IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-9, Sept. 2018 E-ISSN: 2347-2693

Detection of Sensitive Data Leakage for Privacy Preserving

R.J. Patil[1*], Y.S. Borse[2]

[1]Dept. of Computer Engineering, SSBT's College Of Engineering and Technology, Kaviyatri Bahinabai Chaudhari N.M.U, Jalgaon [M.S], India
[2]Dept. of Computer Engineering, SSBT's College Of Engineering and Technology, Bambhori, Jalgaon [M.S],India

*Corresponding Author: ruchapatil@hotmail.com

Available online at: www.ijcseonline.org

Accepted: 27/Sept/2018, Published: 30/Sept/2018

Abstract- According to Risk Base Security, during last few years leakage of sensitive data record has increased. Human mistake is one of the important reason for data exposure. There is an approach in which data is monitored during transmission to detect the inadvertent data leak cause by human mistakes. However it makes the detection process difficult. There is a need of method that support accurate detection without revealing sensitive data. In particular system, Human identity i,e fingerprint is applied to data file for authentication. It is the process in which original fingerprint matrix is compressed using novel down sampling technique. In the technique, original matrix is compressed by calculating arithmetic mean of the sum of the pixel values on each input row matrix to generate a unit input vector for artificial neural network. The fingerprint samples are matched using back propagation technique. The evaluation result shows improved accuracy and detection time.

Keywords- Data leakage, Downsampling, Backpropagation algorithm**.**

I. Introduction

Privacy preserving plays an important role to provide security to data. Data exposure means unauthorized transmission of sensitive data to an unknown destination where the confidentiality of information is compromise. Number of data-leak instances have grown rapidly in research institute and government organization recent years. Data exposure due to inadvertent data leak increases in recent year. Human mistakes are one of the important reason for inadvertent data leak. A common approach is to monitor the data in storage and transmission for expose sensitive information this makes the detection process difficult and detection time to increase. To provide security against the data leakage fingerprint may be the solution for these problem. The objective is to improve accuracy of detection.

In proposed system, fingerprint authentication process is to prevent the data exposure. In which fingerprint is applied to the data file for privacy preserving . Original fingerprint matrix is compressed using novel down sampling technique. The output of downsampling technique is given as input to backpropagation algorithm. The fingerprint samples are matched using backpropagation algorithm. If the backpropagation result is matched then data is safe and user is able to download the data file otherwise data is downloaded but it in encrypted form. The evaluation result show improved accuracy and detection time.

II. RELATED WORK

The structure of literature survey is shown in Figure 1. Security plays an important role
during the data exposure.

Xiaokui Shu et al., in [1] , uses privacy preserving approach for data leak detection. Human mistake are one of the important reason of data loss. Set of digests are use to detect the data leakage problem. In this method sensitive data is provided to user without revealing the plaintext data. The evaluation results show that method can support accurate the detection and strong privacy guarantees with very small number of false alarms under various data leak scenarios. The advantage of this is it provide quantifiable approach to data exposure. The disadvantage of the system is it is not efficient for practical data exposure.

Liu et al., in [2], has proposed the MapReduce framework to provide privacy for detecting sensitive exposure. The algorithm supports data transformation and it is implemented with the hadoop system. MapReduce framework is use to detect sensitive content which is utilize by public resources. Also the algorithm support the transformation it allows to minimize the exposure of the sensitive data during detection. But it is not efficiently work with hybrid scheme.

Jagtap and Mishra, in [3], focuses on fingerprint image for the security of the data.

They use artificial neural network to compress the image so the better result with less time is evolved. Method uses standard backpropagation with multi-layer percepton is use to provide security to data and the method is superior over the previous method use.

Hahn-Ming Lee et al., in [4], focuses on backpropagation algorithm in artificial neural network. data leak detection with high accuracy is achieve by selecting proper feedforword neural network. As a result, it will improve the learning efficiency by enhancing backpropagation algorithm. Sometimes error saturation problem arise which slows down the learning rate of backpropagation. The propose method not only improve the learning efficiency but also prevent the data efficiently. Only the limitation is the error saturation problem.

Stephane, et al., in [5], proposed the method of back propagation neural1 network which is use to provide security to data leak detection. This method uses parameter like mode of learning, normalization, activation function. Back propagation neural network uses image compression technique to provide more security to the data. But while dealing with image with low compression ratio recognizing it will cost great. Also time required is large due to error of longer coverage time. The advantage of the system is it provide security. Disadvantage is it will cost large due to reorganization ratio.

The literature co-relate with all the detection technique of sensitive data exposure. The various research paper of data leak detection found disclosure of password, human mistakes are one of the important reason for inadvertent data leak.

III. Methodology

The propose approach focuses on data exposure due to inadvertent data leak increases in recent year. Human mistakes are one of the important reason for inadvertent data leak. A common approach is to observe the data in storage and transmission for expose sensitive information. Also existing system such as minutiae base extraction, backpropagation with normal pixel this makes the detection process difficult and detection time to increase. In propose system, fingerprint authentication process is use in which original fingerprint matrix is compressed using novel down sampling technique. Furthermore, original matrix is compress by calculating arithmetic mean of the sum of the pixel values on each input row matrix to generate a unit input vector for ANN. The fingerprint samples are matches using backpropagation technique. The evaluation result shows improved accuracy and detection time.

A. *Architecture*

To protect against the misuse of data, practical data-leak detection service is use for privacy
preserving it improves the detection time and accuracy. The architecture of the proposed
system is shown in Fig. 2.

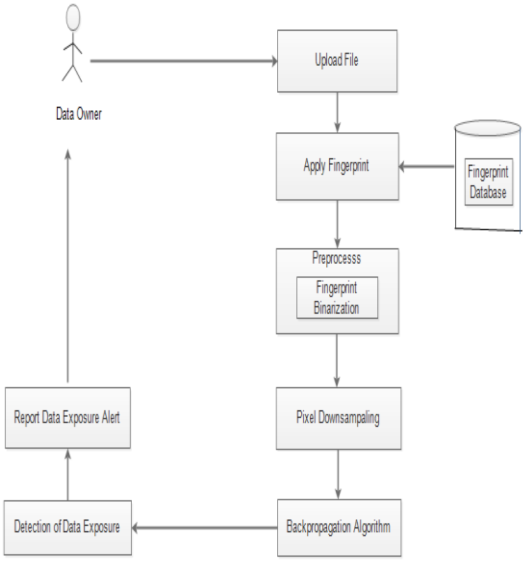
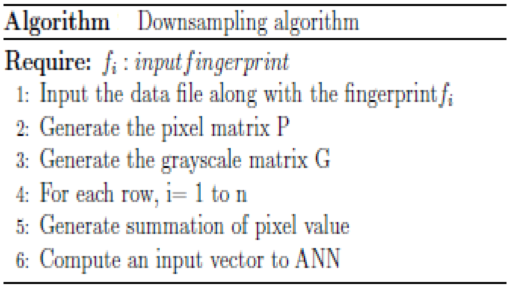


Fig.1 Architecture of the system

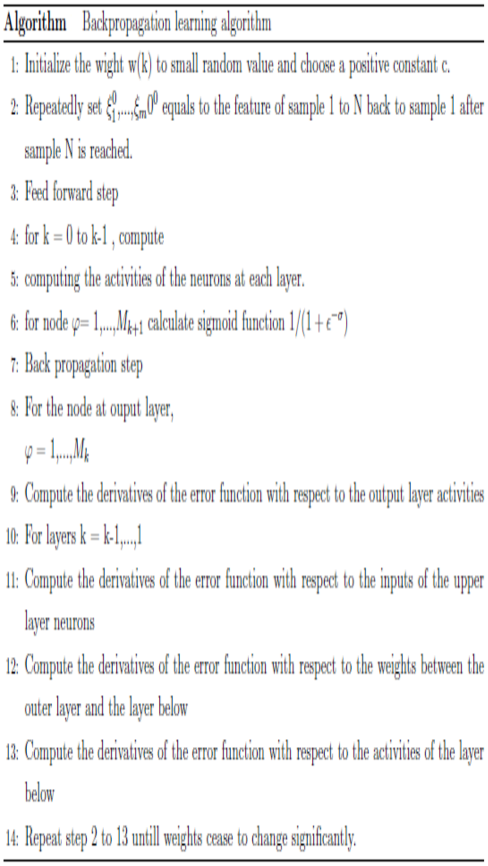
The user first upload data file and apply fingerprint to the data file for authentication.
It will generate pixel matrix for the fingerprint.
It includes preprocesing by binarization i,e conversion of gray scale image into binary number. On the gray scale level each scanned pixel of fingerprint covert white (255) and black (0) binary values of 1's or 0's value.
After fingerprint binarization step, pixel downsampling
method is use in which it will perform summation of pixel
values by dividing with the columns if input vector is row
and vise-verse.



```
Algorithm  Downsampling algorithm
Require: $f_i$ : $input fingerprint$
Input the data file along with the fingerprint $f_i$
Generate the pixel matrix P
Generate the grayscale matrix G
For each row, i= 1 to n
Generate summation of pixel value
Compute an input vector to ANN
```

Fig.2

The output of downsampling is given as input to back-propagation algorithm. To provide security to data exposure they have to confirm the authenticity of fingerprint. As backpropagtion is standard way of training the neural network, it uses down sampled input to neural network. This will check whether the generated output is same as input.



```
Algorithm  Backpropagation learning algorithm
1: Initialize the wight w(k) to small random value and choose a positive constant c.
2: Repeatedly set $\xi_1^0,\ldots,\xi_m 0^0$ equals to the feature of sample 1 to N back to sample 1 after
   sample N is reached.
3: Feed forward step
4: for k = 0 to k-1 , compute
5: computing the activities of the neurons at each layer.
6: for node $\varphi$= 1,...,$M_{k+1}$ calculate sigmoid function $1/(1+\epsilon^{-\sigma})$
7: Back propagation step
8: For the node at ouput layer,
   $\varphi = 1,\ldots,M_k$
9: Compute the derivatives of the error function with respect to the output layer activities
10: For layers k = k-1,...,1
11: Compute the derivatives of the error function with respect to the inputs of the upper
    layer neurons
12: Compute the derivatives of the error function with respect to the weights between the
    outer layer and the layer below
13: Compute the derivatives of the error function with respect to the activities of the layer
    below
14: Repeat step 2 to 13 untill weights cease to change significantly.
```

For privacy preserving in sensitive data exposure system will matched the scanned image with the output of the neural network. Unlike neural network provide shorter time to process and improved accuracy during data exposure.

IV. Results and Discussion

Performance Metrics is used in the detection of sensitive data exposure for finding the accuracy of the system. The proposed algorithm are design in a such a way that it will improve the accuracy. The preliminary about the calculation of precision, recall value in terms of TP, TN, FP, FN as follows.

1. TP = Actual Yes = Predicted Yes

2. TN = Actual Yes = Predicted No

3. FP = Actual No = Predicted Yes

4. FN = Actual No = Predicted No

The accuracy is calculated by following equation of f-measure.

Precision is defined as the ratio of the number of relevant records retrieved to the addition of the number of relevant records retrieved and number of irrelevant records retrieved.

1. Precision = TP/ (TP + TN)

Where, TP : True Positive
TN : True negative

Recall is defined as the ratio of the number of relevant records retrieved to the addition of the number of relevant records retrieved and number of relevant records not retrieved.

2. Recall = TP/(TP + FN)

Where; FN : False Negative

F-Measure is defined as a measure that combines precision and recall is the harmonic
mean of precision and recall.

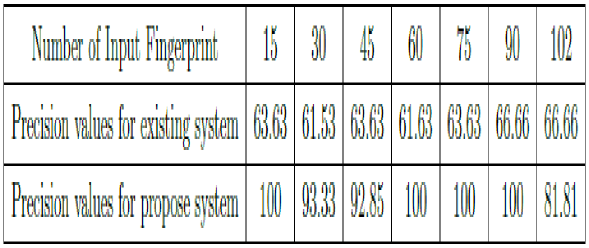
3. F-Measure $=2(Precision \times Recall)/(Precision + Recall)$

Experimental results consists of the value precision, Recall and F-Measure with respect to proposed system as well as existing System. The effectiveness of the proposed system, in which involvement of proposed protocol is proved better by carrying out the experiment.

Results are carried out using excel based calculations.

Table 1 shows the values of precision for the existing and proposed system. Calculating the precision value it use variety of fingerprint applied to data file. The graphs shown in Figure 3 is plotted by considering the average values. The Precision graph shows that the performance of the proposed system.

Table1: Precision Values for Dataset



Number of Input Fingerprint	15	30	45	60	75	90	102
Precision values for existing system	63.63	61.53	63.63	61.63	63.63	66.66	66.66
Precision values for propose system	100	93.33	92.85	100	100	100	81.81

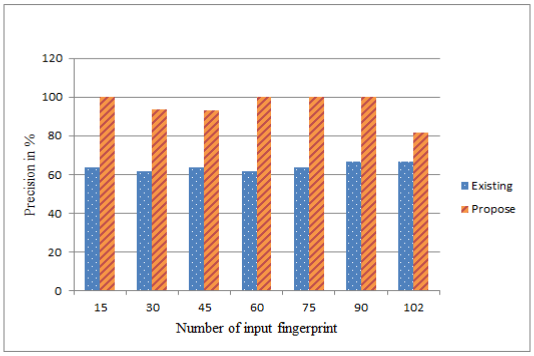


Figure 3. Precision

For Recall Table 2 shows the values of recall for proposed system and existing system. Here for calculating the Recall, variety of fingerprint from dataset is use to data file. The graphs shown in Figure 4 is plotted by considering the average values. The Recall graph shows that the performance of the proposed system.

Table 2: Recall Values for Dataset

Number of Input Fingerprint	15	30	45	60	75	90	102
Recall values for existing system	87.5	88.88	77.77	80	77.77	88.88	66.66
Recall values for propose system	86.66	100	100	93.33	86.66	93.33	90

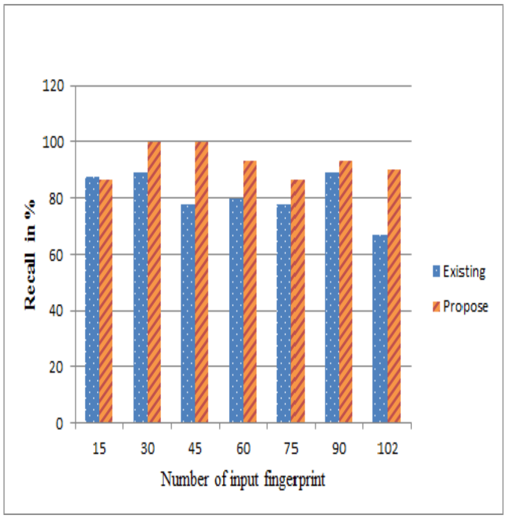


Figure 4. Recall

For F-Measure calculations, Table 3 shows the values of recall for proposed system and existing system. Here for calculating the Recall, variety of fingerprint from dataset is use to data file. The graphs shown in Figure 5 is plotted by considering the average values. The F-Measure graph shows that the performance of the proposed system.

Table3: F-Measure Values for Dataset

Number of Input Fingerprint	15	30	45	60	75	90	102
F-Measure values for existing system	73.67	72.71	69.99	69.62	69.99	76.18	66.66
F-Measure values for propose system	92.85	96.54	96.29	96.54	92.85	96.54	85.70

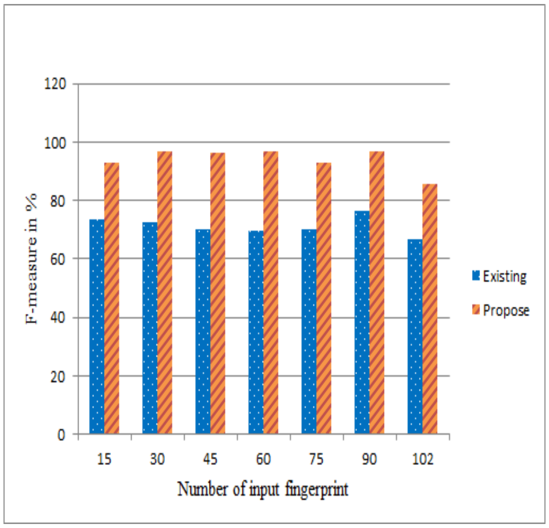


Figure 5. F-Measure

The implementation result of proposed system shows that the precision value are improved and get better accuracy during the detection. Similarity the recall value are gradually improved and f-measure value are changed according to precision and recall. Therefore result of experiment shows improvement in accuracy of the system also provide better security to the system as accuracy is improved.

V. Conclusion and Future Work

The detection of sensitive data exposure require more authentication. The most important issue of authentication method is it is difficult to remember password due to complexity of password. To avoid this the propose system use fingerprint authentication to the data file.

In which arithmetic mean is calculated using effective downsampling technique to provide unit input to backpropagation algorithm. Propose system successfully compress the input. The experimental result of the system improve accuracy and time in detection of sensitive data exposure and provide privacy to data.

The system can improve accuracy, security by exploring host assisted mechanism over an unsupervised network channel for privacy preserving.

REFERENCES

[1] Xiaokui Shu, Danfeng Yao and Elisa Bertino, fellow," Privacy-Preserving Detection of Sensitive Data Exposure" IEEE Trans. on Information Forensics and Security, Vol. 10, No. 5,pp.1092-1103, May 2015.

[2] Liu F., Shu X., Yao D., and A. R. Butt, "Privacy-preserving scanning of big content for sensitive data exposure with MapReduce," in Proc. ACM Conference on Data Application Security and Privacy, pp.195-206, 2015.

[3] Jagtap V. and Mishra S.," Fast efficient artificial neural network for handwritten digit recognition," International Journal of Computer Science and Information Technologies, vol. 5, pp. 2302-2306., 2014.

[4] Hahn-Ming Lee, Chih-Ming Cheb, Tzong-Ching Huang "Learning improvement of back propagation algorithm by error saturation prevention method," Neurocomputing, November 2001, pp. 125-143.

[5] Lei Yu, Mohamed Laaraiedh, Stephane Avrillon, "Fingerprint localisation based on neural networks and ultra-wide band signals," IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Bilbao, pp. 184-189, 14-17 December 2011.

[6] Shu X. and Yao D., "Data leak detection as a service," in Proc. 8th Int. Conf. Secur.Privacy Commun. Netw"., pp. 222-240, 2012.

[7] K. Borders, E. V. Weele, B. Lau, and A. Prakash, "Protecting con_dential data on personal computers with storage capsules," in Proc. 18th USENIX Secur. Symp. , pp. 367-382 ,2009.

[8] A. Nadkarni and W. Enck, "Preventing accidental data disclosure in modern operating systems," in Proc. 20th ACM Conf. Comput. Commun. Secur., pp. 1029-1042, 2013.

[9] Risk Based Security. (Feb. 2017). Data Breach Quick- View: An Executive's Guide to 2013 Data Breach Trends. [Online]. Available: https://www.riskbasedsecurity.com/reports /2016 DataBreachQuickView.pdf, accessed on Oct. 2017.

Authors Profile

Miss. R. J Patil pursed Bachelor of Engineering from Kaviyatri Bahinabai Chaudhari North maharashtra University, jalgaon in 2016. She is Currently pursuing Master of Engineering from Kaviyatri Bahinabai Chaudhari North maharashtra University,jalgon.

Y. S. Borse pursued Bachelor of Engineering and Master in Computer Science and Engineering.She is currently pursuing Ph.D. and currently working as Assistant Professor in SSBT college of engineering Bambhori. She has 7 years of teaching experience.