



SECURING DATA CONFIDENTIALITY IN CLOUD COMPUTING USING IMPROVED BOOSTING TECHNIQUE

Rubal
Department of CSE
Guru Nanak Dev University,
Regional Campus, Jalandhar, India

Sheetal Kalra
Department of CSE
Guru Nanak Dev University,
Regional Campus, Jalandhar, India

Abstract—Data security is a challenging issue in today's fast paced and competitive environment. Many techniques are used to secure data from security attacks that can potentially damage business critical data and services. But deciding data security technique without understanding the security needs is not a technical approach. Before applying any security, it is best to know the sensitivity levels of the data. This paper presents a data classification approach based on improved boosting technique with hybrid encryption. Adaboosting algorithm is used as a classification technique which is modulated in the cloud virtual environment. The aim to use Adaboosting is to classify the data based on their security requirements. The data is classified into sensitive data and non-sensitive (public) datasets. Then hybrid encryption that is RSA with blowfish algorithm is used to encrypt the sensitive dataset to keep it secure.

Keywords— Adaboosting, Blowfish, CloudSim, Decision Stump, RSA, Virtual machine, Weak learners

I. INTRODUCTION

In today's era of competition, organizations are under big pressure to improve efficiency and transform their IT processes to achieve more with less. Businesses need to reduced time-to-market, better agility, higher availability and reduced expenditures to meet the challenging business requirements. All these challenges are addressed by new computing style called cloud computing. It is a complete computing as a service rather than computing as a product, which can be delivered to clients over the internet through big data centers in a Cloud Computing environment. The cost of the resource management is more than the actual cost of the resources. So, it is better to get the required resources by renting instead of purchasing one's own resources.

Cloud Computing is a model for distributed computing in which resources and services can be accessed by scaling up or down as per consumer demands. Cloud Computing providers

typically charge customers on a pay-per-use model. The biggest benefit in moving to Clouds is that developers need not to make large capital expenditures on hardware for deploying their ideas in various Internet based services resulting in operational costs only using Cloud services [1]. These services are available for all users without any data bias. Cloud computing is a technique using which users can browse and select relevant cloud services, such as compute, software, storage, or combination of these resources, via a portal. It helps the organizations and individuals to use IT resources as a service over the network at reduced total cost of ownership [2]. Cloud computing enables consumers to pay only for the resources they use for example CPU hours used, amount of data stored and transferred.

Although there is growing acceptance of cloud computing, but the cloud service consumers have been facing some security related challenges. Storing the data on cloud without knowing its security needs is the main problem [3]. For example, it is not feasible to encrypt 200GB data using security keys if only 10% of data is confidential. Therefore, there should be a mechanism to classify the data according to its sensitivity level. AdaBoosting algorithm is used as a classification technique to classify the data into two categories: sensitive data and non-sensitive data, which is modulated in the cloud virtual environment.

A. AdaBoost Algorithm

The AdaBoost algorithm was developed by Freund and Schapire in 1995. The AdaBoost combines the performance of several weak classifiers in order to obtain a strong classifier. Freund and Schapire have explained in their work that if a weak classifier is a little better than chance, then in that case the error of the final classifier decreases exponentially. This algorithm performs a linear combination of weak classifiers represented as $h(x)$ and results into a strong classifier. AdaBoost is adaptive only for this reason that consequent classifiers of those instances misclassified by previous

classifiers are found weaker. Noise data and the outliers are highly hazardous to AdaBoost.

$$f(x) = \sum_{t=1}^T \alpha_t ht(x)$$

where $ht: X \rightarrow \{1, -1\}$ and $\alpha_t \in \mathbb{R}$

A weak classifier is a very simple model that has slightly better accuracy than a random classifier, which has 50% accuracy on the training dataset [5]. The set of weak classifiers is built iteratively from the training datasets over thousands of iterations. At each iteration, the examples are reweighted in the training data according to how well they are classified. Weights are calculated for the weak classifiers based on their classification accuracy.

Each classifier is voted using the assigned weight D . Classifier with less error rate is given more weight to its vote. This training cycle keeps on repeating. The weight of classifiers those voted for an object of a class is added. Finally, the final class is decided on the basis of higher resultant weight and is introduced as the predictive class for that object. AdaBoost assigns values are based on the error of each weak classifier α to each of the classifiers. This is given by ϵ

$$\epsilon = \frac{\text{number of incorrectly classified examples}}{\text{total number of examples}}$$

Is given by α

$$\alpha = \frac{1}{2} \ln \frac{1 - \epsilon}{\epsilon}$$

After calculating α , update the weight vector D , that means decrease the weight of the examples that are correctly classified and increase the weight of the misclassified examples. D is given by

$$D_i^{(t+1)} = \frac{D_i^{(t)} e^{-\alpha}}{\text{Sum}(D)}$$

If correctly predicted and

$$D_i^{(t+1)} = \frac{D_i^{(t)} e^{\alpha}}{\text{Sum}(D)}$$

If predicted incorrectly.

AdaBoost is a powerful practically successful classification algorithm with wide range of applications, such as in biology, computer applications, and speech processing domain. In contrast to SVM and other similar powerful classifiers, AdaBoost is capable to produce similar classification results involving much reduced tweaking of parameters or settings.

The factors to be chosen by user are:

- (1) The weak classifier that can function possibly best to solve the classification problem in hand.

- (2) The number of boosting rounds to be involved during the training phase.

Finally, the weak classifier operating best at the particular boosting round is selected by the AdaBoost algorithm. Ada Boost uses decision stump as a weak learner. A decision stump is a machine learning model containing a simple one-level decision tree. One-level decision tree means it consists of one internal node which is referred as root and is immediately connected to the leaf nodes. A decision stump makes decision taking a single input feature into consideration.

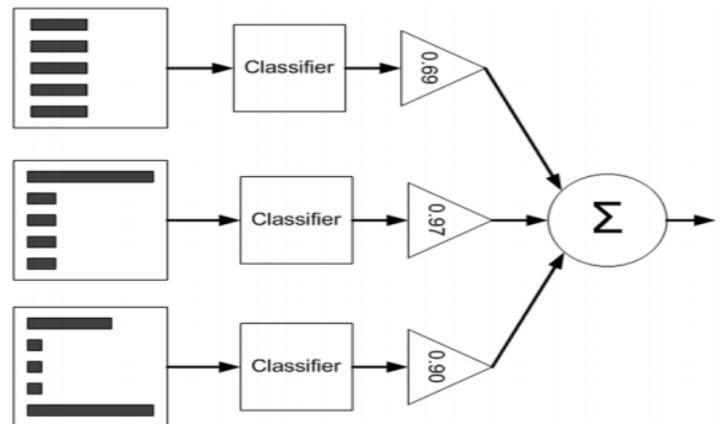


Figure 1: Schematic representation of AdaBoost with the datasets on the left side.

Several variations may exist depending on the type of the input feature. For instance, for nominal features a stump may be built in two ways. One may contain a leaf for each possible feature value or the other may contain two leaves, where one leaf corresponds to some chosen category, and the other leaf corresponds to all the other categories. In machine learning techniques like bagging and boosting, decision stumps are used as components referred as "weak learners" or "base learners".

The best stump is found the same way a node is approached in a decision tree. All the possible features which are subjected to split on are searched; further all possible thresholds are searched for each one. In classification problems, each node in a decision stump tree represents a feature in an instance which is to be classified, and the branch represents a value that the node can take. Instances are classified from the root node and sorting them based on their feature values [6]. AdaBoost proceeds by iterations, in each iteration the weak learner focuses on hard data points that previous weak learners cannot handle well. In each round, it increases the weights of misclassified data, and decreases the weights of correctly classified data. This algorithm classifies the datasets into sensitive and non-sensitive classes.



B. Hybrid Encryption Technique

After classification, next step is to encrypt the sensitive part of data. Encryption and decryption are the two methods to guarantee data security. Encryption is the process to convert original data in to some other anonymous structure using a key which cannot be identified by anyone. Decryption defines the recovery of original data from the encrypted image. Combination of RSA and blowfish algorithm leads to hybridized technique of data encryption. It involves:

- 1) Generation of digital signatures using RSA.
- 2) These digital signatures are then used as keys in the rounds of blowfish algorithm to encrypt the required data.

II. RELATED WORK

M. A. Zardari et al. [1] proposed a data classification cloud model to solve data privacy and security issue in cloud computing environment. Due to which the confidentiality level of data is increased and this model is demonstrated to be more user friendly in terms of cost and memory. It is also useful for the cloud services providers.

An FGPA [5] implementation for image interpretation provides architecture in 2012 for objects classification which is based on adaptive boosting algorithm. The architecture uses the color and texture as input and discriminate the objects in a scene. It uses optimized technique for reusing the texture feature modules and becoming an easy model for discrimination process. It takes into account for real time systems. This technique of reusing allows to increase the information of the object model without decrease the performance or excessively increase the area used on the FPGA. This classifies 30 dense images per second.

P. S. Rawat et al. [7] proposed a tool and then published a report on this in 2012. This Tool provides the sustainable and fault tolerant environment for experimental assessment of cloud based application like social sites and scientific work flow. But in this tool Real time scenario is missing and can be extended to real time cloud.

In 2010, Praveen Ram C et al. [8] have proposed a mechanism for gaining maximum security by leveraging the capabilities of a processor called a cryptographic coprocessor. Since data security is a major task in cloud, with this open source cryptographic coprocessor, the functionality will get increased and cost will get reduced. But the disadvantage is that the extra cost has to be paid up by the consumer, because of the coprocessor that performs encryption and decryption in a

secured way. This Cryptographic coprocessor may cost somewhere from several hundred to thousand U.S dollars.

Dawn Song et al. [9] proposed a new cloud computing paradigm- data protection as a service DPaaS, which is a collection of security primitives offered by a cloud platform. This cloud computing model enforces data security and privacy and provides the information of privacy to data owners, even in the presence of potentially compromised or any malicious applications. Adding protections to a single cloud platform can immediately benefit thousands of applications and hundreds of millions of users.

Thair Nu Phyu [10] gives a comprehensive review of different classification techniques in data mining. The goal of classification algorithms is to generate more certain, precise and accurate system results. The main issue with this approach is lack of security during classification. An improved version of KNN combined with Genetic Algorithm (GA) has been proposed to improve its classification performance by N. Suguna, and Dr. K. Thanushkodi [11]. In this complexity of KNN is reduced and there is no need to consider the weight of the samples. It improves the classification accuracy also.

Data classification is a process that allows individuals and organization to categories different kinds of data according to the security needs. Data is categorized into three levels according to the confidentiality degree: basic, confidential and highly confidential. Manual classification can also be there in which user can specify the confidentiality level of data [13]. Data can also be categorized according to its critical value that is how frequently it must be accessed. Data with larger critical value will be stored on a faster media whereas data that are less critical are stored on slower media. Automatic data classification with cryptographic algorithms provides more data confidentiality and data integrity. Security challenges are still there in public cloud and data outsourcing is a big challenge in it. In data outsourcing, the user cannot be sure about the location of data, security of the stored data and accuracy of data transaction [15].

In 2014 Jin Li et al. [16] proposed a secure dekey which was secure and Less overhead was provided for the realistic environments but the confidentiality can be enhanced further in that case. V. Varadharajan and U. Tupakula[17] have proposed a security architecture which provides a security as a service model that a cloud provider can offer to its multiple tenants and customers of its tenants. security architecture and performance analysis of the proposed system is enhanced.

N. chandel and S. Mishra[18]have demonstrated about a new framework where they create a secure and normal cloud environment of user choice and then apply data mining techniques in the cloud environment and check the



performance. In that framework privacy rules are preserved using cloud environment.

B. Poornima et al. [19] proposed a paper on Improving cloud security by improved HASBE using hybrid encryption scheme introduced the HASBE conceive for recognizing scalable, flexible, and fine-grained get access to command in cloud computing. The algorithm not only carries aggregate attributes due to flexible ascribe set combines, but furthermore accomplishes effective purchaser revocation because of multiple worth assignments of attributes.

III. SECURITY ISSUES IN EXISTING SYSTEM

Nowadays consumers are using cloud services to avoid IT infrastructure purchasing and maintenance cost. Critical data requires privacy and continuous monitoring of its access. A large amount of data can be stored on cloud. If the data moves to a cloud model other than a private cloud, consumers could lose absolute control of their sensitive data. So, security of the critical data is always a challenging threat for quality of services and it also stops the users to adopt cloud services [15]. In cloud storage, all kinds of data are stored on servers through two storage methods. The first method is to store data on servers without encryption. The second method is to encrypt the received data and store them on cloud servers. These methods of data storage can face the issue of data confidentiality. In a cloud environment, data are stored on remote servers which are not physically known by the consumer and there is always high chance of confidentiality leakage. This paper basically focuses on the threat of confidentiality of data when it is stored on a cloud [14]. When a dataset is being stored to cloud, it passes through a security mechanism, such as data encryption without understanding the needs of data or directly being stored on servers without encryption. All data have different sensitivity levels. So, storing data into a cloud without understanding its security needs is not a valid and technical approach.

A. Existing security approach with K-NN classifier

Many techniques are used to encrypt data for security, but to encrypt complete data is very expensive and time consuming tasks and are not affordable by small and medium enterprises and cloud service providers. It would be better to separate the sensitive data from public data and then encrypt only sensitive part. In many cases, k-NN algorithm is used to classify the data instances. It is an easy technique with low complexity. It is basically a machine learning technique used for pattern recognition, prediction and classification. It is an iterative technique to classify unclassified datasets into user specified classes, k. In this technique selecting the value of k is importance because using larger k may include some not so

similar datasets. On the other side, using smaller k may exclude some potential candidates of datasets. In both of these cases, accuracy of classification will decrease [11]. The k-NN algorithm must compute the distance and then sort all of the training datasets distance at each prediction. Another issue with this approach is combining the class labels. It uses the simplest method to take a majority vote, but this can be a problem if the nearest neighbors vary widely in their distance. In this algorithm we specify class for confidential attributes.

The K-NN algorithm has a set of n labeled training sets. This can be represented as:

$$D = \{d_1, d_2, d_3, \dots, d_n\}$$

Where D is the set of total samples and 'n' is the number of data items in the set. The set of n labeled samples can be represented as:

$$D = \{d_1, d_2, d_3, \dots, d_n | C\}$$

Where C is a specified class for target values [1].

k-NN is a supervised machine learning technique in which the classes are already defined whereas in unsupervised learning classes are not already defined but classification is done automatically based on the similarity between items. One another issue with this model is that it uses the RSA encryption technique which is less secure and slow in encryption and decryption of large amount of data [13]. RSA contains a public and a private key therefore the solution is to hybridize the encryption technique.

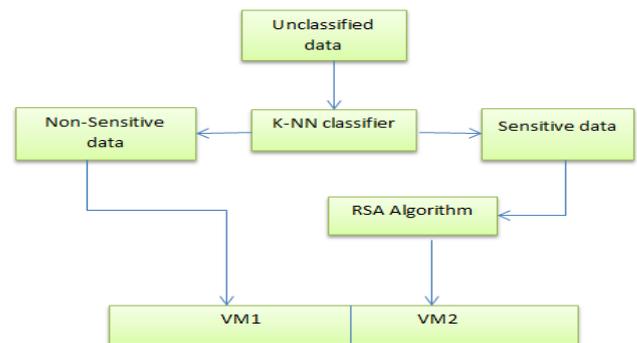


Figure 2: Model with K-NN classification

IV. PROPOSED WORK WITH IMPROVED ADABOOST TECHNIQUE

In the proposed model the limitation of previous model is overcome by using improved boosting technique and RSA with blowfish algorithm to encrypt sensitive datasets that can contain very important data of individuals and organizations. On the other hand side, Public data which is also considered as unrestricted data is directly allocated to virtual machine



without encryption. Unrestricted data contains information which is not critical for an individual and organizations. After that VM will process the data and store it on cloud.

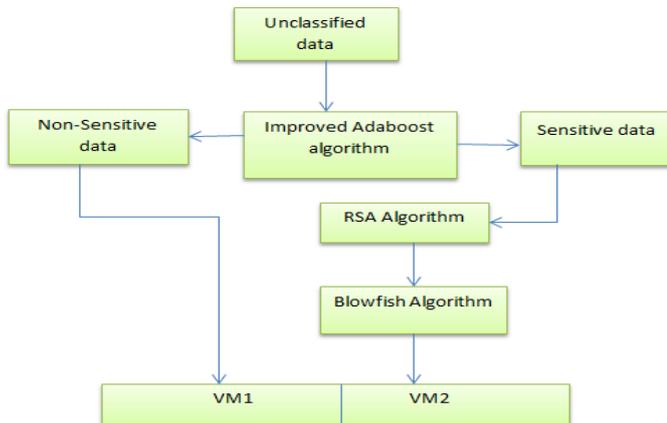


Figure 3: Proposed model

V. METHODOLOGY OF PROPOSED SYSTEM

- 1) Creating virtual cloud environment which includes creation of data centres, brokers, virtual machines, cloudlets and hosts.
- 2) Classifying the dataset by using improved Boosting technique i.e Apply the enhanced AdaBoost algorithm for classification which will
 - a) Replace the weak learner of AdaBoost with hybrid classifier that contains AdaBoost algorithm and Naïve Bayes which are being hybridized on the basis of average of their probabilities.
 - b) Add more decision making conditions while calculating the class for model prediction i.e. on the basis of error rate adds more weight to class that will give better class for the prediction. This will classify the data into 2 parts: sensitive and non- sensitive
- 3) Securing the sensitive data by using hybrid encryption algorithm which contains hybridization of RSA and blowfish algorithm. Digital signature for the sensitive data is generated using RSA. Then these digital signature works as a key to Blowfish algorithm for encrypting the data.

VI. ALGORITHMS

1. AdaBoost Algorithm Pseudo code

a) Model generation

Assign equal weight to each training instance

For 't' iterations:

Apply learning algorithm to weighted dataset, and store resulting model

Compute model's error e on weighted dataset

If $e = 0$ or $e \geq 0.5$:

Terminate model generation

For each instance in a dataset:

If classified correctly by model:

Update the weight of the tuple by

$$W \text{ of the tuple} * \frac{\text{Error of } M(i)}{1 - \text{Error of } M(i)}$$

Normalize weight of all instances by

$$W \text{ of the tuple} * \frac{\text{sum of old weights}}{\text{sum of new weights}}$$

b) Classification

Assign weight = 0 to all classes

For each of the t (or less) models:

For the class this model predicts

Add $-\log e / (1-e)$ to this class's weight

Weight of the classifiers vote

$$W_i = \log \frac{\text{Error of } M(i)}{1 - \text{Error of } M(i)}$$

Return class with highest weight.

2. Enhanced AdaBoost Algorithm Pseudo code

a) Model generation

Assign equal weight to each training instance

For 't' iterations:

Apply hybridized average probabilities of Naïve Bayes and Decision stump learning algorithm to weighted dataset, store resulting model

Compute model's error e on weighted dataset

If $e = 0$ or $e \geq 0.5$:

Terminate model generation

Else if $e > 0.1$ or $e <= 0.3$:

For each instance in dataset:

If classified correctly by model:

Update the weight of the tuple by

$$W \text{ of the tuple} * \frac{\text{Error of } M(i)}{1 - \text{Error of } M(i)} + 0.1$$

Normalize weight of all instances by

$$W \text{ of the tuple} * \frac{\text{sum of old weights}}{\text{sum of new weights}}$$

Else if $e > 0.3$ or $e < 0.5$:

For each instance in dataset:

If classified correctly by model:

Update the weight of the tuple by



$$W \text{ of the tuple} * \frac{\text{Error of } M(i)}{1 - \text{Error of } M(i)} + 0.2$$

Normalize weight of all instances by

$$W \text{ of the tuple} * \frac{\text{sum of old weights}}{\text{sum of new weights}}$$

Else

For each instance in dataset:

If classified correctly by model:

Update the weight of the tuple by

$$W \text{ of the tuple} * \frac{\text{Error of } M(i)}{1 - \text{Error of } M(i)}$$

Normalize weight of all instances by

$$W \text{ of the tuple} * \frac{\text{sum of old weights}}{\text{sum of new weights}}$$

b) Classification

Assign weight = 0 to all classes

For each of the *t* (or less) models:

For the class this model predicts

Add $-\log e / (1-e)$ to this class's weight

Weight of the classifiers vote

$$W_i = \log \frac{\text{Error of } M(i)}{1 - \text{Error of } M(i)}$$

Return class with highest weight.

VII. EXPERIMENTAL RESULTS

The proposed algorithm is implemented through simulation package CloudSim. Java language is used for development and implementation of proposed algorithm for data classification and cloud security. In this, three parameters are evaluated and compared the results of proposed and the previous technique

- ✓ Accuracy
- ✓ Time Taken
- ✓ Encryption Time

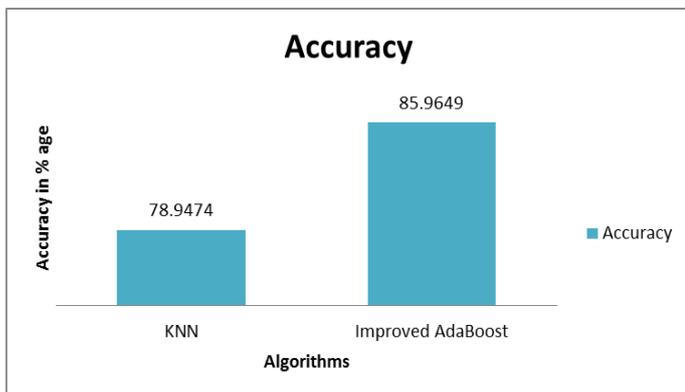


Figure 4: Accuracy Comparison

It is clearly seen from the above figure that the accuracy of the Improved Bagging Algorithm is much better than the existing algorithm KNN.

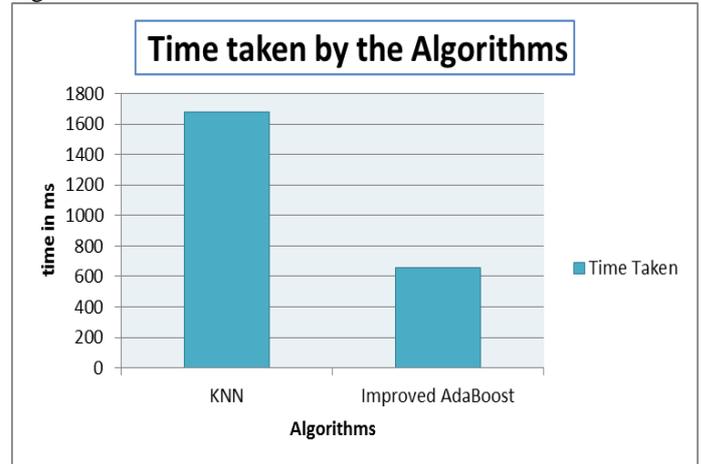


Figure 5: Time Comparison

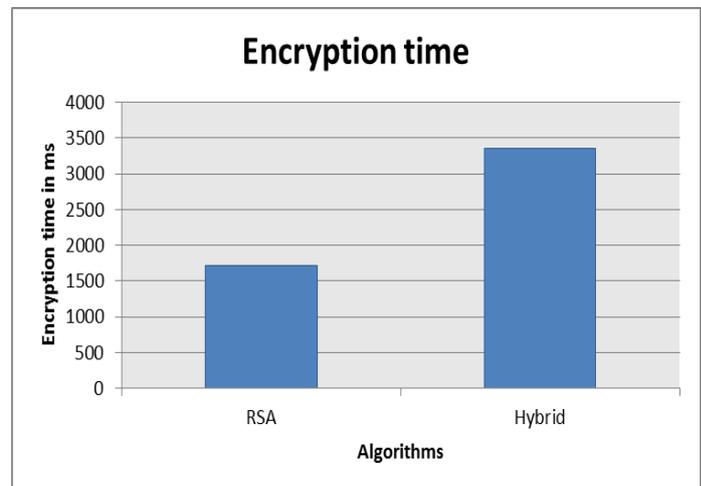


Figure 6: Encryption Time

We can clearly see from the above figure that the proposed Hybrid is taking more time to encrypt when compared with the RSA Algorithm

VIII. CONCLUSION

To compare and analyze the cloud based applications, the setup of the virtual cloud computing environment using cloudSim simulator is required. This paper proposed a data classification model for secure data storage on cloud. For data classification, apply enhanced adaboost algorithm to classify the datasets based on their security needs. Data is classified into sensitive and non-sensitive categories using learning algorithm of classifier. After that only sensitive data is encrypted using hybrid encryption technique which contains hybridization of RSA and blowfish algorithm. Digital



signature for the sensitive data is generated using RSA. Then these digital signature works as a key to Blowfish algorithm for encrypting the data, whereas the non-sensitive data is stored directly on the cloud. This model has been implemented using cloudSim simulator. As compared to k-NN classification model, this improves the confidentiality level of data in a cloud environment. Encryption time, data uploading time and classification time is less in this model

IX. REFERENCES

- [1] M. A. zardari, L. T. Jung, N. Zakaria,” Data Classification Based on Confidentiality in Virtual Cloud Environment”, *Research Journal of Applied Sciences, Engineering and Technology* 8(13): 1498-1509, 2014
- [2] D. Purushothaman, and S. Abburu, “An Approach for Data Storage Security in Cloud Computing”, *IJCSI International Journal of Computer Science Issues*, Vol. 9, Issue 2, March 2012.
- [3] J. W. Rittinghouse, and J. F. Ransome, “Cloud Computing Implementation, Management, Security”, CRC Press 2009 by Taylor and Francis Group, *Journal of high technology law*, 2010.
- [4] D. Catteddu, and G. Hogben, “Cloud Computing: Benefits, risks and recommendations for information security”, *European Network and Information Security Agency (ENISA)*, 2009.
- [5] M.A. I.Manzano, D.L.A.-Ojeda,” An FPGA Implementation for Image Interpretation Based on Adaptive Boosting Algorithm in the Real-Time Systems”, *The 2012 Iberoamerican Conference on Electronics Engineering and Computer Science, Procedia Technology* 3 (2012) 187 – 195.
- [6] De Z. Li, W. Wang , F. Ismail,” A selective boosting technique for pattern classification,” *Neurocomputing* 156 (2015) 186–192, 2015.
- [7] P. S. Rawat, G. P. Saroha, V. Barthwal ,” Quality of Service Evaluation of SaaS Modeler (Cloudlet) Running on Virtual Cloud Computing Environment using CloudSim”, *International Journal of Computer Applications* (0975 – 8887), Vol. 53– No.13, Sept 2012.
- [8] C. P. Ram, and G. Sreenivaasan, “Security as a Service (SaaS): Securing user data by coprocessor and distributing the data,” *Trendz in Information Sciences & Computing (TISC2010)*, pp. 152-155, Dec. 2010
- [9] D. Song, E. Shi, I. Fischer, and U. Shankar, “Cloud Data Protection for the Masses,” *IEEE Computer Society*, pp. 39-45, Jan. 2012.
- [10] T. N. Phyu, “Survey of Classification Techniques in Data Mining”, *Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009*, March 18 – 2009.
- [11] N. Suguna, and K. Thanushkodi, “An improved K-nearest neighbour classification using genetic algorithm”, *IJCSI*, Vol. 7, Issue-4, No 2, July 2010.
- [12] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinbach, D. J. Hand, D. Steinberg, “Top 10 algorithms in data mining”, *Knowl Inf Syst* 14:1–37, 2008.
- [13] L. Tawalbeh1, N. S. Darwazeh, R. S. Al-Qassas and F. AlDosari,” A Secure Cloud Computing Model based on Data Classification”, *First International Workshop on Mobile Cloud Computing Systems, Management, and Security (MCSMS)*, *Procedia Computer Science* 52 (2015) 1153 – 1158.
- [14] R. Velumadhava Rao, K. Selvamani,” Data Security Challenges and Its Solutions in Cloud Computing”, *International Conference on Intelligent Computing, Communication & Convergence (ICCC)*, *Procedia Computer Science* 48 (2015) 204 – 209.
- [15] Renu S, H. Parveen O H,” Biometric Based Approach for Data Sharing in Public Cloud”, *International Journal of Advanced Research in Computer and Communication Engineering* Vol. 4, Issue 2, February 2015.
- [16] J. li, X. Chen, M. Li, Jin. Li, P. P.C. Lee, Wenjing Lou, “Secure deduplication with efficient and reliable convergent key management” *IEEE Transactions on Parallel and Distributed Systems*, Vol.25 , Issue-6, pp-1615-1625, June 2014.
- [17] V. Varadharajan, U. Tupakula, “Security as a service model for cloud environment”, *IEEE Transactions on Network and Service Management* Vol.11 , Issue-1, pp- 60-75, March 2014.
- [18] N. chandel, S. Mishra,” Dynamic secure cloud creation for frequent pattern mining and association”, *Intelligent Systems and Signal Processing (ISSP)*, 2013 *International Conference on, IEEE*, pp- 356-360, March 2013.
- [19] B. Poornima ,T. Rajendran,” Improving cloud security by enhanced HASBE using hybrid encryption scheme”, *Computing and Communication Technologies (WCCCT)*, 2014 *World Congress on, IEEE*, pp- 312- 314, March 2014.