scientific reports

Check for updates

OPEN

# A deep dictionary clustering approach for unsupervised image retrieval using convolutional sparse coding

G. Sucharitha[1], B. Vikas[2], P. V. Bhaskar Reddy[3], Sajja Suneel[4], Naadem Divya[5], Onur Osman[6] & Jawad Rasheed[7,8,9,10]✉

**Medical image repositories have been rapidly growing due to the widespread use of imaging techniques, making manual annotation unfeasible. Efficient image retrieval systems are crucial for diagnosing diseases, planning treatments, and conducting medical research. This paper presents Deep Dictionary Clustering for Image Retrieval (DDicCIR), a novel framework that integrates deep learning with dictionary clustering for unsupervised medical image retrieval. The method employs DenseNet121 to extract image features, followed by a two-level dictionary learning process. In the first dictionary layer, the sparse representations are learned to group similar images, while the second layer refines these representations to capture higher-level abstractions and improve feature discrimination. An iterative clustering mechanism, based on k-means, updates the clusters until convergence, enhancing sparsity, reducing noise, and strengthening category separation. Experimental results on the NIH Chest X-ray and IRMA datasets show that DDicCIR achieves significant improvements in precision, recall, and mean average precision (mAP), demonstrating its effectiveness for medical image retrieval. The code is available on https://github.com/sucharithasu/DDicCIR.**

**Keywords** Deep learning, Dictionary learning, K-means clustering, K-SVD, Medical image retrieval, Deep clustering

Certainly, searching for similar images based on image content from large repositories poses a significant challenge for multimedia systems. Traditional keyword-based searching methods have limitations, such as difficulty in capturing visual similarity and handling large datasets. Content-based image retrieval (CBIR) has emerged as a promising solution for searching similar images in large databases[1–3]. However, CBIR meets several challenges that must be addressed to improve the accuracy of image retrieval systems. These include the need for effective representation of image content, both automatic and manual image annotation, and bridging the semantic gap between low-level image features and high-level semantics. Additionally, resolving image ranking problems presents another substantial challenge. To improve feature extraction and bridge the semantic gap, extensive research efforts are necessary, as they are crucial for enhancing retrieval precision. The mushroom growth of hospitals and imaging techniques like histopathology, Computer Tomography (CT), X-ray, mammography, sonography, etc., is used to diagnose diseases, depositing the images in large amounts in hospital repositories[4]. Consequently, there is a great demand for a compelling image retrieval system to aid clinicians in surfing these massive datasets and improving the diagnostic accuracy. The detailed literature survey on content-based and medical image retrieval is presented in[5–8].

For a decade, the most popular image representations have relied on deep learning and dictionary learning (or sparse representation) methods[9]. Dictionary learning is a powerful technique that has a wide range of applications

[1]Computer Science and Engineering, School of Engineering, Anurag University, Hyderabad, India. [2]Computer Science and Engineering, Sreenidhi University, Hyderabad, India. [3]Computer Science and Engineering, REVA University, Bangalore, India. [4]Computer Science and Engineering (Data Science), Institute of Aeronautical Engineering, Hyderabad, India. [5]Sreenidhi Institute of Science and Technology, Hyderabad, India. [6]Department of Electrical and Electronics Engineering, Istanbul Topkapi University, Istanbul, Turkey. [7]Department of Computer Engineering, Istanbul Sabahattin Zaim University, 34303 Istanbul, Turkey. [8]Department of Software Engineering, Istanbul Nisantasi University, Istanbul 34398, Turkey. [9]Research Institute, Istanbul Medipol University, Istanbul 34810, Turkey. [10]Applied Science Research Center, Applied Science Private University, Amman, Jordan. ✉email: jawad.rasheed@izu.edu.tr

in signal processing, image processing, and neuroscience. It leverages the ideas of sparse representation to capture the essential features of data efficiently. In the context of images, dictionary learning aims to learn a set of atoms (basis vectors) such that a given image can be well approximated by a sparse linear combination of these learned atoms. Many researchers have emphasized the significance of dictionary learning in statistics, machine learning[10-11], signal processing, and computer vision applications[12-13]. Recent advancements have shown that data-driven dictionary learning has significant potential across various applications. Techniques such as K-SVD[14-15] and the Method of Optimal Directions (MOD)[16] have been developed to enable sparse representation for each database entry. Furthermore, the growing field of compressed sensing shows promise for exploiting the inherent sparsity in medical images. Alongside traditional approaches such as dictionary learning, deep learning models have increasingly been adopted in medical image retrieval. With architectural modifications, these models now enable more efficient unsupervised retrieval and offer significant improvements in feature representation and retrieval performance. For example, using a CNN model for feature extraction and the clustering technique for indexing the feature map database, the authors of[17] proposed a content-based medical image retrieval pipeline. Their method incorporated multi-level gain-based feature selection to reduce the dimensionality of feature vectors derived from pre-trained CNN models. Similarly, Weng et al.[18] proposed a novel unsupervised method for medical image retrieval, employing a self-distilling dual network to train a dual encoder for image feature extraction, followed by unsupervised metric learning to enhance feature representation. The subsequent Related Work section elaborates on existing approaches in greater detail and outlines the limitations that motivated this study.

The rest of the article is planned as follows: In "Related work" section, related works, "Proposed framework-DDicCIR" section presents with proposed work, experimental results, and analysis in "Experimental analysis" section, and the article concludes with a conclusion in "Conclusion"section.

## Related work

The objective of this paper is to retrieve the grayscale medical image encompassing different organs, views, and modalities. For a specified number of clusters, we create an equal number of dictionaries to represent them. Each image in the database is linked to a dictionary based on a sparsity criterion. When a query image is provided, the concept of sparsity is used to determine the appropriate cluster, within which relevant images are retrieved. The effectiveness of the proposed method is demonstrated through comprehensive experimental results.

(A) Deep learning models

The great success of pretrained deep learning models in extracting fine-grained features from images has simplified recognition and classification tasks for labeled datasets. Hang et al.[19] surveyed various advanced deep learning models, such as RNN, FCNN, and ResNet, in medical image analysis, including tasks like segmentation, disease diagnosis, registration, retrieval, and more. Medical imaging datasets like Chest X-rays have further enabled large-scale deep learning applications for disease detection and retrieval[20]. Recently, several deep learning models have been proposed to learn fine-grained similarities among images, enhancing retrieval precision. Qiaoping He[21] proposed the Deep-Embedding Global Feature Descriptor (DGFD), which incorporates frequency statistics ranking and embeds global topology features both spatially and channel-wise into deep convolutional features. The Modified Resilient Back-Propagation (MRPROP) technique was employed in[22] to improve CNN training's efficiency and convergence for the image classification task. Zheng et al.[23] introduced a novel system called DCDicL, which combines deep learning and dictionary learning. It learns the priors for both dictionaries and representation coefficients and adaptively modifies the dictionary for every input image according to its content. From these studies, it has been concluded that deep learning models play a significant role in medical image feature extraction across a wide range of applications. In this research, we employ DenseNet121(Dense Convolutional Networks), where each layer is connected to every other layer in a feed-forward manner, allowing for better feature reuse throughout the network. This architecture helps mitigate the vanishing gradient problem, enabling effective training of deep networks. As a result, DenseNet121 is particularly well-suited for extracting fine features from data[24]. DenseNet121 consists of 121 layers, including dense blocks, convolutional layers, pooling layers, and fully connected layers[25]. The depth of this architecture enables it to effectively penetrate medical images, facilitating the extraction of more relevant and detailed features. The model is often pre-trained on large datasets like ImageNet, which allows it to learn generalized feature representations beneficial for various image types, including medical images such as chest X-rays. Due to its robustness to variability in patient positioning, imaging techniques, and pathological conditions, this model is particularly suitable for feature extraction in Chest X-ray images. In this study, 1,024 features are extracted per image, providing a rich representation for further extension through dictionary learning[26]. In summary, prior research underscores the importance of deep learning models in medical image analysis, and DenseNet121 is adopted in this work as a robust feature extractor to generate rich, discriminative representations for subsequent dictionary learning.

(B) Dictionary learning

Naturally occurring signals often contain vast amounts of data, making it challenging to extract relevant information. Sparse coding is a technique that represents data as a sparse linear combination of basis functions or atoms from a dictionary. Its objective is to obtain a concise representation of the data by emphasizing the most relevant atoms while minimizing the number of non-zero coefficients. Dictionary learning, also known as sparse coding or representation learning, is a machine learning technique used to discover a dictionary (a set of basis

functions) that can efficiently represent a given dataset[27–30]. The goal is to obtain a compact and informative representation of the data by capturing its underlying structures and patterns. K-SVD (K-Singular Value Decomposition) is a widely used algorithm for dictionary learning[31], particularly in sparse coding applications. It extends standard singular value decomposition (SVD) by incorporating sparsity constraints. The K-SVD algorithm learns a dictionary from training samples by iteratively updating both the dictionary atoms and their corresponding sparse codes. In the seminal work on K-SVD[32], the dictionary is optimized in two stages. First, a greedy approach called Orthogonal Matching Pursuit (OMP) is applied to estimate the sparse coefficients under the $\ell_0$-norm constraint, which controls the number of non-zero elements in the coefficient matrix. Second, Singular Value Decomposition (SVD) is used to update the dictionary. Subsequent improvements to the $\ell_0$-sparsity constraint have been proposed to optimize the coefficient matrix further, and these have been successfully applied to image retrieval tasks. The ultimate objective is to learn a dictionary that efficiently represents the training data using a sparse linear combination of dictionary atoms[33]. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments (SwAV)[34] introduces a clustering-based contrastive method that jointly assigns and learns features, achieving strong unsupervised representation learning without labeled data. Deep Clustering for Unsupervised Learning of Visual Features (DeepCluster)[35] jointly learns feature representations and cluster assignments by iteratively applying k-means on deep features, enabling powerful unsupervised visual representation learning. In summary, sparse coding and dictionary learning, particularly through the K-SVD algorithm and its extensions, provide powerful mechanisms for deriving compact and discriminative representations of data, which have been effectively applied to tasks such as image retrieval.

(C) Limitations

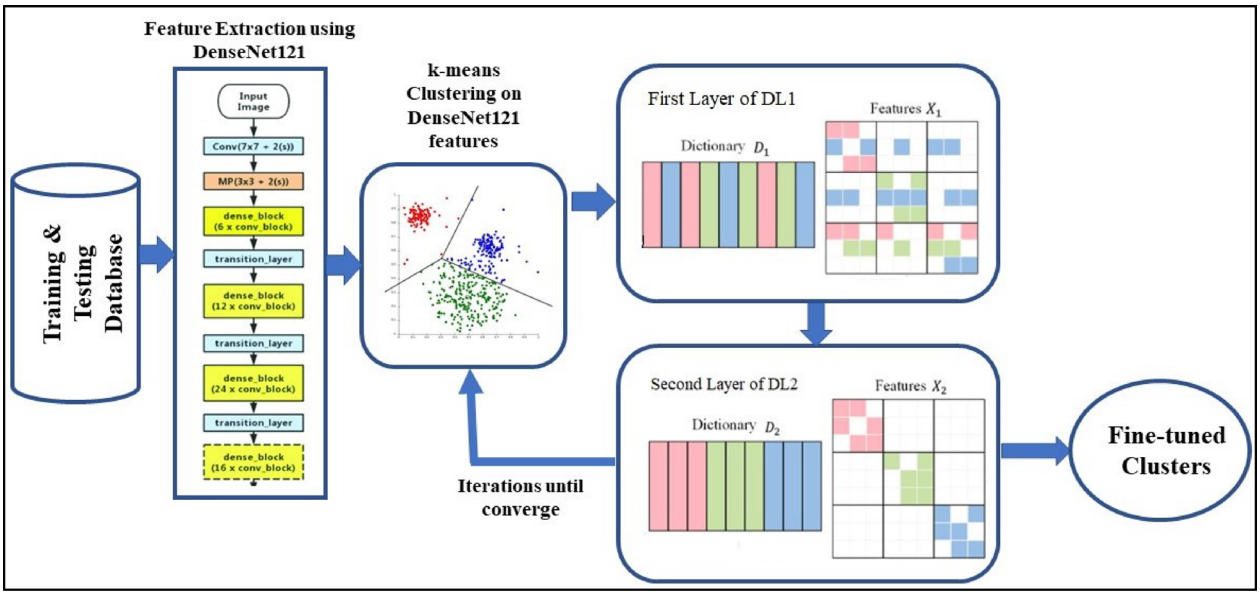
Many sparsity-based approaches have been developed for image classification, exploiting the idea that high-dimensional data such as images can be represented with only a few significant components using an appropriate basis or dictionary. Laplacian group sparse coding and graph-based Laplacian sparse coding[36_37] have been applied to classify images by capturing correlated and specific attributes. However, it is observed that Laplacian sparse coding is significantly affected by high computational complexity and slow convergence. In[37-38], variants of convolutional sparse coding methods were introduced, leveraging their ability to learn localized and spatially invariant features. These methods employ convolutional filters to extract edge and texture information, representing each image as a sparse combination of the learned convolutional filters. The sparsity constraint helps focus on the most important features, leading to better generalization for classification and clustering tasks. Nevertheless, convolutional sparse coding remains computationally intensive and less flexible in updating dictionaries. Sadik et al.,[39] introduced a multivariate dictionary learning model for portfolio selection, demonstrating efficient feature representation in financial data. However, like most conventional dictionary learning approaches, the method struggles with scalability and adaptability when applied to high-dimensional or complex datasets, highlighting the need for more flexible and hierarchical dictionary frameworks.

To overcome the limitations of sparse coding methods, this work proposes a new unsupervised method called Deep Dictionary Clustering for Image Retrieval (DDicCIR). Unlike previous studies that carry out shallow or one-layer dictionary learning, the proposed DDicCIR method utilizes a two-layer adaptive dictionary learning network that continuously updates and enhances the feature representations of clustered image data. The two dictionary layers produce sparse representations by convolutional kernels to maintain local and spatial dependencies in DenseNet121 feature maps, and then refine the fused sparse codes with higher-level discriminative structures. Besides, an iterative interaction process between k-means clustering and dictionary learning process guarantees that both cluster assignments and dictionary atoms are dynamically optimized until convergence, enhancing feature coherence and cluster compactness. The architecture is fully unsupervised, rendering it extremely well-fitted for large unlabeled medical image databases.

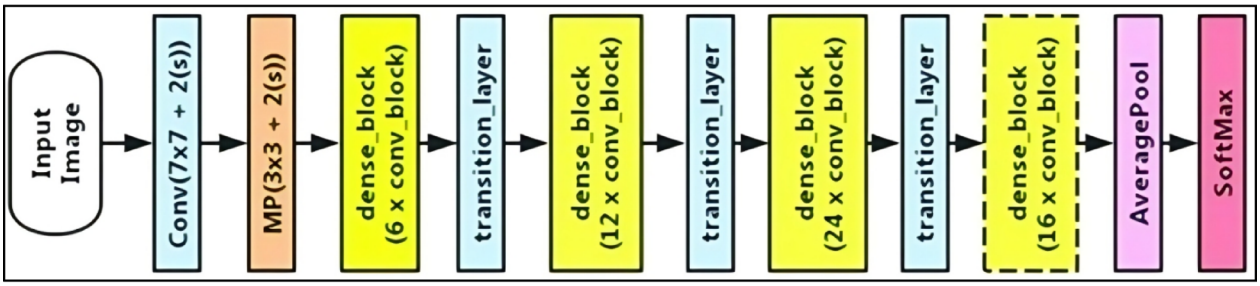
## Research gap and contributions

To address the above-mentioned limitations, this research introduces a novel unsupervised framework called DDicCIR pipeline, that incorporates two-layer K-SVD–based adaptive dictionary learning for fine-grained clustering of unlabelled images. In this approach, image features are first extracted using DenseNet121, a state-of-the-art deep learning model known for its ability to capture deep semantic features across diverse image datasets. DenseNet121 in our framework is used solely as a feature extractor and is not trained using class labels for the retrieval task. Although the model weights are initialized using the standard ImageNet-pretrained checkpoint, no label information is used during training, optimization, or clustering in the DDicCIR pipeline. The network is frozen during feature extraction, and no supervised fine-tuning is performed. These extracted features are then used as input to the k-means clustering algorithm, which performs an initial partitioning of the data based on feature similarity. This serves as the first step in an iterative clustering process. Once the initial clusters are generated, the process advances to the first layer of dictionary learning (DicL1). At this stage, a compact and discriminative dictionary is learned for each cluster, enhancing the quality of the partition. For each image, sparse matrices are generated using different kernels, which are then aggregated into a single sparse matrix by applying average or max pooling. This aggregated sparse representation captures refined relational features from the image features, offering a more robust description of the underlying data structure.

The output from DicL1 is then passed to a second layer of dictionary learning (DicL2), which provides further refinement by capturing finer distinctions that may have been overlooked in the first stage. This two-layer hierarchical strategy ensures that clustering becomes progressively more accurate and detailed with each iteration. The iterative process continues until convergence, when the cluster assignments stabilize and no further improvement in clustering quality is observed. Figure 1a provides a pictorial representation of the proposed framework, and (b) gives the detailed architecture of DenseNet121, illustrating the sequence of steps from




(a)




(b)

**Fig. 1**. (**a**) Schematic diagram of proposed method, including the layers information of DenseNet121 and k-means iterative structure for Dictionary layers features, (**b**) DenseNet121 architecture.

feature extraction to final cluster convergence. By enhancing both the discriminative power and compactness of features, the proposed method is particularly effective for handling complex and subtle variations in medical datasets such as NIH Chest X-rays and IRMA.

- *Convolutional Dictionary Learning (CDicL)*

A common approach in machine and image processing techniques is to represent an image patch vector $y \in \mathbb{R}^m$ as a linear combination of basis atoms, i.e., $y = Dx$, where $D \in \mathbb{R}^{mXd}$ is the dictionary of atoms (a set of learned or predefined basis vectors), and $x \in \mathbb{R}^d$ is the representation coefficient vector. Generally, the image features $y$ can be reconstructed as a weighted sum of atoms in D, with the coefficients for each atom expressed by the entries of *x*. This method is known as sparse coding or dictionary learning.

The DicL model can be expressed with the following equation.

$$min_{D,x} \frac{1}{2} \parallel DX - Y \parallel_2^2 + \lambda_D \psi(D) + \lambda_X \phi(X) \tag{1}$$

$Y \in \mathbb{R}^{mXN}$, represent a set of N training samples, where each column corresponds to a vectorized image patch. The matrix $X \in \mathbb{R}^{dXN}$, denotes the sparse representation matrix of Y over dictionary D. The $\phi(X)$ represents the coefficients X and $\psi(D)$ represents the regularization term imposed on the dictionary D. The parameters $\lambda_X$ & $\lambda_D$ are regularization weights for X & D, respectively, controlling the trade-off between reconstruction accuracy and the sparsity or regularization of the model. The alternative nature of performance on dictionary learning, K-SVD[39,40], is mostly used sparse representation method. It first fixes D, performs sparse coding to compute X, and updates D through singular value decomposition (SVD). Even though the K-SVD is effective but it fails to capture the spatial relationship among the features of an image. This limit can be addressed

by convolutional dictionary learning CDicL which preserves the spatial and local structures using pretrained convolutional filters[41]. Specifically, the function of CDicL can be written in Eq. (2).

$$min_{D,\{X_i\}} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \| \ D \star \ X_i - Y_i \ \|_2^2 + \lambda \ _D \psi \ (D) + \lambda \ _X \phi \ (X) \, , \tag{2}$$

Here, $D_i \star \ X_i = \sum \ _{k=1}^{K} D_k \ \otimes X_{k,i}, \otimes$ is 2D convolution function, and K is the number of kernels; $D = \{D_k\}_{k=1}^{K}$, is the convolutional dictionary and $D_k \in \mathbb{R}^{cXc}$ is the kth 2D dictionary kernel, $X_i = \{X_{i,k}\}_{k=1}^{K}$, is the sparse coefficient of image features $Y_i \in \mathbb{R}^{hXw}$, and $X_{i,k} \in \mathbb{R}^{hXw}$, is the k[th] kernel of $X_i$.

## Proposed framework-DDicCIR

The proposed Deep Dictionary Clustering architecture for image retrieval (DDicCIR), involves an iterative process where k-means clustering is used alongside two layers of dictionary learning. The goal is to progressively refine the feature representations of images and improve the accuracy of clustering. This section outlines the iterative procedure of clustering and dictionary learning, followed by a detailed mathematical explanation of the first dictionary learning layer. The overall framework of the proposed work is given in Fig. 2.

### Iterative clustering and dictionary learning

The iterative procedure in DDicCIR involves the following steps:

(A) Initial Feature Extraction:

First, image features are extracted using a pre-trained deep learning model such as DenseNet121. The image features for all database images are considered as $Y = \{Y_i\}_{i=1}^{N}$, where $Y_i \in \mathbb{R}^{h*w*f}$, characterizes the feature map of the $i^{th}$ image, h and w are the spatial dimensions, and f is the number of channels or kernels used in DenseNet121.

(B) k-means Clustering:

Primarily, the image features database $Y = \{Y_i\}_{i=1}^{N}$, are clustered using k-means into k number of clusters. Let $C_k$ denote the group of image features belonging to the k[th] cluster.

$$\{C_1, C_2, ..C_K\} = K_{means(Y)} \tag{3}$$

Each cluster $C_k$ consists of images that have similar feature representations.

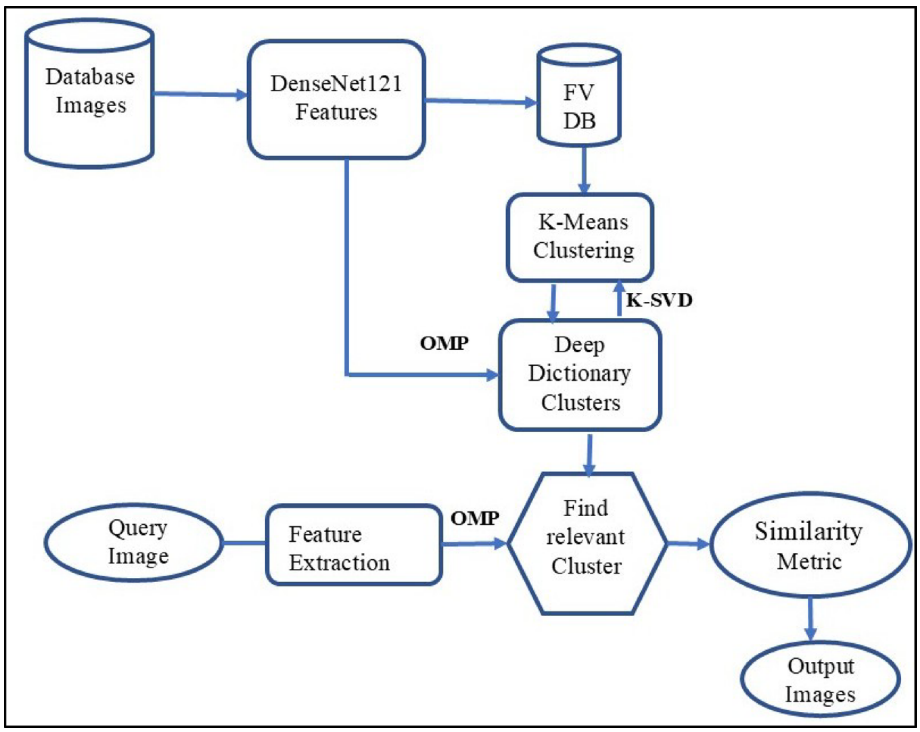



**Fig. 2**. Block diagram of the proposed framework for database images and query image.

(C) Two-layer Dictionary Construction:

The DDicCIR model employs a hierarchical two-layer dictionary structure. The first layer operates on the feature representations extracted from DenseNet121 and performs cluster-specific dictionary learning, generating sparse representations for each cluster independently. The second layer refines these representations by learning a global dictionary over the aggregated sparse matrices (via max-pooling), thereby capturing cross-cluster dependencies and higher-level abstractions.

(D) Relationship Between the Two Layers:

The first layer focuses on intra-cluster refinement, enhancing local feature quality and compactness, while the second layer captures inter-cluster relations to improve global separability. Together, these layers form a coarse-to-fine hierarchical representation, where the output of the first layer acts as the input to the second, creating a progressively refined and discriminative feature space.

(E) First Layer of Dictionary Learning & its Initialization:

The initial dictionary is generated after performing k-means clustering on the DenseNet121-extracted image features. Each cluster produced by k-means is assigned an individual dictionary, initialized using the feature vectors of the images belonging to that cluster. The initial dictionary atoms $D^{(0)} = \{D_1, D_2, \ldots, D_K\}$ are formed by selecting representative cluster features, ensuring that each dictionary captures the intrinsic characteristics of its corresponding cluster. This cluster-specific initialization provides a meaningful starting point for dictionary learning, leading to faster convergence and more discriminative sparse representations.

In this first dictionary learning layer, convolutional dictionary learning is applied to learn sparse representations that preserve local spatial structures. The goal of this layer is to learn a set of convolutional dictionaries that can effectively capture the local spatial structures in the image features extracted from DenseNet121.

For the $i^{th}$ image, the feature map $Y_i \in \mathbb{R}^{h*w*f}$. The first layer of dictionary learning on this feature map is as follows:

$$min_{D^1,\,\{X_i^1\}} \frac{1}{N} \sum_{i=1}^{N} \left( \frac{1}{2} \parallel D^1 * X_i^1 - Y_i \parallel_2^2 + \lambda_{D^1} \psi\left(D^1\right) + \lambda_{X^1} \phi\left(X_i^1\right) \right) \tag{4}$$

In this, the term $D^1 = \left\{D_k^1\right\}_{k=1}^{K}$, is the deep convolutional dictionary with K number of kernels or filters, $D_k^1 \epsilon \mathbb{R}^{c*c*f}$, here $c*c$ is the spatial dimension of the kernel, $f$ is the channels in an image, and the $X_i^1$ denotes the sparse feature maps to the corresponding dictionary atoms. The sparse matrix dimensions are given like $X_i^1 = \mathbb{R}^{h*w*K_1}$, and each channel in $X_i^1\ [:,:,k]$, denotes the activation of the $k^{th}$ dictionary atom for the $i^{th}$ image feature. ${D_k}^1 * X_{i,k}^1$, represents the convolution operation on the $k^{th}$ kernel to the $k^{th}$ sparse feature map. $\lambda_D$ and $\lambda_X$ are regularization parameters for the dictionary and the sparse coefficients, respectively. Here, these hyperparameters were empirically determined based on preliminary grid search evaluations on a held-out validation set during the 5-fold cross-validation procedure. Specifically, for $\lambda_X$ (sparsity regularization parameter), we tested a range of values: {0.001,0.01,0.1,1.0}, in this research the value for $\lambda_X = 0.1$, is chosen. On the other hand, $\lambda_D$ regularizes the dictionary D to prevent it from overfitting. In this research, the value for $\lambda_D = 0.001$ is chosen, which consistently produced the best trade-off between sparsity and reconstruction accuracy, and prevents overfitting across both datasets (NIH Chest X-ray and IRMA) during validation.

To promote smoothness in the kernels and minimize noise, the regularization terms play a vital role. In the above equation, $\psi\left(D^1\right)$ penalize the large dictionary atoms and validate the smoother kernels.

$$\psi\left(D^1\right) = \sum_{k=1}^{K} \parallel D_k^1 \parallel_2^2 \tag{5}$$

The other regularization term $\phi\left(X_i^1\right)$, enhances the sparsity in the $X_i^1$ coefficients by minimizing the noise. This $\mathcal{L}_1$-norm regularization confirms that only a small number of dictionary atoms are active for each image feature. Actually, $\mathcal{L}_1$-norm is a way of measuring the "size" or "magnitude" of a vector by summing up the absolute values of its components.

$$\phi\left(X_i^1\right) = \parallel X_i^1 \parallel_1 \tag{6}$$

Before aggregating the dictionaries, Dictionary Coherence is used to evaluate the similarity or distinctiveness of the basis vectors (atoms) within the dictionaries. Low coherence indicates that the dictionary atoms are diverse and represent a wide range of features in the data, which is essential for representing variability in image features. Conversely, high coherence suggests that the atoms are overly similar, leading to redundancy and potential overfitting. By reducing coherence, we ensure that only distinct, meaningful dictionaries are aggregated, contributing to improved generalization and reduced model complexity.

The coherence $\mu\ (D_{Conv})$ of a convolutional dictionary $D_{Conv} = \{d_1, d_2, \ldots .d_k\}$, is defined as the maximum absolute inner product (cosine similarity) between any two distinct atoms $d_i$ and $d_j$, normalized by their norms:

$$\mu\ (D_{Conv}) = max_{i \neq j} \frac{\left| d_i^T d_j \right|}{\| d_i \| \| d_j \|} \tag{7}$$

Where, $d_i$ and $d_j$ are two distinct dictionary atoms of two different kernels, $d_i^T d_j$ is the dot product between $d_i$ and $d_j$.

(F) Sparse matrices Aggregation:

After the first dictionary learning layer, we have multiple sparse matrices $\left\{ X_i^{(1,k)} \right\}_{k=1}^{K}$, where each $X_i^{(1,k)}$ is the sparse coefficient matrix for the *kth* dictionary atom in the first layer. There are two major ways to aggregate these sparse matrices into a single sparse matrix: concatenation and max-pooling. Upon the advantage of max-pooling over concatenation in terms of precise sparse representation, max-pooling has been observed in this research. The max-pooling operation is applied to select the strongest activation at each spatial location. This results in a more compact and focused representation, which is then passed to the second dictionary learning layer for further refinement. Hence, the max-pooling aggregates the information from each kernel and produces a single sparse matrix called the aggregated sparse matrix $X_i^{aggregated}$, for the second dictionary layer.

For every spatial location $(x_1, x_2)$, in the sparse coefficient matrices, max-pooling selects the maximum value across all the kernel matrices $\left\{ X_i^{(1,k)} \right\}_{k=1}^{K}$. The computation procedure for $X_i^{aggregated}$ is follows:

$$X_i^{aggregated}\,(x_1, x_2) = max_{k=1}^{K} X_i^{(1,k)}\,(x_1, x_2) \tag{8}$$

Here, $X_i^{aggregated} \in \mathbb{R}^{h*w}$, is the ensuing aggregated sparse matrix for an image from one particular cluster. $X_i^{(1,k)}\,(x_1, x_2)$, is the value of sparse coefficient at the given spatial location of $(x_1, x_2)$ for k-kernel.

Second layer of Dictionary learning with Aggregated Matrix:

The single aggregated sparse matrix $X_i^{aggregated}$, from the first dictionary layer produced by the max-pooling operation, serves as the input to the second layer. This layer further refines the sparse representation by solving the following optimization problem:

$$min_{D^2, \{X_i^2\}} \frac{1}{|C_k|} \sum_{i \in C_k} \left( \frac{1}{2} \| D^2 * X_i^2 - X_i^{aggregated} \|_2^2 + \lambda_{D^2} \psi \left( D^2 \right) + \lambda_{X^2} \phi \left( X_i^2 \right) \right) \tag{9}$$

Each dictionary learning layer alternates between sparse coding via Orthogonal Matching Pursuit, OMP and dictionary update via Singular Value Decomposition, SVD. Convergence is achieved when the relative change in reconstruction error between two successive iterations becomes smaller than $10^{-4}$.

(G) k-means iteration:

After the second layer of dictionary learning, the refined sparse matrices for each cluster $\left\{ X_i^2 \right\}$ are used to recalibrate k-means clustering. The updated clusters $\{C_k\}_{k=1}^{K}$ are formed based on the new sparse features. This iterative process between two layers of dictionary learning and k-means clustering ensures progressive feature refinement, leading to improved clustering and image retrieval results. The outer k-means loop iteratively refines clusters based on the updated sparse features obtained from the two-layer dictionary learning. The process stops when the cluster centroids stabilize, measured by the relative centroid movement falling below a threshold of $10^{-3}$ is reached. This ensures that the iterative clustering dictionary refinement process converges efficiently without overfitting or unnecessary computation. The results section that follows demonstrates the effectiveness of the proposed framework compared with existing methods on unsupervised medical image retrieval.

## Experimental analysis

The proposed framework was evaluated on two benchmark databases: NIH Chest X-ray and IRMA. The NIH Chest X-ray database was released by the National Institutes of Health for medical research, particularly to train deep learning models for medical image analysis and the detection and diagnosis of thoracic diseases[42]. This dataset consists of 112,120 frontal-view X-ray images from 30,805 unique patients. It covers 14 disease categories, including "No Findings," and many images are associated with more than one disease. Sample X-ray images for different diseases from the database are shown in Fig. 3. In this experimental setup, we employed a 5-fold cross-validation strategy to ensure robust performance evaluation. Specifically, the dataset was randomly divided into five equal parts and, all images were resized to 224 × 224 pixels to match the input requirements of DenseNet121. In each fold, three parts were used for training, one part for validation (for iterative optimization such as k-means and feature learning), and the remaining part for testing retrieval performance. This process was repeated five times so that each subset served as the test set once, and the average performance across all folds was reported for all evaluation metrics. This strategy helps mitigate overfitting and ensures that our retrieval framework generalizes well across different subsets of data. The same procedure was consistently applied to both

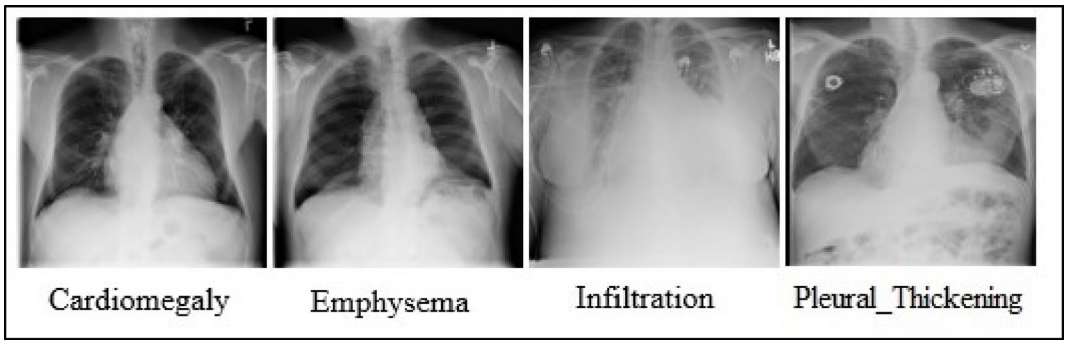



**Fig. 3**. Sample images from NIH Chest X-ray dataset.

| Algorithm | % Precision | %Recall | %F1_Score | %mAP | Cluster Purity |
|---|---|---|---|---|---|
| SCNN | 66.37 | 65.82 | 66.09 | 64.25 | 0.6 |
| SPM | 65.28 | 64.97 | 65.12 | 63.94 | 0.68 |
| ResNet50 | 68.51 | 67.28 | 67.88 | 65.21 | 0.71 |
| DenseNet121 | 71.23 | 69.17 | 70.18 | 69.41 | 0.77 |
| DGFD | 72.49 | 70.89 | 71.68 | 70.56 | 0.78 |
| ResNet50 + DL | 74.54 | 73.28 | 73.9 | 73.86 | 0.81 |
| DenseNet121 + DL | 75.15 | 74.86 | 75.01 | 74.38 | 0.84 |
| DeepCluster (ResNet50 Backbone) | 82.01 | 80.27 | 81.13 | 77.65 | 0.85 |
| SwAV (ResNet50 backbone) | 81.43 | 80.35 | 80.88 | 77.24 | 0.84 |
| DDicCIR | 84.53 | 83.24 | 83.88 | 82.28 | 0.87 |

**Table 1**. Comparative analysis of existing algorithms over the proposed framework for $n = 20$ & no.of clusters k = 15 on NIH-Chest X-Ray dataset.

the NIH Chest X-ray and IRMA datasets. The proposed framework is implemented on NVIDIA Tesla V100 GPU (32GB VRAM), Intel(R) Core(TM) Ultra 7 155 H (1.40 GHz), 32GB RAM, PyTorch 1.12 for DenseNet121 feature extraction, scikit-learn for K-means clustering and evaluation metrics, and SPAMS library for dictionary learning and sparse coding.

Based on prior knowledge of the dataset, 15 clusters were chosen for k-means clustering, and the process was iterated until convergence. Instead of using DenseNet121 with pretrained weights, we trained the model from scratch on 40,000 samples from the NIH Chest X-ray dataset to extract fine-grained features for clustering. Once the model was trained, image features were extracted for all database images. The significance of the proposed framework is demonstrated by comparing it with several popular methods on this dataset. Specifically, we compared our approach with Sparse Convolutional Neural Networks[43], Spatial Pyramid Matching[44], DenseNet121 features[25], ResNet50 features[45], DenseNet121 + Dictionary Learning (DL)[2], ResNet50 + DL, DGFD[21]. While doing comparison all baseline models, including DeepCluster and SwAV, were employed with their original backbone architectures and configurations as proposed in their respective studies. No re-training or architectural modification was performed.

The comparison was carried out in terms of precision, recall, F1-score, and cluster purity for the top 20 retrieved images, using the equations defined in Eqs. (10)–(13). Cosine similarity, as given in Eq. (14), was used to measure similarity between the query feature vector and the feature vectors of all database images, with higher similarity values indicating greater relevance. The comparative analysis of the proposed framework against existing methods is presented in Table 1, with graphical representations shown in Figs. 4 and 5. The choice of k = 15 clusters was made empirically to achieve an optimal balance between over-segmentation and under-representation of features. This value was selected after conducting preliminary experiments with varying cluster counts (ranging from 5 to 30), where k = 15 yielded the best trade-off between retrieval accuracy and computational efficiency[46].

The mathematical representation for the metrics used is as follows:

$$Precision = \frac{No.of\ relevant\ images\ retrieved}{Total\ No.of\ images\ retrieved} \tag{10}$$

$$Recall = \frac{No.of\ relevant\ images\ retrieved}{Total\ No.of\ relevant\ images\ in\ the\ database} \tag{11}$$

$$F1\_Score = \frac{2X\,Precision X\,Recall}{Precision + Recall} \tag{12}$$

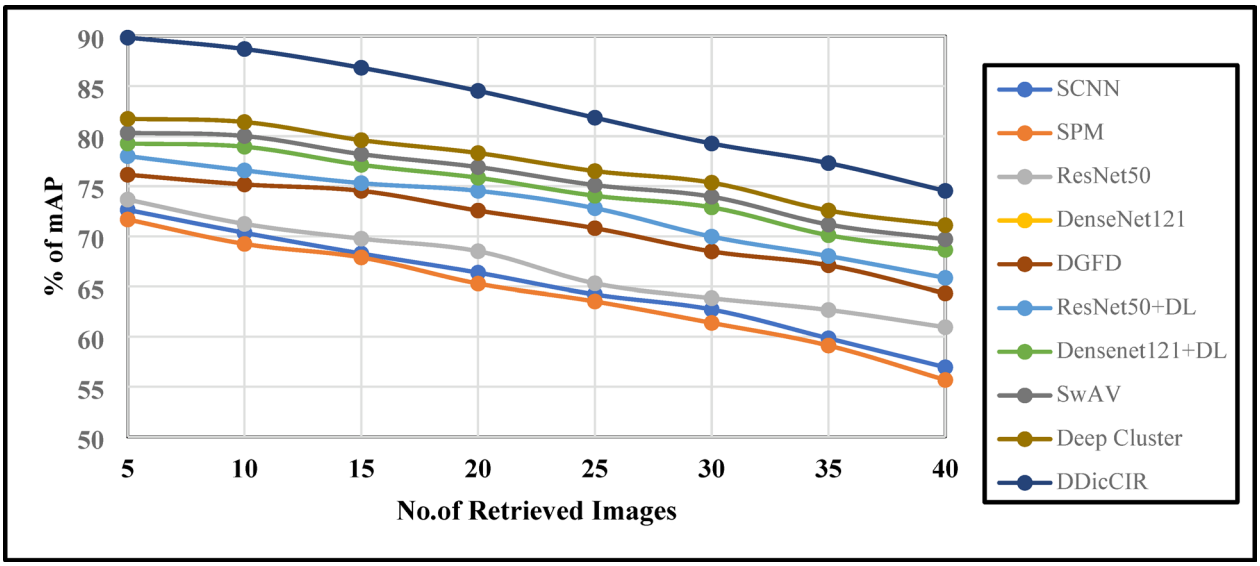



**Fig. 4**. Graphical representation of mAP vs. No.of retrieved images of NIH-chest X-Ray dataset.

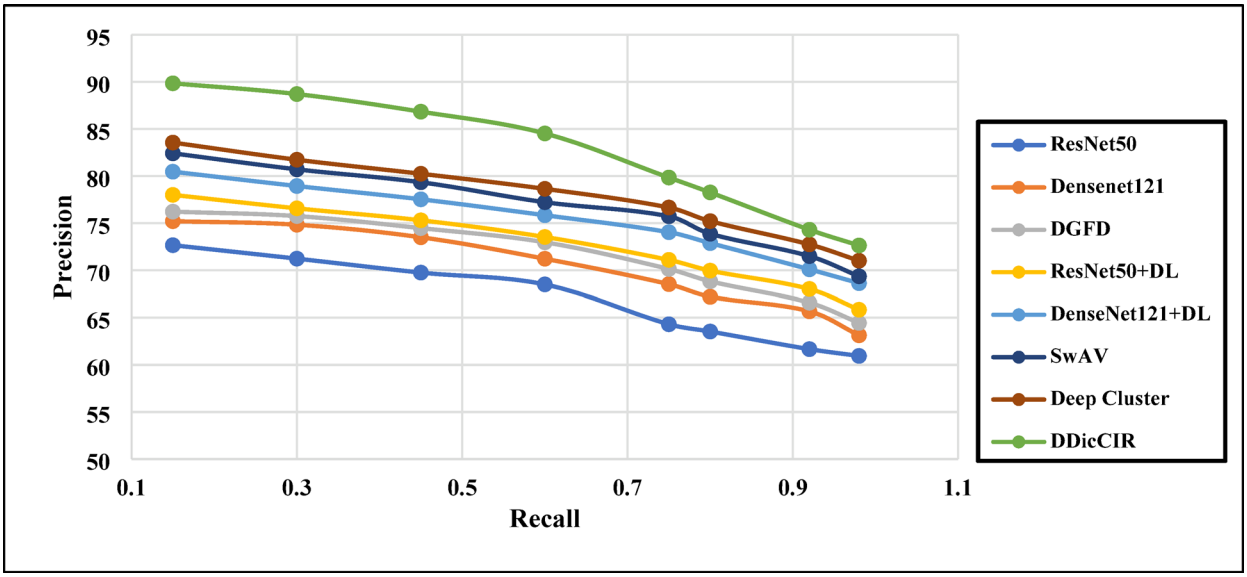



**Fig. 5**. Precision vs. recall for existing approaches to the proposed approach on NIH Chest X-ray dataset.

| Method | Inference Time (ms/query) | Relative accuracy (mAP) |
|---|---|---|
| ResNet50 (feature extraction only) | 8 ms | 0.72 |
| DenseNet121 (feature extraction only) | 10 ms | 0.74 |
| DeepCluster | 11 ms | 0.77 |
| SwAV | 13 ms | 0.76 |
| DDicCIR | 12 ms | 0.86 |

**Table 3**. The inference time of trained models is shown here for a query image.

| Method | nDCG@5 | nDCG@10 |
|---|---|---|
| ResNet50 + DL | 0.71 | 0.68 |
| DenseNet121 + DL | 0.75 | 0.72 |
| DeepCluster | 0.78 | 0.74 |
| SwAV | 0.8 | 0.76 |
| DDicCIR | 0.86 | 0.83 |

**Table 4**. nDCG results on NIH chest X-ray dataset. Signifiacnce value bold.

$$mAP = \frac{1}{n} \sum_{\substack{i \in Test \\ Database}} P_i \tag{13}$$

$$Cosine\ Similarity = \frac{f_{db}.I_q}{\| f_{db} \| \| I_q \|} \tag{14}$$

Here, $f_{db}, I_q$ are feature vectors of the database and the query image feature vector, respectively. In this work, mAP was computed using a top-N truncation strategy, denoted as mAP@N, where N ∈ {5, 10, 15, 20, 25, 30, 35, 40}. This measures the retrieval accuracy within the most relevant subset of results, which is clinically meaningful for medical image retrieval scenarios.

*Normalized discounted cumulative gain (nDCG)* Since the NIH Chest X-ray dataset is multi-label, a graded relevance measure is required. We therefore adopt nDCG, which evaluates both the relevance and ranking position of retrieved results. The DCG at cutoff *k* is defined as.

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{log_2(i+1)} \tag{15}$$

where $rel_i$ is the graded relevance score of the $i^{th}$ retrieved item. The ideal DCG (IDCG@k) is the maximum possible DCG for the given query. The normalized score is then:

$$nDCG@k = \frac{DCG@k}{IDCG@k} \tag{16}$$

For multi-label datasets such as NIH Chest X-ray, the relevance score $rel_i$ for a retrieved image is defined as the normalized overlap between the label sets of the query and the retrieved image:

$$rel_i = \frac{| L_q \cap L_i |}{| L_q |} \tag{17}$$

where $L_q$ and $L_i$ are the label sets of the query and the $i^{th}$ retrieved image, respectively.

This definition ensures that partially relevant images receive fractional relevance values, consistent with graded relevance protocols in multi-label retrieval tasks. A higher nDCG indicates better ranking quality. In our experiments, we report nDCG@5 and nDCG@10 to evaluate retrieval effectiveness at different cutoff levels.

Cluster purity is the metric given in Eq. (18) used to evaluate and validate the quality of clustering algorithm results in terms of containing data points from a single ground truth class and its performance. It helps quantify how well the clustering associates with the original labels or categories.

$$Cluster\ Purity = \frac{1}{N} \sum_{i=1}^{k} |C_i| \cdot \frac{No.of\ points\ in\ C_i\ that\ belongs\ m\,(C_i)}{|C_i|} \tag{18}$$

Here, $C_i$ is $i^{th}$ cluster, $m\,(C_i)$ is the majority of $C_i$ cluster, N is the total no.of data points. The cluster purity value varies from 0 to 1, 1 for pure clustering (i.e., maximum no.of similar data points), 0 is for the worst possible clustering. In Fig. 6, the images of one particular cluster is shown with the no.of overlapping images for the top 20 ranked in that cluster. To demonstrate efficiency and scalability, we have compared the average inference time per query with existing retrieval methods are shown in Table 3. Results show that our method achieves competitive inference time compared to DenseNet121-only retrieval and traditional dictionary learning, while offering significantly higher retrieval accuracy.

As predicted, nDCG@10 rankings are very slightly below nDCG@5 rankings for all methods as shown in Table 4, since a higher number of lower-ranking items generally decreases ranking quality on average. However, ours DDicCIR framework still has the overall highest ranking at both cutoff levels, validating that it can bring highly relevant images forward in the ranked list early a key necessity in medical image retrieval.

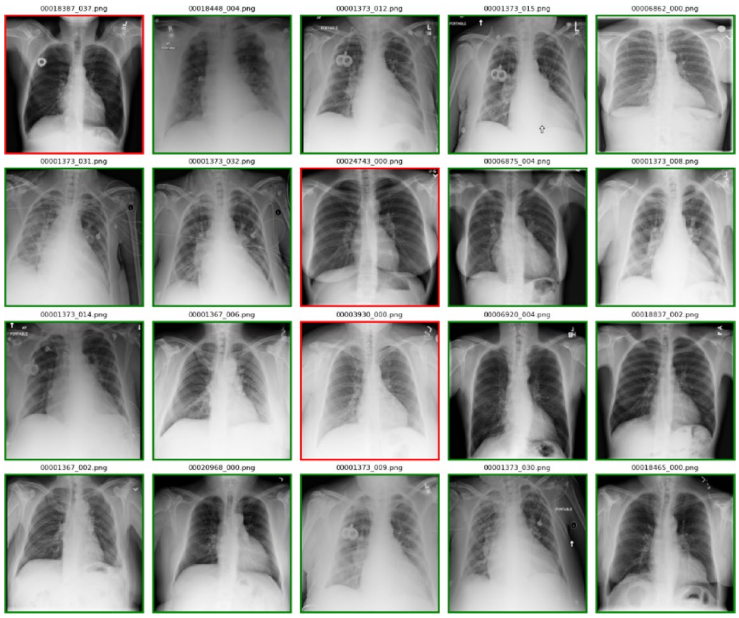


**Fig. 6**. Cluster purity for one particular disease for the top 20 images and identification of non-related images in red boundaries.

| Aggregation Method | %Precision | %Recall | F1_Score | %mAP |
|---|---|---|---|---|
| Concatenation | 66.89 | 64.72 | 65.78 | 68.29 |
| Max-pooling | 76.54 | 72.48 | 74.45 | 74.54 |

**Table 5**. Aggregation approaches vs. metrics over $n = 40$ (No.of retrieved Images).

| No.of retrieved images | SCNN | SPM | ResNet50 | DenseNet121 | DGFD | ResNet50 + DL | DenseNet 121 + DL | SwAV | Deep Cluster | DDicCIR |
|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 72.64 | 71.68 | 73.67 | 75.22 | 76.15 | 78.01 | 79.27 | 82.42 | 83.56 | 89.85 |
| 10 | 70.35 | 69.24 | 71.24 | 74.84 | 75.18 | 76.58 | 78.95 | 80.12 | 81.74 | 88.71 |
| 15 | 68.28 | 67.89 | 69.76 | 73.51 | 74.54 | 75.31 | 77.14 | 79.35 | 80.25 | 86.84 |
| 20 | 66.37 | 65.28 | 68.51 | 71.23 | 72.57 | 74.54 | 75.85 | 77.23 | 77.65 | 84.53 |
| 25 | 64.21 | 63.52 | 65.31 | 69.54 | 70.81 | 72.81 | 74.05 | 75.74 | 76.67 | 81.85 |
| 30 | 62.72 | 61.38 | 63.84 | 67.21 | 68.49 | 69.97 | 72.89 | 73.88 | 74.22 | 79.27 |
| 35 | 59.84 | 59.12 | 62.67 | 65.98 | 67.11 | 68.03 | 70.11 | 71.51 | 72.75 | 77.31 |
| 40 | 56.95 | 55.68 | 60.95 | 63.14 | 64.32 | 65.85 | 68.64 | 69.37 | 71.02 | 74.54 |

**Table 2**. Comparison of existing approaches to proposed framework over % of mAP vs. no.of retrieved images from NIH-Chest X-ray dataset.

In this research to generate an aggregated sparse matrix from the first layer dictionary learning to the second layer of dictionary learning can be done by two approaches, as explained: concatenation and max-pooling. The following Table 5 tries to demonstrate the significance of the max-pooling approach in increasing the accuracy.

To understand the significance of the proposed framework's structure, an ablation study has been conducted, which highlights the systematic improvements leading to the achieved performance. This study explains how each component contributes to the accuracy, as shown in Tables 1 and 2, and compares the framework against existing methods. Specifically, we analyze the effect on accuracy when the following components are excluded:

- *Second Dictionary Layer*: By not including this layer, we observe a lack of refinement in the sparse representation, which negatively impacts cluster purity. As shown in Table 6, the absence of this layer results in a significant decay in performance metrics, and the Table 7 is given with an extended ablation study with respect to number of Kernels, pooling layers and number of clusters.
- *k-means Reiteration*: This iterative procedure updates the clusters based on fine-grained features. We test the framework without this recalibration step to assess its importance. The results reveal the necessity of this iterative process for maintaining clustering accuracy.

| Algorithm | % Precision | %Recall | %F1_Score | %mAP | Cluster purity |
|---|---|---|---|---|---|
| Full Model (DDicCIR) | 84.53 | 83.24 | 83.88 | 82.28 | 0.87 |
| Without Second DL | 78.25 | 76.48 | 77.35 | 75.37 | 0.76 |
| Without k-means iteration | 74.47 | 72.81 | 73.63 | 73.12 | 0.69 |

**Table 6**. Ablation study of the proposed framework to represent the Stepwise accuracy.

| Variation | Setting | %Precision | %Recall | F1-Score | %mAP |
|---|---|---|---|---|---|
| Baseline | K = 64, Max-Pooling, k = 15 | 84.53 | 83.24 | 83.88 | 82.28 |
| Number of Kernels (K) | K = 32 | 81.29 | 80.35 | 80.82 | 80.32 |
| | K = 128 | 82.92 | 81.45 | 82.19 | 81.54 |
| | K = 256 | 84.83 | 83.46 | 84.14 | 83.62 |
| Aggregation method | Average-pooling | 79.81 | 78.32 | 79.05 | 78.39 |
| | Max-pooling | 84.53 | 83.24 | 83.88 | 82.28 |
| Number of clusters (k) | k = 10 | 78.45 | 77.81 | 78.13 | 78.54 |
| | k = 15 | 84.53 | 83.24 | 83.88 | 82.28 |
| | k = 20 | 81.87 | 80.65 | 81.25 | 81.23 |

**Table 7**. Extended ablation study of the proposed method with respect to the various hyperparameters.

| Iteration | $Frobenius\ Norm\ Difference = \parallel D^{(It)} - D^{(It-1)} \parallel_F$ |
|---|---|
| Iteration 1–2 | 15.34 |
| Iteration 2–3 | 12.97 |
| Iteration 3–4 | 8.21 |
| Iteration 4–5 | 5.45 |
| Iteration 5–6 | 3.89 |
| Iteration 7–8 | 1.73 |
| Iteration 8–9 | 0.84 |
| Iteration 9–10 | 0.35 |
| Iteration10-11 | 0.10 |

**Table 8**. Dictionary stability for each iteration.

| Algorithm | % Precision | %Recall | %F1_Score | %mAP | Cluster purity | IRMA error |
|---|---|---|---|---|---|---|
| ResNet50 | 72.38 | 71.27 | 71.82 | 71.41 | 0.76 | 0.26 |
| DenseNet121 | 75.46 | 74.13 | 74.78 | 73.74 | 0.79 | 0.22 |
| DGFD | 76.85 | 76.08 | 76.84 | 75.81 | 0.80 | 0.22 |
| ResNet50 + DL | 78.95 | 77.52 | 78.23 | 77.62 | 0.83 | 0.19 |
| DenseNet121 + DL | 82.57 | 81.33 | 81.94 | 80.34 | 0.86 | 0.18 |
| DDicCIR | 89.83 | 88.45 | 89.13 | 87.18 | 0.91 | 0.15 |

**Table 8**. Comparison of the proposed framework over existing on IRMA dataset for $n = 20$ (No.of images retrieved).

*Dictionary stability* In this unsupervised learning, the number of iterations will be taken as per the convergence of clusters. Dictionary stability is the key indicator of convergence in dictionary learning. If the dictionary is still changing significantly between iterations, the learning process hasn't converged yet. Stabilizing the dictionary implies that the atoms have settled and are effectively representing the input data. To measure dictionary stability, we need to compute Frobenius norm difference between the dictionaries at consecutive iterations during the learning process. A very low difference suggests the dictionary has converged and it is given in Table 8.

$$Frobenius\ Norm\ Difference = \parallel D^{(It)} - D^{(It-1)} \parallel_F \tag{19}$$

Here, $D^{(It)}, D^{(It-1)}$ are dictionary at iteration *It* and *It-1*.

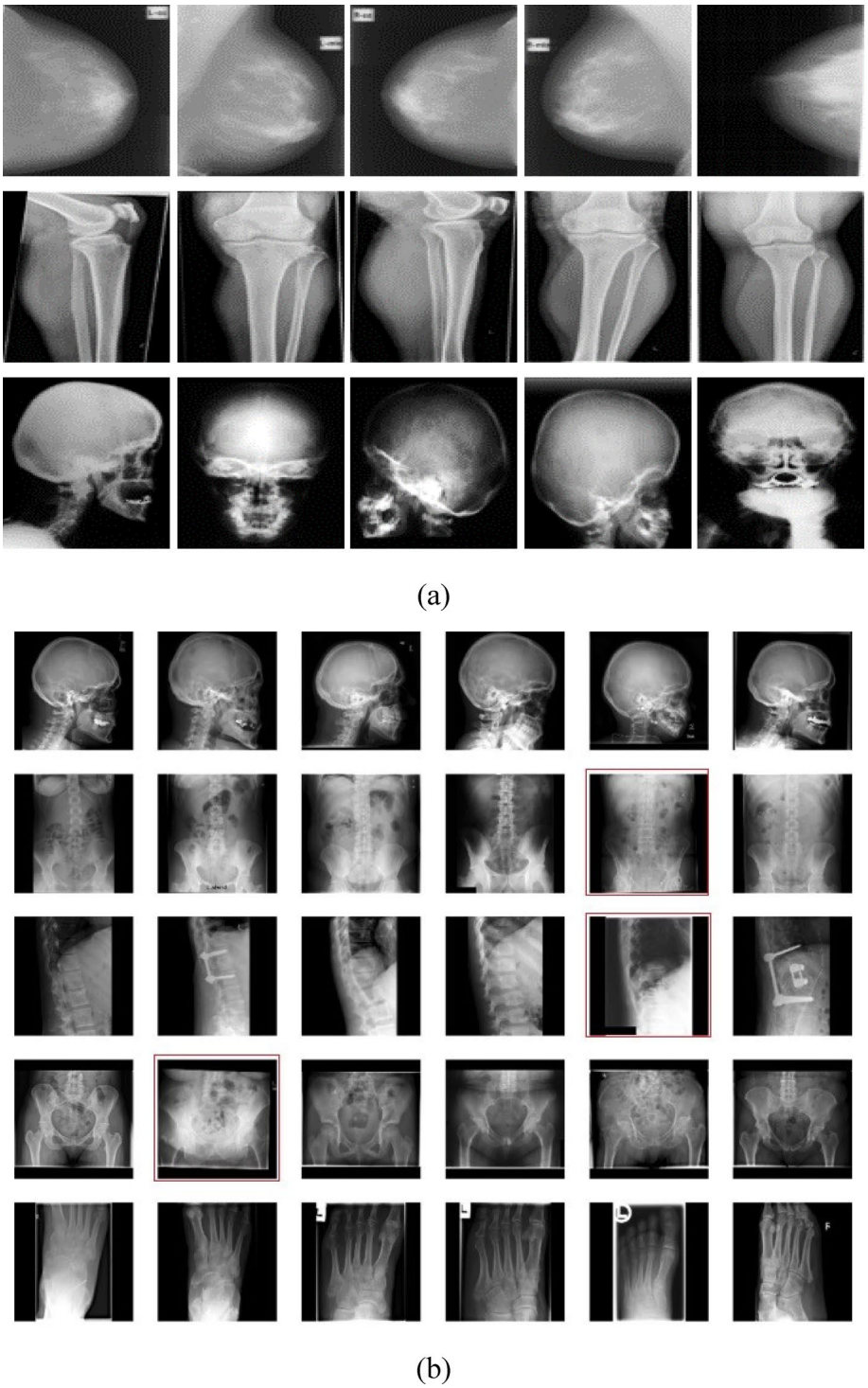


**Fig. 7**. (**a**) Sample images from the IRMA Dataset and retrieved results in (**b**).

The IRMA (Image Retrieval in Medical Applications) dataset is a widely used dataset for research in medical image retrieval[47]. The dataset contains 14,410 x-ray images categorized using a.

hierarchical IRMA code that encodes aspects like imaging modality, anatomical region, pathology, and orientation. The dataset provides a benchmark for testing retrieval systems, making it valuable for your research in dictionary learning and medical image retrieval. The IRMA dataset can help evaluate feature extraction and retrieval models using metrics like precision, recall, mAP, and the IRMA error score. The proposed method's significance over existing methods on IRMA is given in Table 8. The precision and recall are significantly higher as compared to the NIH Chest X-ray dataset due its' nature of the samples. The sample images and retrieved

| Parameter | Description | Final value |
|---|---|---|
| Kernel size ($DicL_1$) | Size of convolutional filters used in first-layer dictionary | 3 × 3 |
| Stride ($DicL_1$) | Stride applied during convolution in first dictionary layer | 1 |
| Padding ($DicL_1$) | Zero-padding to preserve spatial dimensions | 1 |
| Number of kernels ($DicL_1$) | Number of dictionary atoms per cluster in first layer | 128 |
| Kernel size ($DicL_2$) | Convolutional filter size in second dictionary layer | 3 × 3 |
| Stride ($DicL_2$) | Stride used in second-layer dictionary learning | 1 |
| Padding ($DicL_2$) | Padding in second dictionary layer | 1 |
| Number of kernels ($DicL_2$) | Total number of global dictionary atoms in second layer | 256 |
| OMP sparsity level | Maximum number of active atoms in sparse representation | 15 |
| k-means iterations | Maximum iterations for iterative clustering loop | 15 |
| Feature dimension | Output feature size from DenseNet121 (before dictionary learning) | 1024 |
| Image input size | Resized input image resolution | 224 × 224 |
| Regularization parameters | $\lambda_X = 0.1,\ \lambda_D = 0.001$ | |
| Convergence threshold (dictionary update) | Stop when relative change in reconstruction error $< \varepsilon_1$ | $10^{-4}$ |

**Table 9**. Hyperparameters and implementation settings of the proposed DDicCIR Framework.

results are shown in Fig. 7a, b, respectively, and he hyperparameters of the DDicCIR is given in the following Table 9.

*IRMA Error Metric:*

The IRMA error quantifies hierarchical retrieval accuracy on the IRMA dataset by comparing the IRMA codes of the query and retrieved images. Each IRMA code encodes four hierarchical axes: technical (T), directional (D), anatomical (A), and biological (B). Errors in higher-level positions are penalized more heavily than those in deeper levels. The error is defined as Eq. (20):

$$E_{IRMA} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{4}\frac{1}{b_j}\delta\left(l_{ij},\ \widehat{l_{ij}}\right) \tag{20}$$

where $b_j$ is the number of code branches and $\delta\left(l_{ij}, \widehat{l_{ij}}\right)$ equals 1 if the $j^{th}$ code differs and 0 otherwise. The final error value lies in $[0,1]$, with smaller values indicating higher retrieval consistency. In this study, the proposed DDicCIR achieves an IRMA error of approximately 0.15, outperforming recent unsupervised deep clustering methods.

## Conclusion

The proposed Deep Dictionary Clustering for Image Retrieval (DDicCIR) framework offers an unsupervised approach for efficient image retrieval using clustering techniques. Its effectiveness was evaluated on two benchmark datasets: NIH Chest X-ray and IRMA. The two-layer dictionary learning approach refines image features, thereby enhancing both clustering and retrieval performance. First, image features are extracted using DenseNet121, which are then supplied to the two-layer deep dictionary model for fine clustering via k-means. The first dictionary learning layer captures diverse, high-level features using convolutional filters, while the second layer further fine-tunes these representations to extract more detailed and abstract patterns. To eliminate redundancy and promote diversity among the learned atoms (basis vectors) in the first layer, dictionary coherence is applied. On the NIH Chest X-ray dataset, DDicCIR achieved significant improvements in precision, recall, mean Average Precision (mAP), nDCG@5, nDCG@10, and cluster purity, indicating better feature representation and more accurate retrieval of relevant medical images. Similarly, on the IRMA dataset, the framework demonstrated notable improvements across these metrics. By integrating multi-layer dictionary learning with clustering, the DDicCIR framework provides a scalable and robust solution for medical image retrieval. Its ability to capture diverse image features and improve retrieval accuracy makes it a valuable tool for supporting diagnostic decisions, enabling healthcare professionals to quickly and accurately retrieve similar medical cases.

## Data availability

This article used publicly available datasets, no private is used. The web links for the datasets used are given below.https://www.kaggle.com/datasets/nih-chest-xrays/data, https://www.kaggle.com/datasets/raddar/irma-xray-dataset.

Received: 27 June 2025; Accepted: 15 December 2025
Published online: 29 December 2025

## References


1. Jin, Q. et al. Iterative pseudo-labeling based adaptive copy-paste supervision for semi-supervised tumor segmentation. *Knowl. Based Syst.* **324**, 113785 (2025).
2. Jin, Q. et al. Inter-and intra-uncertainty based feature aggregation model for semi-supervised histopathology image segmentation. *Expert Syst. Appl.* **238**, 122093 (2024).
3. Liu, Y. et al. A survey of content-based image retrieval with high-level semantics. *Pattern Recogn.* **40**(1), 262–282 (2007).
4. Lehmann, T. M. et al. Automatic categorization of medical images for content-based retrieval and data mining. *Comput. Med. Imaging Graph.* **29**(2-3), 143–155 (2005).
5. Sucharitha, G. et al. Efficient Image Retrieval Technique with Local Edge Binary Pattern Using Combined Color and Texture Features. *Computational Intelligence for Engineering and Management Applications: Select Proceedings of CIEMA 2022* 261–276 (Springer Nature Singapore, Singapore, 2023).
6. Mizotin, M. et al. Feature-based brain MRI retrieval for Alzheimer disease diagnosis. in *2012 19th IEEE International Conference on Image Processing* (IEEE, 2012).
7. Zeng, X. et al. Uncertainty Co-estimator for improving Semi-Supervised medical image segmentation. *IEEE Trans. Med. Imaging* (2025).
8. Arora, Nitin, G. & Sucharitha Sharma. MVM-LBP: Mean – Variance – Median based LBP for face recognition. *Int. J. Inform. Technol.* **15**(3), 1231–1242 (2023).
9. Tang, H. et al. When dictionary learning meets deep learning: Deep dictionary learning and coding network for image recognition with limited data. *IEEE Trans. Neural Networks Learn. Syst.* **32**(5), 2129–2141 (2020).
10. Wang, H., Li, G. & Chih-Ling, T. Regression coefficient and autoregressive order shrinkage and selection via the Lasso. *J. Royal Stat. Soc. Ser. B: Stat. Methodol.* **69**(1), 63–78 (2007).
11. Du, G. et al. Dual diversity and pseudo-label correction learning for semi-supervised medical image segmentation. *Int. J. Imaging Syst. Technol.* **35**(5), e70194 (2025).
12. Aharon, M., Elad, M. & Bruckstein, A. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal. Process.* **54**(11), 4311–4322 (2006).
13. Liu, H. et al. Sequential Bag-of-Words model for human action classification. *CAAI Trans. Intell. Technol.* **1**(2), 125–136 (2016).
14. Feng, D. et al. The noise attenuation and stochastic clutter removal of ground penetrating radar based on the K-SVD dictionary learning. *IEEE Access.* **9**, 74879–74890 (2021).
15. Lu, H. et al. Efficient Multi-View k-means for image clustering. *IEEE Trans. Image Processing* (2023).
16. Cai, S. et al. A dictionary-learning algorithm based on method of optimal directions and approximate K-SVD. *2016 35th Chinese control conference (CCC)* (IEEE, 2016).
17. Sudhish, D. K. & Latha, R. Nair. Content-based image retrieval for medical diagnosis using fuzzy clustering and deep learning. *Biomed. Signal Process. Control.* **88**, 105620 (2024).
18. Weng, X. et al. Unsupervised visual similarity-based medical image retrieval via dual-encoder and metric learning. *Neurocomputing* **634**, 129861 (2025).
19. Yu, H. et al. Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives. *Neurocomputing* **444**, 92–110 (2021).
20. Wang, X. et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017).
21. He, Q. Unsupervised Deep-Embedding global feature descriptor for image retrieval. *Circuits Syst. Signal. Process.* **43**, 2251–2272 (2024).
22. Rehman, S. et al. Optimization of CNN through novel training strategy for visual classification problems. *Entropy* **20** (4), 290 (2018).
23. Zheng, H., Yong, H. & Zhang, L. Deep convolutional dictionary learning for image denoising. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 630–641 (2021).
24. Chhabra, M. & Kumar, R. A smart healthcare system based on classifier DenseNet 121 model to detect multiple diseases. *Mobile Radio Communications and 5G Networks: Proceedings of Second MRCN 2021* 297–312 (Springer Nature Singapore, Singapore, 2022).
25. Rochmawanti, O. & Utaminingrum, F. Chest X-ray image to classify lung diseases in different resolution size using DenseNet-121 architectures. in *Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology* (2021).
26. Huang, Z. et al. Medical image classification using a light-weighted hybrid neural network based on PCANet and densenet. *IEEE Access* **8**, 24697–24712 (2020).
27. Wei, C. P. & Wang, Y. C. F. Undersampled face recognition via robust auxiliary dictionary learning. *IEEE Trans. Image Process.* **24**, 1722–1734 (2015).
28. Sulam, J. et al. Trainlets: dictionary learning in high dimensions. *IEEE Trans. Signal Process.* **64**, 3180–3193 (2016).
29. Zhang, F. et al. Dictionary pruning with visual word significance for medical image retrieval. *Neurocomputing* **177**, 75–88 (2016).
30. Regan, J. & Khodayar, M. A triplet graph convolutional network with attention and similarity-driven dictionary learning for remote sensing image retrieval. *Expert Syst. Appl.* **232**, 120579 (2023).
31. Madhuri, G. and Atul Negi. Discriminative dictionary learning based on statistical methods. in *Statistical Modeling in Machine Learning* 55–77 (Academic Press, 2023).
32. Arun, K. S. & Govindan, V. K. Optimizing visual dictionaries for effective image retrieval. *Int. J. Multimedia Inform. Retr.* **4**(3), 165–185 (2015).
33. Zhang, L. Low-rank decomposition and laplacian group sparse coding for image classification. *Neurocomputing* **135**, 339–347 (2014).
34. Caron, M., Bojanowski, P., Joulin, A. & Douze, M. Deep clustering for unsupervised learning of visual features. in *Proceedings of the European conference on computer vision (ECCV)* 132–149 (2018).
35. Caron, M. et al. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural. Inf. Process. Syst.* **33**, 9912–9924 (2020).
36. Sha, L., Schonfeld, D. & Wang, J. Graph laplacian regularization with sparse coding for image restoration and representation. *IEEE Trans. Circuits Syst. Video Technol.* **30**(7), 2000–2014 (2019).
37. Nozaripour, A. & Soltanizadeh, H. Image classification via convolutional sparse coding. *Visual Comput.* **39**(5), 1731–1744 (2023).
38. Wang, L. et al. Convolutional sparse representation and local density peak clustering for medical image fusion. *Int. J. Pattern recognit. Artif. Intell.* **34**, 2057003 (2020).
39. Sadik, S. Et-tolba, and Benayad Nsiri. An efficient multivariate approach to dictionary learning for portfolio selection. *Digit. Signal Proc.* **153**, 104647 (2024).
40. Rubinstein, R., Faktor, T. & Elad, M. K-SVD dictionary-learning for the analysis sparse model. *IEEE International Conference on Acoustics, Speech and* Signal Processing *(ICASSP)* (IEEE, 2012).
41. Garcia-Cardona, C. & Wohlberg, B. Convolutional dictionary learning: A comparative review and new algorithms. *IEEE Trans. Comput. Imaging.* **4**(3), 366–381 (2018).
42. Hsu, J., Lu, P. & Khosla, K. Predicting thorax diseases with NIH chest X-rays 4–8. (2017).
43. Lei, Y. et al. High-resolution CT image retrieval using sparse convolutional neural network. in *Medical Imaging 2018: Physics of Medical Imaging*. Vol. 10573 (SPIE, 2018).

44. Karmakar, P. et al. An enhancement to the spatial pyramid matching for image classification and retrieval. *IEEE Access.* **8**, 22463–22472 (2020).
45. Rajpal, S. et al. Using handpicked features in conjunction with ResNet-50 for improved detection of COVID-19 from chest X-ray images. *Chaos Solitons Fractals*. **145**, 110749 (2021).
46. Mehar, A. et al. Determining an optimal value of K in k-means clustering. in *2013 IEEE International Conference on Bioinformatics and Biomedicine* (IEEE, 2013).
47. Lehmann, T. et al. The IRMA code for unique classification of medical images. in *Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation*. Vol. 5033 (SPIE, 2003).

## Author contributions

G.S.- Framework Modelling, execution. J.R.- Result analysis, comparative analysis. V.B.-Literature Survey. P.V.B.R.- Software design. S.S.- Documentation. N.D.- Data Collection, Model Testing and Validation. O.O.- Documentation and verification.

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to J.R.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.