Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Classification of genetic expression in prostate cancer using support vector machine method

To cite this article: S A Komarudin *et al* 2020 *J. Phys.: Conf. Ser.* **1613** 012032

View the article online for updates and enhancements.

Classification of genetic expression in prostate cancer using support vector machine method

S A Komarudin, D Anggraeni, A Riski and A F Hadi

Departmen of Mathematics, Faculty of Mathematics and Natural Science, University of Jember, Jember, Indonesia

E-mail: afhadi@unej.ac.id

Abstract. Prostate cancer has long been a concern of expert's human genetics in health research. However, an explanation of the main causes of prostate cancer cannot be obtained metabolically-biologic, except the most common one of which is heredity. Explanation of the risk of contracting prostate cancer is sought through genetic explanation of prostate cancer cells and healthy prostate cells from DNA sequencing in the form of micro arrays data or in the form of Gleason values. Cancer cell genetic data is high dimensional where the number of variables observed were far more than the individual observed. It's make ordinary multivariate classification techniques fail to handle this data because of the singularity matrix. In addition, the observations number of cancer patients are small since they are rarely found. With these two facts, then in this paper we will use a machine learning approach to study the classification, namely SVM. SVM will be compared with the Naive Bayes Classifier and Discriminant Analysis method to determine the accurate division in distinguishing prostate cancer cells from healthy prostate cells. The sample data used consisted of 102 people with 2135 genetic variables which were then divided into training data and testing data. Based on the results of the study, the classification by the SVM method has an accuracy value of 96% with a precision error in the tumor class of 7%. The Naive Bayes classification has a precision error of 23.5% with a classification accuracy of 84%. While the Discriminant Analysis method produces an accuracy of 92% with a precision error of 13.33%.

1. Introduction

Prostate cancer is a leading cause of death in men in western countries. The cause of cancer is still unknown, but many factors can affect the risk of getting prostate cancer. The most common risk factors are age and heredity. Then an explanation of the risk of developing prostate cancer can be sought from one of the risk factors that cause it is through genetic mutations in prostate cells. Genetic mutations of prostate cells that are not normal, can develop into malignant tumors in the prostate which then causes prostate cancer [1].

Prostate cell genetic mutation data used data type is the value of the degree of malignancy of the prostate with a Gleason score system, which is a prognostic factor for predicting the risk of developing prostate cancer. Therefore, the need for related analysis of contracting prostate cancer risk classification by mutation of genetic expression in prostate cells to be measured from the Gleason score. The form of genetic mutation data is microarray datasets, where the number of observed variables much more than individuals who were observed [2]. Thus, ordinary multivariate analysis classification techniques cannot be performed on this data type because of the singularity matrix. Thus, it will use machine learning approach to solve the case of the classification. There are several methods that can be used to determine the classification of cases, one of which is Support Vector Machine (SVM).

SVM is one method that lately received more attention from researchers because it provides good classification results with a high degree of accuracy [3], as evidenced by several previous studies such as Pratama [4], Damanik [5], and Andari [6]. The concept of SVM explains how a simple effort to find the best separator function (hyperplane) from several alternative dividing lines that might occur in a case. The best separator function is to find the optimal value of the margin of demarcation to each class, to the right at the center dividing line between positive class and negative class [7]. This study aims to provide an accurate dividing function to distinguish normal category prostate genetic cells from genetic cells at tumor risk using SVM and will be compared with the Naive Bayes Classifier and Discriminant Analysis.

2. Material and Method

2.1. Prostate Cancer

The cause of prostate cancer is still not known for certain, but there are several factors that can affect the spread of prostate cancer. These factors include age, genetics, race, heredity, diet, sexual behavior and other factors [8]. The doctors agree that age and heredity are common factors that often occurs in people. In general, prostate cancer often occurs in the range of over 40 years, the incidence of outbreaks is increasing rapidly in the above age. Family history or origin who can also increase the risk of developing prostate cancer, a person who has a family member living with prostate cancer have twice the risk or even exposed to higher prostate carcinoma [9].

Genetic factors may affect the risk of contracting prostate cancer through mutations in the genetic expression of the prostate gland in the male reproductive system. Genetics of the prostate gland can predict a person's risk of contracting prostate cancer whether or not from the data analysis on the results of genetic mutation. Genetic data observed as the result of metabolic processes in cells, the arranged data is a DNA sequencing through the process of transcription and translation of mRNA molecules are then translated in order to determine the nature of an organism. Analysis of mutations in the genetic expression can be used to determine the identification of genes that may have a risk of becoming infected, although no correlation with serum Prostate Specific Antigen (PSA) to be measured from Gleason score [2].

Analysis of prostate cancer focused on gene expression from cancer microarrays data using the classification method. The accuracy of the classification results from the microarray classification is very useful, because the accuracy of the diagnosis using microarrays can help the selection of appropriate therapy [10].

2.2. Classification

Classification is a method of grouping data that will learn training data using a classification algorithm. As for several classification algorithms, including Bayesian Classification, K-Nearest Neighbor, Decision Tree Induction, Case-Based Reasoning, Genetic Algorithms, Discriminant Analysis, and Support Vector Machines [11]. Experimental and evaluation shows that SVM, KNN and NB are traditional classification texts. Experiments and evaluations show valid clarification texts [12].

Measurement of classification performance can describe how well the classifier is in classifying data. Confusion matrix is one method that can analyze how well the classifier recognizes tuples from each class classification [13]. Confusion matrix is a table recording the results of classification work, can be seen in Table 1.

Table 1. Confusion Matrix

Original class	Prediction class	
	Negative	Positive
Negative	True Negative (TN)	False Positive (FP)
Positive	False Negative (FN)	True Positive (TP)

Based on the table, the classification performance of accuracy and precision can be calculated. The accuracy value is the value of how big the accuracy of the data classification results, while precision is the ratio between the true positive class and all positive result classes. Comparing precision and accuracy parameters can be used as a reference in determining the best classification method [14].

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} * 100\% \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} * 100\% \tag{2}$$

2.3. Support Vector Machine

Support Vector Machine (SVM) was first introduced by Vapnik in 1995 as a harmonious series of leading concepts in the field of pattern recognition. SVM is one of the best methods that can be used in classification problems. SVM is a learning machine method that works on the principle of Structural Risk Minimization (SRM) which aims to find the best hyperplane that separates two classes in the input space [15]. Basically SVM works with the principle of linear classifier, then developed to be able to work in non-linear cases using the concept of the kernel in a high-dimensional workspace. Feature selection and parameter adjustment in SVM significantly influence the results of classification accuracy [16].

2.3.1. Linear Support Vector Machine. The SVM concept can be explained simply as an attempt to find the best hyperplane boundary that functions to separate the two classes in the input space [17]. Each of data is denoted as $\vec{x}_i \in R^d, i = 1,2, \dots, n.$ Where n is the number of data. Positive class is denoted as 1, and a negative class as 0. Thus, the data and label each class is denoted as: , $y_i \in \{0,1\}$. It is assumed that these two classes can be separated completely by hyperplane in d-dimensional feature space. The hyperplane is defined as follows:

$$\vec{w}.\vec{x_i} + b = 0 \tag{3}$$

Data $\vec{x_i}$ were classified into negative class that satisfies the following inequality:

$$\vec{w}.\vec{x_i} + b \leq -1 \tag{4}$$

While the data $\vec{x_i}$ were classified into positive class that satisfies the following inequality:

$$\vec{w}.\vec{x_i} + b \geq 1 \tag{5}$$

The margin can be obtained by maximizing the distance between the hyperplane to the nearest point of each class, namely $1/\|\vec{w}\|$, Furthermore, it can be formulated as a Quadratic Programming (QP) problem, by finding the minimum point of the equation (6) and the constraint in equation (7).

$$(w) = \frac{1}{2}\|w\|^2 \tag{6}$$

$$y_i(\vec{w}.\vec{x_i} + b) - 1 \geq 0, \forall_i \tag{7}$$

This problem can be solved by Lagrange Multiplier

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i y_i(\vec{w}.\vec{x_i} + b - 1)\, , (i = 1,2, \dots, l) \tag{8}$$

α_i Lagrange multipliers are zero or positive. Constrained optimization problems can be calculated by minimizing L against $\vec{w}$ and b , and maximizing L against α_i.
Maximize:

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i x_j \tag{9}$$

Subject to:

$$\alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{10}$$

α_i the value is greater than 0 is a support vector, while the rest $(\alpha_i > 0)\alpha_i = 0$ [18].

2.3.2. Non Linear Support Vector Machine. SVM was discovered in 1964 to open a class using hyperplane in pattern recognition. Then in 1992-1995 a generalization was made to construct a non-linear separating function (only in the feature space). In 1995, another generalization was carried out to estimate the function that had real value. Finally, in 1996 a solution was found for non-linear separators with kernel functions [7]. The mapping process requires the calculation of the dot product of two variables on a new vector space. dot product both vectors (x_i) and (x_j) denoted as $\Phi(x_i).\Phi(x_j)$, The value of the second dot product of vectors can be calculated indirectly, without knowing the transformation function Φ, This computational technique called Kernel Trick, is to calculate the dot product of two vectors in a new vector space by using both components of these vectors in a vector space of origin as follows.

$$K(x_i, x_j) = \Phi(\mathrm{x}_i).\Phi(x_j) \tag{11}$$

Various types can be used as a kernel function K, as listed in Table 2 [19].

Table 2. The kernel functions in SVM

Kernel	Definition
Linear	$K(x_i, x_j) = (x_i.x_j)$
Polynomial	$K(x_i, x_j) = (x_i.x_j + 1)^p$
Gaussian RBF	$K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$
Sigmoid	$K(x_i, x_j) = \tanh(\alpha \mathrm{x}_i.\mathrm{x}_j + \beta)$

2.4 Naive Bayes Classifier

Naive Bayes is a simple classification that calculates probabilities by adding up frequencies and combinations of values from a given dataset. Theorem algorithm uses Bayes by assuming all the attributes are independent or not interdependent given by all classes of variables [20]. Naive Bayes is based on the simplification assumption that attribute values are conditionally mutually independent if output values are given. In other words, given the value of output, the probability of observing together is a product of individual probabilities [21].

The equation from the Bayes theorem is [22]:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \tag{12}$$

To explain the Naive Bayes method, it is important to know that the classification process requires a number of clues to determine what class is suitable for the analyzed sample. Therefore, the Naive Bayes method above is adjusted as follows:

$$P(C|F1 \dots F_n) = \frac{P(C)P(F1 \dots F_n|C)}{P(F1 \dots F_n)} \tag{13}$$

Where variable *C* represents class, while variable $F1 \dots F_n$ represents the characteristic instructions needed to classify. Then the formula explains that the probability of entering certain characteristic samples in class *C* (Posterior) is the chance of the emergence of class *C* (before the sample entry, often called prior), multiplied by the chance of the appearance of sample characteristics in class *C* (also called likelihood), divided with the opportunity for the emergence of sample characteristics globally (also called evidence). The more complex the factors that affect the probability value, as a result the calculation becomes difficult to do. Next we use the very high assumption of independence (naive), that each of the instructions is free from one another. With this assumption, the following equation applies:

$$P(F_i|F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i)P(F_j)}{P(F_j)} = P(F_i) \tag{13}$$

For $i \neq j$, then

$$P(F_i|C, F_j) = P(F_i|C) \tag{13}$$

The equation above is a model of the Naive Bayes theorem which will then be used in the classification process. For classification with continuous data the Gauss Density formula is used:

$$P(X_i = x_i|Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(x_i - \mu_{ij})^2}{2\sigma^2 ij}} \tag{14}$$

2.5 Discriminant Analysis

Discriminant analysis is a dependent technique in which the independent variables are non metric. Where the grouping of each object into two or more is based on the criteria of independent variables, which means a statistical technique used to categorize into two or more classes. The purpose of discriminant analysis is to determine the discriminant function to differentiate a group into predetermined categories. The discriminant analysis model is stated by the following formula [23].

$$Z = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_nX_n \tag{15}$$

3. Implementation

3.1. Datasets

The data used is the result of mutation of genetic expression in prostate cells. The characteristics of genetic data type of microarray datasets, where the number of observed variables much more than individuals who were observed. Sample the data in this case consists of 102 people with a genetic variables 2135, which are then divided into training and testing with a proportion 75:25 randomly.

3.2 Application

Stages will be done for classifying prostate cell genetic mutations are as follows:

a. Prepare and divide the data into two parts, training and testing data with a proportion of 75:25.
b. Classifying using the Support Vector Machine method.
    1) Determine kernel functions for modeling of Linear, Polynomial, Gaussian RBF and Sigmoid kernels.
    2) Obtained the best kernel function from the smallest error value.
    3) Forming the results of classification using the best kernel with testing data.
    4) Calculate the performance of classification accuracy.
c. Classify using the Naive Bayes Classifier.
    1) Calculate criteria and probabilities.
    2) Testing the Naive Bayes.
    3) Determine the result of the classification
d. Classifying using the Discriminant Analysis method.
    1) Perform linear discriminant analysis.
    2) Determine the prediction class.
    3) Calculating classification performance.
e. Compare the classification accuracy obtained from Support Vector Machine, Naive Bayes Classifier, and Discriminant Analysis.
f. Make conclusions.

4. Results and Discussion

General description of the research data on genetic expression of prostate cancer will be presented using microarrays data. Microarray data used in the form of 102 individuals with each genetic number of 2135. A total of 50 individuals belong to the normal class, and 52 individuals are included in the tumor class. Genetic expression data that have been obtained from microarrays are processed and converted into matrix form so that they can be processed using packages in the R program. Then the data is divided into training data and testing data with a ratio of 75:25 with the same proportions. Furthermore, classification will be done using SVM, Naive Bayes, and Discriminant Analysis.

4.1 Support Vector Machine

The analysis uses SVM method with linear, polynomial, radial, and sigmoid kernel functions. The first step is determining the kernel functions that will be used for modeling using training data. The output obtained is the smallest error value of each kernel function as in Table 3 below.

Table 3. The cost error value for each kernel

	Kernel			
	Linear	Polynomial	Gaussian RBF	Sigmoid
Cost error	0.1535	0.2303	0.1425	0.1308

Based on the error values that have been obtained, it can be seen that the best kernel function is a sigmoid kernel with 0.1308 as the smallest error value. So for classification modeling on SVM will be done using the sigmoid kernel function. The best kernel function is used to predict classification classes in testing data. To find out the results of classification with testing data using sigmoid kernel, it can be seen using the confusion matrix presented in Table 4 below.

Table 4. SVM confusion matrix in testing data

Original class	Prediction class	
	Normal	Tumor
Normal	11	1
Tumor	0	13

Based on the confusion matrix, it can be seen that there are 2 misclassified data, with the correct classification being 10 normal class data and 13 tumor class data. In this case the discriminant analysis produces a classification accuracy of 92%. The precision generated from 15 data classified as tumor class with a correct classification of tumors of 13 is 86.67% with an error precision of 13.33%.

4.2 Naive Bayes Classifier

Naive Bayes is based on the simplification assumption that attribute values are conditionally mutually independent if output values are given. In other words, given the value of output, the probability of observing together is a product of individual probabilities. In the training data criteria it can be seen from 77 data that there are 38 data as normal classes and 39 data as tumor classes. Whereas the testing data criteria from 25 data contained 12 normal class data and 13 tumor class data. The probability of training data and testing data criteria can be seen in Table 5.

Table 5. Probability of data criteria

Class	Training		Testing	
	Total data	Probability	Total data	Probability
Normal	38	49.35065	12	48
Tumor	39	50.64935	13	52

From the probability value above, it will be tested and solved using the R program so that the results of the classification of testing data are produced which are presented with the confusion matrix as in Table 6 below.

Table 6. Naive Bayes confusion matrix in testing data

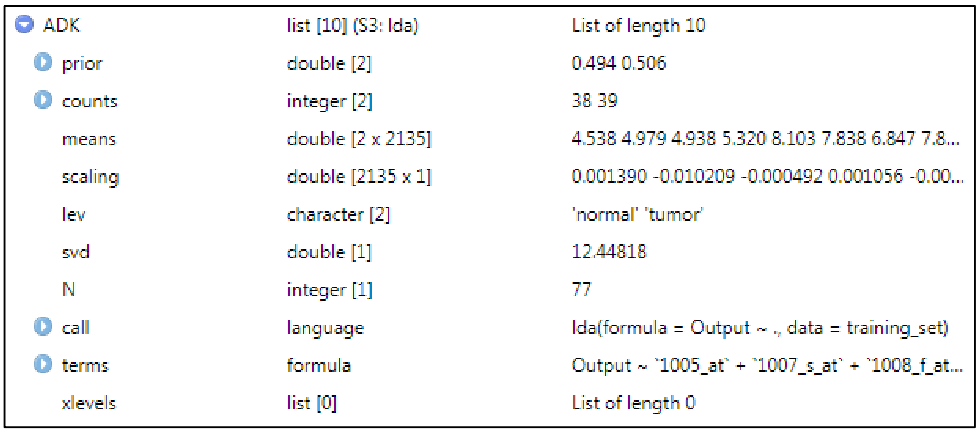
Original class	Prediction class	
	Normal	Tumor
Normal	8	4
Tumor	0	13

The resulting confusion matrix shows correct classification results with 8 data in the normal class and 13 data in the tumor class which results in an accuracy rate of 84%. There are 4 misclassified data on the tumor class (False Positive) which should be included in the normal class. The number of tumor class classification data is 17 data with a 76.5% precision level with a precision error of 23.5%.

4.3 Discriminant Analysis

The test uses discriminant analysis method from the output of normal class and tumor class. After formulating data from genetically independent variables with the dependent variable output next calculates the odds of each class. To determine the opportunity is done by finding prior values through training data.

These priors can represent the chance that an object will be in which class. The prior value of training data from the normal class is 0.494 and the tumor class is 0.506. These results represent that the classification results tend to be slightly toward the tumor class, to prove that it is validated using testing data. Before validating using data testing, a linear discriminant analysis is performed first for the discriminant function approach presented in Figure 1 below.



```
ADK          list [10] (S3: lda)    List of length 10
  prior      double [2]             0.494 0.506
  counts     integer [2]            38 39
  means      double [2 x 2135]      4.538 4.979 4.938 5.320 8.103 7.838 6.847 7.8...
  scaling    double [2135 x 1]      0.001390 -0.010209 -0.000492 0.001056 -0.00...
  lev        character [2]          'normal' 'tumor'
  svd        double [1]             12.44818
  N          integer [1]            77
  call       language               lda(formula = Output ~ ., data = training_set)
  terms      formula                Output ~ `1005_at` + `1007_s_at` + `1008_f_at...
  xlevels    list [0]               List of length 0
```

Figure 1. Linear discriminant analysis

Next, validate the classification results using data testing. These results are presented in the confusion matrix in Table 7 below.

Table 7. Discriminant Analysis confusion matrix in testing data

Original class	Prediction class	
	Normal	Tumor
Normal	10	2
Tumor	0	13

Based on the confusion matrix table, the results show that there are 23 data classified correctly and there is 2 data misclassification. From these results obtained classification accuracy with discriminant analysis using testing data of 92%. The precision given with the results of 13 tumor data was correctly classified from 15 tumor class classification data of 86.67% with a percentage of precision error of 13.33%.

4.4 Comparison of classification results from the three methods

From the evaluation of the classification results in this analysis, comparing the value of classification accuracy to determine the best classification using the Support Vector Machine method, Naive Bayes Classifier or Discriminant Analysis obtained the performance in Table 8 as follows.

Table 8. Classification comparison

Support Vector Machine		Naive Bayes Classifier		Discriminant Analysis	
Accuracy	Precision error	Accuracy	Precision error	Accuracy	Precision error
96%	7%	84%	23.5%	92%	13.33%

From the comparison comparison table, information is obtained that the best classification is to use the SVM method with a classification accuracy of 96% and an error precision of 7%.

5. Conclusion

Based on the analysis that has been done, using microarray data on genetic expression of prostate cancer from 102 people with 2135 variables. Training and testing data of 77 and 25 concluded that using the SVM method produced the highest classification accuracy of 96%. While the accuracy of classification using Naive Bayes Classifier produces an accuracy of 84% and Discriminant Analysis produces 92%. Thus the SVM method can separate the normal class and tumor risk in the genetic expression of prostate cancer very well.

Acknowledgement

Researchers would like to thank the respondents in the University of Jember environment both at the rectorate level, faculties, departments and to LP2M who have helped many research activities.

References

[1] Komite Penanggulangan Kanker Nasional 2013 *Panduan Penatalaksanaan Kanker Prostat* Jakarta: Komite Penanggulangan Kanker Nasional

[2] Singh D, P G Febbo, K Ross, D G Jackson, J Manola, C Ladd, P Tamayo, A A Renshaw, A V D'Amico, J P Richie, E S Lander, M Loda, P W Kantoff, T R Golub, and W R Sellers 2002 Gene Expression Correlates of Clinical Prostate Cancer Behavior *Cancer Cell* **1** pp 203-09

[3] Vapnik V and Cortes C 1995 Support Vector Network *Machine Learning* **20** pp 273-97

[4] Pratama A, R C Wihandika, and D E Ratnawati 2018 Implementasi Algoritma Support Vector Machine (SVM) untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* **2** pp 1704-08

[5] Damanik S M S, D Ispriyanto, and Sugito 2015 Klasifikasi Lama Studi Mahasiswa FSM Universitas Diponegoro Menggunakan Regresi Logistik Biner dan Support Vector Machine *Jurnal Gaussian* **4** pp 23-32

[6] Andari S 2013 *Smooth Support Vector Machine Dan Multivariate Adaptive Regression Splines Untuk Mendiagnosis Kanker Payudara* Surabaya : Institut Teknologi Sepuluh Nopember

[7] Vapnik V 1998 *The Support Vector Method of Function Estimation* Nonlinear Modeling: Advanced Black-Box Techniques, Kluwer Academic Publishers Boston 55–85

[8] Umbas R 2008 Penanganan Kanker Prostat saat ini dan Beberapa Pengembangan Baru *Jurnal Kanker Indonesia* **3** pp 114-19

[9] Basch E, T K Oliver, A Vickers, I Thompson, P Kantoff, H Parnes, D A Lowblaw, B Roth, J Williams, and R K Nam 2012 Screening for Prostate Cancer With Prostate-Specific Antigen Testing: American Society of Clinical Oncology Provisional Clinical Opinion *Journal of Clinical Oncologi* **30**

[10] Glaab E, Bacardit J, Garibaldi J M, and Krasnogor N 2012 Using Ruled-Based Machine Learning for Candidate Desease Gene Prioritization and Sample Classification of Cancer Gene Expression Data *Journal Plos One* **7** pp 1-18

[11] Khan, Aurangzeb, B Baharudin, L H Lee, and K Khan 2010 A Review of Machine Learning Algorithms for Text-Documents Classification *Journal of Advances in Information Technology* **1** pp 4-20

[12] Yao and Zhi-Min 2012 An Optimized NBC Approach in Text Classification *Physics Procedia* **24** pp 1910-14

[13] Sokolova M and G Lapalme 2009 A Systematic Analysis of Performance Measures for Classification Tasks *Information Processing and Management* **45** pp 427-37

[14] Ali J, R Khan, N Ahmad, and I Maqsood 2012 Random Forest and Decision Trees *International Journal of Computer Science Issues* **9** pp 272-78

[15] Chou J S, M Y Cheng, Y W Wu, and A D Pham 2014 Optimizing Parameters of Support Vector Machine Using Fast Messy Genetic Algorithm for Dispute Classification *Expert System with Application* **41** pp 3955-64

[16] Zhao M, C Fu, L Ji, K Tang, and M Zhou 2011 Feature Selection and Parameter Optimization for Support Vector Machine: A New Approach Based on Genetic Algorithm with Feature Chromosomes *Expert System with Application* **38** pp 5197-204

[17] Gunn S R 1998 *Support Vector Machine for Classification and Regression* (Southampton: University of Southampton)

[18] Suykens J A K and J Vandewalle 1999 Least Squares Support Vector Machine Classifier *Journal Neural Processing Letters* **9** pp 293-300

[19] Kecman V 2005 Support Vector Machine – An Introduction *Support Vector Machine: Theory and Applications* **177** pp 1-47

[20] Patil T R and M S Sherekar 2013 Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification International *Journal of Computer Science and Application* **6** pp 256-61

[21] Pattekari S A and A Parveen 2012 Prediction System for Heart Disease Using Naive Bayes *International Journal of Advanced Computer and Mathematical Science* **3** pp 290-94

[22] Frank E, L Trigg, G Holmes, and I H Witten 2000 Technical Note: Naive Bayes for Regression *Machine Learning* **41** pp 5-25

[23] Lachenbruch P A and M Goldstein 1979 Discriminant Analysis *Biometrics* **35** pp 69-85