Breast Cancer detection from Thermograms Using Feature Extraction and Machine Learning Techniques

Vartika Mishra
Department of Computer Science and Engineering
National Institute of Technology Rourkela, Odisha
Rourkela, 769008, India
vartikamishra151@gmail.com

Yamini Singh
Department of Computer Science and Engineering
National Institute of Technology Rourkela, Odisha
Rourkela, 769008, India
singhyamini312@gmail.com

Dr. Santanu Kumar Rath
Department of Computer Science and Engineering
National Institute of Technology Rourkela, Odisha
Rourkela, 769008, India
skrath@nitrkl.ac.in

***Abstract*—Presence of tumors in breasts have lead to possibility of occurrence of cancer in a global level. It's diagnosis is one of the challenging tasks. Researchers have come across a technique named thermography, which overcomes the drawbacks of a conventional technique i.e., mammography. In thermography, the early diagnosis of the breast cancer is carried out by implementing an analytical infrared thermal imaging techniques. This work focuses on different features-based machine learning techniques namely Support Vector Machine (SVM), k-Nearest Neighbour (KNN), Random Forest (RF) and Decision Tree (DT) to classify the images and detect possibility of cancerous image. This study also discusses about the SIFT and SURF features extraction techniques and a critical analysis of performance of various machine learning techniques have been presented.**

***Index Terms*—Breast Cancer, Thermography, Classification Techniques, Features Extraction**

I. Introduction

Our human body consists of the right count of cells of each type. Disturbance in this cell-organization leads to Cancer. The uprooting of cancer begins with the abrupt changes in these cell-organization. This is because, the signals that are generated from the cells decide the controlness and division of cells. Missing or faults in any of these signals lead the cells to grow and multiply too much resulting formation of a lump called tumor. This tumor turns out to be cause of cancer in a good number of cases.

In earlier days, X-ray mammography was the main imaging methodology for recognizing breast malignant growth of cancer cells [1]. But, this technique isn't a healthy one as it exposes the patients to ionizing radiations and also isn't a suitable technique for patients with dense breast. Breast cancer are more likely to occur in dense breast and the sensitivity of detection of breast cancer through mammography in these cases, can be as low as 30%-48%.

Any object above 0°C emits infrared radiations. These radiations are very well captured by thermographic camera [2]. Many researchers have come to the conclusion that the tumour is proportional to its changes in temperature. Infrared Thermography Screening helps in determining the cancer for an individual before occurrence of any side effects [3]. The Infrared Radiations do not require any type of internal contact while examining a patient. The relationship of a surface emitting radiations and the corresponding temperature is measured by the Steffans Boltzmann Law [4]. It is very well suited for the cases having abnormality. Thermography being non-invasive and non-ionizing, is very reliable and due to the emerging technology, thermography has found its place to be above other techniques by efficiently giving the positive results. It helps to discover malignant growth of cancer cells at early stages. If cancer can be detected at an early stage, then survival rate becomes high.

This study intends to detect the presence of cancer cells in breasts from thermographic images, based on various features extraction techniques and subsequently applying machine learning techniques. Data set used in this study is the pre-processed Breast thermograms available at Visual Labs DMR(Database Mastology Research). Figure 1 indicates the step followed in this study. Different features from images are extracted using Scale Invariant Feature Transform (SIFT) and Speeded Up Robust Features(SURF) techniques. These features are further reduced using Principal Component Analysis (PCA) for better interpretation of parameters. These selected features are fed to the different classifiers and different parameters are evaluated such as accuracy, specificity, sensitivity, precision and F-1 score to determine the efficiency of classifiers.

In this paper, section 2 describes different features extraction

techniques. Section 3 gives information on four popular ML techniques applied to a breast cancer data set that are analyzed and compared. Section 4 describes the simulation setup that has been used for carrying out the comparative study. Simulation results and interpretations are provided in section 5. Finally, conclusions and future work are provided in section 6.

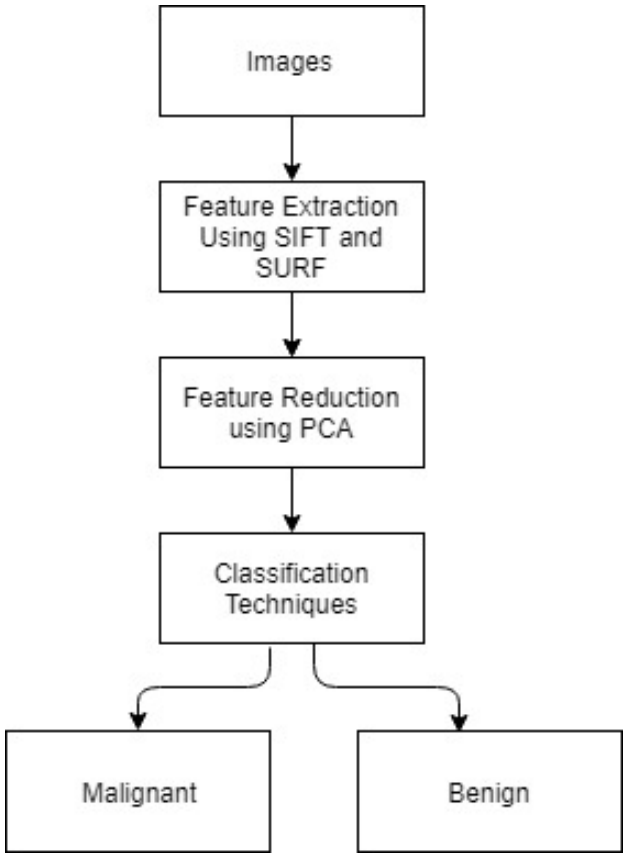

Fig. 1. Proposed block diagram of our work.

II. Feature Extraction

A. Application of Scale Invariant Feature Transform Technique

The Scale Invariant Feature Transform (SIFT) is a feature extraction technique that helps to find the similarity in between the images without affecting the scale(invariant) of the images with descriptors which extract the features [5]. The process of SIFT is divided into four major steps: scale-space extrema detection, keypoint localization, orientation assignment and keypoint descriptor. Scale-space extrema is used to identify location and scales of key points in the DoG (Difference-of-Gaussian) functions with different values of , the DoG function is separated by a constant factor k as in the following equation,

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \quad (1)$$

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \quad (2)$$

where G is the Gaussian function, I is the image and (x,y) are the coordinates of image. In the key point localization step, the low contrast points and the edge response are eliminated. This is obtained by using the Laplacian function.

The orientation assignment projects on appointing a steady introduction to the key points dependent on local image picture properties.

B. Application of Speeded Up Robust Features Technique

Application of Speeded Up Robust Features (SURF) algorithm which is used for the most vision undertakings and for the object detection purpose [5]. SURF falls in the classification of highlight descriptors by extricating keypoints from various areas of a given image and consequently is valuable in discovering likeness between images. The algorithm starts by discovering features / keypoints that are probably going to be found in various images of a similar object. Those features ought to be scaled and rotated invariantly if possible. SURF algorithm, is base on multi-scale space theory and the feature detector is based on Hessian matrix. Since Hessian matrix has good performance and accuracy. In image I, x = (x, y) is the given point, the Hessian matrix H(x,) in x at scale , it can be define as

$$H(x, \sigma) = \begin{bmatrix} Lxx(x, \sigma) & Lxy(x, \sigma) \\ Lyx(x, \sigma) & Lyy(x, \sigma) \end{bmatrix} \quad (3)$$

Where

$$L_(xx)(x, \sigma)$$

is the convolution result of the second order derivative of Gaussian filter

$$\frac{\partial^2 g(\sigma)}{\partial x^2}$$

with image I in point x.

C. Principal Component Analysis for Reducing Dimension

Principal Component Analysis (PCA) intends to find new dimension where the variation between the variables is maximum preserved. Principal components are the eigen vectors of the covariance matrix. The principal components are ordered in such a way that the variance decreases as we go down the list. The first principal component has more retention of the variation than the others. First the data set should be normalized then its covariance matrix is calculated from which eigen vectors can be computed. Then we take the eigen vectors (from the beginning as it has more information i.e., maximum value) and feature vector is generated which is used to get the dimensionally reduced dataset.

III. Classification Techniques for detecting tumors

In medical field, diagnosis has become a very difficult task, but due to algorithmic computations on machines with the increasing trend of technology [6], it has made the analysis very convenient. The main issue of this field is to analyze the images and classify them. So, in this work following techniques are being considered for the biclass classification

of the breast thermogram images to detect as to whether given subject has cancer or not.

A. Support Vector Machine

Support Vector Machine(SVM) is a supervised technique which can be used for both classification and regression problems but is popular in classification [7]. SVM creates a hyperplane which acts a boundary between the classes and helps us to correctly classify the points. SVM gives more accuracy than many algorithms and is best suited for problems with small datasets because training can become complex and time taking for larger datasets. SVM uses subset of training points which make classification more efficient. It cannot perform well if the data has noise.

B. Random Forest

Random forest(RF) is an ensemble of decision trees. Random Forest can give more stability than decision trees [8]. This proves that RF's are insensitive to noise and can also be used for datasets which are imbalanced. We often find that dataset for cancer often has imbalanced dataset and this classification algorithm can work well in those situations. The Random forest randomly selects features and builds decision trees and then it takes average of the result obtained from different decision trees and assigns the label to the data. The Random forest can be used for both classification and regression problems. The problem of overfitting is not much prominent in Random forests.

C. k-Nearest Neighbour

k-Nearest Neighbour(KNN) is a popular supervised classification algorithm which is used in pattern recognition, Intrusion detection and many other fields [9]. The algorithm calculates k nearest neighbors for each of the test data in the training data and the label is assigned to the class where majority of the k points resides. It is a simple algorithm which is easy to understand and the accuracy is also high but it is computationally expensive and the memory requirement is very high as we need to store both test and training data. KNN does not create any model and only uses similarities between data points to calculate distance.

D. Decision Tree

Decision Tree(DT) is a tree like model where decisions, possible consequences and their outcomes are taken into consideration [10]. Each internal node represents a question to be asked (e.g., person is male? or female?) and each branch represents an outcome and the leaf nodes are class labels. When we reach a leaf node the corresponding label is assigned to the sample. Decision Trees can be used for simple problems or when the data set is small. It is easy to understand and implement but one should look out for overfitting problem which is common in decision trees. Decision Trees can give biased outcomes if the data set is imbalanced.

IV. Simulation Setup

A. Data set

The data used in this study is the preprocessed Breast Thermograms available at Visual Labs DMR (Database Mastology Research) (http://visual.ic.uff.br/dmi). In this work we have used 20 images of 30 patients each. For our study purpose, we have used 70% data for training and 30% data for testing to calculate the accuracy.

B. Simulation Software

We have implemented four classification algorithms in Python 3.5 using Anaconda3 platform. The system configuration is of i7 processor with 3.4 GHz clock speed.

V. Results and Discussion

A. Evaluation Parameters

1) Confusion Matrix: A confusion Matrix is a representation technique for the execution of classification models. The confusion matrix demonstrates to us the quantity of accurately and inaccurately classified samples, contrasted with the real results (target value) in the test information. A confusion matrix framework of two class classification is a 2X2 table designed by adding the quantity of the four results of a binary classifier and TP, FP, TN and FN.

A binary classifier predicts a test dataset as either positive class or negative class for all data instances. This prediction (or classification) produces four results, for example, true positive (TP), false negative (TN), false positive (FP) and false negative (FN).

- True positive (TP): correct prediction of a positive class
- False positive (FP): incorrect prediction of a positive class
- True negative (TN): correct prediction of a negative class
- False negative (FN): incorrect prediction of a negative class

2) Accuracy: Accuracy is calculated by dividing the number of all correct predictions with the total number of the dataset. The best accuracy is assumed as 1.0, whereas 0.0 is the worst. We can also calculate it by using the formula as

$$1 - (error\ rate) \tag{4}$$

The value of accuracy is being obtained for all the four classifiers by using the confusing matrix.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{5}$$

3) Specificity: Value of Specificity is obtained by dividing the total number of correct negative predictions with the total number of negatives. Specificity ranges from maximum 1.0 to minimum 0.0.

$$Specificity = TN/(TN + FP) \tag{6}$$

4) Sensitivity: Value of Sensitivity is found by dividing the total number of correct positive predictions with the total number of positives. Sensitivity ranges from maximum 1.0 to minimum 0.0.

$$Sensitivity = TP/(TP + FN) \tag{7}$$

5) Precision: Precision is calculated by dividing the total number of correct positive predictions with the total number of positive predictions. Precision ranges from maximum 1.0 to minimum 0.0.

$$Precision = TP/(TP + FP) \tag{8}$$

6) F1-score: In two class classification, the testing accuracy is measured by F1 score (also F-measure). Recall and precision of the test are considered to calculate the F1-score.

$$F1score = 2TP/(2TP + FP + FN) \tag{9}$$

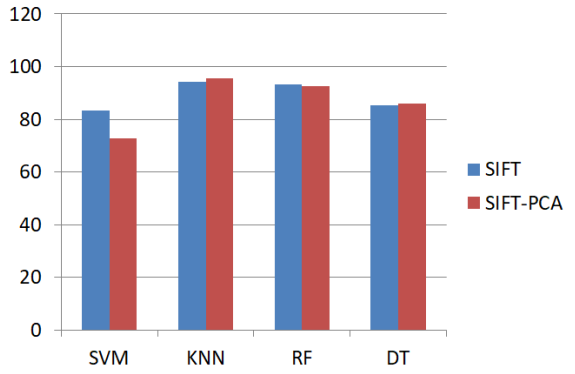

Fig. 2. Comparison between SIFT and SIFT-PCA.

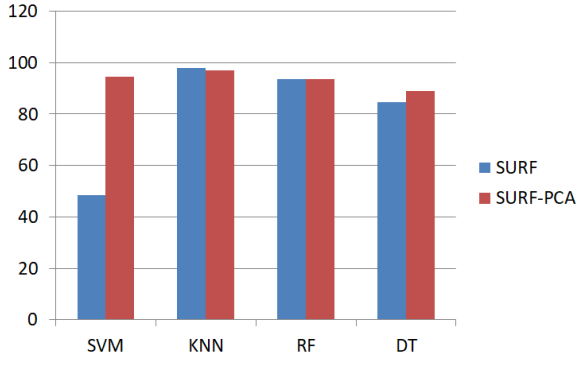

Fig. 3. Comparison between SURF and SURF-PCA technique

B. Results

In this work, features are extracted from breast thermograms using SIFT and SURF technique. SIFT extracted 128 features from the images and SURF 64 features from the images. Further these features are used for classification of images using SVM, KNN, RF & DT technique.

In Table I and Table II, the different parameters are calculated using four classifiers without applying PCA technique on features. This comparison concludes that KNN performs the best among other classifiers with SURF technique.

TABLE I
ACCURACY OF DIFFERENT CLASSIFIERS WITHOUT PCA

Comparison Table of SIFT				
	SVM	**KNN**	**RF**	**DT**
Accuracy	0.8341	0.9430	0.9326	0.8521
Sensitivity	0.6013	0.9608	0.9613	0.7474
Specificity	0.9871	0.9063	0.8740	0.9177
Precision	0.7903	0.9356	0.9396	0.8529
F1-Score	0.8778	0.9519	0.9503	0.8804

TABLE II
ACCURACY OF DIFFERENT CLASSIFIERS WITHOUT PCA

Comparison Table of SURF				
	SVM	**KNN**	**RF**	**DT**
Accuracy	0.4835	0.9802	0.9365	0.8459
Sensitivity	0.001 0.	0.9910	0.9523	0.8378
Specificity	0.00	0.9650	0.9230	0.8535
Precision	0.4835	0.9658	0.9243	0.8435
F1-Score	0.6501	0.9778	0.9508	0.8406

In Table III and Table IV, the features extracted from the SIFT(128 features) and SURF(64 features) techniques are futher reduced using Principal Component Analysis method(10 features) and then classification methods are applied. This result shows that KNN performs best among SIFT and SURF techniques with accuracy, specificity, sensitivity, precision and F1-score.

TABLE III
ACCURACY OF DIFFERENT CLASSIFIERS WITH PCA

Comparison Table of SIFT				
	SVM	**KNN**	**RF**	**DT**
Accuracy	0.7279	0.9559	0.9248	0.8601
Sensitivity	0.01	0.9742	0.9570	0.8558
Specificity	0.2905	0.9281	0.8758	0.8658
Precision	0.6938	0.9537	0.9214	0.8962
F1-Score	0.8192	0.9639	0.9389	0.8755

TABLE IV
ACCURACY OF DIFFERENT CLASSIFIERS WITH PCA

Comparison Table of SURF				
	SVM	**KNN**	**RF**	**DT**
Accuracy	0.9452	0.9609	0.9365	0.8808
Sensitivity	0.9038	0.9615	0.9532	0.9017
Specificity	0.9790	0.9603	0.9572	0.8944
Precision	0.9800	0.9782	0.9147	0.8828
F1-Score	0.9403	0.9698	0.9390	0.8921

Figs. 2 and 3 show the comparison of SIFT v/s SIFT-PCA and SURF v/s SURF-PCA. Here the accuracy of the different classifiers is plotted.

VI. Conclusion and Future Work

In this study, the four important classification models such as SVM, KNN, RF and DT in machine learning have been examined and the performance of the classifiers have been compared. In the four classification models, the results show that KNN gives the highest accuracy among SIFT, SIFT-PCA, SURF and SURF-PCA. It has been found that KNN classifier is found to have better accuracy as compared to others. This work also concludes that SURF is faster than SIFT because it deals with the better feature reduction and takes less time for computation as compared to SIFT. In future, we intend to apply other models from deep learning techniques in order to further improve the performance of classifiers.

References

[1] Islam, Md Shafiul, Naima Kaabouch, and Wen-Chen Hu. "A survey of medical imaging techniques used for breast cancer detection." In EIT, pp. 1-5. 2013.

[2] Usuki, H. "Relationship between thermographic observations of breast tumors and the DNA indices obtained by flow cytometry." Biomedical Thermology 10 (1990): 282-285.

[3] Arora, Nimmi, Diana Martins, Danielle Ruggerio, Eleni Tousimis, Alexander J. Swistel, Michael P. Osborne, and Rache M. Simmons. "Effectiveness of a noninvasive digital infrared thermal imaging system in the detection of breast cancer." The American Journal of Surgery 196, no. 4 (2008): 523-526..

[4] Yang, Wen-Jei, and Paul Yang. "Literature survey on biomedical applications of thermography." Bio-medical materials and engineering 2, no. 1 (1992): 7-18.

[5] Panchal, P. M., S. R. Panchal, and S. K. Shah. "A comparison of SIFT and SURF." International Journal of Innovative Research in Computer and Communication Engineering 1, no. 2 (2013): 323-327.

[6] Miranda, Eka, Mediana Aryuni, and E. Irwansyah. "A survey of medical image classification techniques." In Information Management and Technology (ICIMTech), International Conference on, pp. 56-61. IEEE, 2016.

[7] Acharya, U. Rajendra, Eddie Yin-Kwee Ng, Jen-Hong Tan, and S. Vinitha Sree. "Thermography based breast cancer detection using texture features and support vector machine." Journal of medical systems 36, no. 3 (2012): 1503-1510.

[8] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2, no. 3 (2002): 18-22.

[9] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." Emerging artificial intelligence applications in computer engineering 160 (2007): 3-24.

[10] Kohavi, Ronny, and J. Ross Quinlan. "Data mining tasks and methods: Classification: decision-tree discovery." In Handbook of data mining and knowledge discovery, pp. 267-276. Oxford University Press, Inc., 2002.