Classification of Invasive Lobular Carcinoma (ILC) and Invasive Ductal Carcinoma (IDC) Using the Support Vector Machine (SVM) Method

Anak Agung Ngurah Gunawan [1], Putu Suardana [1], Albert Sulaiman [2],
Anak Agung Ngurah Franky Kusuma Negara [3],
Anak Agung Ngurah Surya Mahendra [1],
Anak Agung Ngurah Frady Cakra Negara [3]

[1] Department of Physics, Faculty of Mathematics and Natural Sciences
Universitas Udayana, Bali, Indonesia

[2] Research Center for Physics-Badan Riset dan Inovasi Nasional (BRIN)
Tanggerang Selatan 15311, Indonesia

[3] Department of Technical Information
Faculty of Mathematics and Natural Sciences
Universitas Udayana, Bali, Indonesia

Corresponding author: Anak Agung Ngurah Gunawan

Abstract

Breast cancer is a deadly disease because of the delay in early detection, so early detection methods are very active research today. One of the early detection methods that have been widely used is an X-ray mammogram. However, this method has a limited amount of data. A practical method for early detection with a limitation database is needed, such as a Support Vector Machine (SVM). The paper used the physical parameters contained in the X-ray mammography results as input parameters of the SVM to classify histopathological types such as Invasive Ductal Carcinoma

(IDC) and Invasive Lobular Carcinoma (ILC). The experimental results show that the SVM method is a practical algorithm for histopathological classification, especially IDC and ILC. The provision of the selection of physical parameters as input variables must be made at the pre-processing stage.

Keywords: breast cancer, invasive ductal carcinoma, invasive lobular carcinoma, supportvectormachine

1. Introduction

Breast cancer is a deadly disease for women all over the world. There is still an urgent need for
developing new techniques for cancer screening, diagnosis, and intraoperative surgical guidance.
One of the early detection methods that have been widely used is an X-ray mammogram. However, this tool is less effective for patients less than 40 years old and less sensitive to small tumors under 1 mm and dense breasts. It can use another instrument for breast cancer detection. For example, ultrasonography (USG) may be used for screening dense breasts [Ozmen et al 2015]. Magnetic Resonance Imaging (MRI) can screen small breasts that cannot be detected by mammography [Raganovid et al 2015]. Emission Positron Tomography (PET) is a very accurate method for visualizing the spread of breast cancer [Xu et al 2015]. Computed Tomography (CT) Scan is the most accurate method to see the spread of breast cancer [Lee et al 2007]. Mammograms can detect tumors because the tumor tissue absorbs more X-rays than the surrounding tissue. Despite its shortcomings, it turns out that mammography can see microcalcifications, which are generally an indication of being a malignant tumor, so it can be used for early detection of breast malignancy [Champaign and Calderboom 2000].

Despite the limitations in the sensitivity of mammography, improvements in technology have been made, including film/screen combinations, Computer-Aided Detection (CAD), and Digital Mammography. However, mammography can still be used with the development of advanced algorithms for early detection. Many algorithms for early detection of the presence of microcalcification have been developed. For instance, texture coding, neural network, support vector machine (SVM), deep learning, extraction, edge detection, adaptive k-means clustering, self-similar fractal, contour description, orthogonal polynomials model, Gabor features, vector quantization technique [Gao et al 2020, Condon et al 2021]. Besides determining the presence of micro-calcification, other researchers use the chest muscle for early detection of breast cancer because it has the same pixel intensity as the lesion [Wong et al 2016].

In developed countries that have a complete database on breast cancer, early detection is easier to do. But in some countries, there is a limitation in the database. A practical method for early detection with a limitation database is needed. One method that is widely used in this problem is SVM. The advantage of this method is still compelling in case the number of dimensions is greater than the number of samples [Rejani and Selvi 2009, Sweilam et al 2010]. The branch of SVM called Adaptive support vector machine (ASVM) successfully classified the point of coarse calcification and false calcification point with an accuracy of up to 94% [Cai et al 2019]. SVM application for detecting breast cancer shows that the SVM ensemble is based on a linear kernel based on the bagging method. The RBF kernel SVM-based ensembles with the boosting method can be better for small-scale datasets. RBF kernel-based SVM ensembles based on upgrades to large-scale datasets perform better than other classifiers [Huang et al 2017, Wu and Hick 2021]. The results of other studies show that the algorithm (SVM) performs better in terms of log-loss and classification accuracy rates of other underlying models. Combining SVM with other methods such as fuzzy c-means (FCM) or Genetic Algorithm clustering improves classification accuracy in cases of malignant tumors [Ghosh et al 2021, Resmini et al 2021].

As mentioned above, the SVM method has enormous potential for early detection of breast cancer. The working principle is based on an integrated grouping where each type of breast cancer has a different intensity pattern from one type to another. For instance, to find IDC and ILC, anatomical pathologists use a fine needle biopsy to detect cancer. While in this study, we used the physical parameters contained in the X-ray mammography results as input parameters of the SVM Method. In previous studies, we detect the type of histopathology (the degree of malignancy, malignant and benign lesions) by using the physical mammography parameters. Using the same parameter, we also determine the contralateral health condition of breast cancer using mathematical modeling methods [Gunawan et al. 2012, Gunawan, 2014]. With mathematical modeling, it is challenging to determine IDC and ILC, so data-driven classification is necessary. This method is what we do in this study with the SVM algorithm. Although the SVM method is successful for early detection of breast cancer, it is not yet known with certainty its success in classifying histopathological types (IDC and ILC). Therefore, the findings of this paper allow future researchers to choose an effective method for classifying the histopathological types of breast cancer.

2. Methodology

2.1. The Model

The SVM learning part of machine learning is a powerful classification tool for pattern recognition algorithm. The SVM method introduced by Vladimir Vapnik is more effective in many pattern recognition problems and provides better classification performance among many other classification techniques [Vapnik 1998, Huang et al.

2017, Dhawan 1996]. SVM performs binary classification, which separates a set of training vectors for two different classes *(x₁, y₁), (x₂, y₂) ,. . ., (xₘ, yₘ),* where the data pount $x_i = \{x_1, x_2, \ldots x_m\} \in R^d$ the data classes: $y_i \in \{-1, +1\}$, d is dimension. Next, we map the input vector to a new higher dimensional feature space $\phi: R^d \rightarrow H^f$ where $d < f$ (see Fig 1). Then, the optimal dividing hyper plane in the new feature space is constructed by the kernel function *K(xᵢ,xⱼ)*, which is the product of the input vector *xᵢ* and *xⱼ* where *K(xᵢ, xⱼ) = F (xᵢ) F (xⱼ)*. Specifically, all vectors lying on one side of the hyper plane are labelled as -1 as IDC, and all vectors lying on the other side are labelled as +1 as ILC.
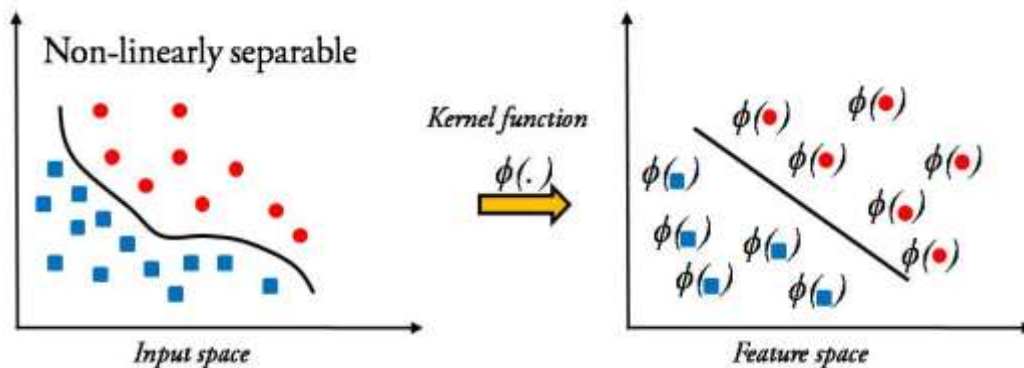
**Figure 1**. Illustration of linear kernel methods based SVM procedure

Physical parameters are physical quantities contained in the mammogram image which are converted into the form of angular second moment, contrast, entropy, inverse difference moment, deviation, correlation, mean, the entropy of *h_diff*, mean *h_diff* and angular second moment (ASM) of *h_diff.* The physical parameters of the mammogram are calculated by the formula [Dhawan 1996, Enda 2009, Aswini and Saroj 2010, Vasantha et al 2010],

2.2.The Data

Data were obtained from Doctor Sutomo Hospital Surabaya. There are two types of breast cancer data from radiology databases: IDC type data and ILC data type, consisting of 156 IDC data samples and 8 ILC data samples. The experimental procedure was carried out as follows. First, collect data from the Radiology database at Dokter Sutomo Hospital Surabaya and divide it into 90% for training data and 10% for testing data. The second step focuses on selecting the physical parameters used as input parameters for the SVM method using the T statistic test. Then do the classification using the SVM method to calculate accuracy, sensitivity, and specificity.

3. Result and Discussion

After selecting the physical parameters as input variables for the SVM method, the statistical results are shown in Figure 2. The results show the T-test results, where the mean parameter is selected as the SVM input variable. Figure 3 shows the mean IDC and ILC data before the SVM classification is carried out, and after the SVM classification, the IDC and ILC mean data are separated, as shown in Figure 4. Figure 5 and Figure 6 show that there are differences in the visualization results of IDC and ILC. In particular, the performance of SVM with the mean as input variable obtained accuracy, sensitivity, and specificity of 76.56%, 76.79%, and 75%, respectively.
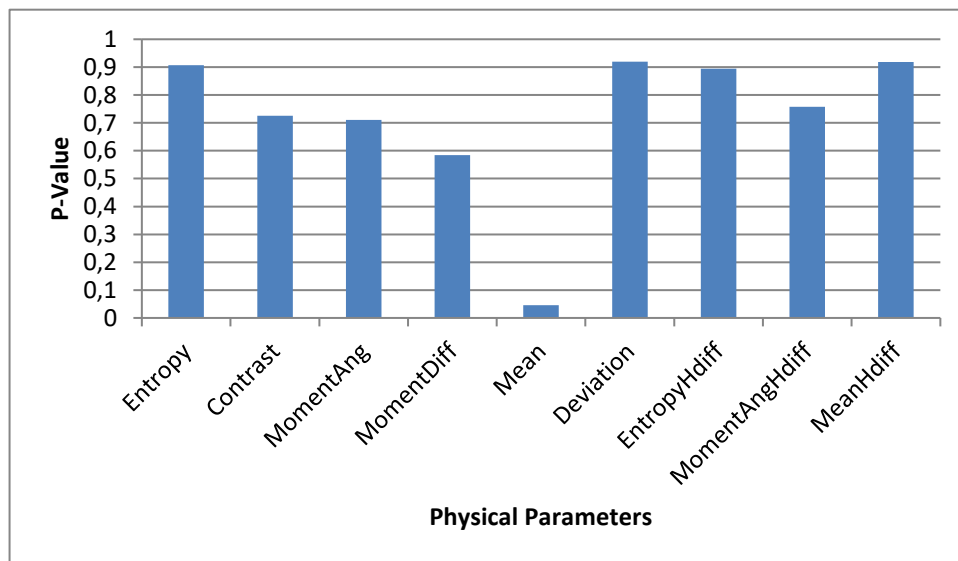

Figure 2. T-test results of physical parameters to differentiate histology of IDC and ILC with a mean P-value of 0.046. This value meets the requirements for the input parameter in the SVM method, which is <0.05 for the mean value.

Of the nine physical parameters in the mammography image, only one can distinguish the histopathological types of IDC and ILC, namely the mean parameters as shown in Figure 2. Furthermore, this means parameter will be used as an input parameter for the SVM method.
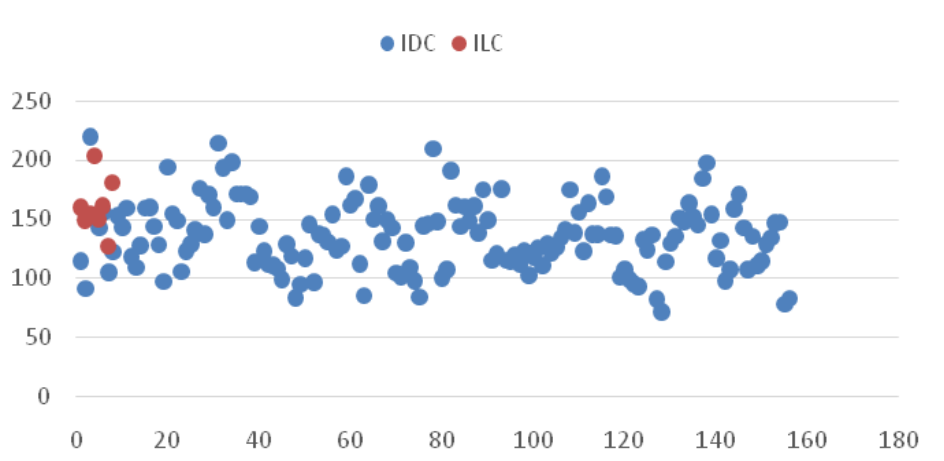
Figure 3. Conditions before SVM classification, the horizontal axis represents the nth data and the vertical axis represents the average value. The red and blue dots cannot be separated linearly because there is no classification separator in these two classes.
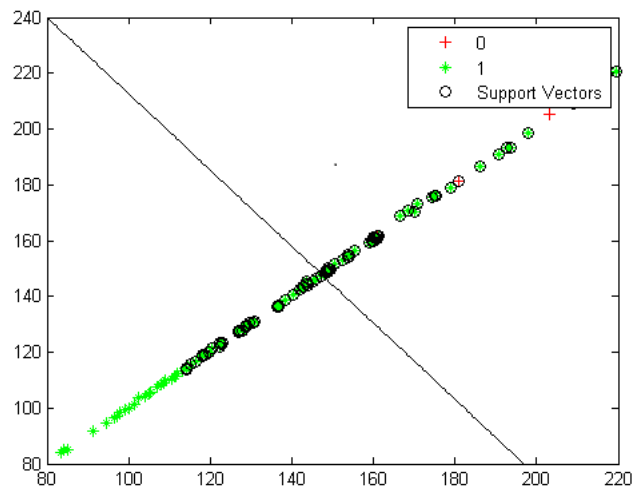

Figure 4. After SVM classification, the horizontal axis represents the nth data and the vertical axis represents the average value. The data is modified using kernel functions. Red and blue classes can now be separated linearly.

Figure 3 shows that the mean data obtained were in clusters that were difficult to separate before using the SVM method. After using the SVM method, the data is separated by a dividing line, as shown in Figure 4.
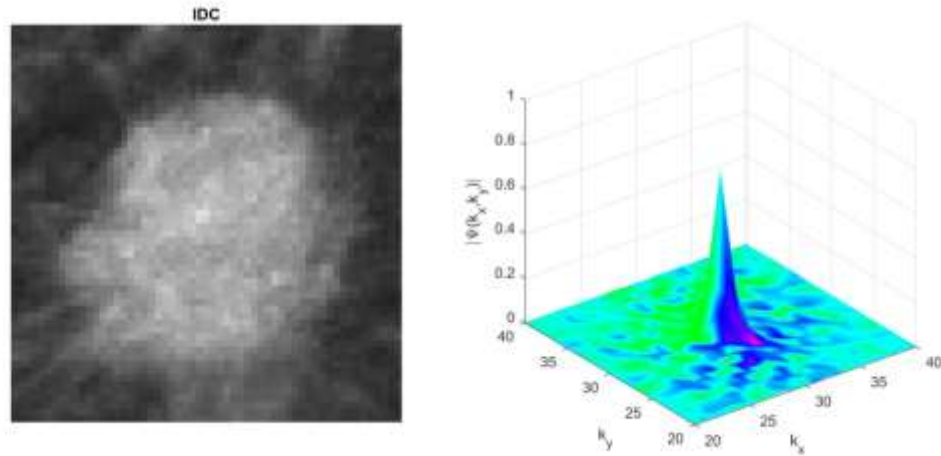
Figure 5. (a) ROI IDC, (b) normalized power spectrum of IDC.
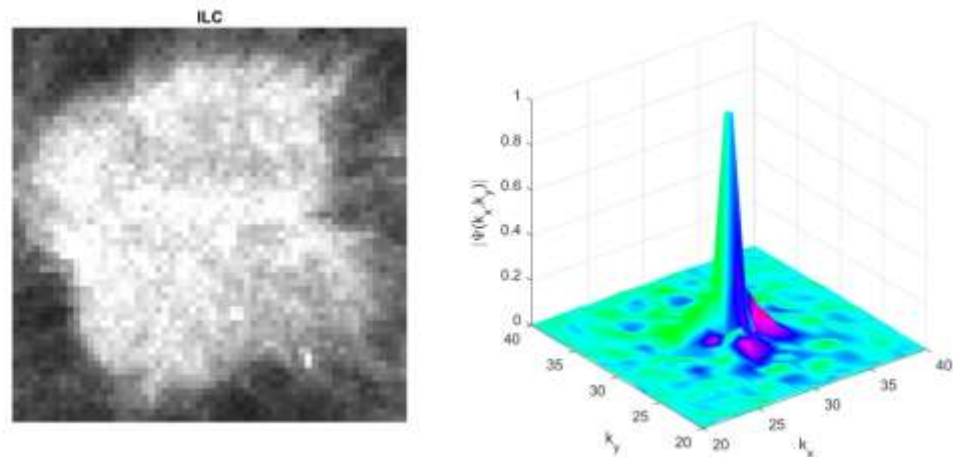
Figure 6. (a) ROI ILC, (b) normalized power spectrum of ILC.

The difference in the graph between IDC and IDC is due to the significantly different mean value between IDC and IDC. ILC has a higher intensity than IDC, where a brighter color characterizes it (see Fig 5 and Fig 6). The IDC spectrum has a low amplitude centered on wavenumber 30/pixel, while the ILC spectrum has a higher amplitude with a broader spectrum.

Previous researchers have carried out the histological determination of IDC and IDL using a mathematical model based on physical parameters [Gunawan et al. 2012]. The classification is based on the optimization of the model. The probability of IDC occurrence follows the Velhuzs distribution (sigmoid function), while the probability of ILC is 1- the probability of IDC. The development of this model is to change the coupling parameters with increasing accuracy from 93.3% to 97.5% [Gunawan 2014]. Several statistical methods were developed, such as K-nearest Neighbourhood, with an accuracy of 70% [Gunawan et al. 2018]. The SVM method was developed to classify breast cancer in normal and abnormal classes with a classification of 86%. The normal-abnormal classification uses the Bilateral Asymmetry method, which has an accuracy of 84%, while the GA+RBF SVM method has an accuracy of 98% [Zeng 2010]. In the microcalcification classification, the Neural Network method has the worst performance. Otherwise, the Orthogonal Polynomials have the best performance [Dheeba and Wiselin 2010]. The Contour method has the poor performance to classify Benign-malignant, while the K-mean method has good performance [Bagwati and Shinha 2010]. Of all the methods mentioned above, no method has been proposed to classify the histopathological types of IDC and ILC. Although the accuracy is not too high, this is the first step for further research.

4. Conclusion

This paper investigates the classification of the histopathological breast cancer types IDC and ILC by using the SVM method. In addition, many physical parameters of mammographic images are compared to get accurate parameters. It can detect the histopathology of breast cancer types IDC and ILC. We compare various methods to classify the accuracy of normal, abnormal types and benign-malignant types. The histopathological classification of IDC and ILC types of breast cancer has never been shown before. The experimental results allowed us to understand the SVM method's predictive performance fully. Better predictive models can be identified as basic classifiers for future studies. In future research, genetic algorithms (GA) and SVM may be used to improve the accuracy of histopathological classification of breast cancer types IDC and ILC.

Author contribution: AANG designing experiments AANG, PS experiment, AANFKN, AANFCN data analysis, AANG,PS,AS,AANFKN, AANSM, AANFCN write paper.

References

[1] K. M. Aswini and K. L. Saroj, Efficient Image Mining Technique for Classification of Mammograms to Detect Breast Cancer, *International Journal of Computer and Communication Technology*, (2012), 44-51.

https://doi.org/10.47893/ijcct.2012.1113

[2] C. P. Bhagwati and G. R. Sinha, Early Detection of Breast Cancer using Self Similar Fractal Method, *International Journal of Computer Applications*, **10** (2010), no. 4, 39-43. https://doi.org/10.5120/1466-1981

[3] H. Byun, S.-W. Lee, A survey on pattern recognition applications of support vector machines, *International Journal of Pattern Recognition and Artificial Intelligence*, **17** (2003), no. 3, 459-486. https://doi.org/10.1142/s0218001403002460

[4] S. Cai, S.Z. Liu, Y.M. Luo, Y.Z. Du, J.N. Tang, Breast Microcalcification Detection Algorithm Basedon Contourlet and ASVM, *Algorithms,* **12** (2019), 135. https://doi.org/10.3390/a12070135

[5] J.L. Champaign, G.J. Cederbom, Advances in breast cancer detection with screening mammography, *The Ochsner Journal*, **2** (2000), 33-35.

[6] Condon, J.J., Oaken-Rayner, L., Hall, K.A., Reintals, M., Holmes, A., Carneiro, G., Palmer, LJ., Replication of an open-access deep learning system for screening mammography: Reduced performance mitigated by retraining on local data, *medRxiv preprint.* https://doi.org/10.1101/2021.05.28.21257892

[7] A.P. Dhawan, Analysis of Mammographic Microcalcifications Using Gray Level Image Structure Feature, *IEEE Trans. Medical Imaging*, **15** (1996), 246-257. https://doi.org/10.1109/42.500063

[8] J. Dheeba and J.G. Wiselin, Detection of Microcalcification Clusters in Mammograms using Neural Network, *International Journal of Advanced Science and Technology,* **19** (2010), 13-22.

[9] M. Enda, *Automated Image Analysis Techniques for Screening of Mammography Images*, National University of Ireland, Galway, 2009, 5-23.

[10] Y. Gao, B. Liu, Y. Zhu, L. Chen, M. Tan, X. Xiao, Y. Buo, Detection and Recognition of Ultrasound Breast Nodules based on semi-supervised deep learning, *medRxiv*. https://doi.org/10.1101/2020.04.24.20078816

[11] S. Ghosh, G. Samanta, M. De la Sen, Multi-Model Approach and Fuzzy Clustering for Mammogram Tumor to Improve Accuracy, *Computation,* **9** (2021), 59. https://doi.org/10.3390/computation9050059

[12] Gunawan, A Novel Model Determination of Breast Cancer Stage Using Physical Parameter, *Far East Journal of Mathematical Sciences (FJMS),* **87** (2014), no. 1, 23-35.

[13] A. A. Gunawan, K. S. P. Suhariningsih, Triyono and B. Widodo, Determination of Physical Parameter Model for the Photo Film Mammographic X-Ray Results on the Breast Cancer Histology Classification*, International Journal of Contemporary Mathematical Sciences*, **7** (2012), no. 45-48, 2236-2244.

[14] Gunawan, K.S.P. Triyono Suhariningsih and Yasin, Conversion of images into numerical models to determine the condition of breast health on Contralateral, *Applied Mathematical Sciences*, **7** (2013), 5185-5191. https://doi.org/10.12988/ams.2013.212669

[15] Gunawan, Supardi and Dharmawan, Readability Increase Of MammographyX-Ray Photos Results In Determining The Breast Cancer Histopathology Types Using Special Pattern Cropping With Physical Parameter*, Advances in Applied Physics*, **2** (2014), no. 1, 43-52. https://doi.org/10.12988/aap.2014.4910

[16] Gunawan, Widana, Determination model of carcinoma and benign types using physical quantities of the ca mammae mammography, *Wulfenia Journal*, **22** (2016), no. 7, 181-188.

[17] M.W Huang, C.W Chen, W.C. Lin, S.W Ke, C.F.Tsai, SVM and SVM Ensembles in Breast Cancer Prediction, *PLOS One*, **12** (2017). https://doi.org/10.1371/journal.pone.0161501

[18] Lee E.W., Grant J.D., Mari C., Incidental PET/CT Detection of Breast Cancer in a Patient with Negative Mammogram and Breast Sonogram, *Radiology Case Reports*, **2** (2007), no. 3, 84. https://doi.org/10.2484/rcr.v2i3.84

[19] N. Ozmen, R. Dapp, M. Zapf, H. Gemmeke, N.V. van Ruiter, K.W. Dongen, Comparing different ultrasound imaging methods for breast cancer detection, *IEEE Tran. Ultrason. Ferroelectr.*, **62** (2015), no. 4, 637–646. https://doi.org/10.1109/TUFFC.2014.006707

[20] Y.I.A. Rejani, S.T. Selvi, Early detection of breast canver using SVM classifier technique, *International Journal on Computer Science and Engineering*, **1** (2009), no. 3, 127-130.

[21] R. Resmini, L. Silva, A.S. Araujo, P. Medeiros, D. Muchaluat-Saade, A. Conci, Combining Genetic Algorithms and SVM for Breast Cancer Diagnosis Using Infrared Thermography, *Sensors*, **21** (2021), 4802. https:// doi.org/10.3390/s21144802

[22] D. Roganovic, D. Djilas, S. Vujnovic, D. Pavic, D. Stojanov, Breast MRI, digital mammography and breast tomosynthesis: Comparison of three methods for early detection of breast cancer, *Bosn. J. Basic Med. Sci.*, **15** (2015), no. 4, 64-68. https://doi.org/10.17305/bjbms.2015.616

[23] N.H. Sweilam, A.A. Tharwat, N.K.A. Moniem, Support vector machine for diagnosis cancer disease, A comparative study, *Egyptian Informatics Journal*, **11** (2010), no. 2, 81-92. https://doi.org/10.1016/j.eij.2010.10.005

[24] V. Vapnik, *Statistical Learning Theory*, John Wiley, 1998.

[25] M. Vasantha, S. Bharathi, R. Dhamodharan, Medical Image Feature, Extraction, Selection and Classification, *International Journal of Engineering Science and Technology*, **2** (2010), no. 6, 2071- 2076.

[26] L. Wang, Early Diagnosis of Breasy Cancer, *Sensors*, **17** (2017), no. 7, 1-20. https://doi.org/10.3390/s17071572

[27] Wong B. Y., Ji E. O., Eun Y. C., Hak H. K., Soo Y. L., and Kwang G. K., Automatic Detection of Pectoral Muscle Region for Computer-Aided Diagnosis Using MIAS Mammograms, *BioMed Research Int.*, **2016** (2016), article ID 5967580. http://dx.doi.org/10.1155/2016/5967580

[28] J. Wu, C. Hicks, Breast Cancer Type Classification Using Machine Learning, *J. Pers. Med.*, **11** (2021), 61. https://doi.org/10.3390/jpm11020061

[29] P. Xu, Y. Peng, M. Sun, X. Yang, SU-E-I-81, Targeting of HER2-Expressing Tumors with Dual PET-MR Imaging Probes, *Med. Phys*., **42** (2015), no. 6, 3260-3260. https://doi.org/10.1118/1.4924078

[30] Zheng Yufeng, Breast Cancer Detection with Gabor Features from Digital Mammograms, *Algorithms*, **3** (2010), 44 - 62. https://doi.org/10.3390/a3010044