C-Support Vector Classification: Selection of Kernel and Parameters in Medical Diagnosis

J. Novakovic* and A. Veljovic**
* Megatrend University, Graduate School of Computer Sciences, Belgrade, Serbia
** University of Kragujevac, Technical Faculty in Cacak, Kragujevac, Serbia
jnovakovic@megatrend.edu.rs
alempije@beotel.rs

***Abstract*—This paper investigates the impact of kernel function and parameters of C-Support Vector Classification (C-SVC) to solve biomedical problems in a variety of clinical domains. Experimental results demonstrate the effectiveness of optimizing parameters for C-SVC with different basic kernel. Without optimizing parameters results for classification accuracy with data sets in medical domains shows the best performance of linear kernel. After optimization of parameters, results of classification accuracy are more consistent for all kernel functions, and we no longer have the dominance of certain kernel functions, or larger variance in the results. The biggest benefits of optimization had those kernel functions, which have a smaller accuracy of classification. Results show that time taken to build model are very high with C-SVC and polynomial kernel, compare with others kernels.**

Keywords: C-SVC, classification accuracy, kernel, parameter selection, SVM.

I. INTRODUCTION

In medical diagnosis, machine learning systems providing computational methods for accumulating, changing and updating knowledge in intelligent systems, and in particular learning mechanisms that will help us to induce knowledge from examples or data. These systems successfully analyze medical data giving results used for further treatment. There are several reasons why machine learning systems should be used in medical diagnosis. In medicine, these systems are useful in cases where algorithmic solutions are not available, there is lack of formal models, or expertise in understanding the complex physiological functions is limited. They have the potential to discover new relationships among concepts and hypotheses by examining the record of successfully solved cases and may aggregate knowledge that has yet to be formalized.

In diagnosis of diseases, machine learning systems have found many valuable applications so that appropriate intervention can be exercised to achieve better outcomes [1, 2, 3, 4, 5, 6, 7, 8, 9]. Machine learning in medical application includes a broad class of methods such as artificial neural network, genetic algorithms, probabilistic models, induction of rules, decision trees, statistical or pattern recognition methods, k-nearest neighbors, Bayesian classifiers and discriminate analysis.

For medical diagnosis, Support Vector Machine (SVM) is a popular due to its excellent generalization performance. SVM introduced by Vapnik [10], employs structural risk minimization whereby a bound on the risk is minimized by maximizing the margin between the separating hyperplane and the closest data point to the hyperplane. As supervised learning methods that analyze data and recognize patterns, it rigorously based on statistical learning theory simultaneously minimizes the training and test errors.

While there are a number of different types of SVM, for the purpose of this research, we suggest C-SVC that can incorporate different basic kernels. The goal of this research is to present different basic kernels for constructing and evaluating systems that learn from experience to make the decisions and predictions.

The rest of this paper is organized into four sections. Section 2 presents C-SVC. Section 3 discusses the results and investigates the performance of the proposed technique. Finally, concluding remarks and future research are discussed in section 4.

II. C-SVC

SVM constructs a hyperplane or set of hyperplanes in a high dimensional space, which can be used for classification, regression, or other tasks. Many hyperplanes might classify the data; the best hyperplane is the one that represents the largest separation, or margin, between the two classes. Generally speaking, the larger the margin it is the lower the generalization error of the classifier. We choose, the maximum–margin hyperplane, such the hyperplane in which the distance from it to the nearest data point on each side is maximized.

Vapnik in 1963 proposed original optimal hyperplane algorithm, which was a linear classifier. In 1992, Boser, Guyon and Vapnik [11] suggested a way to create nonlinear classifiers by applying the kernel trick (originally proposed by Aizerman, Braverman and Rozonoer [12]) to maximum-margin hyperplanes. In this algorithm, every dot product is replaced by a nonlinear kernel function, which allows the algorithm to fit the maximum-margin hyperplane in a transformed feature space.

Often happens that in a finite dimensional space the sets to be discriminated are not linearly separable. To make the separation easier, it was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space. Mapping into a larger space, cross products may be computed easily in terms of the variables in the original space, making the computational load reasonable.

In the following text, for the purpose of this research, one type of SVM is explained, C-SVC that can incorporate different basic kernels. Given training vectors

978-1-4577-1974-5/11/$26.00 ©2011 IEEE

$x_i \epsilon R^n, i = 1, \dots, l$, in the two-class case and the corresponding class labels decision $y_i \epsilon \{1, -1\}$, the statement of C-SVC optimization for classification problems may be the following [13, 14]:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

with constraints:

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \dots, l.$$

The dual problem definition is:

$$\min_{\propto} \frac{1}{2} \propto^T Q \propto - e^T \propto, \ 0 \leq \propto_i \leq C, i = 1, \dots, l,$$

with constraints $y^T \propto = 0$, where e is the vector of all ones, C > 0 is the upper bound, Q is a l by l positive semidefinite matrix, $Q_{ij} \equiv y_i y_j K(x_i, x_j)$, and $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is the kernel. Function ϕ transforms training vectors x_i into a higher (maybe infinite) dimensional space.

The decision function is

$$sgn\left(\sum_{i=1}^{l} y_i \propto_i K(x_i, x) + b\right).$$

The choice of the appropriate kernel for a specific application is often a difficult task. A necessary and sufficient condition for a kernel to be valid is that it must satisfy Mercer's theorem, but other than that, there is really no mathematically structured approach to prefer one kernel to the other. Obviously, we would expect that a non-linear kernel based C-SVC would perform better than the one based on a linear kernel, if the data is known to be not linearly separable. The choice of kernel results in different kinds of C-SVCs with different performance levels. For the purpose of this research, linear, polynomial, radial basis function (RBF) and sigmoid kernels are investigated.

A. Linear kernel

We can define the linear kernel as follows:

$$K(x_i, x_j) = x_i^T x_j$$

This is the simplest kernel. It shows good performance for linearly separable data, but surprisingly, works very well even in cases of non-linear data.

B. Polynomial kernel

We can define the polynomial kernel as follows:

$$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0.$$

where *d* is the degree of the polynomial. For vectors x_i that are linearly dependent on *d* dimensions, the kernel function of order *d* can be used to transform them into linearly independent vectors on those *d* dimensions. Vectors x_i are transformed into the dimension space where they become linearly separable and the linear C-SVC case can handle the classification problem. The performance does depend on the order *d* of the polynomial, since how well the data becomes separable depends on it. Because the principle behind these two kernels is the same and the transformation is to take them to different space, the performance of polynomial kernel is expected to be around the same as that of the linear kernel.

C. RBF kernel

We can define the RBF kernel as follows:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0$$

The RBF kernel is suited best to deal with data that have a class-conditional probability distribution function approaching the Gaussian distribution. This kernel maps such data into a different space where the data becomes linearly separable. We would expect of the RBF kernel to perform much better than either the linear or the polynomial kernel, but this kernel is difficult to design, in the sense that it is difficult to arrive at an optimum γ and choose the corresponding *C* that works best for a given problem. Certain combinations of γ and *C* make the C-SVC highly sensitive to training data and contributes to the error rate of the C-SVC with RBF kernel.

The RBF kernel nonlinearly maps samples into a higher dimensional space so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear. In addition, the linear kernel is a special case of RBF [15] since the linear kernel with a penalty parameter C has the same performance as the RBF kernel with some parameters (C; γ). RBF kernel is often first choice compare to polynomial kernel, because the number of hyperparameters influences the complexity of model selection, and the polynomial kernel has more hyperparameters than the RBF kernel.

However, in some situations the RBF kernel is not suitable, for example, when the number of features is very large, the linear kernel may be better solution. In addition, the RBF kernel has fewer numerical difficulties, one of them is $0 < K_{ij} \leq 1$ in contrast to polynomial kernels of which kernel values may go to infinity $(\gamma x_i^T x_j + r > 1)$ or zero $(\gamma x_i^T x_j + r < 1)$ while the degree is large.

D. Sigmoid kernel

We can define the sigmoid kernel as follows:

$$K(x_i, x_j) = tanh(\gamma x_i^T x_j + r).$$

For classification, sigmoid kernel is not as efficient as are the other three kernels. One of the fundamental requirements on a valid kernel is that it must satisfy Mercer's theorem, and that requires that the kernel be positive definite. The sigmoid kernel is not necessarily positive definite, and the parameters γ and r must be properly chosen. The results may be drastically wrong, so much so that the C-SVC performs worse than chance in cases where the kernel is not positive definite. However, under some parameters [10] the sigmoid kernel is not valid (i.e. not the inner product of two vectors). In addition, the sigmoid kernel for a certain range of values of γ and r behaves as a linear kernel [4], while for a certain other range of values of the same parameters; the sigmoid kernel behaves like RBF [16].

III. Experiment and Results

In this section, we will investigate the impact of selection kernel and parameters on C-SVC in medical domains. Consequences of choosing different kernels with C-SVC are monitored, together with the effects of optimizing parameters. Later on, comparisons of results of measuring the performance of classifiers are presented.

For the purpose of this research, LIBSVM is used, as a library for SVC and regression [17]. LIBSVM provides usefully models: C-SVC, γ-SVC, distribution estimation (one-class SVM), ϵ-support vector regression, and γ-support vector regression. C-SVC with following kernels:

linear, polynomial, RBF and sigmoid is used in experiment.

Nine real data sets in medical domains were used for tests, taken from the UCI repository of machine learning databases [18]. We used these data sets to compare results of classification with C-SVC and different kernels and parameters in medical diagnosis. In the following, we provide the details for the benchmark data sets we have used from UCI repository of machine learning databases.

Cardiotocography: This data set consists of measurements of fetal heart rate and uterine contraction features on cardiotocograms classified by expert obstetricians. Fetal cardiotocograms were automatically processed and the respective diagnostic features measured, and classified by three expert obstetricians and a consensus classification label assigned to each of them. Classification was both with respect to a morphologic pattern and to a fetal state. This data set with 2126 instances and 23 features can be used either for 10-class or 3-class experiments.

Parkinsons: The data set is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease. Discriminate healthy people from those with Parkinson's disease, is possible according to "status" column which is set to 0 for healthy and 1 for Parkinson's disease. There are 195 instances and 23 features.

Hepatitis: The main aim of this data set is to predict whether hepatitis patients will die or not. In this data set, there are two classes: live (123 instances) and die (32 instances). There are 155 instances and 19 features, with missing values.

Liver: In the liver data set, the first five variables are all blood tests, which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. Each row in this data set constitutes the record of a single male individual. There are 345 instances and 7 features, with no missing values.

Pima Indians diabetes: In this data set the diagnostic is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care). Patients included in this investigation live near Phoenix, Arizona, USA. There are 768 instances and 8 features all of which are numeric valued. Negative (tested negative for diabetes) are 500 instances and remaining 268 are positive. There are missing values.

Statlog (Heart): The task is to predict absence or presence of heart disease. This data set contains 13 attributes (which have been extracted from a larger set of 74). There are 270 observations, with no missing values.

Breast cancer: Class of this data set has following values: no-recurrence-events and recurrence-events. There are 201 instances of one class and 85 instances of another class. Each instance is described by nine attributes, some of which are linear and some are nominal. There are missing values in this data set.

Mammographic mass: The task is to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. This data set contains a BI-RADS assessment, the patient's age and three BI-RADS attributes together with the ground truth (the severity field) for 516 benign and 445 malignant masses that have been identified on full field digital mammograms collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006. In the data set, each instance has associated BI-RADS assessment ranging from 1 (definitely benign) to 5 (highly suggestive of malignancy) assigned in a double-review process by physicians.

Lung cancer: This data set described three types of pathological lung cancers. All predictive attributes are nominal, taking on integer values 0-3. There are 32 instances and 56 features, with missing values.

To achieve the goal of C-SVC to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes; the following procedure is used. It consists of transform data to the appropriate format, conduct simple scaling on the data, consider different kernels (RBF kernel $K(x,y) = e^{-\gamma\|x-y\|^2}$; sigmoid $K(x_i,x_j) = tanh(\gamma x_i^T x_j + r)$; linear $K(x_i,x_j) = x_i^T$; polynomial $K(x_i,x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$) each combination of parameter choices for different kernels is checked using 10-fold cross-validation, and the parameters with best cross-validation accuracy are picked to train the whole training set and test.

The experiment was run on AMD Phenom (tm) 9650 Quard-Core Pocessor 2.31 GHz with 4GB RAM. The results of preliminary experiment with C-SVC and four kernels, without optimizing parameters are presented in Table 1. Because C-SVC works only with no missing values and following data sets: breast cancer, mammographic mass, lung cancer and hepatitis have missing values, we implemented filter which replaced all missing values for nominal and numeric attributes in data sets with the modes and means from the training data.

TABLE I.
CLASSIFICATION ACCURACY OF C-SVC AND DIFFERENT KERNELS WITHOUT OPTIMIZING PARAMETERS

Data set	accuracy			
	RBF	sigmoid	linear	polynomial
Cardiotocography	80.8	77.8	98.5	96.3
Parkinsons	79.5	75.4	83.6	91.3
Hepatitis	79.4	79.4	85.2	72.9
Liver	59.4	58.0	70.1	-
Pima Indians diabetes	65.1	65.1	77.5	-
Statlog (Heart)	55.9	55.6	84.4	91.3
Breast cancer	73.1	69.9	70.6	69.9
Mammographic mass	80.9	53.7	83.0	82.2
Lung cancer	71.9	71.9	75.0	78.1

Preliminary results for classification accuracy without optimizing parameters with nine data sets in medical domains shows the best performance of linear kernel. In more than half data sets, we get the best results with linear kernel. With sigmoid kernel, there are no best results; RBF has best results for only data set, and polynomial with tree data sets.

Preliminary results show that time taken to build model are very high with C-SVC and polynomial kernel, compare with others kernels. We stopped experiments with C-SVC and polynomial kernel twice, because model was not built for 48 hours for two data sets: Pima Indians diabetes and liver. Time taken to build model for

mammographic mass is 4833.8 sec and completely results for classification accuracy we got after 24 hours.

Vapnik proved that SVM is capable of finding optimal solutions, but the full success can be achieved only if the adaptive process parameters (C and kernel functions parameters) are set to proper values. With the exception of the linear kernel, all the others have some parameters of free choice. The RBF kernel free parameter is the γ. The sigmoid kernel has also the γ parameter. The polynomial kernel has more hyperparameters than the RBF and sigmoid kernels.

On the other hand, the number of hyperparameters influences the complexity of model selection. Since doing a complete grid-search may still be time-consuming, after preliminary results we decide to use only three kernels: RBF, sigmoid and linear. In that case, there are two parameters for RBF and sigmoid kernel: C and γ, and only C for linear kernel to optimize.

Our implementation is as follows. The training data is separated into 10 subsets of equal size in 10-fold cross-validation. Sequentially one fold is considered as the validation set and the rest are for training. The cross validation accuracy is the average of accuracy on predicting the validation sets. For optimal solution, we considered two parameters for RBF and sigmoid kernel: C and γ, and only C for linear. For a given problem, it is not known beforehand which C or γ is best. Therefore, some kind of parameter search must be done to identify good C and γ so that the classifier can accurately predict unknown data. We use a grid-search on C and γ using 10-fold cross-validation. The grid-search can be easily parallelized because each C and γ is independent. We provide a possible interval of C and γ with the grid space. For C, maximum value is 20.0, minimum value is 5.0 and step is 1.0. For γ, maximum value is 5.0, minimum value is -10.0 and step is 1.0. Then, all grid points of C and γ are tried to see which one gives the highest cross validation accuracy. Finally, we use the best parameter to train the completely training set and generate the final model.

TABLE II.
C-SVC WITH RBF KERNEL AND THE BEST VALUES OF C AND **γ**

Data set	Coordinates	C	γ
Cardiotocography	20.0, -4.0	20.0	1.0E-4
Parkinsons	20.0, -3.0	20.0	0.0010
Hepatitis	20.0, 5.0	20.0	100000.0
Liver	20.0, -4.0	20.0	1.0E-4
Pima Indians diabetes	20.0, -6.0	20.0	1.0E-6
Statlog (Heart)	20.0, -4.0	20.0	1.0E-4
Breast cancer	18.0, -2.0	18.0	0.01
Mammographic mass	20.0, -2.0	20.0	0.01
Lung cancer	20.0, -2.0	20.0	0.01

TABLE III.
C-SVC WITH SIGMOID KERNEL AND THE BEST VALUES OF C AND **γ**

Data set	Coordinates	C	γ
Cardiotocography	20.0, -6.0	20.0	1.0E-6
Parkinsons	19.0, -6.0	19.0	1.0E-6
Hepatitis	20.0, 5.0	20.0	100000.0
Liver	20.0, -5.0	20.0	1.0E-5
Pima Indians diabetes	20.0, -6.0	20.0	1.0E-6
Statlog (Heart)	10.0, -6.0	10.0	1.0E-6
Breast cancer	5.0, -1.0	5.0	0.1
Mammographic mass	12.0, -4.0	12.0	1.0E-4
Lung cancer	20.0, 5.0	20.0	100000.0

The best values of parameters of C-SVC for nine data sets in medical domains are presented in Table 2 for RBF kernel, Table 3 for sigmoid kernel and Table 4 for linear kernel. Only with C-SVC and linear kernel we stopped experiment for Pima Indians diabetes data set, because model was not built for 36 hours.

TABLE IV.
C-SVC WITH LINEAR KERNEL AND THE BEST VALUE OF C

Data set	C
Cardiotocography	5
Parkinsons	5
Hepatitis	5
Liver	5
Pima Indians diabetes	-
Statlog (Heart)	1
Breast cancer	5
Mammographic mass	15
Lung cancer	5

TABLE V.
ACCURACY OF C-SVC WITH RBF KERNEL AND THE BEST PARAMETERS

Data set	accuracy	time
Cardiotocography	97.4	598.1
Parkinsons	98.5	14.8
Hepatitis	79.4	11.0
Liver	71.9	13.5
Pima Indians diabetes	75.9	59.9
Statlog (Heart)	75.2	12.2
Breast cancer	73.1	15.1
Mammographic mass	82.3	45.5
Lung cancer	71.9	9.2

Results of classification accuracy, as a method for measuring the performance of C-SVC with different kernel and the best values of parameters for nine data sets in medical domains, are presented on Table 5, Table 6 and Table 7. Second column in the table presents method for measuring the performance of classifiers, and the last one presents the time taken to build model, in seconds. Setting appropriate parameters for a given data set, the reliability of classification is increased, also the time taken to build the model.

TABLE VI.
ACCURACY OF C-SVC WITH SIGMOID KERNEL AND THE BEST PARAMETERS

Data set	accuracy	time
Cardiotocography	84.9	406.5
Parkinsons	76.9	23.1
Hepatitis	79.3	10.6
Liver	59.4	12.6
Pima Indians diabetes	76.3	34.0
Statlog (Heart)	60.4	18.4
Breast cancer	71.0	8.9
Mammographic mass	72.8	67.2
Lung cancer	78.1	5.7

With RBF kernel this improvement is significantly; more than half data sets have improvement of classification accuracy more than 10% (Cardiotocography, Parkinsons, Liver, Pima Indians diabetes and Statlog (Heart)). The maximum increase accuracy of classification is 19.3% for Statlog (Heart).

With sigmoid kernel, improvement of accuracy is higher, but not significantly. Only two data sets have improvement of classification accuracy more than 10%

(Pima Indians diabetes and Mammographic mass). The maximum increase accuracy of classification is 19.1% for Mammographic mass).

TABLE VII.
ACCURACY OF C-SVC WITH LINEAR KERNEL AND THE BEST VALUE OF C

Data set	accuracy	time
Cardiotocography	98.5	796.3
Parkinsons	83.6	582.9
Hepatitis	85.2	739.2
Liver	70.1	335.0
Pima Indians diabetes	-	-
Statlog (Heart)	84.4	57270.4
Breast cancer	71.3	7.5
Mammographic mass	83.1	2831.2
Lung cancer	75.0	1.1

With liner kernel, this improvement is slightly; more than half data sets have no improvement of classification accuracy. For Pima Indians diabetes data set, we stopped experiment, because model was not built for 36 hours. Only for two data sets, we have improvement (Mammographic mass and Breast cancer). The maximum increase accuracy of classification is 0.7% for Breast cancer.

Optimizing parameters classification accuracy is significantly improve with RBF kernel, higher with sigmoid kernel and slightly with linear kernel. The biggest benefits of optimization had those kernel functions, which have a smaller accuracy of classification. RBF kernel is at least as good as the linear kernel, but only after searching the $(C; \gamma)$ space, which is quick compare with others kernel. After optimization of parameters, results of classification accuracy are more consistent for all kernel functions, and we no longer have the dominance of certain kernel functions, or larger variance in the results.

Fig. 1 presents the time required for optimal modeling C-SVC with different kernel. The computational efficiency we measured in seconds of running time, logarithmic scale, base 10. Each configuration started ten times, and the average running time we used as a result.

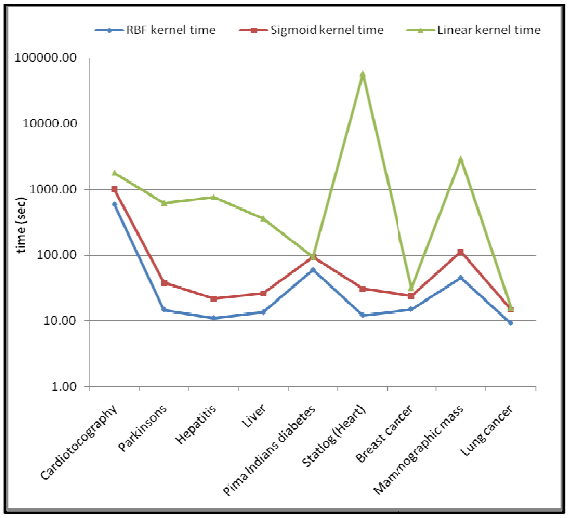

Figure 1. Time taken to build model for optimal modeling C-SVC with different kernel

Time required for optimal modeling C-SVC depends on the kernel. For all data sets, RBF kernel required minimum time and linear kernel required maximum time. Required time for sigmoid kernel is between these two kernel functions.

IV. CONCLUSIONS

C-SVC with different kernels has used on nine medical data sets. These results evaluated and compared choosing different parameters. As we expected the effectiveness of C-SVC depends on the selection of kernel and the parameters. Experimental results proved that after optimization of parameters, results of classification accuracy are more consistent for all kernel functions, and we no longer have the dominance of certain kernel functions, or larger variance in the results.

There are many questions and issues that remain to address and that we intend to investigate in future work. In further research, we will try with others kernel with aim to get improvements of classification accuracy. We will use larger data sets and expect that a complete grid-search may still be time-consuming. In that case, we will use a coarse grid first: after identifying a "better" region on the grid, a finer grid search on that region can be conducted. We will also use appropriate feature selection and reduction methods for large data sets.

In addition, this research could help in future works, like classifying static and no-static medical images into appropriate classes. These results of C-SVC with different kernels can be used for improvement medical diagnosis.

REFERENCES

[1] E. C. Ifeachor, A. Sperduti, and A. Starita, *Neural networks and expert systems in medicine and health care,* Singapore: World Scientific Publishing, 1998.

[2] T. Teodorrescu, A. Kandel, and L. C. Jain, *Fuzzy and neuro-fuzzy systems in medicine,* Boca Raton, FL: CRC Press, 1998.

[3] H. Nazeran, and K. Behbehani, "Neural networks in processing and analysis of biomedical signals", in M. Akay (Ed.), *Nonlinear biomedical signal processing: Fuzzy logic, neural networks and new algorithms,* IEEE Press, 2001, pp. 69-97.

[4] H. A. Abbass, "An evolutionary artificial neural networks approach for breast cancer diagnosis", *Artificial Intelligence in Medicine, 25*(3), 2002, pp. 265-281.

[5] K. Chan, T. W. Lee, P. A. Sample, M. H. Goldbaum, R. N. Weinreb, and T. J. Sejnowski, "Comparison of machine learning and traditional classifiers in glaucoma diagnosis", *IEEE Transactions on Biomedical Engineering, 49*(9), 2002, pp. 963-974.

[6] S. Yagneswaran, M. Baker, and A. Petrosian, "Power frequency and wavelet characteristics in differentiating between normal and Alzheimer EEG", *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society (EMBS/BMES),* 1, 2002, pp. 46-47.

[7] D. Garrett, D. A. Peterson, C. W. Anderson, and M. H. Thaur, "Comparison of linear, non-linear and feature selection methods for EEG signal classification", *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* 11, 2003, pp. 141-147.

[8] W. I. Schöllhorn, "Applications of artificial neural nets in clinical biomechanics", *Clinical Biomechanics,* 19, 2004, pp. 876-98.

[9] J. Kamruzzaman, R. Begg, R. Sarker, "Overview of Artificial Neural Networks and their Applications in Healthcare", *Neural networks in healthcare: potential and challenges*, Idea Group, 2006.

[10] V. N. Vapnik, "The Nature of Statistical Learning Theory", *Information Science and Statistics*, Springer, 1st edition, 1995.

[11] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers", in: *Proceedings of the*

Fifth Annual Workshop of Computational Learning Theory (COLT), 1992.

[12] M. Aizerman, E. Braverman, and L. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning", *Automation and Remote Control* 25: 821–837, 1964.

[13] C. Cortes and V. Vapnik, "Support-vector network", *Machine Learning*, 20:273–297, 1995.

[14] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, NY, 1998.

[15] S. S. Keerthi, and C.-J. Lin, "Asymptotic behaviors of support vector machines with Gaussian kernel", *Neural Computation*, 15(7):1667-1689, 2003.

[16] H.-T. Lin, and C.-J. Lin, "A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods", Technical report, Department of Computer Science, National Taiwan University, 2003, URL http://www.csie.ntu.edu.tw/~cjlin/papers/tanh.pdf.

[17] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: a library for support vector machines", *ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[18] A. Frank, and A. Asuncion, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science, 2010.