International Journal of Inventive Engineering and Sciences (IJIES)
ISSN: 2319-9598 (Online), Volume-5 Issue-5, August 2019

Enhancement of Cloud Stationed Healthcare Information Security by Dimensionality Reduction

S. Gnana Sophia, K. K Thanammal

***Abstract**: The security of healthcare information can be secured by the use of cloud environment, and takes finite estimating power. The security of patient's data shared over the internet can be distressed by healthcare institutions because of growing high popularity. The Eigen decomposition (ED) and Single Value Decomposition (SVD) of a matrix are relevant to maintain the security and the study of Dimension Reduction and its advantages are also applicable. To reduce the data without loss, Principal Component Analysis (PCA) is used. Fast retrieval methods are critical for many large-scale and data-driven vision applications. Recent work has explored ways to embed high-dimensional features or complex distance functions into a low-dimensional space where items can be efficiently searched. However, existing methods do not apply for high-dimensional kernel based data The proposed method covers how to generalize locality-sensitive hashing and the implementation of Kernel PCA based methods for Dimensionality Reduction can be applied to Medical data provides high security and utilize the resources of the cloud to inhibit data efficiently.*

***Keywords:** Principle Component Analysis(PCA), Kernel Principle Component Analysis (K-PCA), Single Value Decomposition (SVD), Eigen decomposition (ED)*

I.INTRODUCTION

The important topic in the field of Computer Technology is the concept of cloud computing. The patient's personal records over the internet are stored in a secure manner in the field of healthcare industries. [1]. In Enterprise Cloud, clouds are finite to an appropriate organization. In Public Cloud, the clouds can be accessible to many organizations. Enterprise cloud and public cloud can be combined together to form Hybrid cloud. The development of cloud computing includes the high competence network, minimum cost computers as well as the acceptance of hardware virtual memory, service oriented architecture and autogenous and on-demand computing. Generally, multidimensional images have very high dimensionality. This data needs the low level performance of data for high dimensional space. This problem of dimensionality reduction was created in cloud systems [2].

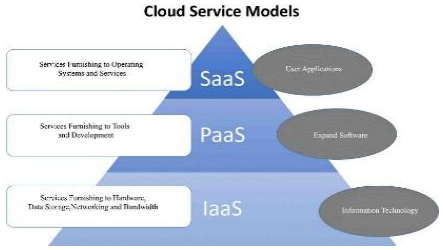

Fig 1: Cloud Services

Revised Manuscript Received on August 20, 2019
*Correspondence Author
S. Gnana Sophia*, M.Sc.,M.Phil., Research Scholar , S.T. Hindu College in Nagercoil, Tamil Nadu, India.
Dr. K. K Thanammal, Associate Professor, Department of Computer Science and Applications S.T.Hindu College, Nagercoil, Tamil Nadu, India.

Dimensionality Reduction have some disadvantages like certain practices can build models from high dimensional data. KPCA technology can reduce the

dimensionality of data for understanding and classification [3]. Power iterative method is used instead of Direct computation method of eigenvectors because of its expensive cost. To describe security and high ability of the protocol, the result verification mechanism is used [4]. To

debug the exact result from the cloud, an efficient verification algorithm is used. [5]. Many new technologies are releasing to keep the cloud services secure and efficient [6]. To maintain the accuracy of the analysis, the Dimensionality reduction is used. The concept of Principal Component Analysis is used to implement on data before clustering which can produce high accuracy and reducing the time [7].

II.RELATED WORK

Zebin Wu et.al in "Parallel and distributed Dimensionality Reduction of Hyperspectral Data on Cloud Computing Architectures". In this paper [8], the author developed a parallel and distributed execution of frequently used technique for hyperspectral Dimensionality Reduction, PCA based on Cloud Computing architecture

Jarin Firose Moon, Shamminuj Aktar and M .M A Hashem et.al in "Securely Outsourcing Large Scale Eigen Value Problem to Public Cloud". Use contrary power method [9] to calculate the eigenvalue with smallest value and calculate if the matrix is no diagonalizable. In this paper the author constructed a process for outsourcing large scale Eigen value problem to a cloud. This can maintain I/O, data security, result checking and client's ability. There are three challenges of this proposed model. The security can be protected by storing the important data in a secure manner. For more security, the client has to first encrypt their confidential data before and after deploying, decrypt the returned result from cloud. The next challenge is to check the returned result whether it is correct or not. Manas A Pathak and Biksha Roy "Privacy preserving protocols for eigenvector computation". The author [10] proposed the protocol for computing the Eigen vector of a collection of private data. The data can be protected and cannot be reconstructed by anyone. "Healthdep an efficient and secure deduplication scheme for cloud assisted e-Health system". The author [11] proposed in the above title that each patient visits the hospital gets a secret key for the treatment. Using this key, the communication between doctor and patient will be done in a secure way. The information about the patient and the treatment given by the doctor are very secret. The doctor encrypts the patient's detail and eliminate the repeated data. So deduplication is done by this process provides more security

III.HEALTHCARE MANAGEMENT SYSTEM

This is not a time to overpass healthcare industry at any cost in today's world. This promises to be the next giant wave in healthcare. It provides all the benefits conceivable by patient and user [12]. Healthcare Management System is patient controlled in some enterprises, some have committed healthcare monitors for maintaining HMS. HMS need to be preserved from hazard at every point. Prior to get an appointment from the hospital each patient has two applications to their devices by the hospital, an application running in the true world and the application running in the secure world because most of the patients have smartphones. After the registration process each patient obtains a treatment key from the hospital and must be kept in a secure manner. This key can be generated whenever needed.

IV.HEALTHCARE INFORMATION SYSTEM IN CLOUD

Extensible, secure and connected cloud system for healthcare information system created to assimilate into scientific system. This will connect to hundreds of medical devices and machines. A cloud based Healthcare Management System is an indispensable requirement of time, as industrial growth is dependent on such advancement and is required by healthcare management because document maintenance and management are the most difficult concern.

There are two types of HMS:

- Cloud based HMS and Dimension Reduction refers
- Client – Server based HMS

Cloud based HMS is more advantageous than other software solutions. The life assumption faster population of age, which provides flourishing of more resources and vast medical care. Caring of medical data was more efficient as the number of patients are increased. These can be maintained by the personal medical record.

V.DIMENSIONALITY REDUCTION

The common problem used in this modern world are, there may be huge amount of large dimensional data, for example, ECG, X ray, Genetic., etc. We have to find the meaningful data in vast amount of information. The methods of Dimensionality Reduction are used to diminish the dimensionality of data and keep most important information. The common methods are Principle Component Analysis (PCA), Random Projection, and Kernel Principle Component Analysis (KPCA). Mostly high dimensional data are multiple and cannot be measured directly. Dimensionality is the process of solving a set of unrestricted variable in a frequency distribution which is used to emulate most of the volatility of a dada set. Dimension reduction is the process of converting the vast dimensions into lesser dimensions. It is the system of preferring a subset of aspects for use in classic structure. And it is the process of diminishing the number of random variables. It consists of three parts: Feature Selection, Feature Extraction and Reduction of Dimension

1) FEATURE SELECTION

The original number of variables in the data set is converted to a smaller number of variables. i.e., select the most important features, remove features with lacking values, remove features with low deviation and unidimensional feature selection

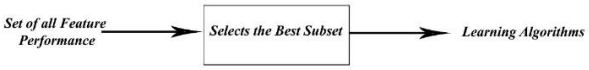

Fig 2: Filter Strategy

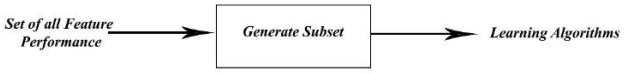

Fig 3: Wrapper Strategy

Example:

- Forward Selection
- Backward elimination
- Recursive feature elimination
- Embedded Strategy

Difference between Filter and Wrapper methods:

- Filter method measures the relevance of features by their correlation with dependent variable whereas, the wrapper method measures the usefulness of a subset of feature by training a model on it.
- Filter method is faster and wrapper method is computationally very expensive
- In filter method, Use statistical methods for evaluation of a subset of features. Wrapper methods use cross validation
- Filter methods fail to find the best subset of features and wrapper methods always provides the best subset of features.

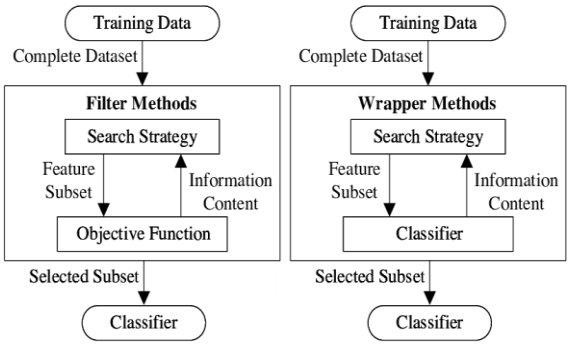

Fig 4: Difference between filter and wrapper method

2) FEATURE EXTRACTION

It mutates the data on high dimensional space to a space of fewer dimensions

- Principal component analysis (PCA)
- Non-negative matrix factorization (NMF)
- Kernel PCA
- Graph-based kernel PCA
- Linear discriminant analysis (LDA)
- Generalized discriminant analysis (GDA)
- Auto encoder

Retrieval Number D0921055419/2019©BEIESP
DOI: 10.35940/ijies.D0921.085519
Journal Website: www.ijies.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

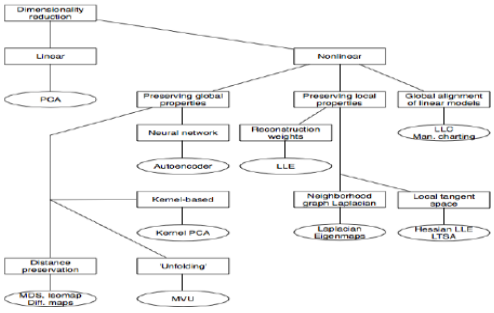

Fig 5: Techniques of Dimensionality Reduction

Eigenvalues and Eigenvectors of symmetric matrices

We are familiar with the matrix Algebra: Addition, subtraction, multiplication transpose and solving linear equations. In a linear algebra, a symmetric matrix is a square matrix that is equal to its transpose. Square matrices can be symmetric because equal matrices have equal dimensions. The entries of symmetric matrices are symmetric with respect to main diagonal

To solve for the Eigen values, λ_i, and the corresponding Eigen vectors X_i of an X_n matrix M.

The methods that reduce the number of variables is called Dimensionality Reduction.

PCA Principle Component Analysis:

In real world, tasks of data analysis can be analyzed as multidimensional data. We can assume the data point A which may be considered as a physical object. As the dimensions of data increases, the anticipating level and the estimation performance also increases. The reasons for reducing the dimensions are

- Catch the mysterious correlations or topics
- Discard the superfluous and noisy features
- Perception and visualization
- Easier repository and processing of the data

To minimize the dimensions of data, we have to dissolve the dimensions and keep the necessary dimensions. It is a popular technique to transform a dataset onto a lower dimensional subspace for visualization. If we reduce the number of dimensions in some data, we can anticipate it because the projection is 2D or 3D space. High dimensional data has problems of High time and space complexity and Overfitting. The objective of PCA is the dimension reduction without much damage of information. PCA finds a new set of dimensions such that all the dimensions are rectangular and rated according to the variance of data. Variance is the evaluation of how spread the data set is. It is the average squared deviation from the mean square.

$$var(x) = \frac{\Sigma(x_i - \bar{x})^2}{N}$$

$$cov(x) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Covariance is the evaluation of the amount to which the related elements from two sets of ordered data change in the equal direction.

There are three types of Covariance

- Positive Covariance : As X increases, Y also increases
- Negative Covariance: As X Decreases Y also decreases
- Zero Covariance : X and Y are not related.

Working of PCA

- Calculate the Covariance Matrix X of data points.
- Compute Eigen values and Eigen vectors
- Arrange the Eigen vectors according to their Eigen values in descending order.
- Choose first k Eigen vectors and that will be the new k dimensions.

Limitations of PCA

- Sources are rarely orthogonal
- Greater interesting sources might not have the largest variance
- PCA is not responsive to hypothesis-testing

The PCA approach works well if the data is linearly separable. In the situation of linearly inseparable data, a non-linear technique is appropriate if the task is to reduce the dimensionality of dataset. Hence we are implementing Kernel PCA in our proposed system which can reduce the dimension of linear data as well as nonlinear data.

Kernel PCA

PCA is linear and cannot classify linear data effectively. KPCA is used to find features in the data and keep this features in the subspace and eliminate the remaining space. We need a method to capture "non-linear data" pattern. Mapping data of lower dimension into a high dimension space makes it linear separable.

It is a nonlinear PCA developed by using kernel method. It uses Kernel trick to find PC in different space. It performs PCA in a new space, and gets higher variance than PCA.

Nonlinear mapping function Ø so that the mapping of a sample X can be written as X -> Ø(X) which is called "kernel function ". Now the term kernel defines a function that computes the dot product of the images of the samples X under Ø

$K(X_i, Y_j) = Ø(X_i)\ Ø(X_j)\ T)$

Goal of Kernel Method:

- To get one kernel function to figure out a certain way of mapping.
- To use Kernel Function K, we know geometry feature in the new space to classify data patterns

$= \phi(X)^T\phi(Z)=(X_1^2,\sqrt{2}X_1X_2,X_2^2)^T\ (Z_1^2, \sqrt{2}Z_1\ Z_2,Z_2^2)$
$= (X^2Z_1^2+X_2^2Z_2^2+2X_1X_2+2X_1X_2Z_1Z_2)$
$= (X_1Z_1+X_2Z_2)^2$
$= (X^r\ Z)^2$
$=K\ (X, Z)$

LSH Feature Mapping Function Φ is always necessarily No! Kernel Function K is positive/Semidefinite. Given a kernel function K, then we can come up with a feature space H.Some popular kernel functions are

Polynomial Kernel Function: $K(x,y)=(x^Ty+c)^d, d\epsilon Z^+$

Polynomial Kernel Function: $K(x,y)=\exp(\frac{-|x-y|}{2a})^2$

- To extend PCA and enable PCA to classify nonlinear data
- To derive kernel PCA

- To project the data to a high-dimension feature space K $\varphi = x_7 \rightarrow H$, $x \in R_d$, $H \in RD$, and $d << D$
- To compute covariance matrix of data in the feature space:

$$C_F = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)\phi(x_i)^T$$

- To compute principal components by solving eigenvalue problem

$$C_F v = \lambda v$$

- The eigenvector can be expressed as linear combination of features

$$v = \sum_{i=1}^{N}\alpha_i\phi(x_i)$$

$$C_F v = \lambda v = \sum_{i=1}^{N}\phi(x_i)\phi(x_i)^T v$$

$$v = \frac{1}{N\lambda}\sum_{i=1}^{N}\phi(x_i)\phi(x_i)^T v = \sum_{i=1}^{N}\frac{\phi(x_i)^T v}{N\lambda}\phi(x_i) = \sum_{i=1}^{N}\alpha_i\phi(x_i)$$

We multiply $\varphi(xi)$ to both sides of $\lambda_v = C_{Fv}$, $\lambda[\varphi(xk)v] = [\varphi(x_k)C_{Fv}]$

$$\lambda\sum_{i=1}^{N}\alpha_i\phi(x_k)\phi(x_i) = \frac{1}{N}\sum_{i=1}^{N}\alpha_i(\phi(x_k)\sum_{j=1}^{N}\phi(x_j))(\phi(x_j)\phi(x_i))$$

We define matrix K as $K_{ij} := (\varphi(xi) \cdot \varphi(xj))$

- Then we will get $N\lambda K\alpha = K^2\alpha$
- which can be solved via the eigen-decomposition.

$$N\lambda\alpha = K\alpha$$

Normalizing the feature space

- Generally, $\varphi(xi)$ may not be zero mean. Thus, we want to center feature.

$$\hat{\phi(x_k)} = \phi(x_i) - 1/N\sum_{k=1}^{N}\phi(x_k)$$

The Corresponding kernel is:

$$\hat{K}(x_i, x_j) = \hat{\phi}(x_i)^T\hat{\phi}(x_j) = (\phi(x_i) - \frac{1}{N}\phi(x_k))^T(\phi(x_i) - \frac{1}{N}\phi(x_k))$$

$$= K(x_i, x_j) - \frac{1}{N}\sum_{i=1}^{N}K(x_i, x_k) - \frac{1}{N}\sum_{i=1}^{N}K(x_j, x_k) + \frac{1}{N^2}\sum_{l,k}^{N}K(x_l, x_k)$$

Thus, we get the centered kernel function $\hat{K}$

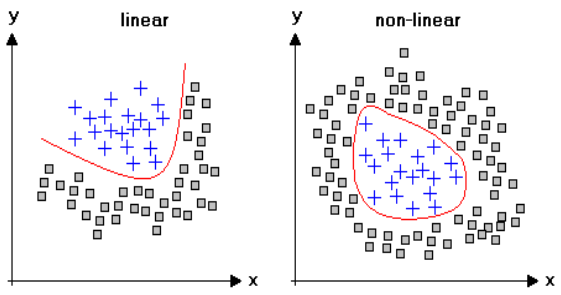

Fig 6: Linear Vs Nonlinear Problems

3) Dimensionality Reduction
K Nearest Neighbor Algorithm

It is the simplest algorithm among all the machine learning algorithm. It is sensitive to the local and it is simple to solve classification problem. K- is the number of neighbors in K nearest neighbor algorithm.

Applications:

It does not learn anything in training period, because KNN algorithm is much faster. It accumulates the training data set and learns from it at the time of making real time predictions. It will not clash the accuracy of the algorithm while new data can be added. It is always easy to implement.

Disadvantages

- Not working with large datasets: In large datasets, it diminishes the achievement of the algorithm when the cost of calculating the Length between the new point and each current point is huge.
- Not working with high dimensions: It becomes difficult for the algorithm to compute the length in each dimensions
- Feature scaling is needed: It needs feature scaling before implementing K nearest neighbor algorithm to any dataset.

The function of nearest neighbors is common. But it is not scalable and this algorithm does not guarantee to give the explicit answer. Hence we are opting the Locality Sensitive Hashing Algorithm to provide greater results.

Locality Sensitive Hashing

This is an effective way of reducing the dimensionality of our data. LSH is used to detect near duplicates. i.e., From the large set of documents eg, text, images web pages, etc., catch the near-duplicates among them. LS found uses in many applications.

- Near-Duplicate detection
- Near-Neighbor Search
- Sketching
- Clustering

Finding Similar documents

We have to find “near duplicate” pairs from a large collection of objects. For this we can break down the *LSH* algorithm in to three broad steps:

1. Shingling
2. Minhash
3. Locality Sensitive hashing

Shingling

We can convert all document into a set of characters of length k to represent each document in our collection as a set of k- shingles.

Example:

Our document(D): “Malayalam “the two shingles are our set{ma, al, la, ay, ya , al , la , am }. The set of three shingles are { mal, ala, lay, aya, yal, ala, lam}

- Similar documents are sharing more shingles
- A small value will produce in many shingles and these are present in many documents.

Retrieval Number D0921055419/2019©BEIESP
DOI: 10.35940/ijies.D0921.085519
Journal Website: www.ijies.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

i. Jaccard Index

This is used to represent each and every document in the mode of shingles. Jaccard index between the two documents A and B is defined as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

It is otherwise called intersection over union (IOU)

Consider there are two documents A: Malayalam B: Malaysia. Then the representation of two shingles will be A: { ma, al, la, ay, ya , al , la , am } and B : { ma, al , la, as, si, ia}

3 in intersection

11 in union

Jaccard Index = 3/ 11

ii. Hamming Similarity

For two n-bit vectors x and y

HS(x , y) = # { I : Xi = yi } / n

For example, disjoint vectors have similarity 0 and HS (x x) = 1

X= 01101 , y = 11001 HS(x , y) = 3/5

1-HS(x , y) is the Hamming distance metric

MinHash

- π on U uniformly at random

VI.DIMENSIONALITY REDUCTION IN MEDICAL DATA

With the improvement of medical field, more specifications are composed to describe the human anatomical position producing huge dimensional clinical data sets. This will boost the complexity of classification, which is utilized in the models that diminishes efficiency. The Locality Sensitive Hashing (LSH) is an admirable algorithm for dimensionality reduction. In the conventional KNN, the work was done only on textual data. The problem of LSH solves for finding the nearest neighbor is a very expensive one both in time and space when operating in large future spaces. It hashes input vectors in a way such that similar vectors are likely to have the same hashes. Because of this property, lookup of near neighbors becomes very efficient operation.

VII.CONCLUSION

In machine learning grouping predicaments, there are habitually too much factors affecting the selection of method. This aspect of fundamentally variables known as features. The greater the number of types, the tougher it procures to envisage the training set and then function on it. Sometimes, most of these functions are correlated, and hence redundant. This is wherein dimensionality reduction algorithms come into play. Dimensionality reduction is the technique of lowering the wide variety of random variables underneath consideration, by way of acquiring a set of most important variables. It may be divided into feature selection and feature extraction. Here Kernel PCA based methods are used with the intention of processing the linear a well as nonlinear data. The proposed scheme gives a high security, reliable and an effective way of reducing the dimensionality of medical data followed by Locality Sensitive Hashing algorithm for providing a good approximation with faster and scalable one.

REFERENCE

1. Mohsen S. Tabatabaei[1], Mostafa Langarizadeh[2*], Mohammad K. Akbari[3], "Security Solutions in Cloud –based Healthcare Information Systems: A Systematic Review" IJCSNS International Journal of Computer Science and Network Security, VOL.18 No.9, September 2018
2. [1]Sangeetha Balodhi, [2]Dr. Bhasker Pant "Dimensionality Reduction Problems Targeting Private Cloud Environments" IJCST Vol. 6, Issue 2, April - June 2015 ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)
3. [1]Sonam Malik, [2]Er. Pooja Narula "Fast Dimensionality Reduction for High Dimensional Datasets Supporting Big Data & Cloud Computing" IJCST Vol. 6, Issue 3, July - Sept 2015 ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)
4. Jarin Firose Moon, Shamminuj Aktar and M.M.A. Hashem, "Securely Outsourcing Large Scale Eigen Value Problem to Public Cloud" 2015 18th International Conference on Computer and Information Technology (ICCIT) DOI: 10.1109/ICCITechn.2015.7488120
5. Lifeng Zhou ; Chunguang Li , "Outsourcing Eigen-Decomposition and Singular Value Decomposition of Large Matrix to a Public Cloud" IEEE Access (Volume: 4) **DOI:** 10.1109/ACCESS.2016.2535103 Pages 869 - 879 25 February 2016
6. Wg Cdr Nimit Kaura, Lt Col Abhishek Lal ,"Survey paper on cloud computing security" March 2017, DOI: 10.1109/ICIIECS.2017.8276134 ,Conference: 2017 4th International Conference on Innovations in Information, Embedded and Communication Systems.
7. G.N.Ramadevi [1] K.Usharani [2] , "Study On Dimensionality Reduction Techniques And Applications" Publications Of Problems & Application In Engineering Research – Paper Vol 04, Special Issue01; 2013
8. Zebin Wu, Member, IEEE, Yonglong Li, Antonio Plaza, Fellow, IEEE, Jun Li, Member, IEEE Fu Xiao, Member, IEEE, and Zhihui Wei "Parallel and Distributed Dimensionality Reduction of Hyperspectral Data on Cloud Computing Architectures" IEEE journal of selected topics in applied earth observations and remote sensing, VOL. 9, NO. 6, JUNE 2016
9. J. F. Moon, S. Aktar and M. M. A. Hashem, "Securely outsourcing large scale Eigen value problem to public cloud," 2015 18th International Conference on Computer and Information Technology (ICCIT), Dhaka, 2015, pp. 490-494. doi: 10.1109/ICCITechn.2015.7488120
10. Manas A. Pathak , Bhiksha Raj, Efficient Protocols for Principal Eigenvector Computation over Private Data, Transactions on Data Privacy, v.4 n.3, p.129-146, December 2011
11. Y. Zhang, C. Xu, H. Li, K. Yang, J. Zhou and X. Lin, "HealthDep: An Efficient and Secure Deduplication Scheme for Cloud-Assisted eHealth Systems," in IEEE Transactions on Industrial Informatics, vol. 14, no. 9, pp. 4101-4112, Sept. 2018. doi: 10.1109/TII.2018.2832251
12. Muneeb Ahmed Sahi1, Haider Abbas [1], (Senior Member, IEEE), Kashif Saleem[2], Xiaodong Yang[3], (Senior Member, IEEE), Abdelouahid Derhab[2] , Mehmet A. Orgun[4,5], (Senior Member, IEEE), Waseem Iqbal[1], Imran Rashid[1], And Asif Yaseen[6], "Privacy Preservation in e-Healthcare Environments: State of the Art and Future Directions" Special Section On Security Analytics And Intelligence For Cyber Physical Systems Received, October 30, 2017, Digital Object Identifier 10.1109/ACCESS.2017.2767561