A Secure Data Classification Model for achieving Data Confidentiality and Integrity in Cloud Environment

Kulwinder Kaur
Research Scholar, Department of Computer Science and Engineering,
Swami Vivekanand college of engineering &Technology, Banur, Rajpura, India
kulwinder.kaur0117@gmail.com

Vikas Zandu
Associate Professor, Department of Computer Science and Engineering,
Swami Vivekanand college of engineering &Technology, Banur, Rajpura, India
vckz8037@gmail.com

Abstract— Cloud computing offers numerous benefits including scalability, availability and many services. It needs to address three main security issues: confidentiality, integrity and availability. It is on demand and pay per use service. But as this new technology expanding it also discovers new risks and vulnerabilities too. Security of the data in cloud computing is still an ultimatum to be achieved. The data owners must place their confidential information into the public cloud servers which are not within their trusted domains. Hence, security and privacy of knowledge is the major concern within the cloud computing. To overcome this, various security aspects of security issues has been analyzed and then a framework to mitigate security issues at the level authentication and storage level in cloud computing is proposed in this paper. Deciding data security approach for the data without understanding the security needs of the data is not a valid technical approach. Before applying any security on data in cloud, it is best to know the security needs of the data. That is, whether that kind of data need security or not. In this a data classification approach based on data confidentiality is proposed. The aim of classification technique is to classify the data based on the security needs into two classes confidential and non-confidential. The proposed approach identifies that part of the dataset which must be encrypted using machine learning algorithm and other doesn't which will save time and cost of privacy preserving. After that efficient security mechanisms should be deployed by means of encryption, authentication, and authorization or by some other method to ensure the privacy of consumer's data on cloud storage.

Index terms— Cloud Computing; security issues, privacy preserving, Integrity, confidentiality, availability, graphical passwords, Data classification, Machine Learning, KNN, Boosting.

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 reduced time-to-market, higher availability, better agility, and reduced expenditures to meet the challenging business requirements. All these challenges are addressed by new computing style called cloud computing.

Cloud Computing is an internet based distributed virtual environment. All computational operations are performed on cloud through the Internet. The cost of the resource management is more than the actual cost of the resources. So, it is often better to get the required resources by renting instead of purchasing one's own resources. Basically, the cloud computing provides all IT resources for rent. The definition of cloud computing is: "A distributed virtual environment provides virtualization based IT-as-Services by rent". Beside all of the services like Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure-as-a- Service (IaaS), cloud also provides storage as a service, in which distributed database servers are available for rent to consumers [1]. These services are available for all users without any data bias. With cloud computing, users can browse and select relevant cloud services, such as computer, software, storage, or combination of these resources, via a portal. Cloud computing automates delivery of selected cloud services to the users. It helps the organizations and individuals deploy IT resources at reduced total cost of ownership with faster provisioning.

As cloud computing helps organizations to sharpen their growth and performance. Besides this, it also hosts many users to provide access to shared resources with less effort. But security problems or threats are still a stumbling block in the success path of cloud computing. Numbers of reasons are the matter. First reason is that users and many organizations store their data on cloud storage, so the primary focus is the data must be secure, and the data are not being lost and tampered while travelling from one place to another over the network. So it is essential that confidentiality, availability and integrity of data should be ensured. Secondly, unauthorized access where an attacker tries to be the impersonator of the legal user. [2]

Security is the number one issue when it comes to any upcoming technology and cloud computing is no exception. Cloud computing poses numerous security risks in distributed cache model. Information security is the primary developing risk for the nature of administrations that prevents the clients to embrace the cloud administrations. In distributed storage, the information is put away on the separates through two cache techniques. The previous is to encode the information and store on the server while the last is to store the information without encryption. These functions can often face confidentiality issue. The data is regularly not of the same sort and may have distinctive properties. As the consumer's data is stored on the remote servers and the consumer has no idea about its physical location, so there is always a risk of confidentiality leakage. [3] This paper concentrates on privacy issue in cloud computing. At whatever point the information is exchanged to the cloud server it experiences a security system i.e. encryption without comprehension the level of sensitivity of the data or the data is essentially put away on cloud server without securing it. [13] All information has diverse sensitivity levels so it is improper to store the information without comprehension its sensitivity level and security necessities. To direct the security requirements of data, we have proposed a data classification model to classify the data according to its sensitivity level and then encrypting the only data which is required to secure using an encryption technique in cloud environment. [14]

Classification of objects is an indispensable field of research and of practical applications in numerous fields like pattern recognition and artificial intelligence, statistics, vision analysis and medicine. A very intelligent technique to secure the data would be to first classify the data into sensitive and non-sensitive data and then secure the sensitive data only. This will help to reduce the overhead in encrypting the entire data which will be exceptionally costly in connection of both time and memory. For encrypting the data many encryption techniques can be used and for classifying the data numerous classification algorithms are available in the field of data mining.

Data classification is a machine learning strategy used to predict the class of the unclassified information. Data mining uses unique instruments to grasp the unknown, legitimate patterns and relationships within the dataset. These tools are numerical calculations, factual models and prediction and evaluation of the data. Consequently, data mining consists of management, collection, prediction and analysis of the data. ML algorithms are described in to 2 categories: supervised and unsupervised. In supervised studying, classes are already outlined. For supervised studying, first, an experiment dataset is defined which belongs to distinctive classes. These lessons are competently labelled with a certain identify. Lots of the data mining algorithms are supervised studying with a detailed goal variable. In unsupervised learning classes are not effectively characterized but rather arrangement of the information is performed automatically. The unsupervised algorithm looks for similarity between two gadgets in order to find whether they are able to be characterized as forming a cluster. In simple words, in unsupervised learning, "no goal variable is identified". The classification of information within the context of confidentiality is the classification of knowledge headquartered on its sensitivity level and the have an impact on to the organization that knowledge be disclosed only licensed users. The data classification helps determine what baseline security requirements/controls are appropriate for safeguarding that data. The information is categorized into two lessons, confidential and non-confidential (non-exclusive) information. The classification of the information depends on the attributes of the information. The values of the sensitive attributes are classified as "confidential" and values of the non-sensitive attributes are categorized as "non-confidential".

II. RELATED WORK

Tawalbeh L et al. [4] proposes a secure cloud computing model based on data classification. The proposed cloud model minimizes the overhead and processing time needed to secure data through using different security mechanisms with variable key sizes to provide the appropriate confidentiality level required for the data. They have stored data using three level- Basic, confidential and highly confidential level and providing different encryption algorithms at each level to secure the data. The proposed model was tested with different encryption algorithms, and the simulation results showed the reliability and efficiency of the proposed framework.

Moghaddam F. et al. [5] proposes a hybrid encryption model using classification indexing, attributes and time based procedures. Data classification is mainly based on attributes. A hybrid ring was used to establish the security between the rings. These securely protected rings perform the re-encryption in order to protect themselves from un-authorized access, time based, data owner request and user revocation. The result analysis shows that the hybrid ring model enhances the reliability and the efficiency of the data protection applications.

Dhamija Ankit et al. [6] proposes cloud architecture which ensures secure data transmission from the client's organization to the servers of the Cloud Service provider (CSP). In this, combined approach of cryptography and steganography is used because it will provide a two-way security to the data being transmitted on the network. First, the data gets converted into a coded format through the use of encryption algorithm and then this coded format data is again converted into a rough image through the use of steganography. Steganography also hides the existence of the message, thereby ensuring that the chances of data being tampered are minimal.

Singh Amritpal et al. [7] proposes an enhanced LSB based Steganography procedure for images bestowing better data security. It exhibits an embedding algorithm for hiding ciphered messages in nonadjacent and irregular pixel areas in edges and smooth regions of images. The edges in the cover-image are detected using improved edge detection filter. The encrypted message bits are then embedded in the least significant byte of randomly selected edge pixels and some specific LSBs of red, green, blue components respectively. Such type of steganography technique ensures least chances of suspicion about message bits hidden in the image and it gets hard to estimate the true message length by standard steganography detection methods. The Proposed approach shows better results in PSNR value and Capacity as compared to other existing techniques.

Mishra R et al. [8] describes a secure data transmission using LZW compression before hiding the secret data inside the image. Compression is done in order to reduce the size of the data so that it can be fit inside any media. For data hiding, encryption algorithms are used in order to provide more secure to the secret data using some keys. Now the edges can be detected using canny edge detection method and the encrypted data is embedding inside those edges. Performance analysis shows that the proposed technique is efficient and more secure with less image distortion.

Gaurav S et al. [9] describes all graphical methods for password authentication system and also proposed an approach which describes that first calculation has been done by server based on user entered username and according to result one set of images will be transferred on user screen, each set contains hundreds of images, and then user has to select two images from given set, whereas server also add two images by its own to form complete password

III. COMPARISON OF VARIOUS PPDM TECHNIQUES

Techniques of PPDM	Merits	Demerits
ANONYMIZATION	This method is used to protect respondents' identities while releasing truthful information. While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure.	There are two attacks: the homogeneity attack and the background knowledge attack. Because the limitations of the k-anonymity model stem from the two assumptions. First, it may be very hard for the owner of a database to determine which of the attributes are or are not available in external tables. Second limitation is that the k-anonymity model assumes a certain method of attack, while in real scenarios there is no reason why the attacker should not try other methods. 3. Highly information loss is there.
RANDOMIZATION	The randomization method is a simple technique which can be easily implemented at data collection time. It has been shown to be a useful technique for hiding individual data in privacy preserving data mining. The randomization method is more efficient.	Randomization technique is not for multiple attribute databases. However, it results in high information loss.
CRYPTOGRAPHIC	Cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. There exists a vast toolset of cryptographic algorithms and constructs to implement privacy preserving data mining algorithms.	This approach is especially difficult to scale when more than a few parties are involved. Also, it does not address the question of whether the disclosure of the final data mining result may breach the privacy of individual records.
PERTURBATION	Independent treatment of the different attributes by the	The method does not reconstruct the original data values, but only distribution, new algorithms

	perturbation approach.	have been developed which uses these reconstructed distributions to carry out mining of the data available.
CONDENSATION	This approach works with pseudo-data rather than with modifications of original data, this helps in better preservation of privacy than techniques which simply use modifications of the original data.	It results in high information loss.

IV. PROPOSED METHOD

The research involves exploring various security issues in cloud environment at with respect to three aspects confidentiality, integrity and availability and analyzes their impacts. Also it explores various data classification algorithms in machine learning like KNN, Naïve Bayes and AdaBoost and analyzes their performance.

The paper proposes a secure data classification model using novel boosting supervised machine learning approach. In this, data is classified according to its sensitivity level. Then encrypting only, the data which is required to be secure using a hybrid privacy preserving based image steganography technique in cloud environment. [10] The proposed work also ensures the privacy and integrity of data using hashing approach [11]

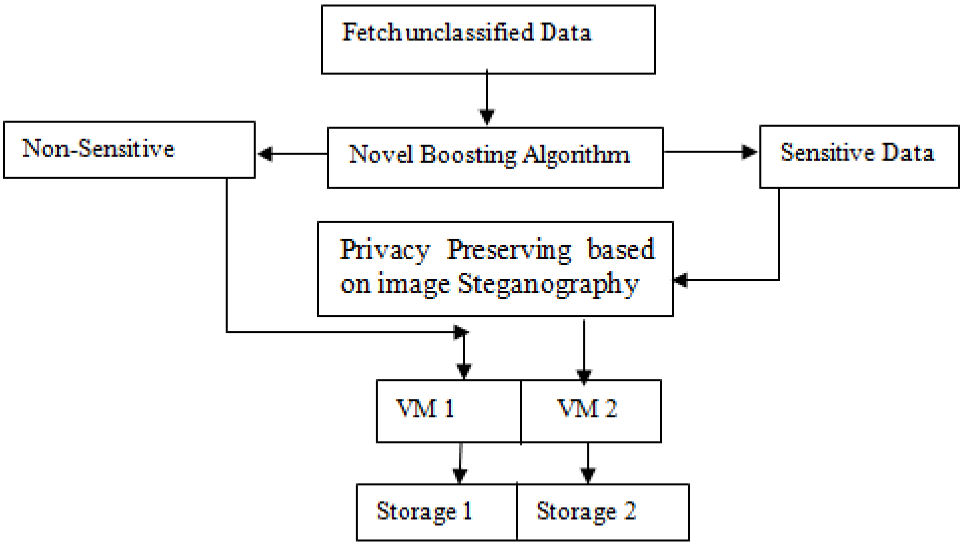

Fig. 1. Overview of the proposed work

Step 1: Image Sequencing Password Authentication

This password is based on the sequences of some images. It is much secure because sequence of images will change each time. This image sequencing password is use for cloud authentication purpose. Only legitimate user will allow entering in cloud, if they enter the correct sequence of image. [12] After authentication, during access of data operations this interface will again ask the user sequence, this time images gets shuffle, based on sequence of images password will also be change.

Step 2: Data Classification

Classifying the dataset by using Novel Boosting Technique.

Algorithm 1: Proposed Boosting algorithm

Input:

D, a set of d class-labeled training tuples;

k, the number of rounds (one classifier is generated per round);

a classification learning scheme which is hybridized average probabilities of Naïve Bayes and Decision stump.

Output: A composite model.

Method:

```
(1) initialize the weight of each tuple in D to 1/d;
(2) for i = 1 to k do // for each iteration:
(3) sample D with replacement according to the tuple weights to obtain Di;
(4) use training set Di to derive a model, Mi;
(5) compute error(Mi), the error rate of Mi using $\sum_{j}^{4}\left(W_j * \mathrm{error}(X_j)\right)$
(6) if error(Mi) > 0.5 then
(7) reinitialize the weights to 1/d
(8) go back to step 3 and try again;
(9) endif
(10) if error(Mi) >0.1 or error(Mi) <=0.3
          (i) for each tuple in Di that was correctly classified do
(ii) multiply the weight of the tuple by error $\mathrm{W\ of\ the\ tuple} * \frac{\mathrm{Error\ of\ M(i)}}{1-\mathrm{Error\ of\ M(i)}}$+0.1 // update weights
          (iii) normalize the weight of each tuple $\mathrm{W\ of\ the\ tuple} * \frac{\mathrm{sum\ of\ old\ weights}}{\mathrm{sum\ of\ new\ weights}}$
(11) else if error(Mi) >0.3 or error(Mi) <0.5
          (i) for each tuple in Di that was correctly classified do
          (ii) multiply the weight of the tuple by error $\mathrm{W\ of\ the\ tuple} * \frac{\mathrm{Error\ of\ M(i)}}{1-\mathrm{Error\ of\ M(i)}}$+0.2 // update weights
          (iii) normalize the weight of each tuple  $\mathrm{W\ of\ the\ tuple} * \frac{\mathrm{sum\ of\ old\ weights}}{\mathrm{sum\ of\ new\ weights}}$
(12) else
          (i)  for each tuple in Di that was correctly classified do
          (ii) multiply the weight of the tuple by error $\mathrm{W\ of\ the\ tuple} * \frac{\mathrm{Error\ of\ M(i)}}{1-\mathrm{Error\ of\ M(i)}}$ // update weights
          (iii) normalize the weight of each tuple $\mathrm{W\ of\ the\ tuple} * \frac{\mathrm{sum\ of\ old\ weights}}{\mathrm{sum\ of\ new\ weights}}$
(13) end if
(14) end for
To use the composite model to classify tuple, X:
(1) initialize weight of each class to 0;
(2)for i = 1 to k do // for each classifier:
(3) Wi=log (Error of M(i))/ (1-Error of M(i))// weight of the classifier's vote
(4)c = Mi(X); // get class prediction for X from Mi
(5) add wi to weight for class c
(6) endfor
(7) return the class with the largest weight.
```

Step 3: Data hiding Architecture

To keep the sensitive data, secure from attackers on the network, a new technique has been proposed which provide the privacy to the individual's data. Instead of sending the actual data in anonymized or encrypted form on to the cloud, in this approach the actual data is not sent on to the cloud. In this an image is used to mask the sensitive data and send the randomized index values in the form of text file on to the cloud.

Hybridized Steganography technique based on Canny Edge Detection

Store all the edge pixel calculated from canny algorithm values with their positions i.e. its row and column into an array.

For example:P (i, j, x)

Where P is the pixel with i= ith row

j= jth column

and x= value of the pixel on that address.

3. 1 Randomization

After finding the edges and pixel values of that edges and store them into an array, randomly select the (pixel value or index value) of that array by using random function generator.

3. 2 Masking

In this phase masking of one bit of message with pixel is done. This is done as follows:

After pixel value is selected using randomization, LSB (Least Significant Bit) Embedding is done. This is described as:

- If the LSB bit of the image pixel values to I (i, j) which is equal to the message bit m, embed that message bit onto the LSB.
  Else
- Again find another pixel value using randomization.

```
        Ls (I, j) =LSB (I (I, j) =1) and m=0,
        ignore that LSB
LSB (I (I, j) =0) and m=1,
        ignore that LSB
LSB (I (I, j) =0) and m=0, then,
((binary value of Pixel & 0xFE) | m)
LSB (I (I, j) =1) and m=1, then,
((binary value of Pixel & 0xFE) | m)
```

For example: let the pixel value be 111 with binary representation 0110111. This pixel value is having LSB 1 and if message bit is also 1 then mask m with its LSB, otherwise again find another pixel value using randomization. This procedure works until all message bits of the dataset are not masked with LSB.

3.3 Reduction

Reduction is done in the way that instead of storing the pixel values onto which our message bit is masked, we would store the corresponding index values of that array into a text file and send this text file to cloud. And at our local end store that array containing edge pixels and their positions.

V. PERFORMANCE PARAMETERS

The evaluation parameters considered for evaluating the performance of the proposed system are:

a) Time taken for classifying the data
b) Accuracy of the classified data
c) Encryption time

VI. RESULTS AND DISCUSSION

The proposed methodology is implemented with the help of Cloudsim and Net beans IDE 8.0. Cloudsim is the library that provides the simulation environment of cloud computing and also provide primary classes describing virtual machines, data centers, users and applications. The classification and data hiding time results have been illustrated in the following figures Figure 2 and Figure 3. And comparison between KNN with RSA and Proposed Boosting with hybrid Image steganography has been made in these figures.

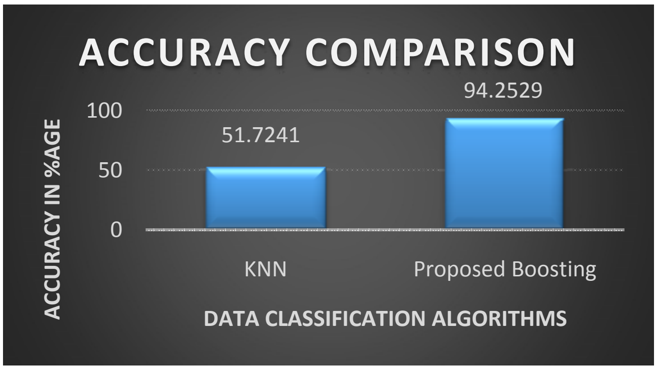

Fig. 2. Performance analysis of data classification algorithms on the basis of accuracy.

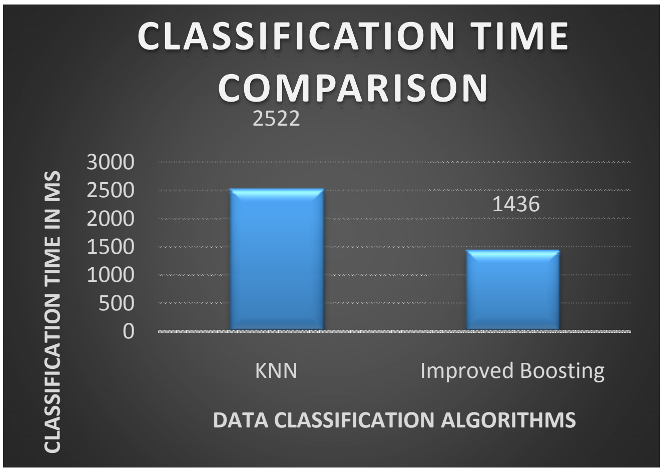

Fig. 3 Performance analysis of data hiding algorithms on the basis of classification time.

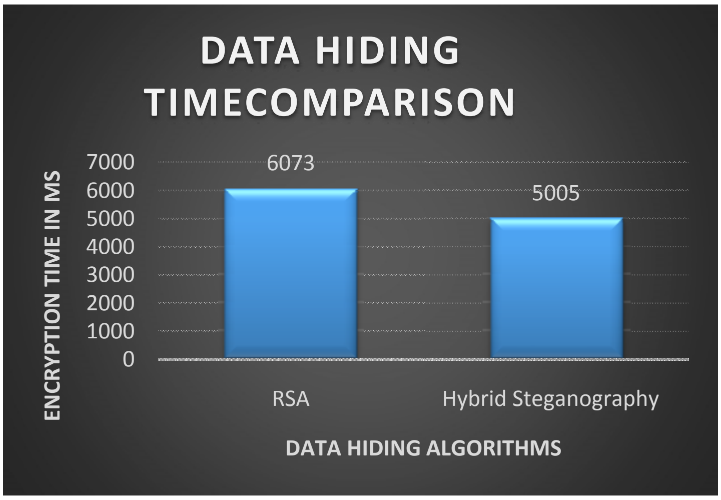

Fig. 4 Performance analysis of data hiding algorithms on the basis of classification time.

The above figure shows the performance analysis of the proposed methodology with the previous method. It is clearly analyzed from the performance graphs that the proposed technique is better than the previous approach. Figure 2 and 3 shows the accuracy and classification time comparison of data classification algorithms KNN and proposed Improved Algorithm. KNN algorithm is having accuracy 51.7241% and improved boosting is having 94.2529% i.e. proposed algorithm has classified data more correctly and performs 42.5288% better than the KNN algorithm. Similarly, figure 4 shows the data hiding time comparison between the proposed and previous RSA approach. Proposed Hybrid steganography technique takes 5005 milliseconds and the RSA algorithm takes 6073 milliseconds to hide the sensitive data. Therefore, in order to reduce the encryption time on cloud data is classified according to its security needs using machine learning algorithms. From the above analysis it is shown that the proposed methodology performs betters in respect to Accuracy and data hiding time.

VII.CONCLUSION AND FUTURE SCOPE

The focus of the research was to characterize the data taking into account the security prerequisites of the information that divides the data into sensitive and non-sensitive using improved machine learning algorithm. The fundamental contribution of this security model is data confidentiality and classification of data using machine learning classification approach. The classified confidential information is then encrypted using hybrid steganography based privacy preserving approach and is stored in the cloud server with hash key to maintain the integrity of the data while the non-confidential data is sent to the cloud environment as public data directly. Furthermore, to enhance the security at the authentication level, image sequencing passwords based on different themes has been used in order to avoid un-authorized access to the cloud environment. The results depict that the proposed technique is more relevant than storing the data without deciding the security needs of the data. Also the results show that the improved boosting technique works better than the K-NN classification technique in terms of both the accuracy and the classification time.

In future, some more security requirements can be taken in account in order to take the classification decision using machine learning algorithm, also the boosting algorithm can be further enhanced using fuzzy logics based decision rules for the classification of data according to the security needs. Authentication level security can be extended to multi-level authentication scheme so that each user will have different access permissions and roles. Availability of the encrypted data can also be improved in future.

ACKNOWLEDGMENT

I would like to place on record my deep sense of gratitude to Associate Professor Mr. Vikas Zandu for his valuable suggestions in my research work. I would like to thank all the people whose encouragement and support has made the fulfilment of this work conceivable.

REFERENCES

[1] Munwar ali zardari, Low Tang Jung, Nordin Zakaria," K-NN Classifier for Data Confidentiality in Cloud Computing", IEEE, pp.1-6, 2014.

[2] Almorsy, M., Grundy, J., & Ibrahim, A. S., "Collaboration- Based Cloud Computing Security Management Framework" IEEE conference of cloud computing, Washington (DC), pp. 364-371,2011.

[3] Song, D., E. Shi, I. Fischer and U. Shankar, "Cloud data protection for the masses", IEEE Computer. Soc., Vol. 45, Issue 1, pp.39-45, 2012

[4] Lo'ai Tawalbeh, Nour S. Darwazeh, Raad S. Al-Qassas and Fahd AlDosari, "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

[5] 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," 2015 IEEE 6th Control and System Graduate Research Colloquium (ICSGRC), pp. 11–16, 2015.

[6] A. Dhamija and V. Dhaka, "A novel cryptographic and steganographic approach for secure cloud data migration," 2015 International Conference Green Computing and Internet of Things (ICGCIoT), pp. 346–351, 2015.

[7] A. Singh and H. Singh, "An improved LSB based image steganography technique for RGB images," 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1–4, 2015.

[8] R. Mishra, A. Mishra, and P. Bhanodiya, "An Edge Based Image Steganography with Compression and Encryption," 2015 International Conference on Computer, Communication and Control (IC4), pp. 2–5, 2015.

[9] S. M. Gurav, L. S. Gawade, P. K. Rane, and N. R. Khochare, "Graphical password authentication: Cloud securing scheme," 2014 International Conference on Electronic Systems, Signal Processing and Computing Technologies, pp. 479–483, 2014.

[10] Chen, D., & Zhao, H., "Data Security and Privacy Protection in cloud computing." IEEE, Hangzhou, pp. 647-651,2012.

[11] Mohammed Faez Al-Jaberi and Anazida Zainal, "Data Integrity and Privacy Model in Cloud Computing" International Symposium on Biometrics and Security Technologies, IEEE, pp.280-284, 2014

[12] Guo, M.-H., Yen, C. T., & Hsiao, L.-L., "Authentication using graphical password in cloud", IEEE, Taipei, pp. 177-181,2012.

[13] Abdullah, A., Hashim, F., & Al-Haddad, S., "A review of cloud security based on cryptographic mechanisms", IEEE, Kuala Lumpur, pp. 106-111,2014.

[14] Abuhussein, A., Bedi, H., & Shiva, S., "Evaluating Security and Privacy in Cloud Computing Services: A Stakeholder's Perspective", IEEE, London, pp.388-395,2012.

AUTHORS PROFILE

Kulwinder Kaur is a student of Swami Vivekanand Institute of engineering and Technology, Punjab. She completed Bachelor of Technology degree in Computer Science and Engineering from Lovely Professional University in 2012. She is now pursuing Master of Technology in Computer Science at SVIET. Her major areas of interest are Cloud Computing, Data Mining and Software Engineering. She has 3 publications in reputed journals.

Vikas Zandu is an assistant Professor in SVIET. He has completed his B. Tech and MTech in Computer Science in 2011 and 2013 respectively. His main areas of interest are networking. He has a numbers of publications in good journals and conferences.