International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:10, No:1, 2016

A General Framework for Knowledge Discovery Using High Performance Machine Learning Algorithms

S. Nandagopalan, N. Pradeep

***Abstract*—**The aim of this paper is to propose a general framework for storing, analyzing, and extracting knowledge from two-dimensional echocardiographic images, color Doppler images, non-medical images, and general data sets. A number of high performance data mining algorithms have been used to carry out this task. Our framework encompasses four layers namely physical storage, object identification, knowledge discovery, user level. Techniques such as active contour model to identify the cardiac chambers, pixel classification to segment the color Doppler echo image, universal model for image retrieval, Bayesian method for classification, parallel algorithms for image segmentation, etc., were employed. Using the feature vector database that have been efficiently constructed, one can perform various data mining tasks like clustering, classification, etc. with efficient algorithms along with image mining given a query image. All these facilities are included in the framework that is supported by state-of-the-art user interface (UI). The algorithms were tested with actual patient data and Coral image database and the results show that their performance is better than the results reported already.

***Keywords*—**Active Contour, Bayesian, Echocardiographic image, Feature vector.

I. Introduction

WHILE knowledge discovery in databases (KDD) and data mining have enjoyed great popularity and success in recent years, there is a distinct lack of a generally accepted framework for data mining. The present lack of such a framework is perceived as an obstacle to the further development of the field [1].

BigData analytics involve processing heterogeneous data from various distributed data sources producing complementary datasets. Hence, the data sets are not only characterized by their extremely large volumes but also by their heterogeneity and the distributed acquisition of data.

Distributed data mining techniques have been proposed in the literature to process such distributed data sets [2]. Our framework, though, does not propose a distributed data mining framework it does address heterogeneous data types such as medical images and general datasets. Efficient algorithms are proposed to extract features using SQL based queries in order to build a unified feature database.

Fig. 1 shows the four layer architecture of our proposed general framework. Each layer is designed to perform specific task(s). The bottom most layer i.e. Physical Storage Layer we store the raw images and general datasets. Object identification layer is responsible for segmenting the images and obtain all relevant features. For obtaining useful knowledge, data mining tasks such as clustering, classification are needed and these are carried out in Knowledge Discovery layer. Finally, user applications interact with the system through UI in the top most layer.

We consider two types of echo images: 2D echo and color Doppler echo (CDF). In addition, general color images are included in the image database. Tags are used to identify the type of image, because each type of image entails special processing for feature extraction.

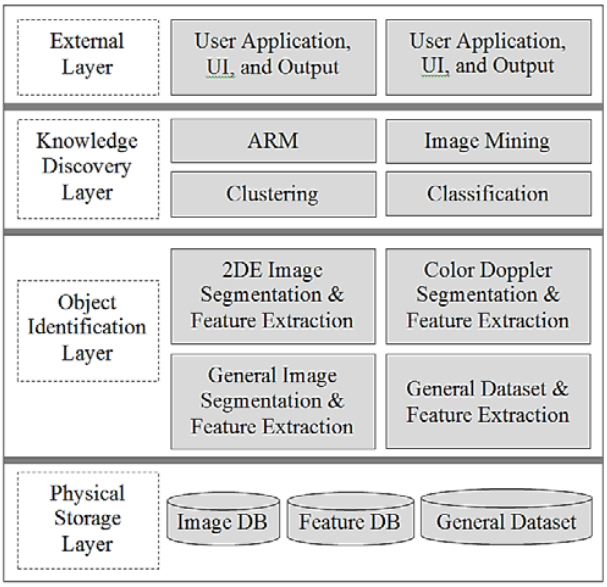


Fig. 1 General Framework Architecture

In addition to non-image datasets, general data sets having specific formats can also be analyzed with our framework. For 2D echo images segmentation, we adopt parallel k-Means clustering algorithm and then to identify the cardiac chamber (mainly Left Ventricle or LV), active contour method is used. Several quantitative features like area, width, height, ejection fraction, etc. are computed automatically. Similarly, for color

Dr. S. Nandagopalan is with the Bangalore Institute of Technology, Dept. of Computer Science & Engineering, Bangalore, India affiliated to Visvesvaraya Technological University, Belgaum, India (e-mail: snandagopalan@gmail.com).

N. Pradeep, Research Assistant and a student is with the Hochschule Bremerhaven University, Bremerhaven, Germany (e-mail: pradeepnandagopalan@gmail.com).

Doppler images qualitative features such as features color, texture, and other statistical parameters are obtained. In the case of normal images, features such as color, texture, edge density, edge histogram descriptor, etc. are extracted. All these features are combined into a feature vector for each type of image and stored in a feature database. These tasks are carried out in the Object identification layer.

Knowledge discovery layer consists of high performance data mining algorithms such as clustering, classification, etc. that are used to find some specific knowledge as per the user requirements. Through the sophisticated UI in top layer we interact with the data mining algorithms for some specific dataset and obtain the results.

II. Previous Works

Weka framework provides a large collection of machine learning algorithms written in Java for data pre-processing, classification, clustering, association rules, and visualization, which can be invoked through a common graphical user interface [3]. ELKI: Environment for Developing KDD-Applications Supported by Index-Structures is another framework which is designed to be easy to extend for researchers and students in this domain [http://elki.dbs.ifi.lmu.de/], [4]. These frameworks offer a number of algorithms for each data mining tasks with the user supplied input features in the .csv file format. However, neither they accept raw input like image data nor present the knowledge in a more readable form.

Parallel k-Means algorithms are suggested by Piotr Kraj et al. and many other authors [5], [6]. The software described in this paper is a high performance multithreaded application that implements a parallelized version of the k-Means Clustering algorithm. For LV contour tracing, Santos has adopted windows adaptive technique, many have used active contour method, and others have shown results based on template matching. A semi-automatic procedure and full automatic for segmentation of echocardiogram images based on histogram and edge detection techniques was discussed in [7].

Image mining has been researched by many authors, but only limited work has been done on echo image datasets. A multifeature based technique to retrieve 2D and color Doppler echo images is attempted by authors in [8]. Again, this system is more specific to echo image domain and does provide a general framework.

III. Proposed Framework's Process Model

In this section, we describe the proposed knowledge discovery model with the help of a block diagram as shown in Fig. 2. The echo images/video frames are gathered for each patient and stored in the image database. Each image is tagged with its type: 2DE/CDF/Normal Image so that object identification and feature extraction can be done by calling appropriate algorithm(s). For instance, 2DE images parallel k-Means algorithm followed by active contour algorithm is invoked to find the LV dimensions. In CDF images, we do not call k-Means as pixel classification method itself is sufficient to accurately find the object. Similarly, for 2DE images, quantification features are extracted, whereas for CDF images qualitative features are important. All normal images require totally different processing. Thus, the feature vectors constructed for different type of images and general dataset are stored in a separate storage called feature database. Now, high performance data mining algorithms are called to find the useful patterns in the feature vector.

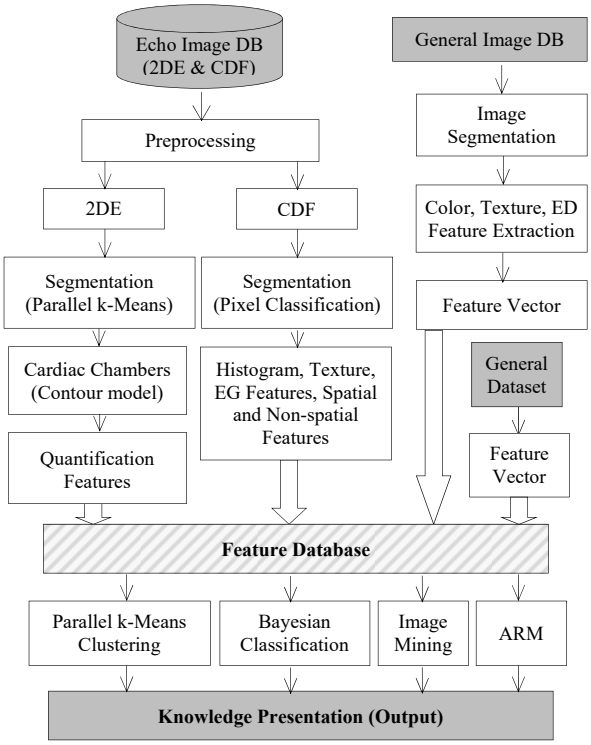


Fig. 2 Block diagram of the proposed model

To find patients with common heart problems can be done by running parallel k-Means clustering algorithm. Similarly, a patient's severity of the heart abnormality can be found out by invoking Bayesian classification algorithm. It is also possible to find a given patient's 2DE and CDF images, whether he/she is normal or abnormal. Given a general query image, we can mine all semantically relevant images (CBIR), etc. This framework concentrates in providing knowledge extracted from the input dataset. In contrast, the existing frameworks have a set of algorithms for each data mining task and the knowledge discovery process is left to the user.

IV. Proposed Algorithms

A number of efficient algorithms have been designed specifically for this research. The following sections explain the methods and algorithms in detail.

A. Parallel K-Means Clustering Algorithm

Parallel k-Means clustering algorithm is the prime task in this proposed framework. This algorithm is run on each frame of the echo video concurrently so that a speedup is achieved.

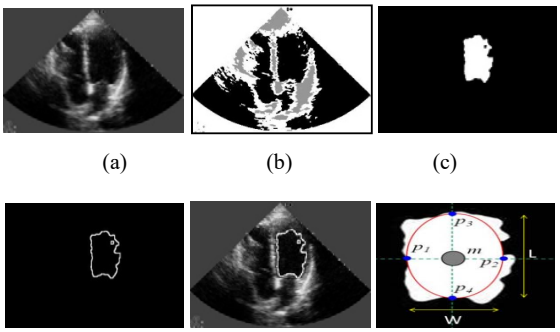
The algorithm of [5] is used for segmenting the 2DE images and is shown in Fig. 3. This algorithm needs a small modification for clustering the objects in the feature vector. That is, instead of a fixed k value, it must be a variable whose value is given as an input by the user.

Algorithm Parallel_KMeans(D, k, m)
// Input: D: The pixel dataset of frame-i, $k = 3$, m - # of iterations
// Output: Three clusters with respective pixel data
1. Select k centroids from D using random number generator
2. p = omp_get_num_proc(); // p - number of processors
3. Divide the dataset D into p subsets, D_1, D_2, ..., D_p
4. **for** $u = 1$ to m **do**
   4.1 Execute the for loop in parallel for every subset and on each p
   #pragma omp parallel shared <variables>
   #pragma omp for private <variables>
   foreach pixel j do
   • Find the nearest cluster center using Euclidean distance metrics
   • Assign/Reassign cluster id for pixel j
   Update the cluster centroids
End.

Fig. 3 Parallel k-Means algorithm

A. Active Contour Model for Quantification

A modified geometric active contour algorithm as presented in [5].



(a) (b) (c)

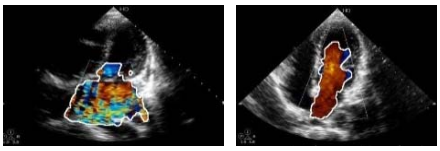
(d) (e) (f)

Fig. 4 Image Quantification (a) Original image (b) k-Means Clustered (c) LV boundary marked (d) Sobel edge detected (e) Final image (f) Quantification

The next step in processing the 2DE image is to extract clinically relevant and useful features like area, volume, ejection fraction, etc. A series of image processing and geometric methods are applied to compute these parameters automatically and is shown in Figs. 4 (a)-(e).

B. Color Doppler Image Segmentation

In the case of color Doppler images, the color object in the entire echo image is important as it provides diagnostic information. For instance, if the object contains uniform blue or red pixels it signifies that the patient is normal. However, if the object contains mosaic color pixels it is due to the turbulent flow of the blood. This type of flow indicates that the patient has stenosis or regurgitation abnormality as shown in Fig. 5.



(a) (b)

Fig. 5 Color Doppler Image Boundary tracing

Two more features, skewness and kurtosis are computed using (1) and (2). Kurtosis characterizes the relative peakedness or flatness of a distribution compared to the normal distribution and provides another useful feature to detect abnormality.

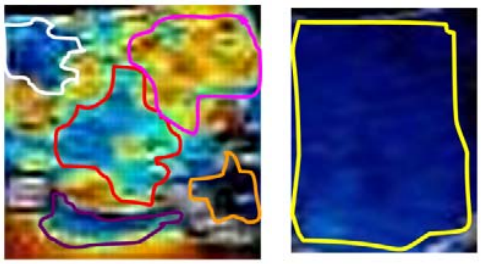
$$Skewness = \frac{\sum_{i=1}^{N}(Y_i-\mu)^3}{N\sigma^3} \quad (1)$$

$$Kurtosis = \frac{\sum_{i=1}^{N}(Y_i-\mu)^4}{N\sigma^4} - 3 \quad (2)$$

where, μ - mean, σ - standard deviation, N – number of samples, $Y_i - i^{th}$ value.

C. Hierarchical Clustering

Hierarchical connections are especially evident in the biological domain [9]. Gene Ontology classifies genes into hierarchies of biological processes and molecular functions [6]. In our work, we use the hierarchical clustering to automatically recognize and group the patterns present in the CDF images.



(a) (b)

Fig. 6 Hierarchical clustering of CDF images (a) Abnormal image (b) Normal image

Agglomerative clustering methods create a hierarchy bottom up, by choosing a pair of clusters to merge at each step. The result is a rooted binary tree. N leaves correspond to input data items called singleton clusters, and N - 1 inner nodes that are clusters correspond to groupings in coarser granularities at higher tree levels. Merge scores correspond to

dendrogram heights. This hierarchy is often used to infer knowledge from cluster statistics, as well as relatedness at varying granularities.

A sample abnormal CDF image is clustered using hierarchical clustering and is shown in Fig. 6 (a); here not all clusters are shown. The clusters are shown in different contours corresponding to the multicolor flow pattern.

Similarly, Fig. 6 (b) shows the hierarchical clustering being applied to a normal CDF image where we find a single cluster due to the laminar flow pattern. In order to discover the flow pattern a simple algorithm is followed as shown below:

Algorithm FlowPatternDiscovery(I)

Step 1: Apply the pixel classification method to extract the biggest color blob.
Step 2: Run the Hierarchical Clustering algorithm and store the spatial data of each cluster.
Step 3: Find the non-spatial features such as color histogram, edge density, etc. of each cluster.
Step 4: Build the feature vector consisting of both spatial and non-spatial data and store it in the feature database.
Step 5: **End**.

For general images, we first compute their texture. A texture is characterized by a set of values called energy, entropy

D. General Image Feature Extraction

For general images, we first compute their texture. A texture is characterized by a set of values called energy, entropy, contrast, and homogeneity. The following formulas are used to calculate the features and are shown in (3)-(6) [10]:

$$\text{Energy} = \sum_i \sum_j P_d^2(i,j) \tag{3}$$

$$\text{Entropy} = -\sum_i \sum_j P_d(i,j) \log P_d(i,j) \tag{4}$$

$$\text{Contrast} = \sum_i \sum_j (i-j)^2 P_d(i,j) \tag{5}$$

$$\text{Homogeneity} = \sum_i \sum_j \frac{P_d(i,j)}{1+|i-j|} \tag{6}$$

Edges convey essential information to a picture, and their accurate detection is of primary importance. To recognize the geometric shapes within an image is done by extracting edge details. The Edge Histogram Descriptor represents the local edge distribution in the image which is obtained by subdividing the whole image into 4 × 4 sub images. For each sub image, the edge density can be calculated using (7). Let (x_1, y_1) and (x_2, y_2) are the top left corner and the bottom right corner of the sub image. Then the edge density f is given by,

$$f = \frac{1}{a_r} \sum_{x=x1}^{x2} \sum_{y=y1}^{y2} e(x,y) \tag{7}$$

All these features are normalized and stored as a feature vector in the feature database [10].

E. General Dataset Processing

For general datasets, a separate segmentation process may not be required. For example, consider the UCI datasets Bag of Words or Diabetes that are used for clustering. The datasets are available as a set of features that can directly be used as input to the clustering algorithms. Our clustering algorithm or classification algorithm can also accept these input data and provide useful information.

F. Content Based Image Mining

Content based image mining is a technique which uses *visual contents* (features) to search images from large scale image databases according to users' requests. In the case of echo images, when the query image is that of an abnormal patient, then all patients with the same abnormality will be retrieved. Alternatively, if it is a non-echo query image, we get all similar images from the corresponding image database using the Euclidean distance measure.

The algorithm shown in Fig. 7 follows the greedy strategy to compare the similarity between the query image and the database images [1], [10].

```
Algorithm ImageMining
  // I[n] – Image DB with n images
  // Q – Query Image
  foreach (Image I in I[n]) do
      // get the segments F of image, I
      F = Hirerarchical_Clustering()
      foreach ( Segment s in F ) do
        if (Euclidean(Q[s], I[s]) < threshold)
          // continue to check other segments
        else break
          // no need to check other segments
End.
```

Fig. 7 Algorithm for Similarity comparison based on greedy strategy

Assuming that each image consists of several segments, F, which can be obtained by running hierarchical clustering algorithm, we compare the segments of the query image and the segments of the database images. To retrieve visually similar images, all the segment distances must be within threshold set. Thus, this method is efficient as the entire image descriptor need not be compared.

V. Feature Vector and Feature Database

Two different feature databases: (1) for 2D echo image and (2) for color Doppler images are used. For 2DE images 35 features are computed for cardiac chambers LV, LA, etc. {LVHeightED, LVDiameterED, LVAreaED, LVVolumeED, LVHeightES, LVDiameterES, LVAreaES, LVVolumeES, EF, FS, …..}. For CDF images a total of 18 features are extracted: {RedMean, GreenMean, BlueMean, RedSD, GreenSD, BlueSD, RedContrast, GreenContrast, BlueContrast, Energy, Entropy, Homogeneity, Skewness, Kurtosis, RedEG, GreenEG, BlueEG, ED}. For non-echo images color, texture, edge density, etc. are some of the features that are computed.

These features for each image are stored as a vector in the feature database. These vectors form the basis for knowledge

discovery process in the data mining algorithms.

VI. Classification

Classification of echocardiographic images and other datasets is an important data mining task that helps hospitals and other applications without transferring the data in any form. In this paper a novel method is proposed to accomplish this task using naïve Bayesian model via SQL [11].

In this model, all attribute values are assumed as continuous because the echo image parameters are all of numeric values. There are two phases involved in this design namely training phase and prediction phase. During the training phase the posterior probabilities and class probabilities are computed and stored in a suitable data structure. In the prediction phase, based on the test patient data, the probability is calculated by referring to the trained probability table and the final class label is obtained.

Algorithm SQL_Bayesian_Classifier(*EDT*)
[Training Phase]
Step 1: [Extract attributes from *EDT* for each *c*]
EClass ← *AClass* ⋈ *EDT*
Step 2: [Find class probability $P(C_i)$, where $i = 2$]
EPC ← *EDTgroup*(*bc*)
Step 3: [Initialize *EPX* probability with 0]
EPX ← *EClass* ⋈ *EPC*
Step 4: [Update EPX with count of tuples]
for i ← 1 to l do
UPDATE EPX ← *EDTi* ⋈ *EPX*
Step 5: [Compute $P(X|C_i)$ of all combinations]
EPXi ← *EPC* ⋈ *EPX*
End.

Algorithm SQL_Bayesian_Prediction(*EPXi*, *Target*)
[Prediction Phase]
Step 1: [Compute probabilitites w.r.t *X*]
EFP ← *EPXi* ⋈ *Target*
Step 2: [Compute $P(X|C_i) * P(C_i)$]
Update *EFP* ← *EFP* ⋈ *EPC*
Step 3: [Normalize the probabilities]
Update *EFP* ← *EFP* ⋈ *EFP*
End.

Fig. 8 Algorithms for SQL-Bayesian Training and prediction

Fig. 8 shows the algorithms for training and prediction phase through SQL approach.

VII. Experimental Setup and Results

The entire framework is designed and implemented using C#.NET environment along with Oracle 11g SQL as the DBMS. The patient images are acquired using a Philips Envisor C HD ultrasound machine with S4-2 adult probe. The training data include 27 normal and 42 abnormal patient images under four-chamber, short-axis, and color Doppler modalities.

For natural images Coral image database of 100 images in 10 categories are considered. All the programs are tested on Dell Optiplex 3010, Intel Core i3 @ 3.3GHz, 4GB RAM machine.

A. Parallel k-Means Performance

The computational performance of parallel k-Means is tested by considering varying data sizes keeping the dimensions to 3 for synthetic data. It is then compared with its sequential performance and is shown in Table I.

TABLE I
Computational Performance of Parallel k-Means versus Sequential k-Means, *P* - number of processors, *D* = 3, number of Dimensions

Data Size (n)	Sequential k-Means (Secs)	Parallel k-Means p=4 (Secs)	Speedup $S_p = \frac{T_s}{T_p}, 4$
1K	0.031	0.058	-
10K	0.393	0.521	-
100K	4.406	2.19	2.0
500K	21.107	8.76	2.4
1000K	48.062	16.65	2.8
10000K	484.187	160.94	3.0

It can be observed that the computation time of parallel k-Means is much less compared to the sequential version.

B. Performance of LV tracing

Fig. 9 shows the Bland–Altman plot to demonstrate the comparison between manual and automatic EF calculation 69 images of normal and abnormal patients manually extracted by experts with ECG gating.

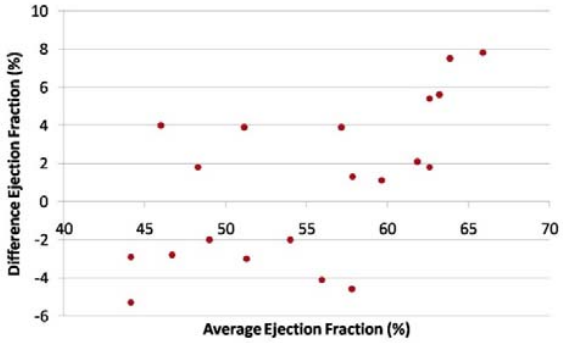


Fig. 9 Bland–Altman plot to show the comparison between manually computed EF with automatically computed EF
Normal value of EF: 60% ± 6.2%

It is observed that EF variation is from -6 to +7 which is acceptable due to large tolerance available for most of the 2D echo values.

The kurtosis feature shows wide difference between normal and abnormal images. The kurtosis data of normal and abnormal cases is as shown in Fig. 10. It is observed that kurtosis of normal patient data is positive (except for 3 cases) and abnormal cases are all negative (except for 2 cases).

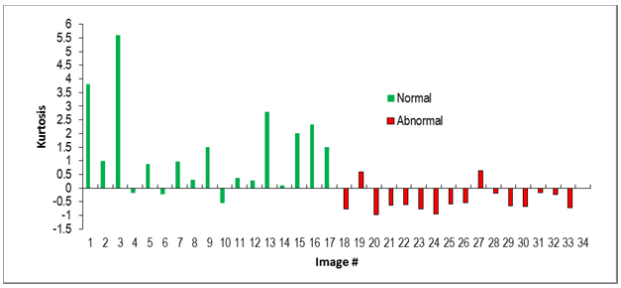


Fig. 10 Bar graph of Kurtosis feature. Thick bars – Normal subjects, Shaded bars – Abnormal subjects
The Correlation Coefficient = -0.43271

C. Performance of Bayesian Classification

Typically, the evaluation of the performance of a classification model is based on the counts of test tuples correctly and incorrectly predicted by the model. Our model is trained with 623 instances of which 200 are positive tuples ('N') and 423 are negative tuples ('A'). The result of the experiment is as shown in Table II.

TABLE II
CONFUSION MATRIX (ATTRIBUTES: CONTINUOUS VALUES)

Patient Echo Image Classification		Predicted Class C1 ('N')	Predicted Class C2 ('A')	Total
Actual Class	C1 ('N')	$tp = 38$	$fp = 0$	$p = 38$
	C2 ('A')	$fn = 12$	$tn = 24$	$n = 36$

Legend: 'N' – Normal and 'A' – Abnormal

For our dataset, the metrics are as follows: accuracy= 83.78%, error Rate= 16.21%, sensitivity= 100.00%, specificity= 66.66%. The results obtained through our framework shows an improved performance than Weka.

D. Performance of Image Mining (CDF Images)

The retrieval efficiency of CDF images, namely recall and precision, are calculated using the equations [12]:

$$precision = \frac{No.of\ relevant\ images\ retrieved}{Total\ No.of\ images\ retrieved}$$

$$recall = \frac{No.of\ relevant\ images\ retrieved}{Total\ No.of\ relevant\ images\ in\ the\ Database}$$

The Recall – Precision curve for the query image based on image ranking is computed and its plot is shown in Fig. 11. In this graph, the curve shown in dotted line is drawn by considering only color histogram feature and the curve shown in thick line is by considering color Doppler image features. The proposed color Doppler image features gives better results as it includes domain specific. Similarly, the retrieval efficiency of general images and synthetic datasets show improved performance over the results reported in the literature.

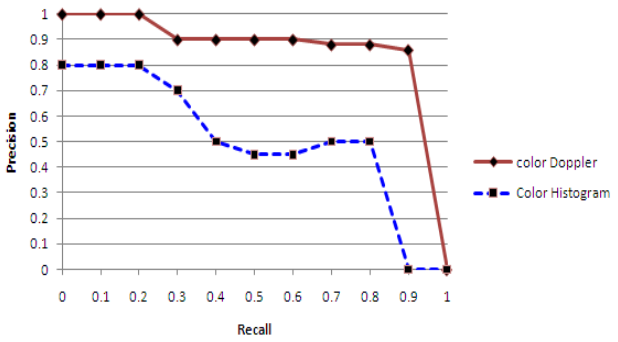


Fig. 11 Recall – Precision curve for the query image based on image ranking (top k images)

VIII. CONCLUSION

The framework is aimed to integrate heterogeneous datasets from where knowledge could be extracted. We have selected two major categories namely eco images and non-echo or general datasets. Novel methods and algorithms are used to preprocess, extract features, and compute metrics for 2DE and CDF, natural images.

Clustering, classification, image retrieval, etc. are some of the data mining tasks that are applied to discover useful knowledge, though no specific format knowledge representation is used. It is believed that this framework is useful for cardiac experts, researchers, clinical lab technicians, etc. Many new algorithms can be plugged-in to the knowledge discovery layer without affecting any other layer for effective and accurate learning and prediction.

REFERENCES

[1] Pedro Domingos and Geoff Hulten, "A general method for scaling up machine learning algorithms and its application to clustering" in ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning, San Francisco, CA, USA, 2001, pp. 106–113.
[2] Yu Zhang, et al., "A Fast Online Learning Algorithm for Distributed Mining of BigData", *ACM SIGMETRICS* Performance Evaluation Review 41 (4), 2014, pp. 90-93
[3] Domenico Talia, Paolo Trunfioy, Oreste Verta, Pedro Domingos Geo, "The Weka4WS framework for distributed data mining in service-oriented Grids", 2008.
[4] http://www. elki.dbs.ifi.lmu.de, "Data mining software framework", 2015.
[5] Piotr Kraj, Ashok Sharma, Nikhil Garge, Robert Podolsky: "ParaKMeans: Implementation of a parallelized k-Means algorithm suitable for general laboratory", BMC Bioinformatics 2008, doi:10.1186/1471-2105-9-200, pp. 1-13.
[6] Wooyoung Kim: Parallel Clustering Algorithms: Survey http://www.cs.gsu.edu/~wkim/index_files/SurveyParallelClustering.pdf, 2009.
[7] S. Nandagopalan, T.S.B Sudarshan, N. Deepak N. Pradeep, "Intelligent Echocardiographic Video Analyzer Using Parallel Algorithms", in Recent Advances in Information and Communication Technology, Advances in Intelligent Systems and Computing, vol. 265, 2014, Springer International Publishing, 2014, pp 157-166.
[8] S. Nandagopalan, B. S. Adiga, TSB Sudarshan, C. Dhanalakshmi, "Multifeature Based Retrieval of 2D and Color Doppler Echocardiographic Images for Clinical Decision Support", in *MySec2011*, Proc. of The Fifth Malaysian Software Engineering Conference – indexed by IEEE and SCOPUS, December 12-14, 2011 Johor Bahru, Malaysia, *IEEXPlore* , pp. 319-324.
[9] Reynaldo J. Gil-García1, José M. Badía-Contelles2 and Aurora Pons-Porrata1 "A General Framework for Agglomerative Hierarchical

Clustering Algorithms," in *IEEE*, The 18th International Conference on Pattern Recognition (ICPR'06), 2006, pp. 569 - 572.
[10] S. Nandagopalan, Dr. B. S. Adiga, N. Deepak, "A Universal Model for Content-Based Image Retrieval", International Journal of Computer Science, Vol. 4, No. 4, Dec 2009, pp. 242–245.
[11] S. Nandagopalan, Dr. B. S. Adiga, Dr. TSB Sudarshan, C. Dhanalakshmi, Dr. C. N. Manjunath, "A Naïve-Bayesian Methodology to Classify Echocardiographic Images through SQL", *Springer-Verlag*, LNCS/LNAI, Berlin Heidelberg, vol. 6746, 2011, pp. 155-165.
[12] Saso D'zeroski, "Towards a General Framework for Data Mining", in *KDID*, Springer-Verlag Berlin Heidelberg 2007, pp. 259–300.