PAPER • **OPEN ACCESS**

Random-Forest (RF) and Support Vector Machine (SVM) Implementation for Analysis of Gene Expression Data in Chronic Kidney Disease (CKD)

To cite this article: Zuherman Rustam *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **546** 052066

View the article online for updates and enhancements.



IOP ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the collection - download the first chapter of every title for free.

This content was downloaded from IP address 218.75.69.50 on 19/09/2019 at 13:09

IOP Conf. Series: Materials Science and Engineering **546** (2019) 052066 doi:10.1088/1757-899X/546/5/052066

Random- Forest (RF) and Support Vector Machine (SVM) Implementation for Analysis of Gene Expression Data in Chronic Kidney Disease (CKD)

Zuherman Rustam[1*], Ely Sudarsono[1], Devvi Sarwinda[1]

[1]Departement of Mathematics, Universitas Indonesia, Indonesia

*Corresponding author: rustam@sci.ui.ac.id

Abstract. The application of mathematics in the field of bioinformatics has been widely developed. For example Support Vector Machines (SVM) and Random Forest (RF) are state of the art for classification of cancer in many applications. One of them is Chronic Kidney Disease (CKD). CKD is one of the kidney diseases that sufferers are increasing and have symptoms that are difficult to detect at first. Later, microarrays in gene expression are important tools for this approach. Microarrays gene expression provides an overview of all transcription activities in biological samples. The purpose of this research is a hybrid model combining Random Forest (RF) and Support Vector Machine (SVM) can be used to classify gene expression data. RF can highly accurate, generelize better and are interpretable and SVM (called RF-SVM) to effectively predict gene expression data with very high dimensions. In addition, from the simulation results on data from the Gene Expression Omnibus (GEO) database, it is shown that the proposed RF-SVM is a more accurate algorithm on CKD data than RFE-SVM.

Keywords: Random Forest (RF), Support Vector Machine (SVM), Chronic Kidney Disease (CKD)

1. Introduction

Chronic Kidney Disease (CKD) has a variety of phenotypic manifestations including structural (such as fibrosis) and functional changes (such as glomerular filtration rate and albuminuria). Reporting on gene expression has recently gained popularity as an important new tool for precision treatment approaches [1] has led to incremental growth in the public databases such as the ArrayExpress [2] and NCBI Gene Expression Omnibus [3].

Microarrays gene expression provides an overview of all transcription activities in biological samples. Unlike most traditional molecular biology tools, which generally allow the study of single genes or small sets of genes, microarrays facilitate completely new discoveries and unexpected functional roles of genes. This advantage has been applied to various applications, including finding new disease subtypes, developing new diagnostic tools, and identifying mechanisms of disease or underlying drugs [4].

Using computational techniques, especially the deep learning method to facilitate, improve cancer detection, and diagnoses are promising and important fields. At present, gene expression data are already widely used to train deep neural networks that are effective for proper cancer diagnosis [5]. Akash

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

(2018) share that RF can highly accurate, generalize better and are interpretable. In addition, there is also research by Huynh, Nguyen, and Do (2018) proposing a hybrid model that combines Deep Convolutional Neural Networks (DCNNs) and Support Vector Machine (SVM) called DCNN-SVM to effectively predict highly dimensional gene expression data. DCNN-SVM trains the DCNN model to select automatically the features of the gene expression microarray data and studies the non SVM model to classify gene expression data. Therefore, in this study we will develop a hybrid model that RF and SVM to classify CKD gene expression data.

2. Support Vector Machines using Random Forest Selection Feature

Support Vector Machine (SVMs) proposed by Vapnik[5] and RF which is belongs to embedded methods. RF are implemented by algorithms that have their own built-in feature selection methods. Some of the advantages are highly accurate, generalize better and are interpretable [6].

Besides, SVM can provide good decision surfaces by maximizing margins using soft-margin approaches [7]. The novelty in this research is to use a hybrid model that combines RF and SVM and Chronic Kidney Disease gene expression data to be used, namely CKD, which will be compared with previous studies using RFE-SVM. So, we propose a hybrid model that combines RF and SVM called RF-SVM to effectively predict highly dimensional gene expression data.

RF-SVM trains the RF model to select automatically the features of the gene expression microarray data and studies the non SVM model to classify gene expression data based on [5-10]. The results that we want to achieve of the study indicate that the proposed RF-SVM is more accurate than the RFE-SVM algorithm. The research methodology that will be carried out: 1) Study the literature on CKD disease, gene expression, SVM, RF, and other topics related to research; 2) Collect data from the Gene Expression Omnibus (GEO) database; 3) Application of algorithms from hybrid models that combine RF and SVM; 4) Simulation using Python software; and 5) Based on the results obtained, conclusions and suggestions are taken which aim to provide input on the development of further research.

3. Evaluation

We implement the RFE-SVM and RF-SVM in python using library SVM, LibSVM [11] anda scikit library [12]. All test were run under Windows 8 on and up to Intel Core i3 processor.

3.1. Experiment Setup and Result

Before conducting classification modeling, gene expression data is divided into training and data testing first. Classification modeling is built using training data, and then the former model will be validated using testing data. It is done randomly with a composition of 75% as training data and 25% as testing data. So there are 36 samples in the training and 12 samples on the testing data as can be seen in Table 1.

Table 1. Divided of gene expression data

	Training	Testing	Total
CKD	18	8	26
Non-CKD	18	4	22
Total	36	12	48

There are two type of classification method we will to do, they are *Support Vector Machine (*SVM) with feature selection using *Recursive Feature Elimination* (RFE) and SVM with feature selection using *Feature Importance* from *Random Forest*.

3.1.1. RFE-SVM. SVM has an equation similar to ordinary regression, as follows:

$$\boldsymbol{w}\boldsymbol{x} + b = 0$$

where $\boldsymbol{w}$ is a vector that is perpendicular to the hyperplane, $\boldsymbol{x}$ is an input vector and b is a constant. In equations, $\boldsymbol{w}$ is commonly referred to chosen as the coefficient of $\boldsymbol{x}$. The RFE method selects variables based on the highest $\boldsymbol{w}$ value. The variable is 1% of the total variables, namely 193 variables out of 19312 total variables available. It can be seen on Figure 1 below.

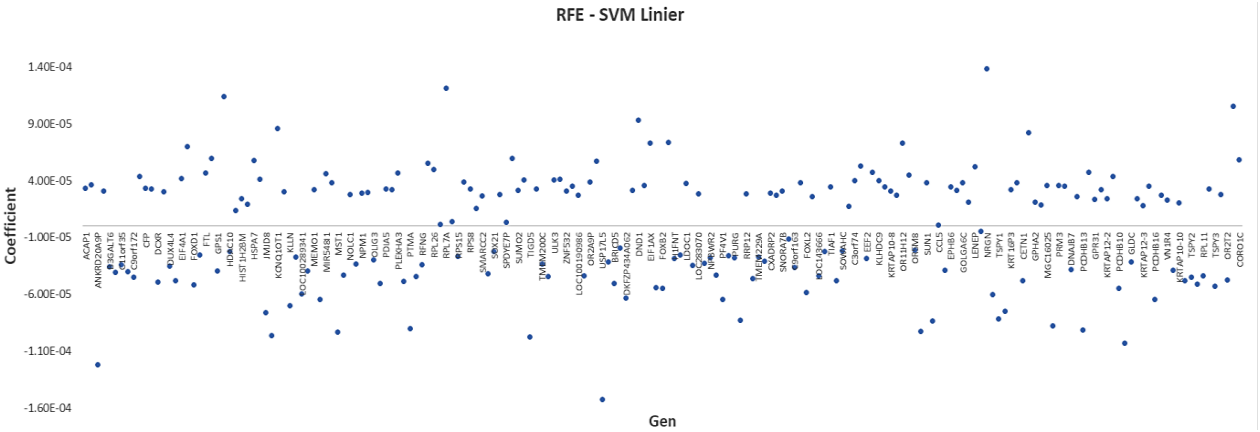

Figure 1. Coefficient value of variable from RFE – SVM

After that, the validation result from RFE – SVM on the testing data provide confusion matrix table (Table 2) as follows,

Table 2. Confusion Matrix RFE – SVM

		Prediction	
		CKD	Non CKD
Actual	CKD	6	2
	Non CKD	1	3

Based on the confusion matrix above, it was found that from 6 out of 8 CKD samples correctly predicted and 3 out of 4 Non CKD samples were also correctly predicted. So that the accuracy obtained is

$$\frac{6+3}{6+2+1+3} = \frac{9}{12} = 0.75$$

or 75%.

3.1.2. RF-SVM. If RFE-SVM. selects a variable based on its coefficient value, then RF-SVM selects variables based on feature importance obtained from Random Forest. Feature importance referred to here is a value that describes how much a variable contributes to reducing the level of impurity of the classification trees formed. The higher of the value is more better. The selected variable is a variable that has a value above the mean value of the overall importance feature. Based on Random Forest, the number of variables that have importance features greater or above the average is 311. It can be seen on Figure 2 below.

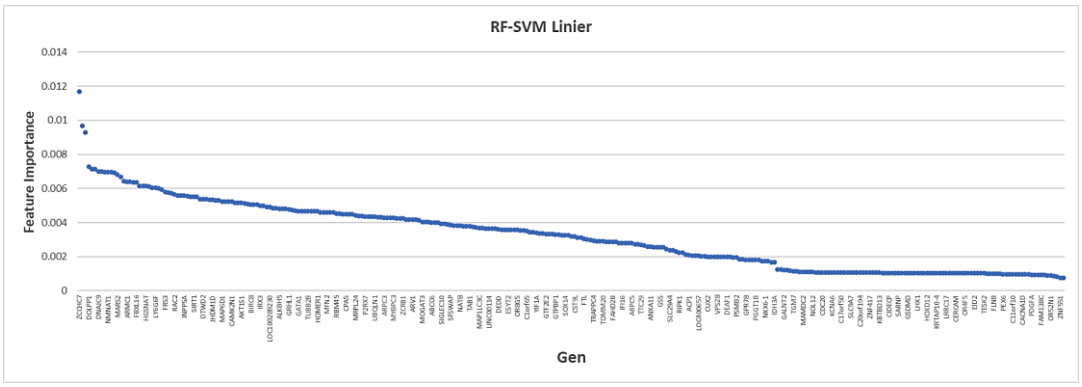

Figure 2. Feature Importance 311 from Random Forest.

After that, the selected variables from the Random Forest are used as input variables or predictors in SVM. The results of the RF - SVM model validation on testing data provide a confusion matrix table (Table 3) as follows,

Table 3. Confusion Matrix RF – SVM

		Prediction	
		CKD	Non CKD
Actual	CKD	8	0
	Non CKD	2	2

based on the confusion matrix above, it was found that from all CKD samples correctly predicted and 2 out of 4 Non CKD samples also predicted correctly. So that the accuracy obtained is

$$\frac{8+2}{8+0+2+2} = \frac{10}{12} = 0.834$$

or 83.4%. If we compare both of approachement, we can see that RF-SVM are more accurate than RFE-SVM.

4. Conclusion

We already make a hybrid model combining RF and SVM based on latest research to classify gene expression of CKD. The numerical test results show that choose RF-SVM is more accurate than the RFE-SVM. The accuracy of RF-SVM alghoritm is 83.4% than RFE-SVM algorithm is 7 %.

In the future, we want to provide more accurate algorithm for classifyng gene expression data on large dataset. Beside that, we want to comparisons with other algorithm or maybe make a new hybrid model using deep learning [13] that can study more stable representations for rare cancers, can improve cancer diagnosis performance even if the expression data is inadequate.

References

[1] Beckerman, Pazit, Chengxiang Qiu, Jihwan Park, Nora Ledo, Yi-An Ko, Ae-Seo Deok Park, Sang-Youb Han, Peter Choi, Matthew Palmer, Katalin Susztak. 2017. Human Kidney Tubule-Specific Gene Expression Based Dissection of Chronic Kidney Disease Traits, EBioMedicine 24, 267-276

[2] Brazma, A., et all. 2003. ArrayExpress a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 31(10), 68-71

[3] Edgar, R., Domrachev, M., Lash, A.E. 2002. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30(1), 207–210

[4] Slonim, Donna K. dan Itai Yanai. 2017. Getting Started in Gene Expression Microarray Analysis,

PLoS Computational Biology, Vol. 5, No. 10, 1-4
[5] Vapnik, V. 1998. Statistical Learning Theory, vol 1. Wiley, New York
[6] Huynh, PH., VH Nguyen., and TN Do. 2018. A Coupling Support Vector Machines with the Feature Learning of Deep Convolutional Neural Networks for Classifying Microarray Gene Expression Data. In: Sieminski A., Kozierkiewicz A., Nunez M., Ha Q. (eds) Modern Approaches for Intelligent Information and Database Systems. Studies in Computational Intelligence, vol 769. Springer, Cham
[7] Dubey, Akash. 2018. Feature Selection Using Random Forest: The Wisdom of Crowds. Towards Data Science
[8] Robin Genuer, Jean-Michel Poggi, Christine Tuleau-Malot. Variable selection using Random Forests. Pattern Recognization Letters, Elsevier, 2010, 31(14),pp.2225-2236 <hal-00755489>
[9] Sandri, Marco, Paola Zuccolotto. 2006. Variable Selection Using Random Forests. ResearchGate, Italy
[10] Nitze, Ingmar , Urs Schulthess. 2012. Comparison of Machine Learning Alghorithms Random Forest, Artificial Neural Networkand Support Vector Machine to Maximum Likelihood for Supervised Crop Type Classification.
[11] Chang, C.C., Lin, C,J. 2012. LibSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 27:1-27:27.
[12] Pedregosa, F., et al. 2011. Scikit Learn: machine learning in pyython. J. Mach. Learn. Res. 12. 2825-2830
[13] Liao, Q. Y. Ding and Z.L. Jiang et al. 2018. Multi-Task Deep Convolutional Neural Network for Cancer Diagnosis, Neurocomputing. https://doi.org/10.1016/j.neucom.2018.06.084