

A Multilayered Ensemble Architecture for the Classification of Masses in Digital Mammograms

Peter Mc Leod and Brijesh Verma

Central Queensland University
Bruce Highway, Rockhampton QLD 4702, Australia
mcleod.ptr@gmail.com, b.verma@cqu.edu.au

Abstract. This paper proposes a technique for the creation of a neural ensemble that introduces diversity through incorporating ten-fold cross validation together with varying the number of neurons in the hidden layer during network training. This technique is utilized to improve the classification accuracy of masses in digital mammograms. The proposed technique has been tested on a widely available benchmark database.

Keywords: Ensemble classifiers, committee of experts, neural networks, digital mammograms.

1 Introduction

Biomedical diagnostic systems are a real world application where pattern recognition systems have great potential due to their ability to repeatedly process large amounts of data without suffering from fatigue [1,2]. Breast cancer can have a high mortality and morbidity rate with survival directly linked to early detection and appropriate treatment. In the United States alone there are expected to be 39,920 deaths in 2012 due to breast cancer [3]. The Gold standard for detection has long been considered to be screening mammography [4] due to its effectiveness and cost efficiency. Recently a shortage of trained radiologists has increased the need for alternative techniques. Researchers have examined this problem for a number of years with numerous solutions being investigated however due to the similarity between benign and malignant tumors as well as processing constraints no suitable automated solution has been developed. Neural networks on the other hand have demonstrated a capacity to adapt to the medical problem domain and have a proven record [5]. The problem however with neural networks like other techniques in this area has been the variable classification accuracy [6]. Ensemble networks while relatively new on the pattern recognition sphere have shown their capabilities in this sphere as the accuracies obtained are higher than for a standard neural network [7]. The creation of an ensemble however relies on conflict between the constituent classifiers in order to produce a more accurate classifier. This conflict is known as diversity and the ability to create diverse classifiers for ensembles is an area of active research [8, 9].

This research is based on the construction of an ensemble network where the constituent classifiers are neural networks that have been created with a different number of neurons in each hidden layer.

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

2 Background

In some cases the complexity of a problem means that it is not easily solvable by a single mechanism (classifier), as the capacity to encapsulate the problem cannot be easily bounded by a single model. In the case of mass breast cancer the similarities between the benign and malignant cases make it hard to make an accurate diagnosis between the two, as the resemblance between the diagnostic features is not clear-cut. The use of many networks combined together can provide a better result than a single classifier [7,10,11] as it is better able to model more complex paradigms. Breast cancer in particular has proven to be an area where pattern recognition has struggled due to the similarities between benign and malignant patterns. Due to the nature of the disease and the practical application of pattern recognition techniques this field has received a lot of interest in recent times [10-16]. Due to the similar characteristics between the benign and malignant classes breast cases have low separability in feature space resulting in classification complexity and variable performance of classifiers.

Although the creation of an ensemble has been shown to have higher accuracy than a single classifier [7,10,11] this is dependent on diversity existing in the constituent classifiers. In a group if everyone follows the lead of a figurehead we have an occurrence which is known as "Group Think" where the decision and hence accuracy is really determined by one individual. Disagreement will lead to potentially a better decision. Diversity is this degree of disagreement and is therefore required to have accuracy higher than that of the best performing constituent classifier(s). The creation of an ensemble however does raise several issues. These issues are:

- Selection of the constituent classifiers (types);
- How many classifiers make up the ensemble;
- Membership criteria (accuracy; diversity; computational speed and or memory requirements / restrictions);
- Desired accuracy;
- Time / memory requirements for solution calculation;
- Nature of the classification problem.

Mechanisms for introducing diversity have been a field of active research [8, 9].

Roselin and Thangavel [13] created a metaheuristic ensemble classifier using ant-miner, as they wanted clarity in relation to the output as many classifiers were seen to be too much of a black box approach. Using anomalies from the MIAS database they achieved a classification accuracy of 83%.

Liu et al. [12] contrasted the performance of a random forest technique (ensemble of classification and regression trees) against a Support Vector Machine (SVM) on

236 regions of interest from the Digital Database of Screening Mammography (DDSM) [17]. The random forest ensemble had few tuning parameters and performed more than twice as fast as SVM however it achieved an accuracy of 79% (Az 0.83) compared to 81% (Az 0.86) for SVM.

Gou et al. [10] developed a Partition Based Network (PBN) boosting technique to handle class imbalanced datasets. They paid more attention to the training of classes that were represented less (in this case the malignant condition) in order to convert the network training into a balanced activity. Experiments were performed on the University of California Irvine (UCI) dataset [18] and achieved a higher accuracy value than for standard network boosting. The PBN network achieved a classification rate of 96.50% compared to a network boost accuracy of 96.25%.

Meena, Arya and Kala [14] achieved a classification accuracy of 96.87% on a UCI dataset [18] by implementing a modular neural network where a Self-Organising Map (SOM) selector selected the neural networks that would classify the breast anomalies.

Their approach involved the training of the SOM selector, as well as the training of the candidate networks. This was to ensure that the best network possible network was chosen for the classification task.

Huang, Monekosso and Wang [15] utilised a clustered ensemble to reduce the variability and improve the accuracy of clustered classifiers, which they called Clustering Ensembles Based on Multi-classifier Fusion (CEMF) that worked by constructing a number of classifiers that were optimal at classifying a subset of data that had been created through partitioning of the data. They had a 10.5% error rate on their breast cancer dataset.

Luo and Cheng [19] used a Decision Tree (DT) that was bagged to gain a classification rate of 83.4% on mass anomalies. They also used a SVM-SMO (Support Vector Machine Sequential Minimal Optimization) from the Waikato Environment for Knowledge Analysis (WEKA) [20] data mining package. They selected mass anomalies from the UCI machine repository database [18]. During their research they utilised feature selection techniques to reduce the number of input features for the classifier from five to four. They employed four BI-RADS® features and found that the mass margin was the most important feature for mass classification. Luo and Cheng concluded that the ensemble that they created was more effective than using a single classifier.

Zhang, Romuro, Furst and Raicu [21] developed an ensemble classifier where they partitioned a mass dataset from the DDSM [17] into four subsets. The partitioning was based on patient age and mass shape category. A number of classifiers were then evaluated and the best performing classifier on each subset was chosen. The classifiers evaluated were SVM, k-nearest neighbor and a DT. The ensemble achieved a classification accuracy of 72%.

3 Proposed Methodology

Devising mechanisms for generating base classifiers that are diverse for the creation of an ensemble is an active area of research for meta-classifier research [8, 22]. Numerous

researchers have utilized techniques on manipulating the training dataset(s) in order to introduce diversity into the ensemble [7,9,22,23]. These techniques have involved training on separate features, training on each classifier on separate datasets as well as other mechanisms. However the number of investigations into using different types of neural networks in ensembles has been low [9] in comparison to manipulation of the training dataset. Neural networks gain their knowledge of the problem domain by being trained on a sufficiently large population sample so that they can then apply that learned knowledge to unseen cases. If the configuration of the classifiers is different the classifier will represent the acquired knowledge in a different way. This is similar to the variation in conceptual understanding that an audience will gain from listening to a lecture. The focus of this research is the demonstration of one mechanism to introduce diversity into the constituent classifiers. Although an explicit diversity inclusion mechanism could have been employed an implicit approach was utilized [9]. Diversity is introduced by varying the number of neurons in the hidden layer. This causes a traversal of feature space that is sufficiently different for each network to result in different weight values that represents the acquired knowledge in the classifier. This introduces diversity into our resultant ensemble [9]. Ten fold cross validation was also utilized as this introduces diversity into the classifier. Partridge and Yates had previously amended the number of neurons to introduce diversity [24] however their research indicated that the impact on diversity and performance was marginal however they only trialed a range of 8 to 12 neurons in the hidden layer. In this research a range from 2 to 150 neurons in the hidden layer has been trialed. This work differs in that the dataset that is utilized also has ten fold cross validation occurring on it, which in itself introduces diversity into the ensemble [9].

The research methodology is represented in Figure 1 below. The process begins with acquiring mammograms. In this research the mammograms have been sourced from a benchmark database called the DDSM [17]. The DDSM provides open access to high quality mammograms that allows for testing of pattern recognition algorithms in order to advance the field of CAD systems for breast cancer diagnosis.

Once the mammograms have been obtained (200 mass type anomalies evenly divided between benign and malignant) the suspicious regions need to be extracted. The process of extracting the anomaly is known as image segmentation and allows a system to more effectively operate as the background noise from normal tissues are removed from the diagnostic or clinical system. Anomalies that are sourced from the DDSM [17] can be extracted by a chain code that exists in order to facilitate the easy extraction of anomalies.

Once the anomalies are extracted it is necessary to obtain the features from the anomalies that are used in feature space to create a mapping to a resultant classification. In this research BI-RADS® are utilised since they have a good diagnostic capability [25-26] and are the same as what a trained radiologist would use to come to a diagnosis. Using the same features helps to increase the utility of this technique for potential clinical usage. The diagnostic features are:

- Density
- Calcification Type / Mass Shape
- Calcification Distribution / Mass Margin

- Abnormality Assessment Rank
- Patient Age
- Subtlety Value

In this research the candidate neural networks that are constructed have six input neurons that correspond directly to the features that are utilised to form a diagnosis.

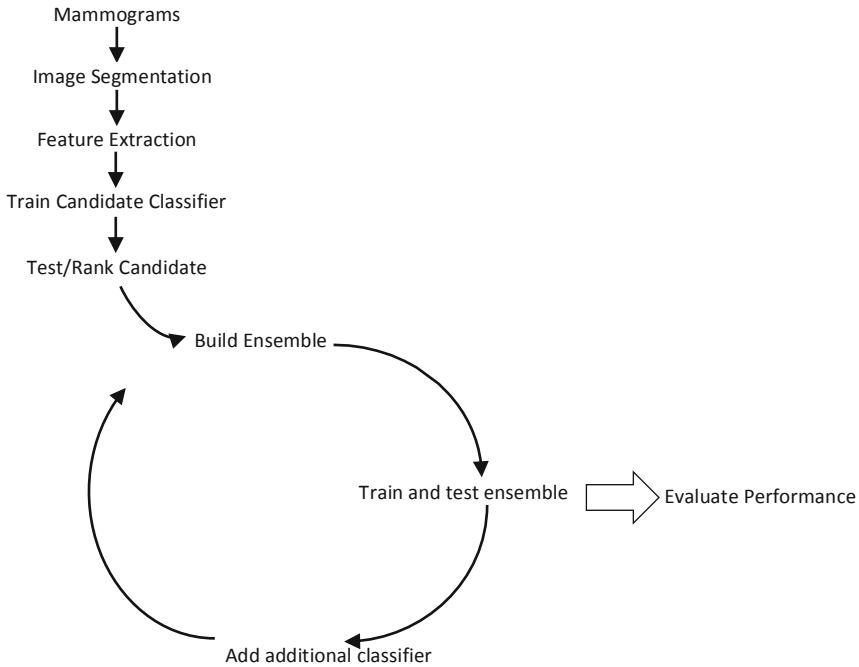


Fig. 1. Proposed research methodology

A number of candidate neural networks were generated for the purpose of this research. These networks were created by varying the number of neurons in the hidden layer from 2 neurons to 150 neurons in order to introduce diversity. The networks were trained and tested using ten-fold cross validation. Once the candidate networks were generated they were ranked based on classification accuracy and then these candidates were utilised to generate our ensemble classifier(s). The ensemble was initially created by using the first three highest performing candidate networks and then was trained and tested (from Table 1 this would be an ensemble that has 24,5 and 15 neurons in each hidden layer). Following this the fourth highest performing candidate classifier was added and then it was trained and tested with the process continuing until a total of the 40 highest performing candidate classifiers had been incorporated into the ensemble and were trained and tested. This was done to evaluate the effect of adding additional classifiers to the ensemble and to attempt to determine the best number of classifiers for the ensemble in terms of achieving a high classification accuracy. The maximum number of constituent classifiers was limited to 40 as

the whole process (load data, train and test candidates, generate and evaluate the ensembles) was to occur in only a few minutes.

In accordance with normal scientific practice ten fold cross validation was utilized for all experiments. The stopping condition for network training was a root mean square (rms) error of 0.001 or (a maximum of 3000 iterations). This ensured that the networks were fully trained for the purposes of our experiments. The other parameters were a learning rate of 0.05 and momentum of 0.7.

Table 1. Selection of candidate neural networks for ensemble inclusion

Accuracy	86	85.5	85.5	85.5	85	85	85	84.5	84.5	...	83
Neurons	24	5	15	32	31	43	50	75	38	...	62
Rank	1	2	3	4	5	6	7	8	9	...	40

From Table 1 the best performing network had 24 neurons, the next 5, followed by 15, then 32 in the hidden layer etc.

The proposed system was implemented in MATLAB 7.12.0. MATLAB was running on a 2.66 Ghz Intel Core i7 Mac Book Pro, with 8 Gig of RAM running Mac OS X 10.6.8. The experiments were carried out on a subset from the DDSM [17] utilising only mass type anomalies. The dataset represented 200 anomalies evenly divided between malignant and benign cases.

4 Results

Accuracy of the individual constituent classifiers ranged from 83% to 86% accuracy (Table 1).

Table 2. Ensemble Performance

# Classifiers	Hidden Neurons	TP	TN	Accuracy (%)
4	24,5,15,32	91	95	93.0
14	13,24,5,15,32,31,43,50,75,38,59,68,79,116	98	98	98.0
18	24,5,15,32,31,43,50,75,38,59,68,79,116,146,14,30,37,95	98	96	97.0
22	24,5,15,32,31,43,50,75,38,59,68,79,116,146,14,30,37,95,103,104,138,140	98	97	97.5
25	24,5,15,32,31,43,50,75,38,59,68,79,116,146,14,30,37,95,103,104,138,140,19,29,33	99	97	98.0
35	24,5,15,32,31,43,50,75,38,59,68,79,116,146,14,30,37,95,103,104,138,140,19,29,33,36,40,46,48,65,74,84,93,127,22	97	96	96.5

The only condition for inclusion in the ensemble networks was accuracy. The ensemble networks that were constructed had from three to forty constituent classifiers. Table 2 provides a summary of the results obtained for the ensemble classifiers.

5 Discussion

The minimum accuracy that the ensemble achieved was 93% while the maximum accuracy was 98%. Numerous classifiers achieved a classification accuracy of 97%. The improvement in classification accuracy supports the claim by previous researchers that ensembles have higher classification accuracy than their constituents providing that the requirement of diversity is met [27]. Ensembles increase accuracy by reducing the variance in prediction errors [1]. In this research a significant improvement in classification accuracy was achieved by varying the number of neurons in the hidden layer as well as utilizing ten-fold cross validation. In order to determine how significant the results of this research are we need to compare the classification accuracy achieved with that of other researchers and also with the best performance achieved for an individual classifier. Table 3 lists the performance of the proposed technique in comparison to current research. The performance achieved by Verma et al. [28] of 93% is on the same dataset and is directly comparable as it is the same dataset as this research.

Table 3. Comparative performance of proposed technique and current research

Researcher / Reference	Technique	Accuracy [%]
Ensemble technique	Proposed ensemble technique	98.00
Base classifier	Neural network	86.00
Verma et al. [28]	Soft clustered neural network	93.00
Liu et al. [11]	SVM	81.00
Liu et al. [11]	Random forest	79.00
Zhang et al. [20]	Partitioned ensemble classifier	72.00

An ANOVA analysis of variance was performed to determine if the improvement in classification accuracy between the neural network and the ensemble was significant. A 5% confidence level was utilized according to standard scientific methodology.

Table 4. ANOVA Summary

Groups	Count	Sum	Average	Variance
NN	38	3195.5	84.09	0.6197
Ensemble	38	3674	96.68	1.3706

Table 5. ANOVA Analysis

Group Variation						
	SS	df	MS	F	P-value	F crit
Between	3012.66118	1	3012.6612	3027.46475	8.8016E-62	3.97
Within	73.6381579	74	0.9951102			
Total	3086.29934	75				

The p-value (Table 5) is significantly below the required 5% confidence level substantiating that the improvement in performance of the ensemble is statistically significant. Of note is the variance in the ANOVA summary (Table 4.) for the ensemble is much higher than that of the neural network. This is to be expected as the classifiers for the neural network were chosen on the basis of being the highest performers from the candidate classifier pool. The ensemble representatives came from a much smaller candidate pool and the variation was therefore higher as all candidates were chosen.

6 Conclusions

Neural networks have been used extensively to solve many classification dilemmas. The problem with such approaches is that a single neural network has not demonstrated the capacity to solve complex problems where the classification dilemma is not straight forward due to similarities between the distinct classes as occurs in breast cancer diagnosis. This has lead to various authors describing neural networks as a monolithic solution [14] and advocating the usage of ensemble and modular networks. In this solution we utilized redundancy through the creation of an ensemble network that introduced diversity by altering the number of neurons in the hidden layer as well as ten-fold cross validation. The results were fused through the majority vote algorithm. The performance of these networks was then contrasted. The results of this study have demonstrated that ensembles perform better than a single constituent classifier. There is no universal classifier that works best on all given problems. It is not easy to determine a suitable classifier for a given problem, but the use of an ensemble allows for combining the best of multiple classifiers in order to improve accuracy and overcome the limitations of a single classifier. The proposed ensemble approach improved the predictive performance for the mammogram dataset. The improvement in classification accuracy was 12% percentage points (98%-87%) over the best performing single neural network classifier. Our future research will involve the automation of the processes through genetic algorithm pruning of network candidates as well as performance and diversity measures to determine a better performing solution. Our research shall also examine the effects of tuning the momentum and learning rate of the neural network and how this impinges on the creation of an ensemble network.

References

1. Oh, S., Lee, M., Zhang, B.: Ensemble Learning with Active Selection for Imbalanced Biomedical Data Classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8(2), 316–325 (2011)
2. Doi, K.: Diagnostic imaging over the last 50 years: research and development in medical imaging science and technology. *Physics in Medicine and Biology* 51(13), R5–R27 (2006)
3. Siegel, R., Naishadham, D., Jemal, A.: CA: A Cancer Journal for Clinicians 62(1) (January/February 2013), doi:10.332/caac.20138
4. Haakinson, D., Stucky, C., Dueck, A., Gray, R., Wasif, N., Apsey, H., Pockaj, B.: A significant number of women present with palpable breast cancer even with a normal mammogram within 1 year. *American Journal of Surgery* 200(6), 712–717 (2010)

5. Brem, R.: Clinical versus Research approach to breast cancer detection with CAD: where are we now? *American Journal of Roentology* 188, 234–235 (2007)
6. Abdelaal, M., Sena, H., Farouq, M., Salem, A.: Using Pattern Recognition Approach for Providing Second Opinion of Breast Cancer Diagnosis. In: *The 7th International Conference on Informatics and Systems (INFOS 2010)*, Cairo, pp. 1–7 (2010)
7. West, D., Mangiameli, P., Rampal, R., West, V.: Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application. *European Journal of Operational Research* 162, 532–551 (2005)
8. Gou, S., Yang, H., Jiao, L., Zhuang, X.: Algorithm of Partition Based Network Boosting for Imbalanced Data Classification. In: *The 2010 International Joint Conference on Neural Networks*, pp. 1–6. IEEE, Barcelona (2010)
9. UCI Machine Learning Database, <http://archive.ics.uci.edu/ml/%3E>
10. Santos, E., Sbourin, R., Maupin, P.: Pareto Analysis for the Selection of Classifier Ensembles. In: *Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation 2008 (GECCO 2008)*, pp. 681–688 (2008), doi:10.1145/1389095.1389229
11. Liu, J., Chen, J., Liu, X., Tang, J.: An Investigate of Mass Diagnosis in Mammogram with Random Forest. In: *Fourth International Workshop on Advanced Computational Intelligence*, pp. 638–641. IEEE, Wuhan (2011)
12. Roselin, R., Thangavel, K.: Classification Ensemble for Mammograms Using Ant-Miner. In: *Second International conference on Computing, Communication and Networking Technologies*, pp. 1–6. IEEE, Karur (2010)
13. Meena, Y., Arya, K., Kala, R.: Classification Using Redundant Mapping in Modular Neural Networks. In: *Second World Congress on Nature and Biologically Inspired Computing*, pp. 554–659. IEEE, Fukuoka (2010)
14. Huang, Y., Monekosso, D., Wang, H.: Clustering Ensemble Based on Multi-Classifer Fusion. In: *2010 IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS)*, pp. 393–397. IEEE, Xiamen (2010)
15. Razavi, A., Gill, H., Åhlfeldt, H., Shahsavar, N.: Predicting metastasis in breast cancer: comparing a decision tree with domain experts. *Journal of Medical Systems* 31, 263–273 (2007)
16. Oza, N., Tumer, K.: Classifier ensembles: Select real-world applications. *Journal of Information Fusion* 9, 4–20 (2008), doi:10.1016/j.inffus.2007.07.002
17. Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, P.: *The Digital Database for Screening Mammography*. In: *IWDM 2000. Medical Physics Publishing* (2001)
18. Luo, S., Cheng, B.: Diagnosing Breast Masses in Digital Mammography Using Feature Selection and Ensemble Methods. *Journal of Medical Systems* 36(2), 569–577 (2010)
19. Witten, I.H., Frank, E.: *Data Mining: Practical Machine Learning Tools with Java Implementations*. Morgan Kaufmann, San Francisco (2000)
20. Zhang, Y., Tomuro, N., Furst, J., Raicu, D.: Building an Ensemble System for Diagnosing Masses in Mammograms. *International Journal of Computer Assisted Radiology and Surgery (CARS)*, 323–329 (2011)
21. Panda, M., Patra, M.: Ensemble of Classifiers for Detecting Network Intrusion. In: *International Conference on Advances in Computing, Communication and Control (ICAC3 2009)*, Mumbai, India, January 23–24, pp. 510–515 (2009)
22. McLeod, P., Verma, B.: Clustered Ensemble Neural Network for Breast Mass Classification in Digital Mammography. In: *IEEE World Congress on Computational Intelligence, WCCI 2012, June 10–15, Brisbane, Australia*, pp. 1266–1271 (2012)
23. Brown, G., Wyatt, J., Harris, R., Yao, X.: Diversity creation methods: a survey and categorization. *Journal of Information Fusion* 6, 5–20 (2005), doi:10.1016/j.inffus.2004.04.004

24. Partridge, D., Yates, W.: Engineering multiversion neural-net systems. *Neural Computing* 8(4), 869–893 (1996), doi:10.1162/neco.1996.8.4.869
25. Alto, H., Rangayyan, R., Desautels, J.: Content-based retrieval and analysis of mammographic masses. *Journal of Electronic Imaging* 14(2), 023016, 1–17 (2005)
26. Andre, T., Rangayyan, R.: Classification of breast masses in mammograms using neural networks with shape, edge sharpness, and texture features. *Journal of Electronic Imaging* 15, 013019 (2005)
27. Kuncheva, L.: *Combining pattern classifiers: methods and algorithms*. Wiley-IEEE Press, New York (2004)
28. Verma, B., McLeod, P., Klevansky, A.: Classification of benign and malignant patterns in digital mammograms for the diagnosis of breast cancer. *Expert Systems with Applications* 37(4), 3344–3351 (2010)