2010 IEEE EMBS Conference on Biomedical Engineering & Sciences (IECBES 2010), Kuala Lumpur, Malaysia, 30th November - 2nd December 2010.

Application of Modeling Techniques to Diabetes Diagnosis

A. M. Aibinu, M. J. E. Salami and A. A. Shafie
Department of Mechatronics Engineering
International Islamic University Malaysia (IIUM)
P. O. Box 53100 Gombak, Malaysia.
E-mail: maibinu@gmail.com

In recent times, the introduction of complex-valued neural networks (CVNN) has widened the scope and applications of real-valued neural network (RVNN) and parametric modeling techniques. In this paper, new expert systems for automatic diagnosis and classification of diabetes using CVNN and RVNN based parametric modeling approaches have been suggested. Application of complex data normalization technique converts the real valued input data to complex valued data (CVD) by the process of phase encoding over unity magnitude. CVNN learn the relationship between the input and output phase encoded data during training and the coefficients of Complex-valued autoregressive (CAR) model can be extracted from the complex-valued weights and coefficients of the trained network. Classification of the obtained CAR or RVAR model coefficients results in required distinct classes for diagnosis purpose. Similar operations can be performed for real-valued autoregressive technique except for CVD normalization. The effect of data normalization techniques, activation functions, learning rate, number of neurons in the hidden layer and the number of epoch using the suggested techniques on PIMA INDIA diabetes dataset have been evaluated in this paper. Results obtained compares favorably with earlier reported results.

Keywords: Complex-Valued Autoregressive (CAR) Model, Complex-Valued Neural Network (CVNN), Diabetes, Neurons, Parametric modeling techniques.

1. Introduction

Biomedical signals classification can be defined as categorization of the input data into distinct classes after the extraction of significant features of the data from a background of other irrelevant detail [1]. In biomedical signals classification, the objective is mainly to group the observed or recorded signals into regions with same properties or characteristics, thus generating decision boundaries to separate the dataset into classes based on the feature vectors. In almost every fields, features classification plays a significant role in facilitating the description of anatomical structures and other regions of interest into similar groups for further analysis and examination. Typical example includes diagnostic check so as to know whether a patient has diabetes disease or not using recorded signals from such a patient. Consider a two-dimensional feature vector shown in Figure 1, where c_h and c_d are the clusters for healthy and disease data set respectively, an optimal linear autoregressive (AR) model decision function $y(n)$ is the perpendicular bisector of the straight line joining c_h and c_d [1].

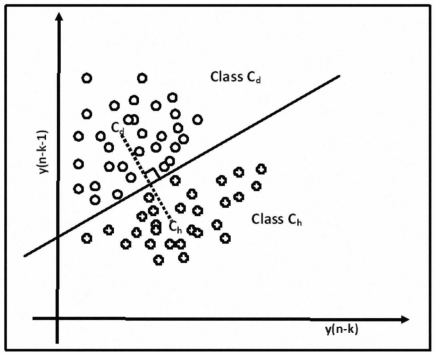

Fig. 1. ANN based Biomedical signals classification technique

A linear AR model decision function is given as

$$y(n) = -a_1 y_1(n) - a_2 y_2(n) - \cdots - a_p y_p(n)$$
$$= -\sum_{k=1}^{p} a_k y_k(n) \quad (1)$$
$$= Y^T A \quad (2)$$

where $Y = [y_1(n), y_2(n), ...y_p(n)]$ is the feature vector, $A = [a_1, a_2 \ldots a_p]$ is the corresponding autoregressive coefficient vector of order p obtained from the trained network and $y(n)$ is the AR model decision function for signal classification. Thus, a two-class biomedical signal pattern classification is given as

$$Y^T A \begin{cases} = 1; & if\ Y \in C_h \\ = 0; & if\ Y \in C_d \end{cases} \quad (3)$$

where C_h and C_d represent the healthy and disease classes respectively. Thus, the AR model output $y(n)$ is a discriminant function that creates the boundary separating the two classes, C_h and C_d.

In this paper, the application of the suggested modeling techniques on PIMA diabetes data diagnosis (using data and some other biomedical signals obtained from female subjects, [8]) have been evaluated. The problem represent a diverse cross section of biomedical signals where the suggested modeling techniques can effectively facilitate clinical diagnosis.

2. Methodology

The block diagram of the suggested artificial neural network (ANN) based autoregressive model classification is shown in Figure 2. Detailed explanation of each of the units and the performance analysis of the proposed technique are subsequently discussed.

978-1-4244-7600-8/10/$26.00 ©2010 IEEE

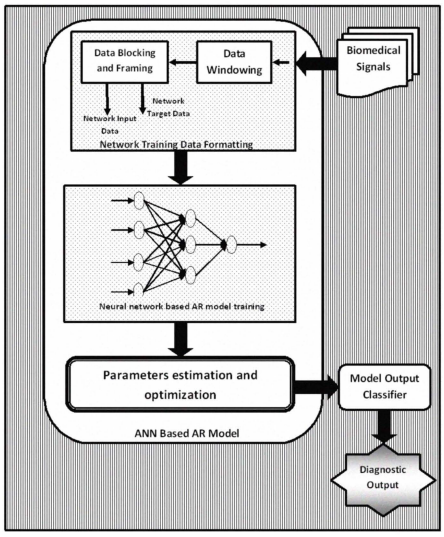

Fig. 2. Flow diagram of Biomedical signals classification using ANN based autoregressive modeling techniques

1) **Data formatting and Normalization**
The first stage in the suggested ANN-based AR signal classification technique is the data formatting and normalization. Data formatting is concerned with structural representation of the data in meeting the suggested model format while data normalization is concerned with band limiting the measured signals or data. Two types of biomedical data have been studied in this work, namely real-valued data (RVD) and complex-valued data (CVD). Application of complex data normalization technique converts RVD to CVD by phase encoding of the input RVD from 0 to π with unity magnitude radius. Other data normalization techniques evaluated in this work include:

- **z-score data normalization**

$$X_{new} = \frac{x - \bar{x}}{\sigma_x^2} \tag{4a}$$

- **Min-Max data normalization**

$$X_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{4b}$$

- **Complex data normalization**

$$\theta = \frac{\pi(x - x_{min})}{x_{max} - x_{min}} \tag{4c}$$

and

$$X_{new} = e^{i\theta} = \cos\theta + i\sin\theta \tag{4d}$$

- **Unitary data normalization**

$$X_{new} = \frac{x}{x_{max}} \tag{4e}$$

where x_{min} and x_{max} are the minimum and maximum values of the data respectively, $\bar{x}$ is the mean value of the dataset and σ_x^2 is the variance of the dataset [5]. The above normalization techniques are henceforth referred to as NT1, NT2, NT3 and NT4 respectively in this paper.
Consider a real value signal x where $x \in \Re^m$, complex data normalization (NT3) involves mapping from real valued domain to complex valued domain, that is $\Re^m \rightarrow \mathbb{C}^m$. Therefore, NT3 data normalization technique maps any RVD within the interval $[x_{min}\ ,\ x_{max}]$ to a complex valued interval $[0\ ,\ \pi]$ using (4c) and (4d), thus variation in RVD value has been transformed to variation in the phase angle θ over the unit circle. NT3 data normalization type retains the relational and spatial properties that exist between two entities in RVD in the transformed complex domain [26]. For example if in RVD, $x_1 < x_2$, the relationship is still valid in the transformed CVD domain, that is $X_{new1} < X_{new2}$. Similarly, the spatial distance between two extremes $[x_{min}\ ,\ x_{max}]$ in RVD is obtainable between the two extremes $[0\ ,\ \pi]$ in CVD.

2) **ANN-based parametric classification technique**
The detailed mathematical derivation, analysis and development of the proposed ANN-based AR model classification technique for biomedical signal classification and diagnosis have been proposed in [2]–[4]. Detail mathematical derivation of the coefficients are well covered in [2]–[4], [6], [7].

3) **Model output classifier**
The main objective of the model output classifier is to group the obtained model coefficients into regions with same property or characteristics. The output of this unit is the required diasnostic results.

4) **Performance measures**
The performance analysis used in this paper are as defined in [9]. Some of which are:

- True Positive(TP): If an output is positive (P) and the network classified it as positive (P), then it is counted as a true positive.
- True Negative (TN): If an output is negative (N) and the network classified it as negative (N), such an output is regarded as true negative.
- False Positive (FP): If an output is negative (N) and the network classified it as positive (P), such an instance is regarded as false positive.
- False Negative (FN): If an output is positive (P) and the network classified it as negative (N), such is regarded as false negative.
- The accuracy of the classifier is one of the most important measures in classification model. It is the degree of closeness of a measured or calculated quantity to its actual (true) value [9]. Accuracy is defined as:

$$Accuracy\ (ACC) = \frac{TP + TN}{P + N}$$

- Time of Completion : The time of completion (TOC) refers to the time taken for the execution of a particular task.

3. Results and Discussion

The main objective of application of the suggested ANN based autoregressive model classification techniques on PIMA India diabetes dataset is to diagnose whether an individual has diabetes or not. The real valued data is based on personal data (age, number of times pregnant) and the results of medical examinations such as blood pressure, body mass index, result of glucose tolerance test performed on the subject. The data has been obtained from UCI learning database (//ftp.ics.uci.edu/pub/machine-learnind autoregressive gdatabases, [8]), it contains 768 instances with a two-class distribution on female subjects.

The dataset features and statical analysis are as contained in Table I and Table II respectively. During training phase,

TABLE I

PIMA Indian diabetes data features

Features	Notation	Unit	Mean	S. deviation
Number of times pregnant	$y_1(n)$		3.8	3.4
Plasma glucose concentration	$y_2(n)$		120.9	32
Diastolic blood pressure	$y_3(n)$	mm Hg	69.1	19.4
Triceps skin fold thickness	$y_4(n)$	mm	20.5	16.0
2-Hour serum insulin	$y_5(n)$	mu $\frac{U}{ml}$	79.8	115.2
Body mass index	$y_6(n)$	$\frac{kg}{m^2}$	32.0	7.9
Diabetes pedigree function	$y_7(n)$		0.5	0.3
Age	$y_8(n)$	Years	33.2	11.8

TABLE II

PIMA Indian diabetes data analysis

Dataset	Total Pattern	Positive (P) C_d	Negative (N) C_h
Training set	460	175	285
Test set	308	93	215
Total	768	268	500

the features vectors $Y = [y_1(n), y_2(n), ..., y_p(n)]$ and classes assigned $[C_h\ ,\ C_d]$ are fed into the suggested ANN-based AR model, upon convergence the network weights and adaptive coefficients of the activation function are extracted for the computation of the required autoregressive coefficients vectors. Only CVNN-based CAR model will be discussed in detailed since similar trends were observed for the real-valued counterpart, moreso CVNN-based approach offers holistic view with capability to process and model RVD.

Evaluated parameters in this work includes the effect of data normalization, number of neurons in the hidden layer, learning rate, number of epoch and activation function on the accuracy of the proposed model. These parameters greatly affect the accuracy of the system hence the need for thorough investigation and analysis. Each entry in the output table represents the mean and variance of 10 different trials using different random initial weights.

1) **Effect of data normalization and Activation function**

The effect of data preprocessing methods and learning rates on the accuracy of the proposed system have been evaluated in this subsection. The proposed system parameters are :

- Model Types :
  CVNN based CAR model and RVNN based AR model govern by equation (2).
- Model order : Number of PIMA data Attributes
- Algorithm : Gradient Descent
- Epoch : 550
- Activation type : ACT1= Hyperbolic Tangent; ACT2 = Sigmoid Function
- Neurons in hidden layer : 5
- Data normalization types : NT1 -NT 4

Typical MSE plot against epoch for the proposed system is shown in Figure 3. The results obtained while evaluating the effect of activation function, learning rate and data normalization on the accuracy of the proposed system are shown in Table III as well as Figure 4. As observed in Table III, the accuracy obtained using hyperbolic tangent is better than that obtained using sigmoid activation function with respect to different data normalization and learning rate. Also noticeable is the excellent performance of

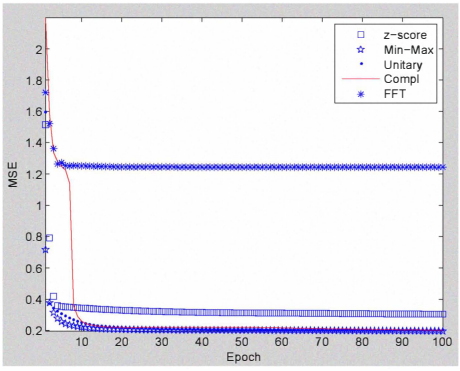

Fig. 3. MSE plot for different normalization methods

TABLE III

Effect of data normalization and learning rate (LR) on PIMA Indian diagnostic accuracy

LR	NT	ACC	$\sigma^2(10^{-3})$	ACC	$\sigma^2(10^{-3})$
		ACT1		ACT2	
0.0101	1	0.71	0.25	0.67	0.18
	2	0.77	0.10	0.69	0.01
	3	0.77	0.08	0.69	0.00
	4	0.77	1.20	0.55	17.24
0.0412	1	0.69	0.17	0.67	1.76
	2	0.76	0.24	0.69	0.00
	3	0.77	0.15	0.69	0.00
	4	0.80	0.13	0.64	13.32
0.0834	1	0.71	2.72	0.68	0.50
	2	0.73	0.30	0.69	0.00
	3	0.72	0.65	0.69	0.00
	4	0.79	0.22	0.69	2.89
0.1194	1	0.71	0.23	0.69	0.13
	2	0.73	0.29	0.69	0.00
	3	0.67	1.53	0.69	0.00
	4	0.77	0.81	0.66	3.63
0.1681	1	0.68	0.65	0.71	0.53
	2	0.66	0.47	0.69	0.00
	3	0.64	0.47	0.69	0.01
	4	0.74	1.00	0.62	13.11

complex domain data normalization technique under different values of learning rate and activation function type 1 (ACT1). This shows that the proposed model would perform better using complex data normalization technique with hyperbolic tangent as the activation function. The use of very low value learning rate increases the time of completion of the algorithm while a high learning rate often leads to instability in the algorithm. Despite the relative poor performance of some data normalization and activation in this section, the accuracy obtained still outperforms some of the reported cases in the literature [15].

2) **Effect of neurons in the hidden layer**

Table IV shows the effect of number of neurons in the hidden layer at learning rate of 0.0412 on the performance of the proposed model. Increasing the number of neurons beyond 5 in the hidden layer has no significant effect on the accuracy of the model for ACT1 using the z-score, complex data and min-max data normalization techniques. Hence, the optimal number of neurons in the hidden layer is selected to be 5. For unitary normalization, increasing the number of neurons in the hidden layer increases the accuracy of the system. Figure 5 shows the effect of increasing the number of neurons in the hidden layer at optimal learning rate and epoch.

3) **Effect of epoch on diagnostic accuracy**

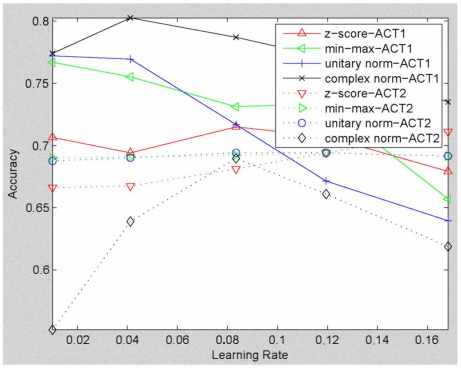

Fig. 4. Accuracy plot for different normalization methods and activation functions (ACT)

TABLE IV

Effect hidden layer neurons on PIMA Indian diagnostic accuracy

Neurons	NT	ACC	$\sigma^2(10^{-3})$	ACC	$\sigma^2(10^{-3})$
		ACT1		ACT2	
2	1	0.76	0.31	0.72	0.17
	2	0.74	0.13	0.69	0.00
	3	0.73	0.27	0.69	0.02
	4	0.73	1.10	0.51	4.65
5	1	0.69	0.17	0.67	1.76
	2	0.76	0.24	0.69	0.00
	3	0.77	0.15	0.69	0.00
	4	0.80	0.13	0.64	13.32
8	1	0.72	0.36	0.68	0.26
	2	0.76	0.21	0.69	0.01
	3	0.77	0.09	0.69	0.00
	4	0.79	0.49	0.66	11.38

Table V shows the effect of epoch on the performance of the proposed model. Based on the obtained results increasing the number of epoch leads to an increase in the time of completion (TOC) of the model without affecting the classification accuracy. Apart from the accuracy, TOC using ACT1 is always greater than TOC of ACT2 for all data normalization techniques. This shows that there is a trade off between accuracy and TOC in using either ACT1 or ACT2.

4. Conclusion

The problem of constructing a plane to separate members of two sets has been reformulated as a parametric AR model problem in which ANN can be used to compute AR model parameters from which the necessary information can be inferred. Though AR modeling approach was originally created to model, reconstruct and represent data sequence in an elegant and parsimonious way. the data is usually

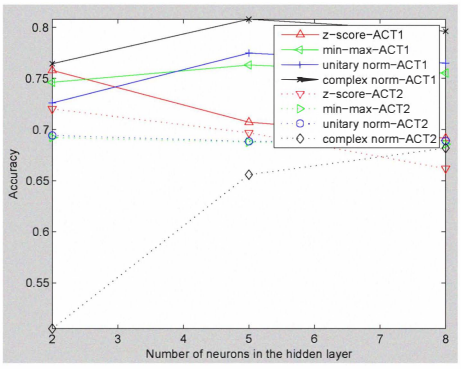

Fig. 5. Accuracy plot for different number of hidden neurons, normalization types and activation functions

TABLE V

Effect of epoch on PIMA Indian diagnostic accuracy

Epoch	NT	ACC	$\sigma^2(10^{-3})$	TOC	ACC	$\sigma^2(10^{-3})$	TOC
		ACT1			ACT2		
150	1	0.71	0.42	0.71	0.67	0.50	0.67
150	2	0.76	0.30	0.76	0.69	0.01	0.69
150	3	0.77	0.17	0.77	0.69	0.00	0.69
150	4	0.81	0.22	0.81	0.66	1.86	0.66
350	1	0.71	1.44	0.71	0.70	0.04	0.70
350	2	0.76	0.23	0.76	0.69	0.01	0.69
350	3	0.77	0.02	0.77	0.69	0.00	0.69
350	4	0.81	0.10	0.81	0.66	10.41	0.66
550	1	0.69	0.17	0.69	0.67	1.76	0.67
550	2	0.76	0.24	0.76	0.69	0.00	0.69
550	3	0.77	0.15	0.77	0.69	0.00	0.69
550	4	0.80	0.13	0.80	0.64	13.32	0.64

considered to be the output excited by white noise or impulse input sequence. However, a novel interpretation and application of modeling technique shows that an optimal linear autoregressive decision function $y(n)$ can be used to construct a perpendicular bisector of the straight line joining two different disjoint classes in a disjointed dataset.

Application of the proposed ANN-based AR model techniques in solving PIMA diabetes diagnosis problem produces results whose accuracy is in the range of $80.65\% - 81.00\%$ whereas application of the same modeling approach to Liver disease classification problem yields result with an average accuracy of 72.83% accuracy . These results compare very favorably with existing classification techniques using these same datasets.

As observed in Table VI, none of the reported work in the literature so far has used the proposed ANN-based modeling approach for diagnosis of diabetes. Though recently there has been an attempt to use CVNN technique for solving this same problem, however the differences in that work with the proposed technique include the use of multilayer network as compared to single layered network reported in [26] and the use of model parameters estimation techniques.

Conventional validation method has been used in this work and has been shown to have significant effect on the accuracy for the same problem [15]. As seen from Table VI, results obtained using this approach with or without optimization performed better than some of the earlier reported cases using the same method of conventional validation (C). Similarly, comparing the results with k-fold cross validation technique (FC), the results in this work are better than most of the reported cases [15]. Thus the results of this work using the modeling approach for PIMA dataset classification is highly encouraging and it has been shown that ANN-based AR can be used to generate decision boundaries to separate PIMA dataset into classes based on the dataset feature vectors.

TABLE VI

PIMA Indian diabetes diagnosis reports

Author	Ref.	Methodology	ACC	Method
Carpenter and Markuzon	[10]	ARTMAP-IC	81.0	C
Kayaer and Yildirim	[12]	GRNN	80.21	C
Kayaer and Yuldirim	[12]	MLNN-GDA	77.60	C
Kayaer and Yuldirim	[12]	MLNN-GDA-MM	76.56	C
Kayaer and Yuldirim	[12]	MLNN-GDA+ADLR	77.60	C
Kayaer and Yuldirim	[12]	MLNN-LM	77.08	C
Termutas et al	[15]	MLNN-LM	82.37	C
Termutas et al	[15]	PNN	78.13	C
This work		RVNN-based AR	80.65	C
This work		CVNN-based CAR	81.00	C
Other results and methods reported in [13]				
Ster and Dobnikar		MLP-GD	76.4	C
Zarndt		MLP-GD	75.8	C
Karol Grudzinski		kNN	75.5	C
Ster and Dobnikar		MLP-GD	76.4	C
Deng and Kasabov	[11]	ESOM	78.40	10FC
Polat and Gunes	[13]	PCA-ANFIS	89.47	10FC
Polat et al	[14]	LS-SNM	78.21	10FC
Polat et al	[14]	GDA-LS-SNM	79.16	10FC

Acknowledgement: This work is supported by Malaysia E-Science Grant : $01-01-08-SF0083$

References

[1] R. M. Rangayyan, " Biomedical Signal Analaysis, A case-study approach", *IEEE Press series*, Canada, 2002.

[2] A. M. Aibinu, M, Nilsson, M. J. E. Salami,, and A. A. Shafie " A new method of Voice Activity Detection Using Real-Valued Neural Network Based Autoregressive modeling Technique ", *submitted for publication in Elsevier Journal, Computer in Biology, Nov 2009.*

[3] A. M. Aibinu, M. J. E. Salami,, and A. A. Shafie "A New Method of Diabetes Diagnosis Using Complex-Valued Neural", *Elsevier,Expert Systems with Applications*, Submitted for Publication, October, 2009.

[4] A. M. Aibinu, M. J. E. Salami,, and A. A. Shafie " New Methods of Teeth Identification Using Neural Network based Autoregressive Technique", *Elsevier,Expert Systems with Applications*, Submitted for Publication, October, 2009.

[5] N. Karthikeyani Visalakshi and K. Thangavel, " Impact of Normalization in Distributed K-Means Clustering," *International Journal of Soft Computing*, 4 (4:) pg 168-172, 2009

[6] A. M. Aibinu, M. J. E. Salami, and A. A. Shafie "Determination of Complex-Valued Parametric Model Coefficients Using Artificial Neural Network Technique", Volume 2010 (2010), Article ID 984381.

[7] A.M Aibinu, M.J.E Salami and A.A.Shafie, "Complex Valued Autoregressive Modeling Technique Using Artificial Neural Network", Biosignal Interpretation, Yale USA, June 2009.

[8] //ftp.ics.uci.edu/pub/machine-learningdatabases.

[9] "T. Fawcett. ROC graphs: Notes and practical considerations for data mining researchers"*Kluwer Academic Publishers.* March 2004.

[10] G. A. Carpenter and N. Markuzon, "ARTMAP-IC and medical diagnosis Instance counting and inconsistent cases", *Neural Networks*, 11, 323336, 1998.

[11] D. Deng, and N. Kasabov, "On-line pattern analysis by evolving self-organizing maps," *Proceedings of the fifth biannual conference on artificial neural networks and expert systems (ANNES)* (pp. 4651), 2001.

[12] K. Kayaer, and T. yaldum, " Medical diagnosis on Pima Indian diabetes using general regression neural networks. In Proceedings of the international conference on artificial neural networks and neural information processing (ICANN ICONIP), (pp. 181184),2003.

[13] K. Polat, and S. Gunes, " An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," *Digital Signal Processing*, 17(4), 702710,(2007).

[14] K. Polat, S. Gunes and A. Asl1an, "A cascade learning system for classification of diabetes disease Generalized discriminant analysis and least square support vector machine", *Expert Systems with Applications*, 34(1), 214221, 2008.

[15] H. Temurtas, N. Yumusak and F. Termutas, "A comparative study on Diabetes disease diagnosis using Neural Network," *Expert Systems with applications,* pp. 8610-8615, 2009.

[16] BUPA Liver Disorders Dataset. UCI Repository of Machine Learning Databases. ftp://ftp.ics.uci.edu/pub/machine-1earning-databases/liverdisorders/bupa.data.

[17] Pham, D. T., Dimov, S. S.and Salem, Z.. Technique for selecting examples in inductive learning. In European symposium on intelligent techniques (ESIT 2000), Aachen, Germany (pp. 119127), 2000.

[18] Cheung, N. . Machine learning techniques for medical analysis. School of Information Technology and Electrical Engineering, B.Sc. thesis, University of Queenland, 2001.

[19] Lee, Y. J., and Mangasarian, O. L. " RSVM: Reduced support vector machines", In Proceedings of the first SIAM international conference on data mining, 2001.

[20] Lee, Y. J., and Mangasarian, O. L. , " SSVM: A smooth support vector machine for classification". Computational Optimization and Applications, 20(1), 522. 2001.

[21] Van Gestel, T., Suykens, J. A. K., Lanckriet, G., Lambrechts, A., De Moor, B., and Vandewalle, J. "Bayesian framework for least squares support vector machine classifiers, Gaussian processes and kernel fisher discriminant analysis," Neural Computation, 14(5), pp 11151147, 2002.

[22] Yalçin, M., and Yildirim, T. . Karacig.er bozukluklar.n.n yapay sinir ag.lar. ile teshisi. In Biyomedikal MuoN hendislig. i Ulusal Toplant.s. (BIYOMUT 2003), Istanbul, Turkey (pp. 293.297, 2003.

[23] Polat, K., Sbahan, S., Kodaz, H., and Gualn S. Biomedicatl analysis,2005.

[24] E. çomak, K. Polat, S. Güneç and A. Arslan, "A new medical decision making system: Least square support vector machine (LSSVM) with Fuzzy Weighting Pre-processing," pp 409-414, 2007.

[25] N. Mehdi,Y. Mehdi yaghobi and N. Mohammad, "Designing an Expert System of Liver Disorder by Using Neural Network and Comparing it With Parametric and Non Parametric System," 2008.

[26] M. FaijulAmin and K. Murase, "Single-layered complex-valued neural network for real-valued classification problems", *Neurocomputing* pp 945-955, 2009.