# Upgrade-Data Security in Cloud by Machine Learning and Cryptography Techniques

**Adnaan Arbaaz Ahmed, M.I.Thariq Hussan, Venkateswarlu Bollapalli**

***Abstract:*** ***The term cloud computing is referred as the shared pool of customizable computer resources and high quantity services which can easily be provisioned with less management endeavours via internet. It transfigured the mode associations reach IT, which enables them to be more perceptive, launch new business models, and minimise the IT costs. These technologies are to be administrated in an interdisciplinary collection of architectures, characterized into various deployment and service models, and can synchronize with other related technologies. The widespread issues with cloud computing are security, reliability, data privacy and anonymity. Cloud computing provides a way to share distributed sources and services that are owned by different organizations or sites. Since it shares distributed resources via network in open environment that results in security issues. In this paper, our aim is to upgrade the security of data in the cloud and also to annihilate the difficulties related to the data security with encipher algorithm. In our proposed plan, some key services of security like authentication and cryptographic techniques are assigned in cloud computing environment.***

***Index Terms:*** ***Cloud computing, cryptography, data classification, decryption, encryption.***

## I. INTRODUCTION

Cloud Computing is a transpiring virtual distributed environment that utilizes the ideas of sharing, power processing, storing, connectivity, and virtualization. Communicating through broad network i.e., Internet cloud facilitate a large pool of resources, storage media and sharing media that helps to supply on-demand services. This will help the end-users to follow the ideas of distribution, security, elasticity and isolation. Security issues are the foremost difficult problems in cloud domain and the vital hurdle for aggrandize of IT based companies that provide users on-demand services. These security issues can be visualized at application phase, network phase, authentication phase, authorization phase and virtualization phaseThere are two reasons for the security concerns in the cloud computing:

Nowadays, many are storing their data on Cloud database. So, the main attention is on the safeguard of user's data and the vital information shouldn't get tampered when shifting across the network. It is necessary that Integrity, Confidentiality and Availability of user data must be ensured.

The unauthenticated user tries to access the authenticated user's data.

**Revised Manuscript Received on October 05, 2019**

**Adnaan Arbaaz Ahmed**, Scholar, Department of Information Technology, Guru Nanak Institutions Technical Campus, Hyderabad, India.

**Dr.M.I.Thariq Hussan**, Professor & Head, Department of Information Technology, Guru Nanak Institutions Technical Campus, Hyderabad, India.

**Venkateswarlu Bollapalli**, Assistant Professor, Department of Information Technology, Guru Nanak Institutions Technical Campus, Hyderabad, India.

We can apply cryptographic algorithms in cloud servers to solve these threats. However when a user is revoked, usage of a single cryptographic algorithm is not adequate to assure the security of data and to manage the Access Control methods in Cloud Computing environment. For data security these techniques are applied on encryption. Encrypting complete data can turn to be very expensive in terms of memory as well as for time. So, to solve this problem it would be better if we first separate our sensitive data and then apply encryption algorithms. It would address reliable results if we reorganize the data depending to its confidentiality level. In machine learning field, the data classification is a method of distinguishing the category of unclassified data sample set with the help of build classifier. The classifier is constructed by constructing a training set of familiar data samples. To develop an appropriate classifier, large range of justified training data samples are required. This advancement invites a new paradigm of services where data classification is offered by servers in a cloud to its various clients/users. Specifically, the server will process the data automatically and hence, categorise the client's data samples present on remote private servers. However un-trusted third party-servers can access the private data. Moreover, any vital detail or training data set specifications may not be disclosed by the servers even if it provides the classification services to its client. Thus, a mechanism that ensures the privacy of the server's training set and client data samples is required. Therefore, a decipher model is essential to forestall the invalidated user from accessing the enciphered information as well as to produce authentic keys for validated users.

## II. INTENTION OF PROPOSED SYSTEM

- To design a system which can render privacy and security in terms of cloud storage.
- To collocate an encipher depended environment for protecting the confidential data on the Cloud and to architect how the user and the storage service provider performs operations on enciphered data.
- To design an architecture where a user can be able to store the data on the Cloud in the encrypted form which is not vulnerable. (In normal cases in Cloud Computing, the data which is stored on the Cloud server is not such assured, it is legible to anyone who can authenticate to access, leaving the data vulnerable).
- To design a method where a user can retrieve the data in the enciphered manner and can be deciphered by the user at the

- same site by using asymmetric key cryptography, both the keys functioning at the user level.

## III. PROPOSED MODEL

Data classification is the task of identifying data sets with respect to its data value. These values are based on usage consumption of data by its users and restrictions on access control methods. KNN (K-Nearest Neighbors) technique is used in machine learning artificial intelligence that helps to categorize the class of unorganized data by using build classifier. It is constructed by employing a training set of familiar data samples. In this proposed work, the modified Ensemble Learning Technique is used to enhance the execution of existing KNN technique. Ensemble learning method comprises a set of different models are group together to improve the prediction and stability power of any model. It has two levels: base level-0 and Meta level-1 as shown in Fig. 1. At base level a no. of algorithms can run i.e. AdaBoost and bagging algorithm. At Meta level, also known as decision making algorithm, random forest tree is used. To utilize the training sets of data which is given by KNN model, the training set is computed with Euclidean distance function. To reduce the computational density, we improved the basic algorithm of ensemble learning.

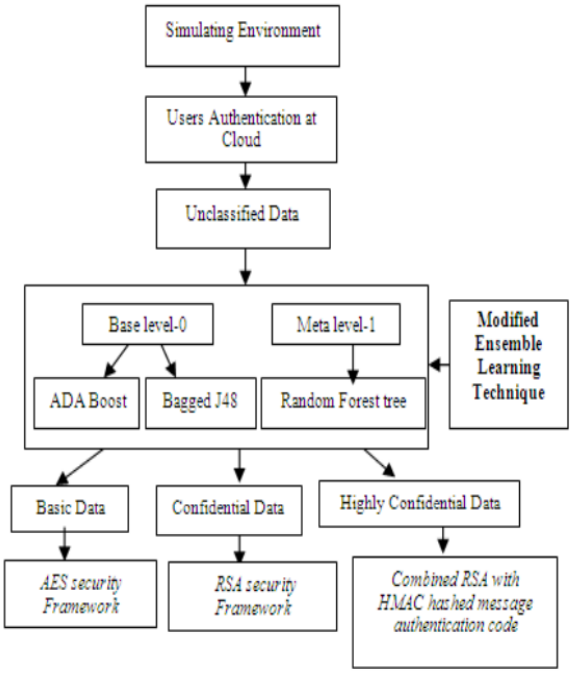


**Fig. 1. Proposed Model**

## IV. WORKING OF THE ALGORITHM

The classifier is evaluated using a cross validation (K-fold) technique. Each layer is trained as: The dataset can be split into two sets: training set and testing set.

- For each base level layers classifiers are generated using Training set.
- The predictions generated by the base level is assembled and forwarded to meta level. At meta level, validation process is performed and provides the final predicted results.
- Finally, by using complete layer's datasets (not the use of simplest out of k folds of the dataset) we can train each layers of the training set.
- Now to secure highly confidential data we combine RSA encryption technique with HMAC. The HMAC( Hashed Message Authentication Code) is a cryptographic checksum. It is stored at local machine as well as appended with the cipher text and sent to the cloud. The HMAC+ cipher text must be matched with the HMAC checksum i.e already stored at the local machine of user. The hacker finds it difficult to guess the HMAC checksum while hacking the data. Thus, results in increasing the security of user's HC data by using RSA encryption technique.
- HMAC is calculated by following formula: HMAC(Key,msg)=HMAC(KeyXORout)||H((Key XORin)||msg))

where Key is the original key, m is the message that needs to be authenticated.

## V. SIMULATION ON CLOUD COMPUTING

The cloud simulation as shown in Fig. 2 is used for the proposed model that solves the problem of confidentiality and organization of data. Firstly we run the simulation then data centres' are building. The virtual machine manager (VMM) can be used to handle the VMs and for allocating VMs to cloudlets i.e. the cloud task. Authentication process is performed so that, only the authenticated users have the access to the data. Then classification KNN is combined with modified ensemble learning technique. This will improve the prediction capabilities and accuracy of existing KNN.

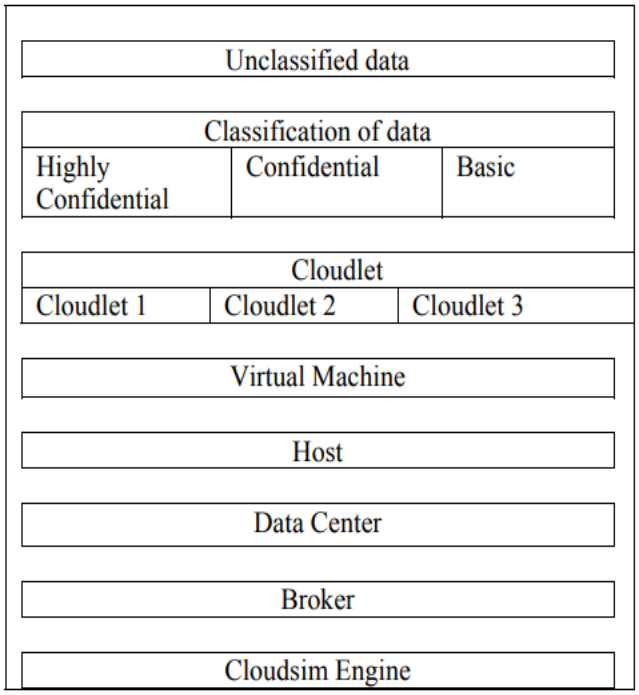


**Fig. 2. Proposed Simulation Environment**

## VI. ALGORITHM FOR DATA CLASSIFICATION

**Step 1: To select Beta and Meta layers**

```
I/p: o.ds: DS, folds: Int
DS←o.ds
AoC as classifiers array that consist AdaBoost and Bagged
Decision Tree
For lr= 0 to 2 do:
For each fold in folds do:
If lr ≠2 do:
AOC←Trn_S.Lr (lr, t.s, folds)
In_New←Classify (test-set,AoC)
Adding In_New to ds [lr+1]Else
Trn_S.Lr (lr, ts, folds)
Lr = lr + 1
Trn-S.Lr (lr 0, o-ds)
```

**Step 2: Single Layer Classification**

```
Trn-S.Lr I/p: Lr
No: Int, ds: DS, folds: Int
O/p: Scr-Ds: DS
Scr DS ← emp Grp
 For each fold in folds do:
 B.C (Lr No, ts)
For each In. in ts
Produce probabilities-vector by applying In. on current
layer's classifiers.
build a new In. from probabilities-vector
Add the new In. to Scr-DS
Ret Scr-DS
```

Where: o.ds: Original dataset, DS: Dataset, Int: Integer, AoC: Array of classifier, Lr: Layer, ts: Training set, In: Instances, Scr: Successor.

## VII. PROPERTIES AND DESCRIPTION OF CLOUD IAAS MODEL

Before starting the simulation, it is necessary to set the characteristics of IaaS(Infrastructure as a Service). The properties are as follows:

**Table 1: The properties of IaaS are associated with the data centres. These data centres are assigned to VMs.**

| Data centre ID | Storage limit | RAM in MB | Architecture of data | Operating System | Bandwidth |
|---|---|---|---|---|---|
| 0 | 100000000 | 4096 | x84 | Linux | 10000 |
| 1 | 100000000 | 4096 | x84 | Linux | 10000 |

## VIII. ENCRYPTION ALGORITHM

This algorithm aims to enhance the ongoing encipher methods by collaborating substitution and transposition ciphers. Both the encipher methods depends on alphabet for the cipher text whereas in our algorithm, the text is converted into respective ASCII code value of each letter. In the ongoing encipher methods, the key value scales between 0 to 25 and key may be string (combination of alphabets), but in this algorithm, key value scales between 0 to 255 to encrypt the user's data in the cloud. As the user has no authority over his data after logging out, the enciphered key gives the primary authentication to the user.

The algorithm is as follows.

- Compute the total number of characters (C) in the text excluding blank spaces.
- Form a square matrix (SXS>=C) by converting the text into corresponding ASCII code and apply the same from left to right direction.
- Split the matrix into three matrices as upper, lower and diagonal matrix.
- Read the values in each matrix in opposite direction.
- Complete the encryption using three different keys one for each matrix as K=K1, K2, K3.
- Alter the enciphered values into the matrix in the same direction.
- Read the enciphered values column-wise in the same order by assigning a key K4.
- Convert the resulted code into the corresponding character value generated by K4.

## IX. RESULTS

The results of our system are the comparison of existing work with the proposed work has been shown. According to the following analysis, it has been observed that the system proposed in this research work is giving improved and reliable results. For higher degree of confidentiality and security different cryptographic techniques are required. The achievements of our system can be evaluated by following parameters i.e., Accuracy, Classification details, Encryption Time and Decryption Time, Error Rates.

### 9.1. Accuracy(A)

It is the measure of properly organized instances compared to total number of properly organized and improperly organized instances. TP: True Positive; TN: True Negative; FP: False Positive; FN: False Negative.

$$A=\frac{TP+TN}{TP+FP+TN+FN}$$

### A. Classification Details

Precision and recall: Both are used for evaluating execution capability and in data analysis field like retrieving data and data mining. Precision measures the correctness and recall measures the completeness of data.

i. $\text{Precision}=\frac{TP}{TP+FP}$

ii. $\text{Recall}=\frac{TP}{TP+FN}$

f-measure: The Harmonic Mean of Precision and Recall is termed as f-measure.

$$\text{f-measure}=\frac{2*Recall*Precision}{Recall+Precision}$$

### B. Encryption Time

Encoding information or any message in such a way that only authorized party can be able to access it.

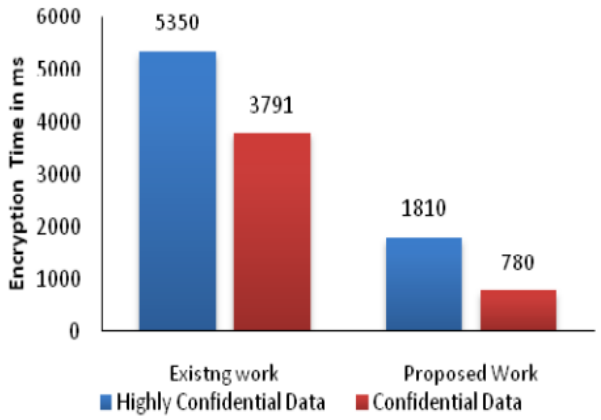


**Fig. 3. Comparison of encryption time between proposed and existing model**

### C. Decryption Time

The time required for reconverting the encrypted/cipher data or message into its original/plain form.

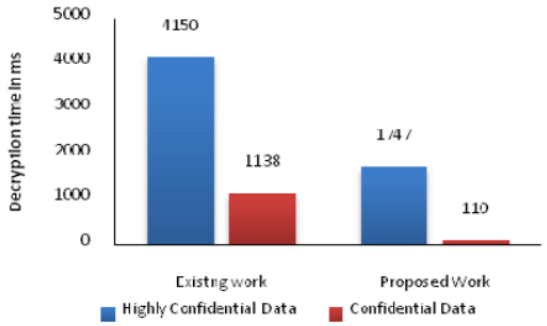


**Fig. 4. Comparison of decryption time between proposed and existing model**

### D. Error rate

The measurement of the effectiveness of a communication channel is calculated by

$$\text{Error rate} = \frac{\textit{no of errornous units of data}}{\textit{total data}}$$

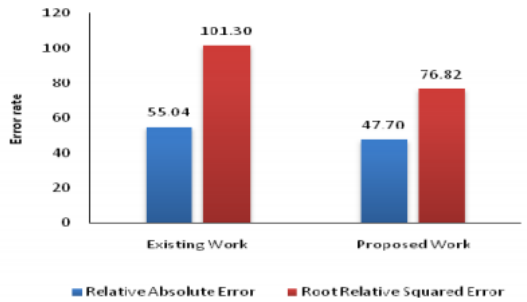


**Fig. 5. Comparison of error rates between proposed and existing model**

## X. ADVANCED TECHNIQUES USED IN SECURING DATA IN THE CLOUD

Below are the three main techniques that are involved in securing data in the cloud. These usually perform with the option of blocking of confidential data. They are:

A. Data Masking
B. Secure Logic Migration and Execution Technology
C. Data Traceability Technology

### A. Data Masking

It is an approach for developing not only architecturally analogous but also spurious format of a company's information that can be used for wide applications viz., software testing and user training. The need of this technique is to shield the original information by providing an operational substitute for various functions where the original information is not needed.

In this technique, the structure of data persists unchanged, but the values are altered. The information may be changed in multiple formats, which includes encipherment, character redistribution and character or word replacement, which results in making reverse engineering unendurable.

Some providers of these products includes Dataguise, Oracle, IBM, and Informatica.

### B. Secure Logic Migration and Execution Technology

For intimate/privileged data which cannot be exposed out from the company, even formed by concealing certain aspects of the data, by solely defining the security level of data, the information gateway transfers the cloud-based application to the in-house sandbox (A protected program execution environment that halts fraudulent tampering of data) for execution. The sandbox will block the pre-authorized access to data or networks, so even applications transferred from the cloud can be safely executed. Application providers are able to confirm if there is any inappropriate use of the data because the performance of applications is recorded.

### C. Data Traceability Technology

The information gateway tracks all the flows and their content can be checked. Information entering into and leaving out of the cloud is monitored by the information gateway. Data traceability technology makes use of the logs obtained on data traffic as well as the features of the related text to visualize the data used in the cloud.

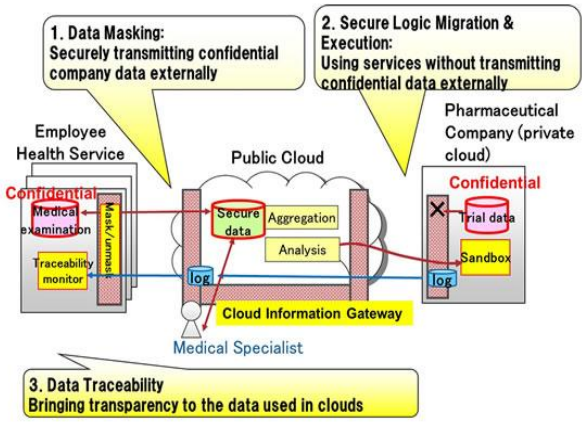


**Fig. 6. Scenario for the usage of newly developed data gateway**

## XI. CONCLUSION

This paper aims at achieving data confidentiality, data access control management and provides an automatic classification of data in cloud. On the contrary, the existing system was lacking in providing automatic classification of data. In this paper the functionality and performance of existing KNN technique is improved with modified Ensemble Learning technique.

The data will be automatically classified by the machine on the basis of data security parameters. Also to emphasis more focus on highly confidential data HMAC (Hash-based Message Authentication Code) function has been appended with the existing RSA encryption algorithm. The proposed system proves to be more accurate and economical. And also saves the user time for encrypting/decrypting different classes of data (basic, confidential and highly confidential).

## REFERENCES

1. http://en.wikipedia.org/wiki/Cloud_computing
2. http://www.roughtype.com
3. F. F. Moghaddam, M. Vala, M. Ahmadi, T. Khodadadi and K. Madadipouya, "A reliable data protection model based on re-encryption concepts in cloud environments," Proceedings of 6th IEEE Control and System Graduate Research Colloquium, pp. 11–16, 2016.
4. N. Surv, B. Wanve, R. Kamble, S. Patil and J. Katti, "Framework for client side AES encryption technique in cloud computing," IEEE International Advance Computing Conference, pp. 525–528, 2015.
5. V. K. Pant, J. Prakash and A. Asthana, "Three step data security model for cloud computing based on RSA and steganography," International Conference on Green Computing and Internet of Things, pp. 490–494, 2015.
6. L. Tawalbeh, N. S. Darwazeh, R. S. Al-Qassas and F. AlDosari, "A secure cloud computing model based on data classification," Procedia Computer Science, vol. 52, no. 1, pp. 1153–1158, 2015.
7. R. Shaikh and M. Sasikumar, "Data classification for achieving security in cloud computing," Procedia Computer Science, vol. 45, pp. 493–498, 2015.
8. M. A. Zardari, L. T. Jung and N. Zakaria, "K-NN classifier for data confidentiality in cloud computing," International Journal of Computing and Information Science, 2014.
9. Dr. A. Padmapriya and P. Subhasri, "Cloud Computing: Reverse Caesar Cipher Algorithm to Increase Data Security", International Journal of Engineering Trends and Technology, vol. 4, Issue 4, pp. 1067-1071, 2013.
10. Quist-Aphetsi Kester, "A Hybrid Cryptosystem Based on Vigenere Cipher and Columnar Transposition Cipher", International Journal of Advanced Technology & Engineering Research, vol. 3, Issue 1, pp. 141-147, 2013.
11. Randolph Barr, "How to gain comfort in losing control to the cloud".

## AUTHORS PROFILE

**Adnaan Arbaaz Ahmed** pursuing Bachelor of Technology in Information Technology, Guru Nanak Institutions Technical Campus, Hyderabad. He is specialized in Networking, Cyber Security, Cloud Computing, Cryptography, DarkWeb, DevOps, Server configuration and maintenance of Linux and Microsoft servers and Data Science. He is crowned with the titles 'Technophyle' and 'Codebrary' and has 3 International publications. He received 'Research Ratna Award' in 2019 by Rula Awards for the research on 'Malware Detection by using Machine Learning Techniques'. He is an active member of Quora contributing his answers to the open source community. His research work mainly focuses on the security issues that are faced in Cloud Management.

**Dr.M.I.Thariq Hussan** has 18 International and 1 National journal publications. He has presented papers in 31 International/National conferences and attended 36 Seminars/Workshops/FDP/QIP. He has published 2 books titled 'System Analysis and Design' and 'Operating Systems'. He has received 'Innovative Technological Research (Communication) and Dedicated Professor Award' from Innovative Scientific Research Professional Malaysia (India Chapter). He also received 'Best Teacher Award-2018' from Institute for Exploring Advances in Engineering accredited by EA-JAS. He has filed 2 patents titled 'Vision Based Safety Seat Belt Monitoring System' and 'Intelligent Detect and Control Cybercrime Device'. He has qualified ESOL certificate in English by Cambridge University. He is a life member of ISTE and nominee member of CSI for knowledge exchange and enhancement.

**Venkateswarlu Bollapalli** pursued Bachelor of Technology (Information Technology) from RVR&JC college of Engineering and Master of Technology (Computer Science and Engineering) from QIS college of Engineering and Technology. He is an Assistant Professor at Guru Nanak Institutions Technical Campus, Hyderabad. He is specialized in Compiler Design, Design and Analysis of Algorithms, Network Security, Data Mining, Discrete Mathematics and Artificial Intelligence. He has published 3 international journals. He is a life member of ISTE and his research work mainly focuses on Machine Learning and Soft Computing.