An automated performance analysis using predictive analysis for the diagnosis of breast tumors

[1]T RADHIKA, [2]Y SANDHYA, [3]RAMYA SAKILAM, [4]GONE SHARNILA
[1,2,3]Assistant Professor, [4]Student, Dept. of Computer Science Engineering, Brilliant Institute of Engineering and Technology, Hyderabad, Telangana, India

ABSTRACT

Breast tumor is the most common non-skin malignancy in women and the second leading cause of female cancer mortality. Breast tumors and masses usually appear in the form of dense regions in mammograms. A typical benign mass has a round, smooth and well circumscribed boundary; on the other hand, a malignant tumor usually has a speculated, rough, and blurry boundary. Computer aided detection (CAD) systems in screening mammography serve as a second opinion for radiologists by identifying regions with high suspicion of malignancy. The ultimate goal of CAD is to indicate such locations with great accuracy and reliability. Thus far, most studies support the fact that CAD technology has a positive impact on early breast cancer detection. There is extensive literature on the development and evaluation of CAD systems in mammography. Most of the proposed system follows a hierarchical approach. Initially the CAD system prescreens a mammogram to detect suspicious regions in the breast parenchyma that serve as candidate locations for further analysis. The algorithm uses several features which are extracted in earlier work. SVM is a learning machine used as a tool for data classification, function approximation, etc, due to its generalization ability and has found success in many applications. Feature of SVM is that it minimizes an upper bound of generalization error through maximizing the margin between separating hyperplane and dataset. SVM has an extra advantage of automatic model selection in the sense that both the optimal number and locations of the basis functions are automatically obtained during training. The performance of SVM largely depends on the kernel.

Keywords: SVM (Support Vector Machine), Tumor, Benign, Malignant, Stratified K fold validation

I. INTROUCTION

The breast is made up of different tissue, ranging from very fatty tissue to very dense tissue. Within this tissue is a network of lobes. Each lobe is made up of tiny, tube-like structures called lobules that contain milk glands. Tiny ducts connect the glands, lobules, and lobes, carrying milk from the lobes to the nipple. The nipple is located in the middle of the areola, which is the darker area that surrounds the nipple. Blood and lymph vessels also run throughout the breast. Blood nourishes the cells. The lymph system drains bodily waste products. The lymph vessels connect to lymph nodes, the small, bean-shaped organs that help fight infection. Groups of lymph nodes are located in different areas throughout the body, such as in the neck, groin, and abdomen. Regional lymph nodes of the breast are those nearby the breast, such as the lymph nodes under the arm.Cancer begins when healthy cells in the breast change and grow out of control, forming a mass or sheet of cells called a tumor. A tumor can be cancerous or benign. A cancerous tumor is malignant, meaning it can grow and spread to other parts of the body. A benign tumor means the tumor can grow but will not spread.

Breast cancer spreads when the cancer grows into adjacent organs or other parts of the body or when breast cancer cells move to other parts of the body through the blood vessels and/or lymph vessels. This is called a metastasis.This guide covers both non-invasive (stage 0) as well as early-stage and locally advanced invasive breast cancer, which includes stages I, II, and III. The stage of breast cancer describes how much the cancer has grown, and if or where it has spread.

Although breast cancer most commonly spreads to nearby lymph nodes, it can also spread further through the body to areas such as the bones, lungs, liver, and brain. This is called metastatic or stage IV breast cancer and is the most advanced type of breast cancer. However, the involvement of lymph nodes alone is generally not stage IV breast cancer.

2. LITERATURE SURVEY

The Deep learning methods for detecting breast cancer include neural networks (Yanet al., 2019; Ting, Tan, & Sim, 2019) for image classification of affected breast tissue. Researchers in this field use images of cancerous breast tissue to build classification algorithms that can predict the stage of development of breast cancer. Neural networks can be combined with feature selection like the ridge and linear discriminant analysis to perform image classification (Toğaçar, Ergen, & Cömert, 2020). These algorithms have wide application in medical screening as they can classify a patient as healthy or sick and determine the type of breast cancer without requiring prior knowledge about the lack or presence of breast cancer (Ting, Tan, & Sim, 2019). These algorithms also achieve high accuracy when predicting the type of cancer (Ting, Tan, & Sim, 2019; Toğaçar, Ergen, & Cömert, 2020) and they can be used as an additional diagnostic method along with medical tests. Machine learning techniques are used to extract text information about typical symptoms of malignant breast cancer by checking medical records of patients. For Instance, Forsyth et al. (2018) built a machine learning algorithm to find the most common symptoms of breast cancer that patients reported. Their algorithm checked symptoms. Zhang et al. (2019) also studied characteristics of breast cancer to predict its occurrence. They achieved a f1-score of 93.53% using neural networks. Such research extends medical knowledge about causes and symptoms of breast cancer dataset and increases chances of correct and timely diagnosis. Along with neural networks, classification methods are another popular machine learning techniques to explore breast cancer. Unlike previously mentioned work, classification methods use quantitative data to predict the type of breast cancer. The most common dataset used is the Wisconsin breast cancer dataset. It contains quantitative data about breast cancer physical characteristics, while the target variable is a categorical variable corresponding to benign or malignant cancer. The classification task is to predict the malignant cases. Many authors have examined this dataset using various approaches. For instance, Liu et al. (2019) devised a novel approach for feature selection on the dataset called IG AGAW-SSVM to improve prediction accuracy. Asri et al. (2016) applied support vector machines, k-nearest neighbors, naive Bayes and decision trees. He achieved an accuracy of 97.13% in predicting the malignant cases.

The purpose of classification methods is to help diagnose breast cancer based on its physical characteristics. The decision tree classifier (Quinlan, 1996) achieved 94.7% accuracy on the Wisconsin breast cancer dataset, while ensemble learning algorithms (Bashir, Qamar, & Khan, 2015) achieved 97.4% accuracy. Although not widely used, a nero-rule based approach (Setiono, 2000) managed to achieve 98.2% accuracy on the dataset. Support Vector Machines (SVMs) is the most commonly used algorithm to achieve high accuracy when predicting malignant breast cancer (Liu et al., 2019; Wanget al., 2018). Previous research (Chaurasia and Pal, 2017) has shown that the SVMs with the RBF kernel is the most suitable algorithm to detect malignant breast cancer. They (Chaurasia and Pal, 2017) achieved accuracy of 96.8%. Other papers also showed that the SVMs may be the most appropriate classification method for detecting malignant cancer (Maldonado et al., 2014). In a previous research we (Vrigazova & Ivanov, 2019) presented a modification of the SVMs called the SVM and achieved accuracy of 97.5%. In this paper, we tune the SVM (Vrigazova & Ivanov, 2019) to increase prediction accuracy on the Wisconsin breast cancer dataset. We show that our version outperforms some current research (Maldonado et al., 2014) and (Khairul Nahar et al., 2019). We also show that our algorithm significantly boosts the results achieved on other breast cancer datasets compared to the classical SVMs. With this discovery, we propose an optimized version of the Support vector machines that can be applied on various breast cancer datasets to detect malignant cancer.

3. PROPOSED WORK

Breast Tumor is very common these days and the early detection results in decrease of mortality. Automatic detection of Breast Tumors is essential in medical industries. Here we propose a computer aided detection model which results in the highest accuracy of 97% to 98% under different validations.
The system uses an SVM classifier for classification of tumors on the breast. The model is a predictive model which uses a Support Vector Machine. The Support Vector Machine uses support Vector or Hyper Planes for classification. Since here we have only two classes so we used the polynomial kernel, linear Kernel and Radial Basis Kernel. We use linear if the data is linearly separable if the data is not linearly separable we use polynomials with degree 3 which points the data in higher dimensions , we can also use Radial Basis which makes the infinitely higher dimension. The SVM increases the dimension.

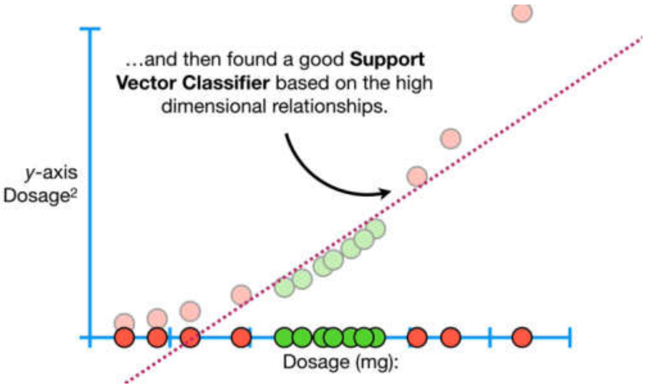


Fig 1: SVM Lower dimension to Higher dimension

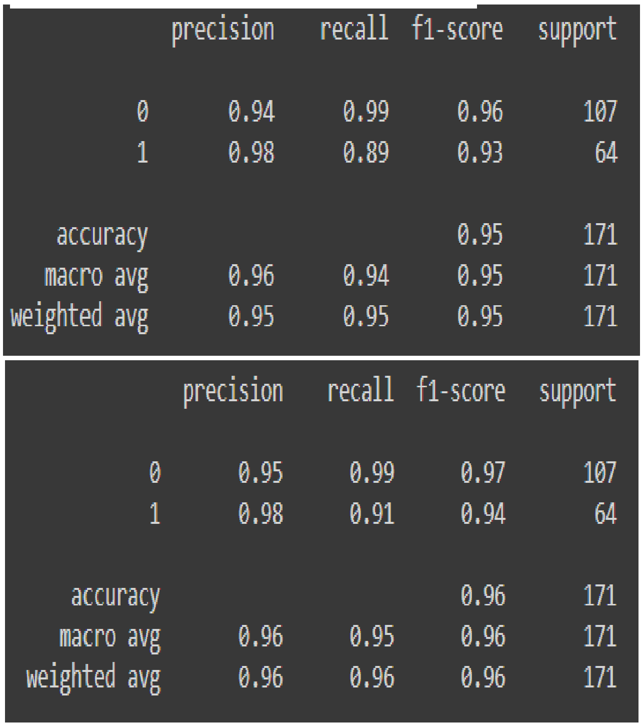
We can easily detect outliers using SVM classification. The SVM used to build a predictive model. It has an elegant way of transforming non linear data so that one can use a linear algorithm to fit a linear model to the data . Kernelized support vector machines are powerful models and perform well on a variety of datasets. SVMs allow for complex decision boundaries, even if data have only few features. They work well in low dimensional and high dimensional data with only a few features, but don't scale very well with the number of samples. Running an SVM on data with up to 10,000 samples might work well, but working with datasets of size 100,000 or more can become challenging in terms of runtime and memory. SVM requires careful preprocessing and tuning parameters. This is why, these days, most people instead use tree-based models such as random forests or gradient boosting (which require low or no preprocessing) in many applications.

SVM models are hard to inspect. It was difficult to understand why a particular prediction was made, and it might be tricky to explain the model to an expert. The most important parameters in kernels SVMs are regularization parameter C, the choice of kernel (linear, radial basis function or polynomial) and kernel specific parameters, gamma and C control the complexity of the model, with large values in either resulting in a more complex model. Therefore good settings for the two parameters are usually strongly correlated, and C and gamma should be adjusted together.
The data is splitted into training data and testing data where 70% of total data used for training and 30% used for testing. The data is splitted is stratified fashion so that every class label is in the same ratio. The classification is done with cross validation hence the training is done by taking all folds except one - referred to as hold out sample. The holdout sample is then thrown back with the rest of the other folds, and a different fold is pulled out as the new holdout sample. Training is repeated again with the remaining folds and we measure performance using holdout samples. This process is repeated until each fold has a chance to be a test or holdout sample. The

expected performance of the classifier, called cross validation error, is simply an average error rate computed on each holdout sample. Feature selection is an important part of the model-building process that you must always pay particular attention to. The system is very effective in tumor detection.

4. RESULTS AND DISCUSSIONS

In statistical modeling and machine learning, a commonly-reported performance measure of model accuracy for binary classification problems is Area Under the Curve (AUC). To understand what information the ROC curve conveys, consider the so-called confusion matrix that essentially is a two-dimensional table where the classifier model is on one axis (vertical), and ground truth is on the other (horizontal) axis, as shown below. Either of these axes can take two values (as depicted).



```
              precision    recall  f1-score   support

           0       0.94      0.99      0.96       107
           1       0.98      0.89      0.93        64

    accuracy                           0.95       171
   macro avg       0.96      0.94      0.95       171
weighted avg       0.95      0.95      0.95       171
```

```
              precision    recall  f1-score   support

           0       0.95      0.99      0.97       107
           1       0.98      0.91      0.94        64

    accuracy                           0.96       171
   macro avg       0.96      0.95      0.96       171
weighted avg       0.96      0.96      0.96       171
```

Fig 2: Scores for 3 fold and 5 fold validation

There are two possible predicted classes: "1" and "0". Malignant = 1 (indicates presence of cancer cells) and Benign = 0 (indicates absence). The classifier made a total of 174 predictions (i.e 174 patients were being tested for the presence of breast cancer). Out of those 174 cases, the classifier predicted "yes" 58 times, and "no" 113 times. In reality, 64 patients in the sample have the disease, and 107 patients do not. (TP+TN)/total = (57+106)/171 = 0.95, (FP+FN)/total = (1+7)/171 = 0.05 equivalent to 1 minus Accuracy also known as Error Rate TP/actual yes = 57/64 = 0.89 also known as "Sensitivity" or "Recall”. FP/actual no = 1/107 = 0.01 .TN/actual no = 106/107 = 0.99 equivalent to 1 minus False Positive Rate. TP/predicted yes = 57/58 = 0.98. actual yes/total = 64/171 = 0.34.

CONCLUSION

In this project we propose the SVM algorithm to increase prediction accuracy on breast cancer datasets like the Breast Tumor dataset (97.6%) . With this method we aim to propose SVM that can improve the quality of detecting breast tumors. The advantages of our proposed algorithms include improvement of the accuracy, reducing the error rate and producing a high enough AUC score. Our algorithm can separate the benign from malignant cells and classify them properly with high accuracy without overfitting. As we show, in some cases our algorithm can be faster than existing ones. Further research can be made into the precision, recall and f1-score measures to observe the model's performance in each class as the accuracy measure is a general measure. Each class performance can be affected differently depending on the resampling procedure. Our previous experiments show that if the ability of the model to correctly classify if the two binary outcomes had not improved, then the accuracy could not be improved. Therefore, we did not show the precision, recall and f1-score measures in our experiments. With the results from this research, we extend the practical applications of the bootstrap procedure for detecting breast cancer as the RBF-SVM provided the smallest error rate, high enough AUC score and could reduce execution time in some cases. With this, we extend existing academic literature and propose optimization of the SVMs for detecting malignant and benign breast cancer that can be applied on various datasets. Moreover, the benefits from our algorithm can be extended to other medical datasets to improve ability to detect other medical Conditions.

REFERENCES

[1] E.C.Fear, P.M.Meaney, and M.A.Stuchly,"Microwaves for breast cancer detection", IEEE potentials, vol.22, pp.12-18, February-March 2003.

[2] Homer MJ.Mammographic Interpretation: A practical Approach. McGraw hill, Boston, MA, second edition, 1997.

[3] American college of radiology, Reston VA, Illustrated Breast imaging Reporting and Data system (BI-RADSTM) , third edition, 1998.

[4] S.M.Astley,"Computer –based detection and prompting of mammographic abnormalities", Br.J.Radiol, vol.77, pp.S194- S200, 2004.

[5] L.J.W. Burhenne,"potential contribution of computer aided detection to the sensitivity of screening mammography", Radiology, vol.215, pp.554-562, 2000.

[6] T.W.Freer and M.J.Ulissey, "Screening mammography with computer aided detection: prospective study of 2860 patients in a community breast cancer", Radiology, vol.220, pp.781- 786, 2001.

[7] C. Cortes, V. N. Vapnik, "Support vector networks", Machine learning Boston, vol.3, Pg.273-297, September 1995.

[8] V. N. Vapnik, "An overview of statistical learning theory", IEEE Trans. Neural Networks New York, Vol. 10, pg. 998- 999, September 19999.

[9] O. Chapelle, V. N. Venice, Y. Bengio, "Model selection for small sample regression", Machine Learning Boston vol.48, pg. 9-23, July 2002.

[10] Y. Liu, Y. F. Zhung, "FS_SFS: A novel feature selection method for support vector machines", pattern recognition New York, vol.39, pg.1333-1345, December 2006.

[11] N. Acir, "A support vector machine classifier algorithm based on a perturbation method and its application to ECG beat recognition systems" Expert systems with application New York, vol.31, pg. 150-158 July 2006.

[12] Girosi f. Jones M. and Poqqio T., "Regularization theory and neural network architectures", Neural computation Cambridge, vol.7, pg.217-269, July 1995.

[13] Smola A. J., Scholkopf B., and Muller K. R., "The connection between regularization operators and support vector kernels", Neural Networks New York, vol.11, pg 637- 649, November 1998.

[14] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, Digital Image Processing Using MATLAB , prentice Hall, New Jersey, USA, 2004.

[15] C. Scott, "TEMPLAR software package, copyright 2001, Rice university", 2001

[16] V. N. Vapnik. The Nature of statistical learning theory Springer, 1995