International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 7 Issue 12 ǁ 2019 ǁ PP. 13-19

Support Vector Machine for Selecting Features for Elimination Breast Cancer

Bhaskar Gupta, Amar Kumar, Subrat Kumar Panda, Sairukmini Sahu
Department of Electronics and Communication Engineering, NM Institute of Engineering and Technology,Bhubaneswar , Odisha
Department of Electronics and Communication Engineering, Raajdhani Engineering College,Bhubaneswar,Odisha
Department of Electronics and Communication Engineering,Aryan Institute of Engineering and Technology Bhubnaeswar , Odisha
Department of Electronics and Communication Engineering,Capital Engineering College,Bhubaneswar,Odisha

Abstract
Deaths caused by cancer are expected to continue to increase, especially for prostate cancer and breast cancer. Both diseases are the most common types of cancer for men and women in the world. The number of deaths can be reduced by the early detection of using machine learning. One of them is the classification of prostate cancer and breast cancer data. Cancer data used has a variety of features, but not all features are essential features. In this study, we used Support Vector Machine-Recursive Feature Elimination (SVM-RFE) and One-Dimensional Naïve Bayes Classifier (1-DBC) as feature selection methods. In both methods, it will get a ranking for each feature. The use of these two methods in the classification of prostate cancer and breast cancer data results in a high level of evaluation. Both of these methods can produce an accuracy rate of 95.61%, the precision of 100%, and recall of 93.61%. In additional evaluation, SVM-RFE has lower running time than 1-DBC.

Keywords: *1-DBC; Cancer; Feature Selection; SVM-RFE; SVM;*

I. INTRODUCTION

Cancer is a disease caused by abnormal cell growth. These cells exist because of the changes in gene expression, then they will be developed into a population of cell that can attack specific tissues [1]. The changes in genes appearance that extend into cells can cause a functional shift of these cells. This is very dangerous because it can cause death. Based on the Global Cancer (GLOBOCAN) statistics [2] part of the International Agency of Research on Cancer (IARC) in 2018, there are 18.1 million cases of cancer in the world and 9.6 million of them have died. In the 18.1 million cases of cancer, the second most common cases experienced by men are prostate cancer cases, while the most common cancer cases experienced by women are breast cancer cases. In total, it is estimated that there are 43.8 cases for patients alive diagnosed with cancer within the past 5 years, which is also known as 5-year prevelance. This data was taken based on 185 countries with 36 types of cancer. Until now, there has not been found a way to treat cancer efficiently and thoroughly.

Prostate cancer is the type of cancer where there is an uncontrolled growth of cancer cells formed in prostate tissue. This type of cancer is the most common cancer in men, and the case will continue to increase in many countries[3]. Based on GLOBOCAN 2018 statistics [2], there are 1.3 million cases of prostate cancer or around 7.1% of all cancer cases in the world. Men aged from 70 to 79 years is the group who suffer the most from this disease.

Breast Cancer is the type of cancer where there is an uncontrolled growth of cancer cells formed in breast tissue. The growth of cancer cells will form lumps that can spread to other tissues within the body, which is also known as malignant tumor. Most cancers in the breast begin to grow in the glands for milk production called lobules, and in the channels that connect the lobules with nipples [4]. Based on GLOBOCAN 2018 statistics [2], there were 2.1 million breast cancer cases, and 626 thousand of them did not survive. In total cases of 5-year breast cancer prevalence, it is estimated that there are 6.9 million cases. Breast cancer is the most popular case of cancer experienced by women in five continents. The number of cases of breast cancer can increase if there is no appropriate treatment.

The number of deaths is expected to increase continuously over time. As many as 30% until 50% of these cancers can prevent in various ways [5]. One of them is early detection using artificial intelligence and machine learning [6]. Artificial intelligence (AI) is an important technology that supports daily social life, economic activities, and also health sector [7]. Machine learning is a branch of science that implements mathematical algorithms into computer programming to identify data patterns and improve performance iteratively. Machine learning applications have solved many problems such as, prediction of cancer patients and predictions of corporate bankruptcy [8,9].

Cancer data has many features that contain information about the cancer itself. However, not all existing features are relevant features. So, the feature selection process is needed. Feature selection is an important part of optimizing classifier performance [10]. Feature selection is the process of determining the best features that can represent data. The benefit of feature selection in machine learning is reducing the amount of data needed to reach the learning stage, increasing the value of predictive accuracy, more concise and easy-to-understand data, and reducing execution time [11].

In the field of health, many methods have been carried out to diagnose prostate cancer and breast cancer. But in this study, we used computational techniques by applying machine learning. The method that is proposed is Supporting Vector Machine-Recursive Features Elimination (SVM-RFE) and One-Dimensional Naïve Bayes Classifier (1-DBC). In this study, we will compare the SVM-RFE and 1 -DBC, as feature selection methods, by using Support Vector Machine (SVM) as a classifier. SVM is known as a binary classification method that can maximize classification results. It is expected that comparing the feature selection methods and classification methods would give significant contribution to the health sector, especially in diagnosing prostate cancer and breast cancer. Previous studies on the classification of prostate cancer and classification of breast cancer have been carried out with various methods such as Logistic Regression and Decision Tree, Convolutional Neural Network [12,13,14]. the SVM method has used for Classification of Schizophrenia, Insurance, and Classification of Hyperspectral Imagery15,16,17.

II. MATERIAL AND METHODS

2.1. Data

The data used in this study were data on prostate cancer and breast cancer, that is obtained from the Kaggle site. First, prostate cancer data consisted of 8 features. The data consisted of 100 observations, in which 62 data were labeled as malignant cancer and 38 data were labeled as benign cancers. Meanwhile, breast cancer data consisted of 30 features. The data consisted of 569 observations, in which 212 cancers were malignant cancer, and 357 were benign cancer. The list of features for each data can be seen in Table 1.

Table 1. The list of features of prostate cancer and breast cancer

Data	**Features**	
Prostate	Radius (16.85;4.88)	Smoothness (0.1;0.01)
	Texture (18.23;5.19)	Compactness (0.12;0.06)
	Perimeter (96.78;23.67)	Symmetry (0.19;0.03)
	Area (702.88;319.71)	Fractal Dimension (0.06;0.008)
Breast	Radius Mean (14.12;3.52)	Compactness Se (0.02;0.001)
	Texture Mean (19.28;4.3)	Concavity Se (0.032;0.03)
	Perimeter Mean (91.96;24.29)	Concave Points Se (0.01;0.006)
	Area Mean (654.88;351.91)	Symmetry Se (0.02;0.008)
	Smoothness Mean (0.09;0.01)	Fractal Dimension Se (0.03;0.002)
	Compactness Mean (0.1;0.05)	
	Concavity Mean (0.08;0.07)	Radius Worst (16.26;4.83)
	Concave Points Mean (0.04;0.03)	Texture Worst (25.67;6.14)
	Symmetry Mean (0.18;0.02)	Perimeter Worst (107.26;33.6)
	Fractal Dimension Mean (0.06;0.007)	Area Worst (880.6;569.3)
		Smoothness Worst (0.13;0.02)
	Radius Se (0.4;0.27)	Compactness Worst (0.25;0.15)
	Texture Se (1.21;0.55)	Concavity Worst (0.27;0.2)
	Perimeter Se (2.86;2.02)	Concave Points Worst (0.11;0.065)
	Area Se (40.33;45.49)	
	Smoothness Se (0.007;0.003)	Symmetry Worst (0.29;0.06)
		Fractal Dimension Worst (0.08;0.01)

Note: (Mean;Std Deviation)

2.2. Support Vector Machine-Recursive Feature Selection

This method is a combination of SVM and RFE. RFE is a method that works by selecting dataset features recursively based on the smallest feature value. With this RFE concept, SVM-RFE works by eliminating irrelevant features in each iteration, namely the lowest weight feature [18]. For this reason, there needs to be a weighting of features starting from features with the highest weight values to features with the smallest weight values.

In general, this method is divided into three stages:

Step 1. Train the dataset using SVM-train to calculate the weight of all features.

Based on [18], to calculate the weight of each feature, we can use SVM-train which produces an α classifier. The weight function is defined

Step 2. Calculate the ranking criterion

To sort features based on their weight, criterion ranking is needed. The function of criterion ranking is defined as follows:

(2)

Step 3. Rank the features based on the weight value and eliminate one feature with the smallest weight value in each iteration

2.3. One Dimensional-Naïve Bayes Classifier (1-DBC)

Naïve Bayes is a simple classification algorithm. Naïve Bayes is a classification technique based on the Bayes theorem with the assumption of independence between its features [19]. In that sense, assuming the presence of certain features in a class is not related to the presence of other features. In principle, the 1-DBC method is a classification method. This method generally classifies a particular data into a particular class. The development of this method is to classify with only one dimension. The point is that the feature used to predict data is only one feature. Each accuracy will be calculated with Naïve Bayes Classifier.

2.4. Support Vector Machines

The basic concept of the SVM method is to form an optimal plane or hyperplane that separates data into each class. The optimal hyperplane is a field that separates data into its class and is located perpendicular to the closest pattern. Patterns are dots that describe a dataset. To find the optimal hyperplane, the maximum margin will be sought. The Margin is the distance between the hyperplane and the closest pattern from each class. The pattern that is located

2.5. Performance Evaluation of Model

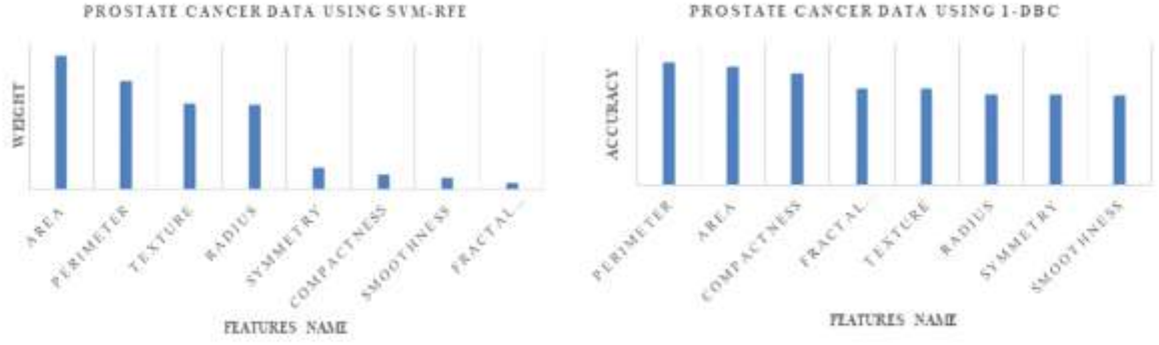
A classification model will map data to prediction classes. There will be four possibilities. If the data has a positive label and classified as positive, then it counts as true positive (TP); if classified negative, count as false negative (FN). If a data has a negative label and is classified as negative, then it counts as true negative (TN); if classified as positive, count as false positive (FP). From a classification model (classifier) and a data set, a 2 × 2 confusion matrix (also called a contingency matrix) can be formed and state the disposition of the data set.

with TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative. Precision is a measure used to calculate how exact a model is predicted to be positive, or how many of them are truly positive. Recall is a measure used to calculate how many real positive things are captured by the model and labeled Positive.

III. RESULTS

In this section, results and analysis of classification of Prostate Cancer and Breast Cancer data with SVM will be covered and the software Python 3.6 was used to form SVM. In this paper, we use the method of Support Vector Machine- Recursive Feature Elimination (SVM-RFE) and One-Dimensional Naïve Bayes (1-DBC) as a feature selection method. From feature selection, a ranking for prostate cancer and breast cancer data will be obtained for each feature selection method. Figure 1(a) shows the results of the ranking score that are obtained from the application of the SVM-RFE method, using Equation (2) for the features selection of prostate cancer. The ranking order of each feature in the dataset can be obtained from the these graphs. The first highest feature is the Area feature that has a weight of 169664594.74, while the feature that has the lowest weight is the

Fractal Dimension feature, which has a weight of only 2.46. Figure 1(b) shows the results of accuracy that are obtained from the application of the 1-DBC method using Equation (3) for the features selection of prostate cancer. The ranking order of each feature in the dataset can be obtained from the these graphs.The first highest feature is the Perimeter feature that gets an accuracy of 77.2%, while the feature that has the lowest rating is the Smoothness feature which only gets an accuracy of 56.4%.

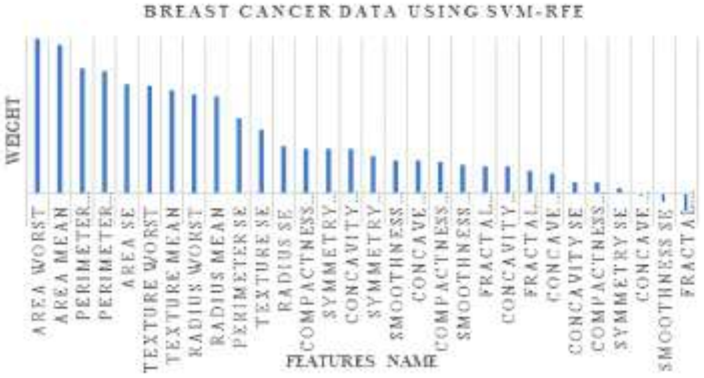


(a) (b)

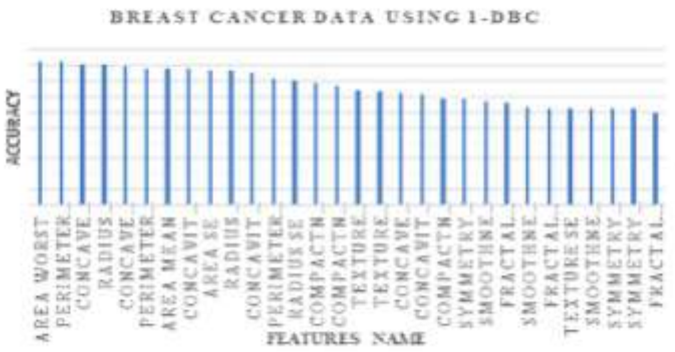
Fig 1. Features Ranking of Prostate Cancer Data with SVM-RFE and 1-DBC Method

While, Figure 2(a) shows the results of accuracy that are obtained from the application of the SVM-RFE method for the features selection of breast cancer. Through these images can be known as the order or ranking of each feature contained in the dataset. The first highest feature is the Area Worst feature that has a weight of 6816715655.41, while the feature that has the lowest rating is the Fractal Dimension feature, which has a weight of only 0.09. Figure 2(b) shows the results of accuracy that are obtained from the application of the 1-DBC method for the features selection of breast cancer. The ranking order of each feature in the dataset can be obtained from the these graphs. The first highest feature is the Area Worst feature, which has an accuracy of 92.8%, while the feature that has the lowest rating is the Fractal Dimension feature, which only has an accuracy of 59.6%.

From the ranking of features for each data with the SVM-RFE and 1-DBC methods will be taken 25%, 50%, 75%, 100% of the best features of each data with each method. The accuracy, precision, and recall will be calculated by using the SVM as the classification method, by splitting test data by 20% and 25%. Later, the results of the comparison of accuracy, precision, and recall will be presented. The comparison of evaluation model for each dataset cancer can be seen in Table 2.



(a)



(b)

Fig 2**.** Features Ranking of Breast Cancer Data with SVM-RFE and 1-DBC Method

Based on Table 2, it can be seen that for the testing data of 20% in prostate cancer data, method 1-DBC produces the best evaluation. While in terms of time, both the 1-DBC and SVM-RFE methods are not much different, while for breast cancer with 20% test data, SVM-RFE and 1-DBC methods produce accuracy, precision, and recall obtained by Equation (9) that is not much different, but for running time the SVM- RFE method has lower running time than 1-DBC for both datasets. Based on Table 3, it can be seen that for the test data of 25% in prostate cancer data, both the 1-DBC and SVM-RFE methods produced the same evaluation. Meanwhile, in terms of time the 1-DBC method is slightly superior. Lastly, for breast cancer data with test data of 25%, the SVM-RFE method produces the best accuracy, precision, recall and time, because SVM-RFE is an embedded feature selection method in which the feature selection process is carried out together with the classification process. While method 1-DBC is a filter type feature selection method where the feature selection process is separate from the classification process.

Table 2. Comparison of Prostate Cancer & Breast Cancer Data with 20% Testing Data

Type of Cancer	Number Of Features	1-DBC				SVM-RFE			
		Accuracy	Precision	Recall	Time	Accuracy	Precision	Recall	Time
Prostate Cancer	**25%**	**90**	**100**	**88.23**	**0.0036**	90	100	88.23	0.0047
	50%	**90**	**100**	**88.23**	0.0040	85	100	82.35	**0.0034**
	75%	**90**	**100**	**88.23**	**0.0034**	85	100	82.35	0.0038
	100%	85	100	82.35	0.0037	85	100	82.35	**0.0035**
Breast Cancer	**25%**	95.61	95.65	93.61	0.0103	95.61	95.65	93.61	**0.0088**
	50%	94.73	95.56	91.48	0.0114	**95.61**	95.65	**93.61**	**0.0100**
	75%	95.61	95.65	93.61	0.0110	95.61	95.65	93.61	0.0110
	100%	95.61	95.65	93.61	0.0189	95.61	95.65	93.61	**0.0116**

Table 3**.** Comparison of Prostate Cancer Data & Breast Cancer Data with 25% Testing Data

Type of Cancer	Number of Features	1-DBC				SVM-RFE			
		Accuracy	Precision	Recall	Time	Accuracy	Precision	Recall	Time
Prostate Cancer	25%	80	100	76.19	0.0051	80	100	76.19	**0.0036**
	50%	80	100	76.19	0.0035	80	100	76.19	0.0035
	75%	80	100	76.19	0.0035	80	100	76.19	0.0038
	100%	80	100	76.19	0.0037	80	100	76.19	0.0040
Breast Cancer	**25%**	94.4	92.45	92.45	0.01	**95.1**	**94.23**	**92.50**	**0.009**
	50%	94.4	92.45	92.45	0.01	**95.1**	**94.23**	**92.50**	**0.010**
	75%	95.1	94.23	92.45	0.01	**95.1**	**94.23**	**92.50**	**0.010**
	100%	95.1	94.23	92.45	0.01	**95.1**	**94.23**	**92.50**	**0.010**

IV. CONCLUSION

The classification of medical data on prostate cancer and breast cancer can be done by selecting features using the SVM- RFE and 1-DBC approaches. SVM-RFE and 1-DBC were able to choose features with a level of accuracy, precision, and recall that was not significantly different. However, the SVM-RFE has a smaller running time than 1-DBC. For further research, it is suggested to use large-scale medical data to see SVM-RFE performance.

REFERENCES

[1]. Ruddon, R. (2007) "Cancer Biology." Oxford: Oxford University Press.

[2]. IARC Global Cancer Observatory. 2018.

[3]. Cuzick J. et al. (2014) "Prevention and Early Detection of Prostate Cancer." *The Lancet Oncology* 15: 484-492.

[4]. NCBI. What is Cancer? Accessed 26 March 2019 see https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

[5]. WHO. Accessed 26 March 2019 see https://www.who.int/cancer/prevention/en/

[6]. H. Lu, Y. Li, T. Uemura, H. Kim, S. Serikawa. (2018) "Low illumination underwater light field images reconstruction using deep convolutional neural networks." *Future Generation Computer Systems* 82: 142-148.

[7]. H. Lu, Y. Li, M. Chen, H. Kim, S. Serikawa. (2018) "Brain Intelligence: go beyond artificial intelligence. Mobile Networks and Applications" 23: 368-375.

[8]. Lynch, C.M. et al. (2017) "Prediction of Lung Cancer Patient of Survival via Supervised Machine Learning Classification Techniques." *International Journal of Medical Informatics* 108: 1-8.

[9]. Rustam, Z., Yaurita, F. (2018) "Insolvency Prediction in Insurance Companies Using Support Vector Machines and Fuzzy Kernel C-Means." *Journal of Physics: Conference Series*.

[10]. Wang, S., Li, D., Song, X., Wei, Y., & Li, H. (2011) "A feature selection method based on improved fisher's discriminant ratio for text sentiment classification." *Expert Systems with Applications* 38; 8696-8702.

[11]. Hall, M.A. (1999) "Correlation-based Feature Selection for Machine Learning." University of Waikato. Thesis.

[12]. Virtanen, A et al. (1999) "Estimation of prostate cancer probability by logistic regression: free and total prostate-specific antigen, digital rectal examination, and heredity are significant variables" 45: 987-994.

[13]. Wu, Chun-Hui. et al. (2009) "Applying Data Mining for Prostate Cancer."

[14]. Alom, M.Z et. al. (2018) “Breast Cancer Classification from Histopathological Images with Inception Recurrent Residual Convolutional Neural Network.” University of Dayton; USA.

[15]. Rampisela, T.P and Rustam, Z. (2018) “Classification of Schizophrenia data using Support Vector Machine (SVM).” University of Indonesia, Indonesia.

[16]. Rustam, Z., Ariantari, N.P.A.A. (2018) “Support Vector Machines for Classifying Policyholders Satisfactorily in Automobile Insurance.” *Journal of Physics: Conf. Series* 1028.

[17]. Z. Chunhui, G. Bing, Z. Lejun, and W. Xiaoqing. (2018) “Classification of Hyperspectral Imagery Based on Spectral Gradient, SVM and Spatial Random Forest.” *Infrared Physics and Technology* 95, 61–69.

[18]. Guyon, I., Weston, J., Barnhill, S., Vapnik, V. (2002) “Gene selection for cancer classification using support vector machines.” Mach. Learn 46: 389–422.

[19]. Bishop, C.M. (2006) “Pattern Recognation and Machine Learning.” New York: Springer.

[20]. Srivastava, D.K and Bhambhu, L. (2005) “Data Classification Using Support Vector Machine.” *Journal of Theoritical and Applied Information Technology*.