An Efficient Scheme for Secure Similarity-Based Medical Image Retrieval using Searchable Symmetric Encryption

Irene Getzi S.
Department of Computer Science (PG)
Jyoti Nivas College Autonomous
Bengaluru, India

ABSTRACT
The knowledge extracted from medical imaging data, amassed from hospital cloud repositories can transform the quality of healthcare. This research paper proposes a secure medical image retrieval scheme using Searchable Symmetric Encryption that performs similarity search over encrypted Chest X-Ray (CXR) dataset. The secure index is built using randomized Histogram of Oriented Gradients (HOG) feature descriptor and is organized as an N-dimensional dense array. The search phase uses a probabilistic token generation approach using the inner product of vectors to prevent statistical attacks and to control leakage. Similarity retrieval implemented using Lloyd's algorithm with seed selection facilitate efficient search. The CXR image dataset of pneumonia affected patients from National Institute of Health, USA is used to test the efficacy of the scheme. The randomization used at various levels such as wavelet approximation, sparse randomized matrices and standardization of the feature vector diffuses the data and affects the reconstruction of the image. The AUC scores of 0.877 using ROC and 0.906 using precision-recall curves show that the secure retrieval accuracy is compatible with the non-secure algorithms. The experimental results and mathematical justification are reported.

General Terms
Searchable Encryption, Content-based Medical Image Retrieval, DICOM Chest X-Ray imaging, Histogram based image features, Dimensionality Reduction Techniques.

Keywords
Histogram of Oriented Gradients, Pneumonia Detection, Random Projection, Searchable Symmetric Encryption, Similarity Search.

1. INTRODUCTION
Medical imaging data forms a major proportion of health care information. The massive computations and storage requirements of medical data prompts hospitals to rely on cloud computing platforms for economic feasibility. With the exhaustive disease samples outsourced to cloud, the CBMIR systems can assist physicians in decision making in diagnosis or an aid to teach medical students. Nevertheless, the sensitive information contained in medical data, requires the deployment of CBMIR solution in cloud to address security requirements of patient's privacy and confidentiality. Encryption of medical imaging data before outsourcing can provide protection of the data. However, the traditional encryption techniques may prevent basic operations such as retrieval or computational tasks.

In recent years, processing over secure image data has seen a significant growth. Prior works on secure retrieval were based on keyword-based searching. The notion of secure searching was introduced by [1]. More expressive search queries, involving multiple keywords and Boolean operations were proposed in [2] and [3]. The searching through keyword-based metadata is studied extensively and described in our previous work [4].

The success of keyword-based search depends mainly on the proper annotation of the content which is not practical with large volume of medical data. Content-based search offers a more flexible approach wherein images with similar visual contents such as color histograms, shape and region descriptors, or saliency points as that of the query image are recognized. However, it is not a trivial task to apply cryptographic primitives to content-based medical image retrieval as traditional cryptographic schemes do not preserve the distance between feature descriptors. Moreover, the feature descriptors of medical images are high dimensional.

Basically, there are two models followed for secured image retrieval. A few systems such as [5, 6, 7] tries to extract features such as histogram, SIFT, LBP and moments directly from the encrypted image itself. Another and protect the images and feature descriptors separately. Such schemes built the secure index using the feature vector.

The latter approach is followed in this work. Different encryption techniques proposed in the literature include Paillier homomorphic encryption [8], randomization [9], bag of visual words and min-hash approach [10], Locality Sensitive Hashing (LSH) [11], compact index using fuzzy Bloom filter [12] and Intel Software Guard Extensions (SGX) platform [13]. The schemes used computationally-intensive procedures for encrypting the index and could not provide necessary distance preservation and thus lead to search errors.

To efficiently search through the large feature space, a few vector space models were proposed. The schemes described in [14] used k-means to build an index tree and [15] proposed a secure index built on a balanced binary clustering tree. There are very few works done in the secure retrieval of images in the medical domain [16, 17, and 18].

Our Contribution: In this work, the challenges in developing a secure retrieval scheme for medical imaging data of high dimensional feature space are addressed. The contributions made in the proposed scheme are highlighted as follows:

- Provides a probabilistic search functionality using inner product of vectors that can prevent statistical attacks and control leakage of information such as search and access patterns.
- Further, the reduction of search space by using index tree and clustering offers sub-linear computation time with less communication overhead with the user.
- Though the HOG feature offers high accuracy in classifying the disease, the high dimension of the feature makes it impractical. In addition, the

probabilistic algorithm for SSE requires the key size to be equal to the feature dimension. This challenge is addressed by wavelet decomposition, random projection and using hybrid features.

- Provides better accuracy and is compatible with non-secure algorithm

The paper is organized as follows. Sections 2 and 3 demonstrate the proposed method. Implementation details and results are discussed in Section 4 and Section 5 summarizes the paper.

2. SYSTEM MODEL AND PRELIMINARIES

2.1 System Model

The secure content-based image retrieval system consists of three types of entities: the data owner (A), authorized data users (B) and cloud server/s (C) as depicted in Figure 1.

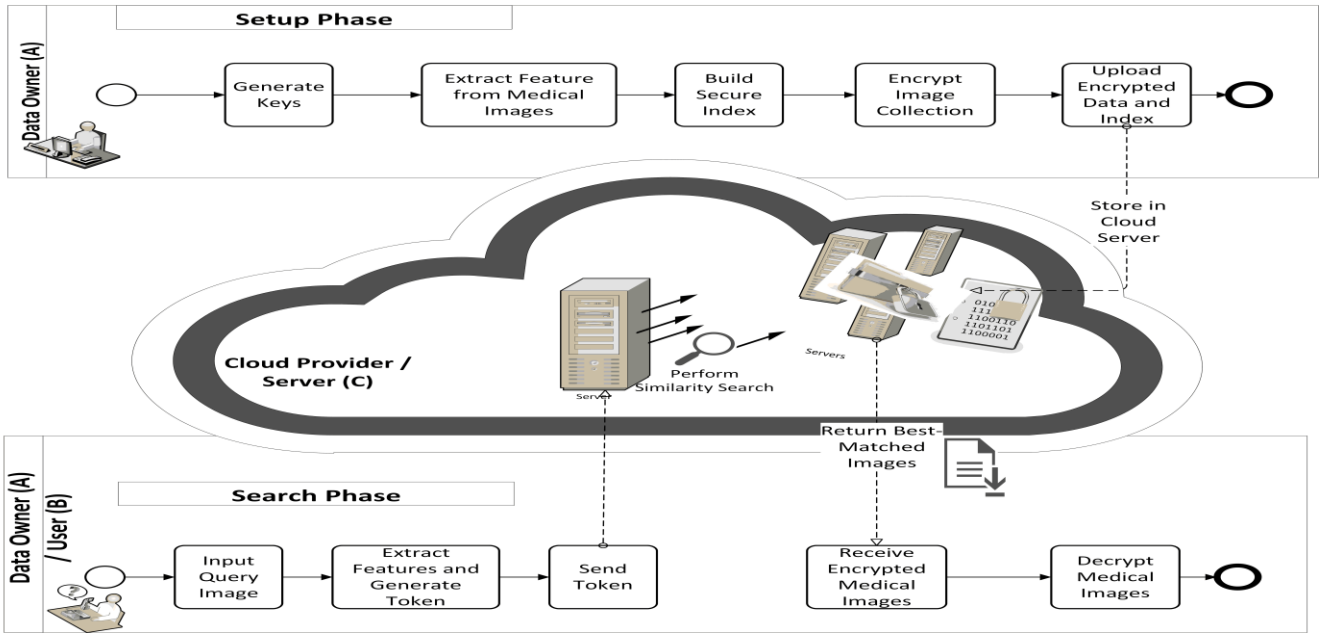

Fig 1: Schematic Diagram of Secure Content-based Image Retrieval over CXR data; The three entities of the scheme are data owner (A), data user (B) and Cloud service provider(C)

To preserve the privacy of the patients and to protect the sensitive information contained in the images, the image collection and the searchable feature index are encrypted before outsourcing to the cloud server.

The data owner or the authorized data user can prepare a searchable token based on the feature vector extracted from the query image and submit it to the cloud server. On receiving the search request, the cloud server computes the similarities with the searchable index and returns 'k' best-matched images encrypted images to the user. The data user can decrypt the received images using the secret keys.

2.2 Preliminaries

SSE Algorithms: Dynamic Searchable Symmetric Encryption Scheme consists of polynomial-time algorithms (KeyGen, BuildSecureIndex, TokenGen, Search, and Update) and the description of these basic SSE algorithms can be found in [4].

Random Projection: Random projection maps the high-dimensional feature vector into lower dimensional by approximately preserving the distance between them according to Johnson-Lindenstrauss lemma mentioned in [19].

For a given feature vector F, the randomization is defined as: $RP(F) = R * F$ which approximately preserves the Euclidean distance L1 between any two projected feature vectors $RP(f1)$ and $RP(f2)$ by a scaling factor √m/2.

The elements $r_{i,j}$ of R are usually transformed by Gaussian distribution. Sparse random matrices as defined by Achlioptas in [20] is a memory efficient alternative that provides similar embedding with faster computations using integer arithmetic.

2.3 Security Requirements and Design Goals

In this model, the data owner and the data users are considered as trusted entities while the cloud server/s is said to be honest-but-curious. The server can attempt to obtain sensitive information about the patient by analyzing the search and access patterns. It is highly essential to provide efficient

Let $M = M_1, M_2, \ldots, M_n$ be the collection of CXR images to be outsourced. The data owner extracts suitable features $F = F_1, F_2, \ldots, F_m$ to identify the disease conditions. A searchable index I is built with the feature descriptors.

To preserve the privacy of the patients and to protect the sensitive information contained in the images, the image collection and the searchable feature index are encrypted before outsourcing to the cloud server.

similarity search over large-scale medical imaging data without compromising on data security.

The following are the design goals of the proposed system.

Privacy Protection: Privacy protection and confidentiality of the image collection, and secure index should be ensured. The leakage of search as well as access patterns need to be controlled.

Accuracy: The search result over the encrypted feature vectors should be as accurate as the search in plaintext data.

Efficiency: The scheme should offer a practical solution for searching over large-scale data with sub-linear search time.

Reconstruction: Given a feature vector, it would be encrypted in such a way that, it is difficult to reconstruct the original image.

The detailed description of the design and implementation of the secure CBMIR scheme is described in the following section.

3. DESIGN AND IMPLEMENTATION

The proposed scheme consists of two phases namely the setup phase during which the data owner builds the secure index and uploads the index and image collection to the cloud server. In Search phase, the user generates a token to search through the encrypted index and decrypts the server responses. The server uses the encrypted feature index to answer queries.

The following Figure 2 describes the detailed flow of the proposed secure similarity-based image retrieval algorithm.

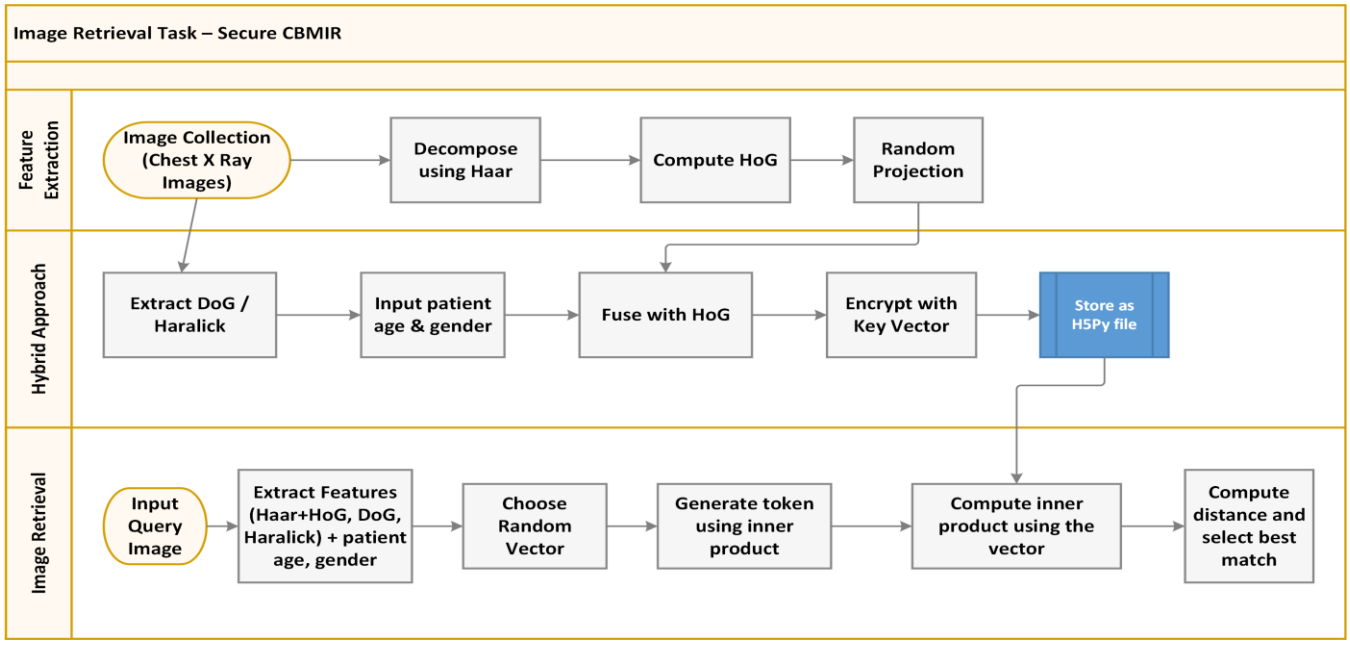

Fig 2: The proposed secure and efficient image retrieval scheme over Encrypted CXR data

3.1 Setup Phase

During the setup phase, the data owner extracts feature vector from each image and build searchable index. It makes use of two algorithms namely:

Key Generation: The key generation algorithm uses a random bit string and generates two secret keys, K_m, K_f for encrypting images and index respectively. The key can be shared using any secure key distribution protocols to the user with whom the owner wishes to share the data. The search phase uses probabilistic encryption algorithm that requires the key size to be the same dimension as that of the feature vector. Thus, $K_f = K_1, K_2, \dots, K_m$ where m is the feature dimension.

Build Secure Index: HOG texture feature is found to provide good accuracy in classifying the pneumonia disease from

CXR images as described in [21]. The high-dimension of the dense HOG descriptor is reduced by wavelet decomposition and randomization to avoid burdensome computations. The construction of the secure index is described in Algorithm 1.

The secure index is organized as a collection of structured arrays stored in HDF5 binary file. A 128-bit K_m is used to encrypt the images. The header and the pixel data array of the DICOM images are encrypted separately using block cipher AES algorithm in CBC mode. The secure index and the encrypted image collection are uploaded to the cloud server.

3.1.1 Algorithm 1 BuildSecureIndex

Require: Secret keys (K_m, K_f) for encrypting index, Medical Image Collection $M = M_1, M_2, \dots, M_n$

Ensure: Encrypted Image collection M_E and a secure Index file I_E

1: for each entry in M do
2: The image M_i of size N*N is down sampled using Haar wavelet and approximation component of level-3 decomposition is selected.
3: HOG feature descriptor is computed by dividing the image into 8*8 cells and by grouping them into blocks of size 2*2.
4: The high dimensional HOG descriptor is reduced by applying Gaussian and Sparse Random projection.

5: The resultant feature vector is encrypted with a Key vector K_f by computing: $C_F = F' + K_f$ that forms an entry of the secure index and is organized as an N-dimensional structured array.
6: end for
For each image M_i a structured array F' is inserted to the secure index I_E.

3.2 Search Phase

During the setup phase, the data owner extracts feature vector In Search Phase, the data owner or the authorized user generates secure searchable token T_f from the query image and send it to the cloud server. The server, then performs similarity search over the encrypted feature vectors (secure index) and return the best-match **'k'** images to the user. The search process is implemented using two algorithms namely:

Token Generation: From the query image, the features are extracted and encrypted as following the steps mentioned in Algorithm 1. In order to generate a probabilistic token, the user chooses a random vector $V = (V_1, V_2, \ldots, V_r)$ where r is the reduced dimension of the feature vector or the size of each entry in secure index. If Q is the feature vector of the query image and K_f is the key, the user, generates the token T_f using inner product of vectors as $T_f = (< V, K >, < V, Q >, V)$ and send to the server along with the vector V.

Similarity Search: To perform exact match, upon receiving the token, the server computes the inner product of secure index with the random vector, $< V, C_F >$ and verifies if:

$$< V, K > + < V, F > == < V, C_F > \quad \text{Eqn. (3.1)}$$

To perform the secure similarity search, we adopt the following strategy:

3.2.1 Algorithm 2 SecureSimilaritySearch

Require: secure Index I_E, Token T_W
Ensure: Images matching the query image
1: Choose the number of clusters '**k**' using a seed selection algorithm.
2: Partition the feature space I_E into n clusters using Lloyd's algorithm.
3: Compute the inner product of vectors as described in Equation 3.1 between the query vector and the cluster centroids and select the nearest cluster.
4: Select all labels in the cluster and compute the minimum distance between the vectors and the query to return the best-matched images.

4. RESULTS AND DISCUSSION

The proposed work has been implemented using Python programming environment and the experimental results are discussed in the following section.

Protocol Summary: The secure CBMIR scheme requires processing at two levels. The medical images and the feature vectors are encrypted independently and uploaded to the server. During search phase, the user generates a probabilistic token from the query image; the server perform similarity search through the secure index and return the best-matched images.

Experimental Platform: The experiments has been implemented using Python OpenCV and run on DELL system, equipped with Intel core i5 6th Gen processor @ 2.8 GHz * 4 cores with 16 GB RAM running Ubuntu 14.04.

Dataset: To test the efficacy of the scheme, the real-world data of 5200 images has been selected from NIH CXR images of pneumonia affected patients. The dataset is available in https://www.kaggle.com/nih-chest-xrays/data. The images are levelled and of uniform dimension of 1024 * 1024.

4.1 Secure Index Construction

The HOG feature was found suitable in identifying the disease pneumonia with 98% accuracy as mentioned in our previous work [21]. When applied HOG over a 1024*1024 image, with cell-size 8*8 and a block size of 2*2, it produces a feature descriptor with a very high-dimension of more than one lakh components. The index construction takes several minutes and causes memory error for larger datasets.

In addition, the proposed SSE algorithm requires the key size to be same as that of the feature vector. Hence, it is necessary to reduce the dimension of HOG without compromising accuracy. Initially, Haar wavelets are used that has better approximation and reduces the high dimensional HOG feature vector to few thousands as mentioned in Table 1 and Figure 3.

Table 1. Effect of Haar Decomposition levels on HOG

Feature	Dimension	Feature Extraction Time (s)	Accuracy (using RF)
HOG	1,20,000	9698.0887	98.00
Haar(L1)+HOG	90000	2634.88	96.04
Haar(L2)+HOG	34596	1175.75	95.01
Haar(L3) + HOG + RP	10	475.2	86.02
Haar(L3) + Hybrid*	10	921.04	93.67

* Hybrid features include HOG with random projection, few Haralick features and DoG

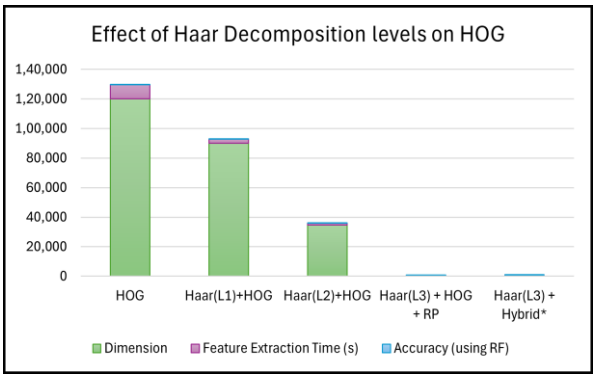

Fig 2: Effect of Haar Decomposition levels on HOG

The feature dimension and computation time is drastically reduced with the increase in the level of decomposition of Haar wavelets as shown in Table 1. The accuracy in classifying the disease is maintained almost at the same level. Random Forest (RF) classification algorithm performs fairly well.

The construction of secure index follows a three-fold approach:

- The computed HOG feature descriptor is standardized using min max scalar function.

- Random projection (RP) helps in further reducing the dimension of the feature vector. Gaussian Random projection and a computationally efficient Sparse there is a reduction in classification performance as the number of projected components decreases.
- The drop in the accuracy level is improved by fusing other features with randomized HOG as shown in row 5 of table 1. The hybrid features consist of first four components of random projection, five Haralick features such as ASM, contrast, correlation, variance and inverse difference moment and total number of Difference of Gaussian (DoG) components.
- The resultant feature vector is encrypted using a random key vector as mentioned in Algorithm 1 and uploaded to the server. The accuracy of the image retrieval scheme is measured using ROC and Precision-Recall curve scores and the same is illustrated in Table 2.

Table 2. The ROC and Precision-Recall Scores

Scheme	F1 Score	Precision-Recall mAP	ROC AUC
HOG (L3 Haar)	0.836	0.924	0.927
HOG (L3 Haar) + RP	0.823	0.861	0.871
HOG (L3 Haar) + Hybrid	0.847	0.906	0.877

4.2 Secure Similarity Search

Once the query image is classified as having pneumonia, it is desirable to perform similarity-based search to retrieve k best-matched images. The proposed secure similarity-based retrieval approach is described in Algorithm 2. K-means clustering combined with Lloyd's algorithm is used to shrink the search space. The selected labels are compared with brute force search using Euclidean distance, the results are ranked, and the best matched images are returned.

Security Analysis: The proposed approach is analyzed under the criteria defined in the design goals. The security of any SSE scheme is characterized by its control over the information leak to the server.

Confidentiality and Privacy Protection: The medical image collection is encrypted with the standard pseudo random function and permutation such as AES in CBC mode with a key size of 128 bits. The secure index construction of the proposed CBMIR scheme uses randomization and permutation at different levels as described as follows.

- The image is transformed into wavelet domain level-3 approximation using Haar wavelet.
- The extracted features are rescaled or normalized to the range of (0, 1) using min-max scalar algorithm.
- The matrices used in random projections are crucial for providing security and are generated using cryptographically secure pseudo random number generator.
- The fusion of Haralick features and DoG components further enhances the security. The resultant random

  Random projection that suits the sparse HOG data are employed. As seen from Table 1,

  projected vector is extended using the randomized key vector.

The probabilistic token generation algorithm utilized during search phase requires a random vector V, which is used to extend the feature vector. The scheme generates different token, every time even for the same query image. This will protect the query or access pattern leakage. It is difficult for the server to perform statistical attack with the known set of tokens and the query results.

Reconstruction: In content-based retrieval system, the security mainly relies on the server or intruder's inability to reconstruct the image from the feature vector. Since Haar wavelet can produce better reconstruction, it is possible to get the image with the feature vector. However, the randomization happened at different levels makes it hard to reconstruct the image from the projected vector. The randomization applied using min-max scaling or standardization, sparse random projection, encryption with a random vector makes it difficult to reconstruct the image from the feature vector.

4.3 Performance Evaluation

The proposed scheme is compared with few existing SSE schemes used in CBMIR domain and the results are tabulated as given in Table 3.

Table 3. Comparisons with a few CBMIR schemes

Scheme	Features Used	Accuracy	Search Time (s)
Paillier [17]	DWT, Histogram	0.43	0.002*N
SGX [13]	SURF	0.68	O(N * L)
Non Secure [18]	Tamura Texture	0.85	O(L* k)
Proposed Scheme	HOG, SSE	0.906	O(L* k/n)

L represents the total number of images
k represents total number of features
N denotes the size of the image
n represents the number of clusters

The proposed SSE scheme with k=10 components offers efficient computation time as well as better accuracy and thus suitable for secure retrieval in large-scale medical imaging data analysis.

5. CONCLUSION AND FUTURE SCOPE

The problem of secure retrieval in content-based medical domain is effectively addressed in this research paper. An efficient and secure CBMIR scheme using HOG feature descriptor is proposed. The dense HOG feature computation takes several minutes. The wavelet level-3 decomposition helps to reduce the dimension as well as provide considerable reduction in the computation time without compromising accuracy. In order to reduce the key size and to diffuse the feature vector, random projection is employed.

The probabilistic token generation through inner product of vectors and randomization offers better security. The drop in accuracy level due to randomization is resolved by using hybrid approach for feature generation. Segmentation of lung region may be considered to improve the accuracy. The feature computation time could be further reduced by using parallel approach.

6. REFERENCES

[1] Song DX, Wagner D, Perrig A, "Practical techniques for searches on encrypted data", In proceedings of 2000 IEEE Symposium on Security and Privacy, p. 44–55, 2000.

[2] Xia Z, Wang X, Sun X, Wang Q, "A Secure and Dynamic Multi-Keyword Ranked Search Scheme over Encrypted Cloud Data", IEEE Transactions on Parallel Distributed Systems, Vol. 27(2), pp.340–352, 2016.

[3] Cash D, Jarecki S, Jutla C, Krawczyk H, Rosu MC, Steiner M, "Highly-scalable searchable symmetric encryption with support for Boolean queries", In proceedings of Advances in Cryptology– CRYPTO 2013, Springer, pp. 353–373, 2013.

[4] Getzi I, Durairaj C. D, "A Dynamic Scheme for Secure Searches over Distributed Massive Datasets in Cloud Environment using Searchable Symmetric Encryption Technique", International Journal of Information Security Science, Vol. 7(3), pp. 126–139, 2018.

[5] Cheng H, Zhang X, Yu J, "A Coefficient histogram-based retrieval for encrypted JPEG images", Multimedia Tools and Applications, Vol. 75(21), pp. 13791–13803, 2016.

[6] Hsu CY, Lu CS, Pei SC, 'Image feature extraction in encrypted domain with privacy-preserving SIFT", IEEE Transactions on Image Processing, Vol. 21(11), 4593–4607, 2012.

[7] Yang T, Ma J, Wang Q, Miao Y, Wang X, Meng Q, "Image Feature Extraction in Encrypted Domain With Privacy Preserving Hahn Moments", IEEE Access, Vol. 6(4), pp.47521– 47534, 2018.

[8] Erkin Z, Franz M, Guajardo J, Katzenbeisser S, Lagendijk I, Toft T, "Privacy-preserving face recognition", In proceedings of International Symposium on Privacy Enhancing Technologies Symposium, Springer, pp. 235–253, 2009.

[9] Lu W, Varna AL, Swaminathan A, Wu M, "Secure image retrieval through feature protection", In proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009, pp. 1533–1536, 2009.

[10] Lu W, Swaminathan A, Varna AL, Wu M, "Enabling search over encrypted multimedia databases", Media Forensics and Security, Vol. 7254, pp. 725418-725429.

[11] Xia Z, Xiong NN, Vasilakos AV, Sun X, "EPCBIR: An efficient and privacy-preserving content-based image retrieval scheme in cloud computing", Information Sciences, Vol. 387, pp. 195–204, 2017.

[12] Wang Q, He M, Du M, Chow SS, Lai RW, Zou Q, "Searchable encryption over feature-rich data", IEEE Transactions on Dependable and Secure Computing, Vol. 15(3), pp. 496– 510, 2018.

[13] Yan H, Chen Z, Jia C, "SSIR: Secure similarity image retrieval in IoT", Information Sciences, Vol. 479, pp. 153–163, 2019.

[14] Yuan J, Yu S, Guo L. Seisa, "Secure and efficient encrypted image search with access control", In proceedings of IEEE Conference on Computer Communications (INFOCOM), pp. 2083–2091, 2015.

[15] Liang H, Zhang X, Cheng H, Wei Q, "Secure and Efficient Image Retrieval over Encrypted Cloud Data", Security and Communication Networks, Hindawi, Vol. 2018, Article ID 7915393, 2018.

[16] Yang M, Trifas M, Chen L, Song L, Aires D, Elston J, "Secure patient information and privacy in medical imaging" J Syst Cybern Inf. Vol. 8(3), pp. 63–66, 2010.

[17] Bellafqira R, Coatrieux G, Bouslimi D, Quellec G, "An end-to-end secure CBIR over encrypted medical database", In proceedings of IEEE 38th Annual International Conference on Engineering in Medicine and Biology Society (EMBC), pp. 2537–2540, 2016.

[18] Xiaoming S, Ning Z, Haibin W, Xiaoyang Y, Xue W, Shuang Y, "Medical Image Retrieval Approach by Texture Features Fusion Based on Hausdorff Distance", Mathematical Problems in Engineering. pp. 1–12, 2018.

[19] Johnson WB, Lindenstrauss J, "Extensions of Lipschitz mappings into a Hilbert space", Contemporary mathematics, Vol. 26(1), pp. 189-206, 1984.

[20] Achlioptas D, "Database-friendly random projections: Johnson-Lindenstrauss with binary coins", Journal of com-puter and System Sciences. Vol. 66(4), pp. 671–687, 2003.

[21] Getzi I, Durairaj C D, Raj J V, "Efficient Image Retrieval approach for Large-scale Chest X Ray data using Hand-Crafted Features and Machine Learning Algorithms", International Journal of Computer Sciences and Engineering, Vol. 6(11), pp. 890–896, 2018.