International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

Differential Privacy based Preserving Data on Cloud Environment

Prachee Atmapoojya
Department of Computer Application
Nit Kurukshetra Haryana, India

Utkarsh Saini
Department of Computer Application
Nit Kurukshetra Haryana, India

Rohit Patidar
Department of Computer Application
Nit Kurukshetra Haryana, India

Rishabh Gupta
Department of Computer Application
Nit Kurukshetra, Haryana

Prof. Ashutosh Kumar Singh
Department Of Computer Application
Nit Kurukshetra, Haryana

***Abstract*:- Cloud Computing offers several profits, including scalability, accessibility, and many services. But with its wide acceptance everywhere in the world, new risks and penetrability have appeared too. Storing the information on the cloud removes one's worries about space considerations, buying new storage gadgets, or managing their data, rather they're ready to operate their data any time from anywhere on the condition that they need internet access. But the rising security problem holds out against the organizations from connecting with cloud computing completely. Hence, security risks have come out because of the main drawbacks of cloud computing. The paper will include descriptions of approaches to information security and strategies used globally to ensure optimal data protection by reducing threats and risks. It is common that the information is always different from data providers for machine learning. Therefore, how to perform machine learning over cloud data from multiple users becomes a new challenge. Traditional differential privacy techniques and encryption methods are not practical for this environment. On the one hand, the information from different users is encrypted with other public keys or noises, making the computation difficult.**

Keywords - Cloud computing, machine learning, data security, privacy techniques, differential privacy, privacy preserving.

I. INTRODUCTION

Cloud computing has been envisioned as the next generation paradigm in computation [1]. The advancement of technology has taken computing to a whole new level, and one of the latest developments in this field is the introduction of cloud computing [2]. It is used for sharing ideas, power processing, storing, connectivity, and virtualization. The Internet cloud offers a vast pool of tools that help to provide on-demand applications, storage media, and sharing media. In the cloud domain, security concerns are the biggest problem and thus the most significant obstacle to the aggravation of IT-based businesses that offer on-demand services to customers. In the network process, the application phase, authentication phase, storage of information and virtualization phase, these security challenges can be visualized. These challenges are still an obstacle within the complete success path of cloud computing. One reason is that consumers and plenty of organizations keep their information on cloud databases. Hence, the main focus is that the user's information should be safe, and the critical information shouldn't drift and tamper when traveling from one place to another across the network. Thus, it is essential that Integrity, Confidentiality, and Availability of user information ought to be ensured. Another reason is that the hacking of data as cloud computing makes it possible for companies to sharpen their Development and achievement. In addition, it also hosts more users with less effort to provide access to shared resources. But security concerns or risks remain a stumbling block in the effective direction of cloud computing. There are a variety of explanations. The first explanation is that consumers and several organizations store their data on cloud storage, so the primary emphasis is that the data must be safe and that when moving across the network from one location to another, the data is not lost and tampered with. It is therefore critical that data security, availability and integrity should be maintained.
Organization: The literature review is discussed in Section II. Section III describes the background study containing service models in the cloud, deployment models, including cloud security threats. Machine learning techniques for cloud data security described in Section IV. Section V and VI described further works and the conclusion.

II. LITERATURE REVIEW

Several resources are consulted to know the fundamentals of cloud computing and storing data security on the cloud. This section provides a review of the literature to lay a foundation for discussing various data security aspects. Srinivas et al. [3] give superb insight into the essential concepts of cloud computing. Several key concepts are explored during this paper by providing samples of applications that can be developed using cloud computing and how they will help the developing world enjoy this emerging technology shared access or multi-tenancy is additionally considered together

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

with the major risks to data in cloud computing Since multiple users are using equivalent shared computing resources like CPU, Storage, and memory, etc.

Table 1. Comparison of techniques for cloud security

References	Proposed model/ work	Used Techniques	Pros	*Cons*
Li et a..[5]	privacy-preserving Naive Bayes learning scheme	Naive Bayes learning,	ε-differential privacy is utilized to preserve the privacy of every owner.	Forge & manipulation of data possible.
Ma et al.[6]	privacy-preserving deep learning model	Deep learning	Reduces storage overhead	Low classification accuracy, high computation cost
R.Yonetani, et al. [7.]	Doubly Permuted homomorphic Encryption	SVM, Semi Supervised learning technique	Reduced high computational cost	Supports one operation at a time
Li et al.[8]	Outsourced privacy-preserving classification services over encrypted data	SVM, Naive Bayes, Logistic regression, least squares	Delegation of remote server	Frequently involved interaction of users.
Li et al..[9]	privacy-preserving machine learning under multiple keys	K-NN, SVM, Random forest, Naive Bayes	Protect the privacy of dataset	High computational cost
Li et al[10]	Privacy-preserving outsourced classification scheme	Naive Bayes, Homomorphic Encryption	Used Fully homomorphic Encryption proxy technique	Storage server no longer applicable in cloud environment
Gao et al.[11]	constructing a privacy-preserving NB classifier	Naive Bayes, privacy -preserving, double blinding technique	Reduce Communication & computation overheads	Disable to protect privacy
Hesamifard et al..[12]	new solutions for running neural network algorithms over encrypted data	Crypto deep learning, homomorphic encryption, neural network	Secures private data well	Non-practical keys are used

It is a threat to not only one user but various users. There's always a risk of personal data accidentally leaking to other users [4]. Multi-tenancy exploits can be exceptionally risky because one fault within the system can allow another user or hacker to access all other data. These sorts of issues are often taken care of by wisely authenticating the users before they will have access to the info Several authentication techniques are in use to avoid multi-tenancy issues in cloud computing. A data protection scheme is proposed by Li et al. [5], which enables a trainer to train a Naive Bayes classifier over the dataset provided jointly by different data owners. ϵ-differential privacy is utilized in this scheme to preserve the confidentiality of every owner. In this approach, collusion is allowed, and adversaries can forge and manipulate the data. To solve the problem of training the model over the encrypted data under multiple keys, a privacy-preserving deep learning model (PDLM) is proposed by Ma et al. [6].

Amritpal et al. [16] suggest an improved steganography technique based on Least Significant Bit (LSB) for photos to provide better data protection. In non-adjacent and distinct pixel regions, it presents an embedding algorithm to hide ciphered messages in confines and smooth image domains. The sides are detected using an enhanced edge detection filter inside the cover image. This steganography technique means that the message bits concealed in the image are less likely to be suspected. Through usual steganography detection methods, it becomes difficult to estimate the true message length. The Proposed approach shows better leads to Peak Signal-to-Noise Ratio (PSNR) value and efficiency compared to other existing techniques.

III. BACKGROUND STUDY

Cloud Computing (CC):

Because of the technology called virtualization, cloud computing is possible. As a platform for providing services over the internet, cloud computing has recently evolved. Without management by the consumer, cloud computing is best for data storage and computing resources. Cloud Computing is subject to confidentiality because service providers can access data at any time. It can alter or delete information. There are several security obstacles to cloud computing that hinder the intense adoption of the computing model.

Service Models of Cloud Computing

Software as a Service (SaaS): On cloud servers, SaaS applications are hosted, and users can access them over the internet. Salesforce, MailChimp, and Slack are all examples of SaaS applications. Security issues with SaaS are certification, acceptance, data privacy, availability, and network safety.

Platform as a Service (PaaS): PaaS dealers offer everything necessary for creating an application having development tools, infrastructure, and operating systems over the internet. Examples include Heroku, Microsoft Azure. The main security concerns are host insecurity, privacy-aware authentication, and fault tolerance [17].

Infrastructure as a Service (I*aaS*): It provides the basic structure through the Virtual Machine, but nowadays, Virtual Machine is not that much popular. Data deletion and disputes may be resolved by determining the time for data deletion by both the client and the cloud provider. Google Compute Engine and OpenStack are among the IaaS providers.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

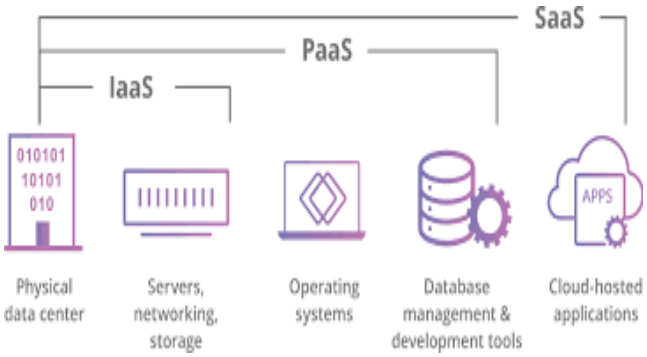


Figure 1. Service Models of Cloud Computing

Deployment Models in Cloud

PUBLIC CLOUD: According to the name, it is available for the general public. This cloud deployment is the first choice for businesses with low privacy concerns. In this third party, run your cloud infrastructure. It is hassle-free infrastructure management, and you can increase your cloud capacity according to your company's needs. Data security and privacy issues raise the concern.

PRIVATE CLOUD: From a technological perspective, there is no noticeable difference between public and private clouds since their architecture is the same. The private cloud provides several chances for customizing the infrastructure to the company's requirements. Only authorized people can access resources from the private cloud. It is costly, and that is the major disadvantage of it.

COMMUNITY CLOUD: It is a private cloud, only the difference in the number of users. Several companies can share the same cloud platform; because of this, sharing cost is reduced but more than the public deployment model.

HYBRID CLOUD: The best attribute of the other three models is the hybrid. In addition to providing security and managing strategically valuable properties, the hybrid cloud does so in a cost and resource-efficient manner. It is priced fairly [18].

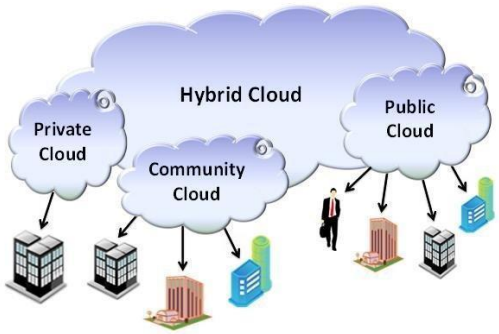


Figure 2. Deployment Models of Cloud

Cloud Security Threats

- Access Management: The primary threat of store in the cloud is not a feature of the system but rather the result of the methods used by companies.
- Data Breaches and Data Leaks.
- Data Loss: Data loss is the most considerable adversity of cloud systems. Not used daily through backups is the most significant threat of the growth of ransomware; it encrypts data and demands payment for returning the data.
- Insecure APIs: The primary tools that allow interconnection with cloud storage systems are application user interfaces (APIs). Security concerns remain with APIs [19].
- Misconfigured Cloud Storage: It is a prevalent threat in cloud systems.

Machine Learning (ML) and Cloud Security: When connected to the cloud, ML applications can be extended. The convergence of ML in the cloud is known as the "intelligent cloud." With Cloud Machine Learning, the ability of both cloud and ML algorithms can greatly increase while the cloud is first used for computing, networking, and storage. For instance, ML is currently a time-consuming task, but with the cloud computing instance, ML tasks can be activated to a great ex-task. As a result, even common statistical tools like R, Octave, and Python have also been transformed into the cloud. Machine Learning is the conceptual analysis of computation and observable models that are used by computer systems to execute a particular effort without using express headings, casual upon models, and acceptance. It is a computerized subset of reasoning. ML in the cloud is so common that any cloud can use ML in the near future.

IV. ML TECHNIQUES FOR THE CLOUD SECURITY

K-Nearest Neighbor(K-NN):- The k-nearest neighbor is the algorithm for pattern recognition. In machine learning techniques, K-nearest neighbor is a basic classifier where the classification is done by defining the nearest neighbor to query examples and then using those neighbors to evaluate the query class. K-NN is the most straightforward ML algorithm for downfall and classification problems. The dedication of this inquiry is the mechanism of information privacy order using ML's K-NN system. By using the RSA calculation, sensitive and private information demands more excellent protection and encryption. In the cloud simulation test system, the implemented data order model for cloud information security has been achieved. Cybersecurity faces many new problems, which are long-lasting. There are two key issues: botnets and intrusion detection and prevention services (IDPS). Botnets assume an important role in distributed denial of service (DDoS) attacks. DDoS success depends on the size of the botnet. For wholesale fraud and data theft, botnets are also used. An IDPS is an invention used by system and framework administrators to differentiate between obstacles. A data taxonomy approach focused on data privacy was suggested by Calderon et al. [24]. They used K-NN so that they could identify the details on the basis of their security criteria. Zardari et al.[20] proposed a data classification approach based on data confidentiality. In this system, the data is divided into two groups, e.g., sensitive and non-sensitive data. Then, the authors use the RSA algorithm to encrypt the sensitive data for protection. These kinds of methods help to evaluate the degree of protection needed for different data.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

The general formula:

d(x,y)=d(y,x)=

$$\sqrt{(y1-x1)2+(y2-x2)2+\ldots+(yn-xn)2}$$

where y1 to yn represent the attribute value for one observation and x1 to xn represent the attribute values for the other observation.

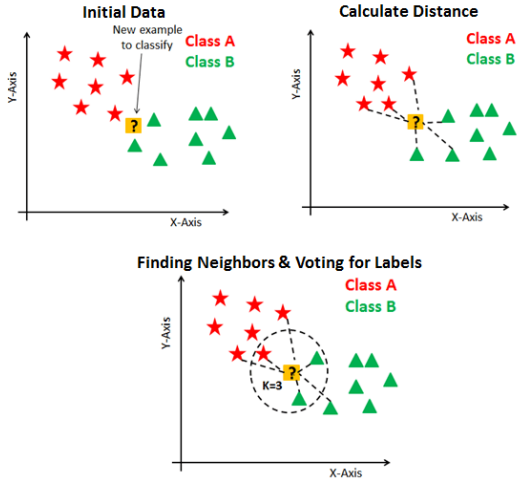


Fig. 3 K-Nearest Neighbor

In Figure 3, we want to classify the yellow new point into a class. Here are two potential classes, red and green. Start by calculating the distance from the yellow point. Here the value of K=6, so we find the distance between the yellow point to the red and green point, then find the nearest neighbors by increasing the distance. The nearest neighbors of the new yellow point are the closest in the dataspace.

Artificial Neural Networking (ANN):- A structure similar to the human brain is an artificial neural network. It involves interactions between neurons that are interrelated. Therefore, researchers tend to use an artificial neural network in problems that include computational tasks, analysis, discovering similarities, and much more. Hussin et al. [25] used this algorithm to solve security problems. Banking organizations used ANN algorithms for resolving security threats. ANNs are applied for renovating performance and learning, neural-functions. By appointing ANN, the authors tried to find out cyberattacks in MCC. They demonstrated that their implemented framework improves the detection of attacks by up to 97.11 percent accuracy. Zamzam et al. [21] explored the use of ML in mobile edge computing (MEC) resource management to solve the problem and improve performance. The authors in [21] discussed cutting-edge AI to facilitate portable edge processing resources. They differentiate the analysis into four types: reducing costs, reducing energy consumption, reducing inactivity, and restricting both latency and energy consumption.

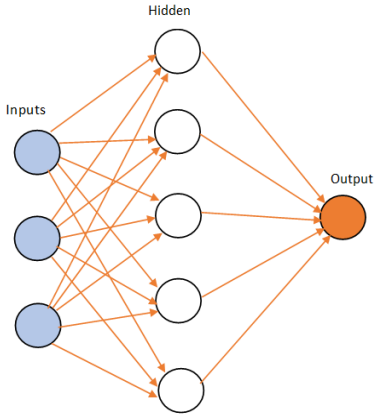
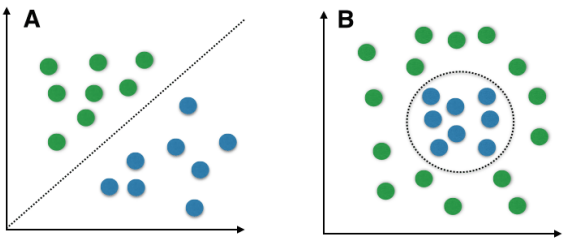


Fig. 4 Artificial Neural Network

Figure 4 shows an Artificial Neural Network, which primarily consists of 3 layers. The input layer accepts inputs in several different formats provided by the programmer. The hidden layer is in between input and output layers. It performs all the calculations to find hidden features and patterns. The input goes through a series of transformations using the hidden layer, which finally results in output conveyed using this layer.

Naive Bayes:- Naive Bayes may also be a classification technique with an assumption of independence among predictors that supports the Bayes theorem. A Naive Bayes classifier may also differently conclude that the existence of a specific function during a category is unrelated to the presence of the other function. The model of Naive Bayes is easy to form and particularly helpful for huge data sets. Naive Bayes is understood to outperform even highly sophisticated classification methods. In ML, Navies Bayes classifiers are a group of fundamental "probabilistic classifiers" that apply the naive freedom assumption of Bayes speculation between the highlights. They are simple Bayesian system models. DDoS attacks become one of the essential risks to defense. Affected by malware, PCs and several machines change into bots (or zombies). As a fundamental move to make a safe and secure environment for Cloud Computing, Hanna et al. [26] explored and analyzed how to achieve quenching security risks for Cloud Computing. An expanded protection structure that has been used in Cloud Computing for intrusion detection has been proposed. To work out the threats, the authors in [26] merged signature and inconsistency-based techniques. They used Naive Bayes and some other algorithms to increase the efficacy of the proposed method.



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

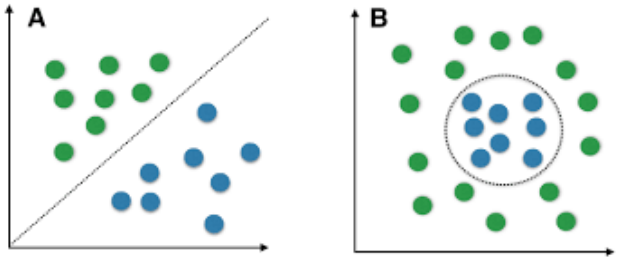


Fig 5. Naive Bayes Classifier

In figure 5, this naive Bayes model can be fit by simply finding the mean and standard deviation of the points within each label. Here the circle represents the Gaussian generative model for each label, with a larger probability closer to the center's origin.

Equation:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Support Vector Machine (SVM):- This is an algorithm for supervised machine learning that can be used for problems with classification or regression. It uses a way to rework the data called the kernel trick and then supports these transformations, finding the optimal boundary between the possible outputs. Simply put, it does some incredibly complicated data transformations, then works out the way to isolate the labels or results you have identified from your data. SVM is an ML methodology that seeks data for union and returns analysis. SVM is a technique of supervised learning that analyzes and classifies knowledge into one of two classes. As far as might reasonably be predicted, an SVM outputs a conductor of the arranged data with the edges between the two. Hou et al. [22] view edge computing systems' investigation of network security using ML to solve the issues. Their research involves the Alibaba ECS' manufactured simulation of a smart home system. Using Tor Hammer as an attacking weapon, the Wani et al.[23] researched a cloud environment and then created a dataset that identified the infraction. They used various ML algorithms for classification, including SVM, Naive Bayes, Random Forest, and found that SVM has the highest accuracy, e.g., 99.7 percent.

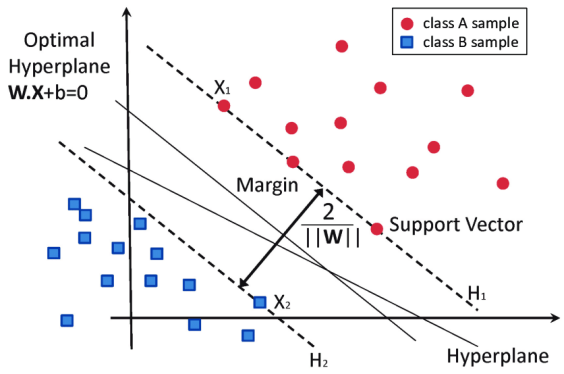


Fig. 6 Classification of data by Support Vector Machine

In figure 6, the support vector machine classification is present. In SVM, the line used to separate the classes (red and blue) is referred to as hyperplane. The data points on either side of the hyperplane closest to the hyperplane are called support vectors used to plot the boundary line (dashed line). Distance between both the dashed lines is a margin which is the sum of the shortest distance from the hyperplane to support vectors on both the sides (positive and negative).

Formula: y_i is the i-th target and $w^T x_i$-b is the i-th output.

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i - b)\right)\right] + \lambda\|\mathbf{w}\|^2,$$

V. FUTURE WORK

Directions that required more researches in future:

- Before including new progressions, a compatible study of overhead should be conducted.
- ML datasets are a series of AI datasets spanning various fields, for which security data sets related to themes, such as spam, phishing, and so on, exist.
- The CSIC HTTP dataset contains a sufficient number of directly produced Web requests and will be used to evaluate Web attack security frameworks.
- Deep neural system disclosure is also an open-source deep neural system endeavor that attempts to discern legally accessible malicious URLs, document methods, and registry keys.
- Datasets within the sample scores are also located within the acquaintance or model's registry. Documents are on it.
- Although the analysis of ML with mob sourcing has been significantly promoted in recent years.

VI. CONCLUSION

The increased use of cloud computing for data storage is undoubtedly accelerating cloud data storage methods. If data stored in the cloud is not adequately secured, it will be at risk. Security risks and assaults are the most complicated problems of cloud computing. Several machine learning algorithms are presented here, such as ANNs, K-NN, Naïve Bayes, SVM to secure cloud data. It proves that the method adopted is more precise and efficient. And also reducing user time to encrypt/decrypt various types of data (basic, confidential, and highly confidential). We have also addressed some research avenues that will need further investigations in the future.

REFERENCES

[1] Farhan, S.; Haider, S. Cloud Computing security risks. In the Internet Technology Proceedings and ICITST (Secured Transactions), Abu Dhabi, UAE, December 11-14, 2011; pp. 214-219
[2] Sun, Y., Zhang, J., Xiong, Y., & G. Zhu. (Year 2014). Data Protection and Cloud Computing Privacy. 1-9, Researchgate,.
[3] J. Srinivas, K. Reddy, and A. Qyser, "Cloud Computing Basics," Build. Infrastruct. Cloud Security., vol. 1, no. September 2011, pp. 3–22, 2014.
[4] L. Rodero-Merino, L. M. Vaquero, E. Caron, A. Muresan, and F. Desprez, "Building safe PaaS clouds: A survey on security in multi-tenant software platforms," Comput. Secur., vol. 31, no. 1, pp. 96–108, 2012.
[5] T. Li, J. Li, Z. Liu, P. Li, and C. Jia, "Differentially private naive Bayes learning over multiple data sources," Inf. Sci., vol. 444, pp. 89–104, 2018
[6] X.Ma, J.Ma, H. Li, Q. Jiang, and S. Gao, "PDLM: Privacy-preserving deep learning model on cloud with multiple keys," IEEE Trans. Serv. Comput., to be published, doi: 10.1109/TSC.2018.2868750

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

[7] R. Yonetani, V. Naresh Boddeti, K. M. Kitani, and Y. Sato, "Privacy Preserving visual learning using doubly permuted homomorphic encryption," in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2040–2050.
[8] T. Li, Z. Huang, P. Li, Z. Liu, and C. Jia, "Outsourced privacy-preserving classification service over encrypted data," J. Netw. Comput. Appl., vol. 106, pp. 100–110, 2018.
[9] P. Li, T. Li, H. Ye, J. Li, X. Chen, and Y. Xiang, "Privacy-preserving machine learning with multiple data providers," Future Gener. Comput. Syst., vol. 87, pp. 341–350, 2018.
[10] Hourani, H.; Cloud Computing: Legal and Security Issues: Abdallah, M. In International Courts Computer Science and Information Technology (CSIT) Meeting, Helsinki, Finland, June 13–14, 2018; Pp. 13 to 16.
[11] Selamat, N.; Ali, F. Using a machine learning algorithm to compare malware detection techniques. Indones. J. Electr. Eng., Eng. Computing. Sci. 2019, 16th, 435th. [The CrossRef].
[12] Shukla, S.; Maheshwari, H. Cloud Computing Security Discerning Risks. Oh, J. Comput. Theor. Theor. Nanosci. Nanosci. 16, 255-261 2019. [The CrossRef]
[13] Yuhong, L.; Yan, S.; Jungwoo, R.; Syed, R.; Athanasios, V. A study of threats to security and privacy in Cloud Computing: Future Paths and Solutions. J. Comput. Sci. Sci. Eng., Eng. 2015, 9, 119-133.
[14] Als olami, E. Security threats and legal issues related to Cloud based solutions. Int. J. Comput. Sci. Netw. Secur. 2018, 18, 156–163.
[15] Shamshirband, S.; Fathi, M.; Montieri, A.; Chronopoulos, A.T.; Palumbo, F.; Pescapè, A. Computational, A. in Mobile Cloud Computing Environments, Intelligence Intrusion Detection Techniques: Analysis, Taxonomy, Issues of open science. Uh, J. Inf. Uh, Secur. Appl. Appl. 1–52, 2019.
[16] Kaur, K., and Zandu, v (2016). A stable model of data classification in the machine learning approach to cloud computing. Computer Science International Journal of Advanced Research, 13-22.
[17] Butt, Umer A., Muhammad Mehmood, Syed B. shah, Rashid Amin, M. W. Shaukat, Syed M. Raza, Doug Y. Suh, and MD. j. Piran. 2020. "A review of Machine Learning Algorithms for Cloud Computing Security." *Electronics* 1397, no. 19 july 2020 (August): 25. 10.3390/electronics9091379.
[18] Zou, Caifeng, Huifang Deng, and Qunye Qui. 2013. "Design and Implementation of Hybrid Cloud Computing Architecture Based on Cloud Bus." *Researchgate* 6, no. December 2013 (December): 6. 10.1109/MSN.2013.72.
[19] Kazim, Muhammad, and Shao Y. Zhu. 2015. "A Survey on top security threats in cloud computing." *International journal of Advanced Computer Science and Applications.* 6 (3): 5. 10.14569/IJACSA.2015.060316.
[20] Zardari, M.A.; Jung, L.T.; Zakaria, N. K-NN classifier for data confidentiality in cloud computing. In Proceedings of the International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, Malaysia, 3–5 June 2014; pp. 1–6.
[21] Zamzam, M.; Tallal, E.; Mohamed, A. Resource Management using Machine Learning in Mobile Edge Computing: A Survey. In Proceedings of the Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt, 8–10 December 2019; pp. 112–117.
[22] Hou, S.; Xin, H. Use of machine learning in detecting network security of edge computing systems. In Proceedings of the 4th International Conference on Big Data Analytics (ICBDA), Suzhou, China, 13–15 March 2019; pp. 252–256.
[23] Wani, A.; Rana, Q.; Saxena, U.; Pandey, N. Analysis and Detection of DDoS Attacks on Cloud Computing Environment using Machine Learning Techniques. In Proceedings of the Amity International Conference on Artificial Intelligence (AICAI), Dubai, UAE, 4–6 February 2019; pp. 870–87
[24] Calderon, R. The Benefits of Artificial Intelligence in Cybersecurity. Available online: https:// digitalcommons.lasalle.edu/ecf-capstones/36 (accessed on 19 July 2020).
[25] Elzamly, A.; Hussin, B.; Basari, A.S. Classification of Critical Cloud Computing Security Issues for Banking Organizations: A Cloud Delphi Study. Int. J. Grid Distrib. Comput. 2016, 9, 137–158. [CrossRef].
[26] Hanna, M.S.; Bader, A.A.; Ibrahim, E.E.; Adel A.A. Application of Intelligent Data Mining Approach in Securing Cloud Computing. Int. J. Adv. Comput. Sci. Appl. 2016, 7, 151–159.