INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 6.078
(Volume 6, Issue 5)
Available online at: https://www.ijariit.com

Prediction of a mammogram mass as benign or malignant

Sanjay Kumar A.
sanjaysan648@gmail.com
St. Joseph's Institute of Technology, Chennai, Tamil Nadu

Sarojini Sharon R. K.
sharon110699@gmail.com
St. Joseph's Institute of Technology, Chennai, Tamil Nadu

ABSTRACT

This paper is about comparing several different supervised machine learning techniques to the "mammographic masses" public data set from the UCI repository, and see which one yields the highest accuracy as measured with K-Fold cross validation (K=10 A lot of needless pain and surgery occurs from the subsequent false positives of mammogram performance. If we can create a better way of understanding them through supervised machine learning, it will make many lives easier .We have applied Decision tree, Random forest, KNN, Naive Bayes, SVM, Logistic Regression and also neural network using Keras. The data was cleaned as many rows were containing missing data, and as there were erroneous data identifiable as outliers as well. Some techniques such as SVM that we have employed require the input data to be normalized first. Many techniques also had "hyperparameters" that had to be tuned. So here we have approached a better way to make it even better by tuning its hyperparameters.

Keywords: *Mammographic Mass, Hyperparameter, Machine Learning, Supervised Machine Learning*

1. INTRODUCTION

Breast cancer is the leading cause of non-preventable cancer-associated death among women. The primary purpose of screening asymptomatic, healthy women for breast cancer is to diagnose it earlier and in so doing reduce the risk of or delay the onset of death from this disease. In addition, detecting cancers when the tumors are smaller benefits many women by permitting the therapeutic option of breast preservation. The importance of mammographic screening is gradually being recognized in the United States, and over the past several years, the number of women screened has increased. Mammography has clearly been shown to be quite sensitive in its ability to detect cancer, but frequently it cannot be used to accurately differentiate benign from malignant lesions.[1] Currently, open biopsy is the only accurate way to determine the benign or malignant basis of a mammographic finding, although fine-needle aspiration cytology and core needle biopsy are being explored as alternatives. Preoperative localization of clinically occult abnormalities detected by mammography, followed by excisional biopsy, represents the gold standard. Because of the lack of morphologic specificity, many biopsies done because of mammographic findings are being performed for what prove to be benign lesions. In the United States, the positive predictive value (PPV) of mammography (the number of cancers diagnosed per number of biopsies recommended) has ranged from approximately 15% to 30% [2-7]; in the European trials, it ranged from 30% [8] to 75% [9]. Some of the reasons for such discrepancies have been discussed [1 0]. Nevertheless, as a consequence of the numbers of biopsies that are now being done, and the anticipated increase in biopsies that are likely to accompany large scale screening, the PPV for mammography is under increasing scrutiny. In this paper we have addressed the issue of predicting the mammographic mass to be either benign or malign. The most effective tool available today for breast cancer screening is mammography. However, the poor positive predictive value of breast biopsy resulting from mammogram interpretation leads to unnecessary biopsies with favourable outcomes of around 70 per cent. In recent years many computer-aided diagnostic (CAD) systems have been introduced to minimise the large number of unnecessary breast biopsies. These programmes help doctors make their decision to perform a breast biopsy on a suspected lesion seen in a mammogram or even perform a short-term follow-up test. This data set can be used to assess the magnitude It contains a BI-RADS classification, the age of the patient and three BI-RADS attributes along with ground truth (the severity area) for 516 benign and 445 malignant masses that were detected on full field digital mammograms obtained at the University of Erlangen-Nuremberg Institute of Radiology between 2003 and 2006[11]. Each instance has an associated BI-RADS evaluation ranging from 1 (definitely benign) to 5 (highly evocative of malignancy) assigned by physicians during a double-check phase. If all cases with BI-RADS measurements are greater or equal to a given value (varying from 1 to 5), are malignant and the other cases can be measured as benign, sensitive to related specificities. These can mean how well a CAD system performs in comparison with radiologists.

In this paper, we have tuned the hyper parameter using the Bayesian hyper parameter model in grid search to increase the accuracy, which increases efficiency and produces better precision. Before we train the models we must set hyperparameters. Hyperparameters are very important in constructing robust and precise models. They help us find the balance between bias and

variance and thus avoid overfitting or underfitting of the model. We need to understand what they mean, and how they modify a model, to be able to alter the hyperparameters. Trying a bunch of hyperparameter values at random will be a boring and never-ending job. Finally the Bayesian hyper parameter model in grid search which enhances optimization and produces better accuracy.

2. RELATED WORKS

Mammography is the effective diagnostic technology to detect Breast cancer. Data mining algorithms may be used to enable doctors to conduct a breast biopsy on suspected lesions in their decisions. Sahar A. Mokhtar et al [12] in their research applied classification algorithms such as Decision Tree, Artificial Neural Network, and SVM on mammographic masses dataset to assess the magnitude of mammographic mass lesion. In that svm predicts breast cancer incidence with the lowest error rate and maximum accuracy of 81.25% . Pedro Ferreira et al [13] in their mammography findings predict that the classifier named SVM trained without mass density has an average performance of 83.8% while the classifier trained with the retrospectively assessed mass has an accuracy of 85.6%.

Also, our other literature survey includes the work of Renato Campanini et al [14] where their presented system achieves an accuracy of 84% in predicting the breast cancer using Wavelet transform with SVM classifier. Gustavo Carneiro et al [15] tested their model on two datasets namely: DDSM-BCRP and breast. The results with a true positive of about 0 96 ± 0 03 at 1 2 on those publicly available datasets. False positive per inbreast image and true positive 0 75 at 4 8 false positives for every single image on DDSM BCRP. Our study also includes the work of Arnau Oliver [16] where they proposed an approach to false positive mammographic mass identification using Local Binary Pattern and SVM for classifying the identified masses using a collection of 1792 dubious regions of interest extracted from the DDSM database. Ireaneus Anna Rejani et al [17] presented a system that detects tumors as suspicious regions and extract features of categorized tumour. They used a mini-MIAS dataset for their testing and achieved an accuracy of 88.75%. Vijayarajeswari et al [18] in their work proposed a system which effectively classifies the abnormal classes of mammogram. Their work states the strategies for classification and feature extraction using Hough transform. Dheeba et al [19] presented an approach using SVM for identifying microcalcification clusters in mammograms where Law's texture energy measures are taken from the image Region of interest in order to improve the classification rate. Dina A. Ragab et al [20] implemented an approach where the feature extraction is done using deep CNN called AlexNet which is fine-tuned to identify classes. The final fully connected layer is integrated to the SVM classifier to yield a better result of 87.2% accuracy.

3. SYSTEM ARCHITECTURE

The below architecture diagram depicts the work flow of our proposed system. Initially, in order to pre-process the breast cancer dataset, the raw data of breast mammogram is fed to the system where data filtering and data cleaning is done in order to avoid the inappropriate data values. Then the processed data need to be normalised. Hence, we normalise the data attributes using a pre-processing method called as Standard Scalar. After data normalisation, our mammographic masses dataset must be divided into two groups of testing and training. Accordingly, 75% of the data was assigned for training set and 25% of which was assigned for testing set. The testing data is subjected to several trained classifiers like Decision Tree, Random forest, KNN, Naive Bayes, SVM with kernels (poly, rbf, etc.), Logistic Regression and also neural network using Keras. All these approaches didn't produce higher accuracy. Hence, we incorporated hyper parameter tuning using bayesian classifier on classifiers such as Support Vector Machine (SVM). Now when the unclassified testing data is tested on the well trained SVM classifier. This produces more accuracy than the previous approaches. Based on the scores obtained, the test data will be classified as benign or malignant**.**

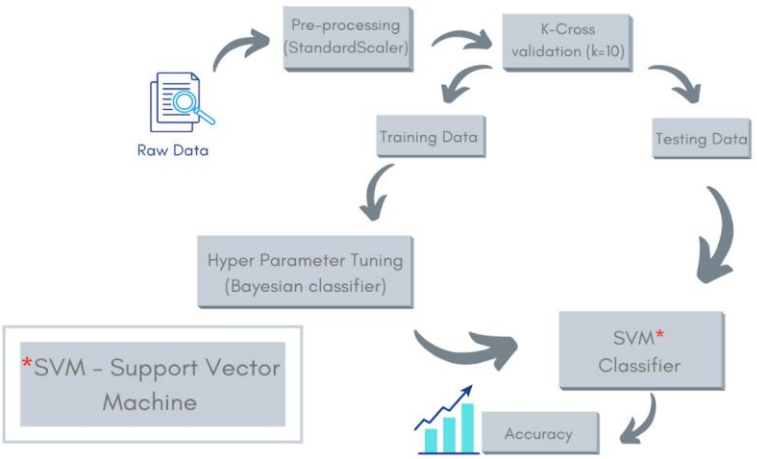

Fig. 1: System Architecture of our SVM model

4. IMPLEMENTATION

Support vector machine or SVM algorithm is based on the 'decision planes' concept, where hyperplanes are used to classify a set of objects in question. Many prefer the support vector machine, as it produces significant precision with less computational power. Support Vector Machine, abbreviated as SVM, can be used for regression tasks as well as classification The aim of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N is the number of features) that distinctly classifies the data points. There are many possible hyperplane that can be chosen to separate the two classes of data points. Our goal is to find a plane with the maximum margin, that is to say the maximum distance between the data points of both classes.

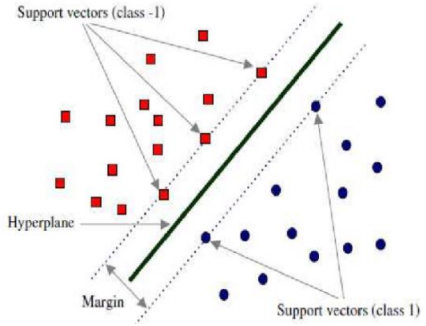

Fig. 2: Support Vector Machine Architecture

Step 1: Read the required data (BI-RADS assessment, Age, Shape, Margin, Density, Severity)
Step 2: Cleaning the data before applying SVM
Step 2.1: There were quite few missing values in the data set. Before we just drop every row that's missing data, we make sure we don't bias our data in doing so
Step 2.2: The missing data seems randomly distributed, so the rows with missing data were dropped.
Step 3: The Pandas dataframes were converted into numpy arrays that can be used by scikit_learn.
Step 4: Provide the data collected UCI repository was divided into as 75% training and 25% as testing.
Step 5: Provide the data from the above steps to train SVM with Bayesian hyperparameter tuning for high accuracy with RBF kernel, where the C and gamma values are pre-assumed values that will help predict the classes of the SVM model.
Step 6: Provide the test data and display the results.
Step 7: Compare the output from step 6 and 7 and show the observations.

5. EXPERIMENTAL RESULTS

The **mammographic masses** public dataset from the UCI repository [11] has been used for experimental analysis. So, after cleaning and normalizing the dataset the SVM algorithm with linear kernal is applied to the dataset where a accuracy of 79% was seen. In order to increase the accuracy we re-visisted the SVM kernsl by replacing it with different kernel functions such as RBF, sigmoid, poly to compare and see which kernel gave the highest possible accuracy. So as you can see in **Figure 3** the RBF kernel gave the highest accuracy of 80%.

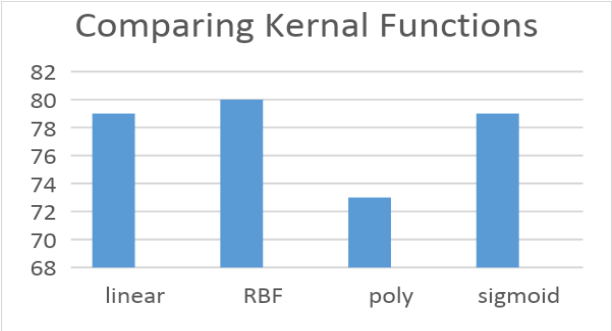

Fig. 3: Comparing kernel Functions Graph

Further we tried to tune our hyper parameter by using Bayesian hyper parameter in grid search. We achieved an accuracy rate of 87% after applying this hyper parameter as you can see in **Figure 4**.

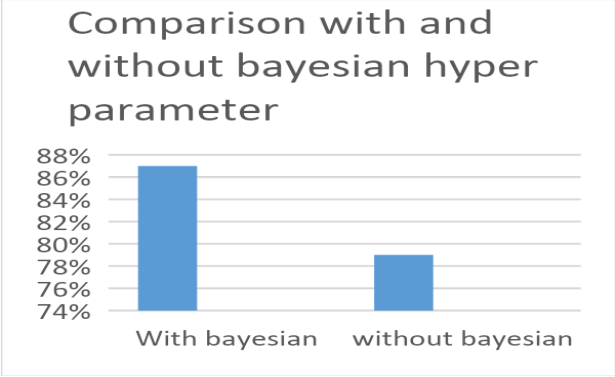

Fig. 4: Comparison graph with and without Bayesian hyper parameter

So we have improved our accuracy rate from 79% to 84% by employing the Bayesian hyper parameter model in our SVM algorithm. We can strongly assure that by this method any dataset that is employed will give higher accuracy as expected. Thus its evident that when SVM and Bayesian hyper parameter model are employed together there will be higher accuracy. Thus by our approach we have achieved a higher percentage of accuracy.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we tried to predict the tumours of breast and classify them using the mammographic mass values. Initially with the classic SVM algorithm we didn't get much accuracy then after comparing with different kernel functions we got a little higher efficiency Then our accuracy value only went higher after our hyper parameter was modified using the Bayesian hyper parameter in grid search model which would estimate the likelihood of the result and search through the grid in all possible ways to estimate the predicted result. In our future work we will try to improve our work further by improving the accuracy by further tuning or different topologies of the multi-level perceptron might make a difference. Also would employ the dataset to other supervised machine learning algorithm and even in deep learning to understand better and produce more efficiency.

7. REFERENCES

[1] Sickles WA. Mammographic features of 300 consecutive nonpalpable breast cancers. AiR 1986;146:661-663
[2] Meyer JE, Kopans DB, Stomper PC, Undfors KK. Occult breast abnermalities: percutaneous preoperative needle localization. Radiology 1984;150:335-337
[3] Yankaskas BC, Knelson MH, Abernathy ML, Cuthno SF, Clark AL. Needle localization biopsy of occult lesions of the breast experience in 199 cases. Invest Radiol 1988;23:729-733
[4] Marrujo G, Jolly PC, McClure HH. Nonpalpable breast cancer: needlelocalized biopsy for diagnosis and considerations for treatment. Am J Surg 1986;151 :599-602
[5] Poole Gv, Choplin RH, Sterchi JM, Leinbach LB, Myers AT. Occult lesions of the breast. Surg Gynecol Obstet 1986;163: 107-110
[6] Rosenberg AL, Schwartz GF, Feig SA, Patchefsky AS. Clinically occult breast lesions: localization and significance. Radiology 1987;162:167-170
[7] Meyer JE, Eberlein TJ, Stomper PC, Sonnenfeld MR. Biopsy of occult breast lesions: analysis of 1261 abnormalities. JAMA 1990;263:2341- 2343
[8] Tinnomans JGM, Wobbes T, Lubbers EC, van der Sluis RF, do Beer HHM. The significance of microcalcifications without palpable mass in the diagnosis of breast cancer. Surgery 1986;99:652-657
[9] Tabar L, Gad A. Screening for breast cancer: the Swedish trial. Radiology 1981;138:219-222
[10] Kopans DB, Swann CA. Observations on mammographic screening and the false-positive mammograms. AIR 1988;150:785-786 11. Walt AJ. Screening and breast cancer: a surgical perspective
[11] Elter M, Schulz-Wendtland R, Wittenberg T. The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. Medical physics. 2007 Nov;34(11):4164-72.
[12] Mokhtar, Sahar & Elsayad, Alaa. (2013). Predicting the Severity of Breast Masses with Data Mining Methods. International Journal of Computer Science Issues (IJCSI) March 2013 Issue (Volume 10, Issue 2)
[13] Pedro Ferreira , Nuno A. Fonseca , Ines Dutra , Ryan Woods, Elizabeth Burnside . Predicting malignancy from mammography findings and image–guided core biopsies. International Journal of Data mining and Bioinformatics (Volume 11, Issue 3)
[14] Campanini R. et al. (2003) A novel approach to mass detection in digital mammography based on Support Vector Machines (SVM). In: Peitgen HO. (eds) Digital Mammography. Springer, Berlin, Heidelberg.
[15] N. Dhungel, G. Carneiro and A. P. Bradley, "Automated Mass Detection in Mammograms Using Cascaded Deep Learning and Random Forests," 2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Adelaide, SA, 2015, pp. 1-8, doi: 10.1109/DICTA.2015.7371234.
[16] Oliver A., Lladó X., Freixenet J., Martí J. (2007) False Positive Reduction in Mammographic Mass Detection Using Local Binary Patterns. In: Ayache N., Ourselin S., Maeder A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007. MICCAI 2007. Lecture Notes in Computer Science, vol 4791. Springer, Berlin, Heidelberg.
[17] Rejani Y, Selvi ST. Early detection of breast cancer using SVM classifier technique. arXiv preprint arXiv:0912.2314. 2009 Dec 11.
[18] Vijayarajeswari R, Parthasarathy P, Vivekanandan S, Basha AA. Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform. Measurement. 2019 Nov 1; 146:800-5.
[19] Dheeba J, Selvi ST. Classification of malignant and benign microcalcification using SVM classifier. In2011 International Conference on Emerging Trends in Electrical and Computer Technology 2011 Mar 23 (pp. 686-690). IEEE.
[20] Ragab DA, Sharkas M, Marshall S, Ren J. Breast cancer detection using deep convolutional neural networks and support vector machines. PeerJ. 2019 Jan 28;7:e6201.

