WCCI 2012 IEEE World Congress on Computational Intelligence
June, 10-15, 2012 - Brisbane, Australia IJCNN

Clustered Ensemble Neural Network for Breast Mass Classification in Digital Mammography

Peter Mc Leod
Central Queensland University
Bruce Highway, Rockhampton QLD 4702, Australia
mcleod.ptr@gmail.com

Brijesh Verma
Central Queensland University
Bruce Highway, Rockhampton QLD 4702, Australia
b.verma@cqu.edu.au

***Abstract*— This paper proposes the creation of an ensemble neural network by incorporating a k-means classifier. This technique is designed to improve the classification accuracy of a multi-layer perceptron style network for mass classification of digital mammograms. The proposed technique has been tested on a benchmark database and the results have been contrasted with current research. The experimental results demonstrate that the accuracy of the proposed technique is comparable with existing systems.**

Keywords; clustering, classifier, neural network ensemble, digital mammograms

I. INTRODUCTION

The classification of images in medical diagnosis utilising intelligent systems is one of the greatest challenges facing Computer Aided Diagnosis (CAD) especially with breast cancer. The similarities that benign anomalies have with their malignant counterparts are that the morphological features do not facilitate such a classification presents a real challenge in this area. Much work has progressed over the last forty years with numerous approaches being utilized. Techniques that have been employed include case based systems, rule based classifiers, support vector machines [1], genetic algorithms [2] and neural networks [3, 4, 5]. Neural networks have distinguished themselves with their capacity to learn and apply that ability to unseen cases [6]. However neural networks potentially have their own problems including local minima, network paralysis, numerous training parameters, post hoc validation as well as the requirements to be trained, which can sometimes take a lengthy period. Fundamentally the problem that all these techniques have been attempting to solve is a high false positive rate meaning that their adoption in the clinical world by radiologists has not been as fast as would otherwise be anticipated [7, 8, 9]. Another problem that exists in terms of adoption of CAD is that such systems often provide little in the way of reasoning as to why a diagnostic decision has been reached and are seen as a black box approach that does not aid the understanding of the radiologist or clinician [10].

The gold standard for breast cancer diagnosis has been x-ray mammography which has been successful due to its accuracy and the fact it is a non invasive technique [11]. CAD systems have been shown to have a higher classification accuracy utilising features that have been derived from other techniques such as Fine Needle Aspirate (FNA) however this then becomes an invasive technique requiring another procedure. The utilization of an invasive procedure increases diagnostic costs as well as adding psychological and physical trauma to the patient.

In order to ensure that the proposed technique remains non invasive radiographic morphological features together with patient age and a subtlety value have been utilized.

The remainder of this paper is organized as follows. Section two presents a review of existing CAD systems for digital mammograms. The proposed technique is described in section three with details of the research methodology presented in section four. Experimental results are detailed in section five with an analysis against other techniques outlined in section six. The conclusions and directions for future research are presented in section seven.

II. REVIEW OF THE CURRENT STATE OF CAD

A range of techniques have been utilized by various researchers to address the variable classification accuracy rate in CAD. Styllanos, Mavroforakis, Georgiou, Dimitropoulos and Theodoridis [1] obtained a classification rate of 85.7% with a SVM classifier on 322 images from the miniMias database. They also used a breast asymmetry technique to detect breast cancer. Verma [3] added an additional neuron for benign and malignant classes to the hidden layer of a multilayer perceptron style neural network and achieved a classification rate of 94% on masses from the DDSM. Verma also replaced the traditional gradient descent mechanism of the network with the incorporation of least squares to avoid network paralysis.

Dheeba and Tamil [4] utilized a radial basis function network to achieve a classification rate of 85.2% on 207 anomalies from the MIAS database. They concluded that CAD systems offer faster detection of tumors than that achieved by radiologists. Meanwhile Vazirani, Kala, Shukla, Tiwari [5] combined a back propagation neural network together with a radial basis function to achieve a classification accuracy of 95.75% on a dataset from the University of California Irvine machine learning repository. The two classifiers were integrated using a probabilistic sum function. The network also fed only part of the input feature set to each classifier in order to reduce training times. This classifier was known as a modular neural network with the authors believing that single training algorithm systems result in slow learning, potentially local minima and over fitting of a neural network. Evaluation of their technique reveals that the contribution of the different

978-1-4673-1490-9/12/$31.00 ©2012 Crown

classifiers is determined by a weight that appears to be determined by post hoc evaluation.

Surendiran and Vadivel [12] used Principal Component Analysis (PCA), ANOVA DA, and stepwise ANOVA analysis to determine those features that resulted in the best classification accuracy for mass anomalies on a dataset from the Digital Database of Screening Mammography (DDSM) [13]. Their work achieved a classification rate of 87.3%. Vanovcanoua, Lehotska and Rauova [14] show that CAD systems in digital mammography have a recall rate of 6.9 to 9.5% less than the radiologists. This substantiates that CAD systems are effective for early stage detection, especially where dense breasts are concerned.

Ensembles have been an area of active research [15] for a number of years with numerous studies demonstrating their advantage over single classifier systems [15-17]. Zhang et al. [17] partitioned their mass dataset obtained from the DDSM [13] into four subsets based on patient age and mass shape category. A number of classifiers were then tested and the best performing classifier on each subset was chosen. They used SVM, k nearest neighbor and Decision Tree (DT) classifiers on the subsets and achieved a combined classification accuracy of 72%. This was higher than any individual classifier and better than the highest classification rate of 56% obtained without partitioning the dataset into subsets. Meanwhile Luo and Cheng [18] used a DT that was bagged to gain a classification accuracy of 83.4% on 961 mass anomalies. During their research they reduced the number of input features to four BI-RADS features (down from five) and found that according to feature selection techniques that mass margin was the most important diagnostic feature. Their research also substantiates the efficacy of ensembles.

III. Proposed Clustered Neural Network Ensemble

The proposed approach of a layered clustered ensemble is based on previous ensemble work by Rahman and Verma [19] where it has been shown that using multiple classifiers can improve the classification accuracy over a single classifier [15-19].

The utilization of an ensemble can be seen as a redundancy technique since it is reliant on redundancy through the use of multiple classifiers that are integrated together to provide an output classification. Ensembles however have the advantage of providing an improvement in classification accuracy by utilizing existing classifiers. Ensembles are conceptually simple to understand, as the resultant classification is the aggregate classification of the individual constituent classifiers. This work is different to a number of techniques that have gone before in that the ensemble networks are created by clustering.

In many classification dilema's the data to be classified contains multiple groups that can be identified by clustering. These clustered groups will represent some groupings that are highly representative of a classification and are easy to map the input features to a subsequent classification (atomic). Other groupings however are more heterogeneous in nature and/or will represent a grouping where the mapping between the input features and subsequent output classification are not as straight-forward. These clusters represent non-atomic members.

Once a clustering operation has been performed and clusters have been identified a classifier (in this case a feed forward back propagation neural network) is trained for each cluster grouping. The problem with this approach is that the clusters could be too variable or have too small a training sample thus it does not provide a suitable degree of diversity for the neural network. K-means has been criticized in that the assignment of a pattern (data set) to a cluster can be different based on the seeding mechanism where the number of k-means clusters is different to the actual number of clusters in the data sample. If multiple clustering operations are performed with different seeding points then a pattern could belong to a different cluster each time. When a new clustering operation is performed with a different initial seeding we call this layering and the clusters form a layer. However this characteristic can be utilized in that each layer will have a particular anomaly belonging to only one cluster. This membership can be different from one layer to the next. This means that a classifier can be trained on these non-atomic clusters for each layer and the result of the classifiers fused together by the majority vote algorithm to create our ensemble. Thus our layers provide a means of introducing diversity into our ensemble and making it easier to classify non-atomic patterns as they belong to clusters that represent harder to classify data.

IV. Research Methodology

An overview of the proposed research methodology employed is detailed in Figure 1 followed by a synopsis of each step in the process.

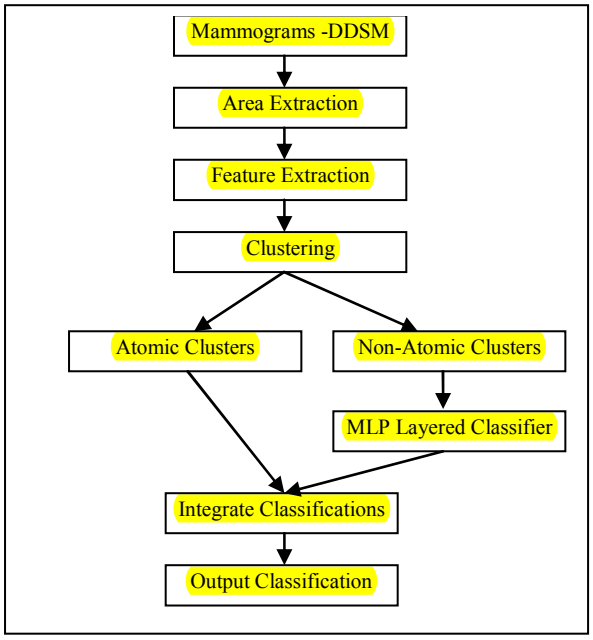

Figure 1. Structure of proposed clustered ensemble classifier.

This research focuses on the classification of breast mass anomalies, as they are harder to classify than micro-calcifications. Each step in the classification process is detailed below:

A. Mammograms

The mass style anomalies utilised in this research have been sourced from the DDSM [13].

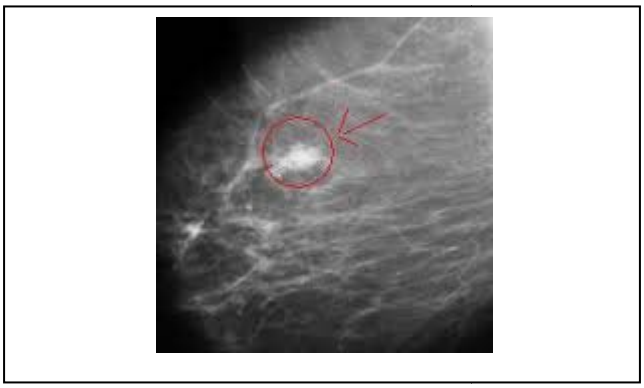

Figure 2. Example of a mass style anomaly in a mammogram.

The DDSM is a publicly available repository on the internet (http://marathon.csee.usf.edu/Mammography/Database.html) that contains a large number of high quality mammographic images as well as pertinent case information (subtlety value, patient age at diagnosis, etc.). Figure 2 shows an example of a mass from a mammogram. This mass is relatively easy to identify however it can be seen that the mammogram is a fairly low contrast medium.

The DDSM [13] has been used as it allows a comparison with other researchers, as it is the premier database for mammographic machine learning.

B. Image Segmentation

Excessive memory and computational expense resulting in long classification times can result when attempting to process full size mammographic images for anomaly classification. The extraction and subsequent processing of the area surrounding an anomaly (called a Region of Interest - ROI) facilitates more rapid and efficient processing. This process does not attempt to classify an anomaly but simply extracts it from the mammographic image.

Anomalies from the DDSM [13] contain chain code information in order to facilitate anomaly and a boundary region to be extracted so that the anomalies can be more readily utilized for research purposes. The chain code is contained in a '.overlay' file and details of chain code usage are available at http://www.marathon.csee.usf.edu/Mammography/DDSM/case_description.html.

C. Feature Extraction

The relationship between features used for diagnostic purposes and the condition or disease in question allows for a diagnosis to be made. The problem in breast cancer diagnosis is that the features used are similar between the benign and malignant conditions. Due to this several features are utilized to provide a better discriminant. In this research the BI-RADS features of density, mass shape, mass margin and abnormality assessment rank are used as they provide good classification accuracy [20]. These are combined with patient age and a subtlety value.

D. Clustering / Classification

The k-means clustering algorithm is used to cluster the anomalies for a labeled dataset (class membership of benign or malignant). This produces both atomic clusters where only one class membership exists and non-atomic clusters where more than one class is represented. A neural network is then trained on the non-atomic clusters and this produces what is termed a layer (a trained classifier based on a particular dataset created through training). This process is repeated for a number of clustering operations where the cluster initialization point is different (seeding) and hence the cluster assignment for a particular anomaly is potentially different. This means that each classifier layer has been trained to recognize a different decision boundary for the non-atomic cluster memberships, thus introducing diversity.

Once the training operation has been completed the network is tested. The ensemble classifier performs an evaluation as to what class an anomaly belongs to in a two-step process. Firstly the cluster membership is determined by examining the anomaly distance to cluster centroids and if membership would belong to an atomic cluster then the class label for that cluster is returned. If membership belongs to a non-atomic cluster then the majority vote is used to return the decision from the layered neural classifier.

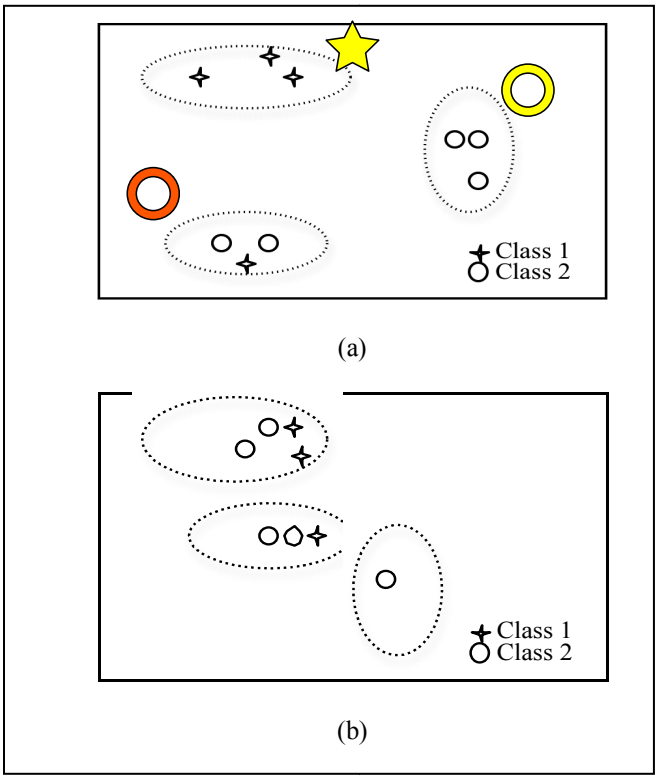

Figure 3. Clustering of the same dataset but at two different cluster layers.

Figure 3 graphically represents the notion of layering that occurs through the seeding of the k-means cluster centres. In the first instance (Figure 3 (a)) three clusters have been defined and the class memberships indicate that two atomic clusters have formed and one non atomic cluster. The non-atomic cluster would be harder to classify and the neural network classifier would be trained on this cluster. The other two clusters are easier to classify and their class label suffices for classifying test patterns. In the second instance (b) the cluster seeding was changed and the layer has again produced three clusters but the cluster memberships are very different to the previous layer. The difference in membership introduces diversity into the neural network training that allows for the development of the ensemble.

The utilization of the neural classifier on the different clusters provides a mechanism where diversity in the output classification of the neural network is achieved due to its exposure to different cluster groupings thus allowing an ensemble to be produced from a partitioning mechanism on the input dataset. The network topology of the neural network classifier consisted of two hidden layers. The input layer had the same number of nodes as the input features for classification purposes (six). The first hidden layer had a variable number of neurons that are recorded in Table II. The output layer consists of two neurons and the hyperbolic tangent sigmoid transfer function tansig is utilized between the layers.

E. Integrate Classifications

The classification outputs from the different neural network topologies (the neural networks trained on each clustered layer) have been integrated by the majority vote as it has been demonstrated to offer a mechanism for improving the aggregate output in comparison to the individual accuracy of each classifier [21-22].

The majority vote works by selecting the output class with the highest number of votes.

V. EXPERIMENTS AND ANALYSIS

Experiments were performed with various parameters being trialed for the classifiers (clusters in the range of 2 to 50, seeds 2 – 21 and hidden nodes 2-111) in order to determine the best configuration that yields the highest classification accuracy. A summary of the results appear in Table I below. The classifier was developed in Matlab version 7.12.0.635 (r2011a). For the purposes of our training and testing the network was trained on 100 masses and was tested on 100 masses. Thus the dataset was comprised of 200 mass anomalies evenly distributed between benign and malignant types.

TABLE I. CLASSIFICATION ACCURACY FOR CLUSTERED ENSEMBLE CLASSIFIER

Clusters	*Seeds*	*Neurons*	*Accuracy*
3	3	25	88
3	6	25	89
4	4	25	90
4	5	25	88
5	5	25	88
8	11	52	90
8	21	20	91
12	9	40	89
14	5	25	88
16	9	30	88
19	2	25	89

In order to determine the efficacy of the ensemble a number of experiments were performed with a neural network of the same configuration as that used in the ensemble to determine the magnitude of the improvement of the ensemble technique. The results of these experiments are tabulated in Table II.

TABLE II. CLASSIFICATION DIFFERENCE FOR NEURAL NETWORK VERSUS ENSEMBLE CLASSIFIER

Neurons	*Ensemble Accuracy %*	*Neural Network Accuracy %*	*Difference %*
10	83	76	7
13	85	80	5
20	91	83	8
25	89	80	9
30	88	79	9
40	89	59	30
52	90	90	0
111	84	79	5

VI. DISCUSSION

The largest improvement in classification accuracy was thirty percentage points, which is a very large improvement in classification accuracy over the baseline neural network. A number of ensemble configurations achieved a classification accuracy of 7-9 percentage points over the equivalent baseline neural network classifier. An ANOVA single factor test of variance was performed on the twenty highest results comparing the ensemble to the equivalent neural network in order to test if the improvement in classification accuracy is statistically significant (H_1). Our null hypothesis is that no statistical variance exists between the ensemble and neural network in terms of classification accuracy (H_0)

TABLE III. SINGLE FACTOR ANOVA SUMMARY

Group	*Count*	*Sum*	*Average*	*Variance*
Ensemble	20	1788	89.4	0.357895
NN	20	1710	85.5	3.947368

TABLE IV. ANOVA ANALYSIS DETAILS

Group Variation						
	SS	*df*	*MS*	*F*	*P-value*	*F Crit*
Between	152.1	1	152.1	70.6577	3.37E-10	4.098172
Within	81.8	38	2.1526			
	239.9	39				

The summary information from the ANOVA analysis (Table III) indicates that the neural network produced classification results that had a significantly higher degree of variability than our ensemble. The variability in classification accuracy has been a major problem with the adoption of CAD by radiologists [8-9]. The improvement in classification accuracy and reduction in variability are key benefits. Examination of the P-value from the ANOVA analysis (Table IV) itself notes that the P-Value is significantly below the standard five percent confidence threshold and thus the null hypothesis can be rejected.

The results that have been obtained need to be compared with accuracies obtained by other researchers. Although this will provide an indication of performance it will not be directly comparable as variations in datasets and features will occur. Also aspects such as configurational complexity (how many parameters need to be optimized), timing and testing times, risk of network paralysis and such are not always disclosed or contrasted as part of the analysis.

The highest achieved classification accuracy that was achieved was 91% that is comparable to existing research. However, it is lower than some of our previous published results using other techniques. Experiments were also performed for the ten highest set of results from our previous experiments using tenfold cross validation to determine how performance would compare. When tenfold cross validation was employed the variability of the neural network classifier was reduced when an ANOVA analysis test of variance was employed (2.89) but was not as good as that achieved by the ensemble (1.39). The highest classification accuracy achieved for the ensemble was 88.5% while the neural network achieved 86.5%. The P-Value however was under our five percent confidence threshold indicating that the null hypothesis could be rejected.

VII. CONCLUSIONS

An ensemble technique was developed and tested on the publicly available benchmark database that demonstrated a significant improvement in classification accuracy over a baseline feed forward network. The proposed technique partitions data into layers based on different seeding points for each cluster in order to generate clusters that have different memberships at each layer. This technique allows a classifier to be trained on each cluster for each layer introducing diversity into the classifiers. The resultant outputs are fused together utilizing the majority vote to create the ensemble clustered network.

The experimental results showed an improvement for the proposed technique (91% highest achieved accuracy) over a neural network (90% highest achieved accuracy). In one instance the performance difference for the ensemble was 89% accuracy for one topology compared with 59% for the neural network. An ANOVA test of variance was performed on the twenty highest results of the two networks and the result was found to be significant at a five percent confidence level. It was noted that classification performance was better at:

- A moderate number of clusters
- Too small a number of clusters and the generalization ability of the ensemble suffered due to the effect of outliers in cluster groupings
- Too high a number of clusters and overtraining of the network occurred as each pattern belonged to its own cluster.

Our future research will undertake more experiments with different network topologies and examine a mechanism to improve classification accuracy through self-tuning capabilities.

REFERENCES

[1] D. Styllanos, M. Mavroforakis, H. Georgiou, N. Dimitropoulos and S. Theodoridis, "A fully automated scheme for mammographic segmentation and classification based on breast density and asymmetry," Computer Methods and Programs in Biomedicine, 2011, April 2011, vol. 102, no. 1, pp. 47-63.

[2] B. Verma and P. Zhang, "A novel neural-genetic algorithm to find the most significant combination of features in digital mammograms," Applied Soft Computing, vol. 7, no. 2, 2007, pp. 612-625.

[3] B. Verma, "Novel network architecture and learning algorithm for the classification of mass abnormalities in digitized mammograms," Artificial Intelligence in Medicine, vol. 42, no. 1, 2008, pp. 67-79.

[4] J. Dheeba and S. Tamil, "Screening Mammogram images for Abnormalities using Radial Basis Function Neural Network," Communication Control and Computing Technologies (ICCCT), 2010 IEEE International Conference, pp. 554-559, doi:10.1109/ICCCT.2010.5670778.

[5] H. Vazirani, R. Kala, A. Shukla and R. Tiwari, "Diagnosis of Breast Cancer by Modular Neural Network," Proceedings of 3rd International Conference on Computer Science and Information Technology (ICCSIT'10), pp. 115-119, Chengdu, doi:10.1109/ICCSIT.2010.5564054.

[6] R. Brem, "Clinical versus research approach to breast cancer detection with CAD: where are we now?", American Journal of Roentology, vol. 188, 2007, pp. 234-235.

[7] M. Abdelaal, H. Sena, M. Farouq and A. Salem, "Using Pattern recognition Approach for Providing Second Opinion of Breast Cancer Diagnosis," 2010, The 7th International Conference on Informatics and Systems (INFOS'10), Cairo, pp. 1-7.

[8] R. Nishikawa, M. Kallergi, "Computer-aided detection, in its present form, is not an effective aid for screening mammography", Medical Physics, vol. 33, 2006, pp. 811-814.

[9] A. Malich, S. Schmidt, D. Fischer, M. Facius, W. Kaiser, "The performance of computer-aided detection when analysing prior

mammograms of newly detected breast cancers with special focus on the time interval from initial imaging to detection", European Journal of Radiology, 2008, doi:10.1016/j.ejrad.2007.11.038.

[10] R. El hamdi and M. N. M. Chtourou, "Breast Cancer Diagnosis Using a Hybrid Evolutionary Neural Network Classifier," in proceedings of 18th Mediterranean Conference on Control and Automation, Marakesh, Morocco, 2010, pp. 1308-1315.

[11] D. Roder, N. Houssami, G. Farshid, G. Gill, P. Downey, K. Beckmann, P. Iosifidis, L. Grieve and L. Williamson, "Population screening and intensity of screening are associated with reduced breast cancer mortality: evidence of efficacy of mammography screening in Australia", Breast Cancer Research and Treatment, vol. 108, 2008, pp. 409-416.

[12] B. Surendiran and A. Vadivel, "Feature Selection using Stepwise ANOVA Discriminant Analysis for Mammogram Mass Classification," International Journal of Recent Trends in Engineering and Technology, vol. 3, no. 2, pp. 55-57, doi:01.IJRTET.03.02.195.

[13] M. Heath, K. Bowyer, D. Kopans, R. Moore and P. Kegelmeyer, "The Digital Database for Screening Mammography," 2001, IWDM-2000, Medical Physics Publishing.

[14] L. Vanovcanova, V. Lehotska and K. Rauova, "Digital Mammography a new trend in carcinoma diagnosis," Bratislava Medical Journal, 2010, vol. 111, no. 9, pp. 510-513.

[15] L. Kuncheva, "Combining pattern classifiers: methods and algorithms", 2004, Wiley-IEEE Press, New York.

[16] J. Rodriguez, L. Kuncheva and C. Alonso, "Rotation forest: a new classifier ensemble method", Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2006, vol. 28, no. 10, pp. 1619-1630.

[17] Y. Zhang, N. Tomuro, J. Furst and D. Raicu, "Building an ensemble system for diagnosing masses in mammograms", International Journal of Computer Assisted Radiology and Surgery (CARS), 2011, June 2011, doi: 10.1007/s11548-011-0628-7.

[18] S. Luo and B. Cheng, "Diagnosing Breast Masses in Digital Mammography Using Feature Selection and Ensemble Methods", 2010, Journal of Medical Systems, May 2010, doi: 10.1007/s10916-010-9518-8.

[19] A. Rahman and B. Verma, "A Novel Layered Clustering based Approach for Generating Ensemble of Classifiers", Neural Networks, IEEE Transactions on, 2011, vol. 22, no. 5 (May 2011), pp. 781-792. doi: 10.1109/TNN.2011.2118765.

[20] M.P. Sampat, A.C. Bovik and M.K. Markey, "Classification of mammographic lesions into BI-RADS Shape Categories using the Beamlet Transform," In proceedigns of the SPIE, Medical Imaging: Image Processing 2005, pp. 16-25.

[21] L. I. Kuncheva, C.J. Whitaker, C.A. Shipp and R.P.W. Duin, "Limits on the Majority Vote Accuracy in Classifier Fusion," Pattern Analysis and Applications, 2003, vol. 6, pp. 22-31.

[22] S. Omar, Z. Saad, M.K. Osman, I. Isa and J.M. Saleh, "Improved Classification Performance for Multiple Multilayer Perceptron (MMLP) Network using Voting Technique," in proceedings of Fourth Asia International Conference on Mathematics/Analytical Modelling and Computer Simulation, 2010, pp. 247-252.