International Journal of Scientific & Engineering Research, Volume 13, Issue 2, February 2022.
ISSN 2229-5518

Analysis of Training Optimization Algorithms in the NARX Neural Network for Classification of Heart Sound Signals.

Sara Khaled, Mahmoud Fakhry, Hamada Esmail, Ahmed Ezzat, Ehab K. I. Hamad

Abstract—A trained neural network classifier is commonly used to predict cardiac abnormalities by the classification of heart sound signals, also known as phonocardiogram (PCG) signals. On the other hand, the best training optimization algorithm for this variety of classification problem is nevertheless up for discussion. In this study, we explore the use of the Nonlinear autoregressive networks with exogenous inputs (NARX) network for the classification of many different features extracted from labelled PCG signals. The classification performance of the trained NARX model is explained in terms of three separate optimization algorithms that are used to train the classifier. The specified results on testing PCG signals confirm that the NARX classifier is better when trained with the Bayesian regularization (BR) algorithm than when trained with the Levenberg-Marquardt (LM)or Scaled Conjugate Gradient (SCG)optimization algorithm. significantly, this classification model performs outperforms a standard approach.

Index Terms— Heart sound signals, Phonocardiogram (PCG), NARX, training optimization, LM, SCG, BR.

1 INTRODUCTION

Automatic detection of disease is based on the establishment of flexible and efficient non - invasive techniques. Cardiovascular diseases (CVDs) are also known as cardiac diseases. In terms of cardiac disease, one of the major causes of death around the world, medical experts are used to evaluate the heart health using a medical stethoscope to hear its sound. This assessment approach implies the acquisition of skills over a long period of time. In this method, it was the beginning of thinking of the automated examination of the heart's health through a computer-assisted analysis of sound recordings. Alongside, the electrocardiogram (ECG) [1] and the photo-plethysmogram (PPG) [2], the phonocardiogram (PCG), which records the heart's sounds and murmurs, its capable of being used efficiently to monitor the heart's health. The ECG and PCG are closely interconnected signals and are suspected of having more information than the PPG signal. The PCG signals allow for the registration, retailing, and interpretation of heart sounds as part of a comprehensive medical test.

The major reason for audible sounds during phonocardiography (PCG) recordings is the mechanical activity of the heart muscle. [3]. One cardiac cycle of the PCG signal contains two distinct heartbeats called $S_1$and S_2, and the systolic and diastolic regions. The heartbeats $S_1$and $S_2$are defined for normal heart and sound activities, called murmurs, arise in the systolic and diastolic regions in case that there is heart abnormality.

- *Sara Khaled is with the Electronics and Communication Department, High institute of Engineering and Technology, Luxor, Egypt. Sara Khaled is also with the Electrical Engineering Department, Faculty of Engineering, Aswan University, Aswan 81542, Egypt. (e-mail: sarakhaled2904@gmail.com)*
- *Mahmoud Fakhry is with the Electrical Engineering Department, Faculty of Engineering, Aswan University, Aswan 81542, Egypt. (e-mail: m.fakhry@aswu.edu.com)*
- *Ahmed Ezzat is with the Electronics and Communication Department, High institute of Engineering and Technology, Luxor, Egypt. (e-mail: engahmedezzat7l@gmail.com)*
- *Hamada Esmaiel, and Ehab K. I. Hamad are with the Electrical Engineering Department, Faculty of Engineering, Aswan University, Aswan 81542, Egypt. (e-mail: h.esmaiel@aswu.edu.com, e.hamad@aswu.edu.eg).*

The cardiac cycle is the time interval between the start of one heartbeat and the start of the next one. It can be described as the time between the start of $S_1$and the start of the following S_1in the next cycle. The systolic region is the time interval between the end of $S_1$and the beginning of S_2, and the diastolic region is the time interval from the end of $S_2$to the beginning of S_1in the next cycle.

We offer the use of the nonlinear autoregressive networks with exogenous inputs (NARX) to classify the heart sound signals whether they are normal or abnormal. There are two different architectures of NARX network, namely, open-loop and closed-loop. The open-loop architecture of NARX network is used throughout the training phase due to the availability of true past values of the time series. After that, this trained open-loop of NARX network is converted to a closed-loop architecture, which is useful for multi-step-ahead prediction in the testing phase. The objective of training a neural network is to minimize a large-scale cost function. This problem is handled with an optimization technique that searches through a space of possible values for the neural network weights for a set of weights that results in good performance on the training dataset. A certain training optimization algorithm may be appropriate for one issue but ineffective in another.

In this paper, we present an experimental comparison analysis of three optimization algorithms used for training the NARX network for the task of binary classification of PCG signals. Three different training optimization algorithms will be compared. Scaled Conjugate Gradient (SCG), Levenberg-Marquardt (LM), and Bayesian regularization (BR) are used to evaluate the NARX model for the identification of the perfect fit of the PCG signal training optimization algorithm with the best results.

The remainder of this work is arranged in the following manner. Section2 reports the related works. The proposed methodology is summarized in Section 3, Section 4 presents the experimental results, and the work is concluded in section5.

International Journal of Scientific & Engineering Research, Volume 13, Issue 2, February 2022
ISSN 2229-5518

2 Related Work

Trained artificial neural network (ANN) have shown good superiority in the computerized classification of heart sound signals. ANN models mainly include deep neural networks (DNN), feed-forward neural network (FFNN), convolution neural networks (CNN), recurrent neural networks (RNN), etc.

Multilayer perceptron (MLP) was created using a public PCG database, given by Physio net/CinC 2016. Then 38 features were extracted from those signals using time-domain statistical characteristics of the signal, Mel-frequency spectral coefficients (MFCC) and DWT detail, and approximation coefficients to investigate heart sounds and classified them as normal or abnormal based on the efficiency of the ensemble classifier as listed in [4].

The FFNN model is one of the most associated classifiers that was used for the classification of PCG signals. One network that uses a back-propagation FFNN with 324 features, is discussed in [5] and a second one that implements FFNN with 90 features is presented in [6]. For training and testing the network, the feature vectors are extracted from the time representation, the frequency representation, and the time-frequency representation of PCG signals. In [7], an ensemble of 20 FFNN is created for anomaly and reliability detection of 3454 PCG label records, that are provided by Physio net/Computing in Cardiology Challenge. In this approach, 40 features in the time, frequency, and time-frequency domains were extracted.

The implementation of deep learning has developed significantly in the classification of PCG signals, particularly the deep convolutional neural network (CNN). The wavelet coefficients are used as features for the classification of PCG signals using deep CNN in [8], and the Mel-frequency spectrum coefficients (MFCCs) in [9]. Researchers use a combination of time-frequency heat map representations and deep CNN for such a classification model in [10]. The statistics map of PCG signals is constructed using MFCC's and one-dimensional time series to obtain the time-frequency distribution of signal energy. Convolutional neural network (CNN) was indicated for heart sound classification without segmentation with the benefit of the designed CNN architecture, the features of the heart cycles with different start positions are fused in the network the proposed approach is implemented on standard datasets from the PASCAL classifying heart sounds challenge as mentioned in [11]. As part of the Physio Net/Computing in Cardiology Challenge 2016, CNNs are learned to distinguish normal/abnormal labels from 5-second samples taken from a recording rather than from the actual recording. The overall classification results are calculated for its segments using a voting system. The extracted features include Spectrograms and Mel frequency Cepstrum coefficients are our characteristics as in [12].

New varieties of ANNs are established to diagnose the sound heart into various types of valve-physiological heart disease, these are the multilayer perceptron (MLP), Elman neural network (ENN), and Radial Basis Function (RBF) network with a backpropagation training algorithm. Training feature vectors are constructed based on the wavelet decomposition of sound signals, which are divided into natural heart sound and the other six valve physiological heart categories as listed in [13].

To model the dynamic characteristics between sequential heart sound signals, a newly updated feature extraction based on MFCCs are selected to train a deep convolutional and recurrent neural network (CRNN) for future classification [14]. The proposed deep learning model provided the superiority of the embedded local characteristics extracted from the convolutional neural network (CNN) and the long-term associations collected by the recurrent neural network (RNN). Classification of PCG signals is later identified by applying the recurrent neural network (RNN) in many papers such as in [15]. The authors explained the classification results of four models of the network, i.e., the long short-term memory (LSTM), the bidirectional LSTM (B-LSTM), the gated recurrent unit (GRU), and the bidirectional GRU (B-GRU) based on Mel Frequency Cepstral Coefficient (MFCCs). A combination of two networks, i.e., B-LSTM and CNN, is explained in [16]. Moreover, a technique is suggested to learn visual, and time dependent characteristics of murmur based on spectrogram and MFCCs of PCG signals.

A DNN model by using Physio Net dataset used to categorize cardiac audiences based on features composed of fractional Fourier transform and MFCCs as mentioned in [17]. A DNN was used with MFCCs, and discrete wavelets transform (DWT) features from the heart sound signal to detect a database of 5 categories of PCG signals from various sources that contain one normal and 4 abnormal categories [18]. One of the popular classifiers used to identify the PCG signal is a multilayer perceptron Neural Network, which has been used to classify heart sounds using the discrete wavelet reduction, 250 cardiac periods from the heart sound model were used to implement the theoretical process as in [19].

In [20] the nonlinear autoregressive network with exogenous inputs (NARX) is exploited for the diagnosis of heart abnormality with spectral, temporal, and statistical classification features.

3 METHODOLOGY

The nonlinear autoregressive network with exogenous inputs (NARX) network is used for the binary classification, i.e., normal, and abnormal, of PCG signals. A feature vector of length 27 is used to train and test the NARX network. The vector is extracted from PCG signals of labelled normal and abnormal heart sound signals. This vector is composed of different entries, including deterministic coefficients and statistical parameters. The proposed methodology is depicted in Figure 1.

The methodology starts with the extraction process for a feature vector of PCG signals, then the classification of the extracted features by using the NARX model for the identification of cardiac abnormalities. When it comes to PCG signal classification, there are several factors to take into account. We explore the effects of using a particular optimization algorithm for training the NARX model. Among the various algorithms that can be used to train the NARX model, we select Scaled Conjugate Gradient (SCG), Levenberg-Marquardt (LM), and Bayesian regularization (BR) algorithms.

International Journal of Scientific & Engineering Research, Volume 13, Issue 2, February 2022
ISSN 2229-5518

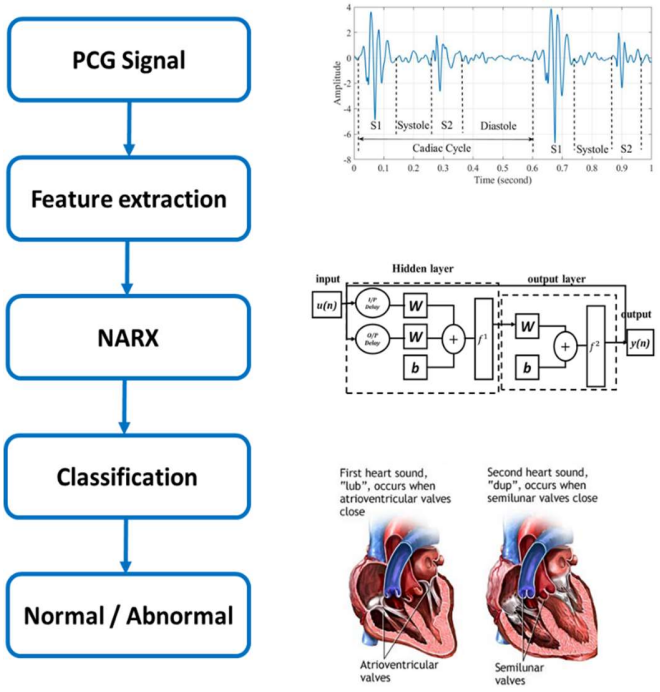

Fig. 1. The flow graph of proposed methodology

3.1 Feature Extraction

In this phase, parameters are assembled and used to classify signals, which can distinguish between the PCG signal categories, which can be a frequency domain, time domain, or statistics parameters [21]. A total of 27 features are extracted and categorized as shown in Table I.

TABLE 1
FREQUENCY DOMAIN AND STATISTICAL-BASED PARAMETER FEATURES

Frequency domain features	Statistical domain features
1 - MFCC1: MFCC13	1 - mean
2 - Dominant frequency value	2 - medians
3 - Dominant frequency magnitude	3 - variances
4- Dominant frequency ratios	4 - standard deviation
5 - Entropy	5 - mean absolute deviation
	6 - skewness
	7 - kurtosis
	8 - 25% percentile
	9 - 75% percentile
	10 - IQR

- Mean:

The mean of a signal is represented as the sum of all amplitude of the signal divided by the number of them. The mean of a signal is represented as the average amplitude of the signal over the total time. Taking amplitude elements at the signal as {x1, x2, x3. . .Xn}; mean (μ) of the amplitude for the signal under consideration is developed as given in Eq. (3), where xi is the value of the amplitude of the signal at its instance. The mean value in abnormal signals is, generally, higher than the normal signals.

$$\mu = \frac{1}{N}\sum_{i=0}^{N-1} x_i$$

- Median:

Arranging the amplitude values {x1, x2, x3. . .Xn} of a signal in ascending order and considering the middle element from a sorted list of N elements, its position is computed using the formula given as Eq. (4)

$$i = (N+1)/2$$

- Standard deviation (SD):

Standard deviation: measure variability and consistency of the signal. The low standard deviation shows that data points tend to be close to the average, while a higher standard deviation indicates that data points are widely distributed from signal values.

$$SD = \sqrt{\frac{1}{N}\sqrt{(xn-\mu)^2}}$$

- Mean Absolute Deviation (MAD):

(MAD): It describes variations in a dataset and provides a glimpse of spread out of value. For a waveform, MAD is an average of the summation of the difference of each amplitude point on the wave, with the overall mean of the signal.

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \mu|$$

- Quantile25 (C25):

The first 25% of the elements from a series of amplitude values of the signal, which are arranged in an ascending order $X_1 < X_2 < X_3, \ldots \ldots < X_n$ relative to their magnitude. These are selected by dividing the total number of elements by 4. The value, which is determined after performing this calculation is the 25th percentile element of the signal.C25 is represented by Quantile and expressed as Quantile 25 = N/4; where N = total number of values in a signal (from $X_1$to X_n, where $X_n$has maximum amplitude).

- Quantile75 (C75)

By arranging the first 75% of the items from a set of amplitude values of the signal, that are arranged in ascending order, $X_1 < X_2 < X_3, \ldots \ldots < X_n$ with relate to their magnitude. These are specified by dividing the total number of elements by 4 and multiplying the result by 3. The value, that obtained after administering this analysis the 75th percentile item of the signal.C75 refer Quantile75 and is defined as C75 = 3N/4; where N = total number of values in a signal (from $X_1$to X_n, where $X_n$has maximum amplitude).

International Journal of Scientific & Engineering Research, Volume 13, Issue 2, February 2022
ISSN 2229-5518

- Inter-quartile range (IQR):

The IQR presents a measure of spread in a data set and is also known as various dispersion measurements. It uses the definition of the median instead of using the mean. To produce the difference between the quarter and three-quarters value, the median of the lower half (or lower quartile) is calculated and deducted from the median of the upper half (or upper quartile). It shows data distributed to either side of the median. It is the difference between Quantile75 and Quantile25 of a signal developed as IQR= Quantile75 - Quantile25

- Spectral entropy:

Entropy can be interpreted as a measure of uncertainty about an event at frequency. Spectral entropy uses the Fourier transformation method, in which the power spectral density (PSD) can be obtained. The PSD represents the distribution of power as a function of frequency.

- Skewness:

The skewness is the degree of asymmetry of a particular distribution. When the data distribution is symmetrical, skewness is nearby zero. Positive skewness suggests an asymmetric tail extension distribution-Positive value toward. Negative skewness means a distribution with an asymmetric tail that extends to a more negative value. Equation (5) is used for calculating the skewness.

$$\frac{\frac{1}{n}\sum_{n=1}^{N}(x_n-\mu)^3}{\sigma^3}$$

- Kurtosis:

Kurtosis is the relative peak- ness or flatness of a distribution compared with normal distribution. Kurtosis to normally distributed data Are zero. Good kurtosis shows a comparatively small distribution. Negative kurtosis indicates a relatively flat distribution. As with skewness, if the value of kurtosis is too big or too small, there is concern about the normality of the distribution. The formula for computing kurtosis is given in Eq. (6) as

$$\frac{\frac{1}{n}\sum_{n=1}^{N}(x_n-\mu)^4}{\sigma^4}$$

- Dominant frequency analysis.

The most common application is the use of dominant frequency (DF) analysis for estimating atrial activation rates. The dominant frequency is the frequency of the sinusoidal signal with maximum amplitude. This sinusoidal waveform also is the one that best approximates a signal. If the signal is strongly periodic, an approximate waveform in morphology, the dominant frequency will in most cases be associated with the rate of the signal. The DF includes three sub-parameters as following:

1- DF value is the frequency at which the maximum of the spectrum occurs (Hz).
2- DF magnitude is the amplitude value of the DF value.
3- DF ratio is the ratio of the energy of the maximum to the total energy.

- Mel Frequency Cepstral Coefficients (MFCC).

Mel Frequency Cepstral Coefficient (MFCC): is a common and efficient technique for signal processing. The cause of using the first 13 of MFCC (lower dimensions) representing the envelope of spectra. The discarded higher dimensions express the spectral details. For different sounds, envelopes are enough to represent the difference, so we can recognize phonemes through MFCC.

3.2 NARX neural network

In this study, NARX is a nonlinear autoregressive exogenous input neural network (NARX model) was used. The internal architecture perceived as the Multi-Layer Perceptron (MLP) is used in NARX neural network model. The MLP provides a powerful structure that enables us to learn any type of continuous nonlinear mapping. NARX is a reliable predictor of time series [22][23][24]. Such as any neural network, the NARX networks consist of an input layer, a hidden layer, and an output layer. In addition to some efficient component neurons, activation or transfer functions, scaler weights, bias, feedback connections, and tapped delay lines (i.e., memory) [26][27]. Based on theory, NARX networks can be used instead of recurrent networks with comparable computational efficiency. It is used efficiently in time series modeling because it has a flexible structure that combines simplicity and time series prediction. Moreover, they have proven to be much more efficient than other neural networks, to converge more rapidly and to spread more efficiently.

For the purpose of achieving the perfect performance of the NARX neural network for nonlinear time series prediction, it is necessary to use its memory ability using the past values of predicted or true-time series, there are two ways of prediction using the NARX model as shown in Figure 4. The first one based on the actual values of output, which is called is series-parallel architecture or (open-loop network). Furthermore, the other one is based on the estimated values of output, which is called parallel architecture or (close-loop network). They are represented mathematically, respectively, as follows:

$$\hat{y}(t+1) = f\left(y(t), y(t-1), \dots, y(t-n_y), x(t+1),\ x(t), x(t-1), \dots, x(t-n_x)\right) \quad (1)$$

$$\hat{y}(t+1) = f\left(\hat{y}(t), \hat{y}(t-1), \dots, \hat{y}(t-n_y), x(t+1), x(t), x(t-1), .., x(t-n_x)\right) \quad (2)$$

where $f(\cdot)$ is the mapping function of the neural network, yˆ (t + 1) is the output of the NARX at time t for the time t + 1 (it is the predicted value of y for the time t + 1). yˆ(t), yˆ (t − 1) . . .Yˆ (t − n_y) are the past outputs of the NARX. Y(t), y (t − 1) . . .Y (t − n_y) are the true past values of the time series, called also desired output values. X (t + 1), x(t). . .X (t − n_y) are the inputs of the NARX. n_x is the number of input delays and n_y is the number of output delays.

In the series-parallel architecture, the future value of the time series y (t − 1) is predicted from the present and past values

International Journal of Scientific & Engineering Research, Volume 13, Issue 2, February 2022
ISSN 2229-5518

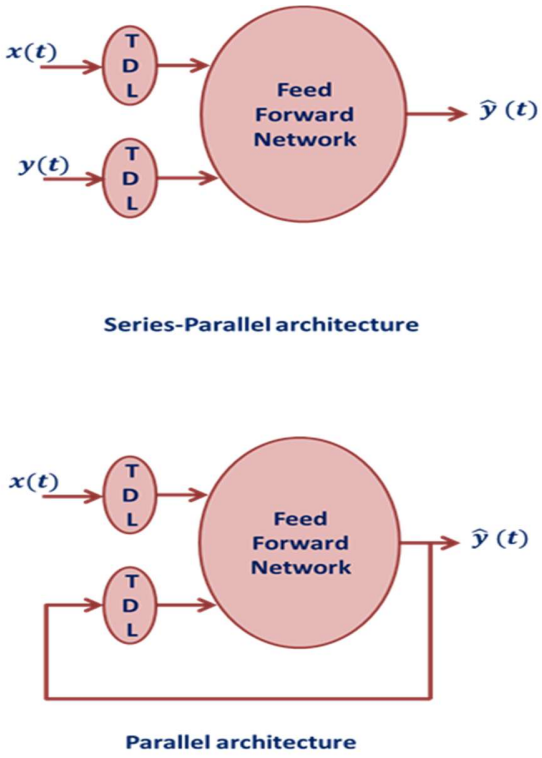

Fig. 2. NARX Architectures

of x(t) and the true past values of the time series y(t). In the parallel architecture, the prediction is performed from the present and past values of x(t) and the past predicted values of the time series y(t).

The open -loop network is used in this study during the training process because of the availability of the true past values of the time series. The application of the open-loop network has two benefits. The First, the use of true values for the feedback network input is greater accurate. The second benefit is a network architecture that is pure network Usual Multi-Layer Perceptron (MLP) network-training algorithms can be used feedforward. The neural NARX networks are converted to closed- loop network, which is beneficial for multi-step-ahead prediction after the training process [28,29].

The mapping function $f(.)$ is initially undefined and it is applied during the training process of the prediction. In the NARX neural network model, the internal architecture that implements this approximation is the Multi-Layer Perceptron (MLP). The MLP makes it possible to learn some form of continuous nonlinear mapping because it provides an effective structure. The input and output delays (feedback delay) have played an important role in enhancing the performance of the NARX model for the classification of the heart sound.

3.3 NARX Training optimization algorithms

The training process is applied to determine the appropriate weights and bias values. This is performed to minimize the overall error function between the network's output and the desired target with associated weights for the our NARX model. The training task is similar to reducing a loss function, which is a measure of how often the NARX model works in a classification test. The intuitive method of training the model consists of three phases: (1) startup of weights and biases, (2) model evaluation based on estimated weights and biases, and the loss function, and (3) updating of estimated weights and biases in the path of loss function reductions. If the loss function's limits as small as possible, it will be moderate. In this instance, our network performs brilliantly. The training optimization algorithm uses backpropagation to produce gradients, which are then used by the training optimization algorithm to minimize the loss function. Due to the fact that there are a variety of loss functions, they are all the same inherently reward us depending on the distance between a specific value's anticipated value and the actual value in our dataset. A popular variety of loss function is the mean squared error (MSE). This error range may be simply calculated by adding up all of the errors, dividing their lengths, and calculating the average. Scaled Conjugate Gradient (SCG), Levenberg-Marquardt (LM), and Bayesian (BR) are the three training optimizations employed as following;

Scaled Conjugate Gradient (SCG): In comparison with the Conjugate Gradient Descent algorithms, the gradient descent algorithm updates the weights and biases along the steepest descent path but is typically associated with a low convergence rate [30]. The Conjugate Gradient (CG) algorithm is the modified variant of the steepest descent algorithm. In the conjugate gradient algorithm, a search is performed along such a direction that produces a faster convergence than the steepest descent direction, while preserving the error minimization achieved in all previous steps. This direction is called the conjugate direction. In most of the CG algorithms the step-size is adjusted at each iteration. A search is made along the conjugate gradient direction to determine the step size, which will minimize the performance function along that line.

Levenberg Marquardt (LM): the LM algorithm first learning algorithm was initially introduced by Kenneth Levenberg and Donald Marquardt. The main reason to use this algorithm is to reduce the error function more effective and it is said to converge rapidly with substantial use in neural network fields. LM training algorithm is highly effective when training networks reach a few hundredweights. The Levenberg-Marquardt algorithm combines two minimization methods: the gradient descent method and the Gauss-Newton method [31] [32] [33]. In the gradient descent method, the sum of the squared errors is reduced by updating the parameters in the steepest-descent direction. In the Gauss-Newton method, the sum of the squared errors is reduced by assuming the least-squares function is locally quadratic and finding the minimum of the quadratic. The Levenberg-Marquardt method acts more like a gradient-descent method when the parameters are far from their optimal value and acts more like the Gauss-Newton method when the parameters are close to their optimal value.

Bayesian Regularization (BR): the BR-training algorithm is updating the weights and bias values according to LM optimization and introduces network weights into the training objective function. It minimizes a combination of squared errors and weights and then determines the correct combination to produce a network that generalizes well. [34]. The BR-training

International Journal of Scientific & Engineering Research, Volume 13, Issue 2, February 2022
ISSN 2229-5518

algorithm is one of the best approaches to overcome the over-tendencies in (ANNs) synthesis so that its prediction accuracy can be further enhanced for invisible data. This approach reduces the problem of over-fitting by considering the goodness of the appropriate fit as well as the structure of the network [35][36].

4 Experimental Results and Discussion

A NARX model is developed as a binary classifier to identify the classification of PCG signals if it was normal or abnormal by using NARX series-parallel architecture (open-loop network) for training and NARX parallel architecture (close-loop network) for testing in addition to studying the effect of three training optimization, which are Levenberg – Marquardt (LM), Bayesian regularization (BR) and Scaled Conjugate Gradient (SCG) techniques. Efficient combination of parameters that can be assessed for the classification process are essential for the implementation of the proposed model. Table 2. describes three different combinations of parameters selected for the investigation of NARX network for such classification task.

4.1 Network Setup

When implementing ANNs for the training process, the first step is to find an effective combination of activation function and optimization algorithm. Among different variations that can be assessed for the classification challenge, we have listed the following one (see table II).

Activation function: the hyperbolic tangent (*tansig*) activation function in the term of neural networks, is like a bipolar sigmoid function with 1 to +1 output range. It is mathematically equivalent to tanh. Although sigmoid runs faster than tanh, there are very small numerical differences between them. As a matter of tradeoff between the speed and accuracy in the network, sigmoid is preferred where speed is more significant than the precise shape of the activation function.

Loss function: the performance of training can be evaluated using several parameters, including recognition accuracy, speed of training, correctness. Among these parameters, the mean squared error is the most important one and it is defined by

$$MSE = \frac{1}{N}\sum_{n=1}^{N} T(n) - Y(n)$$

where T(n) is the target and y(n) is the predicted output

Optimization algorithm: there are several types of learning algorithms available in ANNs. The optimization-training techniques are used to obtain a small error using the backpropagation training algorithm. Three training optimizations are used, which are Scaled Conjugate Gradient (SCG), Levenberg-Marquardt (LM), and Bayesian regularization (BR).

These algorithms find the minimum of a multivariate function that can be expressed as the sum of squares of non-linear real-valued functions.

According to this algorithm, an iterative process is used to reduce the performance function in each iteration. Due to this iterative property train, BR is considered the fastest training algorithm for networks with moderate size.

The number of neurons and ($\boldsymbol{n_x}$, $\boldsymbol{n_y}$): The combinations of hidden neurons and delays produce the lowest average mean.

TABLE 2
DETAILS OF THE PROPOSED MODEL

Network	Train BR	Train LM	Train SCG
Activation Function	TANSIG	TANSIG	TANSIG
Performance function	MSE	MSE	MSE
No. hidden layers	1	1	1
No. neurons	20or 30	20 or 30	20 or 30
No. delays (n_x, n_y)	4or 6	4or 6	2,4or 6

4.2 Performance metrics

In the evaluation of classification systems, classification performance is perhaps the most important factor. Essentially, it should be a count of how many signal instances were successfully classified against how many were mistakenly labelled. There are two types of errors that can arise. A false negative (FN) refers to the total number of PCG signals from pathological hearts that are categorized as normal, whereas a false positive (FP) relates to the total number of PCG signals from normal hearts that are identified as abnormal. in just the same way that FP and FN are defined, the true positives (TP) are clearly diagnosed abnormal heart sounds, while true negatives (TN) are correctly classified normal heart sounds. The classification performance of PCG signals is commonly assessed using the above parameters by computing the sensitivity (Sens.), the specificity (Spec.), and the accuracy (Acc.) as [37].

(SE) = TP / (TP + FN).
(SP) = TN / (TN + FP).
(ACC) = (TP + TN) / (TP + FP + FN + TN).

4.3 Dataset

The datasets that have been discussed in this paper, as well as a few other studies listed, are taken from the Physio Net 2016 challenge, which is publicly available on the website [38].

The total number of extracted 27-length feature vectors of the recordings is 6316 with 3158 vectors of healthy hearts and 3158 of unhealthy hearts. 80/20 and 70/30 percentages of the total number of feature vectors were used to train/test the three networks.

4.4 Results

Table 3 reports the classification results of testing PCG signals for the percentage of training/testing split of data, of 70/30. The NARX architecture model with the number of neurons equals 30 and the number of input and output delays is 6, performs better than the other architecture with the number of neurons and delays ($\boldsymbol{n_x}$, $\boldsymbol{n_y}$) is equal to 20 and 4, respectively.

In general, the NARX classifier provides comparable performance when trained with the BR and LM algorithms. However, it gives low performance values when trained with the SCG algorithm. The optimum performance records achieved by the NARX model are 0.9382, 0.8924, and 0.9153 of sensitivity, specificity, and accuracy, respectively, for the BR training algorithm with number of neurons is 30 and number of delays ($\boldsymbol{n_x}$, $\boldsymbol{n_y}$) equals 6.

International Journal of Scientific & Engineering Research, Volume 13, Issue 2, February 2022
ISSN 2229-5518

TABLE 3
THE CLASSIFICATION RESULTS

Network	# neurons = 30, n_x= n_y=6			# neurons = 20, n_x= n_y=4		
	Se	SP	ACC	Se	SP	ACC
Train BR	0.9382	0.8924	0.9153	0.9100	0.8631	0.8865
Train LM	0.91310	0.8700	0.900	0.8521	0.8321	0.8365
Train SCG	0.8400	0.8109	0.8254	0.7354	0.7136	0.7245

Figures 3 and 4 show the confusion matrices for the training/testing split of data, of 80/20. The confusion matrix is a technique of summing up the classification performance. If you have an unequal number of observations in each class or if you have more than two classes in your dataset, classification accuracy alone can be misleading.

	Normal	Abnormal	
Normal	True Negative 630	False Positive 2	Specificity (SP) 0.99684
Abnormal	False Negative 1	True Positive 631	Sensitivity (SE) 0.99842

(A) BR algorithm

	Normal	Abnormal	
Normal	True Negative 618	False Positive 14	Specificity (SP) 0.97784
Abnormal	False Negative 12	True Positive 620	Sensitivity (SE) 0.98101

(B) LM algorithm

	Normal	Abnormal	
Normal	True Negative 591	False Positive 41	Specificity (SP) 0.93512
Abnormal	False Negative 35	True Positive 597	Sensitivity (SE) 0.94462

(C) SCG algorithm

Fig.3 Confusion matrix at # neurons = 30, # delays ($\boldsymbol{n_x}$, $\boldsymbol{n_y}$) = 6.

	Normal	Abnormal	
Normal	True Negative 618	False Positive 14	Specificity (SP) 0.9778
Abnormal	False Negative 12	True Positive 620	Sensitivity (SE) 0.9810

(A) BR algorithm

	Normal	Abnormal	
Normal	True Negative 612	False Positive 20	Specificity (SP) 0.9683
Abnormal	False Negative 17	True Positive 615	Sensitivity (SE) 0.9731

(B) LM algorithm

	Normal	Abnormal	
Normal	True Negative 586	False Positive 46	Specificity (SP) 0.9272
Abnormal	False Negative 42	True Positive 590	Sensitivity (SE) 0.9335

(C) SCG algorithm

Fig.4 Confusion matrix at # neurons = 20, # delays ($\boldsymbol{n_x}$, $\boldsymbol{n_y}$) =4.

Calculating the confusion matrix will give you a clearer understanding of what's going right in your classification process and what types of errors it makes. As observed in the figures, the classification performance for the training/testing split of data of 80/20 follows the same trend of the classification performance for the training/testing split of data of 70/30.

Figure 5. depicts the performance accuracy of the NARX model when trained with the selected optimization algorithms, in the case of dividing the training/testing data into a ratio of 80/20. The classification accuracy of the model with the number of neurons equals 30 and the number of input and output delays ($\boldsymbol{n_x}$, $\boldsymbol{n_y}$) is 6, is much better than with the number of neurons equals and the number of input and output delays is 20 and 4, respectively. Moreover, the accuracy of the NARX classifier when trained with the BR is comparable to when trained with LM algorithm. However, lower accuracy is regarded when the classifier when is trained with the SCG algorithm.

International Journal of Scientific & Engineering Research, Volume 13, Issue 2, February 2022
ISSN 2229-5518

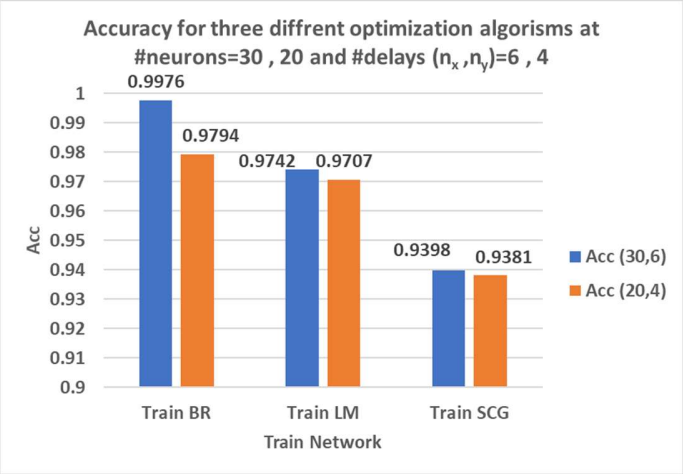

Fig.5 the performance accuracy of the NARX model.

5 Conclusion

In this paper, we suggested an empirical study on optimization algorithms for training the NARX network for the classification of heart sound signals. We focus on three alternative algorithms, namely the Scaled Conjugate Gradient (SCG), the Levenberg- Marquardt (LM), and the Bayesian Regularization (BR). In this approach, the NARX classifier showed better performance when trained with the BR algorithm in comparison to the other two algorithms. This study was carried out with the help of heart sound recordings were collected from Physio Net 2016 data. After the classification result conducted, it was shown that when the number of neurons and delays ($\boldsymbol{n_x}$, $\boldsymbol{n_y}$) equal 30 and 6, respectively, it was better than at they equal 20 and 4, in the case of dividing data into a ratio of 80:20 and a ratio of 70:30.

References

[1] F. Castells, P. Laguna, L. Sornmo, A. Bollmann, J. Millet-Roig, Principal ¨ component analysis in ecg signal processing, EURASIP Journal on Advances in Signal Processing 2007 (2007) 1–21.

[2] L. M. Sepulveda-Cano, E. Gil, P. Laguna, G. Castellanos-Dom´ ´ınguez, Selection of nonstationary dynamic features for obstructive sleep apnoea detection in children, EURASIP Journal on Advances in Signal Processing 2011 (2011) 1–10.

[3] G. D. Clifford, C. Liu, D. B. Springer, B. Moody, Q. Li, R. C. Abad, J. Millet, I. Silva, A. E. W. Johnson, and R. G. Mark, "Classification of normal/abnormal heart sound recordings: the physionet/computing in cardiology challenge 2016," in Computing in Cardiology, CinC 2016, Vancouver, Canada, September 11-14, 2016.

[4] A.F. Gündüz and A. Karci "Heart Sound Classification for Murmur Abnormality Detection Using an Ensemble Approach Based on Traditional Classifiers and Feature Sets"No:1 2020.

[5] H. Tang, H. Chen, T. Li, and M. Zhong, "Classification of normal/abnormal heart sound recordings based on multi-domain features and back propagation neural network," in 2016 Computing in Cardiology Conference (CinC). Computing in Cardiology, sep 2016.

[6] M. Abdollahpur, A. Ghaffari, S. Ghiasi, and M. J. Mollakazemi, "Detection of pathological heart sounds," Physiological Measurement, vol. 38, no. 8, pp. 1616–1630, jul 2017.

[7] M. Zabihi, A. B. Rad, S. Kiranyaz, M. Gabbouj and Aggelos K. Katsaggelos "Heart Sound Anomaly and Quality Detection using Ensemble of Neural Networks without Segmentation".

[8] M. Tschannen, T. Kramer, G. Marti, M. Heinzmann, and T. Wiatowski, "Heart sound classification using deep structured features," in 2016 Computing in Cardiology Conference (CinC). Computing in Cardiology, sep 2016.

[9] V. Maknickas and A. Maknickas, "Recognition of normal–abnormal phonocardiographic signals using deep convolutional neural networks and mel-frequency spectral coefficients," Physiological Measurement, vol. 38, no. 8, pp. 1671–1684, jul 2017.

[10] J. Rubin, R. Abreu, A. Ganguli, S. Nelaturi, I. Matei, and K. Sricharan, "Recognizing abnormal heart sounds using deep learning," CoRR, vol. abs/1707.04642, 2017.

[11] Zhang W. J., Han J. Q. Towards heart sound classification without segmentation using convolutional neural network. 2017 Computing in Cardiology; 2017; Rennes, France. IEEE;

[12] T. Nilanon, J. Yao, J. Hao, S. Purushotham and Y. Liu "Normal / Abnormal Heart Sound Recordings Classification Using Convolutional Neural Network "Computing in Cardiology 2016; VOL 43.

[13] O. Mokhlessi, H. M. Rad, N. Mehrshad, A. Mokhlessi "Application of Neural Networks in Diagnosis of Valve Physiological Heart Disease from Heart Sounds" " 2012 Scientific & Academic Publishing.

[14] M. Deng, T. Meng, J. Cao, S. Wang, J. Zhang and H. Fan "Heart sound classification based on improved MFCC features and convolutional recurrent neural networks "Epub 2020 Jun 23.

[15] S. Latif, M. U. Usman, J. Qadir, and R. Rana, "Abnormal heartbeat detection using recurrent neural networks," CoRR, vol. abs/1801.08322, 2018.

[16] S. Alam, R. Banerjee, and S. Bandyopadhyay, "Murmur detection using parallel recurrent & convolutional neural networks," CoRR, vol. abs/1808.04411, 2018.

[17] Abduh Z., Nehary E. A., Wahed M. A., Kadah Y. M. Classification of heart sounds using fractional fourier transform based mel-frequency spectral coefficients and stacked autoencoder deep neural network. Journal of Medical Imaging and Health Informatics. 2019;9(1):1–8. doi: 10.1166/jmihi.2019.2568.

[18] Yaseen, Son G. Y., Kwon S. Classification of heart sound signal using multiple features. Applied Sciences-Basel. 2018;8(12):p. 2344. doi: 10.3390/app8122344.

[19] M.R.Hadi, M.Y.Mashor and M.S.mohamed" Classification of Heart Sounds using Wavelets and Neural Networks"ICEEE.2008.

[20] S. Khaled, M. Fakhry, A. Mubarak, Classification of pcg sigsnals using a nonlinear autoregressive network with exogenous inputs (narx), Proceedings of 2020 International Conference on Innovative Trends in Communication and Computer Engineering, ITCE 2020 (2020) 98–102.

[21] A. Lerch, An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics, Wiley-IEEE Press, 2012.

[22] Mohanty, S.; Patra, P.K.; Sahoo, S.S. Prediction of global solar radiation using nonlinear autoregressive network win exogenous inputs (narx). In Proceedings of the 39th National System Conference (NSC), Noida, India, 14–16 December 2015.

[23] Pisoni, E.; Farina, M.; Carnevale, C.; Piroddi, L. Forecasting peak air pollution levels using NARX models. Eng. Appl. Artif. Intell. 2009, 22, 593–602. [CrossRef].

[24] Ruiz, L.G.B.; Cuéllar, M.P.; Calvo-Flores, M.D.; Jiménez, M.D.C.P. An Application of Non-Linear Autoregressive Neural Networks to Predict Energy Consumption in Public Buildings. Energies 2016, 9, 684. [CrossRef].

[25] S. A. Billings, Nonlinear system identification NARMAX methods in the time,

International Journal of Scientific & Engineering Research, Volume 13, Issue 2, February 2022
ISSN 2229-5518

frequency, and spatio-temporal domains. Chichester, West Sussex Wiley, 2013.

[26] P. Sugunnasil, S. Somhom, W. Jumpamule, and N. Tongsiri, "Modeling a neural network using an algebraic method," ScienceAsia, vol. 40, no. 1, pp. 94–100, 2014.

[27] Cadenas, E.; Rivera, W.; Campos-Amezcua, R.; Heard, C. Wind Speed Prediction Using a Univariate ARIMA Model and a Multivariate NARX Model. Energies 2016, 9, 109. [CrossRef].

[28] Ferreira, A.A.; Ludermir, T.B.; Aquino, R. Comparing recurrent networks for time-series forecasting. In Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN), Brisbane, Australia, 10–15 Junuary 2012.

[29] Buitrago, J.; Asfour, S. Short-Term Forecasting of Electric Loads Using Nonlinear Autoregressive Artificial Neural Networks with Exogenous Vector Inputs. Energies 2017, 10, 40. [CrossRef].

[30] M. F. MpUer. A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks. To be pubhshed.

[31] M.I.A. Lourakis. A brief description of the Levenberg-Marquardt algorithm implemented by levmar, Technical Report, Institute of Computer Science, Foundation for Research and Technology - Hellas, 2005.

[32] K. Madsen, N.B. Nielsen, and O. Tingleff. Methods for nonlinear least squares problems. Technical Report. Informatics and Mathematical Modeling, Technical University of Denmark, 2004.

[33] D.W. Marquardt. "An algorithm for least-squares estimation of nonlinear parameters," Journal of the Society for Industrial and Applied Mathematics, 11(2):431-441, 1963.

[34] Demuth, H.B.; Beale, M.H.; De Jess, O.; Hagan, M.T. Neural network design. 2nd edition. Martin Hagan, 2014.

[35] MacKay, D.J.C. A Practical Bayesian Framework for Backpropagation Networks. Neural Comput. 4(3), 1992, 448–472. doi:10.1162/neco.1992.4.3.448.

[36] Burden, F.; Winkler, D. Bayesian Regularization of Neural Networks. In Methods Mol. Biol. Humana Press, 2008, 23–42. doi:10.1007/978-1-60327-101-1 3.

[37] C. Liu, D. J. Springer, Q. X. Li, B. Moody, R. Juan, F. J. Chorro, F. Castells, J. M. Roig, I. Silva, A. E. W. Johnson, Z. Syed, S. E. Schmidt, C. D. Papadaniil, L. Hadjileontiadis, H. Naseri, A. Moukadem, A. Dieterlen, C. Brandt, H. Tang, M. Samieinasab, M. R. Samieinasab, R. Sameni, R. G. Mark, G. D. Clifford, An open access database for the evaluation of heart sound algorithms., Physiological measurement 37 12 (2016) 2181–2213.

[38] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, ˆ A. de Mendonc¸a, Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests, BMC Research Notes 4 (2011) 299 – 299.