Computers and Electrical Engineering 96 (2021) 107527

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Machine learning with big data analytics for cloud security

Abdul Salam Mohammad [a], Manas Ranjan Pradhan [b,*]

[a] School of Business, Skyline University College, Sharjah, United Arab Emirates
[b] School of Information Technology, Skyline University College, Sharjah, United Arab Emirates

ARTICLE INFO

Keywords:
Big data
Cloud computing
Cloud security
Data security
Data management
Data storage
Machine learning

ABSTRACT

The amount of data generated and transmitted more quickly, particularly with the demand for action in real-time, has greatly increased with the growing number of internet-connected devices. With the rising diversity of data and need for data integrity, it is more challenging to accomplish this processing on time. However, cloud and edge computing pose security problems and time delays, as computing models are being applied rapidly. Hence in this paper, a Machine Learning-Assisted Cloud Computing Model (ML-CCM) with big data analytics has been proposed to increase security and improve data transmission rates. The simplest approach for storing large volumes of data is cloud storage. Big data can manage or store large amounts of distributed data in clouds. The ML algorithms analyze supervised and unsupervised training used to solve cloud protection problems. The experimental results show that ML-CCM has a data transmission rate of 96.4%, effective data management of 94.3%, computational time of 35.2%, accuracy of 91.7%, and performance of 95.2%.

Introduction of machine learning and big data analytics for cloud security

The provision of various services via the internet, including data storage, databases, servers, applications, and networking, is called cloud computing (CC) [3]. Cloud computing is an accessible, end-user resource available without the client's immediate, precise organization and, in particular, the storing and processing capacity of information [15]. The cloud offers infrastructure as a utility and forms a distributed data center. It contains vast resources and includes processes for delivering, re-equilibrating, and tracking these resources [1]. It serves as a unified resource unit or more while providing service users/applications access without specific detail [2]. Cloud-based hosting allows files to be saved and viewed on demand from a remote archive.

Data is distributed from one device to another in analog or digital format. In essence, data transmission enables machines or components to communicate with each other [11]. Data processing consists of safe, reliable, and cost-effective data storage, conservation, and usage. More critical than ever is a useful data processing strategy, as businesses increasingly rely on intangible assets to generate value [12]. Cloud data management is the practice of maintaining a company's records on an offsite server usually owned and monitored by a cloud service. In data collection, information is collected and measured systematically, allowing someone to answer stated research questions, test hypotheses, and evaluate results.

Cloud data protection is the practice of securing the information of an enterprise in a cloud environment, wherever it is located, be it in rest or in motion, internally managed or externally managed by a third party. Distributed computing is a standard articulation

This paper is for special section VSI-spbd. Reviews processed and recommended for publication by Guest Editor Dr. Marimuthu Karuppiah.

* Corresponding author.
E-mail address: manas.pradhan@skylineuniversity.ac.ae (M.R. Pradhan).

https://doi.org/10.1016/j.compeleceng.2021.107527
Received 30 December 2020; Received in revised form 28 April 2021; Accepted 1 October 2021
Available online 11 October 2021

involving numerous items for various people. It provides users with public and personal data on a single internet network [14].

Data security refers to securing data from unauthorized access and data manipulation during the entire life cycle [4]. Safe data protection includes data encryption, prevention from hacking, control of access, and critical management activities that secure data across devices and networks [5]. The data is an essential asset of any company, so it is necessary to protect it from cyber attackers. Personal computers, tablets, and mobile devices, which could be the next targets of cybercriminals, are being secured for their records [6]. The data is generated, processed, stored, and shared by the establishment of a precious resource. Protecting the organization from offense and unwanted domestic or foreign access protects it from direct losses, damage to the image, disruption of customer trust, and degradation [7]. Cloud security—known as cloud computing safety—comprises a collection of regulations that work together to secure cloud-based services, data, and networks [8]. Cloud security can be optimized to precisely meet the needs of the enterprise by authenticating access to filtering data.

Data can be lost or stolen from cloud storage due to several factors, such as data breaches, malware attacks, and the above rise in cloud usage. Transfer of data refers to transferring between two or more digital devices [9, 10]. The servers must guard against threats. Cloud security means online data protection from theft, leakage, and deletion via cloud computing platforms. Firewalls, penetration tests, obstruction, tokenization, and virtual private networks are included in cloud safety methods, and public internet links are avoided. With web security in the cloud, traffic gets to the cloud rather than to the servers. The cloud analyzes the traffic and gives only legitimate users access to it. Any traffic not approved by the cloud is prevented from accessing the server.

Most staff member do not need access to all applications, pieces of information, or files. A plan must be established to ensure that each employee's correct degree of permission can only view or manipulate the applications or data needed to do their job. Assigning access control not only prevents an employee from accidentally editing information, but it also actually safeguards the data from hackers who have stolen credentials from that employee. Data manipulation refers to the data adjustment process, which makes it easier to read and arrange data. The attacker alters the database by adding a few other accounts and configures credentials to the same one. Unauthorized parties can then easily access and keep an eye on their business's sensitive information.

Cloud data storage delivers automatic recovery policy, technical assistance, and quick access everywhere [13]. It has several applications, including backup of data, data sharing and resources. It can offer standardized interfaces for additional network services. All nodes are physically connected, with wide-ranging clouds, regional clouds, and local clouds. Knots on any two adjacent levels have a relationship between parents and children, in which the knot can be seen as a particular parent-node user. Edge computing with CC implementation used for time-sensitive data collection delivers distributed computing energy at the system's edge to device designers and service providers [16]. Current edge computing expands this approach to simplify the transmission and execution of a wider variety of applications on edge servers by server virtualization technologies.

The conception of this model of distribution indicates a shift in the protection plans. However, CC has many security problems, such as vulnerability to clients, which delay the computing model's fast implementation. Cloud computing provides near-limitless access, from information storage to software services administration. The cloud paradigm embraces and facilitates massive hardware and infrastructure implementations. ML algorithms are used to solve security challenges and handle data more effectively. ML focuses on advancing computer applications that can find an acceptable speed to think about themselves. According to the models, the approach to learning begins from expectations or facts, such as models, direct understanding, or heading, to channel knowledge systems and make better choices about the topic.

The CC and Big Data Analytics (BDA) paradigm has been developed to solve data-oriented challenges. BDA offers the ability to process distributed queries in several datasets using commodity computation. Using Hadoop (a family of distributed data processing systems), CC provides the underlying mechanism. Instead of local storage on a server or electronic system, BDA uses distributed storage technologies based on cloud computing. Big data measurement focuses on fast-rising, virtualized cloud-based technologies in a cloud environment, a standard CC system, and big data mining. It permits the collection of vast volumes of parallel data sets in the cluster.

The main contributions of the study are

- designing ML-CCM with big data for effective prediction of cloud security events and data management;
- analyzing the supervised and unsupervised learning to predict threat detection; and
- performing numerical results. The proposed ML-CCM uses big data for data transmission rates, effective data management, computational time, accuracy, and performance when compared to other methods.

The remaining papers are structured by the following sectors:

1 introduction
2 literary works
3 ML-CCM method for increasing data transmission and computational time,
4 numerical results
5 conclusions

In this analysis, the software-defined SDP (Software-Defined Perimeter) framework in the internet of things (IoT) applications is recognized for message queuing telemetry transport (MQTT). Refaey et al. [17] suggested the proposed SDP framework for MQTT with IoT applications. In reality, the SDP offers an additional layer of protection with or without SSL/TLS, replacing it with a Single Package Authorization (SPA) method with the conventional login protocol (username/password). There are also integrating tests built into the codes. The device is proven safe from denial-of-service and off-line dictionary forms of attacks, even by standard login credentials.

Mohammad et al. [18] introduced the Trustworthy Evaluation Framework (TrustE-VC) for industrial connected vehicles in the cloud. A significant field of study has been the convergence of cloud computing and ad hoc vehicle networks, namely vehicle clouds (VCs). This incorporation was suggested to speed up the introduction of smart transport networks. This pattern involves a security assessment system that guarantees data safety, information confidentiality, and resource availability. In general, this article paves the way for a stable VC using appropriate safety assessments, underscoring the VC community's guidelines and challenges. This article stresses the value of protection when building industrial VCs, emphasizing several levels of defense.

Nguyen [19] discussed the security quantification of IoT infrastructures using hierarchical modeling and analysis framework (HMF) for availability. To determine its efficiency and safety characteristics, it is essential to build an IoT infrastructure. Modern IoT infrastructures have a diverse and heterogeneous architecture, so it is a challenge for system professionals and engineers to take care of both the architecture and the operative specifics of IoT infrastructure. In this respect, they suggest a hierarchical modeling framework for IoT infrastructure availability and security quantification.

Yousafzai et al. [20] studied IoT-supported mobile edge/cloud computing using a process migration-based computational off-loading framework. Smart devices have become an integral part of IoT. However, these devices' resource constraints are in f battery power and storage, processing, and the hampering of computer-intensive application patch deployment, often requiring large bandwidth, stringent response time, long battery life, and high computing power. The proposed architecture does not require binaries for network edge, and it seamlessly migrates native applications. An experimental testbed is used to assess the proposed system.

A blockchain-based access control framework (BCBACF) with privacy protection in the cloud was explored by Yang et al. [21]. A cloud, being a programming model that shares and facilitates computer-friendly, on-demand connectivity, provides new information processing and resources for many business markets, substantially lowering user computation and storage costs while increasing ease of use. To resolve reliability issues, they proposed an auth privacy chain (APC) BCBACF. Finally, their findings revealed that APC cannot deter hackers and administrators from accessing unauthorized tools, yet it still preserved privacy. The auth privacy chain is focused on the EOS utility.

A BCBACF for azure cloud infrastructure is suggested by Liu et al. [22]. IoT, being a detailed-use scenario, is a driving force behind the formation of a digital society. Public cloud providers cannot sufficiently overcome potential problems, as their activities are primarily aimed at applications, and networks can hardly be applied to end devices. They proposed a Data-Centric IoT Framework (DCIoTF), which includes three existing networks—WiFi, thread, and LoRaWAN-layer protocol—that account for extensive IoT- and

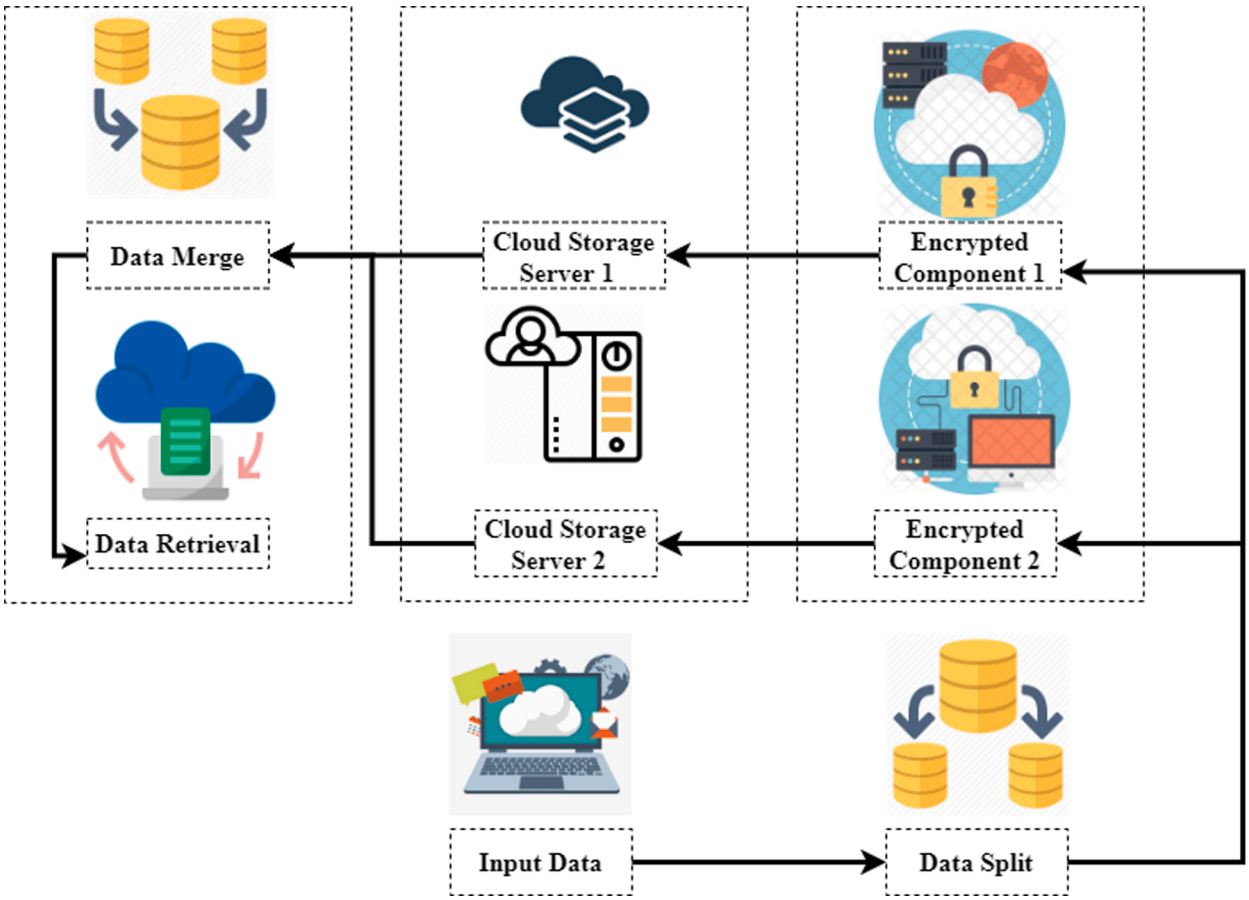

Fig. 1.. Cloud data distribution system.

networking IoT-use scenarios on the local, personal, and commercial networks.

Li et al. discuss an integration of security and reputation approaches for enhancing cloud-based IoT security through trustworthy cloud services [23]. IoT has become a qualified service platform, allowing heterogeneously distributed networks to be built ever more broadly. To accomplish this, a novel trust assessment framework (NTAF) was proposed for cloud services protection and credibility. This system requires a confidential review of cloud providers to ensure the IoT context's integrity by integrating confidence appraisal approaches focused on security and credibility.

Secure and efficient multi-authority access control for IoT cloud storage (SEM-ACSIT) for security was deliberated by Xiong et al. [24]. The data access control scheme for club storage was perceived by the ciphertext policy attribute-based encryption (CP-ABE) approach as a promising strategy for improved efficiency and security. As an IoT cloud management system, SEM-ACSIT provides a modern, reliable, and efficient multi-authorization access control system that maintains both reverse and reverse authentication in the event of revoking a user attribute.

Based on the survey, there are some challenges in the existing model. In this paper, the ML-CCM model is proposed to predict data transmission threats in the cloud environment. Big data usage helps to overcome data management difficulties and issues. The following section discusses the proposed model.

3. Machine learning assisted cloud computing model (ML-CCM) with big data

The cloud infrastructure has evolved and made cloud computing resources a priority for many unique cloud attacks. Cloud storage stability is important for cloud vendors. Service providers need to deploy a robust cloud security scheme. Hence in this paper, ML-CCM has been proposed for cloud security and increased data management using two-factor authentication when provided with the option. In addition to one's password, anyone signing up would be required to provide additional information. Common authentication methods include answering a secret question, providing a personal PIN, or entering a code sent by the cloud provider. The authenticator software can indeed be downloaded. Not all accounts automatically request creating secondary ID checks.

A threat classification model has been developed that uses machine learning algorithms as threat detection tools. They considered three requirements in the model: a) type of algorithm for machine learning, b) data input, and c) cloud computing. Two forms of instruction are considered under the first criterion: regulated and unattended. The second criterion is the data on system efficiency and network traffic, which are added into all learning forms. This study identifies the influences on technological and system challenges, such as data diversity, data storage, data integration, data collection, and resources management.

Figure 1 shows the cloud data distribution system. This paper studies the exploitation of cloud operators and seeks to prevent data released by cloud customers. Additionally, it proposes an ML-CCM designed to achieve adequately distributed mass storage and high-security safety. The mechanism is designed to encrypt all data and store it on the various cloud servers without creating significant overhead and latencies. User data is separated into two sections, which are delegated to be stored in cloud A and cloud B services. The suggested implementations complete the spilling-data process. The use of a protected and reliable data sharing mechanisms is designed to spill data to avoid the expense of locating confidential information on the cloud side.

The operation's implementation involves two supporting algorithms—encryption and decryption of data. Information is partitioned into functional units before the data is stored. The solution is to break down the confidential data into two encrypted components for distributed cloud storage. The principal challenge faced by ML-CMM is to prohibit employees of cloud vendors from accessing the data without cutting down on performance of the original data entry and cloud computing servers. An additional challenge is defining a solution that can successfully save data on the cloud storage without raising the processing times while guaranteeing that cloud operators cannot access the data. The on-site data storage transfer service and activity flows offer two high-performance, cloud-storage paths with the necessary scalability and speed to simplify the data transfer processes. The transfer appliance is offered as an offline data transmission storage server located in the customer's data center and then transmitted to the ingest location, where the data is then uploaded into the cloud storage.

Inputs are the original data composed of a string of confidential information packets. In some cases, confidentiality refers to the fact that internal actions with colleagues are not discussed. In other cases, it relates to the fact that business secrets or other information are not shared with competitors, the press, or anybody else outside a company. User data is stored in cloud computing on remote servers operated by others and accessed via the internet. Confidentiality means preventing unauthorized access to data and guaranteeing that the authorized user can only access that data. The outputs are two distinct data packets that are sent to multiple cloud storage centers. The newly created data packets must conceal confidential information completely that cloud operators cannot read even though the data is available. The main advantage of cloud encryption is the same as in any encrypted application: only authorized parties who access decryption keys can view encrypted data. Encrypting data ensures that it is not useful if it falls into the wrong hands. The implementation of the data-conversion method requires a substantial low execution time. These two steps represent two main data transfer procedures. Every confidential document should be stored in locked file cabinets or rooms only available for people who have a "need for knowledge" business. All information should be secured electronically via firewalls, encryption, and passwords.

Individual companies can possess their own clouds, which employees and customers usually access through the internet and on their private networks, called private clouds. One processes data to split the input data strings into two distinct data chains. An alternative is to combine the information to retrieve the original data. As shown in the diagram, the partition input D into two different components before sending it to the Cloud. Two encrypted parts A and B are present, as seen in the diagram. This is completed a few measures. Next, create a random data C parameter to produce new D-C data packets. This key must be kept on the user's side in a separate file. Finally, all encrypted packets are shipped to independent servers in the cloud.

Figure 2 shows the cloud control distribution network. Encrypted residual data is shielded from unwanted access to disclosure.

Cloud service providers (CSPs) usually offer computing facilities with encryption capability. The corresponding encryption keys ensure reliable encryption. CSPs have a combination of CSP or consumer-managed keys for customers. CSP-managed keys are convenient and therefore do not require the user to verify where and how keys are stored. User-managed keys place the responsibility of key management on the consumer for better control. Thus, the CSPs provide cloud-based hardware security modules for stable key management. CSPs guarantee comprehensive protection against recurrent data loss. However, no infrastructure is flawless, and big cloud companies have destroyed user data unintentionally. Cloud users can make mistakes, which can lead to data loss and CSP errors. CSPs may provide consumer configuration services to carry out other backup and recovery tasks. To ensure consistency, CSPs repeat results. The development of virtual servers, infrastructure, appliances, and computing resources is virtualization. This technique modifies relationships between hardware and software. It is one of the basic elements of cloud computing technology that uses cloud computing capabilities.

As seen in the above figure, confidential data can be obtained by logging and tracking systems, backups, information delivery services, and elsewhere during device operations. Cloud implementation should be scrutinized to identify where confidential data may be copied or stored. Media machines malfunction periodically and must be replaced. If the device itself has crashed, the device holds user data. The CSP manages the removal from the output of storage media. The CSP controls the user's networks and resources and does not monitor the systems and software where customers use its services. The CSP offers market reporting information about the use of services by customers. CSP surveillance data is used as the first monitoring line for improper access to and unforeseen activity use of devices and software or their users. The monitoring information should be used as a first monitoring route. The data supplied by the CSP should be clearly distinguished from the data obtained in the on-site surveillance. Therefore, the CSP must learn to secure its cloud-based services with new data. It is essential to understand the nature of the CSP results, decide the normal cloud usage, and use CSP tools to recognize anomalies. It is also important usage of the data generated by the CSP to incorporate further cloud-based tracking to the extent possible.

Of note is that on-site monitoring techniques cannot operate in the cloud. For instance, the virtual routers do not associate with the span ports that all network traffic operates on, complicating flow-based control. Consumers must cautiously design and incorporate all extra monitoring to ensure it is completely compatible with cloud automation. Distributed denial-of-service attacks are increasing, and a top cloud computer security solution concentrates on actions aimed at stopping huge quantities of traffic on a cloud server. It includes

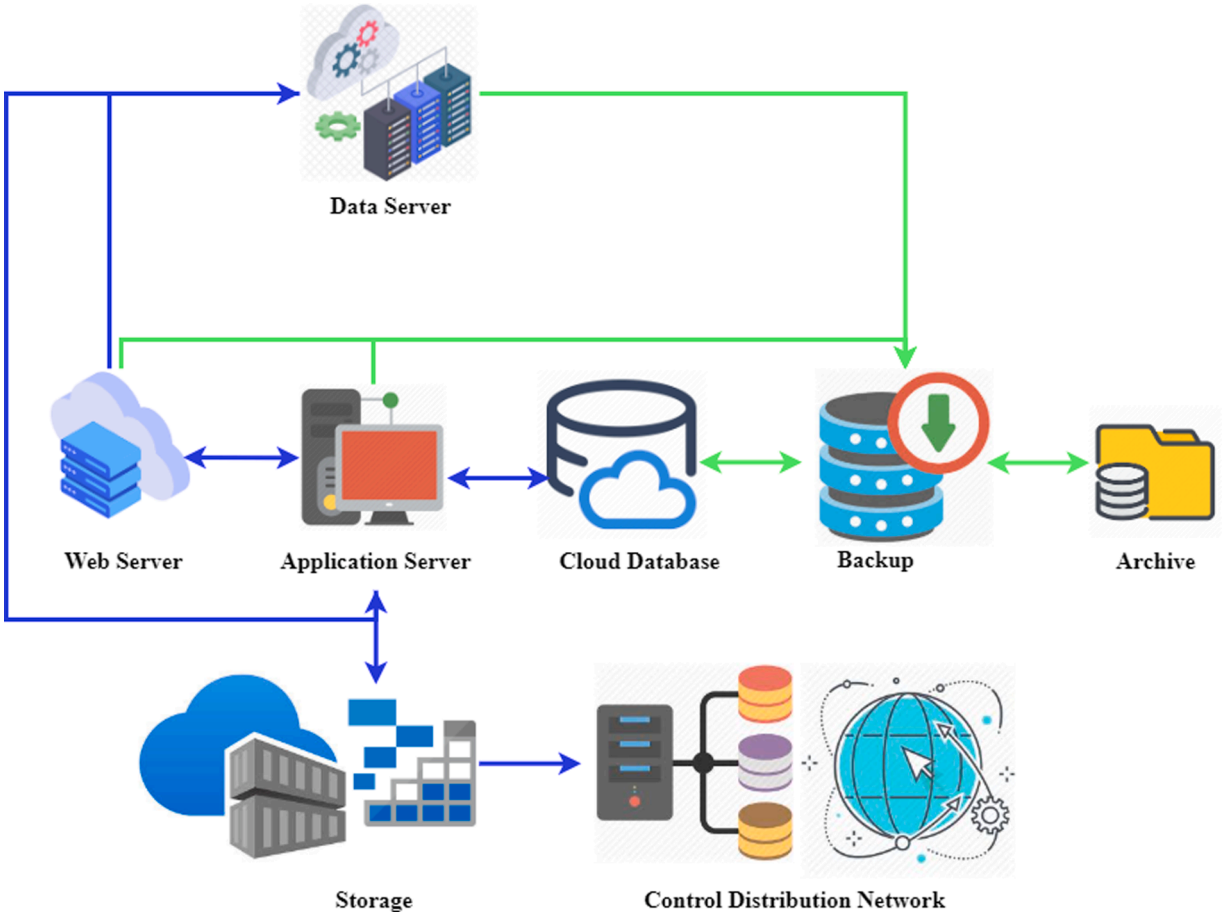

Fig. 2.. Cloud control distribution network.

monitoring, absorption, and distributing data to minimize risk. In the ever-increasing age of data breaches, security protocols must be in place to protect sensitive information and transactions using the cloud security solution. The protocols must prevent a third party from waiving or manipulating the transmission of data.

The combination of CSP monitoring information, customer cloud monitoring information, and on-site information from customers

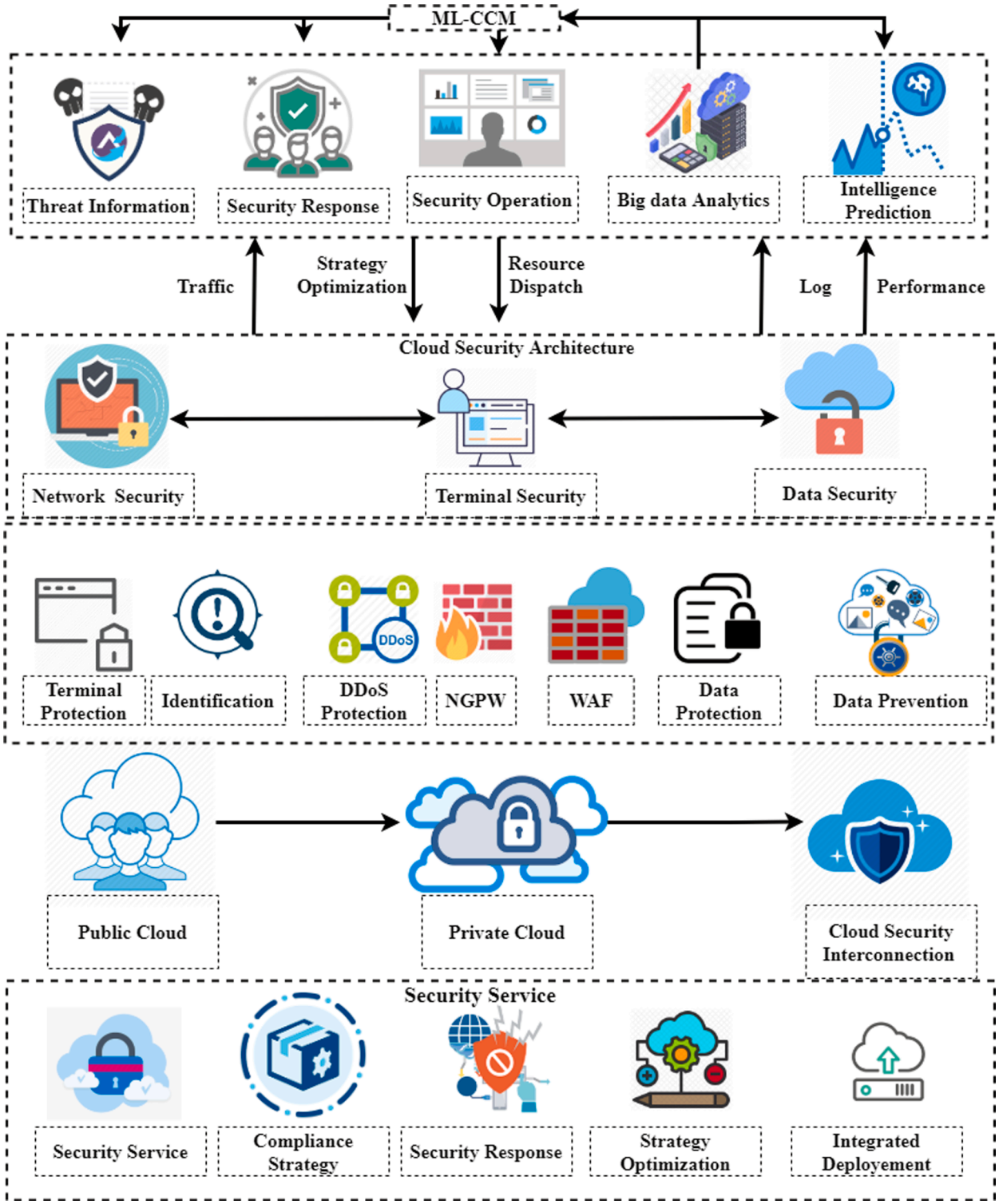

Fig. 3.. Proposed ML-CMM.

are important for the customer's hybrid deployment. This setup transfers some resources to the CSP and conserves various resources to create a full image of the organization's role. The above figure shows a cloud-based analysis and tracking enclave, which incorporates all three surveillance data sources. Although this enclave may be installed in or on the cloud premises, cloud deployment does have benefits. First, for data transmission to and from their facilities, CSPs usually charge fees. CSPs charge more for cloud backups than for cloud transfers to facilitate and hopefully expanded usage of their services. Due to the amount of data concerned, transferring data from on-site surveillance to the cloud can be cheaper than moving cloud-based monitoring data to an enclave on site. Second, storage for large data volumes can be less costly in the cloud, particularly for the retention and non-active archived data use.

Figure 3 shows the proposed ML-CMM. Advanced information technology represents the exponential growth of emerging technologies such as big data, and a digital revolution has powered the public cloud. The accompanying safety issues have become exceedingly serious. The business model is becoming more complicated with the popularity of cloud storage and micro-service modes. Furthermore, there have been major shifts in the pattern of network security attacks. Security officers must meet a lower level of attack, larger attack areas, and more attack methods. Network attacks are no longer the patent of hackers, with the low cost, modulus, and

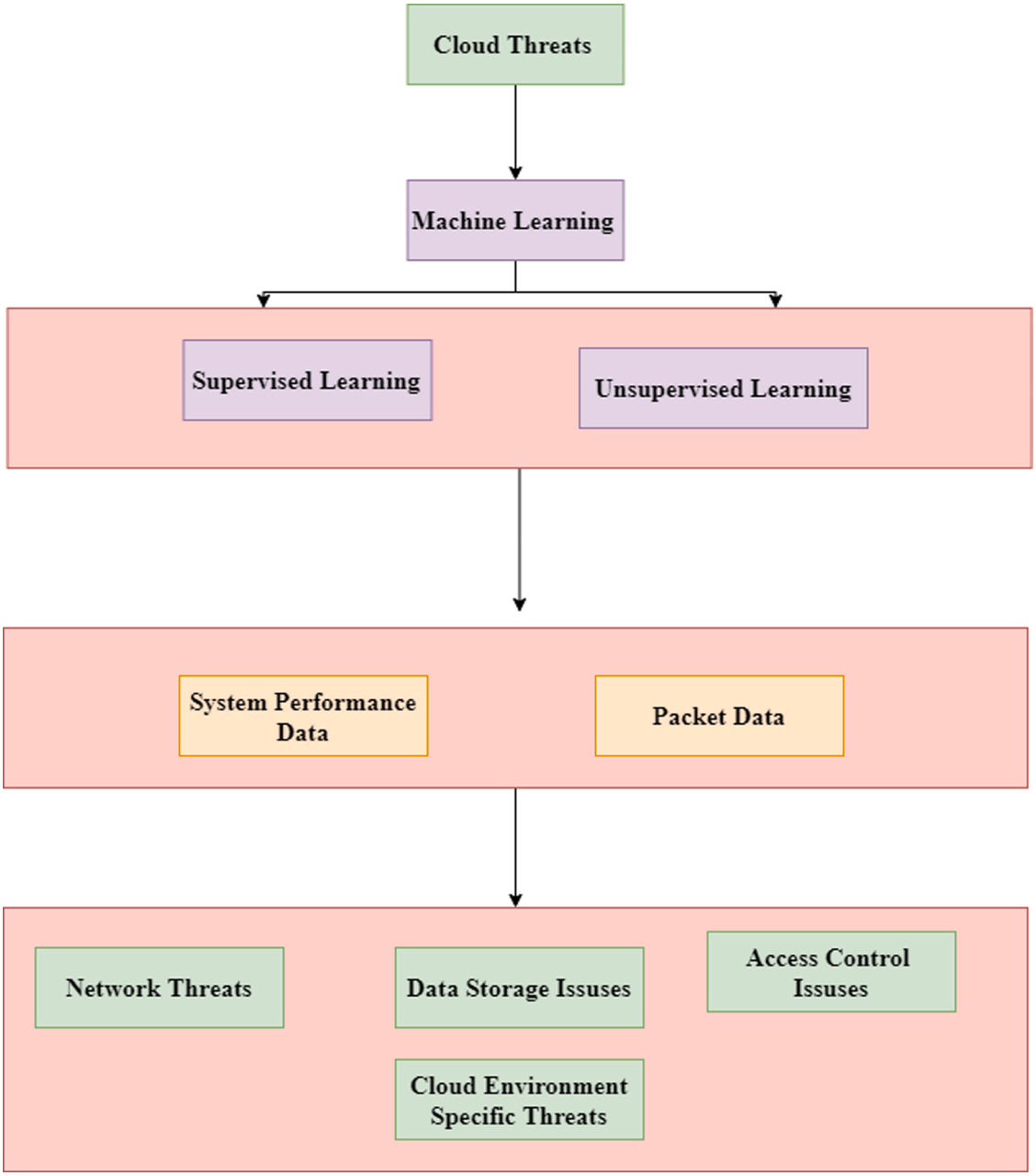

Fig. 4.. Cloud computing threats classification model.

weapon patterns of attack software. Any individual or organization with additional motivations will use these resources to saturate and target businesses, contributing to immense digital property and economic losses.

Unnecessary safety threats will drastically affect the enterprise's value and even compromise its sustainability. Even a minor potential risk could threaten its brand name, customer data, and intellectual property. Security network events can break out, and companies frequently face extortion software. Suspension of operations costs and data loss must afford enormous decryption payoffs. Protection-line databases have been fragmented. Hotel check-in logs, credit card numbers, personal health information, and other personal details have been selling everywhere through illegal channels. The risks to protection trigger tremendous damages to businesses. At the same time, cloud storage stability continues to be a topic of argument for several organizations.

The hybrid world in which hardware and software coexist often requires higher application security specifications to accomplish the overarching goals of a corporate safety strategy: (1) to design and enhance user information systems in areas such as terminal protection, network security, computer security, cloud security, and security services; and (2) to improve the visualization of the internal information ecosystem continually and with the support of the ML-CCM model to improve the overall security and maintenance standard. This approach includes implementing a terminal safety scheme that incorporates machine learning, big data behavior analysis, and simulator technologies to defend against zero-day attacks to cope with unknown risks efficiently. The network helps businesses deal with distributed denial-of-service (DDoS) attacks, application attacks, and web application attacks. It must take into full consideration increasingly encrypted traffic by conducting smart analyses and requiring detailed identification. The protection of the data system is critical as the collecting point of information. The approach recommends implementing a network protocol analytics system and user behavior analysis to help organizations identify data breaches and discourage information leakage.

Cloud and big data processing and hot places have major consequences for the digital transformation of enterprises. Network information management technologies will expand organizational security protection to public and private clouds to the same degree as the data center properties protected in the cloud. Simultaneously, in place of the big data processing approach, the technologies must gather information about threats and vulnerabilities, examine and address the total security threats of organizations in general, carry out security emergencies for main enterprises, and correctly manage security issues.

Figure 4 shows the Cloud Computing Threats Classification Model. This model belongs to the surface cloud computing security risks

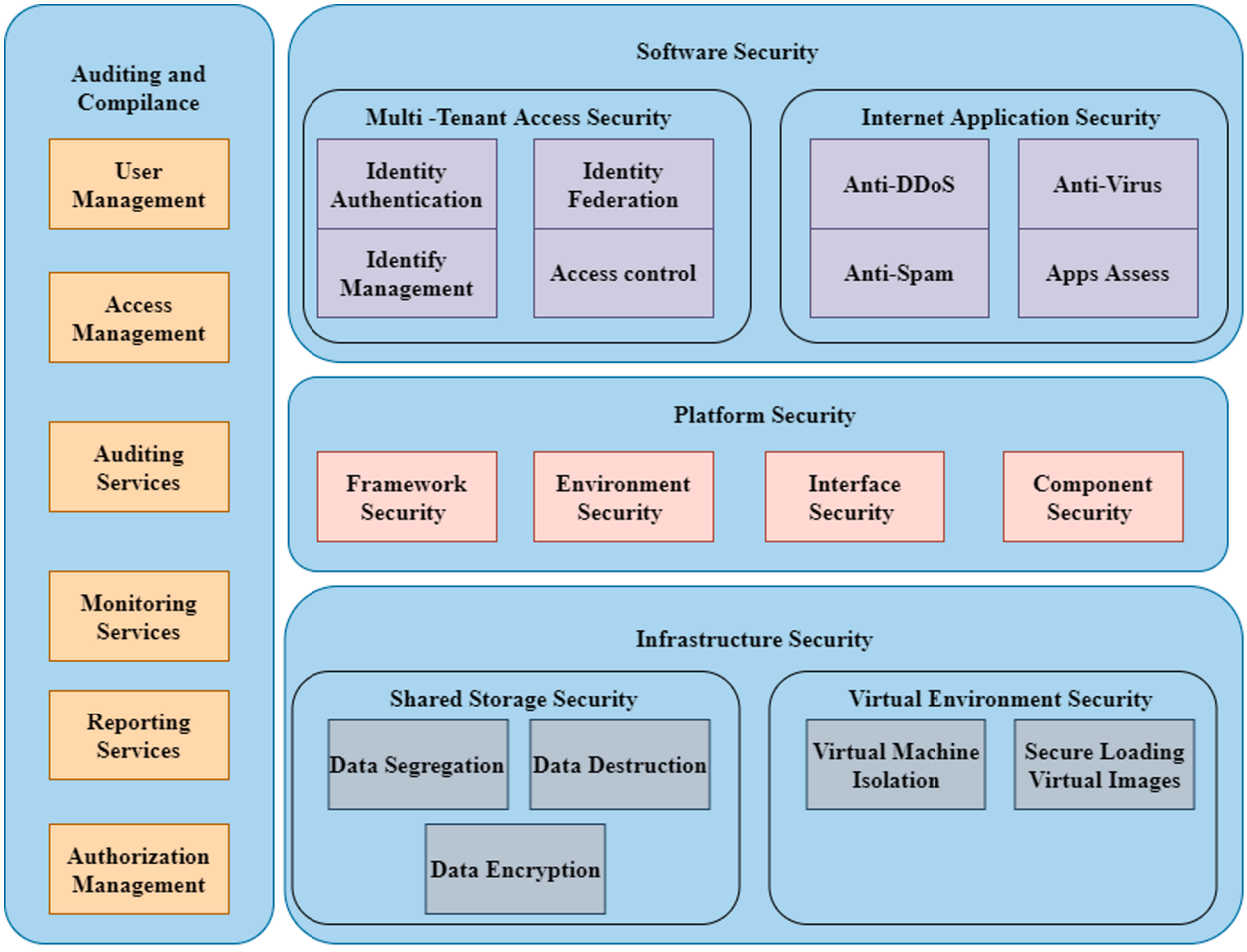

Fig. 5.. Cloud computing security architecture.

mainly through a computer study model or classrooms. Classification learning is supervised where the actual result for an example of training is given, although the example of training results is not specified in unregulated learning. Supervised learning is powerful for well-defined tasks, and it can be used less than where a comprehensive data set is available. Uncontrolled learning works very well in identifying patterns in data, which can detect malfunctions or behavioral changes without depending on labels. Classification learning success can be measured by testing the independent findings with established, true classification results. If there are no defined results, objects are clustered according to related items. A critical step is to choose a suitable model algorithm.

Algorithms are dependent on numerous parameters, and predictive accuracy is usually chosen. Selecting a learning algorithm is critical for computational time, high dimensional values, or over-fitting. Each instance supplying an entry to machine-learning algorithms will contain many values in each attribute. Consequently, cloud attacks have characteristics that classify them and input them on machine learning algorithms. The smart transport system aims to reduce traffic problems to increase efficiency. It aims to reduce travel times and improve the safety and convenience of passengers. Thus, the intention is for the easing of traffic congestion and the efficient use of road safety and infrastructure.

Economic classification is a significant consideration for calculating proper input data for the learning of algorithms. However, it is not easy to classify input characteristics. Much work is required because the classification performance and accuracy are complicated and often contain noise and incomplete details to categorize essential features. It is necessary to detect the pattern of activities by analyzing machine parameters' change by considering device output data. A packet server is used for network data to collect all packets received during attack execution. The service also requires filtering to delete irrelevant and redundant features. This helps computer algorithms to run more rapidly and effectively.

Cloud is an internet-based computer that provides consumer-based metering services. Multiple techniques and algorithms are employed to minimize cloud power consumption. Virtualization and other techniques are used for cloud installations to separate software from physical servers' characters. This technique allows for redesign servers to optimize them and decrease costly energy requirements. In addition to the entire schema, there are two broad categories of the cloud attack level: network threats and cloud computation threats. When designing cloud solutions, network protection is essential because of it being a sensitive area for multiple attacks. Network attacks are intended to control the network and offer valuable information, to alter the content, and to corrupt or consume network resources. Common cloud network attacks are (1) account hijacks or server captures that occur when attackers target users' processes and transactions and (2) DoS attacks that block users from accessing the services normally. Unique cloud computing threats are also target other, larger clouds. Infrastructure risks involve vulnerable interfaces and application programming interface (API) software that might be used to communicate with cloud applications, mutual technology bugs, or disruptive insiders who deliberately exploit the organization's networks, systems, or other tools. Both device and network traffic details may identify network-related challenges for cloud computing. In addition, system performance and network traffic data are entered into the learning of all types, supervised and unsupervised, by considering various parameters such as implementation, sophistication, computing time, managing missed values, overfitting, and the selection of machine-learning algorithms.

Figure 5 shows a cloud computing security architecture. Cloud computing security is an emerging sub-domain of computing security, network security, and information security more generally. It refers to a wide variety of regulations, technologies, and controls to secure the files, software, and related cloud-storage infrastructure. Note that cloud computing protection does not include cloud-based security programs, such as cloud-based anti-viruses, anti-spam, and anti-modulating systems. Privileged user access, compliance with laws, data location, data segregation, rehabilitation, study support, and long-term sustainability. Some of the leading cloud companies' protection and privacy policies have been analyzed in three main areas: safety and privacy.

The Cloud Security Alliance (CSA) brings together vendors, non-profits, and specialists to address best practices for cloud information security for the present and future. Investigations and detailed analyses have been undertaken concerning the validation processes of cloud computing security problems for each security issue. Safety concerns have been discussed from various viewpoints, including cloud infrastructure, application management structures, cloud features, and cloud security issues. They highlight some potential cloud protection possibilities. According to the serial peripheral interface (SPI) models, service delivery models, and implementation models, security problems occur in all system areas at the network, host, and device levels. Private data recognition depends on and is the primary task of security in a given situation.

Processing quality is an essential measurement of the performance used to assess different frameworks' efficiencies for resource management. It is a common statistic for the maximum number of I/O operations in disc or memory or for the rate of data transfer between the cluster computing speed within a certain time. Based on extensive data, the average processing speed, expressed by $\overline{n}$, calculated after n iterations, run the maximum amount of data operations conducted over time in equation (1):

$$\overline{n} = \frac{\sum_{j=1}^{m} n_j}{\sum_{j=1}^{m} S_j} \tag{1}$$

As described in equation (1), the average processing speed has been derived. Performances for Apache Hadoop, Spark, and Flink are demonstrated during a series of tests on a multi-core cluster setup. Efficiency results for Flink and Apache Spark prove to be beneficial platforms for Hadoop's execution of benchmarks that were not sorted. Services such as word counts suggest that Spark has better performance results, while Flink has a better algorithm result. Hadoop Apache and Spark experimentally compare output statistics. Experimental package, required runtime statistics, operating nodes, and dataset scales are compared.

In case of a malfunction of one or more components, fault tolerance is the feature that causes a system to continue to operate. High-performance calculation implementations require hundreds of nodes and are interconnected for the performance of a particular process. System resistance is represented as TOL_{ft} for fulfilling its requirements after a break is the ratio of time to finish activities

without observing any fault occurrences overall running time when such fault occurrences have been found. The system condition is reversed to the following in equation (2):

$$TOL_{ft} = \frac{t_y}{t_y + \delta^2} \tag{2}$$

As discussed in equation (2), fault tolerance has been computed. The symbol t_y is the average correct time of execution obtained from a program execution that is supposed to be default-free, or by combining time of execution from many applications delivering a known accurate performance, and δ^2 reflects the time variation for the implementation of the program due to error events. For HPC, consisting of a series of computer-intensive functions, TOL_{ft} is computed as TOL_{α_m}. The general application durabilityTOL can be measured as $TOL_{\alpha_m}$for any task. For HPC, TOLhas been represented in equation (3):

$$OL = \sqrt[m]{TOL_{\alpha_1} \cdot TOL_{\alpha_2} \cdot \ldots \cdot TOL_{\alpha_m}} \tag{3}$$

As inferred in equation (3), the durability of tolerance has been calculated. In comparison, Spark conducted storm as opposed to massive data handling. However, the speed-up results deteriorated with the growing data collection, and Hadoop quickly outperformed Spark for large data sets. Spark again stated to crash, with Hadoop still doing its job for massively large datasets.

Scalability means reacting to big loads or adjusting sizes and workloads while supplying services on time. This process can be classed as upscaling or downscaling. One of the vital needs of businesses is to process vast data to resolve practical market issues, which can be classified either as upscaling or by adding new nodes. Dynamic resource scalability enables companies to perform significant parallel computational time, thereby minimizing time, cost, and effort overall. Let Z be the workload size before enhancing system resources; the execution working load that profits from improving system resources are β. The component that would not be increased in resources is $1 - \beta$ in equation (4):

$$Z^{'} = \delta Z + (1 - \beta) mZ \tag{4}$$

As shown in equation (4), scalability has been derived. Parallel time for executing a scaled workload is known as scaled workload acceleration on m processors, as shown in equation (5):

$$T^{'} = \frac{Z^{'}}{Z} + \frac{\delta Z + (1 - \beta) mZ}{Z} \tag{5}$$

As deliberated in equation (5), scaled-work load speed-up has been derived. Apache Spark and Flink scalability and efficiency comparisons of various theoretical knowledge requirements use feature selection frameworks. The comparisons test the scalability of data and dimensions to illustrate an insightful benchmark that reflects real-world machine learning criteria. The model has distributed algorithms for maximizing supervised and unsupervised learning algorithms. The implemented machine-learning algorithms for supervised learning via the database select the common K-means clustering algorithm for scaling analysis. Flink's average runtime is comparatively low for resource-restricted applications, while Spark had a distinct advantage until sufficient primary resources became available when additional device nodes are introduced.

Strongly connected are big data and low latency. Big data systems give organizations real benefits at a computational time. When the efficient big data platform is cloud computing, an essential prerequisite is to have a high-speed network to reduce connectivity latency. In comparison, large data systems typically require a centered architecture where the scheduler assigns all tasks from a single node, significantly affecting latency when the data is large.

The average latency expressed as $\lambda(Z, M)$ for workload, size is defined as the average overall time needed for each processor to complete the task: The overhead time between software start-up and finish computational time. The time program delivery in a distributed system and is the successful execution time, and a λ_j the amount of total idle units with the processor fixed Mprocessor time in equation (6):

$$\lambda(Z, M) = \frac{S_{elapsed} - S_j + \lambda_j}{M} \tag{6}$$

As discussed in equation (6), average latency has been computed. A detailed study of latency and performance of Apache Storm, Flink, and Spark has been discussed. Instead of the conventional HPC environment, where data is being processed on-site, big data systems are increasingly being implemented in the cloud in order to view or archive data users' confidential information quickly. While data confidentiality and security issues are not a concern in the distributed computer market, the massive deployment of cloud storage platforms in the big data arena raises their importance.

The data sets are available to other users for numerous reasons and can result in security and privacy threats. Let M be a list of categories for security framework which can be given. A system can use an encryption mechanism for security to provide data protection and a checklist of access to authentication services. Let max$Z_j$be the cumulative weight allocated to the protection category from a list of M categories and Z_j be the credit score of a specific resource management system in equation (7).

$$Security_{Score} = \frac{\sum_{j=1}^{M} Z_j}{\sum_{j=1}^{M} \max\left(Z_j\right)} \tag{7}$$

As introduced in equation (7), the security score has been calculated. To manage authentication and authorship mechanisms, Spark uses a password-based, shared-key setup and access control lines. Flink's stream brokers offer various providers with authentication

mechanisms. In addition, in many situations, this will result in program failures and an inefficient use resources, attributed to a loss of time and money unless big data systems are optimized to achieve greater data sets. Due to their specific design features and implementation criteria, each big data architecture has been created. The results of seven factors in the star score, $R_{RF} \in [0, T]$, which is the highest appraisal level, are outlined based on the primary factors and their objective evidence for assessing the big data frames. For each resource framework (RF), the general equation to generate the rating is given in equation (8):

$$R_{RF} = \| \frac{T}{\sum_{j=1}^{7}\sum_{i=1}^{M}\max(Z_{ji})} * \sum_{j=1}^{7}\sum_{i=1}^{M}(Z_{ji}) \| \tag{8}$$

As is shown (8) the resource framework has been calculated. Where the cumulative number of trials for the same set of measurement criteria is M max(Z_{ji})for each sample of literature, the relative normalized weight allocated to each literature is max(Z_{ji}) $\in$ [0, T], and (Z_{ji})is the experimental test-bed frame score computed from any study. With other implementations, Hadoop is highly compatible and extensive. The device provides a long-term, stable fault tolerance mechanism to supply a flawless platform. In low-cost design, Hadoop will operate. The proposed Ml-CCM and big data cloud data storage capabilities improve cloud performance while achieving a higher data transmission rate, more effective data management, and better computational time, accuracy, and performance.

3. Result and discussion

The proposed ML-CCM and big data predict the threat, and the results have been performed based on data transmission rate, effective data management, computational time, accuracy, and performance.

i) Computational Time Ratio

Machine learning algorithms can benefit from this and evolve. This study shows that device and network traffic knowledge can explain network-related threats and specific cloud computing threats. Input to all types of learning is tracked and regulated by system output data and network traffic data. This paper determines the machine learning algorithm-based constraints by considering multiple parameters, such as the program's complexity and computing time. A great deal of data is stored in the cloud. This data serves as a source for algorithms used in machine learning. Usage of ML to store, network, and exchange cloud data. Figure 6 shows the computational time ratio.

i) Data Transmission Rate

Data transmission proceeds in different phases of the data cycle, such as data processing (from distribution centers to cloud storage), data integration (from several data centers), data management (for the transition of consolidated data to cloud systems), and data analysis (for the shift from storage to analysis). Intelligent preprocessing methods and data compression algorithms are required to efficiently minimize the data size before transmitting the data into the cyber-infrastructure network effectiveness models, with

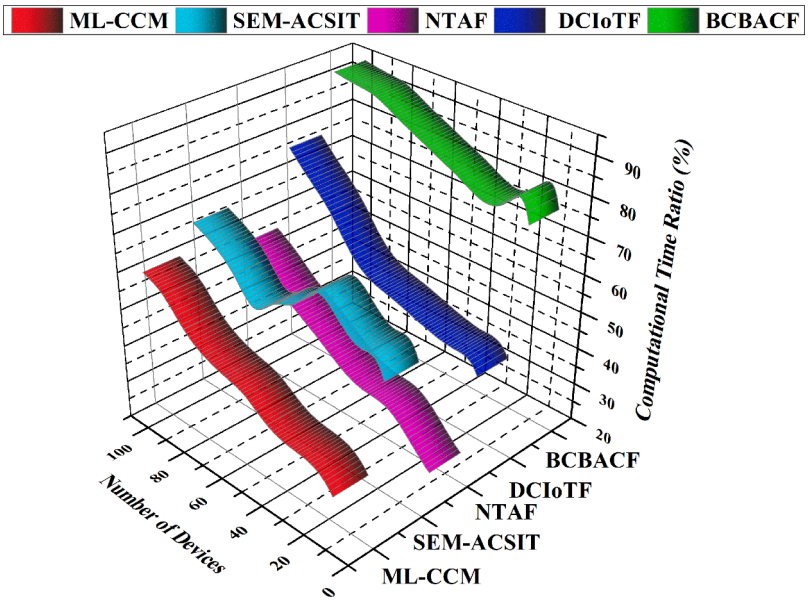

Fig. 6.. Computational time ratio.

various data compressor technologies relaying geospatial data. Furthermore, transfers involve shifting big data to the local data enters to the cloud platforms. Powerful machine learning algorithms are developed to suggest the best cloud provider (location) and optimal data transfer speeds, based on ML-CCM. Figure 7 shows the data transmission rates.

i) **Effective Data Management Ratio**

Cloud-based data management capabilities are required to collect data from an unparalleled volume. To utilize cloud computing for big data processing, the cloud offers a platform to access and incorporate distributed data in a heterogeneous environment. The plan must include a data management system for the private cloud technology handling big data. Likewise, the big data infrastructure for collecting and interpreting data is commonly used for huge, cloud-based database management, processing, monitoring, and sharing. The proposed ML-CCM architecture supports multiple data processing models such as Map-Reduce in-memory computation and agent-based programming to handle massive data collections in the cloud. A multi-source, heterogeneous data management, processing, and analysis infrastructure, architecture, or application is needed to promptly utilize the current data storage, computation, and technology. Successful regulations on data access control and protection management should be investigated and implemented with modern data management systems and storage structures. When data proprietors monitor virtualized storage, data security, integrity, and availability are simplified in the cloud era. Figure 8 shows the effective data management ratio.

i) **Accuracy Ratio**

The development of new safety systems for anomaly detection is proposed with machine learning algorithms because they demonstrate the fast processing of real-time predictions. The machine learns to protect the network from malicious nodes while alleviating the imbalanced data problem. The proposed method combines a supervised machine-learning algorithm with the past knowledge of the network node and a particular iterative algorithm designed to increase the accuracy of seldom detectable attacks. The algorithm for machine learning is used to construct a classifier that distinguishes between attacks. This classifier is kept in a private archive by the system. Except in the classes with a comparatively low number of training entrances, the proposed model improves overall identification precision and maximizes the number of correctly detectable in a cloud environment. Figure 9 shows the accuracy ratio.

i) **Performance Ratio**

The cloud storage system will serve as an efficient database server platform for the large-scale processing of data. Data analyses connect cloud computing to modern architecture to provide computing technologies with a major framework for all types of cloud-based services. There currently different cloud technologies because it becomes increasingly difficult for big data to work with concurrent processing. It enables the recovery in clusters of huge numbers of parallel data sets. The map reduce is a simple example of large-scale cloud computing. CPU resources, data storage, and network cluster computing are of high quality in the distributed device environment. It successfully incorporates strategies, including the ability to offer the requisite features and high efficiency of modern

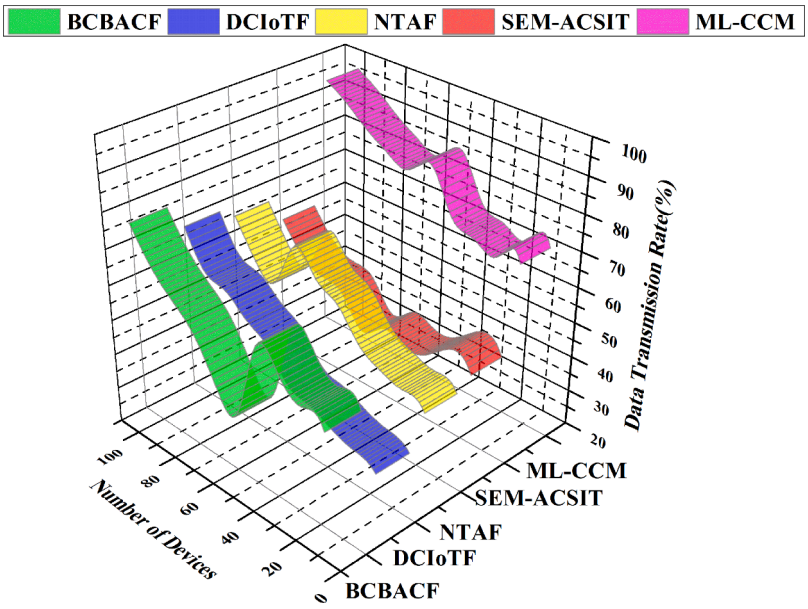

Fig. 7.. Data transmission rate.

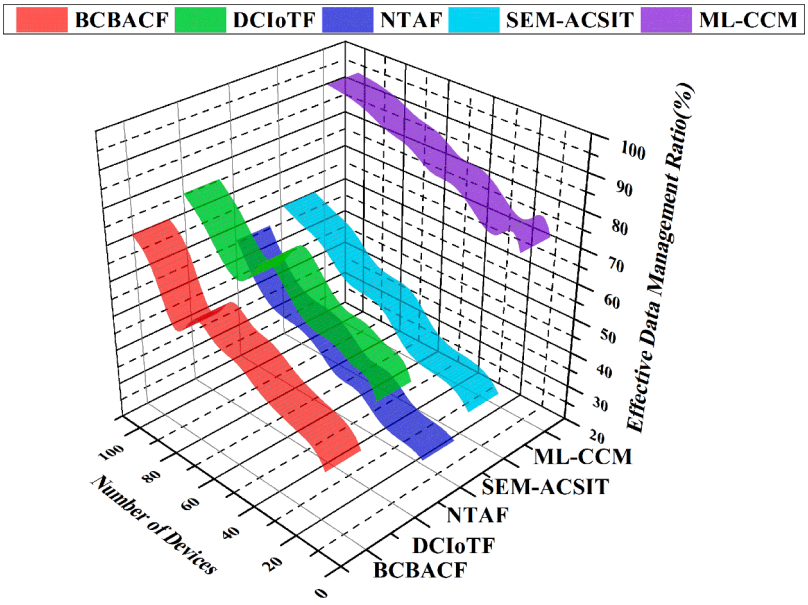

Fig. 8.. Effective data management ratio.

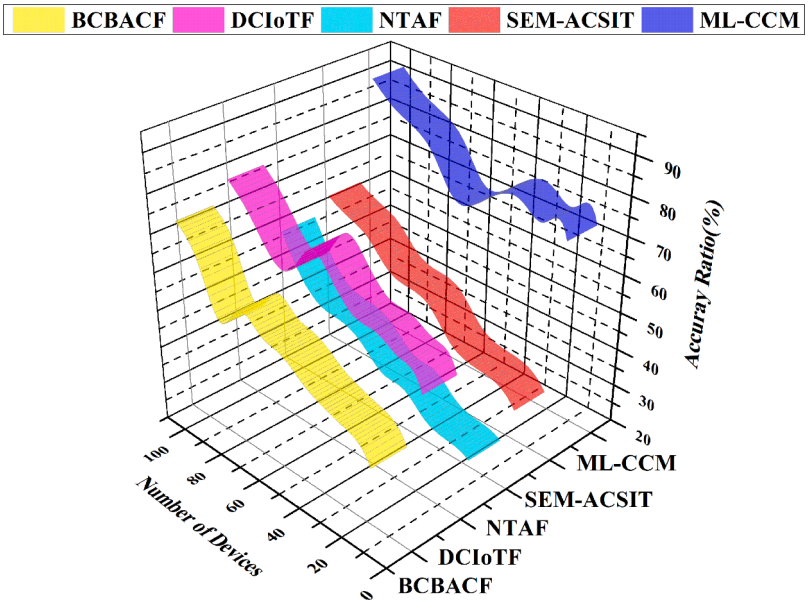

Fig. 9.. Accuracy ratio.

cloud technologies. It addresses data optimization of existing file systems for the amount of data mining applications needed and the ways in which data can be processed quickly and migrated across servers. Figure 10 shows the performance ratio.

The proposed ML-CCM improves the cloud security and reduces computational time to achieve a high data transmission rate, effective data management, accuracy, performance when compared to BCBACF, NTAF, and SEM-ACSIT.

4. Conclusion

This paper discussed cloud security based on machine learning and big data. Cloud is a technology that is delivered or accessible via the internet. Cloud computing is a device that supplies resources over the internet. Cloud computing succeeds at placing a large amount of data and its users together on a single network. Integrating machine learning methods into the cloud will thereby increase performance. This paper discusses cloud storage to provide data processing between servers with a wide group of servers having

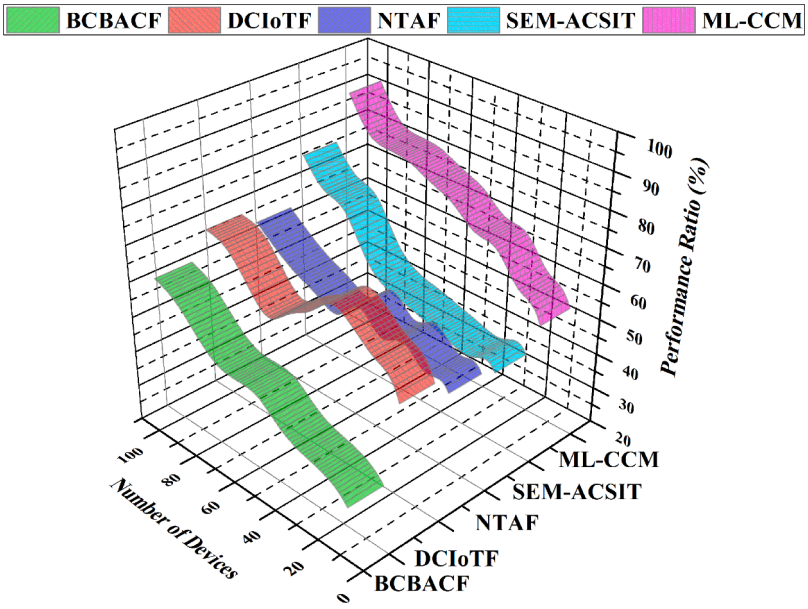

Fig. 10.. Performance ratio.

specialized connections. Hence, in this paper proposes the ML-CCM with big data to improve data management and cloud data security. Big data can be analyzed or stored in clouds of large volumes of distributed data. This article suggests an ML-CCM designed to achieve sufficient distributed mass storage and high-security safety. The mechanism is intended to encrypt all data and store data on different cloud servers without creating important overhead and latencies. The simplest approach for storing large volumes of data is cloud storage. The classification model for cloud computing focused on the feasibility of detecting machine learning algorithms. Thus, as stated at the beginning of this paper, the experimental results show that ML-CCM has a data transmission rate of 96.4%, effective data management of 94.3%, computational time of 35.2%, accuracy of 91.7%, and performance of 95.2%.

Author statement

Conception and design of study, acquisition of data: Abdul Salam Mohammad
Drafting the manuscript: Manas Ranjan Pradhan

Declaration of Competing Interest

None

Reference

[1] Yu Y, Li H, Chen R, Zhao Y, Yang H, Du X. Enabling secure intelligent network with cloud-assisted privacy-preserving machine learning. IEEE Network 2019;33(3):82–7.
[2] Manogaran G, Rawal BS, Saravanan V, Kumar PM, Martínez OS, Crespo RG, et al. Blockchain based integrated security measure for reliable service delegation in 6G communication environment. Comput Commun 2020;161:248–56.
[3] Noor U, Anwar Z, Malik AW, Khan S, Saleem S. A machine learning framework for investigating data breaches based on semantic analysis of adversary's attack patterns in threat intelligence repositories. Future Generation Computer Systems 2019;95:467–87.
[4] Manogaran G, Lopez D. A Gaussian process based big data processing framework in cluster computing environment. Cluster Computing 2018;21(1):189–204.
[5] Patil R, Dudeja H, Modi C. Designing an efficient security framework for detecting intrusions in virtual network of cloud computing. Computers & Security 2019;85:402–22.
[6] Manogaran G, Baskar S, Hsu CH, Kadry SN, Sundarasekar R, Kumar PM, et al. FDM: Fuzzy-optimized Data Management Technique for Improving Big Data Analytics. IEEE Trans Fuzzy Syst 2020.
[7] García ÁL, De Lucas JM, Antonacci M, Zu Castell W, David M, Hardt M, Alic AS. A Cloud-Based Framework for Machine Learning Workloads and Applications. IEEE access 2020;8:18681–92.
[8] Maram B, Gnanasekar JM, Manogaran G, Balaanand M. Intelligent security algorithm for UNICODE data privacy and security in IOT. Service Oriented Computing and Applications 2019;13(1):3–15.
[9] Wang Y, Meng W, Li W, Liu Z, Liu Y, Xue H. Adaptive machine learning-based alarm reduction via edge computing for distributed intrusion detection systems. Concurrency and Computation: Practice and Experience 2019;31(19):e5101.
[10] Zhu S, Saravanan V, Muthu B. Achieving data security and privacy across healthcare applications using cyber security mechanisms. The Electronic Library 2020.
[11] Alam T. Internet of Things: A Secure Cloud-based MANET Mobility Model. IJ Network Security 2020;22(3):514–20.
[12] Alabady SA, Al-Turjman F, Din S. A novel security model for cooperative virtual networks in the IoT era. Int J Parallel Program 2020;48(2):280–95.
[13] Abbasi H, Ezzati-Jivan N, Bellaiche M, Talhi C, Dagenais MR. Machine learning-based EDoS attack detection technique using execution trace analysis. J Hardw Syst Secur 2019;3(2):164–76.

[14] Qadri YA, Ali R, Musaddiq A, Al-Turjman F, Kim DW, Kim SW. The limitations in the state-of-the-art counter-measures against the security threats in H-IoT. Cluster Computing 2020:1–19.
[15] Mehrabi MA, Doche C, Jolfaei A. Elliptic curve cryptography point multiplication core for hardware security module. IEEE Trans Comput 2020;69(11):1707–18.
[16] Jolfaei A, Kant K. Data security in multiparty edge computing environments. Temple University Philadelphia United States; 2019.
[17] Deep S, Zheng X, Jolfaei A, Yu D, Ostovari P, Kashif Bashir A. A survey of security and privacy issues in the Internet of Things from the layered context. Trans Emerg Telecommun Tech 2020:e3935.
[18] Elhoseny M, Yuan X, El-Minir HK, Riad AM. An energy efficient encryption method for secure dynamic WSN. Secur Commun Netw 2016;9(13):2024–31.
[19] Darwish A, Hassanien AE, Elhoseny M, Sangaiah AK, Muhammad K. The impact of the hybrid platform of internet of things and cloud computing on healthcare systems: opportunities, challenges, and open problems. J Ambient Intell Hum Comput 2019;10(10):4151–66.
[20] Abdelaziz A, Elhoseny M, Salama AS, Riad AM. A machine learning model for improving healthcare services on cloud computing environment. Measurement 2018;119:117–28.
[21] Saravanan V, Alagan A, Woungang I. Big data in massive parallel processing: A multi-core processors perspective. Handbook of Research on Big Data Storage and Visualization Techniques. IGI Global; 2018. p. 276–302.
[22] Refaey A, Sallam A, Shami A. On IoT applications: a proposed SDP framework for MQTT. Electron Lett 2019;55(22):1201–3.
[23] Aladwan MN, Awaysheh FM, Alawadi S, Alazab M, Pena TF, Cabaleiro JC. TrustE-VC: Trustworthy Evaluation Framework for Industrial Connected Vehicles in the Cloud. IEEE Trans Ind Inf 2020;16(9):6203–13.
[24] Nguyen TA, Min D, Choi E. A hierarchical modeling and analysis framework for availability and security quantification of IoT infrastructures. Electronics 2020;9(1):155.

Abdul Salam Mohammed- School of Business, Skyline University College, Sharjah, United Arab Emirates

Manas Ranjan Pradhan- School of Information Technology, Skyline University College, Sharjah, United Arab Emirates