Expert Systems With Applications 46 (2016) 139–144

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Breast cancer classification using deep belief networks

Ahmed M. Abdel-Zaher*, Ayman M. Eldeib

Department of Systems and Biomedical Engineering, Cairo University, Giza, Egypt

ARTICLE INFO

Keywords:
Breast cancer diagnosis
CAD
Classification
Deep learning based classifier
Pattern recognition

ABSTRACT

Over the last decade, the ever increasing world-wide demand for early detection of breast cancer at many screening sites and hospitals has resulted in the need of new research avenues. According to the World Health Organization (WHO), an early detection of cancer greatly increases the chances of taking the right decision on a successful treatment plan. The Computer-Aided Diagnosis (CAD) systems are applied widely in the detection and differential diagnosis of many different kinds of abnormalities. Therefore, improving the accuracy of a CAD system has become one of the major research areas. In this paper, a CAD scheme for detection of breast cancer has been developed using deep belief network unsupervised path followed by back propagation supervised path. The construction is back-propagation neural network with Liebenberg Marquardt learning function while weights are initialized from the deep belief network path (DBN-NN). Our technique was tested on the Wisconsin Breast Cancer Dataset (WBCD). The classifier complex gives an accuracy of 99.68% indicating promising results over previously-published studies. The proposed system provides an effective classification model for breast cancer. In addition, we examined the architecture at several train-test partitions.

1. Introduction

Breast cancer is the most common cancers among women with nearly 1.7 million new cases diagnosed in 2012 (Centers for disease control and prevention, cancer prevention control, 2014) (World cancer research fund, 2014). Breast cancer represents 18.3% of the total cancer cases in Egypt. A percentage of 37.3% of breast cancer could be fully healed especially in case of early detection (Salama, Abdelhalim, & Zeid, 2012). In Egypt and Arab countries, the breast cancer targets women in the age of 30 and represents 42 cases per 100 thousand of the population (Salama et al., 2012).

An accurate classifier is the most important component of any CAD scheme that is developed to assist medical professionals in early detecting mammographic lesions. CAD systems are designed to support radiologists in the process of visually screening mammograms to avoid miss-diagnosis because of fatigue, eyestrain, or lack of experience. The use of an accurate CAD system for early detection could definitely save precious lives. In this study, back propagation neural network initialized by weights from a trained deep belief network with similar architecture (DBN-NN) was used to diagnose the breast cancer. Our data source is the Wisconsin Breast Cancer Dataset (WBCD) taken from the University of California at Irvine (UCI) machine learning repository (Wisconsin breast cancer dataset (WBCD) (original), 2014).

2. Background

A variety of classification techniques were developed for breast cancer CAD systems. The accuracy of many of them was evaluated using the dataset taken from the UCI machine-learning repository. For example, Goodman, Boggess, and Watkins, tried different methods that produced the following accuracies: optimized learning vector quantization (optimized-LVQ) method's performance was 96.7%, big-LVQ method reached 96.8%, and the last method, they proposed AIRS, which depending on the artificial immune system, obtained 97.2% of classification accuracy (Goodman, Boggess, & Watkins, 2002).

Quinlan reached 94.74% classification accuracy using 10-fold cross validation with C4.5 decision tree method (Quinlan, 1996). Abonyi and Szeifert used Supervised Fuzzy Clustering (SFC) technique and obtained 95.57% accuracy (Abonyi & Szeifert, 2003). Salama et al. (2012) performed an experiment on WBC dataset and results showed that the fusion between MLP and J48 classifiers with feature selection (PCA) is superior to the other classifiers.

Hamilton, Shan, and Cercone (1996) with RIAC method obtained 96% accuracy. Polat and Günes (2007) examined the robustness of the least square Support Vector Machine (SVM) by using classification accuracy, analysis of sensitivity and specificity, k-fold cross-validation method, and confusion matrix. They obtained classification accuracy of 98.53%.

Nauck and Kruse (1999) obtained 95.06% with neuro-fuzzy techniques. Pauline and Santhakumaran used Feed Forward Artificial Neural Networks and back propagation algorithm to train the network (Pauline, 2011).The performance of the network is evaluated

* Corresponding author. Tel.: +20 1114488419.
E-mail addresses: ahmedallah.m.s@gmail.com (A.M. Abdel-Zaher), eldeib@ieee.org (A.M. Eldeib).

http://dx.doi.org/10.1016/j.eswa.2015.10.015
0957-4174/

using Wisconsin breast cancer dataset for various training algorithms. The highest accuracy of 99.28% is achieved when using Levenberg Marquardt algorithm.

The accuracy obtained by Pena-Reyes & Sipper (1999) was 97.36% using fuzzy-GA method. Akay (2009) combined SVM with feature selection obtaining highest classification accuracy (99.51%) for SVM model that contains five features. Moreover, Setiono (2000) was reached 98.1% using the Neuro-rule method. Übeyli (2007) used SVM and obtain 99.54% accuracy at 37% train and 63% test partition.

Mert, Kılıç, Bilgili, and Akan (2015)., explored features reduction properties of independent component analysis (ICA) on breast cancer decision support system. They proofed that a one-dimensional features vector obtained from (ICA) causes Radial Bases Function Neural Network (RBFNN) classifier to be more distinguishing with the increased accuracy from 87.17% to 90.49%.

Nahato, Nehemiah, and Kannan (2015), used a rough set indiscernibility relation method with back propagation neural network (RS-BPNN). This work has two stages. The first stage handles missing values to obtain a smooth data set and to select appropriate attributes from the clinical dataset by indiscernibility relation method. The second stage is classification using back propagation neural network. The accuracy obtained from the proposed method was 98.6% on breast cancer dataset.

Dheeba, Singh, and Selvi (2014), investigated a new classification approach for detection of breast abnormalities in digital mammograms using Particle Swarm Optimized Wavelet Neural Network (PSOWNN). The proposed abnormality detection algorithm is based on extracting Laws texture energy measures from mammograms and classifying the suspicious regions by applying a pattern classifier. They achieved 93.671%, 92.105% and 94.167% for accuracy, specificity, and sensitivity, respectively.

In our study, we applied deep belief network (DBN) in an unsupervised phase to learn input features statistics of the original WBCD dataset. Then, we transferred the obtained network weight matrix of DBN to back propagation neural network with similar architecture to start the supervised phase. In supervised phase, we tested both conjugate gradient and Levenberg-Marquardt algorithm for learning back propagation neural network.

3. From back propagation (BP) to deep belief network (DBN)

In 1985, the second-generation neural networks with back propagation algorithm have emerged. However, the learning algorithm struggle to adjust network weights so that output neurons state y represent the learning example t. A common method for measuring the discrepancy between the expected output t and the actual output y is using the squared error measure:

$$E = (t - y)^2 \tag{1}$$

The change in weight, which is added to the old weight, is equal to the product of the learning rate and the gradient of the error function, multiplied by −1:

$$\Delta w_{ij} = -\frac{\partial E}{\partial w_{ij}} \tag{2}$$

where almost all data is unlabeled. However, back propagation neural network requires a labeled training data. Therefore, the biggest issue with back propagation NN appears as its possibility to get stuck in poor local optima and the learning time is huge with multiple hidden layers.

In 1963, Vapnik et al. invented the original support vector machine (SVM) algorithm. Boser, Guyon, and Vapnik (1992) suggested a way to create nonlinear classifiers by applying the kernel trick to maximum-margin hyperplanes. In classification task, the weight of each feature

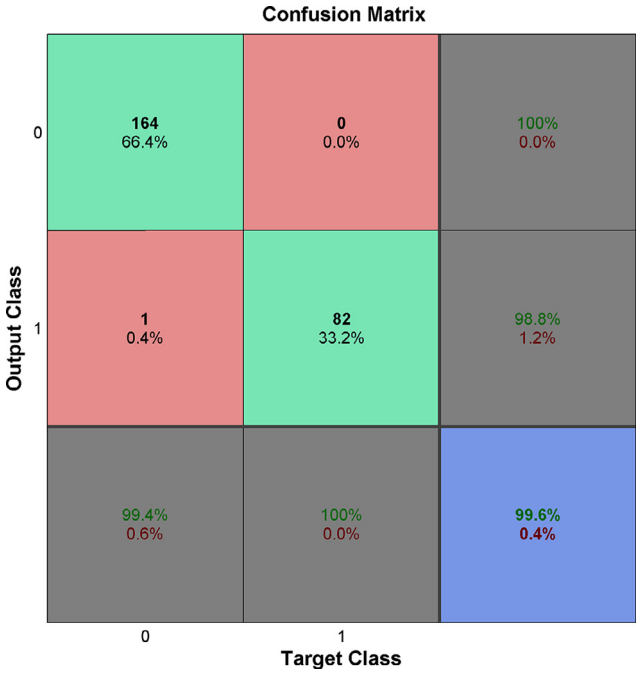

Fig. 1. Confusion matrix of DBN-NN.

is computed by optimization technique. In non-linear classification, SVMs can efficiently perform the task using what is called the kernel trick by mapping their inputs. The non-linear classification task converted to linear classification problem in high-dimensional feature spaces. The biggest limitation of SVM approach lies in choice of the kernel. In practice, the most serious problem with SVMs is the high algorithmic complexity and extensive memory requirements of the required quadratic programming in large-scale tasks (Suykens, Horvath, Basu, Micchelli, & Vandewalle, 2003).

In recent years, the attention has shifted to deep learning. Deep learning is a set of algorithms in machine learning that attempts to model high-level abstractions in data by using model architectures composed of multiple non-linear transformations (Bengio, Courville, & Vincent, 2013; Schmidhuber, 2014). Restricted Boltzmann Machine (RBM) is a generative stochastic artificial neural network that can learn a probability distribution over its set of inputs. On the other hand, Deep Belief Network (DBN) is a generative graphical model, or alternatively a type of deep neural network, composed of multiple layers of latent variables ("hidden units"), with connections between the layers but not between units within each layer (Hinton, 2009b).

From Hinton's perspective, the DBN can be viewed as a composition of simple learning modules each of which is a restricted type of RBM that contains a layer of visible units. This layer represents the data. Another layer of hidden units represents features that capture higher-order correlations in the data. The two layers are connected by a matrix of symmetrically weighted connections (W) and there are no connections within a layer (Hinton, 2009b).

The key idea behind DBN is its weight (w), learned by a RBM define both $p(v|h, w)$and the prior distribution over hidden vectors $p(h|w)$ (Hinton, 2009b). The probability of generating a visible vector, can be written as

$$p(v) = \sum_{h} (p(h|w)\ p(v|h, w)) \tag{3}$$

As the learning of DBN is a computational intensive task, Hinton showed that RBMs could be stacked and trained in a greedy manner to form the DBN (Hinton, Osindero, & Teh, 2006). He introduced a fast algorithm for learning DBN. The weight update between visible v and

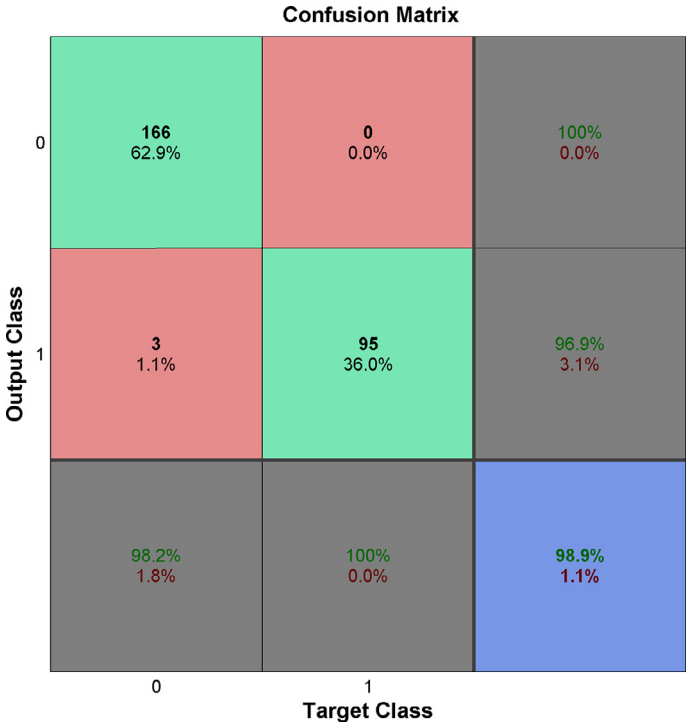

Fig. 2. Confusion matrix of RIW-BPNN.

hidden h unites simply as:

$$\Delta w_{ij} = \varepsilon(\langle v_i.h_j\rangle^0 - \langle v_i.h_j\rangle^1) \quad (4)$$

where 0 and 1 in the above equation designate to the network data and reconstruction state, respectively.

DBN is competitive for five reasons: DBN can be fine-tuned as neural networks, DBN has many non-linear hidden layers, DBN is generatively pre-trained beside it can act as non-linear dimensionality reduction for input features vector, and finally the network teacher is another sensory input. In addition, the real performance of DBN encourages using DBN. For example, in pattern recognition application, Hinton reported that the generalization performance of the DBN is 1.25% errors on the 10000 digits of the MNIST handwritten digits database (Hinton, 2009a). The DBN's performance beats the 1.5% error achieved by the best back propagation nets (Hinton et al., 2006). Platt obtained 1.6% using back propagation while K-Nearest Neighbor produced 3.3% classification error (Lecun, LeCun, Bottou, Bengio, & Haffner, 2001). It is still better than the 1.4% errors reported by Decoste and Schölkopf (2002) for SVM on the same task.

In addition, Hinton discussed the possibility of unsupervised training DBN following by back propagation pass. This will give a better accuracy if we had good data priors (Hinton, 2009a). In 2010, an experiment performed by Erhan et al. (2010) suggested that unsupervised pre-training in prior to supervised learning tasks guides the learning towards basins of attraction of minima that support better generalization from the training dataset. The evidence from these results supported a regularization explanation for the effect of pre-training.

4. Experiment conditions and methodology

We used Matlab 2014a and Palm DBN implementation (Palm, 2012). After the deep belief network fully trained, we transferred its weights matrix to native Matlab back propagation neural network with similar architecture, i.e. same number of input-hidden-output neurons, and then we performed several supervised back-propagation paths. We applied this approach on Wisconsin breast cancer original database with nine features and two classes (benign, malignant). The accepted samples are 690 from 699. Nine samples were rejected for incomplete features. We reduced the used samples down to 683 entries to compare our results easier with others, e.g. Akay (2009) used 683 samples.

The Sampling is repeated randomly Sub-sampling validation of (train+validate) quanta relative to test quanta at different (train+validate) to test partitions, which varied from (0.5–99.5%) to (80–20%) while train–validate partition is fixed at 70–30%. Our methodology is to calculate misclassified sample percentage from the confusion matrix of Randomly Initialized Weight Back-Propagation Neural Network (RIW-BPNN) side by side with back propagation neural network initialized by weights obtained from a trained deep belief network with similar architecture (DBN-NN) at different (train+validate) to test partitions.

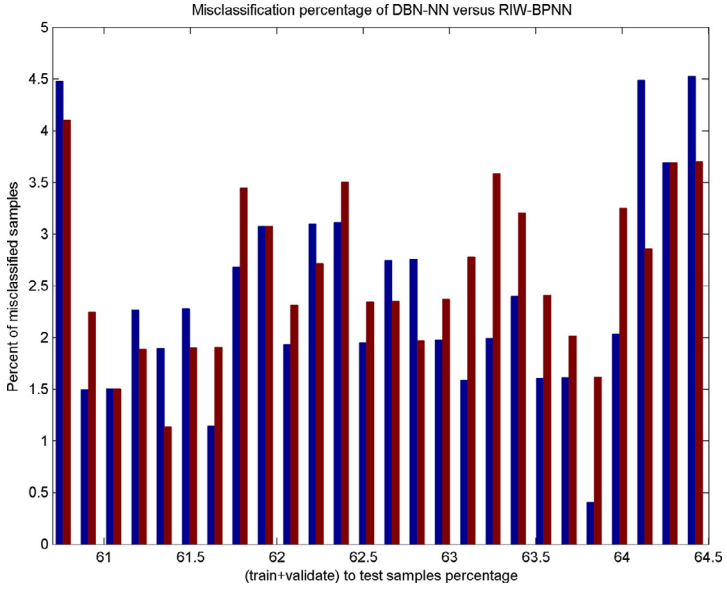

Fig. 3. Dual vertical bar represents misclassified sample percentage. Therefore, shorter bar is the better. The first bar (left/blue) for DPN-NN and the second (right/red) is for RIW-BPNN (conjugate gradient case). The horizontal axis concentrate on percentage from (60.5–39.5%) to (64.5–35.5%) of (train+validate) to test partitions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

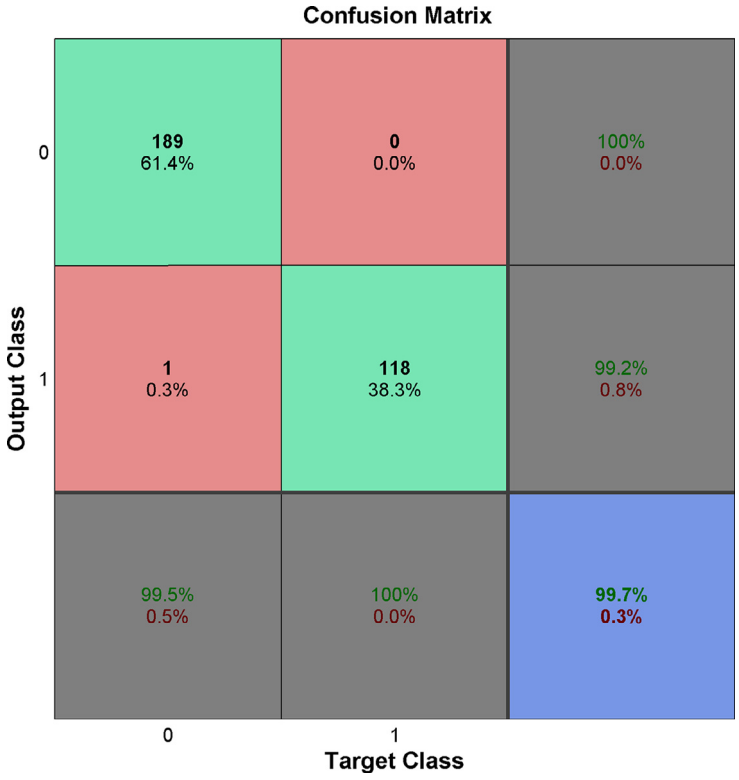

Fig. 4. Confusion matrix of DBN-NN.

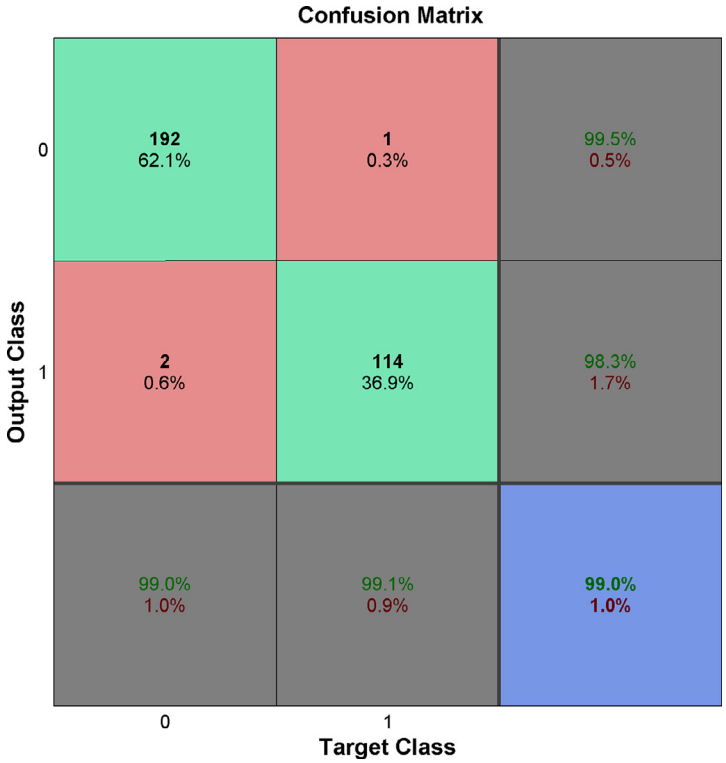

Fig. 5. Confusion matrix of RIW-BPNN.

The RIW-BPNN and DBN-NN architecture is nine inputs – four hidden – two hidden - one output. To increase classifier performance for both architectures, we test conjugate gradient back propagation and Levenberg–Marquardt in neural network learning phase. The experiment conducted by Pauline and Santhakumaran indicating Levenberg–Marquardt learning algorithm gives better classifier accuracy when used with back-propagation neural network (Paulin, 2011).

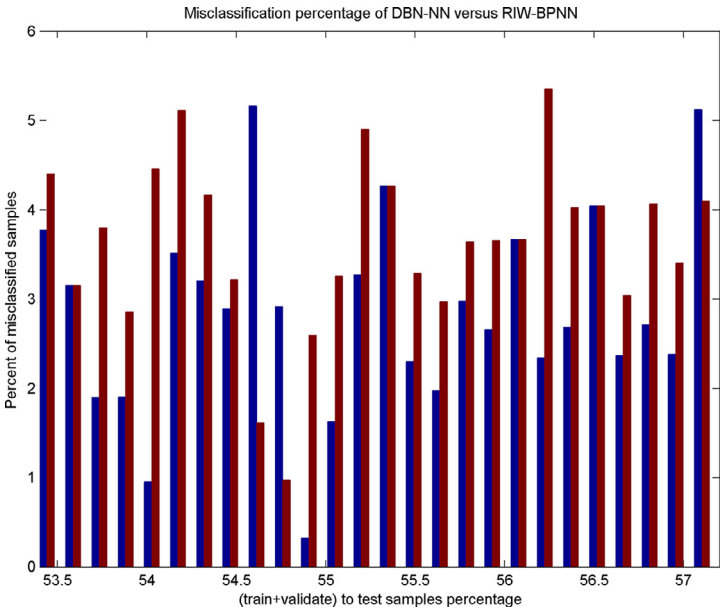

Fig. 6. Dual vertical bar represents misclassified sample percentage. Therefore, shorter bar is the better. The first bar (left/blue) for DPN-NN and the second (right/red) is for RIW-BPNN (Levenberg-Marquardt case). The horizontal axis concentrate on percentage from (53.5–46.5%) to (57.2–42.8%) of (train+validate) to test partitions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Summary table of classifiers performance.

Classifiers	Classifier performance			
	Accuracy (%)	Sensitivity (%)	Specificity (%)	Train+validate to test partition (%)
Ls-SVM, Akay (2009) paper Wisconsin 683 entries	99.51	100	97.91	80–20%
Übeyli (2007) - SVM	99.54	–	–	37–63%
Dheeba et al. (2014)) - PSOWNN	93.67	94.17	92.11	–
Mert et al. (2015)) - ICA-RBFNN	90.49	–	–	–
Nahato et al. (2015)) - RS-BPNN	98.6	–	–	–
Our tested - RIW-BPNN Wisconsin - 683 entries conjugate gradient back propagation	98.86	100	98.22	61.35–38.65%.
Our tested - DBN-NN Wisconsin 683 entries conjugate gradient back propagation	99.59	100	99.39	63.84–36.16%
Our tested - RIW-BPNN Wisconsin - 683 entries Levenberg–Marquardt	99.03	99.13	98.97	54.76–45.24%
Our tested - DBN-NN Wisconsin 683 entries Levenberg–Marquardt	99.68	100	99.47	54.9–45.1%

5. Results

5.1. Scaled conjugate gradient back propagation

Figs. 1 and 2 show the confusion matrix obtained from experiment as (train+validate) to test partitions varied from (0.5–99.5%) to (80–20%), while train–validate partition is fixed at 70–30%. The results show that the best classifier accuracy was 99.59% for DBN-NN complex at (train+validate) to test partition equals to 63.84–36.16%. In comparison, the best accuracy of RIW-BPNN was 98.86% reached at ((train+validate) to test partition equals to 61.35–38.65%.

At the best accuracy of DBN-NN, the total test samples ((1−0.63836) * 683)) =247 samples and

True positive (TP) = 82, True negative (TN) = 164,
False Positive (FP) = 1, False negative (FN) = 0
Sensitivity = 100 * TP/(TP+FN) = 100%
Specificity = 100 * TN/(TN+FP) = 100 * 164/165 = 99.39%

At the best accuracy of RIW-BPNN, the total test samples ((1−0.6134) * 683)) = 264 samples and

True positive (TP) = 95, True negative (TN) = 166,
False Positive (FP) = 3, False negative (FN) = 0
Sensitivity = 100 * TP/(TP+FN) = 100%
Specificity = 100 * TN/(TN+FP) = 100 * 166/169 = 98.22%

Fig. 3 demonstrates part of the sample partition domain and shows the accuracy reach 99.6% for DBN-NN complex at (train+validate) to test partition equals to 63.84–36.16%. In comparison with RIW-BPNN, best accuracy 98.86% reached at (train+validate) to test partition equals to 61.35–38.65%.

5.2. Levenberg–Marquardt

Figs. 4 and 5 show the confusion matrix obtained from experiment at several (train+validate) to test partitions, which varied from (0.5–99.5%) to (80–20%), while train - validate partition is fixed at 70–30%.

The best classifier accuracy of DBN-NN complex was 99.68% obtained at (train+validate) to test partition equals to 54.9–45.1% with misclassified sample percentage reached 0.32%. In comparison, RIW-BPNN best accuracy reached 99.03% with misclassified sample percentage 0.97% obtained at (train+validate) to test partition equals to 54.76–45.24%.

At the best accuracy of DBN-NN, the total test samples ((1−0.549) * 683)) = 308 samples and

True positive (TP) = 118, True negative (TN) = 189,
False Positive (FP) = 1, False negative (FN) = 0
Sensitivity = 100 * TP/(TP+FN) = 100%
Specificity = 100 * TN/(TN+FP) = 100 * 189/190 = 99.47%

At the best accuracy of RIW-BPNN, the total test samples were ((1−0. 548) * 683)) = 309 samples and

TP = 114, TN = 192, FP = 2, FN = 1
Sensitivity = 100 * TP/(TP+FN) = 99.130%
Specificity = 100 * TN/(TN+FP) = 100 * 192/194 = 98.969%

From Fig. 6, we can observe that the accuracy of DPN-NN complex reached 99.68% obtained at (train+validate) to test partition equals 54.9–45.1%.

Table 1 summarizes different classifier performance including our technique.

6. Conclusion

In this research, we presented an automatic diagnosis system for detecting breast cancer based on DBN unsupervised pre-training phase followed by a supervised back propagation neural network phase (DBN-NN). The pre-trained back propagation neural network with unsupervised phase DBN achieves higher classification accuracy in comparison to a classifier with just one supervised phase. The rationale behind this enhancement could be that the learning of input statistics from input feature space by DBN phase initializes back propagation neural network to search objective function near a good local optima in supervised learning phase.

From our experiment at the specified network architecture, DBN-NN complex accuracy outperforms RIW-BPNN when back propagation neural network uses conjugate gradient algorithm for learning. DBN-NN still outperforms RIW-BPNN when we use Levenberg-Marquardt for training in back propagation neural network phase. The enhancement of overall neural network accuracy is reaching 99.68% with 100% sensitivity and 99.47% specificity in breast cancer case. Results show classifier performance improvements over previous studies.

Although Hinton developed fast algorithm for training DBN, DBN's learning process still require substantial computational effort on legacy hardware. Therefore, the main limitation/challenge of our approach is to build a CAD scheme based on DBN using commercial hardware to assist medical professionals in the early detection process of breast abnormality.

Future research effort should be allocated for evaluating such classifier complex for auto diagnosis of other abnormalities such as epilepsy based on EEG dataset, cardiac arrhythmia, and diabetic retinopathy (DR). Further, the presence of general-purpose computing on graphics processing units (GPGPU) and the distribution nature of DBN may also encourage the developments of efficient parallel algorithm for learning such classifier.

References

Abonyi, J., & Szeifert, F. (2003). Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recognition Letters, 24*, 2195–2207.

Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems With Applications, 36*, 3240–3247.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*, 1798–1828.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th annual ACM workshop on computational learning theory* (pp. 144–152). ACM Press.

Centers for disease control and prevention (2014) <http://www.cdc.gov/cancer/dcpc/data/women.htm>. *Cancer prevention control.* Accessed 03.09.14.

Decoste, D., & Schölkopf, B. (2002). Training Invariant Support Vector Machines. *Machine Learning, 46*, 161–190.

Dheeba, J., Singh, N. A., & Selvi, S. T. (2014). Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach. *Journal of Biomedical Informatics, 49*, 45–52.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., & Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research, 11*, 625–660.

Goodman, D. E., Boggess, L. C., & Watkins, a. B. (2002). Artificial immune system classification of multiple-class problems. In *Proceedings of the intelligent engineering systems* (pp. 179–184). ASME.

Hamilton, H.J., Shan, N., & Cercone, N. (1996). RIAC: A Rule Induction Algorithm Based on Approximate Classification.

Hinton, G.E.(2009a).*Deep belief nets*. <http://www.cs.toronto.edu/~hinton/nipstutorial/nipstut3.pdf> Accessed 03.09.14.

Hinton, G.E.(2009b). *Deep belief networks*. <http://www.scholarpedia.org/article/Deep_belief_networks> Accessed 03.09. 14.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18*, 1527–1554.

Lecun, Y., LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (2001). Gradient-based learning applied to document recognition. *Intelligent Signal Processing* (pp. 306–351). IEEE Press.

Mert, A., Kılıç, N. Z., Bilgili, E., & Akan, A. (2015). Breast cancer detection with reduced feature set. *Computational and Mathematical Methods in Medicine*, 1–11.

Nahato, K. B., Nehemiah, H. K., & Kannan, A. (2015). Knowledge mining from clinical datasets using rough sets and backpropagation neural network. *Computational Mathematical Methods in Medicine*, 1–13.

Nauck, D., & Kruse, R. (1999). Obtaining interpretable fuzzy classification rules from medical data. *Artificial Intelligence in Medicine, 16*, 149–169.

Palm, R.B.(2012). *Deep learning toolbox*. <https://github.com/rasmusbergpalm/DeepLearnToolbox> Accessed 03.09.14.

Paulin, F. (2011). Classification of breast cancer by comparing backpropagation training algorithm. *Intenational Journal on Computer Science and Engineering, 3*, 327–332.

Pena-Reyes, C. A., & Sipper, M. (1999). A fuzzy-genetic approach to breast cancer diagnosis. *Artificial Intelligence in Medicine, 17*, 131–155.

Polat, K., & Günes, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing, 17*, 694–701.

Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research, 4*, 77–90.

Salama, G. I., Abdelhalim, M. B., & Zeid, M. A. (2012). Breast Cancer diagnosis on three different datasets using multi-classifiers. *International Journal of Computer and Information Technology, 1*, 36–43.

Schmidhuber, J. (2014). Deep learning in neural networks: An overview. *Neural Networks, 61C*, 85–117.

Setiono, R. (2000). Generating concise and accurate classification rules for breast cancer diagnosis. *Artificial Intelligence in Medicine, 18*, 205–219.

Suykens, J. A. K., Horvath, G., Basu, S., Micchelli, C., & Vandewalle, J. (2003), *Advances in Learning Theory: Vol. 190* p. 392. IOS Press.

Übeyli, E. D. (2007). Implementing automated diagnostic systems for breast cancer detection. *Expert Systems With Applications, 33*, 1054–1062.

Wisconsin breast cancer dataset (WBCD) (2014). (original). Accessed 03.09.14 <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29>.

World cancer research fund. (2014). <http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics> Accessed 03.09.14.