A Performance Enhancement of Breast Cancer Detection Model using Ensemble Classifier

Aya Hossam
Electrical Engineering Department
Faculty of Engineering (Shoubra), Benha University
Cairo, Egypt
Aya.ahmed@feng.bu.edu.eg

Islam Hany M Harb
Virginia Tech
iharb@vt.edu
Virginia, USA

Hala M. Abd El Kader
Electrical Engineering Department
Faculty of Engineering (Shoubra), Benha University
Cairo, Egypt
hala.mansour@gmail.com

***Abstract*—Breast cancer is one of the most common malignancies among women in the world, which may lead to death. Survivability rate of breast cancer can be improved if it is identified in its early stage. Breast thermography plays an important role in early detection of breast cancer, since it couples the physiological information with the anatomical features of woman breast. Typically, thermograms are visually analyzed by physicians for breast cancer early diagnosis. But it is very challenging, since it is hard to provide objective and quantitative *analysis*. Therefore, Computer Aided Detection (CAD) systems are used to improve the diagnostic accuracy by providing a comprehensive analysis on these Thermograms. One of the important factors that impact CAD system's performance is the classifier used for the classification of breast thermograms. However, problems such as low rate of accuracy and poor self-adaptability still exist in traditional classifiers. In this paper, a hybrid approach consists of support vector machine (SVM) classifier, with feature selection and boosting ensemble method, is proposed to enhance the performance of SVM classifier. AdaBoost algorithm is used as our boosting ensemble method. The experimental results show that the proposed hybrid approach achieves a better performance compared to the base classifier SVM alone and the base SVM classifier coupled with feature selection method. An accuracy of 99.24% is obtained using our hybrid approach.**

Keywords—Breast Cancer, Thermography, Support Vector Machine, Boosting, Ensemble.

I. INTRODUCTION

Breast cancer is the most commonly diagnosed cancers among the middle aged women. It is considered as the second most common cause of cancer death among females [1]. The early detection and accurate diagnosis of breast cancer can lead to successful treatment and improve the survivability rate of the patients [2]. Breast thermography is a relatively new imaging tool used for early detection of breast cancer. It is based on temperature that might be produced by a tumor. It is a non-invasive functional imaging test, low-cost, harmless, fast and sensitive method. It also can be utilized for women of all ages, and suitable for women with dense breast tissues where mammography is less efficacious [3-5]. Early detection needs a precise and reliable breast diagnosis procedure that allows physicians to distinguish between normal breast thermograms and abnormal ones [6]. For this purpose, there are various CAD systems to serve as the breast diagnosis procedure and help the radiologists in detecting the abnormal regions present in the breast. These systems act as a second reader, while the final decision lies with the radiologist [7-9]. Breast cancer diagnosis benefits from the advancements in data mining techniques in CAD.

In the domain of data mining and CAD systems, classifier plays an important role to classify breast thermogram images into normal and abnormal cases. Classification is considered as one of the most important factors that impacts the performance/accuracy of CAD system. Classification approaches are divided into two categories which are supervised and unsupervised. Generally speaking, supervised classification techniques are more suitable for classification of breast thermograms than unsupervised techniques. Ssupervised classifiers include, but not limited to, Support Vector Machines (SVM) [10], Artificial Neural Networks (ANN) [10] [11], Naive Bayes classifier [12] [13] and Fuzzy classifier [14] [15]. The classification performance of the breast cancer detection model depends on many factors such as ROI extraction, selected features, and model structure. One of the tough challenges that faces researchers is to build a breast cancer detection model with high accuracy, high reliability and robustness [16].

Ensemble learning is one of the most popular methods that can be used to improve the classification accuracy. These methods are a class of highly successful machine learning algorithms which combine several models to obtain an ensemble which is, supposedly, more accurate than its individual members [17]. This paper proposes a hybrid approach which consists of a SVM classifier with feature selection and boosting ensemble method to enhance the accuracy of SVM classifier and compensate for any limitations. This proposed ensemble model includes SVM as a base classifier after applying the feature selection algorithm based on particle swarm optimization technique as a search method, named PSO-FS algorithm in [18]. Then, SVM classifier is adopted to be hybridized as an ensemble learning model for breast cancer detection. This paper used the AdaBoost method as a boosting algorithm [19]. The proposed breast cancer model is validated and evaluated based on the breast cancer data set represented in [20] which are extracted from breast thermograms. This data set contains around twenty five attributes and about 450 instances.

II. THE PROPOSED SYSTEM

In this section, we will discuss the data and methods used in this work to build the proposed hybrid. This approach is a SVM-based ensemble algorithm. The two main components in the proposed methodology are: SVM classifier, and ensemble method. The input breast cancer data set to our model was extracted from ~225 (120 abnormal and 105 normal) thermograms that are presented in [20]. The performance of base model using just the SVM classifier is studied to show the impact of model parameters on model accuracy. Then, the performance of SVM classifier is evaluated once more, but first the feature selection method (i.e., PSO-FS algorithm) is appplied, which is proposed in [18]. Finally, AdaBoost method is applied as an ensemble method to compensate the weakness and enhance the accuracy of individual SVM classifiers. The SVM method along with AdaBoost can be applied on balanced data, as well as, imbalanced data. Usually, Boosting is done at the end so that all the output weak learners are clubbed to form a strong learner. Boosting focuses more on the misclassified examples, or on the examples that have higher prediction errors. Figure 1 shows the flowchart of the proposed hybrid system.

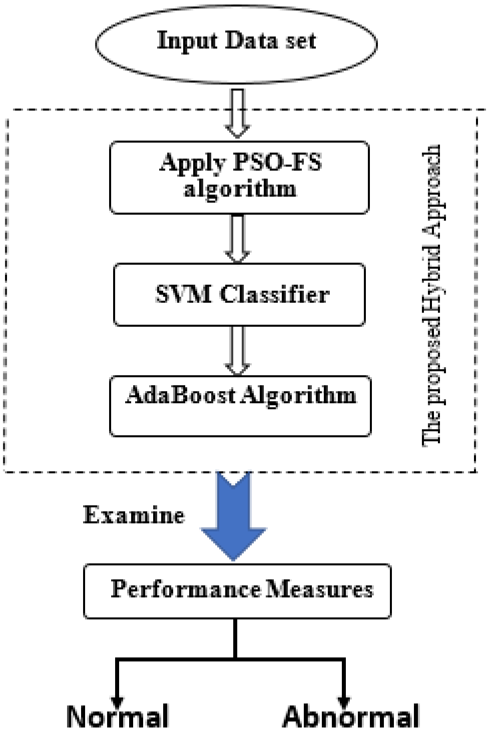


Fig.1. Flow chart of the Proposed Hybrid System.

A. *Feature Selection Process*

In machine learning and data mining domains, the feature selection (FS) process is considered as an active research area for decades. It can be defined as the process of reducing the total large number of attributes in any dataset by choosing the optimal feature subset from them based on a certain criteria [21] [22]. In this paper, the PSO-FS algorithm presented is applied to select the optimal feature subset from the overall input features. This algorithm used the PSO technique as a search method through the FS process.

In PSO-FS algorithm, the search space dimensionality is n, where n is the total number of features in the Breast Cancer (BC) dataset. Therefore, the search space size is 2n. Each particle is randomly initialized in terms of both the number of features and the combination of individual features. The position value of each particle i in the dth dimension (i.e., X_{id}) is in the interval [0, 1]. A threshold θ, 0 < θ < 1, is required in order to determine whether a feature will be selected or not. If $X_{id} > \theta$, then the corresponding feature d is selected. Otherwise, feature d is not selected.

B. *Support Vector Machine (SVM) Classifier*

SVM is a powerful supervised classification algorithm that differentiates between two classes by finding a hyperplane that separates between them [23][24]. It performs classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories. This hyperplane can be presented as a wide line or two parallel lines with maximum distance where there is no data point between them. Figure 2 shows the concept of SVM Operation. Although there can be a lot of possible separating lines for a given set of objects, not all the separating lines are equally good. Among the possible hyperplanes, SVM searches for the one that maximizes the distance between the two support vectors (called a margin) as shown in figure 2. The Support Vectors are the points closest to the separating hyperplane. So, the best hyperplane is the one that provides the biggest distance between the closest members of both classes. The criteria increases the probability of classifying the data point to its correct corresponding class. This is called the linear classifier.

Referring to figure 2, let {x_1, ..., x_m} are the values of the data set and yi ∈ {1,-1} is the class label of x_i. Suppose that a pair of variables (w, b) defines a hyperplane which has the following equation:

$$w.x_i + b = 0 \qquad (1)$$

where w is the normal weight vector which is perpendicular to the hyperplane and b is an intercept term that represents a shift of the hyperplane from the origin of the coordinate system. Also, the training data can be described by the following equations:

$$w.x_i + b \geq +1 \qquad \text{if } y_i = +1 \qquad (2)$$

$$w.x_i + b \leq -1 \qquad \text{if } y_i = -1 \qquad (3)$$

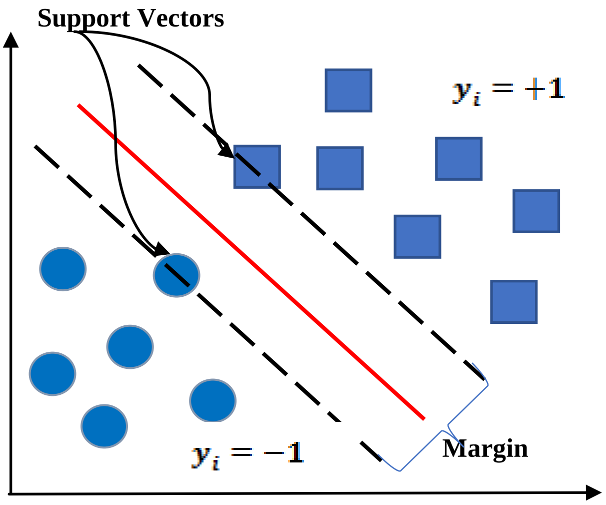


Fig.2. The concept of Support Vector Machines Operation.

C. *Ensemble Classification*

Ensemble classification, or ensemble learning, is the process of combining multiple classifiers to get an improved performance of a single classifier [25] [26]. The concept of ensemble classification return to the nature of information processing in the brain, which is modular. As the individual functions can be subdivided into functionally different sub-process or subtasks without mutual interference [27]. The ensemble based systems produce most favorable results than single-expert systems under a variety of scenarios for a broad range of applications. There are many ensemble-based algorithms such as Bagging, Boosting, Stacked Generalization and Hierarchical Mixture of Experts.

The ensemble based systems can help to improve the confidence with which making the right decision through a process in which various opinions are weighed and combined to reach a final decision. Some of the reasons for using these systems are as follows [28]:

- Divide and Conquer: An individual or base classifier is unable to solve specific problems. In some cases, the decision boundary for different classes may be very complex. Due to that, the complex decision boundary can be estimated by combing different classifiers appropriately.
- Data Fusion: A single classifier has not the ability to learn information contained in data sets with heterogeneous features. Ensemble based approaches are most suitable for which called data fusion applications. In this applications, data from different sources are combined to make more informed decision.
- Large Volumes of data: The amount of data is too large to be analyzed effectively by a single classifier.
- Too little data: Resampling techniques can be used to overlap random subsets of inadequate training data and each subset can be used to train a different classifier.
- Statistical Reasons: The risk of selecting a poorly performing classifier can be reduced by combining the outputs of several classifiers by averaging.

In this paper, Boosting ensemble classification is used in our proposed model. In boosting, each classifier is trained using a different training set. However, the T classifiers are trained in a ssequential order. Nowadays, AdaBoost (or Adaptive Boosting) is the most common boosting learning algorithm used in pattern recognition and CAD systems. This paper used AdaBoost classifier as a boosting ensemble classifier.

AdaBoost classifier is an algorithm for constructing a ”strong” classifier from "simple/weak” classifiers [19]. In this algorithm, there are three main steps as follows. ***Sampling step:*** in this step, some samples (St) are selected from the training set, where St is the set of samples in the iteration t. ***Training step:*** in this step, different classifiers are trained using St, and the error rates (εi) for each classifier are calculated. ***Combination step:*** all trained models are combined at this step. The AdaBoost algorithm’s steps are shown in the following Algorithm (1).

Algorithm (1): AdaBoost (Adaptive Boosting) Classifier

```
Given a training set $X = (x_1, y_2)$,….., $(x_N, y_N)$ where $y_i$ represents the label of sample $x_i \in$ X and N denotes the total number of samples in the training set.
Initialize the parameters of AdaBoost classifier.
**for** t = 1 **to** T **do**
    Take a sample St from X using distribution wt.
    Use the distribution St to train the weak learner (Lt) with a minimum error ($\varepsilon$t).
    **while** $\varepsilon_t$ >= 0.5 **do**
        Reinitialize the weights to $w_j^t = \frac{1}{N}$, j = 1, . . . , N.
        Recalculate $\varepsilon_t$.
    **end while**
    Compute the weight of each weak learner ($\alpha_t$)
    Update the weights of the training samples.
**end for**
Final AdaBoost classifier:
    $L_{final} = Sign(\sum_{t=1}^{T} \alpha_t L_t(x))$.
```

III. PERFORMANCE EVALUATION AND RESULTS

In this section, the performance of the proposed hybrid approach of SVM-based AdaBoost classifier has been analyzed. The accuracies of the three methods, (1) SVM algorithm, (2) SVM with PSO-FS, and (3) SVM with PSO-FS and Boosting, are evaluated and presented in this section.

A. *Breast Cancer Dataset Description*

In this paper, the collected data set presented in [20] is adopted in our approach. In [20], the authors used a breast thermograms that were selected from an open online data base PROENG (http://visual.ic.uff.br/) which called DMR-IR database [29]. Also, the authors used an automatic segmentation method to extract the region of interest only from thermograms and remove the other parts. This segmentation helped to get an accurate attributes' values in the extracted data set. This data set contains ~450 instances with about twenty five features. The given dataset is divided into 70% training and 30% testing sets based on the10-fold cross validation strategy.

B. *Classification Results*

This paper used a WEKA software to apply SVM-based AdaBoost classifier on the breast cancer data set. The data set was given as inputs to three approaches to compare between them which are SVM classifier, SVM with PSO-FS method and finally the proposed SVM with PSO-FS and AdaBoost classifier. The classifier(s) will classify the data into two classes, namely Normal and Abnormal, which are corresponding to normal and abnormal cases respectively. The reason to choose the SVM classifier is that it was proved to be the best classifier for breast cancer diagnosis in our previous study [18] [20]. To enhance the accuracy of the SVM classifier, PSO-FS method is used. The results show that SVM with FS enhances the accuracy compared to SVM alone. In this study, a hybrid approach of applying, SVM with PSO-FS and a boosting ensemble method, on the breast cancer dataset. The procedure of the hybrid approach is done as follows: Apply the PSO-FS method on breast cancer data set. Once the optimal reduced data sets are obtained, then experiments, of applying both the boosting algorithm with the SVM algorithm on the data set, are conducted. The accuracy of this hybrid approach is tested on the selected breast cancer dataset. The comparison of the accuracies of three methods, SVM algorithm, SVM with PSO-FS and our hybrid approach is represented in Table 1.

In order to evaluate our proposed approach, there are several criteria that need to be considered, such as, Classification Accuracy, Root mean squared error (RMSE), Kappa statistic, True Positive Rate (TP-Rate), False Positive Rate (FP-Rate), Precision, Recall, F-Measure etc. These several standard terms have been defined for the two classes confusion matrix (Normal and Abnormal):

The Accuracy is used to determine the overall correctness of the model and it is calculated as the ratio of the sum of correct classifications and the total number of classifications, as determined using the equation:

$$\mathrm{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

The Precision is the proportion of the positive cases that were predicted correctly, and is calculated using the equation:

$$\mathrm{Precision} = \frac{TP}{TP+FP} \tag{5}$$

The FP-Rate is the proportion of negative cases that were classified as positive incorrectly and can be calculated using the equation:

$$\mathrm{FP-Rate} = \frac{FP}{FP+TN} \tag{6}$$

The Recall or TP-Rate is the proportion of the correctly identified positive cases, and is calculated using the equation:

$$\mathrm{Recall} = \mathrm{TP-Rate} = \frac{TP}{TP+FN} \tag{7}$$

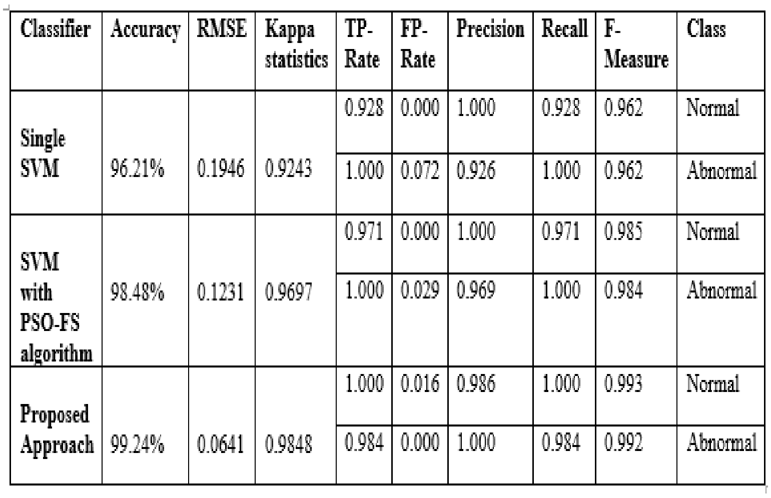
In some cases, higher precision may be very important, and in other cases higher recall may be very important. But, in most cases, both values are needed to be improved.

F-Measure is defined as the combination of these values, and in the most common form, it is the harmonic mean of the both:

$$\mathrm{F-measure} = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{8}$$

According to the comparison results presented in table 1, it is clear that by applying hybrid approach the classification accuracy is enhanced and reached to 99.24%, which is better than using a SVM classifier alone or using of the SVM classifier with feature selection method. By considering the aforementioned facts, we conclude that boosting algorithm is recommended to the breast cancer data classification along with SVM and FS over the SVM as the only classifier. The graphical representation of these results is shown in figure 3.

TABLEI THE RESULTS OF VARIOUS PERFORMANCE MEASURES OF SVM CLASSIFIER.



Classifier	Accuracy	RMSE	Kappa statistics	TP-Rate	FP-Rate	Precision	Recall	F-Measure	Class
Single SVM	96.21%	0.1946	0.9243	0.928	0.000	1.000	0.928	0.962	Normal
				1.000	0.072	0.926	1.000	0.962	Abnormal
SVM with PSO-FS algorithm	98.48%	0.1231	0.9697	0.971	0.000	1.000	0.971	0.985	Normal
				1.000	0.029	0.969	1.000	0.984	Abnormal
Proposed Approach	99.24%	0.0641	0.9848	1.000	0.016	0.986	1.000	0.993	Normal
				0.984	0.000	1.000	0.984	0.992	Abnormal

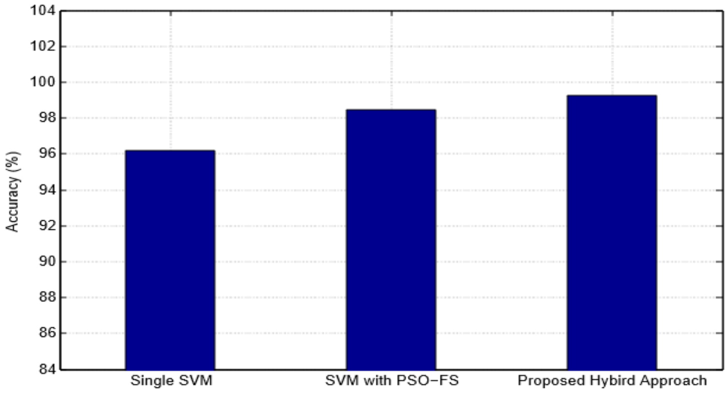


Fig.3. Accuracy (%) of SVM algorithm, SVM with PSO-FS Method and Hybrid Approach.

IV. CONCLUSION

In this paper, a hybrid approach of SVM with PSO-FS and Boosting ensemble classifier is proposed for breast cancer detection. AdaBoost classifier was applied as a boosting ensemble learning method for the diagnosis and classification of the breast cancer based on optimal reduced set of features. In the proposed approach, the PSO-FS method was applied to breast cancer data set, at first, to select the best feature subset from the whole original data. Then, boosting algorithm with the SVM algorithm were applied on the optimal reduced data set to achieve the highest accuracy. The accuracy of this hybrid approach is tested on the selected breast cancer dataset. The performance of proposed SVM-based AdaBoost classier was compared to the performance of SVM as a single classifier and also the performance of SVM with PSO-FS method. The experimental results show that the classification accuracy was enhanced and reached to 99.24%, which outperforms the SVM classifier alone and the SVM classifier with feature selection method.

REFERENCES

[1] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet- Tieulent, and A. Jemal, " Global cancer statistics, 2012," CA: a cancer journal for clinicians, vol. 65, no. 2, pp. 87-108, 2015.

[2] I. Harirchi, M. Ebrahimi, N. Zamani, S. Jarvandi, and A. Montazari, “Breast cancer in Iran: a review of 903 case records,” Public Health, vol. 114, no. 2, pp. 143-145, 2002.

[3] E.Y.K. Ng, "A review of thermography as promising non-invasive detection modality for breast tumor", International Journal of Thermal Sciences, vol. 48, no. 5, pp. 849-859, 2009.

[4] S.V. Francis, M. Sasikala, and S. Saranya, "Detection of breast abnormality from thermograms using curvelet transform based feature extraction," Journal of Medical Systems, vol. 38, no. 4, pp. 1-9, 2014.

[5] J. F. Head, F. Wang, C. A. Lipari, and R. L. Elliott, "The important role of infrared imaging in breast cancer," IEEE Engineering in Medicine and Biology Magazine, vol. 19, no. 3, pp. 52–57, 2000.

[6] T. Subashini, V. Ramalingam, and S. Palanivel, “Breast mass classification based on cytological patterns using RBFNN and SVM,” Expert Systems with Applications, vol. 36, no. 3, pp. 5284-5290, 2009.

[7] U. Bottigli, P. Cerello, P. Delogu, M.E. Fantacci, F. Fauci, G. Forni, B. Golosio, P.L. Indovina, A. Lauria, E. Torresand R. Magro, G.L. Masala, P. Oliva, R. Palmiero, G. Raso, A. Retico, A. Stefanini, S. Stumbo, and S. Tangaro, “A Computer Aided Detection System for Mammographic Images Implemented on a GRID Infrastructure,”. In Proceedings of the 13th IEEE-NPSS Real Time Conference, Montreal, Canada, pp. 18-23, 2003.

[8] K. Doi, “Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential,” Available at http://www.sciencedirect.com, (last accessed on 5th February 2016), 2007.

[9] L. Silva, A. Augusto, S.M.D. Santos, R. Bravo, A. Silva, D. Saade, and A. Conci, “Hybrid Analysis for Indicating Patients with Breast Cancer using Temperature Time Series,” Computer Methods and Programs in Biomedicine, vol. 130, no. 1, pp. 142-153, 2016.

[10] J. Tan., E. Ng, R. Acharya, L. Keith, and J. Holmes, “ Comparative Study on the Use of Analytical Software to Identify the Different Stages of Breast Cancer using Discrete Temperature Data,” Journal of Medical Systems, Vol. 33, No. 2, pp. 141-153, 2009.

[11] S. Pramanik, D. Bhattacharjee, and M. Nasipuri, “Wavelet based Thermogram Analysis for Breast Cancer Detection,” In International Symposium on Advanced Computing and Communication (ISACC) IEEE, Silchar, India, pp. 205-212, 2015.

[12] A. Lashkari, F. Pak, and M. Firouzmand, “Full Intelligent Cancer Classification of Thermal Breast Images to Assist Physician in Clinical Diagnostic Applications,” Journal of Medical Signals and Sensors, vol. 6, no. 1, pp. 12-24, 2016.

[13] C. Nicandro, M. Efren, A. Yaneli, M. Enrique, A. Gabriel, P. Nancy, G. Alejandr, H. Jesus, and B. Rocio.” Evaluation of the Diagnostic Power of Thermography in Breast Cancer Using Bayesian Network Classifiers,” Computational and Mathematical Methods in Medicine, vol. 5, pp. 1-10, 2013.

[14] G. Schaefer, M. Zavisek, and T. Nakashima, “Thermography based Breast Cancer Analysis using Statistical Features and Fuzzy Classification,” Pattern Recognition, vol. 47, no. 6, pp. 113-137, 2009.

[15] T. Z. Tan, C. Quek, G. S. Ng, and E. Ng,” A Novel Cognitive Interpretation of Breast Cancer Thermography with Complementary Learning Fuzzy Neural Memory Structure,” Expert Systems with Applications, vol. 33, pp. 652-666, 2007.

[16] Z. Bichen, Z. Jinghe, W. Y. Sang, S. L. Sarah, K. Mohammad, and P. Srikanth, " Predictive modeling of hospital readmissions using metaheuristics and data mining," Predictive modeling of hospital readmissions using metaheuristics and data mining, Expert Systems with Applications, vol. 42, no. 20, pp. 7110-7120, 2015.

[17] M. Sri Bala, and G. V. Rajya., "Efficient Ensemble Classifiers for Prediction of Breast Cancer," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6, no. 3, pp. 5-9, 2016.

[18] H. Aya, Hany M. Harb, and Hala M. Abd El Kader, "A Sub-Optimum Feature Selection Algorithm for Effective Breast Cancer Detection Based On Particle Swarm Optimization," IOSR Journal of Electronics and Communication Engineering, vol. 13, no. 3, pp. 01-12, 2018.

[19] Y. Freund and R.E. Schapire, "Decision-theoretic Generalization of Online Learning and an Application to Boosting," Journal of Computer and System Sciences, vol. 55, no.1, pp.119-139, 1997.

[20] H. Aya, Hany M. Harb, and Hala M. Abd El Kader, "Automatic image segmentation method for breast cancer analysis using thermography," Journal of Engineering Sciences, Vol. 46, No. 1, pp. 12-32, 2018.

[21] M. Dash, and H. Liu, "Feature selection for classification," Intelligent data analysis, vol. 1, no. 3, pp. 131-156, 1997.

[22] MIT Lincoln Laboratory. URL: http://www.ll.mit.edu/IST/ idaval/.

[23] C. Cortes, and V. Vapnik, “Support-vector networks,” Machine learning, vol. 20, no. 3, pp. 273–297, 1995.

[24] A. Laura and A. M. Rouslan , “Support Vector Machines (SVM) as a technique for solvency analysis,” in DIW Wochenbericht, no. 811, pp. 1-18, 2008.

[25] L. Rokach, "Ensemble-based classifiers," Artificial Intelligence Review, vol. 33, pp.1–39, 2010.

[26] M. Woznia, M. Grana, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," Information Fusion, vol. 16, no. 1, pp. 3–17. 2014.

[27] B. L. Happel, and J. M. Murre, "The design and evolution of modular neural network architectures," Neural Networks, vol. 7, no. (6–7), pp. 985–1004, 1994.

[28] P. Robi, "Ensemble Based Systems in Decision Making," IEEE Circuits and Systems Magazine, 2006.

[29] PROENG (2012). Image processing and image analyses applied to mastology. http://visual.ic.uff.br/en/proeng/.