Classification of Breast Masses using Tactile Imaging System and Machine Learning Algorithms

Vira Oleksyuk[1], Firdous Saleheen[1], Dina F. Caroline[2], Suzanne A. Pascarella[2], and Chang-Hee Won[1]
Email: {vira.oleksyuk, f.saleheen, cwon}@temple.edu, {dina.caroline, suzanne.pascarella}@tuhs.temple.edu
[1]Department of Electrical and Computer Engineering, Temple University, Philadelphia, PA 19122, USA
[2]Department of Radiology, School of Medicine, Temple University Hospital, Philadelphia, PA 19140, USA

***Abstract*—In this study, we used Tactile Imaging System (TIS) and machine learning algorithms to classify breast masses *in vivo* as malignant or benign. When the silicone probe at the front end of TIS is compressed against the breast mass, the indentation profile of this waveguide is captured by a CCD camera. Then TIS algorithm determines the size and stiffness of inclusions based on the acquired tactile images. The size and stiffness results are then used as the input features for breast tumor classification algorithms. We compared three tumor classification algorithms: k-nearest neighbor, support vector machine, and Naïve Bayes, which are known to work well for limited data set. We tested these algorithms on twelve human breast tumors. The results were evaluated using the leave-one-out cross validation technique. Among the three algorithms, k-nearest neighbor classifier performed the best with sensitivity of 86% and specificity of 100%.**

***Index Terms*— breast cancer screening, breast tissue stiffness, breast tumor classification, tactile breast imaging.**

I. Introduction

Some breast cancer screening techniques revolve around the fact that cancerous masses tend to be stiffer than benign masses [1], [2]. One such method is tactile imaging, which is capable of characterizing breast masses *in vivo*. Tactile Imaging System (TIS) is based on optical sensor technology [3]. Therefore it is noninvasive and harmless. It utilizes a thin, flexible, and transparent silicone waveguide and a charge coupled device (CCD) to capture tactile images (Fig. 1). The TIS algorithm estimates mechanical properties, such as size and stiffness. These properties are used to differentiate between malignant and benign breast masses [2], [4].

Recently, it is becoming increasingly popular to use machine learning algorithms for computer-aided clinical diagnosis. It helps to analyze medical data more efficiently and with fewer human-related errors [5], [6]. K-nearest neighbor (KNN) is a non-parametric classification technique and it is used for breast cancer application in [7]. Samir, Al-Absi and Kassoul classified images from the Mammographic Image Analysis Society (MIAS) dataset. The authors combined KNN method with K-means method in their classification. They achieved 98.2% classification accuracy for normal/abnormal region and 100% classification accuracy in malignant/benign classification. Support vector machine (SVM) method is a two-class classifier, which found uses in multiple medical applications. SVM was used for breast cancer classification by [5], [8]. Both groups of researchers used the Wisconsin breast cancer dataset (WBCD) and achieved 98.5% and 99.5% accuracy, respectively. Naïve Bayes (NB) is a parametric technique, which is used for malignant/benign of breast masses classification in [9]. They achieved classification accuracy of 93% [9].

The objective of this work is to employ TIS and the machine learning algorithms to classify benign/malignant breast masses *in vivo*.

In our previous work [10], it was shown that TIS estimate size and stiffness of inclusions. In this paper, we analyze the pilot dataset of 12 human breast tumors. We use the machine learning classification algorithms to discern malignant and benign tumors.

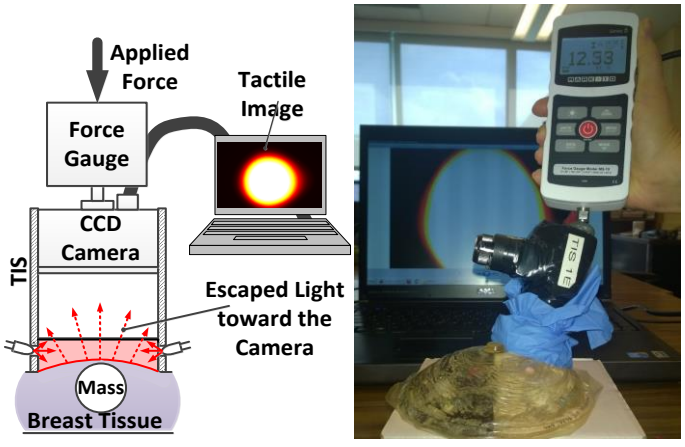

Fig. 1. TIS in breast cancer application.
TIS is applied from the top of a breast mass. Tactile images are captured by the CCD camera. They are saved on the computer with corresponding force information for the further analysis.

II. Tactile Imaging System

This section describes the TIS design and algorithm for size and stiffness estimation.

A. TIS Design and Principle

TIS has two main components: a thin silicone waveguide as the tactile sensor, and a charge coupled device (CCD) camera as the detector. Fig. 1 shows the design and operating principle of TIS.

Our system utilizes the total internal reflection principle of light inside the flexible and transparent silicone waveguide, which is a layer of polydimethyl siloxane (PDMS). The four white LEDs homogeneously illuminate the waveguide. The external force gauge (Mark-10, Long Island, NY) attached on top of the CCD camera.

When TIS is compressed against a breast mass, the waveguide deforms. Due to the deformation, the internally scattered light within the transparent silicone probe escapes toward the camera. The captured light by the camera forms the tactile image. These images along with the corresponding applied forces are the inputs for TIS algorithm.

B. TIS Algorithms for Size and Stiffness Estimation

In the previous work [10], we described the methods to calculate size and stiffness of tissue inclusions. The 3D interpolation method estimates the size of masses. To estimate stiffness of a mass or tumor, TIS uses the conventional tensile tests [11]. TIS estimates Young's modulus by capturing the indentation of the soft silicone probe during compression. Young's modulus describes the stiffness of the tested material in its elastic region. It is calculated as the stress applied to the region divided by the strain in the test region [10].

Breast tissue phantoms are developed to test TIS performance [10]. Polyvinyl chloride (PVC) is used to create the phantom, which mimicked breast tissue mechanical properties. The phantoms of breast masses were either polyacrylo nitrile (acrylic) spheres, or made from PDMS. We varied the stiffness of the phantom breast masses by changing the ratio of the PDMS mix (the two-component silicone material) during fabrication. Young's moduli of the fabricated breast tissue and inclusion phantoms were measured using Instron 4442 (MA, USA).

III. Computer-Aided Classification

To classify breast masses using the TIS output (size and stiffness), we used three classifiers: KNN, SVM, and NB. These methods are known to work well with small datasets. In addition, we assigned prior probabilities for the classification based on the knowledge that malignant tumors compose approximately 20% to 30% of all biopsies [12]. The priors were 0.30 for malignant and 0.70 for benign cases. Figure 2 presents the general procedure of the method. Next, we describe KNN, SVM, and Naïve Bayes classification algorithms.

A. KNN Classification

KNN is a non-parametric classifier that can work with arbitrary distributions [13]. It uses available data to classify new data points based on a similarity measure. KNN uses a distance function, as the similarity measure, to define *k* nearest neighbors among the data points. The class of majority neighboring points will define the class assignment of new data point.

We used *k* number of neighbors in the range from 1 to 9, and applied Euclidian, Mahalanobis, and Chebyshev distances as the metric for the algorithm. Euclidian, Mahalanobis, and Chebyshev distances were calculated using the following,

$$d_{xy\_E} = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}\,, \tag{1}$$

$$d_{xy\_M} = \sqrt{(x_i - y_i)' C^{-1} (x_i - y_i)}\,, \tag{2}$$

$$d_{xy\_C} = \max_k |x_i - y_i|\,, \tag{3}$$

where x_i and y_i were two groups of points, and C is the covariance matrix [14].

B. SVM Classification

SVM classifier finds optimal decision surface for linearly separable data. The goal of SVM training is to search for the decision surface with the largest margin. The classifier will have better generalization of the dataset if its margins are greater [14]. If the data is not linearly separable, it is projected to the high dimensional feature space using kernel functions, where this data becomes linearly separable. For SVM, we used polynomial and Gaussian kernel options following [15]. The polynomial kernel is the following,

$$K(x_1, x_2)_{Poly} = (1 + x_1' x_2)^d\,, \tag{4}$$

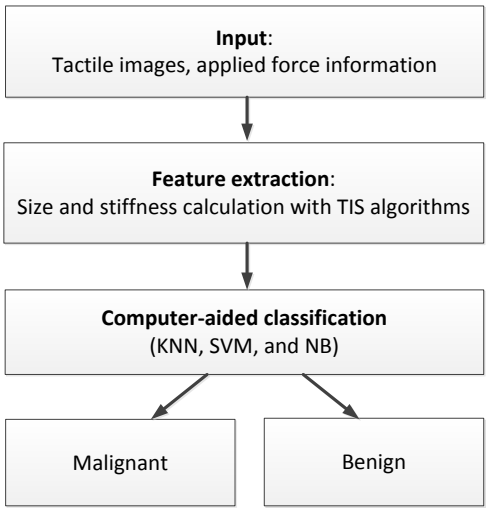

Fig. 2. Computer-aided classification of breast masses with TIS.

where *d* is a degree of the polynomial. The polynomial order was in the range from 1 to 9. The Gaussian option is the following,

$$K(x_1, x_2)_{Gauss} = \exp(-(x_1 - x_2)'(x_1 - x_2)/(2\sigma^2)), \quad (5)$$

for some positive number σ, as the standard deviation.

C. Naïve Bayes Classification

NB is the parametric technique. It assumes independence of the dataset features [14] and searches for the most likely class in the given dataset. Using NB method, we applied several distributions for smoothing the data: normal and kernel. In the case of the normal distribution option, the classifier finds mean and standard deviation of the training set to estimate Gaussian probability distribution for each class in the training set. When the kernel option is selected, the kernel density will be calculated for each class in the training dataset. We used box, triangle, Gaussian and Epanechnikov kernels. The following equations present one dimensional kernel functions. For the bivariate data, the same smoothing kernel function applied to each variable.

$$K(u)_{Box} = \frac{1}{2}, if \ \ |u| < 1, \quad (6)$$

$$K(u)_{Tri} = 1 - |u|, if \ \ |u| < 1, \quad (7)$$

$$K(u)_{Gauss} = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2), \quad (8)$$

$$K(u)_{Epa} = \frac{3}{4}(1 - \frac{1}{5}u^2)/\sqrt{5}, \ if \ \ |u| < \sqrt{5}, \quad (9)$$

were *u* is the difference between the estimated density function at *x* and the observations at x_i.

IV. Validation of the Method

We calculated sensitivity and specificity of the tumor classification method to evaluate its performance. The leave-one-out cross validation (LOOCV) was used in this work [14]. The idea behind LOOCV is to use all available data to validate the classification method. We take *N-1* data points from the N point dataset at a time to train the classifier. One point, which was not included in the training process, is used to test the classification result. All points are tried as a test point exactly once. Then the average error for all trials is computed.

The sensitivity and specificity were calculated as follows,

$$Sensitivity = \frac{TP}{TP+FN}(\%), \quad (10)$$

$$Specificity = \frac{TN}{FP+TN}(\%). \quad (11)$$

TP denotes true positive result of the classification, where the malignant mass was classified as malignant. TN is the true negative result of the classification, where the benign mass was classified as benign. FP denotes false positive result, where the benign mass was classified as malignant. FN is the false negative result, where the malignant mass was classified as benign.

V. Results

This section presents the performance of the classifiers on TIS dataset obtained from human patients.

A. Human Dataset

The pilot TIS dataset consisted of 12 human patients (IRB# 13661 Temple University). All 12 females were scheduled for biopsy. The patients for our experiments were selected by the radiologists. The patients' age ranged from 24 to 84 years. We received the clinical pathology results for each patient, which specified 7 malignant and 5 benign cases.

B. Classification Results

All algorithms were implemented in MATLAB R2014b (The MathWorks Inc., USA). We applied the cross-validation method with leave-one-out technique to estimate the sensitivity and specificity. Table I shows the sensitivity and specificity results for the classification of tumors for the pilot dataset of 12 human patients. The best performance of each classifier is highlighted in each column of the table. For the KNN classifiers, the value of *k* was varied from 1 to 9 with Euclidian, Mahalanobis, and Chebyshev metrics. Among the Euclidian and Mahalanobis metrics, KNN with *k=1* produced the highest sensitivity of 86% with a specificity value of 80%

TABLE I
MALIGNANT/BENIGN CLASSIFICATION RESULTS FOR HUMAN DATA (SENSITIVITY/SPECIFICITY)

KNN				SVM			NB		
k	**Euclidian**	**Mahalanobis**	**Chebyshev**	**Order/Shape**	**Polynomial**	**Gaussian**	**Kernel**		**Gaussian**
1	0.86/0.80	0.86/0.80	0.71/1.00	**1**	0.00/1.00	0.71/0.80			0.29/0.80
2	0.71/1.00	0.71/1.00	0.57/1.00	**2**	0.29/0.60	0.71/0.80	**Box**	0.86/0.60	
3	0.43/1.00	0.57/1.00	0.43/1.00	**3**	0.71/0.80	0.86/0.80	**Triangle**	0.71/0.60	
4	0.71/1.00	0.71/1.00	0.86/1.00	**4**	0.86/0.80	0.86/0.60	**Gaussian**	0.71/0.40	
5	0.43/1.00	0.43/1.00	0.43/1.00	**5**			**Epanechnikov**	0.71/0.60	
6	0.29/0.60	0.43/0.60	0.29/0.60	**6**		0.86/0.40			
7	0.00/0.60	0.00/0.60	0.00/0.60	**7**	0.86/0.60	0.86/0.60			
8	0.00/1.00	0.00/1.00	0.00/1.00	**8**					
9				**9**					

for both distance metrics. For the Chebyshev metric, KNN with *k=4* performed the best with 86% sensitivity and 100% specificity. The SVM algorithm with polynomial option of order 5 provided the highest sensitivity and specificity values of 86% and 80%. With the SVM algorithm with Gaussian option, the third order produced the same results as the fifth polynomial order. The NB classifier with the "box" kernel provided the highest sensitivity and specificity values of 86% and 60%, respectively. Overall, the KNN classifier (*k=4* and Chebyshev distance) showed the best classification performance among the classifiers.

VI. Conclusions and Discussions

TIS and its algorithms together with machine learning techniques are capable of differentiating malignant and benign tumors *in vivo*. Size and stiffness of tumors are estimated using TIS. The performances of the three classification algorithms (KNN, SVM, and NB) were compared using sensitivity and specificity.

We tested the developed method using the pilot TIS dataset of 12 human patients. We achieved best performance of 86% sensitivity and 100% specificity with LOOCV using KNN classification. These primary results demonstrate the clinical feasibility of TIS, and provide us information for selecting the classifier for the TIS dataset. Note that the pilot dataset is not large enough for a comparative study of the classifiers with the published clinical results in the literature. A full scale clinical test is left as a future work.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. DGE-1144462, and the Tobacco Formula Fund of Pennsylvania Department of Health. The authors thank Amrita Sahu for supporting our data collection. We also thank Dr. Joseph Picone for the helpful discussions. We are thankful to Dr. Rouzbeh Tehrani and Dr. Gangadhar Andaluri from Civil & Environmental Engineering Department, Temple University for letting us use their lab facilities. We also thank Anthony Lowmen, PhD, and Samuel Laurencin from the College of Engineering, Rowan University, NJ for letting us use their instruments.

References

[1] P. Wellman, E. P. Dalton, and D. Krag, "Tactile Imaging of Breast Masses," *Arch. Surg.*, vol. 136, no. 2, pp. 204–208, 2001.

[2] T. Krouskop, T. Wheeler, F. Kallel, B. Garra, and T. Hall, "Elastic Moduli of Breast and Prostate Tissues Under Compression," *Ultrason. Imaging*, vol. 20, no. 4, pp. 260–274, 1998.

[3] J.-H. Lee and C.-H. Won, "The tactile sensation imaging system for embedded lesion characterization," *IEEE J. Biomed. Heal. Informatics*, vol. 17, no. 2, pp. 452–458, 2013.

[4] A. Itoh, E. Ueno, E. Tohno, H. Kamma, H. Takahashi, T. Shiina, M. Yamakawa, and T. Matsumura, "Breast disease: clinical application of US elastography for diagnosis.," *Radiology*, vol. 239, no. 2, pp. 341–350, 2006.

[5] K. Polat and S. Güneş, "Breast cancer diagnosis using least square support vector machine," *Digit. Signal Process.*, vol. 17, no. 4, pp. 694–701, 2007.

[6] E. A. Krupinski, "Computer-aided detection in clinical environment: benefits and challenges for radiologists.," *Radiology*, vol. 231, no. 1, pp. 7–9, 2004.

[7] B. B. Samir, H. R. H. Al-Absi, and K. Kassoul, "Breast cancer classification using cluster k-nearest neighbor," vol. 382, no. May 2015, pp. 382–385, 2012.

[8] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Syst. Appl.*, vol. 36, no. 2 PART 2, pp. 3240–3247, 2009.

[9] S. Kharya, S. Agrawal, and S. Soni, "Naive Bayes Classifiers : A Probabilistic Detection Model for Breast Cancer," vol. 92, no. 10, pp. 26–31, 2014.

[10] V. Oleksyuk, F. Saleheen, Y. Chen, and C. H. Won, "Tactile Imaging System for inclusion size and stiffness characterization," *2015 IEEE Signal Process. Med. Biol. Symp. - Proc.*, pp. 1–6, 2015.

[11] "Instron Model 4400 Universal Testing System Operator's Guide," Norwood, MA, 1995.

[12] C. Y. Wu, R. H. Patt, M. T. Freedman, and S. K. Mun, "Diagnosis of breast cancer by MRI: A 3D computer visualization and analysis system," in *Proceedings of SPIE - The International Society for Optical Engineering. Vol. 2707*, 1996, pp. 87–96.

[13] R. Gonzalez and R. Woods, *Digital Image Processing. Upper Saddle River, N.J: Prentice Hall; 2002.*, Third. Pearson/Prentice Hall, 2008.

[14] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd. New York: Wiley, 2001.

[15] Mathworks Inc., "Nonlinear Transformation with Kernels," 2016. [Online]. Available: http://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html. [Accessed: 01-Jan-2016].