Performance evaluation of naive Bayes and support vector machine in type 2 Diabetes Mellitus gene expression microarray data

by Lawi Armin

Submission date: 19-Feb-2020 11:51AM (UTC+0700)
Submission ID: 1259987169
File name: Ramdaniah_2019_J._Phys.__Conf._Ser._1341_042018.pdf (1,008.89K)
Word count: 4398
Character count: 22689

Performance evaluation of naive Bayes and support vector machine in type 2 Diabetes Mellitus gene expression microarray data

Performance evaluation of naive Bayes and support vector machine in type 2 Diabetes Mellitus gene expression microarray data

Ramdaniah[1*], A Lawi[2*] and S Syarif[1*]

[1]Department of Electrical Engineering, Universitas Hasanuddin, Jl. Poros Malino KM.6, Bontomarannu, Gowa 92171, Sulawesi Selatan, Indonesia.
[2]Department of Computer Science, Universitas Hasanuddin, Jl. Perintis Kemerdekaan KM.10, Makassar 90245, Sulawesi Selatan, Indonesia.

*Email: ramdaniahh@gmail.com, armin@unhas.ac.id , ssyariftuh376@gmail.com

Abstract. Type 2 Diabetes Mellitus (T2DM) is a metabolic disorder that the number of diabetics increases every year. So that prevention is needed by knowing the trigger of T2DM. Gene expression microarray data contains information of gene that can be used to determine the causes of T2DM. It is necessary to use certain techniques to analyze gene expression microarray data because it has a large amount of data and attributes. This study aims to evaluate the performance of algorithms in classifying gene expression microarray data. Algorithms that were used in this study were Naive Bayes, and Support Vector Machine (SVM). SVM used many kernels function such as Linear, Radial Basis Function (RBF), Polynomial, and Sigmoid. Information gain was used to select the features in GSE18732 dataset by choosing top 10, 20, 30, 40, and 50 features. Performance of algorithms was evaluated and compared by using 30% testing set and 20% testing set. The results of the study indicated that SVM using Polynomial kernel had a high performance if it was compared to other algorithms. It achieved 98.15% accuracy using 30% testing set and achieved 100% accuracy using 20% testing set.

1. Introduction

Type 2 Diabetes Mellitus (T2DM) is a metabolic disorder caused by the pancreas not producing enough insulin or the body cannot use the insulin that is produced effectively so that there is an increase in the concentration of glucose in the blood[1]. In 2013, the International Diabetes Federation (IDF) estimated the number of diabetics in the world to be 382 million and estimated that diabetics in 2035 would continue to increase to 592 million people[2].

Microarray technology is a technology in biomedicine that can be used to assess gene expression, genomic structure analysis, identification of genetic polymorphisms or detection of viruses, bacteria and pathogenic fungi[3]. So that gene expression data from these technologies can be used to analyze genes that affect T2DM disease and classify T2DM samples with normal samples. However, analyzing datasets from microarray gene expression data is a challenge in data mining techniques because the data is very large and consists of thousands of attributes[4]. It is necessary to use the right method to analyze very large data.

Research on disease classification using microarray gene expression data has been done before. The study was conducted by identifying T2DM disease in microarray gene expression data using the

Biomarker Module (32 genes) with an accuracy of 84.79% and comparing it with the Support Vector Machine (SVM) method combined with Recursive Feature Elimination (RFE) with an accuracy of 73. 24%[5]. Al-Sabti et.al [6] classified T2DM in several gene expression data, one of them were GSE18732 data. The results showed that the accuracy of the classification using Identified Biomarkers achieved 84.71%, Module Biomarker (32 genes) achieved 92.39%, SVM-RFE achieved 67.39%, and Pathway Activity inference using Condition-responsive genes (PAC) achieved 84.78%. Kourou et. al [7] classified oral cancer in gene expression data using several methods. The results showed that the classification accuracy using the Naïve Bayes method achieved 70.80%, Bayesian Network (BN) achieved 83. 20%, SVM achieved 78.40%, Artificial Neural Network (ANN) achieved 83.90%, Adaboost achieved 82.5%, and the Random Forest achieved 89.30 %. Widiawati et. al [8] classified specific organ expression in maize using gene expression data. The researcher classified the data using K-Nearest Neighbor (KNN). The results showed that the accuracy using the proposed classifier achieved 93.33%.

Previous studies that related to the classification of gene expression data have not yet achieved high performance, so that we proposed and evaluated the system that can classify Type 2 Diabetes Mellitus in gene expression data using several classifiers such as Naïve Bayes, and Support Vector Machine (SVM) and use Information Gain (IG) to select features in gene expression microarray data to achieve high performance in classifying data.

2. Materials and Methods

Several stages and algorithms are needed in classifying GSE18732 data. Based on Figure 1, the data that used was GSE18732 gene expression data and then the data was processed at the pre-processing stage. Data transformation was carried out during the pre-processing stage using logarithmic transformations. Furthermore, the data were normalized using quantile normalization. In the feature selection stage, the algorithm that used was Information Gain. Then the data entered the resampling stage using the SMOTE algorithm.

The method that used in classification were Naïve Bayes, and SVM using Linear, RBF, Polynomial, and Sigmoid kernels. Furthermore, the performance of the algorithms was evaluated by calculating the value of accuracy, sensitivity, and specificity. The stages carried out in the study are as follows:

2.1. Dataset

The data that used is publicly accessible data provided by the National Center for Biotechnology Information (NCBI) [9]. Data can be accessed at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18732. The input data that used consisted of 118 samples and 25770 features.

2.2. Logarithmic Transformation

The method used in the pre-processing stage was Logarithmic transformation. Logarithmic transformation was used to change the scale of measurement of original data into other forms so that data can fulfill the assumptions underlying the various analysis. Logarithmic transformation (usually base 2) is often applied to microarray data. Equation (1) is a logarithmic transformation based on previous research[10].

$$X' = \log_2 X \quad (1)$$

Where the original data is X and X' is the data from the logarithmic transformation.

2.3. Quantile Normalization

Quantile normalization method was used in the pre-processing stage after processing data in transformation. The purpose of the quantile method is to make the distribution of the probe intensity for each array in the same set of arrays. This method is derived from the idea that quantile-quantile plots

show that the distribution of two data vectors is the same if the plot is a straight diagonal line and is not the same if in addition to the diagonal line[11]. Equation (2) is a formula for quantile normalization.

Given $q_k = (q_{k1}, \ldots, q_{kn})$ for $k = 1, \ldots, p$ is a vector from quantile kth for all n array $q_k = (q_{k1}, \ldots, q_{kn})$ and $d = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$ is a diagonal.

$$proj_d q_k = \left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n} q_{kj}, \ldots, \frac{1}{\sqrt{n}}\sum_{j=1}^{n} q_{kj}\right) \quad (2)$$

2.4. Feature Selection

In the feature selection stage, the algorithm used was Information Gain. Information Gain is a method for selecting features by ranking attributes. The formula for calculating entropy can be used using equations (3) and (4) based on the explanation in the study[12].

$$Entropy(X) = -\sum_x p(x) log_2 p(x) \quad (3)$$

$$Entropy(X|Y) = -\sum_Y p(x) log_2 p(x) \quad (4)$$

Where $p(x)$ represent probability function from X. Equation (5) is a formula for calculating information gain. Information Gain value is obtained from *Entropy(X)* minus *Entropy(X|Y)*.

$$InfoGain(X;Y) = Entropy(X) - Entropy(X|Y) \quad (5)$$

2.5. Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE is used to solve classification problems. SMOTE is useful for balancing datasets to improve the learning algorithm that will be used. Equation (6) is used to added new sample into dataset [13].

$$x_{new} = x + rand * (y[i] - x) \quad (6)$$

Where $i = 1, 2, \ldots, \mathrm{N}, \mathrm{rand}$ represents random number between 0 and 1. x_{new} represents new sample. x represents sample and $y[i]$ representing a sample of neighbors ith.

2.6. Naïve Bayes

Naïve Bayes is a classification method that predicts future opportunities based on previous experience. Naïve Bayes classifier is assumed that certain characteristics in the class does not have relation with other class[14]. Equation (7) is a formula of Naïve Bayes[15].

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (7)$$

Where $P(A|B)$ is posterior probability, $P(B|A)$ is likelihood, $P(A)$ is prior probability, and $P(B)$ is the marginal likelihood.

2.7. Support Vector Machine (SVM)

SVM aims to find the best classification function to distinguish examples from two classes in training data. In the SVM there is a hyperplane $f(x)$ in the middle of the class and separates two classes. After the f (x) function is found, new data items x_n can be classified by checking the function sign $f(x_n)$. x_n will be in a positive class if $f(x_n)$>0. Equation (8) is used to calculate the maximum margin of hyperplane[16]

$$L_p = \frac{1}{2}\|W\| - \sum_{i=1}^{t} \alpha_i y_i (w.x_i + b) + \sum_{i=1}^{t} \alpha_i \quad (8)$$

Where t is the number of training samples, y_i where $i = 1 \ldots t$ is a positive number, α_i is *multipliers Lagrange* and L_p is Lagrangia, b is constant and identify hyperplane.

The kernel functions that used are Linear kernel, Radial Basis Function (RBF), Polynomial, and Sigmoid. The following equation is kernel used.

$$k(x,y) = x^T y + c \quad (9)$$

$$k(x,y) = exp\left\{\frac{\|x-y\|^2}{2\sigma^2}\right\} \quad (10)$$

$$k(x,y) = (\alpha x^T y + c)^d \quad (11)$$

$$k(x,y) = \tanh(\alpha x^T y + c) \quad (12)$$

Equation (9) is a linear kernel that has the simplest function. Equation (10) is RBF kernel that σ is parameter that can be adjusted and used to calculate the performance of classifier. Equation (11) is the Polynomial kernel where x and y are feature vectors, α is *parameter* that can be adjusted, c is constant, d is degree of Polynomial. Equation (12) is Sigmoid kernel where a and c are *parameter* that can be adjusted.

2.8. Performance Evaluation

The confusion matrix is one way to analyze the performance of the classifier used. The confusion matrix is a two-dimensional matrix consisting of the actual class of an object and class as a result of classification. Confusion matrix often uses positive and negative[17]. Calculating accuracy, sensitivity, and specificity are needed to evaluate performance of the algorithms. Equation (13) is the formula to calculate accuracy, while equation (14) is used to calculate sensitivity and equation (15)is used to calculate specificity.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{(TP+TN+FP+FN)}} \quad (13)$$

$$\text{Sensitivity or Recall} = \frac{\text{TP}}{\text{(TP+FN)}} \quad (14)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{(TN+FP)}} \quad (15)$$

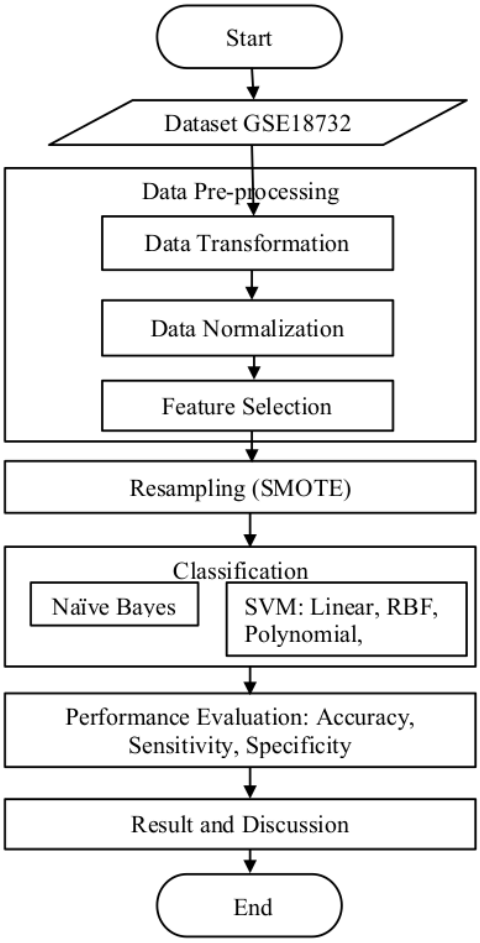

Figure 1. Research Methodology

3. Experiment Results

3.1. Feature Selection

The feature selection was done after the data was processed in the pre-processing stage. Information gain was used to select the best features of 25770 features. Then the 50 best features were selected based on the highest Information Gain value. Then the dataset was divided into 5 types, namely dataset1 which consists of 10 best features, dataset2 consists of 20 best features, dataset3 consists of 30 best features, dataset4 consists of 40 best features and dataset5 consists of 50 best features.

3.2. Resampling using SMOTE

The GSE18732 data consisted of 118 rows and had two classes, namely Type 2 Diabetes Mellitus and normal class. But the amount of data in both classes is different. Data on T2DM class consisted of 46 rows while data in normal class consisted of 72 rows. It can affect the performance of the classifier so it was necessary to use the SMOTE algorithm so that the amount of data in both classes was balanced.

The number of rows in each class after using the Smote algorithm was 92 rows. So that the total row of data after using the SMOTE algorithm was 184 rows.

4. Performance Evaluation and Discussion

Data was divided into two types of testing sets, namely data with 30% testing sets and data with 20% testing sets before classifying data. Both types of testing sets were used to compare the performance of the classifier that was used.

4.1. Algorithms Performance using 30% Testing Set

Table 1 shows the performance of Naïve Bayes. Based on table 1, the highest accuracy value using Naïve Bayes is 88.89% by classifying dataset1 which consists of 10 best features.

Table 1. Performance of Naïve Bayes using 30% Testing Set

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Dataset1	**88.89**	**88.89**	**88.89**
Dataset2	66.67	85.19	48.15
Dataset3	79.63	77.78	81.48
Dataset4	81.48	85.19	77.78
Dataset5	77.78	77.78	77.78

Table 2 shows the SVM performance using a Linear kernel with parameter cost = 10. The classification results using the Linear kernel achieved an accuracy of 92.59%, a sensitivity of 96.30% and a specificity of 88.89% in classifying datasets with 50 features.

Table 2. Performance of SVM Linear Kernel using 30% Testing Set

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Dataset1	79.63	92.59	66.67
Dataset2	79.63	96.30	62.96
Dataset3	79.63	66.67	92.59
Dataset4	83.33	100	66.67
Dataset5	**92.59**	**96.30**	**88.89**

Table 3 shows SVM performance using RBF kernels with cost = 1 and gamma = 1. Based on table 3, the highest accuracy achieved 94.44%, sensitivity achieved 88.89, and specificity achieved 100%.

Table 3. Performance of SVM RBF Kernel using 30% Testing Set

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Dataset1	**94.44**	**88.89**	**100**
Dataset2	94.44	96.30	92.59
Dataset3	79.63	100	59. 26
Dataset4	77.78	100	55.56
Dataset5	85.19	100	70.37

Table 4 shows the SVM performance using the Polynomial kernel with parameters cost = 10, gamma = 0.1, and degrees = 2. Table 4 shows that the highest accuracy using the Polynomial SVM kernel achieved 98.15% in classifying datasets with 30 features.

Table 4. Performance of SVM Polynomial Kernel using 30% Testing Set

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Dataset1	83.33	85.19	81.48
Dataset2	92.59	96.30	88.89
Dataset3	**98.15**	**100**	**96.30**
Dataset4	88.89	96.30	81.48
Dataset5	88.89	96.30	81.48

Table 5 shows the performance of SVM with the Sigmoid kernel using parameters cost = 1 and gamma = 0.1. The highest accuracy using the SVM Sigmoid kernel was 83.33% in classifying dataset1 which consists of 10 best features.

Table 5. Performance of SVM Sigmoid Kernel using 30% Testing Set

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Dataset1	**83.33**	**92.59**	**74.07**
Dataset2	53.70	62.96	44.44
Dataset3	59.26	59.26	59.26
Dataset4	64.81	74.07	55.56
Dataset5	72. 22	81.48	62.96

Figure 3 shows a comparison of the algorithms used in the dataset consisting of 70% training sets and 30% testing sets. The classification results show that SVM polynomial kernel produces the highest accuracy of 98.15%.

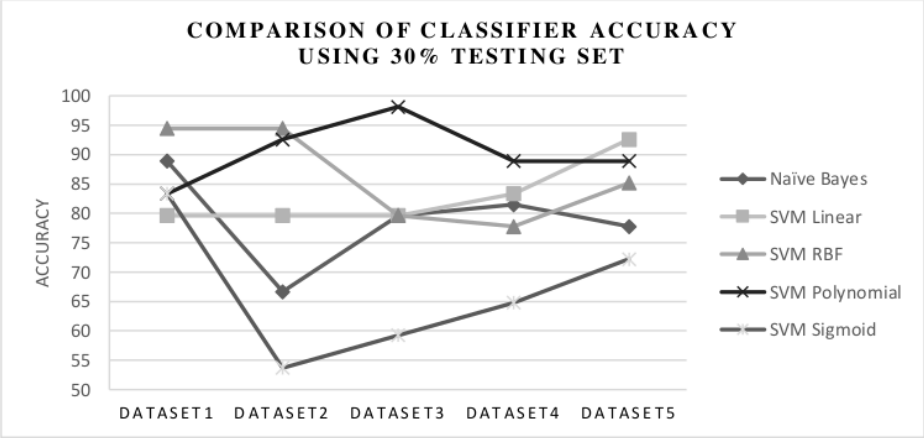

Figure 3. Comparison of Classifier Accuracy using 30% Testing Set

Figure 4 shows a comparison of the algorithms used in the dataset consisting of 70% training sets and 30% testing sets. SVM using polynomial kernel functions has the highest sensitivity of 100%.

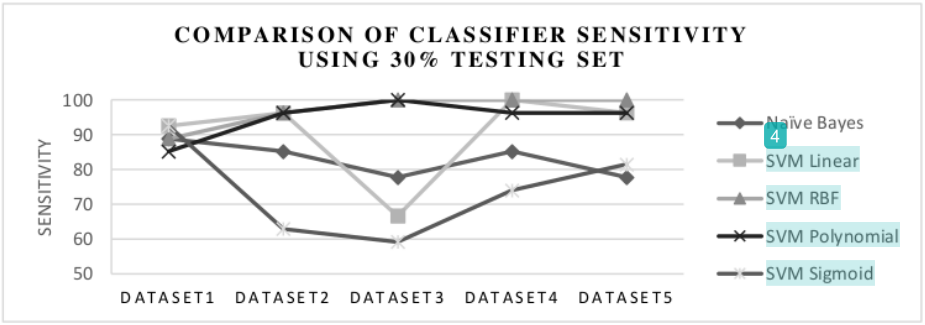

Figure 4. Comparison of Classifier Sensitivity using 30% Testing Set

Figure 5 shows a comparison of the algorithms used in the dataset consisting of 70% training sets and 30% testing sets. SVM using polynomial kernel functions has a specificity of 96.30%.

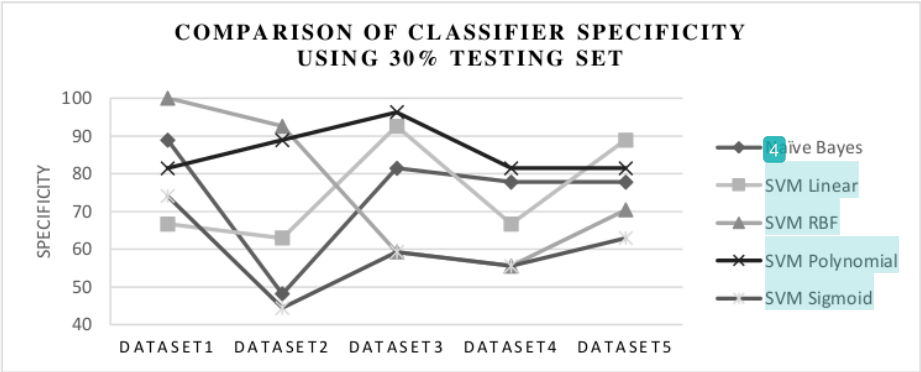

Figure 5. Comparison of Classifier Specificity using 30% Testing Set

4.2. Algorithms Performance using 20% Testing Set

The following is the performance of the algorithms that were used in classifying data using 20% testing sets. Table 6 shows the performance of Naïve Bayes on data with 20% testing sets.

Table 6. Performance of Naïve Bayes using 20% Testing Set

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Dataset1	88.89	88.89	88.89
Dataset2	80.56	100	61.11
Dataset3	**91.67**	83.33	100
Dataset4	86.11	88.89	83.33
Dataset5	80.56	88.89	72.22

Table 6 shows that the highest accuracy using Naïve Bayes on data with 20% testing set is 91.67%. Table 7 shows the classification results using the SVM Linear kernel. The highest accuracy using Linear kernel SVM is 94.44% with parameter cost = 10.

Table 7. Performance of SVM Linear Kernel using 20% Testing Set

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Dataset1	**94.44**	100	88.89
Dataset2	80.56	94.44	66.67
Dataset3	91.67	88.89	94.44
Dataset4	80.56	77.78	83.33
Dataset5	88.89	88.89	88.89

Table 8 shows SVM RBF kernels performance has accuracy 97.22% using cost = 10 and gamma = 0.1 parameters by classifying dataset that consists of 10 features.

Table 8. Performance of SVM RBF Kernel using 20% Testing Set

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Dataset1	91.67	88.89	94.44
Dataset2	94.44	100	88.89
Dataset3	88.89	88.89	88.89
Dataset4	**97.22**	94.44	100
Dataset5	97.22	100	94.44

Table 9 shows performance of SVM Polynomial kernel accuracy using cost = 10, gamma = 0.1, and degree = 2. It achieves 100% accuracy in classifying dataset with 40 features.

Table 9. Performance of SVM Polynomial Kernel using 20% Testing Set

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Dataset1	88.89	83.33	94.44
Dataset2	97. 22	100	94.44
Dataset3	94.44	100	88.89
Dataset4	**100**	100	100
Dataset5	94.44	100	88.89

Table 10 shows the performance of SVM Sigmoid kernel achieves 94.44% accuracy using cost = 1 and gamma = 0.1 parameters.

Table 10. Performance of SVM Sigmoid Kernel using 20% Testing Set

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Dataset1	**94.44**	100	88.89
Dataset2	58.33	66.67	50
Dataset3	75	66.67	83.33
Dataset4	66.67	83.33	50
Dataset5	61.11	61.11	61.11

Figure 6 shows the difference in classifier accuracy. Based on Figure 7, SVM using the Polynomial kernel has the highest accuracy in classifying data with 20% Testing Sets. Polynomial kernel SVM achieves 100% accuracy, 100% sensitivity, and 100% specificity in classifying datasets consisting of 40 best features.

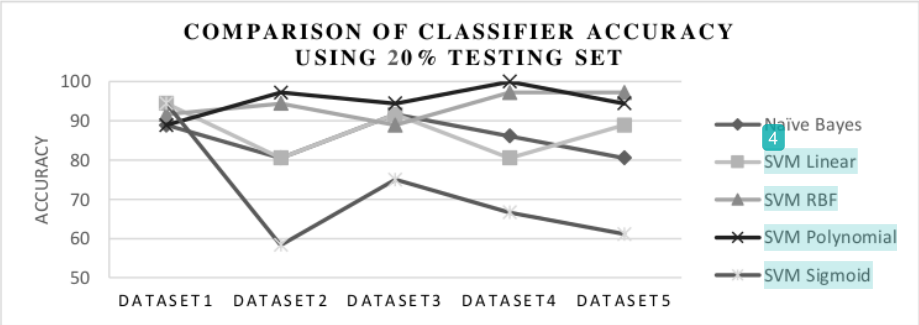

Figure 6. Comparison of Classifier Accuracy using 20% Testing Set

Figure 7 shows the difference in classifier sensitivity. SVM Polynomial kernel has the highest sensitivity of 100%.

Figure 7. Comparison of Classifier Sensitivity using 20% Testing Set

Figure 8 shows the difference in the specificity of the classifier. SVM Polynomial kernel has the highest sensitivity of 100%.

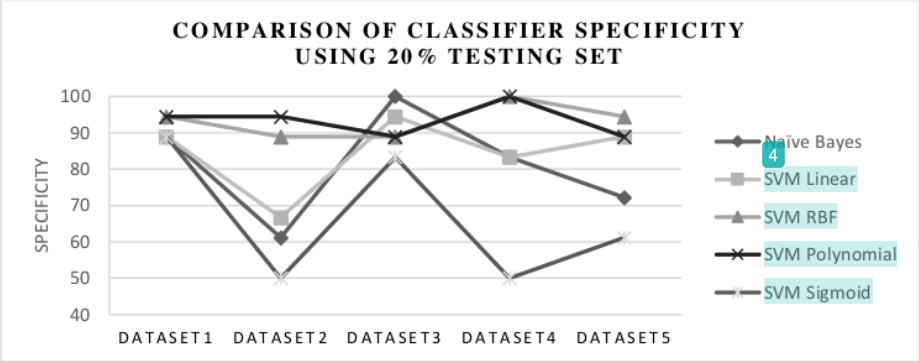

Figure 8. Comparison of Classifier Specificity using 20% Testing Set

4.3. Discussion

Performance evaluation of algorithms in classifying gene expression microarray data uses 30% testing set and 20% testing set. Table 11 shows the difference in classifier accuracy using different testing sets.

Table 11. Comparison of Algorithms Performance

	Accuracy (%)	
	30% Testing Set	20% Testing Set
Naïve Bayes	88.89	91.67
SVM Linear	92.59	94.44
SVM RBF	94.44	97.22
SVM Polynomial	**98.15**	**100**
SVM Sigmoid	83.33	94.44

Based on table 11, the performance of the algorithm was increased when using 20% testing sets. SVM with a polynomial kernel has the highest performance compared to other algorithms. Polynomial SVM achieves 98.15% accuracy using 30% testing set and 100% accuracy using 20% testing set.

5. Conclusion

This study evaluated the algorithms to classify gene expression microarray data into Type 2 Diabetes Mellitus class and Normal class. Information gain was used to select features that most affect T2DM. The datasets were divided into 5 datasets that use top 10 features, 20, 30, 40 and 50 features. Comparison of classifier performance was evaluated by using different testing sets, namely data with 30% testing set and 20% testing set. The results showed that the classification used 30% testing set and using Naïve Bayes achieved 88.89% accuracy, SVM Linear kernel achieved 92.59% accuracy, SVM using RBF kernel achieved 94.44% accuracy, SVM Polynomial kernel produced the highest accuracy of 98.15% and SVM using the kernel sigmoid achieved an accuracy of 83.33%. The classification that used 20% testing set and used Naïve Bayes achieved 91.67% accuracy, SVM Linear kernel achieved 94.44% accuracy, SVM uses RBF kernel achieved 97.22% accuracy, SVM using Polynomial kernel had 100% accuracy and SVM Sigmoid kernel achieved 94.44% accuracy. In other words, the classification of GSE18732 gene expression data achieved the highest accuracy by using the SVM method with the Polynomial kernel.

References

[1] W. Chen, S. Chen, H. Zhang, and T. Wu, "A hybrid prediction model for type 2 diabetes using K-means and decision tree," *Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS*, vol. 2017-Novem, no. 61272399, 2018.

[2] F. Aguirre *et al.*, *IDF Diabetes Atlas : sixth edition[J]*. International Diabetes Federation, 2013.

[3] J. Clausen Mork, *Microarray Technology*, vol. 1368, no. 1. New York, NY: Springer New York, 2016.

[4] X. R. Jenifer and R. Lawrance, "An adaptive classification model for microarray analysis using big data," *2016 Int. Conf. Comput. Technol. Intell. Data Eng.*, pp. 1–5, 2016.

[5] X. Zhang, L. Gao, Z. P. Liu, and L. Chen, "Identifying module biomarker in type 2 diabetes mellitus by discriminative area of functional activity," *BMC Bioinformatics*, vol. 16, no. 1, 2015.

[6] A. Al-Sabti, M. Zaibi, and S. Jassim, "An Integrative Omics Approach to Identify Sub-Network Biomarker in Type 2 Diabetes Mellitus," *Proc. - UKSim-AMSS 11th Eur. Model. Symp. Comput. Model. Simulation, EMS 2017*, pp. 53–58, 2017.

[7] K. Kourou, C. Papaloukas, and D. I. Fotiadis, "Identification of differentially expressed genes through a meta-analysis approach for oral cancer classification," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 3876–3879, 2017.

[8] I. Fitria Widiawati, H. Nugrahapraja, and R. Fajriyah, "K-Nearest Neighbor (KNN) Analysis on Genes Expression Datasets of Maize Nested Association Mapping (NAM) Showed Confident Classification on Organ-specific Expression," *Proc. - 2018 1st Int. Conf. Bioinformatics, Biotechnol. Biomed. Eng. BioMIC 2018*, vol. 1, pp. 1–3, 2019.

[9] A. M. Goldstein, "The NCBI Databases : an Evolutionist ' s Perspective," pp. 451–455, 2010.

[10] X. Cui, M. K. Kerr, and G. A. Churchill, "Transformations for cDNA Microarray Data," *Stat. Appl. Genet. Mol. Biol.*, vol. 2, no. 1, 2005.

[11] B. Bm, I. Ra, Astr, and S. T. P. M, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 2, pp. 185–193, 2003.

[12] M. Liang and W. Ahmad, "Breast Cancer Intelligent Diagnosis based on Subtractive Clustering Adaptive Neural Fuzzy Inference System and Information Gain," *2017 Int. Conf. Comput. Syst. Electron. Control*, no. x, pp. 152–156, 2017.

[13] Y. Weng, F. Deng, and G. Yang, *Smart Computing and Communication*, vol. 10135. Springer International Publishing, 2017.

[14] L. Marlina, M. Muslim, and A. P. U. Siahaan, "Data Mining Classification Comparison (Naïve Bayes and C4 . 5 Algorithms)," *Int. J. Eng. Trends Technol.*, vol. 38, no. 7, pp. 380–383, 2016.

[15] I. D. Dinov, *Data Science and Predictive Analytics: Biomedical and Health Applications Using R*. 2018.

[16] S. Chidambaram and K. G. Srinivasagan, "Performance evaluation of support vector machine classification approaches in data mining," *Cluster Comput.*, vol. 0123456789, pp. 1–8, 2018.
[17] A. C. Kakas, *Encyclopedia of Machine Learning and Data Mining*. 2016.

Performance evaluation of naive Bayes and support vector machine in type 2 Diabetes Mellitus gene expression microarray data

ORIGINALITY REPORT

16%	8%	14%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Yusran, Sultan, J A Isnam, Y S Akil. "The potential of electrical power generation based on organic waste utilization at Tamangapa landfill Makassar", Journal of Physics: Conference Series, 2019 Publication	5%
2	N A P Mangarengi, A Zubair, M A Abdurrahman. "Analysis of infrastructure needs and operational systems for traditional market solid waste management (A case study on Makassar – Niaga Daya traditional market)", IOP Conference Series: Earth and Environmental Science, 2020 Publication	1%
3	B.M. Bolstad, R.A Irizarry, M. Astrand, T.P. Speed. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias", Bioinformatics, 2003	1%

Publication

4	Shahnoor C. Eshan, Mohammad S. Hasan. "An application of machine learning to detect abusive Bengali text", 2017 20th International Conference of Computer and Information Technology (ICCIT), 2017 Publication	1%
5	digilib.uinsby.ac.id Internet Source	1%
6	D R Prehanto, A D Indriyanti, K D Nuryana, S Soeryanto, A S Mubarok. "Use of Naïve Bayes classifier algorithm to detect customers' interests in buying internet token", Journal of Physics: Conference Series, 2019 Publication	1%
7	Subekti, Aan Budi Setiawan, Abdul Hammid. "Simulation of Robot Arm for Diabetes Mellitus Patients", Journal of Physics: Conference Series, 2019 Publication	1%
8	Submitted to BMJ Group Student Paper	1%
9	Zhongxian Xu, Zhiliang Wang. "A Risk Prediction Model for Type 2 Diabetes Based on Weighted Feature Selection of Random Forest and XGBoost Ensemble Classifier", 2019	<1%

	Eleventh International Conference on Advanced Computational Intelligence (ICACI), 2019 Publication	
10	www.jci.org Internet Source	<1%
11	Advances in Intelligent Systems and Computing, 2014. Publication	<1%
12	Submitted to University of Macau Student Paper	<1%
13	Submitted to University of Ghana Student Paper	<1%
14	Mikko J. Alava, Kent Bækgaard Lauritsen. "Chapter 43 Branching Processes", Springer Science and Business Media LLC, 2009 Publication	<1%
15	mafiadoc.com Internet Source	<1%
16	Submitted to University of Stellenbosch, South Africa Student Paper	<1%
17	f1000research.com Internet Source	<1%
18	Adi L. Tarca, Roberto Romero, Sorin Draghici. "Analysis of microarray experiments of gene	<1%

expression profiling", American Journal of Obstetrics and Gynecology, 2006
Publication

19	journals.sagepub.com Internet Source	<1%
20	Lecture Notes in Computer Science, 2015. Publication	<1%
21	Submitted to CSU Northridge Student Paper	<1%
22	www.ijitee.org Internet Source	<1%
23	link.springer.com Internet Source	<1%
24	"PRICAI 2014: Trends in Artificial Intelligence", Springer Science and Business Media LLC, 2014 Publication	<1%
25	Submitted to Universiti Putra Malaysia Student Paper	<1%
26	Mustakim, Siti Syahidatul Helma, Ulya Ramadhani, GS Achmad Daengs, Rice Novita, Nuryanti, Sri Rahmawati Fitriatien. "Data Sharing Technique Modeling for Naive Bayes Classifier for Eligibility Classification of Recipient Students in the Smart Indonesia Program",	<1%

Journal of Physics: Conference Series, 2019

Publication

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		