

Feature Analysis for Detection of Breast Cancer Thermograms Using Dimensionality Reduction Techniques



Vartika Mishra, Subhendu Rath, and Santanu Kumar Rath

Abstract Breast cancer is observed as the most dominant disease availed in women across the globe. It has been studied that if diagnosed early, the mortality rate can be reduced. Among different modalities such as mammography and ultrasound, thermography is proving to be a successful one which helps to detect the tumor at an early rate. It functions by recording the surface temperature of the required area, and with the help of images consisting of different colors, it helps the doctors to interpret the tumor region. Feature reduction process helps in removing the multicollinearity between the different features thus resulting in less computation time to process. In this study, unsupervised-based nonlinear feature analysis is performed for which different methods are applied based on feature reduction, viz., t-SNE, LLE and KPCA. Different classification models such as random forest and decision tree are applied for classifying between healthy and unhealthy breast. Thus, features obtained by kernel principal component analysis (KPCA) method and classified with random forest give a good accuracy of 89.49%.

Keywords Breast cancer · Classification · Feature reduction · Thermography

1 Introduction

Among many diseases, cancer is observed to be one such type where the cells in the human body grow abruptly [1]. It can occur in any part of the human body. Among many cancers, one of the most occurring is breast cancer. It is analyzed that it occurs more often in women as compared to men. The cells form a tumor which invades

V. Mishra (✉) · S. K. Rath
NIT Rourkela, Rourkela, Odisha 769008, India
e-mail: vartikamishra151@gmail.com

S. K. Rath
e-mail: skrath@nitrl.ac.in

S. Rath
VCU, Richmond, VA 23219, USA
e-mail: subhendu.rath@vcuhealth.org

its surrounding tissues [2–4]. Mortality rate can be reduced for breast cancer, if the cancer is diagnosed early. However, many different modalities have been taken into consideration for the detection of breast cancer such as ultrasound, mammography, magnetic resonance imaging (MRI,) X-ray and thermography. These modalities have helped in reducing the mortality rate from 30 to 70% by assisting the radiologists and physicians in diagnosing the abnormalities [5].

Mammography is among one of the modalities which helps in detection of breast tumor. It may help in detecting the cancer for an interval between one and half years to four years [6]. It may increase the future chance of growth in cancer. It does not work well for the dense tissues [7, 8]. Ultrasound is considered as an adjunct to the mammography which is done for women having dense breast tissue. Several research has claimed that ultrasound breast imaging in women with dense breasts and negative mammograms has elated cancer detection rate [9]. Studies have observed that there is lower sensitivity rate for women with dense breast by using mammography. It has been observed that the application of ultrasound screening results in detection of otherwise occult cancer, irrespective of the women having high risk with dense breast or not [9].

Among all the above mentioned modalities, with a decade research, it has been reported that thermography has proved to be one of the early diagnostic modalities. It helps in early detection by 10 years which leads to an immense increment in the survival rate [10]. With the thermal camera, breast thermography detects the tumor by recording the temperature of the surface. By placing the thermal camera in front of the patient, the images are captured appropriately as shown in Fig. 1. This modality is contact-free, non-ionizing, non-invasive and radiation-free in nature [11]. This is suitable for women breast implants, with every age, all breast densities and after surgery monitoring. The vascular heat radiated by that breast surface is measured by the infrared radiation. The increase in local vascularization, the tumor is indicative of high metabolic rate. The normal and abnormal thermogram can be distinguished based on the asymmetric heat pattern between the left and the right breast [12–14]. Physicians analyze the subjects for different abnormalities, but sometimes it becomes difficult to detect the same through naked eyes. So, different computer-aided detection (CAD) systems provide additional assistance in finding the information from the thermograms [15, 16].

Thus, in this study, the work has been carried out on the nonlinear feature reduction techniques for detection of breast cancer between healthy and unhealthy breast which reduces the computational cost as compared to the linear reduction techniques efficiently and hence improving the predictions. Here, while preserving the original distances between the data points, the lower dimensional representation is attained. The transformed images obtained are segmented into left and right breast manually. Further, from these segmented images, gray-level run length matrix (GLRLM) and gray-level co-occurrence matrix (GLCM)-based features are extracted. From the matrices, a total of twenty-seven features are extracted. Further, feature reduction techniques are applied, viz., t-SNE, LLE and KPCA for the features extracted. The different features obtained by applying the reduction techniques are taken for the classification by applying two classifiers, viz., random forest and decision tree.



Fig. 1 A thermal camera capturing image of the patient and stored as a thermogram

Further, the work is organized as: related work described in Sect. 2. Section 3 describing breast cancer methodology and Sect. 4 discusses experimental result analysis. Section 5 thus concludes the proposed work.

2 Literature Survey

For applications of data analytics, large amount of day results in worsening the results; hence, redundant data must be minimized [17, 18]. So, this section emphasizes on the state of the art by different researchers for the detection of breast cancer using thermograms.

Borchatt et al. have detected presence of cancer by using the temperature range on region of interest (ROI), eight-level posterization by quantizing the higher tone, standard deviation and the mean temperature. Further, feature extraction method was applied from the breast quadrants and entire image as well by extracting the aforementioned features. Further, support vector machine (SVM) was applied which obtained 86% accuracy [19]. Milosevic et al. had computed twenty GLCM-based texture features. Further, Naïve Bayes, k -nearest neighbor (KNN) and support vector machine (SVM) classifier were applied. The classification performance for breast cancer detection was computed by fivefold cross-validation. Among different classifiers, KNN classifier gave better result as compared to the other two classifiers, viz., SVM and Naïve Bayes [20].

Ali et al. have extracted texture features based on the statistical features first-order and second-order (GLCM) features. They performed and analyzed the accuracy of their model by applying the SVM classifier for normal and abnormal breast. They applied four different scenarios and compared their results. Every scenario consists of different proportions of training and testing images. It was observed that the first

scenario performed better than other scenarios. They applied different kernel functions for the different scenarios and obtained a highest accuracy of 85% for statistical features for linear and quadratic kernels. By applying quadratic and polynomial kernel functions for GLCM-based features, they obtained a highest accuracy of 80% [21]. Jakubowska et al. have applied wavelet transform on the breast images. Further, they have applied PCA and LDA for dimensionality reduction methods for reducing the number of features. These transformed images with dimensionality reduction methods are further classified with the multi-layer perceptron classifier [22].

Resmini et al. applied dimensionality reduction technique, viz., principal component analysis (PCA). A total of sixteen features were chosen and further classified with the SVM classifier implementing on WEKA tool. Their methodology attained an accuracy of 82.14% with specificity and sensitivity of 25 and 91.7% [23]. Gogoi et al., applied Mann Whitney Wilcoxon test and further obtained 84.75% of the highest accuracy by applying support vector machine (SVM) [24].

In this work, significant features are extracted from the three feature reduction techniques. The t-SNE technique gives three most important features; LLE gives ten features, and KPCA gives 7 most important features. These obtained set of features are further classified with random forest and decision tree classifiers.

3 Breast Cancer Methodology

3.1 Dataset

In this study, the dataset used is from an online source available at Database Mastology Research (DMR), Brazil [25]. The dataset consists of 56 subjects each with 20 different positional temperature matrix among which 19 are healthy and 37 are unhealthy. The Thermal camera FLIR SC-620 is used for capturing the thermal image of the subject.

3.2 Proposed Approach

The breast cancer detection between healthy and unhealthy breast thermograms is carried out as shown in Fig. 2. Primarily, the breast images are obtained from the dataset available in the form of temperature matrix as shown in Fig. 3 in preprocessing. The images are further segmented between left and right breast using the manual segmentation method. Further, by applying GLCM and GLRLM matrix, the features are extracted from the images [26].

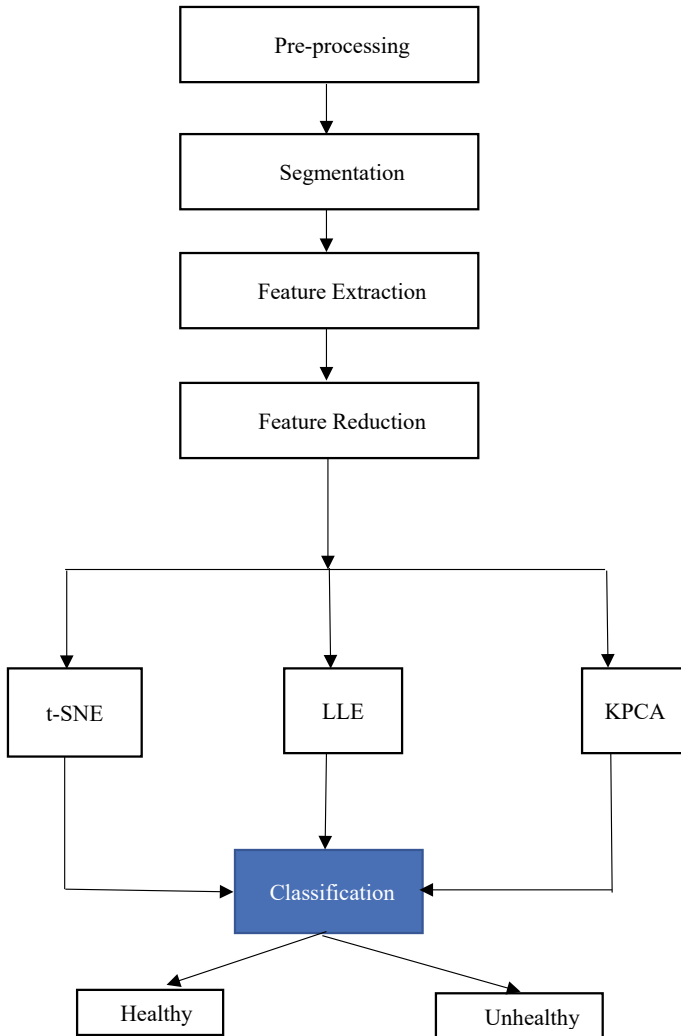


Fig. 2 Steps of the proposed work for detecting healthy and unhealthy breast thermograms

3.2.1 Feature Reduction

Feature extraction often yields the redundant data resulting in worse performance of the models leading to curse of dimensionality. So, this calls for the removal of the features that are less important or irrelevant. This is done by feature reduction techniques.

The resultant of applying dimensionality reduction techniques is a meaningful representation of the reduced data which is obtained by transformation of the high-dimensional data into relevant dimensionality. This reduced dimensionality accounts

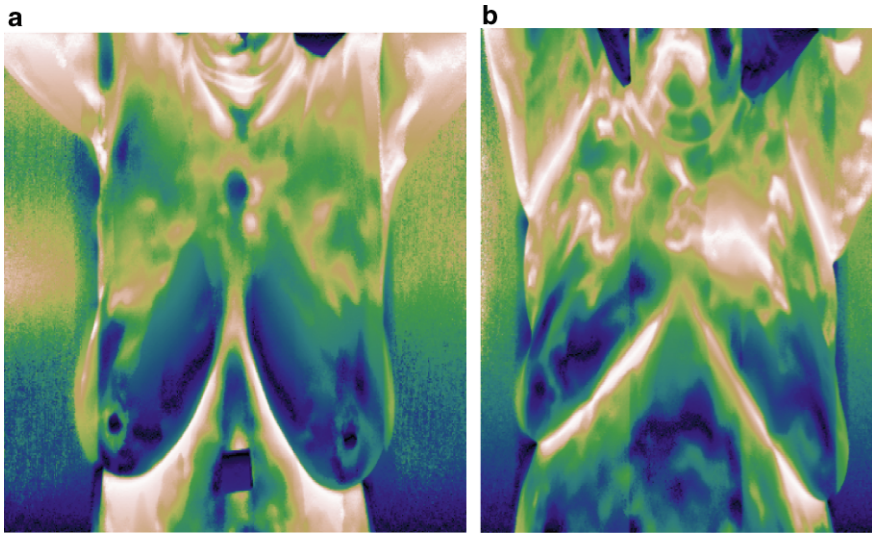


Fig. 3 **a** A subject with healthy breast thermogram and **b** a subject with an unhealthy breast thermogram

for minimum parameters which are needed for observing the properties of the required data [27]. It facilitates among different fields, viz., visualization, compression of high-dimensional data and classification. This work concentrates on three-dimensionality reduction techniques, viz., t-SNE, LLE and KPCA. Each of them is described as follows:

a. t Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE technique simultaneously retains local and global structure of the data. For both, high-dimensional and low-dimensional space, probability similarity points are computed. It is observed that, results obtained from applying t-SNE method are good for linear and nonlinear data [28]. It makes the visualization of the data better and clear. In this work, n components with 3 are taken into consideration for feature reduction.

b. Locally Linear Embedding (LLE)

LLE starts with finding the nearest neighbors for all sample and further computes set of weights. It is an unsupervised dimensionality reduction method, which preserves the originality of the local features. Further, for finding the lower dimensional embedding points, it applies the eigenvector-based technique [29].

c. Kernel Principal Component Analysis (KPCA)

It is a nonlinear form of PCA, which helps to solve the more complicated correlation among the different features. The decision boundaries are described by the nonlinear functions. It helps to project the high-dimensional data from the original data with less computational time. This is also known as nonlinear feature mapping [30].

3.3 Classification and Performance Parameters

The classification of breast thermograms is analyzed by applying the tree-based learning algorithms. It is observed that tree-based learning algorithms enhance the models with easy interpretation, stability and high rate of accuracy. It works well for nonlinear relationships and is adaptable for solving problem efficiently.

3.3.1 Random Forest

It is a supervised learning algorithm and consists of different trees which leads to ensemble learning method. It measures the importance of feature relatively while making the predictions. The various input variables are recorded enabling the higher number of observations for making predictions [31].

3.3.2 Decision Tree

It is a supervised learning method based on conditions for every feature-based split present in the dataset. The decisions are resulted from the tree-like structure. The process starts from the root node and ends with the leaves giving the predictions after computing all the features [31].

3.3.3 Performance Parameters

The different parameters of a classification algorithm are visualized by a table known as confusion matrix. It describes the relationship for a set of test data given the true values. The values are determined with the help of four data elements, viz., as described below (Fig. 4).

True Positive (TP): It predicts the true value for the actual true value.

Fig. 4 A confusion matrix to evaluate the performance metrics

		Predict	
		Yes	No
Actual	Yes	True Positive	False Negative
	No	False Positive	True Negative

Table 1 Different parameters for observing the performance of the random forest classifier

Classifiers	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
t-SNE	86.91	62.50	81.96	84.36
KPCA	89.58	89.32	94.17	91.82
LLE	91.42	61.72	84.19	87.65

Table 2 Different parameters for observing the performance of the decision tree classifier

Classifiers	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
t-SNE	63.83	79.77	62.76	63.29
KPCA	89.10	82.28	89.51	89.30
LLE	89.07	59.11	74.77	76.38

Table 3 Accuracy parameter for different set of features obtained from two different classifiers

Classifiers	Decision Tree (%)	Random Forest (%)
t-SNE	74.22	78.66
KPCA	86.57	89.49
LLE	76.38	82.22

True Negative (TN): It predicts a false value for an actual true value.

False Positive (FP): It predicts a true value for an actual false value.

False Negative (FN): It predicts a false value for an actual false value.

The values of performance parameters, i.e., accuracy, specificity, *F1*-score, sensitivity, specificity and precision are computed based on two different classifier models, i.e., random forest and decision tree as shown in Tables 1, 2 and 3.

4 Experimental Result Discussion

In this proposed approach, feature reduction-based work is carried out between three techniques mainly: t-SNE, LLE and KPCA. Each technique has reduced the features to different dimensions: three features from t-SNE, seven features from KPCA and ten features from LLE. These features are further applied individually for classifying among the healthy and unhealthy breast. Decision tree and random forest classification methods are applied, for each set of features obtained from the abovementioned methods for classification.

In Table 1, the accuracy of the two different classifiers is shown which gives an observation that KPCA has given better set of features when classified with random forest classifier with an accuracy of 89.49% as compared to other two dimensionality reduction techniques.

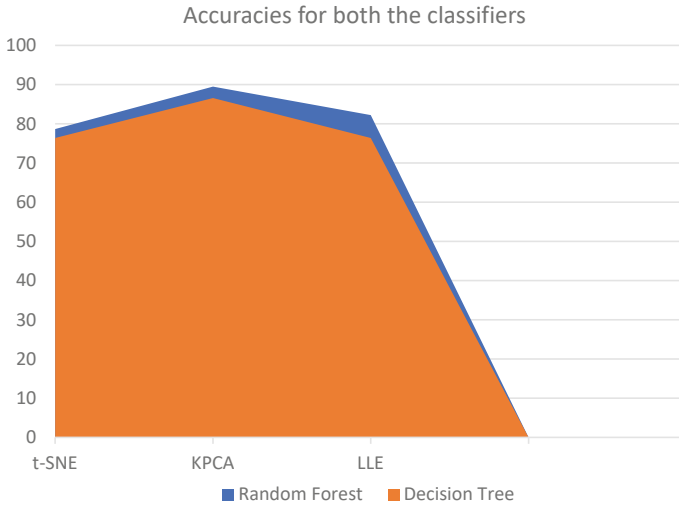


Fig. 5 A chart representing the accuracies obtained for both the classifiers

Among all the four parameters calculated for the random forest classifier, it is observed that the precision value of the set of features obtained from KPCA reduction technique gives a highest value of 94.17% as mentioned in Table 1. This gives an analysis that the true positive rate calculated from KPCA set of features is good as compared to the other two reduction method. It selects the most significant eigenvectors and eigenvalues giving the low-dimensional representation of the data objects. The sensitivity rate for t-SNE and LLE method is high which determines that the false-positive rate is more. It also accounts for catching good number of actual cases of the disease.

The precision rate for decision tree is high, i.e., 89.51% as compared to other performance parameters for the KPCA set of features obtained. The sensitivity rate for both the other techniques is higher viz. t-SNE and LLE as compared to KPCA set of features as shown in Table 3 (Fig. 5).

5 Conclusion

From the above proposed work, it is concluded that kernel principal component analysis feature reduction technique gives better set of features when compared with other two techniques, i.e., t-SNE and LLE with the highest precision rate of 94.17% for decision tree classifier and the highest accuracy of 89.49% for random forest classifier for classifying between healthy and unhealthy breast. In future work, deep learning methods will be applied to obtain better prediction for detection of breast cancer.

References

1. Cancer statistics, <https://www.cancer.gov/about-cancer/understanding/what-is-cancer>
2. Breast Cancer, <https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html>
3. Siegel RL, Miller KD, Jemal A (2017) Cancer statistics, CA. *Cancer J Clin* 67(1):7–30
4. Sree SV, Ng E-K, Acharya RU, Faust O (2011) Breast imaging: a survey. *World J Clin Oncol* 2(4):171–178
5. Schneider M, Yaffe M (2000) Better detection: improving our chances, digital mammography. In: 5th International workshop on digital mammography
6. Tabár L, Vitak B, Chen H-HT, Yen M-F, Duffy SW, Smith RA (2001) Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. *Cancer: Interdisc Int J Am Cancer Soc* 91(9):1724–1731
7. EtehadTavakol M, Vinod Chandran EYK, Ng, and Raheleh Kafieh, (2013) Breast cancer detection from thermal images using bispectral invariant features. *Int J Therm Sci* 69:21–36
8. Brown J, Bryan S, Warren R (1996) Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. *BMJ* 312(7034):809–812
9. Lee CH, David Dershaw D, Kopans D, Evans P, Monsees B, Debra Monticciolo R, Brenner J et al (2010) Breast cancer screening with imaging: recommendations from the Society of Breast Imaging and the ACR on the use of mammography, breast MRI, breast ultrasound, and other technologies for the detection of clinically occult breast cancer. *J Am Coll Radiol* 7(1):18–27
10. Etehadtavakol M, Ng EYK (2013) Breast thermography as a potential non-contact method in the early detection of cancer: a review. *J Mech Med Biol* 13(2):1330001
11. Tan TZ, Quek C, Ng GS, Ng E-K (2007) A novel cognitive interpretation of breast cancer thermography with complementary learning fuzzy neural memory structure. *Expert Syst Appl* 33(3):652–666
12. Sathish D, Kamath S, Rajagopal KV, Prasad K (2016) Medical imaging techniques and computer aided diagnostic approaches for the detection of breast cancer with an emphasis on thermography—a review. *Int J Med Eng Inf* 8(3):275–299
13. Lee H, Chen Y-P (2015) Image based computer aided diagnosis system for cancer detection. *Expert Syst Appl* 42(12):5356–5365
14. Lipari CA, Head JF (1997) Advanced infrared image processing for breast cancer risk assessment. In: Proceedings of the 19th annual international conference of the IEEE engineering in medicine and biology society. Magnificent milestones and emerging opportunities in medical engineering' (Cat. No. 97CH36136), vol 2. IEEE, pp 673–676
15. Arakeri M, Reddy G (2015) Computer-aided diagnosis system for tissue characterization of brain tumor on magnetic resonance images. *SIViP* 9(2):409–425
16. Moghbel M, Mashohor S (2013) A review of computer assisted detection/diagnosis (CAD) in breast thermography for breast cancer detection. *Artif Intell Rev* 39(4):305–313
17. Rao A, Noushath S (2010) Subspace methods for face recognition. *Comput Sci Rev* 4(1):1–17
18. Tomakova R, Komkov V, Emelianov E, Tomakov M (2017) The use of Fourier descriptors for the classification and analysis of peripheral blood smears image. *Appl Math* 8(11):1563
19. Borchardt TB, Resmini R, Conci A, Martins A, Silva AC, Diniz EM, Paiva A, Lima RCF (2011) Thermal feature analysis to aid on breast disease diagnosis. In: Proceedings of 21st Brazilian Congress of Mechanical Engineering—COBEM, pp 24–28
20. Milosevic M, Jankovic D, Peulic A (2014) Thermography based breast cancer detection using texture features and minimum variance quantization. *EXCLI J* 13:1204
21. Ali MAS, Sayed GI, Gaber T, Hassanien AE, Snaes V, Silva LF (2015) Detection of breast abnormalities of thermograms based on a new segmentation method. In: 2015 Federated conference on computer science and information systems (FedCSIS). IEEE, pp 255–261
22. Jakubowska T, Wiecek B, Wysocki M, Drews-Peszynski C, Strzelecki M (2004) Classification of breast thermal images using artificial neural networks. In: The 26th annual international conference of the IEEE engineering in medicine and biology society, vol 1. IEEE, pp 1155–1158
23. Resmini R (2011) Análise de imagens térmicas da mama usando descritores de textura. Computer Institute, Master's Thesis, Federal Fluminense University

24. Gogoi UR, Bhowmik MK, Ghosh AK, Bhattacharjee D, Majumdar G (2017) Discriminative feature selection for breast abnormality detection and accurate classification of thermograms. In: 2017 International conference on innovations in electronics, signal processing and communication (IESC). IEEE, pp 39–44
25. Dataset Retrieved from <http://visual.ic.uff.br/en/proeng/thiagoelias/>
26. Mishra V, Rath SK (2021) Detection of breast cancer tumours based on feature reduction and classification of thermograms. *Quant InfraRed Thermography J* 18(5):300–313
27. Fukunaga K (1990) Introduction to statistical pattern recognition. Academic Press Professional Inc., San Diego
28. Devassy BM, George S (2020) Dimensionality reduction and visualisation of hyperspectral ink data using t-SNE. *Forensic Sci Int* 311:110194
29. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326
30. Schölkopf B, Smola A, Müller K-R (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput* 10(5):1299–1319
31. Nguyen C, Wang Y, Nguyen HN (2013) Random Forest classifier combined with feature selection for breast cancer diagnosis and prognostic