RESEARCH ARTICLE

Evolutionary Wavelet Neural Network ensembles for breast cancer and Parkinson's disease prediction

Maryam Mahsal Khan*, Alexandre Mendes, Stephan K. Chalup

Interdisciplinary Machine Learning Research Group (IMLRG), School of Electrical Engineering and Computing, The University of Newcastle, Callaghan, NSW 2308, Australia

* maryammahsal.khan@uon.edu.au

Abstract

Wavelet Neural Networks are a combination of neural networks and wavelets and have been mostly used in the area of time-series prediction and control. Recently, Evolutionary Wavelet Neural Networks have been employed to develop cancer prediction models. The present study proposes to use ensembles of Evolutionary Wavelet Neural Networks. The search for a high quality ensemble is directed by a fitness function that incorporates the accuracy of the classifiers both independently and as part of the ensemble itself. The ensemble approach is tested on three publicly available biomedical benchmark datasets, one on Breast Cancer and two on Parkinson's disease, using a 10-fold cross-validation strategy. Our experimental results show that, for the first dataset, the performance was similar to previous studies reported in literature. On the second dataset, the Evolutionary Wavelet Neural Network ensembles performed better than all previous methods. The third dataset is relatively new and this study is the first to report benchmark results.

Introduction

Breast cancer is the second leading cause of cancer-related deaths in Australian women [1], while Parkinson's disease is the second most common neurological condition in Australia [2]. The identification and assessment process for both diseases is multi-staged, that is tedious, time-consuming, and challenging where data needs to be manually labeled. Such assessments might also lead to misdiagnosis. In medical practice, in order to reduce the risk of misdiagnosis, opinions from multiple doctors (or specialist doctors) are taken into account. A similar approach is used in the computational intelligence domain, where performance of prediction models (or specialist models) is improved by combining multiple models, thus creating an ensemble of classifiers [3].

Ensemble classifiers and their use have been an active area of research for the past two decades, with Bagging [3] and Boosting [4] being two popular techniques, particularly in the field of applied statistics, pattern recognition and machine learning [5–7]. Many of the prediction models have been improved by using ensembles of support vector machines [8, 9], latent

class analysis (LCA) [10], artificial neural networks [11], k-nearest neighbour [12], and even combinations of these classifiers [6].

Wavelet Neural Networks (WNN) are complex machine learning algorithms that use wavelet analysis and neural networks to generate prediction and control models. WNNs have been applied before in several areas, including time-series prediction and control [13, 14]. Evolutionary Wavelet Neural Networks (EWNN) are a recently proposed method for training WNNs and have been used to generate models for breast cancer and Parkinson's disease classification [15]. However, there have been no studies on the prediction performance of an ensemble of EWNN classifiers, yet.

The motivation of this research is to evaluate the performance of EWNN and ensembles of EWNNs (EWNN-e) and compare them with other ensemble techniques used on the same data reported in literature. The findings of this paper aim to provide future researchers an alternative and effective model to compare with. Moreover, this study also investigates a newly published Parkinson's disease dataset with multiple speech recordings.

The paper is organized as follows. Background provides an overview to Wavelet Neural Networks and its structure, EWNN and its response to a two-spiral task, related work on pruning ensembles, and description of some of the performance measures used in our study. The biomedical datasets, proposed mechanism and the experimental setup are described in the Experimental Methodology section. Results & Discussion presents the outcome of experiments and compares the method's effectiveness with other techniques reported in literature. That section is then followed by the conclusions and future work.

Background

Wavelet Neural Networks

Wavelet Neural Networks are a class of neural networks that combine the theory of wavelets and neural networks [16]. In standard neural networks, weights and biases are the only parameters that are trained and the most common activation functions used are sigmoid, hyperbolic tangent and linear functions. The activation functions found in WNNs are those that belong to the family of wavelet basis functions, with the most common being the *Morelet* and *Mexican hat*. In addition to weights and biases, three other parameters are used in WNNs: translate, dilate and rotate. The use of standard gradient methods to adjust WNN parameters, in particular the weights, biases, the translate and dilate parameters, often resulted in premature convergence [16, 17]. For that reason, global optimization approaches, such as genetic algorithms and evolutionary programming techniques, have been used in applications such as air and ground traffic flow [18, 19], energy consumption [20], large scale function estimation [21], function approximation [22] and power transformer monitoring [23]. A diagram of a WNN is shown in Fig 1. WNNs generally have a feed-forward structure, with one hidden layer having m wavelons (ψ_m) and a neuron in the output layer. There are also n shortcut connections from the inputs to the output neuron.

Evolutionary Wavelet Neural Networks

EWNNs were first proposed by Khan et al. [15] as a method for optimizing all WNN parameters concurrently. The method was tested successfully on both simulated and real datasets [15]. For a detailed description of EWNN characteristics and performance, we refer the reader to reference [15].

Fig 2 is an example of the EWNN applied to a standard benchmark two-spiral task shown in Fig 2(a). Two-spiral is a non-linear task with two spirals (shown as black and white dots) each with 97 sample data points in a 2D Cartesian space [24, 25]. The two-spiral task is fairly a
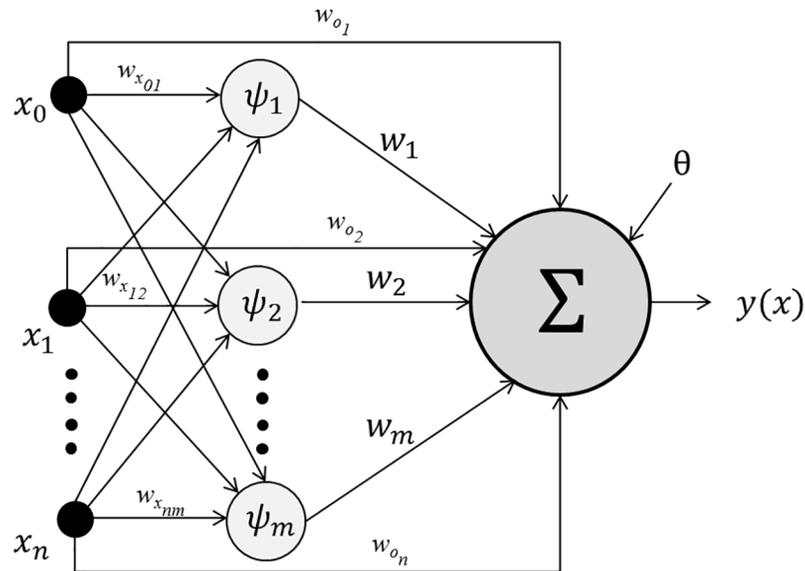
Fig 1. Structure of a Wavelet Neural Network. The network has n inputs in the input layer, m wavelons in the hidden layer and one neuron in the output layer. A bias θ is added to the WNN output response. Also, notice the n shortcut connections from the inputs to the output neuron.

https://doi.org/10.1371/journal.pone.0192192.g001

challenging problem where for an Artificial Neural Network (ANN) with architecture 2-5-5-5-1 took 10,000–20,000 epochs in [24]. While in [26] a 2-50-1 ANN was trained by employing a second-order Newton optimization method where training took only 650 epochs. In contrast, for EWNNs with a wavelet activation function of Morelet shown in Fig 2(b), the optimum response of the EWNNs was achieved within 9 generations and with two wavelons only. This indicates its potential to separate non-linear classes effectively and efficiently.

Classifier ensembles and pruning

The role of a classifier C is to learn how to map the feature set to a set of class label(s). The data samples are divided into training U and test V sets. The C is first trained on U where it learns the mapping process and then the performance of the C is measured on V. A multiple classifier system, or ensemble (Ens), is composed of a set of base classifiers that are trained on the same training dataset, and combined in a manner that improves the classification performance of the system. There are two main methods for creating an ensemble: averaging and voting [27]. Averaging is normally used for classifiers with numeric outputs. While voting is used for
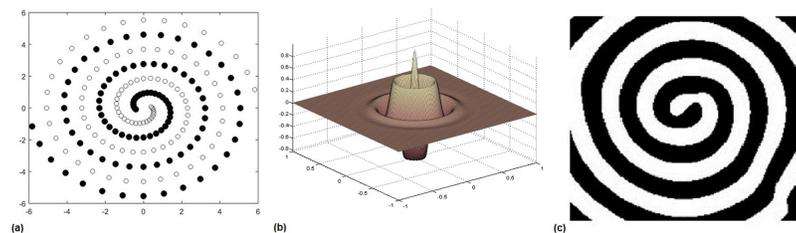

Fig 2. Training of an EWNN on a two-spiral task. (a) two-spiral classification task, each spiral consisting of 97 data points in the 2D Cartesian space. (b) Morelet wavelet activation function and (c) Optimal response on the task where a EWNN with Morelet wavelet activation function has separated the two classes successfully.

https://doi.org/10.1371/journal.pone.0192192.g002

categorical outputs (e.g. binary), and is used in the present study. Each sample is classified independently by the *k* classifiers that constitute the ensemble. The final outcome of classification will be the most represented class labels. It is the one that received the most votes. The ensemble *Ens* classification for a sample *V* is described in Eq 1 (for the binary classification case).

$$Ens(V) = \begin{cases} 1, & \text{if } \sum_{i=1}^{k} C_i(V) > \frac{k}{2} \\ \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Ensemble pruning, selective ensembles, ensemble selection and ensemble thinning are all different names given to the same task—reducing ensemble sizes. Pruned ensembles exhibit better performance and robustness with lower computational and memory costs [28], compared to traditional ensemble techniques [29, 30]. The three most popular ensemble pruning techniques are ranking, clustering and optimization [31], and this study focuses on the latter. Among the optimization techniques for ensemble pruning the most commonly used are evolutionary algorithms, semi-definite programming and hill climbing [32–34].

GASEN-b was one of the earliest algorithms for ensemble pruning, and was introduced by [32]. The ensemble is represented as a bit string, with each decision tree model using a bit. The bit string representation provides a direct mechanism of adding or removing classifiers, as opposed to a weighting mechanism with a predefined threshold. A similar approach was also used in [6] to select/remove classifiers from a heterogeneous pool of networks.

Zhang et al. [33] chose a quadratic integer programming approach for pruning. The weights were kept binary and the size of the final ensemble was prefixed. In terms of computational complexity, the algorithm could run in polynomial time.

Hill climbing methods generally use either forward selection or backward elimination of classifiers, and include various performance measures, e.g. diversity, weighted accuracy [35–39]. More recently, human-like foresight has been used as a measure to prune ensembles via hill climbing [34].

In this study, a pool of optimized EWNNs is pruned using genetic algorithms so that a better prediction model is obtained. The approach follows the GASEN-b mechanism [32] of pruning classifiers directly through bits so that to reduce the amount of parameter tuning. Our method introduces a fitness function that involves the sum of two accuracy measures: the accuracy of each individual classifier; and the ensemble accuracy using the voting method.

Network performance measures

There are many performance measures for binary classification problems available in the literature. Power [40] investigated those measures and generalized them for multiclass problems. Next, we present the measures used in this work:

- Training Accuracy (Tr_{acc}): fraction of correctly classified samples in the training set U.

- Test Accuracy (Te_{acc}): fraction of correctly classified samples in the test set V. This is also known as the classification accuracy, and expressed as $Te_{acc} = (TP + TN)/(P + N)$. TP represents true positive cases, i.e. accurate classification of control (non-diseased) samples; TN represents true negative cases, i.e. accurate classification of diseased samples; and $(P + N)$ is the total number of positive and negative test samples.

- Sensitivity (*Sens*): measurement of the fraction of true positive cases, mathematically expressed as $Sens = TP/(TP + FN)$. FN is the number of false negatives and reflects the more serious mistake of classifying a disease sample as control.

- Specificity (*Spec*): measurement of the fraction of true negative cases, mathematically represented as $Spec = TN/(TN + FP)$. FP reflects the misclassification of control samples as diseased ones.

- Mathew's Correlation Coefficient (*MCC*): is a balanced measure of quality for binary classification problems, normally used if classes are unbalanced. The measure was introduced in [41] and is expressed as:

$$MCC = W1/\sqrt{W2}, \text{where}$$

$$W1 = TP \cdot TN - FP \cdot FN$$

$$W2 = (TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)$$

Experimental simulations

This section provides a description of the three biomedical datasets, references to some related studies and the experimental settings for the proposed approach. An overview of the datasets' characteristics is given in Table 1.

Datasets

Digital Database for Screening Mammography (DDSM). The DDSM is an online repository of mammographic images (available at: http://marathon.csee.usf.edu/Mammography/Database.html) with different resolutions and obtained from various hospitals [46, 47]. The suspicious areas on the mammograms were manually marked by two experienced radiologists. For analysis, these markings are represented as chain codes and hence can be extracted easily. In the dataset used by [48], 200 mammographic images scanned by a HOWTEK scanner at 43.5 micron per pixel spatial resolution were downloaded and extracted via the chain code. That dataset had an equal number of benign and malignant samples. Even though [48] derived 25 features from the extracted region, only 6 of the features were actually investigated in the present study, in order to provide a fair comparison with previous works that used the same dataset [11, 49]. Among those 6 features, there are 4 BIRADS (Breast Imaging Reporting and Data System established by [50]) lexicon features: *mass shape, mass margin, assessment, breast density*, specified by an expert radiologist; and 2 features: *Patient age* and *subtlety*, that were extracted from the individual mammographic records.

Little's Parkinson's Dataset (LPD). This dataset (available at: http://archive.ics.uci.edu/ml) was acquired from the online machine learning database repository from the University of California at Irvine (UCI) [51, 52]. It is a challenging, imbalanced dataset that has been investigated previously by several researchers [9, 53–55]. It contains 195 samples, each with 22 different biomedical voice measurements. These voice measurements were taken from 31 individuals, where 23 had Parkinson's disease. Each patient has between 6 and 7 records in the data set, totalling 195 samples.

Sakar's Parkinson's Dataset (SPD). The dataset by Sakar et al. [45] is a recent entry (from 2014) in the UCI database (available at: http://archive.ics.uci.edu/ml) [43]. The dataset contains multiple speech recordings that include sustained vowels (a, o, u), numbers from 1 to 10, four short rhyming sentences and nine turkish words from 40 individuals. These

Table 1. Datasets' characteristics used in this study.

Datasets (References)	Abbrev.	Class distribution (control,diseased)	Number of Features
Digital Database for Screening Mammography [42]	DDSM	(100,100)	6
Little's Parkinson's Disease [43, 44]	LPD	(48,147)	22
Sakar's Parkinson's Disease [45]	SPD	(520,520)	26

recordings sum up to 26 records per individual. Half of the individuals are diagnosed with Parkinson's disease and the other half represents control subjects.

Training and test sets

For all the datasets, the data was divided into 90% training and 10% test data. The proposed approach is divided into two main phases as shown in Fig 3. Phase I creates optimal EWNNs from cross validation folds conducted on the 90% training and validation data and the average classification accuracy Te_{acc} for the EWNNs was reported. The optimal EWNNs were then used by next phase. Phase II uses genetic algorithm to prune the optimal EWNN classifiers where the separate test set was used and a final ensemble classification accuracy ETe_{acc} was then reported.

In both LPD and SPD datasets, individuals had different numbers of records. Thus, if more than half of the individual's records are classified as Parkinson's disease, then the individual itself is classified as Parkinson's disease (diseased). This approach was adopted from [45, 54] in order to avoid over-fitting, as the frequency response of the records of the same patient are potentially very similar.

Approach

Phase I: Generating optimized EWNNs.

1. EWNN initialization: An EWNN genome requires the initialization of the number of wavelons, the different parameters of each wavelon, and the wavelet function(s).

- The number of wavelons is critical as too many wavelons would likely result in over-fitting and too few would not capture the variability of the data [56]. The three datasets have been investigated in detail under different parameter settings and those are reported in [15]. The best configurations from that study were adopted here. The number of wavelons used for each dataset is shown in Table 2.

- Selection of an appropriate activation function depends on the data itself, but the Mexican hat wavelet has performed satisfactorily in many applications [56]. For the DDSM, the present study uses a heterogeneous WNN with four possible activation functions. For the remaining case studies, we used a homogeneous WNN that uses the Mexican hat wavelet as activation function.

- Each wavelon is represented by matrices of inputs $x_n \in [1, Feat]$; switches $c_n \in \{0, 1\}$ where 0/1 indicates non-connected/connected features, respectively; input weights $w_{x_{nm}} \in [-1, +1]$; scale parameters $\alpha_{nm} \in [0, 1]$; translation parameters $\beta_{nm} \in [-\infty, +\infty]$; rotation parameters $R_{nm} \in [-1, +1]$; as well as categorical values representing the type of wavelet function $\psi_m \in [1, number_{waveletFunctions}]$; wavelon weights $wt_m \in [-1, +1]$; and active neurons $ot_m \in \{0, 1\}$, where 0/1 represents an inactive/active hidden neuron, respectively. The
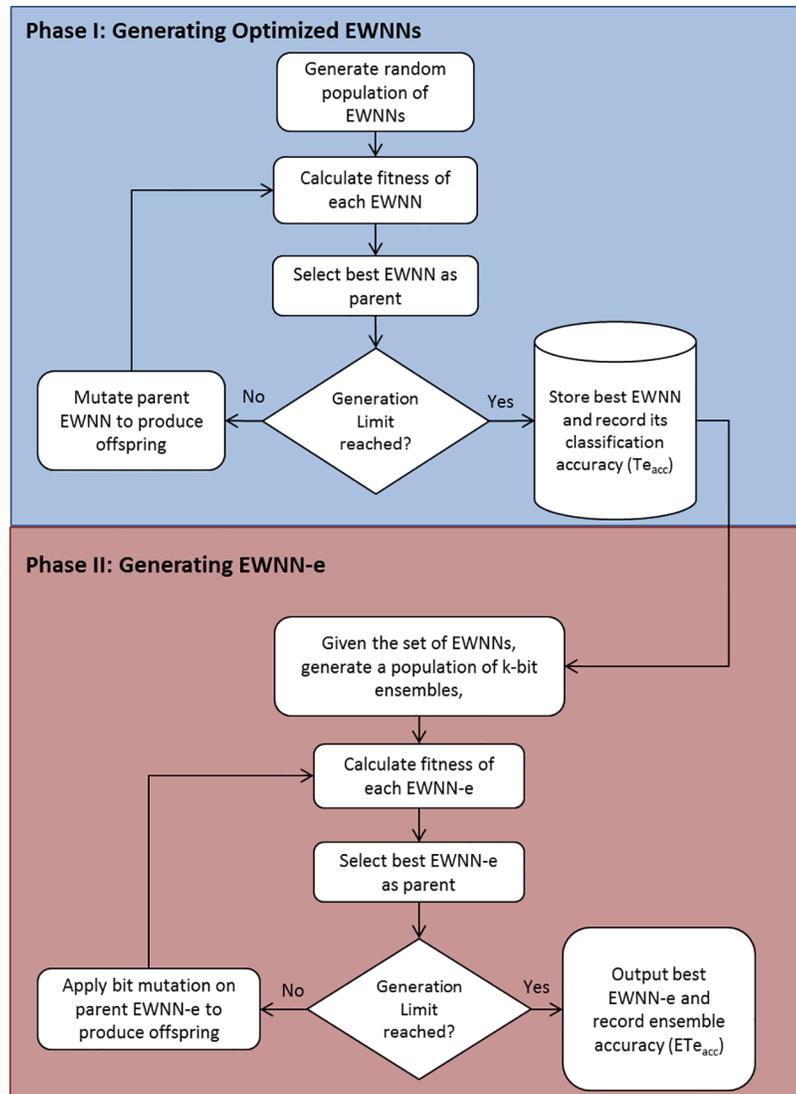
Fig 3. Flowchart of the two phases of the approach. Phase I is the process of generating optimized EWNNs. Phase II uses the optimized EWNNs to generate the ensemble of EWNNs.

parameters of each wavelon are initialized uniformly at random, within the corresponding ranges of possible values.

2. Population Size: There are two basic types of evolutionary strategies: (μ, λ)-ES and $(\mu + \lambda)$-ES [57]. μ represents the parent population and λ refers to the number of offspring produced in a generation. In (μ, λ)-ES, offspring replaces the parents as the μ fittest are selected

Table 2. Main parameter settings of the evolutionary wavelet neural networks for the different datasets.

Parameters	DDSM	LPD	SPD
Wavelons	4	5	5
Input to wavelons	6	6	26
$(\mu + \lambda)$-ES [57]	(1+25)	(1+25)	(3+20)

from λ, while in (μ + λ)-ES, the μ fittest are selected from both parents and offspring for the next generation. The value of μ and λ used for the different case studies are shown in Table 2.

3. Fitness evaluation: All individuals in the population are evaluated and sorted based on their accuracies and mean square error where the best individual is promoted as parent to the next generation. The purpose of using two dimensional sorting is to promote networks with uncorrelated evaluation metrics in generations ahead.

4. Mutation: A 1% mutation rate is used to generate new EWNNs, similar to [15]. Mutation occurs in three different ways. For continuous parameters, such as input weights, wavelon weights, translation, rotation, dilation parameters and the bias, values are perturbed by adding a small percentage of the current value. For binary parameters, e.g. switch, the value is inverted from 0 to 1 or 1 to 0. For the third type of mutation, a network input is randomly changed to another input feature in the feature list, or similarly, a wavelet function is randomly changed to another wavelet function in the list.

5. Termination condition: The simulations stop at 2,000 generations. We observed that this value is sufficient for the evolutionary process to converge to a high-quality solution. The optimal EWNNs are later used in Phase II to create the ensembles. A total of 50 independent evolutionary runs were executed for each of the cross-validation folds.

Phase II: Genetic algorithm-based ensemble.
Given the set of optimized EWNN ensembles, the next step is to prune them. This stage uses another genetic algorithm as follows:

1. Chromosome *Chr* representation: A *k*-bit string is used to represent an ensemble with the optimized EWNNs. A bit value of 1 indicates that the classifier is actively used in the ensemble; 0 otherwise.

2. Population size: After a number of preliminary tests, we decided for an (μ+λ)-evolutionary strategy with μ = 3 parents and λ = 25 offspring in each generation. For ensemble pruning, having 3 parents considerably reduced the risk of premature convergence and at the same time kept the evolutionary process under a reasonable selective pressure.

3. Fitness evaluation: The fitness value of each chromosome is evaluated as in Eq 2. It is an average of the individual accuracies Tr_{acc} of the active EWNNs and their ensemble training accuracy *Ens(U)*, where the objective is to maximize the average accuracies.

$$FE(Chr) = 0.5(Tr_{acc} + Ens(U)) \qquad (2)$$

4. Mutation: After pilot tests, mutation rate was set to 1% for all simulations, and the strategy used was bit-swap.

5. Termination condition: The limit for the number of generations was set at 1,000. Ensemble accuracy was found not to improve after few hundred generations.

The program starts with random chromosomes that are evaluated based on the fitness function in Eq 2. The best individuals are selected as parents and thus preserved for the next generation—all other individuals are removed. Then, λ offspring are produced by mutating the parents. Every offspring is evaluated and added to the next generation. The process continues

until the number of generations limit is reached. The best parent's ensemble accuracy ETe_{acc} on the test set is then reported.

Results and discussion

Did the ensemble of EWNNs perform better? The performance of the evolutionary ensemble method is shown in Table 3. Classification accuracy Te_{acc}, ensemble classification accuracy ETe_{acc}, sensitivity *Sens*, specificity *Spec* and Mathew correlation coefficient *MCC* are reported for the three datasets. The ensemble approach improves the classification accuracy by up to 23.7 percentage points (Te_{acc} vs. ETe_{acc}), compared to individual EWNN classifiers. For the DDSM dataset, the ensemble approach improved the performance of the network from 89.0% to 95.5%. An MCC score of 91.0% also indicates a very high classification accuracy. For the LPD dataset the accuracy increased from 92.9% to 100%, and for the SPD dataset it increased from 66.3% to 90.0%.

What were the significant features identified by the process? Fig 4 is the averaged connected features for all datasets, across 50 independent runs in EWNN, and the number of active classifiers in the EWNN-e. In a standard WNN all features are connected to every wavelons in the hidden layer. While in EWNNs (from Fig 4), there is some variability in how often these features are connected. This indicates the flexibility of pruning features (during training) at the hidden layer, as opposed to the input layer, for which many feature reduction methods already exist. For the DDSM dataset, *mass margin*, *patient age*, *mass shape* and *assessment* were the top four features that had an impact on performance—similarly to [48]. For the LPD dataset, *spread1* and *D2* were the top two features—similarly to [53]. The trend of feature selection was found to be same for both the EWNN and EWNN-e networks for all datasets except SPD. For the SPD dataset, $Shimmer_{apq3}$ is the top feature in the ensemble network, whereas $Shimmer_{dda}$ is the top one for the individual EWNNs. This drift in frequency of feature selection indicates possible significance of the feature in the ensemble domain.

Should every wavelon be fully connected? The connectivity or dimensionality of a wavelon is determined by the number of active or connected inputs. Fig 5 displays the sum of the wavelons' dimensions for each dataset, over 50 independent runs and, over the number of active classifiers in the final EWNN-e, across the 10 folds. The frequency of each wavelon dimension is lower in the ensemble network, as classifiers are pruned. The ensemble networks exhibited different trends, depending on the dataset. Interestingly, for the DDSM dataset we observed a reduction in the number of 6-dimensional wavelons, thus indicating that fully connected EWNNs were not part of the ensemble network. The frequent occurrence of wavelons with lower dimensions indicates that WNNs should be given the flexibility to adjust their input, in contrast to a standard WNNs, where all inputs are connected [16].

How many classifiers are necessary to create an effective ensemble? The average number of EWNNs in the ensemble networks for the datasets is shown in Table 4. The ensemble networks combine around 1/3 (14-17) of the 50 available EWNNs, and they improved both speed and performance, compared to the non-ensemble approach.

From Table 5, it can be concluded that the proposed method generated either competitive or better results in comparison to existing techniques. An advantage of EWNN-e is that it does not require pre-processing for feature pruning, which is present in some of the comparison methods. Given the results, it can be stated that the ensemble version of EWNN classifiers is a suitable approach for predictive analysis. Just for clarification purposes, and to put the results into context, for the DDSM dataset, the accuracy reported for NN-e was achieved with an ensemble of 127 classifiers [11], as opposed to the average of only 14.50 in the proposed

Table 3. Performance of the ensemble EWNN on the different case studies. Notice the increase in accuracy of the classifiers when an ensemble approach is adopted (second column).

Datasets	Te_{acc}%	ETe_{acc}%	Sens.%	Spec.%	MCC%
DDSM	89.0	95.5	95.0	96.0	91.0
LPD	92.9	100.0	100.0	-	100.0
SPD	66.3	90.0	93.0	97.0	87.0

method. That is, NN-e has a better performance for that dataset, but the classifier is much more complex than the classifiers obtained by our approach.

Conclusion

Ensemble approaches aim at combining the classification power of individual classifiers ultimately improving the overall performance of the system. The current study contributes to the
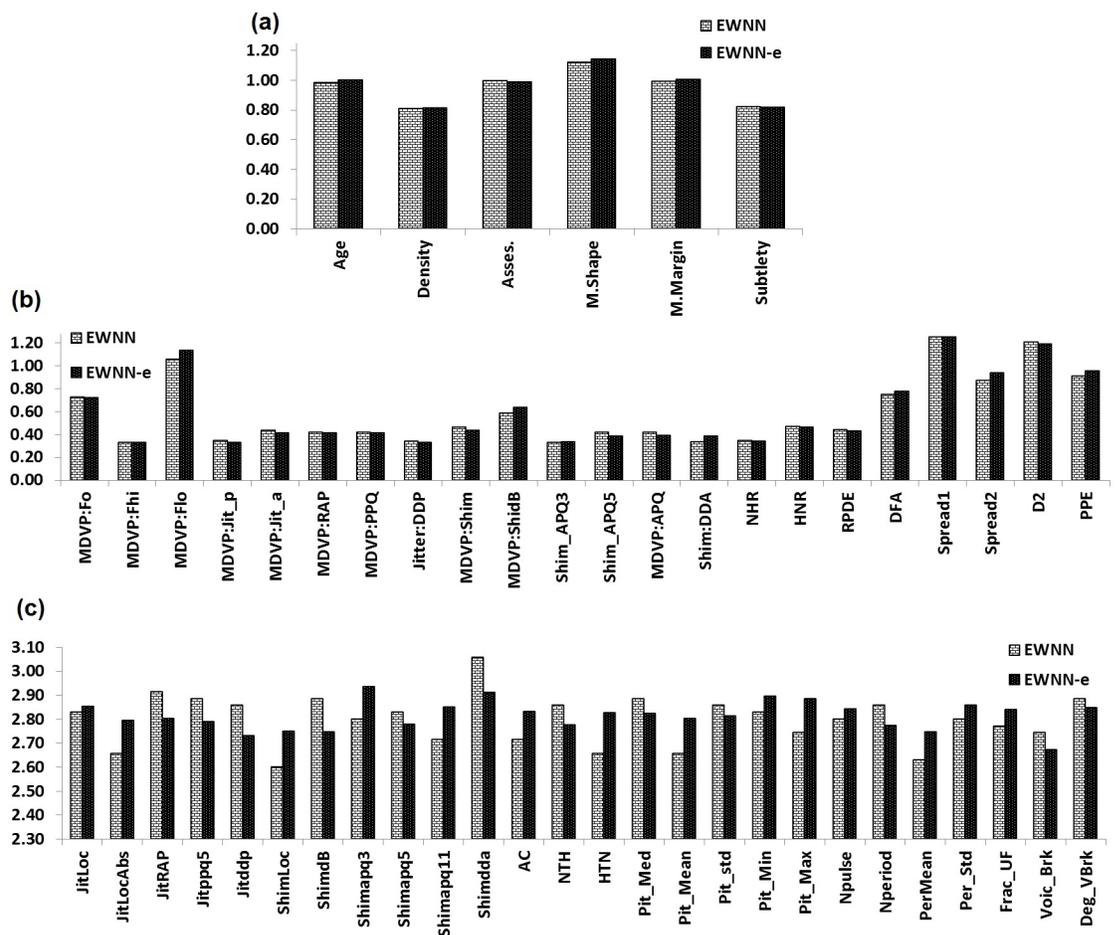

Fig 4. Identification of significant features. The figure shows the average number of connections per feature within the EWNN (over 50 independent evolutionary runs) and its ensemble EWNN-e (over the active classifiers), for the datasets: (a) DDSM [42], (b) LPD [44], and (c) SPD [45]. For all three datasets, and for all features, the average is higher than zero indicating that no feature should be completely removed from analysis. For illustration purposes, consider the example of feature Age in (a). The correct way to interpret the values is that the feature is connected to 1 wavelon on average, considering the 50 runs of EWNN. Details on the features can be found in the referenced papers [42, 45, 44].
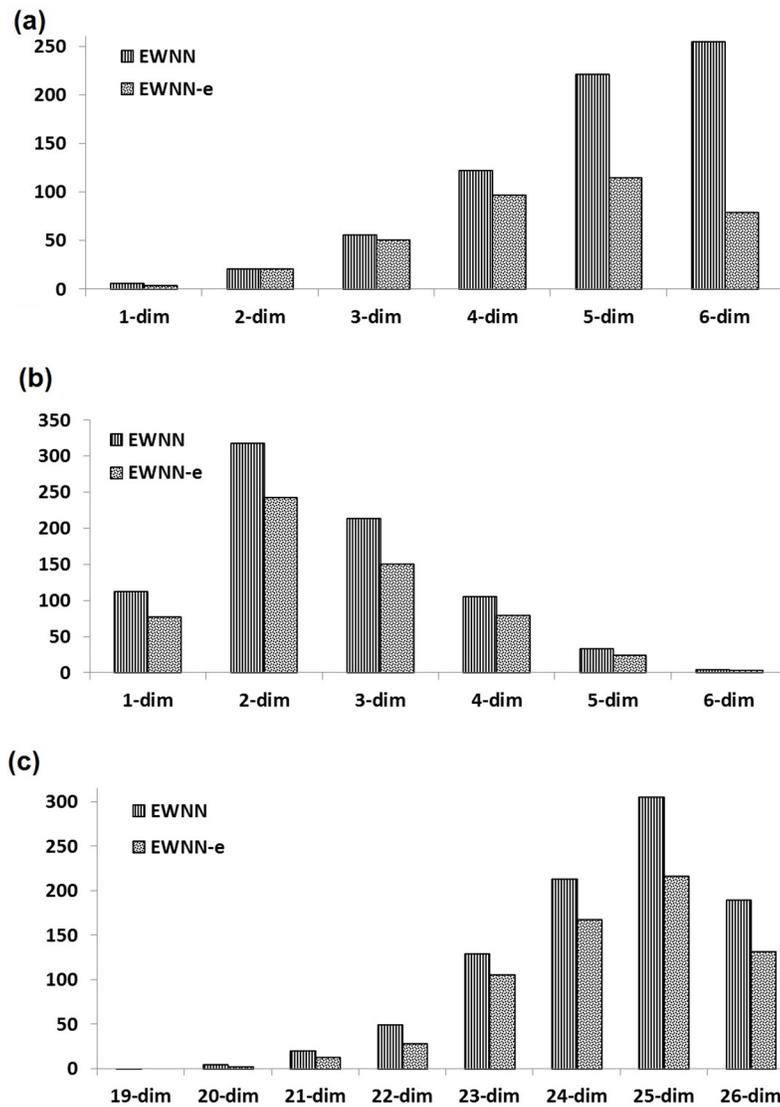
(a)

(b)

(c)

Fig 5. Should every wavelon be fully connected? Summation of wavelons' dimensionality for individual EWNNs (over 50 independent evolutionary runs); and for the ensemble EWNN-e (over the number of active classifiers), for the three datasets (a) DDSM, (b) LPD and (c) SPD. Note the overall increasing trend for DDSM, with a larger number of high-dimensionality wavelons (except for EWNN-e which shows a decrease in the number of 6-dimensional wavelons). For LPD we see a concentration between 2- to 4-dimensional wavelons for both individual and ensemble EWNNs. Finally, for SPD, we see a concentration at the higher dimensions. These results indicate that having the features connected to all wavelons is not necessarily the most appropriated choice.

https://doi.org/10.1371/journal.pone.0192192.g005

Table 4. Average number of active EWNNs in the ensembles, using 10-fold cross-validation, and across the three datasets. The average is calculated over 50 independent runs.

Datasets	Average EWNNs
DDSM	14.50
LPD	16.80
SPD	16.80

https://doi.org/10.1371/journal.pone.0192192.t004

Table 5. Comparison between EWNN/EWNN-e and the different classifiers found in the literature for the DDSM, LPD and SPD datasets. In the case of LPD, EWNN-e outperformed all methods reported in literature, reaching a test accuracy of 100%.

DDSM	
Algorithm	**Test accuracy (%)**
SCBDL [58]	97.50
SCNN [59]	94.00
MCSVM [60]	94.50
NN-e [11]	99.00
LCA-e [10]	94.00
EWNN-e	95.50
LPD	
Algorithm	**Test accuracy (%)**
SVM+RBF-e [9]	88.71
SVM+LADTree-e [9]	92.82
SVM+J48-e [9]	92.3
SVM+KSTAR-e [9]	96.41
SVM+IBk-e [9]	96.93
EWNN-e	100.00
SPD	
Algorithm	**Test accuracy (%)**
EWNN	66.3
EWNN-e	90.00

https://doi.org/10.1371/journal.pone.0192192.t005

literature of ensemble classifiers by proposing an ensemble of evolutionary wavelet neural networks (EWNN-e).

The performance of the EWNN-e has been validated on three biomedical datasets. The pruned EWNN-e used less than 1/3 of the available EWNNs and resulting in better performance. For one of the datasets, the method achieved a testing accuracy of 100%, whereas the best approach reported in literature to date had reached 96.9% only.

Each EWNN used all features available, but features were not connected to every wavelon in the network. In other words, the proposed method prunes features at the hidden layer level, instead of at the input layer level.

The dimensionality of the wavelons is represented by the number of active inputs. The trend of the average sum of wavelons' dimensionality in the Parkinson's disease datasets was same for both EWNNs and EWNN-e. While for the Breast Cancer dataset (DDSM) the wavelons' dimensionality of a fully connected wavelon were reduced in the EWNN-e. This indicates that WNNs should be provided with the flexibility to adjust their network inputs, as opposed to a conventional WNNs, where all inputs are forced to be connected.

Author Contributions

Conceptualization: Maryam Mahsal Khan.

Investigation: Maryam Mahsal Khan.

Methodology: Maryam Mahsal Khan.

Software: Maryam Mahsal Khan.

Supervision: Alexandre Mendes, Stephan K. Chalup.

Validation: Maryam Mahsal Khan.

Visualization: Maryam Mahsal Khan.

Writing – original draft: Maryam Mahsal Khan.

Writing – review & editing: Alexandre Mendes, Stephan K. Chalup.

References

1. Breast Cancer Care WA; Cited 17 June 2014;. Available from: http://www.breastcancer.org.au/.

2. Parkinson's Australia, Living with Parkinson's Disease Update—October 2011; Deloitte Access Economics Pty Ltd; Cited August 2014. Available from: http://www.parkinsonsnsw.org.au/.

3. Breiman L. Bagging Predictors. Machine Learning. 1996; 24(2):123–140. https://doi.org/10.1023/A:1018054314350

4. Breiman L. Bias, Variance, and Arcing Classifiers. Technical Report 460, Statistics Department, University of California; 1996.

5. Ho TK, Hull JJ, Srihari SN. Decision Combination in Multiple Classifier Systems. IEEE Trans Pattern Anal Mach Intell. 1994; 16(1):66–75. https://doi.org/10.1109/34.273716

6. Haque MN, Noman N, Berretta R, Moscato P. Heterogeneous Ensemble Combination Search Using Genetic Algorithm for Class Imbalanced Data Classification. PLoS ONE. 2016; 11(1):1–28. https://doi.org/10.1371/journal.pone.0146116

7. Gang Y, Hualin Z, Fei C, Chang S, Chih-Min L, Changle Z. Integration of classifier diversity measures for feature selection-based classifier ensemble reduction. Soft Comput. 2015; p. 2995–3005.

8. Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF. SVM and SVM Ensembles in Breast Cancer Prediction. PLoS ONE. 2017; 12(1):1–14.

9. Ozcift A. SVM feature selection based rotation forest ensemble classifiers to improve computer-aided diagnosis of Parkinson disease. J Med Syst. 2012; 36(4):2141–2147. https://doi.org/10.1007/s10916-011-9678-1 PMID: 21547504

10. Pour SG, Mc Leod P, Verma B, Maeder A. Comparing Data mining with ensemble classification of breast cancer masses in digital mammograms. In: Second Australian Workshop on Artificial Intelligence in Health: AIH 2012. Aachen, Germany: CEUR-WS, Sun SITE Central Europe operated under the umbrella of RWTH Aachen University; 2012. p. 55–63.

11. Mc Leod P, Verma B. Effects of Large Constituent Size in Variable Neural Ensemble Classifier for Breast Mass Classification. In: 20th International Conference on Neural Information Processing (ICONIP'2013). vol. 8228; 2013. p. 525–532.

12. Gul A, Perperoglou A, Khan Z, Mahmoud O, Miftahuddin M, Adler W, et al. Ensemble of a subset of kNN classifiers. Advances in Data Analysis and Classification. 2016; p. 1–14.

13. Liao GC. Application a Novel Evolutionary Computation Algorithm for Load Forecasting of Air Conditioning. In: Asia-Pacific Power and Energy Engineering Conference; 2012. p. 1–4.

14. Vazquez LA, Jurado F, Alanis AY. Decentralized Identification and Control in Real-Time of a Robot Manipulator via Recurrent Wavelet First-Order Neural Network. Mathematical Problems in Engineering. 2015; 2015:1–12. https://doi.org/10.1155/2015/451049

15. Khan MM, Mendes A, Zhang P, Chalup SK. Evolving multi-dimensional wavelet neural networks for classification using Cartesian Genetic Programming. Neurocomputing. 2017; 247:39–58. https://doi.org/10.1016/j.neucom.2017.03.048

16. Zhang Q, Benveniste A. Wavelet networks. IEEE Trans Neural Netw. 1992; 3(6):889–898. https://doi.org/10.1109/72.165591 PMID: 18276486

17. Zhang J, Walter GG, Miao Y, Lee WNW. Wavelet neural networks for function learning. IEEE Trans Signal Process. 1995; 43(6):1485–1497. https://doi.org/10.1109/78.388860

18. Qiu F, Li Y. Air traffic flow of genetic algorithm to optimize wavelet neural network prediction. In: IEEE International Conference on Software Engineering and Service Science (ICSESS'2014); 2014. p. 1162–1165.

19. Yang HJ, Hu X. Wavelet neural network with improved genetic algorithm for traffic flow time series prediction. Optik. 2016; 127(19):8103–8110. https://doi.org/10.1016/j.ijleo.2016.06.017

20. Zhao H, Liu R, Zhao Z, Fan C. Analysis of Energy Consumption Prediction Model Based on Genetic Algorithm and Wavelet Neural Network. In: 3rd International Workshop on Intelligent Systems and Applications (ISA'2011); 2011. p. 1–4.

21. Sahoo D, Dulikravich GS. Evolutionary Wavelet Neural Network for Large Scale Function Estimation in Optimization. In: 11th Multidisciplinary Analysis and Optimization Conference (AIAA/ISSMO); 2006. p. 1–11.

22. Xu J. A Genetic Algorithm for Constructing Wavelet Neural Networks. In: International Conference on Intelligent Computing (ICIC'2006). vol. 4113 of Lecture Notes in Computer Science; 2006. p. 286–291.

23. Huang YC, Huang CM. Evolving wavelet networks for power transformer condition monitoring. IEEE Trans Power Deliv. 2002; 17(2):412–416. https://doi.org/10.1109/61.997908

24. Lang KJ, Witbrock MJ. Learning to tell two spirals apart. In: 1988 Connectionist Models Summer School. Touretzky D., Hinton G. and Sejnowski (eds), Morgan Kaufmann, Los Altos, CA; 1988. p. 52–59.

25. Chalup SK, Wiklendt L. Variations of the Two-spiral Task. Conn Sci. 2007; 19(2):183–199. https://doi.org/10.1080/09540090701398017

26. Osowski S, Bojarczak P, Stodolski M. Fast Second Order Learning Algorithm for Feedforward Multilayer Neural Networks and its Applications. Neural Netw. 1996; 9(9):1583–1596. https://doi.org/10.1016/S0893-6080(96)00029-9 PMID: 12662555

27. Zhou ZH. Ensemble Methods: Foundations and Algorithms. 1st ed. Chapman & Hall/CRC; 2012.

28. Zaki FW, Abd el Fattah AI, Enab YM, El-Konyaly SH. An ensemble average classifier for pattern recognition machines. Pattern Recognit. 1988; 21(4):327–332. https://doi.org/10.1016/0031-3203(88)90046-5

29. Kuncheva LI, Whitaker CJ. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. Mach Learn. 2003; 51(2):181–207. https://doi.org/10.1023/A:1022859003006

30. Cavalcanti GDC, Oliveira LS, Moura TJM, Carvalho GV. Combining diversity measures for ensemble pruning. Pattern Recognit Lett. 2016; 74:38–45. https://doi.org/10.1016/j.patrec.2016.01.029

31. Tsoumakas G, Partalas I, Vlahavas I. An Ensemble Pruning Primer. In: Applications of Supervised and Unsupervised Ensemble Methods. vol. 245 of the series Studies in Computational Intelligence. Springer Berlin Heidelberg; 2009. p. 1–13.

32. Zhou ZH, Tang W. Selective Ensemble of Decision Trees. In: 9th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. vol. 2639 of Lecture Notes in Artificial Intelligence. Springer Berlin Heidelberg; 2003. p. 476–483.

33. Zhang Y, Burer S, Street WN. Ensemble Pruning Via Semi-definite Programming. J Mach Learn Res. 2006; 7:1315–1338.

34. Taghavi ZS, Sajedi H. Human-inspired ensemble pruning using hill climbing algorithm. In: AI Robotics and 5th RoboCup Iran Open International Symposium (RIOS), 2013 3rd Joint Conference of; 2013. p. 1–7.

35. Banfield RE, Hall LO, Bowyer KW, Kegelmeyer WP. Ensemble diversity measures and their application to thinning. Inf Fusion. 2005; 6(1):49–62. https://doi.org/10.1016/j.inffus.2004.04.005

36. Caruana R, Niculescu-Mizil A, Crew G, Ksikes A. Ensemble Selection from Libraries of Models. In: Proceedings of the Twenty-first International Conference on Machine Learning (ICML); 2004. p. 18–27.

37. Fan W, Chu F, Wang H, Yu PS. Pruning and Dynamic Scheduling of Cost-sensitive Ensembles. In: Eighteenth National Conference on Artificial Intelligence. Menlo Park, CA, USA: American Association for Artificial Intelligence; 2002. p. 146–151.

38. Ioannis P, Grigorios T, Ioannis V. Focused Ensemble Selection: A Diversity-Based Method for Greedy Ensemble Selection. In: Proceedings of the 2008 Conference on ECAI 2008: 18th European Conference on Artificial Intelligence; 2008. p. 117–121.

39. Ioannis P, Grigorios T, Ioannis V. An ensemble uncertainty aware measure for directed hill climbing ensemble pruning. Mach Learn. 2010; 81(3):257–282. https://doi.org/10.1007/s10994-010-5172-0

40. Powers DMW. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. J Mach Learn Tech. 2011; 2(1):37–63.

41. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)—Protein Structure. 1975; 405(2):442–451. https://doi.org/10.1016/0005-2795(75)90109-9

42. Digital Database for Screening Mammography; 2014 [cited 2014 May 5]. Available from: http://marathon.csee.usf.edu/Mammography/Database.html.

43. UCI Machine Learning Repository; Parkinson Speech Dataset with Multiple Types of Sound Recordings Data Set; 2014 [cited 2014 May 5]. Available from: http://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with++Multiple+Types+of+Sound+Recordings.

44. Little MA, McSharry PE, Hunter EJ, Spielman JL, Ramig LO. Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease. IEEE Trans Biomed Eng. 2009; 56(4):1015–1022. https://doi.org/10.1109/TBME.2008.2005954 PMID: 21399744

45. Sakar BE, Isenkul ME, Sakar CO, Sertbas A, Gurgen F, Delil S, et al. Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings. IEEE J Biomed Health Inform. 2013; 17(4):828–834. https://doi.org/10.1109/JBHI.2013.2245674 PMID: 25055311

46. Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer WP. The Digital Database for Screening Mammography. In: Proceedings of the Fifth International Workshop on Digital Mammography; 2001. p. 212–218.

47. Bowyer K, Kopans D, Kegelmeyer WP, Moore R, Chang K, Kumaran SM. Current status of the Digital Database for Screening Mammography. In: Proceedings of the Fourth International Workshop on Digital Mammography; 1998. p. 457–460.

48. Zhang P, Kumar K. Analyzing Feature Significance from Various Systems for Mass Diagnosis. In: IEEE International Conference on Computational Intelligence for Modelling Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMA-IAW-TIC'2006); 2006. p. 141–146.

49. Zhang P, Kumar K, Verma B. A Hybrid Classifier for Mass Classification with Different Kinds of Features in Mammography. In: Second International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'2005). vol. 3614; 2005. p. 316–319.

50. American College of Radiology; Cited July 2017. Available from: https://www.acr.org/.

51. Lichman M. UCI Machine Learning Repositry; 2014 [cited 2014 May]. Available from: http://archive.ics.uci.edu/ml.

52. Little MA, McSharry PE, J RS, Costello DAE, Moroz IM. Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection. Biomed Eng Online. 2007; 6(23). https://doi.org/10.1186/1475-925X-6-23 PMID: 17594480

53. Sakar CO, Kursun O. Telediagnosis of Parkinson's Disease Using Measurements of Dysphonia. J Med Syst. 2010; 34(4):591–599. https://doi.org/10.1007/s10916-009-9272-y PMID: 20703913

54. Caglar MF, Cetisli B, Toprak IB. Automatic Recognition of Parkinson's Disease from Sustained Phonation Tests Using ANN and Adaptive Neuro-Fuzzy Classifier. Journal of Engineering Science and Design. 2010; 1(2):59–64.

55. Astrom F, Koker R. A parallel neural network approach to Prediction of Parkinson's Disease. J Expert Syst Appl. 2011; 38(10):12470–12474. https://doi.org/10.1016/j.eswa.2011.04.028

56. Alexandridis AK, Zapranis AD. Wavelet neural networks: A practical guide. Neural Netw. 2013; 42:1–27. https://doi.org/10.1016/j.neunet.2013.01.008 PMID: 23411153

57. Beyer HG, Schwefel HP. Evolution strategies: A comprehensive introduction. Nat Comput. 2002; 1 (1):3–52. https://doi.org/10.1023/A:1015059928466

58. Verma B, Mc Leod P, Klevansky A. Classification of benign and malignant patterns in digital mammograms for the diagnosis of breast cancer. Expert Syst Appl. 2009; 37:3344–3351. https://doi.org/10.1016/j.eswa.2009.10.016

59. Verma B, Mc Leod P, Klevansky A. A novel soft cluster neural network for the classification of suspicious areas in digital mammograms. Pattern Recognit. 2009; 42:1845–1852. https://doi.org/10.1016/j.patcog.2009.02.009

60. Mc Leod P, Verma B. Multi-Cluster Support Vector Machine Classifier for the classification of suspicious areas in digital mammograms. Int J Comput Intell Appl. 2011; 10:481–494. https://doi.org/10.1142/S1469026811003203