Acta Scientific COMPUTER SCIENCES

Volume 3 Issue 6 June 2021

Research Article

Support Vector Machine Based Classification Model for Breast Cancer Diagnosis

Tsehay Admassu Assegie*

Department of Computer Science, Aksum University, Ethiopia

***Corresponding Author:** Tsehay Admassu Assegie, Department of Computer Science, Aksum University, Ethiopia.

Received: April 15, 2021

Published: May 06, 2021

Abstract

Getting insight and making data driven decisions with predictive model are of paramount importance in breast cancer diagnosis. The key idea of using data driven model is to automate breast cancer diagnosis by learning specific patterns from data. In recent technological advancement, we all require machines to tell us when a person will need for screening or further test on health condition. While human intelligence and expertise is expensive and rarely available specially in developing nations, without data driven models and intelligent systems we cannot solve the real-world problem of breast cancer diagnosis at huge scale with efficiency. One of the methods for achieving better efficiency in medical dataset classification is the application of preprocessing to the original dataset. In this study, we conducted an empirical investigation with extensive experimental test on supervised learning algorithm namely, support vector machine (SVM) on Wisconsin breast cancer dataset. Experimental result shows that, with preprocessing methods such as scaling and transformation and parameter tuning model performance significantly improves the efficiency of support vector machine. Overall, we have proposed the state of the-art machine learning model for automated breast cancer detection with predictive accuracy of 96.71%.

Keywords: Breast Cancer; Support Vector Machine; Breast Cancer Diagnosis; Model Optimization

Introduction

In recent years, breast cancer have become one of the major health problem in the world with the increased case of breast cancer and shortage of medical expertise in the field for earlier diagnosis specially, in developing nations such as Ethiopia [1-33]. Breast cancer affects roughly 1.7 million women in the world every year [2]. Moreover, the breast cancer is one of the most common type of cancer disease causing the highest number of deaths every year. Optimizing of the performance of predictive model on breast cancer detection for medical decision support system based on support vector machine is crucial in reducing the medical costs, errors during breast cancer diagnosis by various physicians with different experience and practice on detection of breast cancer.

In traditional healthcare systems, diagnosis of breast cancer depends on the oncologist's decision and knowledge for detecting the breast cancer as the most likely because based on the symptoms. Likewise, automated diagnosis models suing machine learning algorithm could be used to support the decision making process of the human expert or oncologist during breast cancer diagnosis. The use of data driven decision support system reduce errors, and decrease the variation in experience, replace human experts where there is lack of oncologist and ultimately improves breast cancer diagnosis result.

Discovering insight from large volume of breast cancer data through patter recognition and visualization of breast cancer data plays significant role in not only diagnosis but also driving the facts from the large volume of data such as identifying the risky factors causing breast cancer, the relationship among other disease [3].

The contributions of this work are the following

- A succinct review on recent articles on breast cancer classification have been presented.
- An optimized state of the-art support vector machine based model has been implemented for breast cancer diagnosis.
- We have analyzed the effect of preprocessing and feature selection on the performance of support vector machine.
- Implement sequential feature selection method and remove irrelevant features to optimize model performance.

Literature Review

Numerous research work have been conducted on breast cancer detection using machine-learning model [4-32]. The researchers have applied different machine learning algorithms such as support vector machine, Naïve Bayes algorithm for prediction of breast cancer. However, the researches focus on developing predictive model with machine learning algorithm and the effect of feature magnitude such as high nonlinear variation in input feature value and pre-processing as well as feature selections is largely neglected and rarely researched in model optimization for improving the predictive accuracy of machine learning model on breast cancer.

Researchers have conducted many experiments on different machine learning algorithms for breast cancer detection. For example in [4], researchers have applied random forest and Naïve Bayes algorithm for breast cancer detection. The researchers have compared the performance of the models and result appears to prove that the random forest model outperforms the Naïve Bayes model.

Support vector machine have also been applied for breast cancer diagnosis [5]. Experimental result shows that support vector machine is the most powerful machine-learning algorithm for breast cancer diagnosis. Support vector machine is the most widely used machine-learning model for breast cancer diagnosis [6]. Thus, based on preliminary literature review, we selected support vector machine for implementation of automated system for breast cancer diagnosis.

In [8], the researchers compared the performance of different machine learning models such as support vector machine, Naïve Bayes and random forest algorithm. The researchers employed Wisconsin's breast cancer dataset. The experimental result shows that support vector machine outperforms the random forest and Naïve Bayes model in terms of prediction accuracy. Thus, this research addresses the observed gap in machine learning model optimization for improving model performance for breast cancer diagnosis. We have implemented pre-processing with scaling input feature vector, feature selection for removing irrelevant features and reduced input feature set avoiding the model complexity and reduced computational time without compromising model performance.

Materials and Methods

We have employed Wisconsin breast cancer dataset collected by Dr. William H originally obtained from university of Wisconsin hospital. This study is conducted following the general approach for problem solving using machine learning. First, we have collected medical dataset training and testing support vector machine. We have employed, preprocessing approaches such as dimensionality reduction with principal component (PCA) to transform the original dataset. Lastly, we have tuned the model for optimization and then tested with tuned hyper-parameter to get optimal result on breast cancer diagnosis. To implement the proposed model and conduct the experiment we have employed support vector machine and python programing language and used the scientific learning kit under python.

Support vector machine

Support vector machine of SVM is the most widely employed and the most powerful classification model in medical diagnosis [13]. The support vector machine is defined in terms of hyperplane dividing data points in n-dimensional space. The hyperplane dividing data points is defined as follows:

$$y = \alpha_0 + \alpha_1 X1 + \alpha_2 X2 \ldots + \alpha_n Xn - - - - - \quad (1)$$

Where, α_1, α_2.... denotes hypothetical values and X1, X2... are data points in sample space of n-dimension.

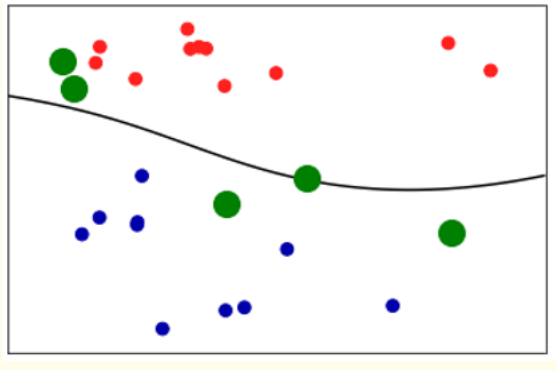

Figure 1: Maximum and minimum values of feature.

The decision boundary is shown in black line. The green dots indicate support vectors. To make prediction, the distance between support vectors and a data point is measured. A classification decision is then made based on the distance to support vectors that was learned during training. To measure the distance between data point and support vector, Gaussian kernel is used which is shown as follows:

$$krbf(x1, x2) = \exp(\gamma||x1 - x2||2) \text{ -----(1)}$$

Where x1 and x2 are data points, x1-x2 is the Euclidean distance between the data points. The Gaussian kennel is used to measure the distance between data point and support vector, and gamma controls the Gaussian kernel.

Dataset description

We have collected Wisconsin's breast cancer dataset originally provided by Wisconsin university, which consists of clinical measurement of breast cancer tumour. Each observation or tumour measurement is labeled as benign, non-cancerous, or malignant for cancerous tumour. The dataset consists of 569 data point or observations and 30 features characterize each observation or data point. The dataset consists of 212 malignant tumour and 357 benign or non-cancerous tumour.

The minimum and maximum values for each feature in breast cancer dataset is demonstrated in figure 2.

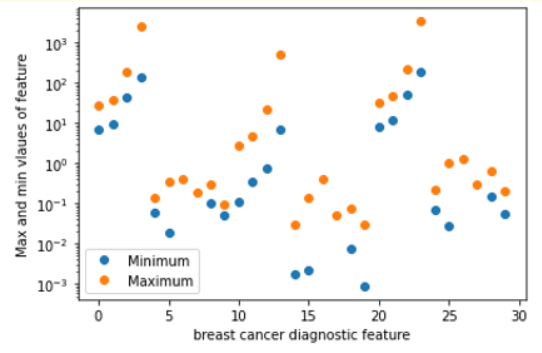

Figure 2: Maximum and minimum values of feature.

As shown in figure 2, the magnitude of features in the breast cancer dataset are different order of magnitude. The higher variation in magnitude between minimum and maximum values in the breast cancer dataset features shows a devastating on the performance of support vector machine. Thus, we employed min max scaler for preprocessing to solve the problem of the higher difference between the magnitudes of features in the breast cancer dataset. The min max scaler rescales the feature magnitude such that all features are between zero and one.

The min max scaler is pre-processing method that replaces every values of input feature in a column to a new value. The min max scaler is defined by using the following formula shown in equation (2).

$$X = \frac{(x - xmin)}{(xmax - xmin)} \text{ --(2)}$$

Where X is the new value, x is the original column value, xmin is the minimum value of the column and xmax is the maximum value of the column in the original dataset.

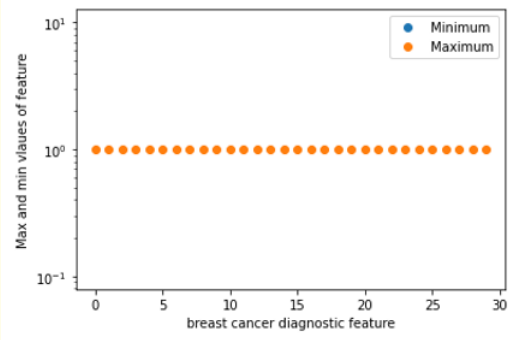

Figure 3: Feature data point after rescaling.

Parameter tuning

We have employed a grid search technique with cross validation to improve support vector machine model generalization with parameter tuning. The parameter tuning with grid search we implemented a simple for loop over the two parameters namely, the gamma value and C, training and evaluating model performance for each combination. We split the dataset into three folds for implementing grid search in order to avoid overfitting of the parameters and validation set. The three folds are explained as follows, one fold is named training set used for model fitting, the other fold is the validation set used for parameter tuning with grid search, the

third fold is the test set used for evaluation of the model trained on best feature set after parameter selection.

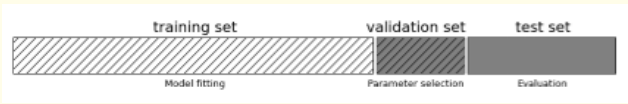

Figure a

We employed a grid search, simplest of the hyper-parameter optimization methods. In this method, we will specify the grid of values (of hyper-parameters) that we want to try out and optimize to get the best parameter combinations.

Then we will build models on each of those values (combination of multiple parameter values), using cross-validation of course, and report the best parameters' combination in the whole grid. The output will be the model using the best combination from the grid. Although it is quite simple, it suffers from one serious drawback that the user has to manually supply the actual parameters, which may or may not contain the most optimal parameters. In addition to the grid search, we have employed randomized search for parameter tuning and result comparison. Grid search is a very popular method to optimizing hyper-parameters in practice. It is due to its simplicity and the fact that it is embarrassingly parallelizable. This becomes important when the dataset we are dealing with is of a large size. However, grid search suffers from some major shortcomings, the most important one being the limitation of manually specifying the grid. This brings a human element into a process that could benefit from a purely automatic mechanism. On the other hand, randomized parameter search is a modification to the traditional grid search. Randomized search takes input for grid elements as in normal grid search but it can also take distributions as input. For example, consider the parameter gamma whose values we supplied explicitly in the last section instead we can supply a distribution from which to sample gamma. The efficacy of randomized parameter search is based on the proven (empirically and mathematically) result that the hyper-parameter optimization functions normally have low dimensionality and the effect of certain parameters are more than parameters. We control the number of times we want to do the random parameter sampling by specifying the number of iterations we want to run (n_iter). Normally a higher number of iterations mean better parameter search but does not find the better parameter setting as compared to the grid search. However, grid search takes higher computation time compared to randomized search.

Feature selection

We have employed sequential feature selection algorithm for selecting the optimal feature subset that could produces highest possible accuracy for proposed support vector machine.

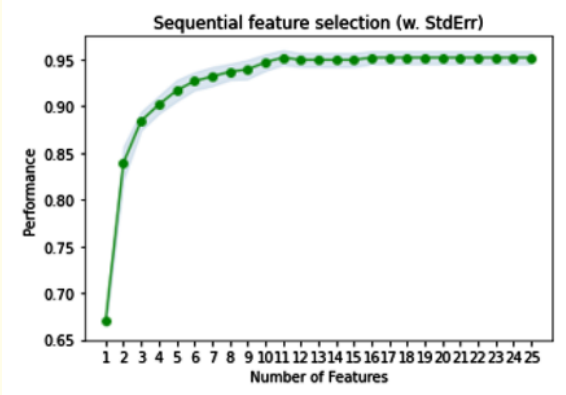

Figure 4: Number of features and cross validation accuracy.

Best combination features (Highest Accuracy: 0.955): (0, 1, 2, 3, 7, 8, 9, 10, 20, 21, 22, 24, 26, 27, 28). The highest classification accuracy is achieved when 29 features are used for training.

Results and Discussion

The performance of support vector machine is evaluated with predictive accuracy as performance measure. We have tested the model performance on unscaled data and then transformed the data with min max scaler. Result appears to prove that the support vector machine is highly sensitive to the magnitude of features. Performance improves with scaled data compared with a scaled data points.

The confusion matrix shown in figure 7 demonstrates the correct and incorrect prediction on breast cancer test set consisting of 171 observations of which 107 malignant and 64 benign observation. The predictive model correctly identified 163 observations and incorrectly predicted 8 0bservations. Moreover, the prediction accuracy on TN or benign class is better as compared to the malignant class. The predictive accuracy is determined by dividing the

number of observations in the test set to the number of correct prediction made by the model. Thus, predictive accuracy for the model is obtained by dividing 163 to 171, which is equal to 95.32%.

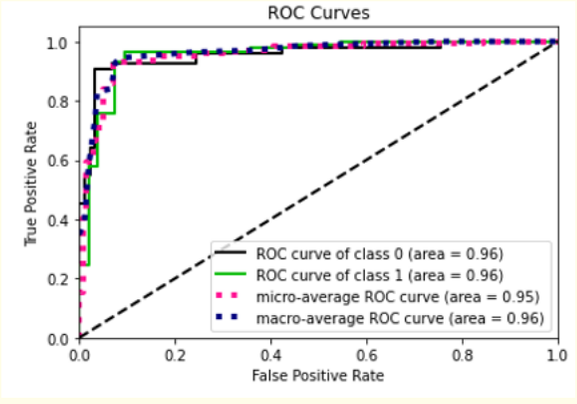

Figure 5: Receiver Operating characteristic curve.

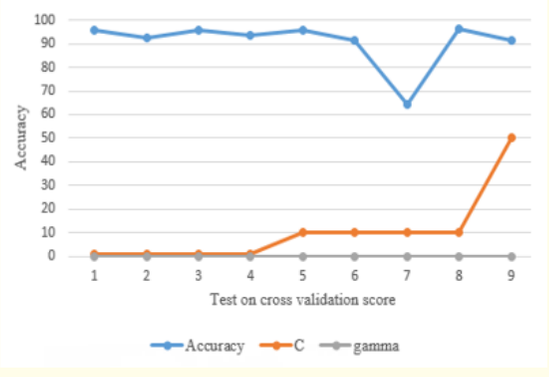

Figure 6: The effect of gamma and C on performance of SVM model.

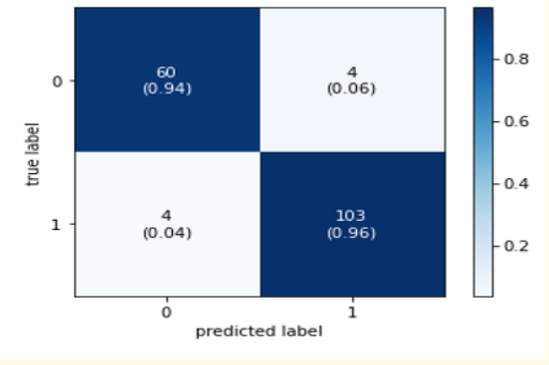

Figure 7: Confusion matrix for the SVM model.

Conclusion

In this research, we conducted an extensive experiment on support vector machine using the popular Wisconsin's breast cancer dataset. We have conducted experiment on support vector machine using different methods for optimization from pre-processing such as scaling to parameter tuning with grid searching, a heuristic search for selecting best parameter setting for support vector machine. Moreover, the performance of model with grid search and randomized search is compared. Experiment on grid and randomized search reveals that grid search have higher cross validation accuracy than randomized search. However, grid search requires higher computational time as compared to randomized search. Overall, this study have proposed the state of the-art machine learning model for breast cancer detection with predictive accuracy of 96.24% and Mathews correlation coefficient 0.90. Thus, the predicted outcome has higher correlation to the actual or real observation in the breast cancer dataset.

Conflict of Interest

The author does not have conflicts of interest.

Bibliography

1. MarcAubreville Christof., *et al*. "A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research". *Scientific Data* 7 (2020).

2. "Optimizing the Performance of Breast Cancer Classification by Employing the Same Domain Transfer Learning from Hybrid Deep Convolutional Neural Network Model". *Electronic* (2020).

3. Anita Paneri and Mayank Patel. "An Improved Model for Breast Cancer Classification Using SVM with Grid Search Method". *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* 8.8 (2019).

4. Ze-Wei Guo., *et al*. "Breast cancer detection via wavelet energy and support vector machine". Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication, Nanjing, China, August 27-31 (2018).

5. Anusha Bharat., *et al*. "Using Machine Learning algorithms for breast cancer risk prediction and diagnosis". IEEE Third International Conference on Circuits, Control. Communication and Computing (2018).

6. Sivapriya J., *et al*. "Breast Cancer Prediction using Machine Learning". *International Journal of Recent Technology and Engineering (IJRTE)* (2019).

7. Nasser H Sweilam., *et al*. "Support vector machine for diagnosis cancer disease: A comparative study". *Egyptian Informatics Journal* 11.2 (2020): 81-92.

8. Arpita Joshi and Dr Ashish Mehta. "Comparative Analysis of Various Machine Learning Techniques for Diagnosis of Breast Cancer". *International Journal on Emerging Technologies* (2017).

9. S Vasundhara., *et al*. "Machine Learning Approach for Breast Cancer Prediction". *International Journal of Recent Technology and Engineering (IJRTE)* 8.1 (2019).

10. Dana Bazazeh and Raed Shubair. "Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis". IEEE (2016).

11. Badal Soni., *et al*. "RFSVM: A Novel Classification Technique for Breast Cancer Diagnosis". *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* 8.12 (2019).

12. K Prasuna., *et al*. "Application of Machine Learning Techniques in Predicting Breast Cancer – A Survey". *International Journal of Innovative Technology and Exploring Engineering (IJITEE*) 8.8 (2019).

13. Ahmet Saygılı. "Classification and Diagnostic Prediction of Breast Cancers via Different Classifiers". *International Scientific and Vocational Journal* (2018).

14. Milon Islam., *et al*. "Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors". Humanitarian Technology Conference (R10-HTC) 21 - 23 (2017).

15. Madeeh Nayer Elgedaw. "Prediction of Breast Cancer using Random Forest, Support Vector Machines and Naïve Bayes". *International Journal of Engineering and Computer Science* 6.1 (2017).

16. S Murugan., *et al*. "Classification and Prediction of Breast Cancer using Linear Regression, Decision Tree and Random Forest". International Conference on Current Trends in Computer, Electrical, Electronics and Communication, IEEE (2017).

17. Bin Dai., *et al*. "Using Random Forest Algorithm for Breast Cancer Diagnosis". International Symposium on Computer, Consumer and Control, IEEE (2018).

18. Guillaume Lemaıtre., *et al*. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning". *Journal of Machine Learning Research* 1 (2017).

19. Shubham Sharma., *et al*. "Breast Cancer Detection Using Machine Learning Algorithms". IEEE (2018).

20. Assegie TA and Sushma S J. "A Support Vector Machine and Decision Tree Based Breast Cancer Prediction". *International Journal of Engineering and Advanced Technology (IJEAT)* 9.3 (2020).

21. Assegie TA. "An optimized K-Nearest Neighbor based breast cancer detection". *Journal of Robotics and Control (JRC)* 2.3 (2020).

22. Assegie TA and Nair PS. "The Performance Of Different Machine Learning Models On Diabetes Prediction". *International Journal of Scientific and Technology Research* 9.1 (2020).

23. Royida A., *et al*. "Efficient method for breast cancer classification based on ensemble hoffeding tree and naïve Bayes". *Indonesian Journal of Electrical Engineering and Computer Science* 18.2 (2020).

24. Assegie TA. "Support Vector Machine And K-Nearest Neighbor Based Liver Disease Classification Model". *Indonesian Journal of Electronics, Electro Medical, and Medical Informatics (IJEEEMI)* 3.1 (2021).

25. Assegie TA., *et al*. "Breast cancer prediction model with decision tree and adaptive boosting". *IAES International Journal of Artificial Intelligence (IJ-AI)* 10.1 (2021).

26. Omar Graja., *et al*. "Breast Cancer Diagnosis using Quality Control Charts and Logistic Regression". IEEE (2018).

27. Tina Elizabeth Mathew. "A Logistic Regression with Recursive Feature Elimination Model for Breast Cancer Diagnosis". *International Journal on Emerging Technologies* (2019).

28. Ebru Aydındag Bayrak., *et al*. "Comparison of Machine Learning Methods for Breast Cancer Diagnosis". IEEE (2019).

29. H Yusuf., *et al*. "Breast cancer analysis using logistic regression". *IJRRAS* 10.1 (2012).

30. Chelvian Aroef., *et al*. "Comparing random forest and support vector machines for breast cancer classification". *TELKOMNIKA Telecommunication, Computing, Electronics and Control* 18.2 (2020).

31. Moh'd Rasoul Al-hadidi., *et al*. "Breast Cancer Detection using K-nearest Neighbor Machine Learning Algorithm". International Conference on Developments in E-Systems Engineering, IEEE (2016).

32. Amin Ul Haq., *et al*. "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms". *Hindawi Mobile Information Systems* (2018).

33. Isra'a Ahmed Zriqat., *et al*. "A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods". *International Journal of Computer Science and Information Security (IJCSIS)* 14.12 (2016).

Volume 3 Issue 6 June 2021