Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 03, March-2022

Data Privacy Preservation Model using Noise Concept in Cloud

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, India

Sakshi Chhabra
Department Of Computer Application
Nit Kurukshetra
Haryana, India

Ashutosh Kumar Singh
Department of Computer Application
Nit Kurukshetra
Haryana, India

I. ABSTRACT

PRACHEE ABSTRACT

The authoritative, valuable data need to be shared with numerous parties in a cloud environment for analysis, storage, and data utilization. However, Establishing security and maintaining privacy while effectively exchanging data across several parties has become inextricably linked tasks. We offer a methodology that allows several parties to safely exchange their data for different objectives using noise and machine learning classifiers. In the proposed model, different noise is added and then shared before transferring the data. In this model, we use Gaussian noise with various classifiers. This model minimizes the risk of data leakage. The experimental findings show that the suggested model's efficiency provides good accuracy.

Index Terms - cloud computing, data security, machine learning, data privacy.

• II. INTRODUCTION

In the field of information technology, cloud computing is a new paradigm. Cloud computing technology is already in widespread usage and is being developed in numerous forms. Cloud computing influences the people, processes, and technology of a company. Despite the benefits of cloud computing, such as productivity, flexibility, ease of setup, and overall cost savings in IT [1] it also poses a threat to privacy and confidentiality. "Not all forms of cloud computing provide the same security and privacy threats. Some predict that in the future, most of the computing activity that is currently carried out solely on machines owned and managed locally by people will migrate to the cloud" [2]. Instead of connecting to different servers on business premises, users connect to the cloud, which seems to be a single entity to the user, as with traditional computing. Public-private partnerships are becoming increasingly popular as a way for governments to address the different requirements of their population while also ensuring that these services are of high quality. Cloud computing may potentially act as a conduit for collaboration between the public and commercial sectors. In such circumstances, an external organization may be involved in delivering Cloud Services, with partial control over data storage, processing, transfer, and the application of privacy rules. [3]. Cloud computing has essential ramifications for personal information privacy as well as commercial and government information secrecy. When utilizing the cloud to store information, perhaps the primary choice is to make the issue of whether to utilize an outsider cloud supplier or build an interior hierarchical cloud. National security data or extremely confidential future product information, for example, are often too essential data kept on a public cloud. This type of information can be highly sensitive, and the effects of revealing it on the internet can be catastrophic [4]. This form of data can be highly sensitive, and exposing it to the public cloud can have significant implications. In such instances, using an internal organizational cloud to store data is strongly recommended. This technique will help secure data by creating on-premises data usage regulations [5]. However, because many firms are not compliant, total data protection and privacy cannot be guaranteed. On the other hand, several firms lack the knowledge to apply all layers of protection to sensitive data, resulting in insufficient data security and privacy. This work aims to secure cloud data sharing by using multiple categorization algorithms to create noise. It delves into the potential threats to data in the cloud, as well as the data-protection methods employed by various service providers [6].

Organization: The literature review is discussed in section III. Section IV introduces problem formulation, including system model, and proposed model, Section V introduces performance evaluation which includes experimental setup and accuracy of the classification model. This section describes different noise with different classifiers in other datasets. Section VI is the conclusion part.

• III. LITERATURE REVIEW

Roy et al. [7] proposed a Privacy Protection Decentralized Information Flow Control system (PPDIFC). In this scheme,

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 03, March-2022

the symmetric key encryption technique was utilized to encrypt the data. However, the proposed scheme is not able to decrypt data after encryption. The Key Management Interoperability Protocol (KMIP) was introduced in [8], to address the previous problem, which improved data security while also lowering costs across different products. Lo et al. [9] proposed an efficient confidentiality-based cloud data storage framework (ECCSF). This framework increases the processing time, confidentiality, and probity through data classification based on classified data using Transport Layer Security (TLS), App Service Environment t(AES), and Secure Hash Algorithms (SHA). This framework included an automatic data classification and algorithms such as RSA, Elliptic curve cryptography, and the asymmetric public key that provide a higher degree of security and confidentiality [10]. The proposed cloud model reduces the overhead and processing time necessary to protect data by employing various security techniques with varying key sizes to give the data the appropriate level of secrecy. It offers a variety of encryption techniques at various levels to protect data. But this should have a more efficient framework by considering other aspects. Yonetan et al. [11] proposed a Doubly Permuted Homomorphic Encryption (DPHE) for learning visual classifiers securely over distributed data. Support vector machine (SVM) and semi-supervised learning techniques are used in this model. This model reduces the high computational cost, DPHE can efficiently encrypt any sort of sparse high-dimensional data, making it useful for a variety of jobs. It can only support one operation at a time, though. Gao et al. [12] suggested a Privacy-preserving Naive Bayes classifier (PPNBC) that is immune to a simple but difficult-to-detect attack., which we call the substitution-then-comparison (STC) attack. Naive Bayes and double-blinding techniques were the key techniques. This model reduced communication and computational overhead. But it was not very effective for protecting privacy. Zibouh et al. [13] proposed a multi-key encryption plan to get the mystery of touchy cloud information. A fully homomorphic encryption algorithm was used in this model. This strategy searches for information on a cloud worker and positions the discoveries dependent on the information's consistency and level of encryption [14]. The method has the advantage of protecting the data without exposing the user's sensitive details. Li et al. [15] proposed an information security system that requires an educator to prepare an Innocent Bayes classifier over a dataset given altogether by various information proprietors. $\boldsymbol{\varepsilon}$-Differential insurance is utilized in this framework to ensure every proprietor's security. The agreement is allowed in this system, and adversaries can possibly misrepresent and abuse records. Li et al. [16] proposed over encrypted data, outsourced privacy-preserving categorization services are available. The proposed schemes allow the model owner (MO) to train its learning model and obtain the optimal coefficient vector based on the dataset owned by the data owner (DO) using SVM, Naive Bayes, Logistic Regression, and least square methods with the help of cloud server (CS), trusty decryption server (TDS), and key conversion server (KCS). This architecture allowed for remote delegation, but it also required a lot of user input. Hesamifard et al. [17] proposed a new method for operating neural networks on encrypted data. In this approach, deep learning cryptography, homomorphic encryption, and neural networks are employed. This model secures private data but it uses non-practical keys which is not beneficial for the model. Moghaddam et al. [18] proposed a half breed encryption model dependent on arrangement ordering, traits and time sensitive. The bulk of data classification is based on attributes. To determine the protection between the rings, a hybrid ring was used [19]. These tightly secured rings re-encrypt data to shield themselves against unwanted entry, time-based queries, data owner requests, and user revocation. The examination of the results uncovers that the blend ring model improves the steadfastness and, in this manner, the introduction of information security applications. Ma et al. [20] proposed a privacy-preserving deep learning model (PDLM) to resolve the issue of preparing the model over encoded information with a few keys. It has a low classification accuracy and a high calculation cost, but it has a lower storage overhead.

• IV. PROBLEM FORMULATION

This section explains the entities involved in the model and their assigned roles, as well as all of the potential dangers that may arise in the protocol, the annoyance, and the layout goals.

SYSTEM MODEL

This model comprises the four entities Data Provider (*DP*) , Cloud Platform *(CP*), Cloud Service Provider (*CSP*), Classifier (*CF*), and Cloud Server (*SC*) that are described as follows.

1. *DP*: Data Provider is an entity that is registered to provide information and requesting services from *CP*. It provides the information that they have to store. Since it's widely believed that a data provider cannot leak its very own data[21], it can also leak the data of the other owner; therefore, the data provider is seen as an untrustworthy organization.
2. *CP* : A cloud platform refers back to the running machine and hardware of a server in an Internet-primarily based totally facts center. It collects all the data from *DO*, transforms it, performs certain computations over it, and compares the calculated accuracy outcomes for secure sharing among the model [22]. *CP* trains the obtain information using machine learning algorithms. The cloud platform is divided into two clouds in our system paradigm, with cloud1 including Cloud Storage and Cloud Service Provider and cloud containing the Classifier (*CF*).
3. *Gaussian Noise*: Gaussian noise, often known as the Gaussian distribution, is statistical noise with a probability density function (PDF) equal to that of the normal distribution. It is named after Carl Friedrich Gauss[23]. To put it another way, the noise's possible values are distributed Gaussian-style.
4. A Gaussian random variable $\boldsymbol{\mathcal{Z}}$ probability density function is given by:
   PG(z)=1/σ√2π e^(-(z-μ)2/22)
   Where $\boldsymbol{\mathcal{Z}}$ represents the gray level, $\boldsymbol{\mu}$ the mean gray value and $\boldsymbol{\sigma}$ its standard deviation.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 03, March-2022

V. PROPOSEDMODEL

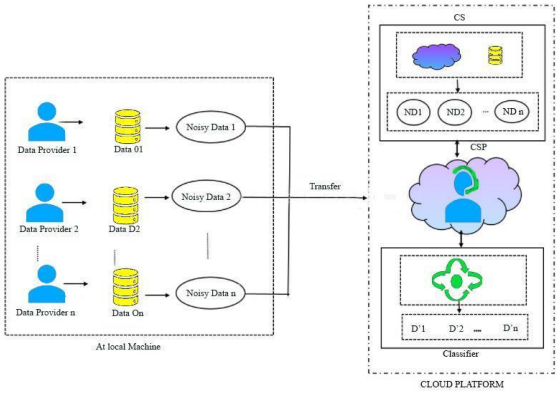

Fig. 1 Proposed Model

The model's architecture is shown in Fig. 1. It displays the entities involved, their communication, as well as the critical blocks and the required flow between them. Let the Data Provider $DP = \{DP_1, DP_2, \ldots, DP_n\}$ owns the data $D = \{D_1, D_2,\ldots,D_n\}$ where the data object $D_i \in D$ is an independent data object that can be of any form and size. *DP* has to share data *D* with registered parties such as Cloud Platform (*CP*) and different users for storage, computation, performance optimization, and the addition of the noise into the data provided by the *DP*. Different noises are added to the data then it transfers to the Cloud Service Provider (CSP) who arranges the data. In *CP* the Classifier (*CF*) here to check the Accuracy of the data is shared with , to make the data private and secure, $DP_1, DP_2, \ldots DP_n$ procures the different types of noisy data $ND = \{ND_1 , ND_2 ,\ldots,ND_n \}$ by applying a different technique, *CSP* shares $ND_1, ND_2,\ldots,ND_n$, which it stores in Cloud Storage (*CS*) and transmits to Classifier*(CF)* for computation. The entity *CSP* interacts with CF, which collects CM's results and delivers them back to CSP. Afterward, CSP delivers the noisy data to the corresponding entity $ND_1 , ND_2 ,\ldots,ND_n$ .The entity *CF* uses $ND_1, ND_2,\ldots,ND_n$ to perform computations such as accuracy, precision, and so on, and produces a Classification Model (*CM*). Any question can be rendered via *CSP* by $ND_1, ND_2,\ldots, ND_n$. *CSP* interacts with CF, which receives the results from *CM* and forwards them to *CSP*. After all the computations are done, the accuracy is measured correctly with all the noises then CSP chooses the appropriate data, which is more secure. Following that, *CSP* sends the obtained results to the corresponding entity $D_1, D_2,\ldots,D_n$. that we have to store it in the cloud.

VI. PERFORMANCE EVALUATION

1. *Experimental setup*

A series of experiments have been conducted over four different datasets Hill Valley, Old wives, Housing, and Bank marketing taken from Kaggle to train classification models using machine learning algorithms. The three different classifiers, K-Nearest Neighbor(K-NN), Naive Bayes and Support Vector Machine (SVM) have been used to train classification models over the training data. These experiments are performed on Intel Core i5-8250u CPU @ 1.60GHz 1.80GHz eight-core processor with Windows operating system, 8GB RAM machine using Python 2.73 for adding noise and machine learning.

2. *Accuracy of the classification model*

8/10 of the data from the entire dataset is utilized as training data, while the remainder is used as testing data.

Both actual and noisy data are used in the machine learning process. We utilized the Gaussian noise and randomly generated noise mechanisms with a privacy level of 0.1 to produce the noisy data. The noisy data's output is compared to the actual data to find variations. As you can see in table no. 1. Furthermore, a comparison is made between Gaussian noise data to determine which is best. The Classification Accuracy (CA) is calculated once the CM result is determined using the testing data. Over actual, Gaussian noisy data and also depicts the comparison among Old wives, Hill valley, Housing, and Bank Management datasets for Naive Bayes, SVM, and KNN and classifier, respectively. CA of noisy data is lower than that of real data in all three classifiers due to the noise addition, although CA is almost equal for noisy data and also provides stronger security than genuine data. Furthermore, in all four classifiers, the gaussian noisy data outperforms out of the two noisy added data. In descending order, the performance of datasets and classifiers is Housing, Bank marketing, Old wives and Hill valley and SVM, K-NN, and Naive Bayes, respectively. For all three classifiers, the Housing dataset beats the other three datasets. The SVM classifier outperforms the other two for actual and Gaussian noise data, whereas Naive Bayes beats the other classifiers for Random noise data. The SVM classifier beats the other two classifiers in our suggested model in aggregate. As a result, performance improves.

Table 1: Results Comparison between proposed model (with noise) and w/o noise

SR.NO.	Data Set	Classifiers	Accuracy (%)		Precision (%)		F1-Score (%)		Recall (%)	
			Without noise	With noise	Without noise	With noise	Without noise	With noise	Without noise	With noise
1.	Hill Valley	K-NN	57	52	57	53	57	50	57	52
		Naïve Bayes	52	52	55	55	45	45	52	52
		SVM	52	52	55	55	43	43	52	52
2.	Old Wives	K-NN	50	52	50	52	49	52	50	52
		Naïve Bayes	43	43	42	42	41	41	43	43
		SVM	57	55	58	55	56	53	57	55
3.	Housing	K-NN	86	89	91	88	88	89	86	89
		Naïve Bayes	72	71	86	86	78	77	72	71
		SVM	93	93	87	87	90	90	93	93
4.	Bank	K-NN	89	89	87	87	88	88	89	89

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 03, March-2022

	marketing	Naïve Bayes	87	87	87	87	87	87	87	87
		SVM	89	89	88	87	87	87	89	89

VI. CONCLUSION

This paper presented a new model called data privacy preservation model using noise concept in cloud, which provides adequate data protection in a real-world cloud environment. In this model we use different machine learning algorithms such as K-NN, Naive Bayes, and SVM classifiers for training and testing data. The noise (such as Gaussian noise) is added to data before sharing the data, which is more secure. It demonstrates that the chosen procedure is more precise and effective. Additionally, it reduces the time it takes for users to encrypt and decode various sorts of data (basic, confidential, and highly confidential). This paper examined various data privacy strategies for data protection in cloud computing environments, focusing on information storage and usage within the cloud to build trust between cloud service providers and customers.

REFERENCES

[1] Sultan, N. (2010), Cloud computing for education: A new dawn? International Journal of Information Management, Volume 30, Issue 2, April 2010, Pages 109–116.
[2] Gellman, R. (2009), Privacy in the Clouds: Risks to Privacy and Confidentiality from Cloud Computing, World Privacy Forum, USA.
[3] Ruiter, J. and Warnier, M. (2011). Privacy regulations for cloud computing, compliance and implementation in theory and practice. In Gutwirth, S., Poullet, Y., de Hert, P., and Leenes, R., editors, Computers, Privacy and Data Protection: an Element of Choice, chapter 17, pages 293–314. Springer.
[4] French, Robert M. "Catastrophic forgetting in connectionist networks." *Trends in cognitive sciences* 3.4 (1999): 128-135.
[5] Al-Issa, Yazan, Mohammad Ashraf Ottom, and Ahmed Tamrawi. "eHealth cloud security challenges: a survey." *Journal of healthcare engineering* 2019 (2019).
[6] Vacca, John R., ed. *Cloud computing security: foundations and challenges*. CRC Press, 2016.
[7] Roy I, Ramadan HE, Setty STV, Kilzer A, Shmatikov V, Witchel E. "Airavat: Security and privacy for MapReduce," In: Castro M, eds. Proc. of the 7th Usenix Symp. on Networked Systems Design and Implementation. San Jose: USENIX Association, 2010. 297.312.
[8] "OASIS Key Management Interoperability Protocol (KMIP) TC", http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=kmip
[9] Lo"ai Tawalbeh, Nour S. Darwazeh, Raad S. Al-Qassas and Fahd AlDosari (2015)."A Secure Cloud Computing Model based on Data Classification", First International Workshop on Mobile Cloud Computing Systems, Management, and Security, Elsevier pp. 1153 – 1158,2015.
[10] Athena, J., and V. Sumathy. "Survey on public key cryptography scheme for securing data in cloud computing." *Circuits and Systems* 8.3 (2017): 77-92.
[11] 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.
[12] C.-z. Gao, Q. Cheng, P. He, W. Susilo, and J. Li, "Privacy-preserving naive Bayes classifiers secure against the substitution-then-comparison attack," Inf. Sci., vol. 444, pp. 72–88, 2018.
[13] Zibouh, O., Dalli, A. and Drissi, H., 2016. CLOUD COMPUTING SECURITY THROUGH PARALLELIZING FULLY HOMOMORPHIC ENCRYPTION APPLIED TO MULTI CLOUD APPROACH. Journal of Theoretical & Applied Information Technology, 87(2)
[14] Vianney, D. Maria Manuel, M. Aramudhan, and G. Ravikumar. "Effective binary cuckoo search optimization based cloud brokering mechanism on cloud." *2017 International Conference on IoT and Application (ICIOT)*. IEEE, 2017.
[15] 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
[16] 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.
[17] Shukla, S.; Maheshwari, H. Cloud Computing Security Discerning Risks. Oh, J. Comput. Theor. Theor. Nanosci. Nanosci. 16, 255-261 2019
[18] F. Moghaddam, M. Vala, M. Ahmadi, T. Khodadadi, and K. Madadi Pouya, "A reliable data protection model based on re-encryption concepts in cloud environments," 2015 IEEE 6th Control and System Graduate Research Colloquium (ICSGRC), pp. 11–16, 2015.
[19] Zhao, Han, et al. "Topological hybrid silicon microlasers." Nature communications 9.1 (2018): 1-6.
[20] Ma, Xindi, et al. "PDLM: Privacy-preserving deep learning model on cloud with multiple keys." IEEE Transactions on Services Computing 14.4 (2018): 1251-1263.
[21] Desale, Ms Madhuri. "NOVEL METHOD FOR DELEGATABLE PROOFS OF STORAGE TO PREVENT DATA LEAKAGE IN CLOUD."
[22] Vyazovkin, Sergey, et al. "ICTAC Kinetics Committee recommendations for performing kinetic computations on thermal analysis data." *Thermochimica acta* 520.1-2 (2011): 1-19.
[23] Lawrence, Andy. "Combining Many Factors: The Gaussian Distribution." *Probability in Physics*. Springer, Cham, 2019. 69-92.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)