BREAST CANCER PREDICTION USING MACHINE LEARNING

Ramik Rawal
School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Gorbachev Road, Vellore, Tamil Nadu 632014, India.

1. ABSTRACT

Women are seriously threatened by breast cancer with high morbidity and mortality. The lack of robust prognosis models results in difficulty for doctors to prepare a treatment plan that may prolong patient survival time. Hence, the requirement of time is to develop the technique which gives minimum error to increase accuracy. Four algorithm SVM, Logistic Regression, Random Forest and KNN which predict the breast cancer outcome have been compared in the paper using different datasets. All experiments are executed within a simulation environment and conducted in JUPYTER platform. Aim of research categorises in three domains. First domain is prediction of cancer before diagnosis, second domain is prediction of diagnosis and treatment and third domain focuses on outcome during treatment. The proposed work can be used to predict the outcome of different technique and suitable technique can be used depending upon requirement. This research is carried out to predict the accuracy. The future research can be carried out to predict the other different parameters and breast cancer research can be categorises on basis of other parameters.

Keywords — Breast Cancer, machine learning, feature selection, classification, prediction, KNN , Random Forest, ROC.

2. INTRODUCTION

The second major cause of women's death is breast cancer (after lung cancer). 246,660 of women's new cases of invasive breast cancer are expected to be diagnosed in the US during 2016 and 40,450 of women's death is estimated. Breast cancer is a type of cancer that starts in the breast. Cancer starts when cells begin to grow out of control. Breast cancer cells usually form a tumour that can often be seen on an x-ray or felt as a lump. Breast cancer can spread when the cancer cells get into the blood or lymph system and are carried to other parts of the body. The cause of Breast Cancer includes changes and mutations in DNA. There are many different types of breast cancer and common ones include ductal carcinoma in situ (DCIS) and invasive carcinoma. Others, like phyllodes tumours and angiosarcoma are less common. There are many algorithms for classification of breast cancer outcomes. The side effects of Breast Cancer are – Fatigue, Headaches, Pain and numbness (peripheral neuropathy), Bone loss and osteoporosis. There are many algorithms for classification and prediction of breast cancer outcomes. The present paper gives a comparison between the performance of four classifiers: SVM , Logistic Regression , Random Forest and kNN which are among the most influential data mining algorithms. It can be medically detected early during a screening examination through mammography or by portable cancer diagnostic tool. Cancerous breast tissues change with the progression of the disease, which can be directly linked to cancer staging. The stage of breast cancer (I–IV) describes how far a patient's cancer has proliferated. Statistical indicators such as tumour size, lymph node metastasis, and distant metastasis and so on are used to determine stages. To prevent cancer from spreading, patients have to undergo breast cancer surgery, chemotherapy, radiotherapy and endocrine. The goal of the research is to identify and classify Malignant and Benign patients and intending how to parametrize our classification techniques hence to achieve high accuracy. We are looking into many datasets and how further Machine Learning algorithms can be used to characterize Breast Cancer. We want to reduce the error rates with maximum accuracy. 10-fold cross validation test which is a Machine Learning Technique is used in JUPYTER to evaluate the data and analyse data in terms of effectiveness and efficiency.

2.1 MOTIVATION

Breast Cancer is the most affected disease present in women worldwide. 246,660 of women's new cases of invasive breast cancer are expected to be diagnosed in the U.S during 2016 and 40,450 of women's death is estimated. The development in Breast Cancer and its prediction fascinated. The UCI Wisconsin Machine Learning Repository Breast Cancer Dataset attracted as large patients with multivariate attributes were taken as sample set.

3. RELATED WORK

The cause of Breast Cancer includes changes and mutations in DNA. Cancer starts when cells begin to grow out of control. Breast cancer cells usually form a tumour that can often be seen on an x-ray or felt as a lump. There are many different types of breast cancer and common ones include ductal carcinoma in situ (DCIS) and invasive carcinoma. Others, like phyllodes tumours and angiosarcoma are less common. Wang, D.; Zhang and Y.-H Huang (2018) et al. [1] used Logistic Regression and achieved an Accuracy of 96.4 %. Akbugday et al., [2] performed classification on Breast Cancer Dataset by using KNN, SVM and achieved accuracy of 96.85%. KAYA KELES et al., [3] in the paper titled "Breast Cancer Prediction and Detection Using Data Mining" used Random Forest and achieved accuracy of 92.2%.Vikas Chaurasia and Saurabh Pal et al., [4] compare the performance criterion of supervised learning classifiers; such as Naïve Bayes, SVM-RBF kernel, RBF neural networks, Decision trees (J48) and simple CART; to find the best classifier in breast cancer datasets. Dalen, D.Walker and G. Kadam et al. [5] used ADABOOST and achieved accuracy of 97.5% better than Random Forest. Kavitha et al., [6] used ensemble methods with Neural Networks and achieved accuracy of 96.3% lesser than previous studies. According to Sinthia et al., [7] used backpropagation method with 94.2 % accuracy. The experimental result shows that SVM-RBF kernel is more accurate than other classifiers; it scores accuracy of 96.84% in Wisconsin Breast Cancer (original) datasets . We have used classification methods like SVM, KNN, Random Forest, Naïve Bayes, ANN. Prediction and prognosis of cancer development are focused on three major domains: risk assessment or prediction of cancer susceptibility, prediction of cancer relapse, and prediction of cancer survival rate. The first domain comprises prediction of the probability of developing certain cancer prior to the patient diagnostics. The second issue is related to prediction of cancer recurrence in terms of diagnostics and treatment, and the third case is aimed at prediction of several possible parameters characterizing cancer development and treatment after the diagnosis of the disease: survival time, life expectancy, progression, drug sensitivity, etc. The survivability rate and the cancer relapse are dependent very much on the medical treatment and the quality of the diagnosis. As we know that data pre-processing is a data mining technique that used for filter data in a usable format. Because the real-world dataset almost available in different format. It is not available as per our requirement so it must be filtered in understandable format. Data pre-processing is a proven method of resolving such issues. Data pre-processing convert the dataset into usable format for pre-processing we have used standardization method.

The following is the summary of the existing works on the given domain:

Table (1):

AUTHOR	DATASET USED	TOOL USED	TECHNIQUE USED	ADVANTAGES	ACCURACY	ERROR RATE
1. Wang et al. [1]	Electronic health records	WEKA	Logistic regression	5-year survivability prediction using logistic regression	96.4 %	0.33

2. V Chaurisya & S Paul [4]	Wisconsin breast cancer	WEKA	Statistical Feature Selection	Patient features sorted out from data materials are statistically tested based on the type of individual feature. Then 51 attributes or features are selected out, a feature's importance score is calculated. XGBoost lgorithm is done by repeating 10-fold cross validation.	92.3 %	0.3%
3. Akbugday [2]	Breast Cancer Wisconsin dataset	WEKA	KNN and SVM	Optimal k-Value for a k-NN classifier, g k-NN is a lightweight, lazy learning algorithm with very short build times.	KNN- 96.85% NAÏVE BAYES - 95.99% SVM – 96.85%	0.66%
4. Keles, M. Kaya, [3]	Wisconsin Diagnostic Breast Cancer dataset	Python	SVM vs KNN, decision trees and Naives bayes	SVMs map the input vector into a feature space of higher dimensionality and identify the hyperplane that separates the data points into two classes. The marginal distance between the decision hyperplane and the instances that are closest to boundary is maximized.	up to 96.91%	0.33
5. KELES et al., (2019) [3]	Wisconsin Dataset	WEKA	RANDOM FOREST	Each dataset is generated with displacement from the original dataset. Then, trees are developed using a random selection feature, but are not pruned.	92.2 %	
6. Chauraisa et. al [4]	UC Irvine machine learning repository	WEKA	Naive Bayes, J48 Decision Tree and Bagging algorithm	Decision tree (C5) is the best predictor on the holdout sample (this prediction accuracy is better than any reported in the literature	96.5%	
7. Delen at al. [5]	Cancer Society	WEKA	ADABOOST	Low in error rate, performing well in the low noise data set. The advantage of this algorithm is that it requires less input parameters and needs little prior knowledge about the weak learner	97.5 %	

8. Kavithaa et. at [6]	Cancer Society	MATLAB	Ensemble method with Logistic and Neural Network	Multiple Learners are combined giving higher accuracy.	96.3 %	
9. Sinthia et al. [7]	Wisconsin Diagnosis Breast Cancer BCI dataset	CAD System	Logistic Regression and the Backpropagation neural Network		94.2 %	
[10] Chaurasia et. at [8]	Wisconsin Breast Cancer (original) datasets	WEKA	SVM	It gives the most optimal hyperplane to distinguish between two classes	97.13%	
[11] Khuriwal et.al [9]	Haberman's Survival dataset	WEKA	NAÏVE BAYES AND SVM	Helps in marginalizing the hyper-parameters and differentiating classes.	74.44%	
[12] Khourdifi et al. [10]	Wisconin breast cancer dataset	WEKA	Fast Correlation-Based Filter with SVM, Random Forest, Naive Bayes, K-NN and MLP	Attributes are reduced by deleting irrelevant and redundant attributes, which have no meaning in the classification task techniques.	96.1%	0.0404
[13] Mohana,et. al [11]	WISCONIN BREAST CANCER DATASET	WEKA TOOL	DECISION TREE	Helps in Splitting and choosing the best attributes	96.3 %	
[14] Shravya at al. [12]	UCI repository	Spyder	SVM	Hyperplane separates two classes which helps in higher accuracy.	92.7%	
[15] Wang et. al [13]	Wisconsin Breast Cancer Database (1991) and Wisconsin Diagnostic Breast Cancer (1995)	WEKA	PCA	dimension reduction Wang and Yoon technique, manifests some advantages in terms of prediction accuracy and efficiency.	Eight PCs are chosen based on the scree plot, which explains 92.6% of total correlation. And ten PCs are selected based on 95% correlation	
[16] Bellaachia et. al [14]	SEER Public-Use Data.	WEKA	Naïve Bayes	Gives a probabilistic model for classification Helping in classification.	96.3 %	

[17] Kim et. al [15]	The breast cancer survivability dataset (1973–2003) from SEER	WEKA	SSL, and SSL-Co(Semi-Supervised Learning)	SEMI-SUPERVISED CO-TRAINING SSL may be a good candidate to use as a predictive model for cancer survivability, particularly when the available dataset for model learning has an abundance of unlabelled patient cases.	0.84 %	
[18] Bellaachia et. al [14]	SEER database.	WEKA	Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree	clear and fast classifier [10]. It has been called 'Naïve' due to the fact that it assumes mutually independent attributes	84.5 % - Naïve Bayes, 86.7% – NN, 81.3% –C4.5	0.57
[19] Khuriwal and Mishra (2018) [16]	Wisconsin Diagnosis Breast Cancer dataset. UCI open database	DWT tool	Ensemble Voting Method, Logistic Regression	useful for predicting the class a binomial target feature.	98.50%	0.99%
[20] AMRANE et al. (2018) [17]	Breast Cancer Dataset University of California, Irvine (UCI)	WEKA	KNN and NAÏVE BAYES	KNN classifiers are ranked first in terms of accuracy and duration.	0.975109-KNN 0.961932	
[21] Khuriwal et. al [16]	Wisconsin Breast Cancer database.	WEKA	Deep Learning Neural Network(convolutional neural network)	Number of hidden layers	99.67%	0.0246
[22] Al-hadidi et al. [18]	General Sample	MATLAB	Logistic Regression and Backpropagation neural Network	BPNN is easy to implement and has been used widely for classification purposes. LR needs a hypothesis and a cost function which optimizes performance.	Greater than 93.7%	Less than 0.07
[23] Kibeom et. al [19]	Gene Expression Dataset Collection	WEKA	C4.5, Bagging and ADABOOST Decision trees	Ensemble Method helps to combine multiple learners.	Single C4.5 – 95.6%, Bagging C4.5 – 93.29%,ADABOOST C4.5 – 92.62%	Sensitivity – 56% and 72%
4] Cruz et. al [20]	Pubmed (biomedical literature), the Science Citation Index	MATLAB	SVM, NAÏVE-BAYES	Helps to form a decision boundary and helps in classification.	97.3%	
[25] Medjahed et. al [21]	Wisconsin breast cancer dataset	WEKA	Decision Trees	Helps in splitting	96.1 %	

4. PROPOSED METHODOLOGY

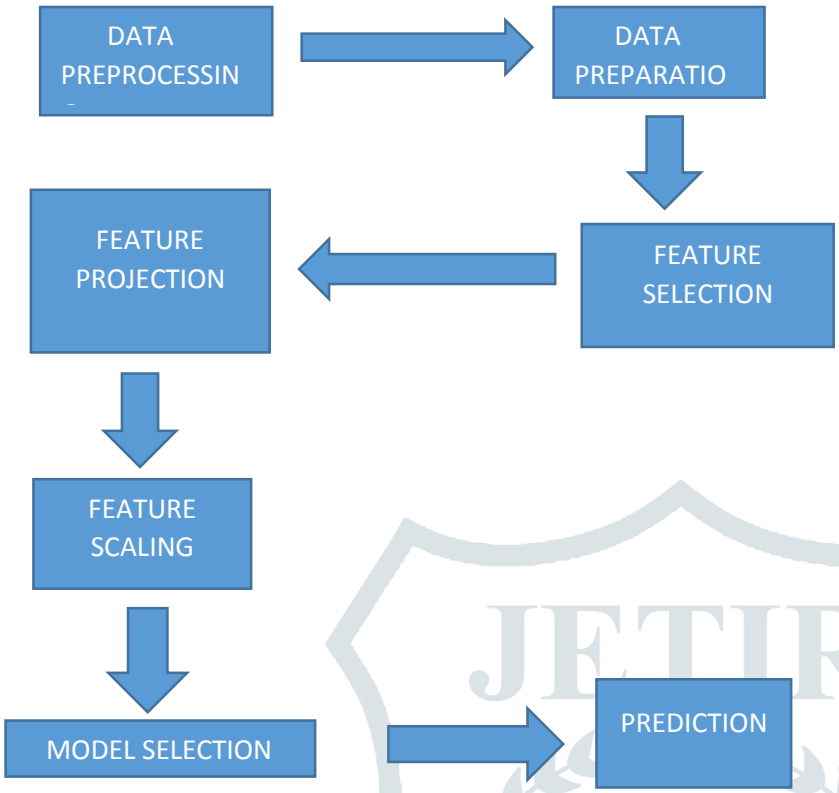

Fig. (1) Phases of Machine Learning consists of seven phases, the phases are elaborated as given below: -

Phase 1 - Pre-Processing Data

The first phase we do is to collect the data that we are interested in collecting for pre-processing and to apply classification and Regression methods. Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real world data is often incomplete, inconsistent, and lacking certain to contain many errors. Data pre-processing is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing. For pre-processing we have used standardization method to pre-process the UCI dataset. This step is very important because the quality and quantity of data that you gather will directly determine how good your predictive model can be. In this case we collect the Breast Cancer samples which are Benign and Malignant. This will be our training data.

Phase 2 - DATA PREPRATION

Data Preparation**,** where we load our data into a suitable place and prepare it for use in our machine learning training. We'll first put all our data together, and then randomize the ordering.

Phase 3 - FEATURES SELECTION

In machine learning and statistics, feature selection, also known as variable selection, attribute selection, is the process of selection a subset of relevant features for use in model construction.

Data File and Feature Selection Breast Cancer Wisconsin (Diagnostic):- Data Set from Kaggle repository and out of 31 parameters we have selected about 8-9 parameters. Our target parameter is breast cancer diagnosis – malignant or benign. We have used Wrapper Method for Feature Selection. The important features found by the study are: Concave points worst, Area worst, Area se, Texture worst, Texture mean, Smoothness worst, Smoothness mean, Radius mean, Symmetry mean.

We have used Wrapper Method for Feature Selection. The important features found by the study are: 1. Concave points worst 2. Area worst 3. Area se 4. Texture worst 5. Texture mean 6. Smoothness worst 7. Smoothness mean 8. Radius mean 9. Symmetry means.

Attribute Information:

ID number 2) Diagnosis (M = malignant, B = benign) 3–32)

Phase 4 - Feature Projection

Feature projection is transformation of high-dimensional space data to a lower dimensional space (with few attributes). Both linear and nonlinear reduction techniques can be used in accordance with the type of relationships among the features in the dataset.

Phase 5 - Feature Scaling

Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidian distance between two data points in their computations. We need to bring all features to the same level of magnitudes. This can be achieved by scaling.

Phase 6 - Model Selection

Supervised learning is the method in which the machine is trained on the data which the input and output are well labelled. The model can learn on the training data and can process the future data to predict outcome. They are grouped to Regression and Classification techniques. A regression problem is when the result is a real or continuous value, such as "salary" or "weight". A classification problem is when the result is a category like filtering emails spam" or "not spam". Unsupervised Learning: Unsupervised learning is giving away information to the machine that is neither classified nor labelled and allowing the algorithm to analyse the given information without providing any directions. In unsupervised learning algorithm the machine is trained from the data which is not labelled or classified making the algorithm to work without proper instructions. In our dataset we have the outcome variable or Dependent variable i.e. Y having only two set of values, either M (Malign) or B (Benign). So Classification algorithm of supervised learning is applied on it. We have chosen three different types of classification algorithms in Machine Learning. We can use a small linear model, which is a simple.

Phase 7 - PREDICTION

Machine learning is using data to answer questions. So Prediction**,** or inference, is the step where we get to answer some questions. This is the point of all this work, where the value of machine learning is real.

METHODS USED

(1) Logistic Regression

Logistic regression was introduced by statistician DR Cox in 1958 and so predates the field of machine learning. It is a supervised machine learning technique, employed in classification jobs (for predictions based on training data). Logistic Regression uses an equation like Linear Regression, but the outcome of logistic regression is a categorical variable whereas it is a value for other regression models. Binary outcomes can be predicted from the independent variables.

The general workflow is:

(1) get a dataset

(2) train a classifier
(3) make a prediction using such classifier

(2) k-Nearest Neighbour (k-NN)

K-Nearest Neighbour is a supervised machine learning algorithm as the data given to it is labelled. It is a nonparametric method as the classification of test data point relies upon the nearest training data points rather than considering the dimensions (parameters) of the dataset.

ALGORITHM

(1) Input the dataset and split it into a training and testing set.

(2) Pick an instance from the testing sets and calculate its distance with the training set.

(3) List distances in ascending order.

(4) The class of the instance is the most common class of the 3 first trainings instances (k=3).

(3) Support Vector machine

Support Vector Machine is a supervised machine learning algorithm which is doing well in pattern recognition problems and it is used as a training algorithm for studying classification and regression rules from data. SVM is most precisely used when the number of features and number of instances are high. A binary classifier is built by the SVM algorithm. In an SVM model, each data item is represented as points in an n-dimensional space where n is the number of features where each feature is represented as the value of a coordinate in the n-dimensional space.

Here's how a support vector machine algorithm model works:

(1) First, it finds lines or boundaries that correctly classify the training dataset.
(2) Then, from those lines or boundaries, it picks the one that has the maximum distance from the closest data points.

(3) Random Forest

Random forest, like its name implies, consists of many individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

RESULT AND DISCUSSION OF PROPOSED METHODOLOGY

The work was implemented on i3 processor with 2.30GHz speed, 2 GB RAM, 320 GB external storage and all experiments on the classifiers described in this paper were conducted using libraries from Anaconda machine learning environment. In Experimental studies we have partition 70-30% for training & testing. JUPYTER contains a collection of machine learning algorithms for data pre-processing, classification, regression, clustering and association rules. Machine learning techniques implemented in JUPYTER are applied to a variety of real-world problems. The results of the data analysis are reported. To apply our classifiers and evaluate them, we apply the 10-fold cross validation test which is a technique used in evaluating

predictive models that split the original set into a training sample to train the model, and a test set to evaluate it. After applying the pre-processing and preparation methods, we try to analyse the data visually and figure out the distribution of values in terms of effectiveness and efficiency.

We evaluate the effectiveness of all classifiers in terms of time to build the model, correctly classified instances, incorrectly classified instances and accuracy.

Table No. 2

Algorithms	Accuracy	Sensitivity/ Recall	Specificity	Precision	F1-Score	ROC
Logistic Regression	0.9624413145539906	TP / (TP + FN)	TN / (FP + TN)			
Decision Tree	1.0					
Random Forest	0.9929577464788732	TP/(TP + FN)	TN/(FP + TN)	(TP + TN) / (TP + FP + TN + FN)		
SVM						

TP- True positive

FN- false Negative

TN- True negative

FN- False Negative

In order to better measure the performance of classifiers, simulation error is also considered in this study. To do so, we evaluate the effectiveness of our classifier in terms of: x Kappa statistic (KS) as a chance-corrected measure of agreement between the classifications and the true classes, x Mean Absolute Error (MAE) as how close forecasts or predictions are to the eventual outcomes, x Root Mean Squared Error (RMSE), x Relative Absolute Error (RAE), x Root Relative Squared Error (RRSE).

In this section, the results of the data analysis are reported. To apply our classifiers and evaluate them, we apply the 10-fold cross validation test which is a technique used in evaluating predictive models that split the original set into a training sample to train the model, and a test set to evaluate it. After applying the pre-processing and preparation methods, we try to analyse the data visually and figure out the distribution of values in terms of effectiveness and efficiency

EFFECTIVENESS

In this section, we evaluate the effectiveness of all classifiers in terms of time to build the model, correctly classified instances, incorrectly classified instances and accuracy. The results are shown in Table 2 and Fig. 1.

The ROC space is defined with true positives and false positives as the x and y coordinates, respectively. ROC curve summarizes the performance across all possible thresholds. The diagonal of the ROC graph can be interpreted as random guessing, and classification models that fall below the diagonal are considered worse than random guessing.

- You have not implemented Naïve however below figure is Naïve Bayes.

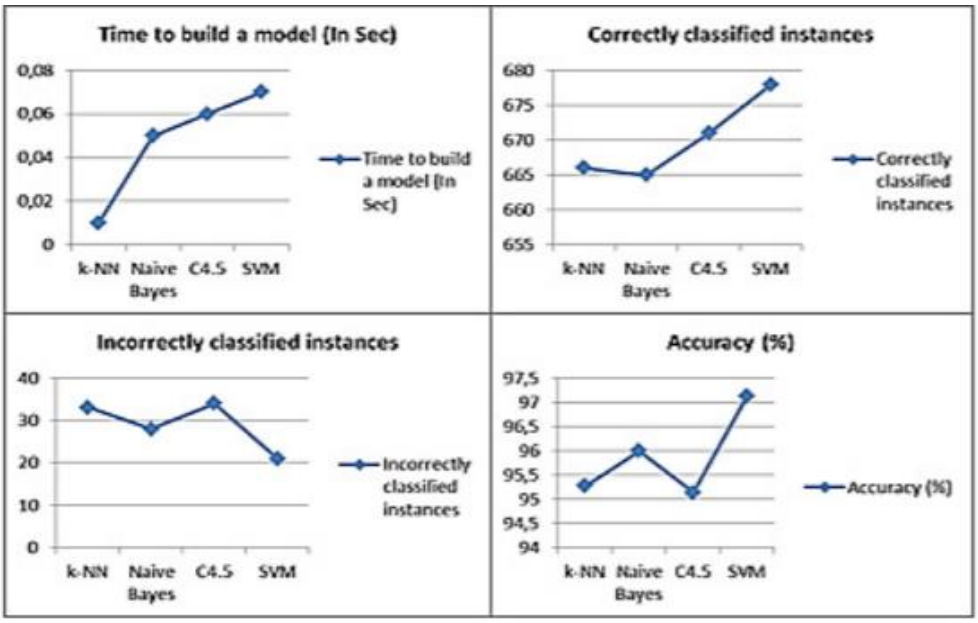

Fig. 1. Comparative graph of different classifiers.

After creating the predicted model, we can now analyse results obtained. SVM and C4.5 got the highest value (97%) of TP for benign class but KNN correctly predicts 97 % of instance that belongs to malignant class

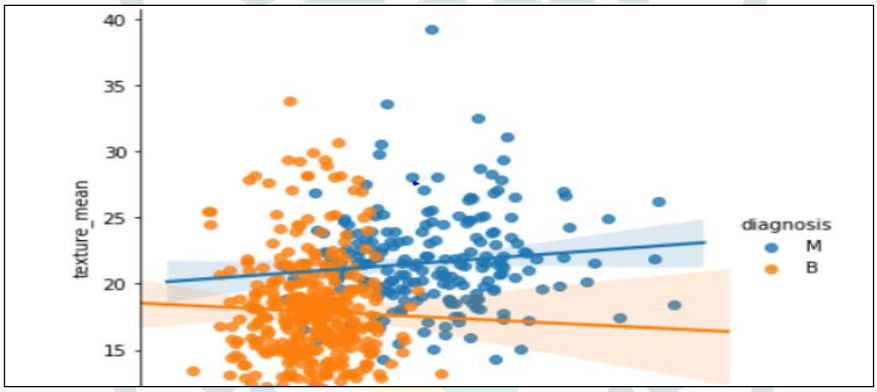

FIG 1:- Comparison of compactness mean vs smoothness mean where orange represents Benign and Blue represents Malignant

CONCLUSION AND FUTURE WORK

We can notice that SVM takes about 0.07 s to build its model unlike k-NN that takes just 0.01 s. It can be explained by the fact that k-NN is a lazy learner and does not do much during training process unlike others classifiers that build the models. In other hand, the accuracy obtained by SVM (97.13%) is better than the accuracy obtained by C4.5, Naïve Bayes and k-NN that have an accuracy that varies between 95.12 % and 95.28 %. It can also be easily seen that SVM has the highest value of correctly classified instances and the lower value of incorrectly classified instances than the other classifiers.

After creating the predicted model, we can now analyse results obtained in evaluating efficiency of our algorithms. SVM and C4.5 got the highest value (97 %) of TP for benign class but k-NN correctly predicts 97% of instance that belong to malignant class. The FP rate is lower when using SVM classifiers (0.03 for benign class and 0.02 for malignant class), and then other algorithms follow: k-NN, C4.5 and NB. From these results, we can understand why SVM has outperformed other classifiers

In summary, SVM was able to show its power in terms of effectiveness and efficiency based on accuracy and recall.

FUTURE WORK

The analysis of the results signifies that the integration of multidimensional data along with different classification, feature selection and dimensionality reduction techniques can provide auspicious tools for inference in this domain. Further research in this field should be carried out for the better performance of the classification techniques so that it can predict on more variables. We are intending how to parametrize our classification techniques hence to achieve high accuracy. We are looking into many datasets and how further Machine Learning algorithms can be used to characterize Breast Cancer. We want to reduce the error rates with maximum accuracy.

REFERENCES

[1] Wang, D. Zhang and Y. H. Huang "Breast Cancer Prediction Using Machine Learning" (2018), Vol. 66, NO. 7.

[2] B. Akbugday, "Classification of Breast Cancer Data Using Machine Learning Algorithms," 2019 Medical Technologies Congress (TIPTEKNO), Izmir, Turkey, 2019, pp. 1-4.

[3] Keles, M. Kaya, "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study." Tehnicki Vjesnik - Technical Gazette, vol. 26, no. 1, 2019, p. 149+.

[4] V. Chaurasia and S. Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability", IJCSMC, Vol. 3, Issue. 1, January 2014, pg.10 – 22.

[5]] Delen, D.; Walker, G.; Kadam, A. Predicting breast cancer survivability: A comparison of three data mining methods. Artif. Intell. Med. 2005, 34, 113–127.

[6] R. K. Kavitha1, D. D. Rangasamy, "Breast Cancer Survivability Using Adaptive Voting Ensemble Machine Learning Algorithm Adaboost and CART Algorithm" Volume 3, Special Issue 1, February 2014

[7] P. Sinthia, R. Devi, S. Gayathri and R. Sivasankari, "Breast Cancer detection using PCPCET and ADEWNN", CIEEE' 17, p.63-65

[8] Vikas Chaurasia and S.Pal, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis" (FAMS 2016) 83 (2016) 1064 – 1069

[9] N. Khuriwal, N. Mishra. "A Review on Breast Cancer Diagnosis in Mammography Images Using Deep Learning Techniques", (2018), Vol. 1, No. 1.

[10] Y. Khourdifi and M. Bahaj, "Feature Selection with Fast Correlation-Based Filter for Breast Cancer Prediction and Classification Using Machine Learning Algorithms," 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Rabat, Morocco, 2018, pp. 1-6.

[11] R. M. Mohana, R. Delshi Howsalya Devi, Anita Bai, "Lung Cancer Detection using Nearest Neighbour Classifier", International Journal of Recent Technology and Engineering (IJRTE), Volume-8, Issue-2S11, September 2019

[12] Ch. Shravya, K. Pravalika, Shaik Subhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-6, April 2019.

[13] Haifeng Wang and Sang Won Yoon, "Breast Cancer Prediction Using Data Mining Method", Proceedings of the 2015 Industrial and Systems Engineering Research Conference,

[14] Abdelghani Bellaachia, Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques"

[15] Juhyeon Kim, Hyunjung Shin, Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data, Journal of the American Medical Informatics Association, Volume 20, Issue 4, July 2013, Pages 613–618.

[16] N. Khuriwal and N. Mishra, "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm," 2018 IEEMA Engineer Infinite Conference (eTechNxT), New Delhi, 2018, pp. 1-5.

[17] M. Amrane, S. Oukid, I. Gagaoua and T. Ensarİ, "Breast cancer classification using machine learning," 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, 2018, pp. 1-4.

[18] M. R. Al-Hadidi, A. Alarabeyyat and M. Alhanahnah, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm," 2016 9th International Conference on Developments in eSystems Engineering (DeSE), Liverpool, 2016, pp. 35-39.

[19] Kibeom Jang, Minsoon Kim, Candace A Gilbert, Fiona Simpkins, Tan A Ince, Joyce M Slingerland "WEGFA activates an epigenetic pathway regulating ovarian cancer initiating cells" Embo Molecular Medicines Volume 9 Issue 3 (2017)

[20] Joseph A. Cruz and David S. Wishart "Applications of Machine Learning in cancer prediction and prognosis Cancer informatics" 2(3):59-77 · February 2007

[21] SA Medjahed, TA Saadi, A Benyettou "Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules" International Journal of Computer Applications 62 (1), 2013