© 2017 IJSRST | Volume 3 | Issue 8 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section : Science and Technology
doi : https://doi.org/10.32628/IJSRST

Breast Cancer Diagnosis using Support Vector Machines and Feature Selection

Tohida Rehman
Department of Computer Science, Surendranath College (C.U), Kolkata, India

Abstract: This study investigates the use of support vector machines (SVMs) in conjunction with feature selection for the purpose of breast cancer diagnosis. The aim of the research is to select the most relevant features that contribute to accurate classification and to use them to train the SVM model. The proposed approach is evaluated using a publicly available Wisconsin breast cancer dataset (WBCD) [2] and compared with other classification methods. The results show that the SVM model with feature selection outperforms other classification methods in terms of accuracy and provides a promising approach for breast cancer diagnosis. The study's findings demonstrate that the SVM model, when combined with feature selection, achieved a high classification accuracy of 98.25%. This model utilized only nine features. The accuracy achieved by the SVM model also indicates its potential to assist in the early detection of breast cancer, which is crucial in improving patient outcomes.

Keywords : Breast cancer diagnosis; Support vector machines; Feature selection; Wisconsin breast cancer diagnosis data

I. INTRODUCTION

Breast cancer is a significant health concern globally, with an increasing incidence rate. Despite this, advancements in diagnostic techniques and treatment methodologies have led to improved outcomes for breast cancer patients. Early detection and accurate diagnosis play a crucial role in the success of breast cancer treatment, particularly when the cancer has not metastasized [1]. Women diagnosed with breast cancer can survive for many years with proper treatment, highlighting the importance of continued research and innovation in breast cancer diagnosis and treatment.

Breast cancer diagnosis is a critical task in healthcare, and accurate and early detection is crucial for successful treatment. Support vector machines (SVMs) [3] have emerged as a popular machine learning method for breast cancer diagnosis due to their ability to handle high-dimensional data and non-linearity. Feature selection is a technique that involves selecting a subset of relevant features that contribute to accurate classification. The Wisconsin breast cancer dataset (WBCD) was downloaded from the University of California at Irvine (UCI) machine learning repository and was used in this investigation. Academics who employ machine learning techniques for breast cancer classification commonly use this dataset, enabling us to compare our system's performance with that of other studies on this topic.

We have used SVMs with feature selection for predicting breast cancer. It has shown promising results in improving the accuracy of breast cancer diagnosis. By selecting the most significant features,

the classification process is optimized, which leads to improved performance.

The rest of the paper is organized as follows. Section 2 explore literature survey on breast cancer classification problem, Section 3 describe methodology in details with used dataset, Section 4 highlight the results and analysis. At the end added conclusion and future works.

II. LITERATURE SURVEY

Breast cancer is a major health concern for women worldwide. According to the World Health Organization, breast cancer is the most common cancer among women globally, both in developed and developing countries. Early detection is critical in improving breast cancer outcomes and reducing mortality rates.

Advancements in technology have significantly improved early detection methods. Mammography, ultrasound, magnetic resonance imaging (MRI), and other imaging techniques are used to detect breast cancer at its early stages, when it is most treatable. Additionally, genetic testing and biomarker analysis can help identify women at high risk of developing breast cancer, enabling them to take preventive measures or undergo enhanced screening.

Burke et al. [4] evaluated the 5-year prediction accuracy of several statistical models against the predictive accuracy of artificial neural networks in a study (ANNs). The study found that ANNs outperformed all other statistical models in terms of predictive accuracy. The authors suggest that this is because ANNs are able to capture complex relationships between input variables in a way that traditional statistical models cannot. However, they also note that ANNs are more computationally intensive and require more data than traditional statistical models.

Pathological TNM staging model [5] used principal component analysis, classification and regression trees, and logistic regression were among the statistical methods for breast cancer prediction. This is a widely used system for staging cancers based on the size and extent of the primary tumor (T), involvement of nearby lymph nodes (N), and the presence of metastasis (M). This staging model helps in predicting the prognosis of the disease and determining appropriate treatment options. They used Principal Component Analysis (PCA), Classification and Regression Trees (CART) and Logistic Regression methods for predicting cancer.

Pendharker et al. [6] conducted a study to investigate trends in breast cancer using data mining techniques. The authors used a variety of methods, including association rule mining, clustering, and decision trees, to analyze a dataset of breast cancer patients.

Abbass et al. [7] used an evolutionary artificial neural network (EANN) to diagnose breast cancer. An EANN is a type of machine learning algorithm that combines artificial neural networks (ANNs) with evolutionary algorithms (EAs) to optimize the architecture and parameters of the ANN. In this study, the EANN was used to analyze a dataset of breast cancer patients and identify patterns in the data that could be used to accurately diagnose breast cancer. The results of the study showed that the EANN was able to achieve high levels of accuracy in diagnosing breast cancer, demonstrating the potential of this approach for medical diagnosis and other applications.

Abu-Hanna and Keizer et al. [8] developed a hybrid model for predicting mortality in critically ill patients. The model combines logistic regression and classification trees to achieve better accuracy in predicting patient outcomes.

Delen et al. [9] proposed a prediction model using a huge dataset, two popular data mining algorithms

(artificial neural networks and decision trees) were combined with a widely used statistical method (logistic regression) for breast cancer survivability.

Polat et al. [10] proposed a machine learning-based mechanism to predict breast cancer using the least square support vector machine (LS-SVM) algorithm. The LS-SVM algorithm is a variant of the support vector machine (SVM) algorithm that aims to minimize the mean squared error instead of maximizing the margin between classes. The proposed mechanism involves preprocessing the input data, feature selection, and classification using LS-SVM.

Akay et al. [11] proposed a mechanism to diagnose breast cancer using support vector machines (SVM) with feature selection. SVM is a machine learning algorithm that is commonly used for classification tasks. In this study, SVM was used to classify breast cancer patients as either malignant or benign based on the features extracted from mammogram images. Feature selection is an important step in machine learning that involves selecting the most relevant features from the available set of features. In the context of breast cancer diagnosis, this involves selecting the features that are most indicative of malignancy or benignity.

III. METHODOLOGY

Overview of Support Vector machine(SVM) method

SVM is a widely used supervised learning algorithm in machine learning, and it is commonly used for classification problems. The basic idea of SVM is to find a hyperplane in a high-dimensional space that separates the classes in the data.

SVM works by creating a decision boundary that maximizes the margin between the two classes. The margin is the distance between the decision boundary and the closest data points from each class. The hyperplane that maximizes the margin is chosen as the decision boundary, as it is considered to be the most robust and accurate.

SVM can be used for both linearly separable and non-linearly separable data by using different types of kernels. A kernel is a function that takes in two inputs and outputs a measure of similarity. By using a kernel, SVM can transform the original feature space into a higher-dimensional space, where the data may be linearly separable.

Overall, SVM is a powerful and versatile machine learning algorithm that can be used for a wide range of classification problems.

We consider a set of n pairs of training data given below:

$$\{(x_i, y_i)\}_{i=1}^{n}, (x_i, y_i) \in \mathbb{R}^p \times \{-1, 1\}$$

Our goal is to learn a linear function to obtain a classification rule of the data: class(x) = sign(f(x) = wx + β).

When the data points are linearly separable (the two classes can be separated by a hyperplane), we aim at finding the optimal separating hyperplane that maximizes the margin M between the two classes [12]. Consequently, we consider the following maximization problem:

$$\max_{w \in \mathbb{R}^p, \beta \in \mathbb{R}} M \text{ subject to } \forall i,\ y_i(x_i^T w + \beta) \geq M$$

Using the scalability of the solutions, we set ||w|| 2 = 1/M we obtain the equivalent problem:

$$\min_{w \in \mathbb{R}^p, \beta \in \mathbb{R}} \frac{1}{2} \| w \|_2^2 \text{ subject to } \forall i,\ y_i(x_i^T w + \beta) \geq 1$$

When the data is not linearly separable, we still want to maximize M by allowing some points to be in the wrong size of the margin. Hence, we define the support vector machine primal problem:

$$\min_{w\in\mathbb{R}^p,\beta\in\mathbb{R}} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\max\left(0, 1-y_i(x_i^T w+\beta)\right)$$

C is a penalization parameter which controls the trade-off between the classification error and the norm of the estimator. We also consider the dual of the SVM problem:

$$\max_{\alpha\in\mathbb{R}^n} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{1\le i,j\le n}\alpha_i\alpha_j y_i y_j x_i^T x_j$$
$$\text{subject to } \begin{cases} \forall i,\ 0\le\alpha_i\le C \\ \sum_{i=1}^{n}\alpha_i = 0 \end{cases}$$

The representer theorem guarantees that the solution can be written as a linear combination of the training data:

$$w_* = \sum_{i=1}^{n}\bar{\alpha}_i y_i x_i$$

We call support vectors the vectors such that $\alpha_i \neq 0$.

Utilization of Feature Selection

Feature selection is an important technique used in machine learning to select the most relevant features or attributes from a dataset. Here are some of the benefits of feature selection:

1. Improves Predictive Accuracy: By selecting only the most relevant features, the model becomes more focused and can better identify patterns in the data, leading to better predictive accuracy.

2. Enhances Model Comprehensibility: Feature selection can help to simplify the model by removing irrelevant features, making it easier to understand and interpret.

3. Increases Learning Efficiency: When a model is trained on a dataset with many irrelevant or redundant features, it can take longer to train and require more computational resources. Feature selection can help to reduce the number of features, making the model more efficient to train.

4. Reduces Model Complexity: By selecting only the most important features, the resulting model can be simpler and more compact, making it easier to deploy and maintain.

IV. Experimental setup

Used dataset

The WBCD (Wisconsin Breast Cancer Diagnostic) dataset [2] is a well-known dataset for binary classification issues in machine learning. It uses features taken from digital pictures of breast mass fine needle aspirate (FNA) samples to predict whether the mass is benign or cancerous\Malignant. The dataset is available on the UCI Machine Learning Repository, and it contains 569 samples, where 212 are malignant and 357 are benign. There are 30 features that describe each sample, including mean, standard error, and worst values for various characteristics of the cell nuclei in the image. Attributes area as follows: ID number, Diagnosis (M = malignant, B = benign), Ten real-valued features are computed for each cell nucleus: radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension. We have used 70% for training and 30% testing of the data examples.

Data Processing

In our experiment consider nine attributes. Firstly, normalize the data using standard normal distribution.

	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
0	0.197905	-0.702212	-0.741774	-0.639366	-0.555608	-0.698853	-0.181827	-0.612927	-0.348400	0
1	0.197905	0.277252	0.262783	0.758032	1.695166	1.772867	-0.181827	-0.285105	-0.348400	0
2	-0.511643	-0.702212	-0.741774	-0.639366	-0.555608	-0.424217	-0.181827	-0.612927	-0.348400	0
3	0.552679	1.583204	1.602192	-0.639366	-0.105454	0.125054	-0.181827	1.354008	-0.348400	0
4	-0.156869	-0.702212	-0.741774	0.059333	-0.555608	-0.698853	-0.181827	-0.612927	-0.348400	0
5	1.262227	2.236180	2.271896	1.806080	1.695166	1.772867	2.269251	1.354008	-0.348400	1
6	-1.221191	-0.702212	-0.741774	-0.639366	-0.555608	1.772867	-0.181827	-0.612927	-0.348400	0
7	-0.866417	-0.702212	-0.406921	-0.639366	-0.555608	-0.698853	-0.181827	-0.612927	-0.348400	0
8	-0.866417	-0.702212	-0.741774	-0.639366	-0.555608	-0.698853	-0.998853	-0.612927	1.961862	0
9	-0.156869	-0.375724	-0.741774	-0.639366	-0.555608	-0.698853	-0.590340	-0.612927	-0.348400	0

Figure 1: Example of Processed data

Hyperparameter tuning

Hyperparameter tuning can improve model performance, but it is important to ensure that the hyperparameters are tuned on a validation set and not the test set, to avoid overfitting and obtaining an overly optimistic estimate of the model's performance. In SVMs, the choice of kernel and its parameters can have a significant impact on the accuracy of the classification. The kernel function maps the input data to a higher-dimensional space where it becomes easier to separate the data into different classes. The choice of kernel depends on the type of data and the problem being addressed.

Common kernel functions include linear, polynomial, radial basis function (RBF), and sigmoid. The RBF kernel is commonly used because of its ability to model complex, nonlinear relationships between the input variables. However, the choice of kernel function and its parameters should be based on the specific problem at hand and the characteristics of the data.

The selection of kernel parameters, such as the regularization parameter and the kernel width for RBF kernels, also plays an important role in the accuracy of SVM classification. The regularization parameter controls the tradeoff between achieving a low training error and a low testing error, while the kernel width determines the shape and flexibility of the decision boundary.

Therefore, it is important to carefully select the appropriate kernel and its parameters to achieve the best possible classification performance. This selection process often involves experimenting with different kernel functions and parameter values and evaluating their performance on a validation set or through cross-validation techniques.

The parameters that should be optimized for the RBF kernel are the penalty parameter C and the kernel function parameter γ: {'C': 1, 'gamma': 0.01}.

V. RESULT AND ANALYSIS

We used the WBCD[2] to test the efficacy of our method. The F-score is used to quantify the importance of each feature, and grid search is used to optimize the SVM parameters. Grid search involves exhaustively trying different combinations of hyperparameters and selecting the one that gives the best performance on a validation set. A confusion matrix[13] contains information regarding a classification system's actual and expected classifications. **Table 1** Shows the confusion matrix for binary class classifier. **Table 2** Shows the shows the accuracy of our experiment with different training and testing splits. Training-Testing split of 75%-25% achieves the highest F-Measure Score 98.25%.

Table 1 : Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	True positive (TP)	False negative (FN)
Actual Negative	False positive (FP)	True negative (TN)

Table 2 : Comparison of accuracy

Cases	**Training-Testing Splits**	**Obtained F-score (%)**
1	70%-30%	95.61
2	85%-15%	97.09
3	90%-10%	97.10
4	80%-20%	97.56
5	75%-25%	98.25

The values from the previously constructed list of test accuracies are reshaped into a 2-dimension numpy array and plotted using a heat map to indicate the influence of pairs of hyperparameter (C and gamma) values on test accuracies. That are shown using different Heat Map graph. Heat Map of Test Accuracies are shown in Figures [2-6]. Heat Map for Train-Test(70%-30%) split shown in Figure 2 and achived accuracy is 95.61%. Heat Map for Train-Test(85%-15%) split shown in Figure 3 and achived accuracy is 97.09%. Heat Map for Train-Test(90%-10%) split shown in Figure 4 and achived accuracy is 97.10%. Heat Map for Train-Test(80%-20%) split shown in Figure 5 and achived accuracy is 97.56%. Heat Map for Train-Test(75%-25%) split shown in Figure 6 and achieved highest accuracy is 98.25.

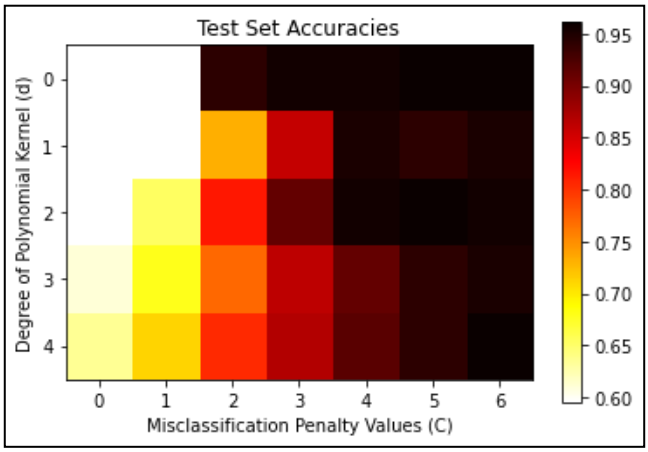

Figure 2 : Heatmap for train-test (70-30) % split cases

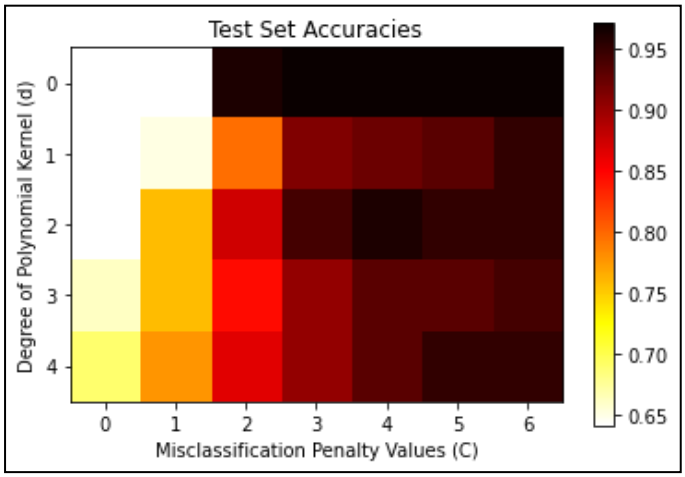

Figure 3 : Heatmap for train-test (85-15) % split cases

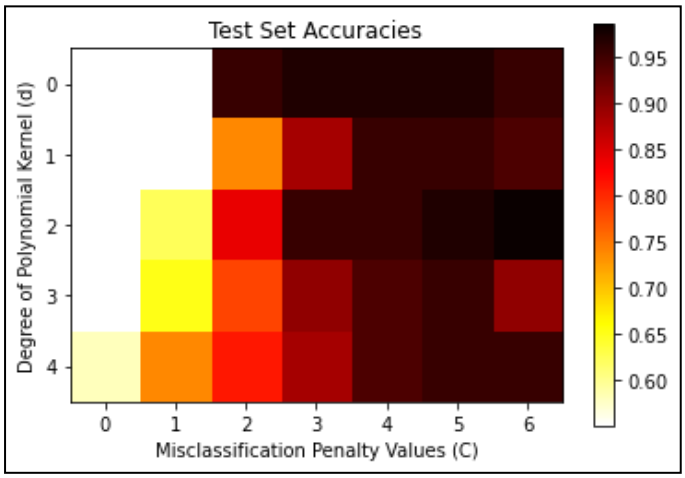

Figure 4 : Heatmap for train-test (90-10) % split cases

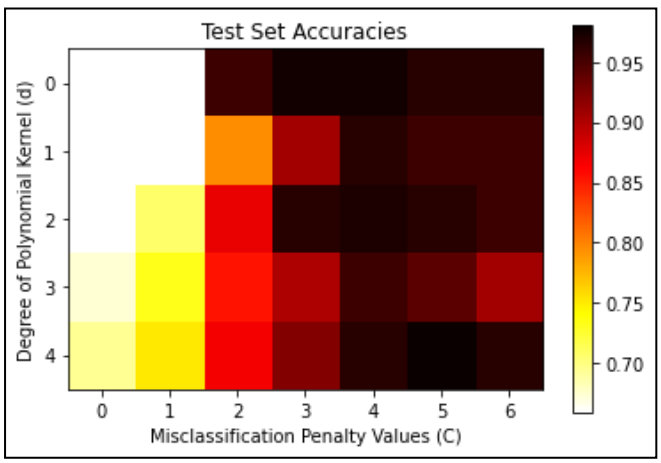

Figure 5 : Heatmap for train-test (80-20) % split cases

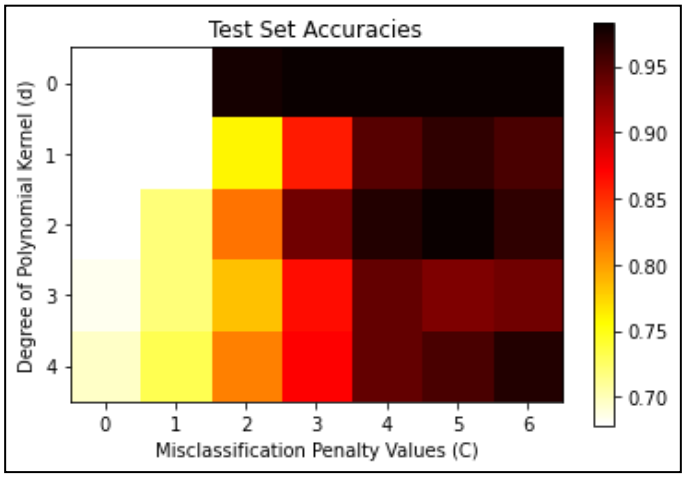

Figure 6 : Heatmap for train-test (75-25) % split cases

VI. CONCLUSION

The application of SVM combined with feature selection on the WBCD dataset for breast cancer diagnosis has been shown to produce high classification accuracies. The experiments conducted on different portions of the dataset show the robustness of the proposed method. The high classification accuracies suggest that the selected subset of features used by the SVM model is highly relevant for breast cancer diagnosis.

Overall, the application of SVM combined with feature selection on the WBCD dataset for breast cancer diagnosis is a promising approach. The high classification accuracies obtained suggest that the proposed method can be useful in the diagnosis of breast cancer, which can lead to earlier detection and improved patient outcomes.

Further, we want to incorporate another method to improve the performance in all cases. In future want to explore the dataset details to enhance the model performance.

VII. REFERENCES

[1] West, D., Mangiameli, P., Rampal, R., & West, V. (2005). Ensemble strategies for a medical diagnosis decision support system: A breast cancer diagnosis application. European Journal of Operational Research(162), 532–551.I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[2] ftp://ftp.ics.uci.edu/pub/machine-learning-databases, last accessed August 2006.

[3] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

[4] Burke, H. B., Rosen, D. B., & Goodman, P. H. (1994). Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival. Advances in neural information processing systems, 7.

[5] Burke, H. B., Rosen, D. B., & Goodman, P. H. (1994). Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival. Advances in neural information processing systems, 7.

[6] Pendharkar, P. C., Rodger, J. A., Yaverbaum, G. J., Herman, N., & Benner, M. (1999). Association, statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Systems with Applications, 17(3), 223-232.

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

[8] Abu-Hanna, A., & de Keizer, N. (2003). Integrating classification trees with local logistic regression in intensive care prognosis. Artificial Intelligence in Medicine, 29(1-2), 5-23.

[9] Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a

comparison of three data mining methods. Artificial intelligence in medicine, 34(2), 113-127.

[10] Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. Digital signal processing, 17(4), 694-701.

[11] Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications, 36(2), 3240-3247.

[12] Dedieu, Antoine.(2016) "MIT 9.520/6.860 Project: Feature selection for SVM."

[13] Kohavi, R. (1998). Glossary of terms. Special issue on applications of machine learning and the knowledge discovery process, 30(271), 127-132.

Cite this Article

Tohida Rehman, "Breast Cancer Diagnosis using Support Vector Machines and Feature Selection", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 3 Issue 8, pp. 2111-2118, November-December 2017.

Journal URL : https://ijsrst.com/IJSRST52310231