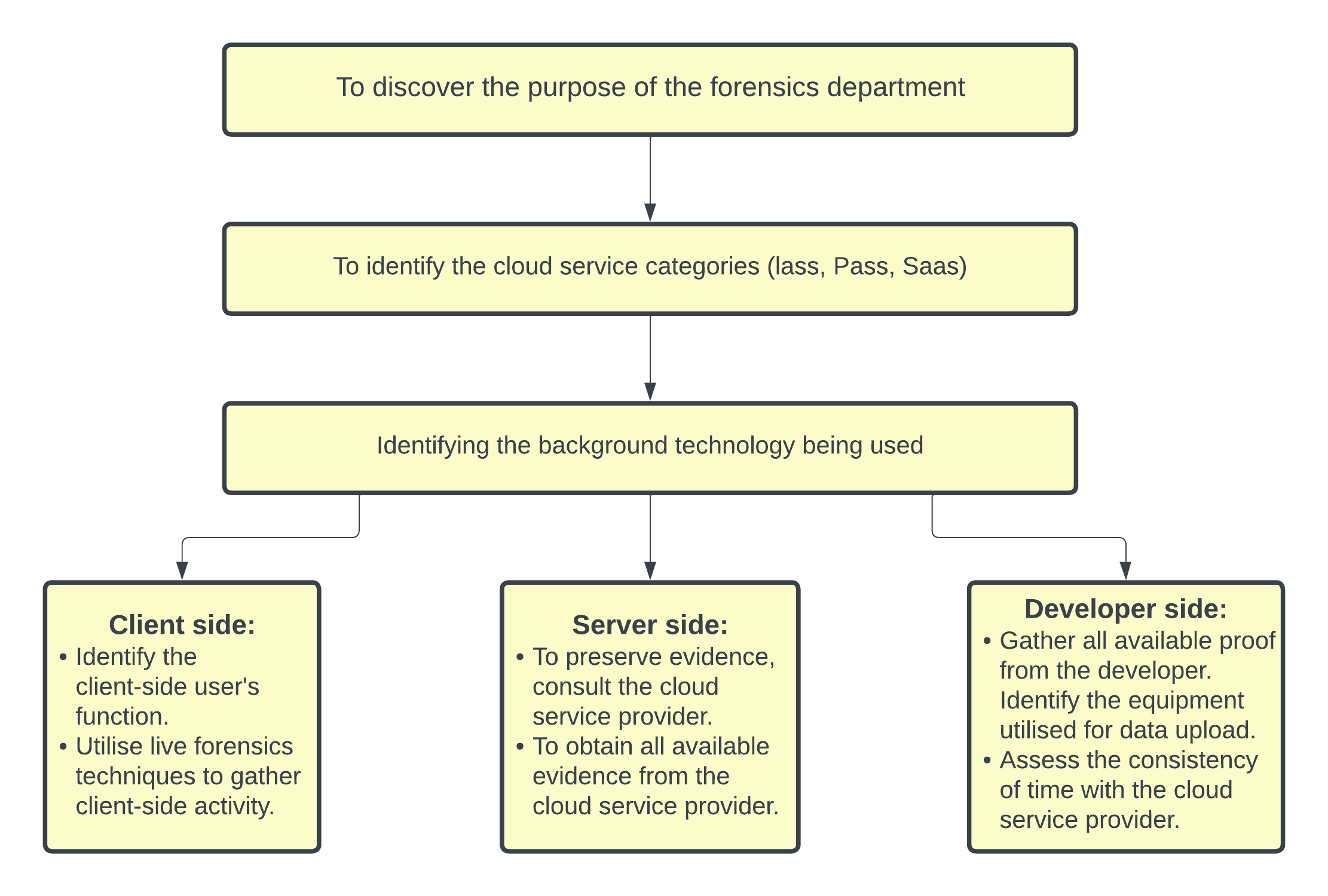


Fig 1. Forensic Process

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.)

# VI . 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.

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. 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 centers 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.

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.

It is sometimes necessary for the detectives to recreate the scene of the crime in order to look into the harmful behavior. In conventional computer forensics, the investigators can quickly determine the number of devices utilized 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. 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 jeopardizing the privacy of other users of the infrastructure.

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.

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.

Storage and data protection are quite often handled with by IaaS providers like Google Cloud, Amazon Web Services, Microsoft Azure, among others. However, for storage encrypted data and archiving, several services employ standard keys. So in order to retreive data from remote locations one can use cloud backup, remote syncing, network attached storage. And to address the complications of data acquisition and storage we can follow this model suggested in [51] which will store volatile data of each tenant in a shared persistant storage which is solely dedicated to volatile data storage and retrieval even from remote systems.

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.

The trust issues, confidentiality issues, authenticity issues, encryption issues, key management issues, data splitting issues, and multitenancy issues are often the biggest obstacles in cloud forensics [13]. and each level of it comes with its own set of difficulties for cloud forensic investigators. For example, during the identification stage, data retrieval might be difficult due to data encryption, and because the information is distributed, this creates synchronisation issues. It may also be difficult to maintain the integrity of information during the preservation stage, and problems with data availability, trust, and other factors may arise during the gathering stage. Moreover, coordinated and encrypted data are used during the examination and analysis stage[3].

# VII. Case Studies

This section mainly deals with the cases that are solved using digital forensics.

[61] This study focuses on forensic analysis of hubiC, a well-known cloud platform that utilizes Microsoft Windows 8.1, in a case that involves cloud storage. The research in this paper focuses on recovering data that the hubiC cloud left on Windows 8.1. The authors concentrated on examining the data that remained as a trace after using the hubiC cloud storage service. To get the data remnants, the authors ran numerous tests on a variety of Virtual Machines. They have independently generated virtual machines (VMs) for each of the following operations: installing, uploading, downloading, uninstalling, downloading, and deleting desktops. This is done to find and examine any signs of data traces that the hubiC application may have left behind. Once the tests have been completed, the results of each VM are examined. After doing their analysis, they concluded that file fragments were found while using the hubiC web browser using Chrome Cache Viewer. The username and password for the current user were also recorded in hubiC's process memory. It is also mentioned that after deletion, the deleted files were also located. Also, it was discovered that, given TLSVI is being used for security, there was communication between the hubiC client and server during which the IP addresses of hubiC servers were transmitted. These are the few risks identified in the hubiC system's forensic examination during installation and uninstallation.

This [58] paper also involves forensic examination of Google Drive. On a PC hard disc and an Apple iphone3G, the traces that are most likely to be left behind after using this cloud storage were discovered. The examiner conducted the investigation in accordance with the steps of identification, preservation, analysis, and reporting. They initially began their study by determining the user information and cloud service. Additionally, it is discovered that the trace can be used to locate Google Drive's account. It is also possible to access a Google Drive account without using a login or password. These kinds of records might likewise be found on an Apple iPhone with Google Drive installed.

A case study on the forensic examination of WhatsApp on Android mobile phones is [59]. With the use of an internet connection, WhatsApp is an application used for instant messaging. This essay primarily focuses on gathering data from WhatsApp and conducting research using that data. Investigation reveals that WhatsApp allows users to view both recent and deleted communications. However, recovering the destroyed data or artefacts is outside the scope of this work.

There is a case study [60] that discusses the forensic analysis of communication records obtained through messaging and email apps found in desktop, web-based, and mobile applications. The examiner found that communication records can be shown in a common format even when data is obtained from a particular application. The study of digital artefacts across numerous applications is then done using this format. The authors examined the communication data that were stored in physical memory using RAMAS, a forensic tool.

# VIII. Role of Machine Learning In Cloud Forensics

The priorities of modern man have fundamentally shifted because of digitization. Due to the sharp rise in cybercrimes, there is an increasing demand for digital forensics [25]. However, as of now, forensic analysis must cope with massive data [26]. Big data includes a huge amount of content. Additionally, with such a large amount of data, it's nearly impossible to manually do error-free analysis. Today's hottest technologies are artificial intelligence (AI) and machine learning (ML). Automation has been the way where everything is going. DF must thus advance with the digital age to analyze the data more effectively [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-labelled 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.

Fig 2. Papers published on cloud forensics with machine learning

The above figure shows the number of research papers published from 2012 – 2022 on Cloud Forensics with the application of Machine Learning.

# IX. Machine Learning Algorithms

Machine learning (ML) tactics are often flexible statistical methods for inferring conclusions or categorizing 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]. Computers are programmed to use machine learning to optimize performance criteria based on prior knowledge [47]. There are three different ways to learn. A situation where the gained knowledge is to be deployed to the unseen necessitates supervised learning since such experience contains sensitive details that is missing [48]. Unsupervised education There are no labels on the data used as input or for training. By inferring pre-existing patterns or clusters from the datasets, a classifier is created. Both labelled and unlabeled 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 1[49]

| S. No | Most Common Machine Learning Algorithms | | |
| --- | --- | --- | --- |
| Algorithm | Strength | Weakness |
| 1 | KNN | efficiency, competitive classification performance, and simplicity across numerous domains | poor run-time performances after a large training set. |
| 2 | DECISION TREE | Simple to grasp and might be used in conjunction with other decision-making processes. | Unstable |
| 3 | SVM | able to simulate nonlinear decision limits and resistant to overfitting. | They require a lot of memory and perform poorly with bigger datasets. |
| 4 | RANDOM FOREST | use huge datasets effectively while maintaining accuracy. | Random forests are biased in favor of qualities with more levels when it comes to variables with no varied number of levels. |
| 5 | LINEAR REGRESSION | Simple to comprehend and explain, with the ability to be regularized to prevent overfitting. | when there are non-linear connections, performs badly. not adaptable enough to recognized intricate patterns |
| 6 | ANN | Processing in parallel and fault tolerance | The network's lifetime is unclear, and hardware reliance |
| 7 | DNN | simultaneous calculations and automated deduction features | big data volume, high training costs owing to complicated data models |
| 8 | K-MEAN | quick, easy, and adaptable | The number of clusters must be specified; however, it is challenging to do so. |
| 9 | C4.5 | Handle properties that are both continuous and discrete | develops bare branches and is sensitive to noise |

# X. Datasets Used In Digital Forensics

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 2.[27]

| S. No | Datasets Used in Various Fields | | |
| --- | --- | --- | --- |
| Source | DF type | Samples |
| 1 | Virus Share | Malware Analysis | Present it has 37,309,072 samples of malware |
| 2 | VX Heaven | Malware Analysis | 271092 samples |
| 3 | Comodo Cloud Security Center | Malware Analysis | 37,930 samples of malware |
| 4 | Pascal VOC 2012 | Image Forensics | 20 classes, roughly 6,929 segmentations in 11, 530 pictures, and 27, 450 objects |
| 5 | MS-COCO | Image Forensics | 2,500,000 occurrences over 330 thousand photos, 80 different item types |
| 6 | IMDB-WIKI | Image Forensics | 523,051 instances |
| 7 | Karina | Video Forensics | 16 videos |
| 8 | Image Net | Image Forensics | 14,197,122 instances |
| 9 | YFCC100M | Video Forensics | 100 million |
| 10 | CAIDA | Network Forensics | 33 datasets |
| 11 | Bot-IoT | Network Forensics | 9543 benign + 73360900 instances of network attacks |
| 12 | Real Data Corpus | Memory Forensics | 6748 GB Corpus |
| 13 | 2007 INEX Wikipedia | Files/ Memory Forensics | 75047 files |

TABLE 3. Open-Source ML Tools [27]

| ***S. No*** | Tool | Description |
| --- | --- | --- |
| 1 | WEKA | an open-source tool with a wide selection of ML algorithms. |
| 2 | Python WEKA Wrapper | a program that connects Python and WEKA libraries |
| 3 | RapidMiner | a machine learning and data mining tool |
| 4 | LIBSVM | Open-source C++ library that supports SVM for classification and regression analysis. |
| 5 | Dlib | ML toolkit in C++ that supports several algorithms |

# XI. Comparative Analysis

TABLE 4.

| ***Literature*** | Objective | Type | Algorithms | Performance metrics |
| --- | --- | --- | --- | --- |
| [29] | Used to observe frequent pattern analysis in forensics | Supervised machine learning, Artificial neural network | k-nearest algorithm,  KNN, MLP | 99.93% accuracy |
| [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. |
| [31] | supervise user behaviour across a cloud platform and detect malicious actions | Machine learning | utilized K-nearest neighbor (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 | --------- |
| [33] | classify related zero-day attacks using behavioural analysis related | 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 |
| [35] | to identify a network-based intrusion | neural networks | KNN classification, MLP algorithm, reinforcement algorithm | a TP rate of  99% and a false positive rate of 1%. |
| [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. |
| [39] | DDoS attack detection and prevention | Machine learning | Naive Bayes, Random Forest | Naïve bayes was more efficient than RF. |
| [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. |

# XII . Conclusion And Future Work

An integrated review for machine learning-based cloud forensics was built in this review study. To summarize, MLP produces the most accurate results when combined with other algorithms such as the k-nearest method. Several hybrid models are built with MLP, which has an accuracy of 99.93%[29] and a fusion method with MLP that has an accuracy of 98.6%[30]. Random forest is the next best among all the other machine learning methods, including forward neural network (FFNN), support vector machine (SVM), Random forest (RF), classification and regression trees (CART), and naive bayes (NB). Yet, when it comes to identifying DDoS assaults, naive bayes surpasses random forest. As a result, where hybrid cloud forensic frameworks are employed, MLP comes into play. 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 emphasized a number of academic papers that back ML-based cloud forensics. We have also presented the numbers of papers that were published in the recent years on cloud forensics through graphs. In all the papers, authors have used more than one type of machine learning methods, they tested their efficiency and they found that the most popular algorithm among researchers, deep learning is considered to play a significant role in cloud forensics. Datasets play a crucial role for generating precise outcomes. An important concern in the overall ML-based cloud forensic process is how to use a dataset as efficiently as possible. In general, future improvements to the cloud forensics process will mostly concentrate on enhancing the effectiveness of the investigation process and more effectively integrating new technologies and approaches into the models.

##### References.

1. Naaz, Sana, and Faisal Ahmad. "Comparative Study of Cloud Forensics Tools." Communications on Applied Electronics 5, no. 3 (2016): 24-30.
2. Purnaye, Prashant, and Vaibhav Kulkarni. "A Comprehensive Study of Cloud Forensics." Archives of Computational Methods in Engineering 29, no. 1 (2022): 33-46.
3. Alenezi, Abdulrahman, Nor H. N. Zulkipli, Hassan F. Atlam, Richard J. Walters, and Gary B. Wills. "The Impact of Cloud Forensic Readiness on Security." In CLOSER 2017 - Proceedings of the 7th International Conference on Cloud Computing and Services Science, 511-517, 2017.
4. Zawoad, Shahriar. "SECURING THE CLOUD Digital Forensics in the Cloud." Accessed March 21, 2023.
5. Le-Khac, Nhien-An and Mark Scanlon. "Evaluation of Digital Forensic Process Models with Respect to Digital Forensics as a Service." Accessed March 21, 2023.
6. Shahzad, Farrukh, Abdul Rauf Javed, Zafar Jalil, and Faisal Iqbal. "Cyber Forensics with Machine Learning." In Encyclopedia of Machine Learning and Data Science, edited by Dao Hoang Thang Phung, Geoffrey I. Webb, and Chuanren Liu Sammut. New York, NY: Springer, 2022.
7. Aditya, K., Grzonkowski, S., & Lekhac, N. "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/BigDataSE2018, (2018): 1250-1255.
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. "2020 International Conference on Computer Communication and Informatics: January 22-24, 2020, Coimbatore, India."
9. Shah, J. J., & Malik, L. G. "Cloud forensics: Issues and challenges." International Conference on Emerging Trends in Engineering and Technology, ICETET, (2013): 138-139.
10. Aldawibi, O. O., Sharf, M. A., & Obaid, M. M. "Cloud Computing Privacy: Concept, Issues And Solutions." 2022 IEEE Symposium on Industrial Electronics & Applications (ISIEA), (2022): 1-4.
11. Zawoad, S. "SECURING THE CLOUD Digital Forensics in the Cloud."
12. Guo, H., Jin, B., & Shang, T. "Forensic investigations in Cloud environments." Proceedings - 2012 International Conference on Computer Science and Information Processing, CSIP 2012, (2012): 248-251.
13. 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/BigDataSE2018,1250–1255.
14. 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.
15. Shah, J. J., & Malik, L. G. (2013). Cloud forensics: Issues and challenges. International Conference on Emerging Trends in Engineering and Technology, ICETET, 138–139.
16. 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.
17. Zawoad, S. (n.d.). SECURING THE CLOUD Digital Forensics in the Cloud.
18. 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.
19. 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.
20. Agarwal, A., and Agarwal, A. "The security risks associated with cloud computing." International Journal of Computer Applications in Engineering Sciences 1, no. Special Issue on (2011): 257-259.
21. Svantesson, Dan, and Roger Clarke. "Privacy and Consumer Risks in Cloud Computing." Computer Law & Security Review 26, no. 4 (2010): 391-397
22. Dobrosavljević, Vladimir, Mladen Veinović, and Ivan Barać. "Standard Implementation in Cloud Forensics." Singidunum University, Danijelova 32, Belgrade, Serbia, 2015.
23. Dorey, P. G., and A. Leite. "Commentary: Cloud Computing – A Security Problem or Solution?" Information Security Technical Report 16, no. 3-4 (2011): 89-96. Elsevier.
24. S. D. Wolthusen, “Overcast: Forensic Discovery in Cloud Environments,” Fifth International Conference on IT Security Incident Management and ITForensics,pp. 3-9, 2009.
25. National Institute of Standards and Technology, "Guide to Interating Forensic Techniques into Incident Response", 2006
26. Cloud Security Alliance, "Security Guidance for Critical Areas of Focus in Cloud Computing", December 2009, [1] http://en.wikipedia.org/wikilElectronic \_identity\_card, 2010.
27. Scott Zimmerman and Dominick Glavach, "Cyber Forensics in the Cloud", IAnewsletter Vol 14 No 1, 2011.
28. Sen, J., 2013. Security and Privacy Issues in Cloud Computing. Architectures and Protocols for Secure Information Technology, (iv), p.42.
29. Chou, T., 2013. Security Threats on Cloud Computing Vulnerabilities. International Journal of Computer Science and Information Technology, 5(3), pp.79–88.
30. 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.
31. S. Iqbal and S. A. Alharbi, “Advancing Automation in Digital Forensic Investigations Using Machine Learning Forensics”. Digital Forensic Science, 2020
32. 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,
33. 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
34. S. Iqbal and S. A. Alharbi, “Advancing Automation in Digital Forensic Investigations Using Machine Learning Forensics”. Digital Forensic Science, 2020
35. 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).
36. 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).
37. 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.
38. C. Szegedy, A. Toshev and D. Erhan, "Deep neural networks for object detection",Proc. Adv. Neural Inf. Process. Syst., 2013.
39. 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
40. 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.
41. 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)
42. 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.
43. Ahmed, M., Mahmood, A.N., Hu, J.: A survey of network anomaly detection techniques. J. Netw. Comput. Appl. 60, 19–31 (2016)
44. Ahmed, M., Mahmood, A.N.: Novel Approach for Network Traffic Pattern Analysis using Clustering based Collective Anomaly Detection, pp. 111–130. Springer, Berlin (2015).
45. 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
46. Yadav, R. M. (2019). Effective analysis of malware detection in cloud computing. Computers and Security, 83, 14–21.
47. 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
48. 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
49. Radha Rani, D., & Geethakumari, G. (2083). A framework for the identification of suspicious packets to detect anti-forensic attacks in the cloud environment.
50. 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.
51. 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.
52. 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.
53. E. Alpaydın, Introduction to Machine Learning Second Edition. .
54. Ben-david, Understanding Machine Learning : From Theory to Algorithms. 2014.
55. Das and R. N. Behera, “A Survey on Machine Learning : Concept ,” pp. 1301–1309, 2017, doi: 10.15680/IJIRCCE.2017.
56. 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.
57. N. Maurya and P. Bhatnagar, "Data Preservation Techniques for Cloud Forensics," International Journal of Computer Science and Engineering, vol. 4, no. 4, pp. 117-124, 2015.
58. Quick, Darren N., and Kim-Kwang Raymond Choo. "Google Drive: Forensic Analysis of Data Remnants." Digital Investigation 13.1 (2015): 1-12.
59. Sahu, Shubham. "An Analysis of WhatsApp Forensics in Android Smartphones." Bachelor of Engineering thesis, Chhattisgarh Swami Vivekananda Technical University, 2019.
60. Barradas, Diogo, Tiago Brito, David Duarte, Nuno Santos, and Luís Rodrigues. "Forensic Analysis of Communication Records of Messaging Applications from Physical Memory." Digital Investigation 29 (2019): 102-116.
61. B. Blakeley, C. Cooney, A. Dehghantanha and R. Aspin, "Cloud Storage Forensic: hubiC as a Case-Study," 2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom), Vancouver, BC, Canada, 2015, pp. 536-541, doi: 10.1109/CloudCom.2015.24.