Identification of Bio-Markers for Breast Cancer Detection through Data Mining Methods

R. Geetha Ramani, G. Sivagami

ABSTRACT--- Worldwide, breast cancer is the leading type of cancer in women accounting for 25% of all cases. Survival rates in the developed countries are comparatively higher with that of developing countries. This had led to the importance of computer aided diagnostic methods for early detection of breast cancer disease. This eventually reduces the death rate. This paper intents the scope of the biomarker that can be used to predict the breast cancer from the anthropometric data. This experimental study aims at computing and comparing various classification models (Binary Logistic Regression, Ball Vector Machine (BVM), C4.5, Partial Least Square (PLS) for Classification, Classification Tree, Cost sensitive Classification Tree, Cost sensitive Decision Tree, Support Vector Machine for Classification, Core Vector Machine, ID3, K-Nearest Neighbor, Linear Discriminant Analysis (LDA), Log-Reg TRIRLS, Multi Layer Perceptron (MLP), Multinomial Logistic Regression (MLR), Naïve Bayes (NB), PLS for Discriminant Analysis, PLS for LDA, Random Tree (RT), Support Vector Machine SVM) for the UCI Coimbra breast cancer dataset. The feature selection algorithms (Backward Logit, Fisher Filtering, Forward Logit, ReleifF, Step disc) are worked out to find out the minimum attributes that can achieve a better accuracy. To ascertain the accuracy results, the Jack-knife cross validation method for the algorithms is conducted and validated. The Core vector machine classification algorithm outperforms the other nineteen algorithms with an accuracy of 82.76%, sensitivity of 76.92% and specificity of 87.50% for the selected three attributes, Age, Glucose and Resistin using ReleifF feature selection algorithm.

Index Terms — Biomarker, Breast Cancer, Core Vector Machine, Data Mining, Feature Selection.

I. INTRODUCTION

Breast cancer [24] is now the most common cancer in most cities in India, and second most common in the rural areas. The point worth noting is that breast cancer accounts to 25% to 32% of all female cancers in all major cities. It implies that one-fourth of all female cancer cases are breast cancers. More and more number of patients are being diagnosed with breast cancer to be in the younger age groups [21].

According to health ministry of India breast cancer ranks as the number one cancer among Indian females with rate as high as 25.8 per 100,000 women and mortality of 12.7 per 100,000 women. India continues to have a low survival rate for breast cancer, with only 66.1% women diagnosed with the disease between 2010 and 2014 surviving, a Lancet study found. The US and Australia had survival rates as high as 90%, according to the study [21].

The major reason are Lifestyle changes such as bearing a child late in life, lack of breastfeeding, medical use of hormones, menarche occurring in younger people, lack of awareness of early signs of breast cancer and screening methods, secondly non- availability of diagnostic centres and knowledgeable oncologists. The domains that need attention include primary prevention, secondary prevention (early detection), and diagnostic modalities including pathology, treatment, palliative care, and translational research including biomarkers. There need to be systematic efforts at researching, preserving, and promoting those factors that "protect" Indian women from breast cancer [21].

Clinical Data mining [8] is the application of data mining techniques using clinical data. It has three objectives. Understanding the clinical data, assist healthcare professionals and develop a data analysis methodology suitable for medical data. It [6] involves the conceptualisation, extraction analysis and interpretation of available clinical data for practice knowledge-building, clinical decision making and practitioner reflection.

A detailed study on breast cancer prediction system using the anthropometric data is analysed. The predictors or features is computed for twenty different classification models in order to get a remarkable classification accuracy levels. The challenge lies in choosing of minimum features that does not pay off the accuracy. This paper is channelized to achieve this objective.

This paper is organised in the following manner. The Section II covers the previous work carried out, Section III focus on the proposed system design which in detail discuss on the dataset used, the training and test design. The Section IV covers the feature selection algorithms used in this paper for feature selection and reduction process. The Section V is mainly focussed on the preview of the classification algorithms used in this study. The section VI is the sketch of results obtained and concluded.

II. RELATED WORK

To Miguel et al [13] had come up with biomarkers for breast cancer prediction. Their experimental study focussed on machine learning algorithms, logistic regression, random forest and support vector machine. And concluded that support vector machine predicts the presence of breast cancer with sensitivity ranging from 82 to 88% and

Revised Manuscript Received on July 10, 2019.
Dr. R. Geetha Ramani, India
G. Sivagami,India (E-mail: shiv.haida@gmail.com)

specificity ranging between 85% to 90%. Also they suggested the attributes Glucose, Age, Resistin and BMI as the effective biomarker for predicting the breast cancer.

Performance Evaluation of Machine Learning methods for Breast Cancer Prediction [18] authored by Yixuan Li and Zixuan Chen did an experimental study with the breast cancer Coimbra dataset and the Wisconsin breast cancer dataset. The analyzed with five different classification algorithms like Decision Tree, Random Forest, Support Vector Machine, Neural Network and Logistic Regression. They concluded that Random Forest as the primary classification model based on the evaluation parameters.

Shamona et al, [16] in their paper used the Wisconsin Prognostic Breast Cancer (WPBC) data. They tried various feature selection and classification algorithms to train the system. And concluded C4.5 and Random tree algorithms gives 100% accuracy .

Vikas et al, [17] in their prediction evolved the Naïve Bayes as an efficient classifier with an accuracy of 97.36% on the hold out sample method using 10 cross fold validation. They compared three algorithms namely Naïve Bayes, RBF network and J48 on the breast cancer Wisconsin dataset.

Ahmad et al, [2] in their paper used three classification techniques to predict the recurrence of the breast cancer and concluded SVM as the best classifier predictor than Artificial Neural Network and Decision tree. They used the ICBC dataset from the National Cancer Institute of Tehran .

Sardouk et al, [Classification of Breast Cancer Using Data Mining] aimed at finding a suitable classifier and concluded that RBF and MLP outperforms other classifiers.

This work attempts in classifying the unhealthy and the controlled individuals by the following proposed methodology.

III. PROPOSED SYSTEM DESIGN

3.1 Dataset Description

The dataset used in the experimental study is the Breast Cancer Coimbra dataset from the UCI Machine Learning Repository. The clinical predictors or the parameters are retrieved from the routine blood sample analysis. There are nine quantitative attributes that predict the presence or absence of breast cancer. The number of instances in this dataset is 116. The attribute description is described below in Table I.

Table 1: Dataset description

S No.	Attributes	Description
1	Age	The age of the individual in years
2	BMI	Body Mass Index expressed in kilogram per metre square
3	Glucose	Glucose measured in milligrams per decilitre
4	Insulin	Insulin measured in microunit per millilitre
5	HOMA	Homeostatic Model Assessment for assessing Beta cell function and Insuling Resistance
6	Leptin	Figures out the fat tissue in the body. Measured in nanogram per millilitre
7	Adiponectin	Protein hormone in regulating the glucose levels. Measured in microgram per milliliter
8	Resistin	Indicator of bad cholesterol level. Measured in nanogram per millilitre
9	MCP-1	monocyte chemoattractant protein 1 measured in picogram per decilitre

3.2 Proposed System Design

The upcoming paragraph narrates the overview of the proposed system design of the research work depicted in Fig 1.

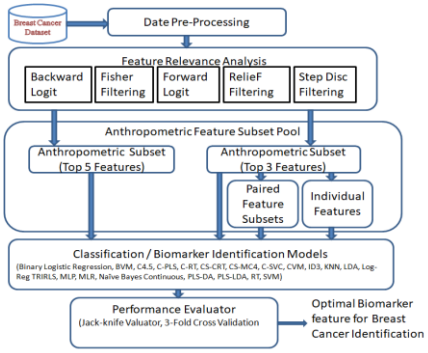

Fig 1: Proposed System Design

The breast cancer dataset is loaded in the machine learning software and ensured for its correctness. This is done in the data visualization phase. Then the classification models such as Binary Logistic Regression, KNN, C4.5, Random Tree, SVM, ID3, CVM, Multilayer Perceptron, Naïve Bayes, C-SVC, LDA etc. are executed with all nine attributes using the evaluator Jack-knife and 3-fold cross validation method. Several Model Evaluation techniques ensure how accurately a predictive model will perform in actual scenario with respect to real time data provided. Among nine features, a subset of best features that predicts the binary dependent variable is detected using the feature selection algorithms. With the selected features given by the each individual feature selection algorithm, all the classification algorithms are executed and the results are compared to find the best feature selection algorithm and the best classification model.

3.3 Data Visualization

The Breast Cancer COIMBRA dataset has 10 predictors, all quantitative and a binary dependant variable indicating the presence or absence of breast cancer. This dataset converted to attribute relation file format and loaded in the TANAGRA, a free data mining software tool. For ease, the binary dependent variable is converted to discrete. Value 1 is converted as “Healthy Controls” and value 2 is converted to “Patients”. Since there are no missing values in the dataset, the data is loaded and checked for its successfulness. The below Table 2 gives the meta data information.

Table 2: Meta-data Information

S.No	Attribute	Category
1	Age	Continuous
2	BMI	Continuous
3	Glucose	Continuous
4	Insulin	Continuous
5	HOMA	Continuous
6	Leptin	Continuous
7	Adiponectin	Continuous
8	Resistin	Continuous
9	MCP-1	Continuous
10	Classification	Discrete(2 Values)

IV. FEATURE SELECTION ALGORITHM

Some of the core idea for performing feature selection is to reduce the training time. It is the best weapon against the Curse of Dimensionality and it is a powerful defense against over fitting, increasing generalizability. It [12] is also called variable selection or attributes selection. It is the automatic selection of attributes in data (such as columns in tabular data) that are most relevant to the predictive modeling problem.

In this paper, five feature selection algorithms are considered for feature relevance analysis. They are Backward and Forward Logit, Fisher Filtering, ReliefF filtering and Stepwise Discriminant Analysis. The ReliefF feature selection attributes / features produced a better classification accuracy and hence it is discussed here.

ReliefF Filtering

Relief calculates the features score for each features and then ranking applied to select the top scoring features. The scores are applied as feature weights. Relief feature scoring is based on the identification of feature value differences between nearest neighbour instances.

V. CLASSIFICATION ALGORITHM

Data analysis is performed in order to extract the models which best describes or predict the future trends. The classification model predicts the class labels that are categorical in nature. In this paper, twenty classification models are considered inappropriate to pick the right classifier. The Core Vector Machine produces a better accuracy results compared with other classifier models. Overview of this algorithms is given below.

Kernel methods [7], such as the support vector machine (SVM), are often formulated as quadratic programming (QP) problems. However, given m training patterns, a naive implementation of the QP solver takes O(m 3) training time and at least O(m2) space. Hence, scaling up these QPs is a major stumbling block in applying kernel methods on very large data sets, and a replacement of the naive method for finding the QP solutions is highly desirable. Recently, by using approximation algorithms for the minimum enclosing ball (MEB) problem, the core vector machine (CVM) algorithm that is much faster and can handle much larger data sets than existing SVM implementations. However, the CVM can only be used with certain kernel functions and kernel methods. The generalized CVM algorithm can now be used with any linear/nonlinear kernel and can also be applied to kernel methods and the ranking SVM. Moreover, like the original CVM, its asymptotic time complexity is again linear in m and its space complexity is independent of m. Experiments show that the generalized CVM has comparable performance with state-of-the-art Support Vector Machine and Support Vector Regression implementations, but is faster and produces fewer support vectors on very large data sets.

VI. EXPERIMENTAL RESULTS

The study is carried out in a procedural fashion. The dataset is verified for its missing values. Since there are nil occurrence, the entire 116 data is considered for our study purpose. As mentioned in the Data Visualization process, the class label is converted into Discrete variables, either as "Healthy Controls" for the given class label value "1" or "Patients".

The experimental procedure is mainly aimed at establishing a predictor system for predicting the breast cancer disease. To commence, twenty diverged Classification models are selected with which the data can be analyzed. The statistical procedure for checking the effectiveness of the classification models is required and necessary. To ensure the performance of the classification model, a 3-fold cross validation and jack-knife re-sampling techniques are utilized to check for the bias estimation. The classification models accuracy values are computed for these two model evaluation techniques.

Experiment 1: Classifier Model Analysis by considering All attributes

As a first phase of the experiment, all the attributes ie. Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin and MCP-1 for the prediction of breast cancer disease are considered. The values are tabulated in Table 3. In this the CVM and the Log-Reg TRIRLS classification models had provided an accuracy percentage of 76.72% using jack-knife validation technique.

Table 3 Classification model accuracy % with all 9 attributes

Classification Models	Accuracy %	
	3 fold Cross Validation	Jack-knife cross validation
Binary Logistic Regression	75.44	75.86
BVM	72.81	75
C4.5	65.79	75.86
C-PLS	71.05	68.1
C-RT	64.04	71.55
CS-CRT	64.04	71.55
CS-MC4	67.54	75.86
C-SVC	68.42	71.55
CVM	73.68	76.72
ID3	55.26	55.17
KNN	73.68	70.69
LDA	73.68	73.28

Log-Reg TRIRLS	75.44	76.72
MLP	73.68	72.41
MLR	75.44	75.86
NB Continuous	65.79	68.1
PLS-DA	65.79	66.38
PLS-LDA	65.79	66.38
RT	69.3	60.34
SVM	57.02	64.66

Since the results produced are not pretty good, an analysis is tried, if any features out of the 9 is helping for this low accuracy prediction. So a feature analysis is performed to enable the prediction system with better predictors.

Experiment 2: Feature Relevance Analysis

Feature selection algorithms aims for finding the relevant features that can be used as a better predictor. Having irrelevant features can significantly decrease the accuracy of the classification model. Therefore selecting the subset of relevant features effectively increases the model accuracy. The feature selection algorithms used in this paper are Backward-Logit, Fisher Filtering, Forward-Logit, ReliefF filtering and stepwise discriminant analysis. The significant of the attributes selection are performed based on the uni-variate metric that helps in ranking of the attributes.

The backward logit regression model starts by considering all the attributes initially. The probability value (p-value) is calculated and greater the value, least significant the attribute. This is an iterative process until a threshold is reached. A reverse process is the forward approach where the model keeps on adding the significant attributes by considering the lowest residual sum of squares. The fisher filtering algorithm ranks the attributes according to the relevance calculated. The ReliefFfeature selection algorithm finds hit and miss instances using the Manhattan to find the relevant attributes. The Stepdisc algorithm is based on the lambda probability distribution. The subset of the attributes are framed still a stopping rule is achieved.

The feature selection algorithms ranks the features based on its significant values and are tabled in 4.

Table 4 Feature Relevance analysis

S.No	Feature Selection Algorithm	Relevant Features
1	Backward Logit	Glucose, BMI, Resistin, Insulin, Age , MCP-1, Leptin, HOMA. Adiponectin.
2	Fisher Filtering	Glucose, HOMA, Insulin, Resistin, BMI, MCP-1, Age, Adiponectin, Leptin
3	Forward Logit	Glucose, BMI, Resistin, Leptin, Age, Insulin, HOMA, MCP-1, Adiponectin.
4	ReliefF	Age,Glucose, Resistin, Insulin, HOMA, BMI, Adiponectin, Leptin, MCP-1
5	Stepdisc	Glucose , BMI, Resistin, Age, Insulin, Adiponectin, HOMA, MCP-1, Leptin

By analyzing the various feature selection algorithms, it can be concluded that Glucose plays a vital role in predicting the

Breast cancer disease. Also BMI, Insulin, Resistin and Age attributes plays an essential role in anticipating the class label. In the forthcoming experiments, if these feature relevance plays a substantial role.

Experiment 3: Accuracy of Classification model based on based on the top 5 features

A close observance of the feature relevance results gives us an impression that top five features are nearly redundant, explicitly showing its importance. And hence as an succeeding procedure, the top 5 features are selected and the accuracy is computed for the models.

Table 5 picture out the accuracy percentage results for the 3-fold cross validation method for the top 5 selected attributes and Table 6 for the jack-knife cross validation method for the same criteria.

Table 5 Accuracy of classification models By considering top 5 features using 3 FOld cross validation

Accuracy % using 3 Fold cross validation (Top 5 features selected)					
Algorithm	**Backward-Logit**	**Fisher Filtering**	**Forward-Logit**	**ReliefF**	**Step Disc**
Binary Logistic Regression	74.56	74.56	73.68	70.18	74.56
BVM	73.68	78.07	71.05	77.19	73.68
C4.5	67.54	68.42	67.54	68.42	67.54
C-PLS	67.54	71.05	66.67	62.28	67.54
C-RT	66.67	64.91	66.67	60.53	66.67
CS-CRT	66.67	64.91	66.67	60.53	66.67
CS-MC4	67.54	64.91	67.54	71.05	67.54
C-SVC	70.18	70.18	71.05	67.54	70.18
CVM	72.81	78.07	72.81	78.07	72.81
ID3	55.26	55.26	55.26	55.26	55.26

KNN	71.05	66.67	65.79	78.07	71.05
LDA	67.54	71.05	70.18	67.54	67.54
Log-Reg TRIRLS	74.56	74.56	73.68	71.93	74.56
MLP	75.44	74.56	69.3	70.18	72.81
MLR	74.56	74.56	73.68	70.18	74.56
NB Continuous	64.91	66.67	67.54	63.16	64.91
PLS-DA	66.67	70.18	70.18	70.18	66.67
PLS-LDA	66.67	70.18	70.18	70.18	66.67
RT	73.68	69.3	71.05	63.16	73.68
SVM	60.53	59.65	54.39	59.65	60.53

The 3-fold cross validation technique produce an accuracy of 78.07% for the top 5 features using the fisher filtering and relief attributes.

Table 6 Accuracy of classification models By considering top 5 features using jack-knife cross validation

Accuracy % using Jack-knife cross validation (Top 5 features selected)					
Algorithm	**Backward-Logit**	**Fisher Filtering**	**Forward-Logit**	**ReliefF**	**Step Disc**
Binary Logistic Regression	77.59	73.28	76.72	70.69	77.59
BVM	79.31	75	77.59	81.9	79.31
C4.5	77.59	74.14	76.72	73.28	77.59
C-PLS	70.69	72.41	68.97	61.21	70.69
C-RT	70.69	67.24	64.66	68.97	70.69
CS-CRT	70.69	67.24	64.66	68.97	70.69
CS-MC4	77.59	70.69	76.72	76.72	77.59
C-SVC	75	72.41	69.83	66.38	75
CVM	80.17	77.59	77.59	81.9	80.17
ID3	55.17	55.17	55.17	55.17	55.17
KNN	75.86	71.55	74.14	76.72	75.86
LDA	68.1	73.28	68.97	68.1	68.1
Log-Reg TRIRLS	77.59	73.28	77.62	69.83	77.59
MLP	75.86	76.72	76.72	73.28	74.14
MLR	77.59	73.28	76.72	70.69	77.59
NB Continuous	72.41	70.69	71.55	64.66	72.41
PLS-DA	68.1	67.24	74.14	64.66	68.1
PLS-LDA	68.1	67.24	74.14	64.66	68.1
RT	75	67.24	73.28	68.1	75
SVM	68.1	72.41	65.52	57.76	68.1

From the Table 6, it can be observed that the classification model Core Vector machine had raised the accuracy percentage to 81.9%. This gives a significant raise in the result by considering the features Age, Glucose, Resistin, Insulin and HOMA (ReliefF Feature selection algorithm top 5 features).

Still an exhaustive search is done to find out if still more accuracy can be increased.

Experiment 4: Accuracy of Classification model based on based on the top 3 features

From the previous findings, it is inferred that feature relevance shows its profoundness. So, the next focus is to reduce the attribute subset size to half. The feature selection algorithms Backward Logit, Forward Logit and Step Disc all project out the same subset i.e. Glucose, BMI and Resistin attributes. The classification accuracy is worked out for the three feature selection algorithms Backward Logit, Fisher Filtering and ReliefF for the two model evaluation techniques. The results are tabulated in Table 7 and Table 8.

Table 7 Accuracy of classification models By considering top 3 features using 3-Fold cross validation

Accuracy % using 3 Fold cross validation (Top 3 features selected)			
Algorithm	**Backward-Logit**	**Fisher Filtering**	**ReliefF**
Binary Logistic Regression	73.68	67.54	70.18
BVM	77.19	68.42	78.95

C4.5	72.81	67.54	66.67
C-PLS	66.67	65.79	60.53
C-RT	68.42	64.91	66.67
CS-CRT	68.42	64.91	66.67
CS-MC4	64.91	67.54	71.05
C-SVC	65.79	65.79	66.67
CVM	78.95	71.05	78.07
ID3	55.26	55.26	55.26
KNN	70.18	71.05	75.44
LDA	71.93	69.3	65.79
Log-Reg TRIRLS	73.68	68.42	70.18
MLP	77.19	70.18	70.18
MLR	73.68	67.54	70.18
NB Continuous	72.81	66.67	68.42
PLS-DA	72.81	65.79	67.54
PLS-LDA	72.81	65.79	67.54
RT	69.3	63.16	65.79
SVM	52.63	60.53	56.14

The Core Vector Machine (CVM) classification model classifies with an accuracy of 78.95%, which is the highest for the 3-fold cross validation technique.

Table 8 Accuracy of classification models By considering top 3 features using jack-knife cross validation

Accuracy % using Jack-knife cross validation (Top 3 features selected)			
Algorithm	**Backward -Logit**	**Fisher Filtering**	**ReliefF**
Binary Logistic Regression	72.41	68.97	71.55
BVM	77.59	67.24	81.03
C4.5	74.14	59.48	76.72
C-PLS	71.55	62.93	64.66
C-RT	66.38	71.55	62.07
CS-CRT	66.38	71.55	62.07
CS-MC4	72.41	59.48	77.59
C-SVC	72.41	66.38	66.38
CVM	**79.31**	**70.69**	**82.76**
ID3	55.17	55.17	55.17
KNN	71.55	63.79	75
LDA	72.41	68.1	67.24
Log-Reg TRIRLS	72.41	68.97	71.55
MLP	71.55	65.52	76.72
MLR	72.41	68.97	71.55
NB Continuous	73.28	67.24	70.69
PLS-DA	75	66.38	71.55
PLS-LDA	75	66.38	71.55
RT	71.55	55.17	71.55
SVM	66.38	54.31	68.1

An accuracy difference of 0.86 is achieved from the ReliefF attributes for the classification model Core Vector Machine using Jack-knife validation technique. Careful findings of the results deduce that Glucose and Resistin helps in achieving a greater accuracy when combined with Age attribute. Age, Glucose and Resistin act as a good predictors for predicting the breast cancer dataset with an accuracy of 82.76% using jack-knife validation technique.

The graphical representation is given for the top five classification model accuracy using the model evaluation technique jack-knife cross validation in Fig 2.

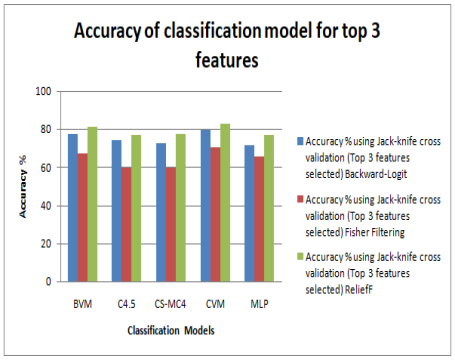

Fig 2: Bar chart representation of accuracy percentage for five algorithms using top 3 features

Still an through analysis for all permutations of the three attributes can be done for an accuracy rise. This is carried out as the last phase of the experiments.

Experiment 5: Accuracy of Classification model based on paired and individual top 3 features for the best 5 selected classification models

The best five algorithms that give a higher accuracy are chosen for each evaluation techniques and the permutations of attributes are considered.

The backward logit algorithm attributes are Glucose, BMI and Resistin. The classification model accuracy are computed by considering the paired combination and individual attribute. The accuracy of 75.44% is obtained through CVM for the 3-fold Cross validation technique using the Glucose and Resistin attribute combination.

And for the Fisher Filtering feature selction algorithm, the attributes selected are Glucose, HOMA and Insulin. 3-fold technique for the Glucose and Insulin combination yielded an highest accuracy results of 71.93% among all feature combination.

Similarly the ReliefF features out Age, Glucose and Resistin. The CVM accuracy is 75.86% resulted for {Age, Glucose} combination and also the model CS-MC4 resulted with the same accuracy for {Age, Resistin} attribute combination and both uses jack-knife validation.

Similarly individual feature classification accuracy results is carried out for the top 3 features selected by the five feature selection algorithm. Glucose, BMI, HOMA, Insulin, Age and Resistin are the selected attributes. All these single attributes are computed for the best five classification models and the results are checked for better accuracy. The accuracy percentage resulted in the range of 46.49% to 72.41%. These results helps in

concluding, the three feature combination produced a better result than paired or individual feature computational results.

This through going of analysis emphasis that the three attributes Age, Glucose and Resistin act as a dependable predictor for the breast cancer disease diagnostics.

To summarize, the ReliefF feature selection algorithm featured out Age, Glucose and Resistin attributes. The supervised algorithms are worked and identified the Core vector machine algorithm using the Jack-knife cross validation evaluation technique yielded and accuracy of 82.76%. The Recall measure for Healthy Controls is 76.92% and for the Patients is 87.50%.

VII. CONCLUSION

In this paper we have considered the UCI machine learning repository dataset, COIMBRA for breast cancer which uses the anthropometric data. Using this dataset we can able to identify healthy versus breast cancer diseased data using the normal blood sample analysis. Feature relevance analysis is performed to find the significant features. A gradual decrement in the features is considered for the classification model accuracy for the Jack-knife and 3-Fold cross validation technique.

According to our findings, the ReliefF feature selection projected out with higher accuracy percentage for the Core Vector Machine classifier. The classification accuracy of 82.76% is achieved by considering Age, Glucose and Resistin attributes with sensitivity of 76.92% and specificity of 87.50%. When feature selection algorithm is considered there is an approximately 6% rise in the accuracy percentage compared to when all attributes are considered. This makes the medical specialists to classify any new breast cancer patient whether they are healthy or affected with breast cancer based on the routine blood sample report, which makes it less complicated than mammographic analysis. Therefore the Age, Glucose and Resistin act as an effective biomarker for Breast cancer Prediction.

REFERENCES

1. Anne-Laure Boulesteix, Korbinian Strimmer; Partial least squares: a versatile tool for the analysis of high-dimensional genomic data, Briefings in Bioinformatics, Volume 8, Issue 1, 1 January 2007, Pages 32–44, https://doi.org/10.1093/bib/bbl016
2. A. LG and E. AT, “Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence,” J. Heal. Med. Informatics, vol. 04, no. 02, pp. 2–4, 2013.
3. Binary Logistic Regression, https://en.wikipedia.org/wiki/Logistic_regression
4. B. Venkata Ramana, M. S. P. Babu, and N. . Venkateswarlu, “A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis,” Int. J. Database Manag. Syst., vol. 3, no. 2, pp. 101–114, 2011.
5. Frontline Solvers, "Classification Tree" ,https://www.solver.com/classification-tree
6. I. Epstein, “Clinical Data-Mining: Integrating Practice and Research,”
7. I. W. H. Tsang, J. T. Y. Kwok, and J. M. Zurada, “Generalized core vector machines,” IEEE Trans. Neural Networks, vol. 17, no. 5, pp. 1126–1140, 2006
8. J. Iavindrasana, G. Cohen, A. Depeursinge, H. Müller, R. Meyer, and A. Geissbuhler, “Clinical Data Mining: a Review,” Yearbook of Medical Informatics, vol. 18, no. 01. pp. 121–133, 2009.
9. J. Starkweather and A. K. Moske, “Multinomial Logistic Regression (lecture notes),” vol. 51, no. 6, pp. 404–410, 2011.
10. Linear Discriminant Analysis for Machine Learning, https://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning/
11. M. J. Zheng et al., "Large-Scale Data Classification Based on Ball Vector Machine", Applied Mechanics and Materials, Vol. 312, pp. 771-776, 2013.
12. M. Kuhn and K. Johnson, “An Introduction to Feature Selection,” Applied Predictive Modeling. pp. 487–519, 2013.
13. M. Patrício, J. Pereira, J. Crisóstomo, P. Matafome, M. Gomes, R. Seiça, and F. Caramelo, “Using Resistin, glucose, age and BMI to predict the presence of breast cancer,” BMC Cancer, vol. 18, no. 1, 2018, pp. 1–8.
14. O. F.Y, A. J.E.T, A. O, H. J. O, O. O, and A. J, “Supervised Machine Learning Algorithms: Classification and Comparison,” Int. J. Comput. Trends Technol., vol. 48, no. 3, pp. 128–138, 2017.
15. S. G. Jacob and R. G. Ramani, “Efficient Classifier for Classification of Prognostic Breast Cancer Data through Data Mining Techniques,” Proceedings of the World Congress on Engineering and Computer Science 2012 Vol I WCECS 2012, October 24-26, 2012, San Francisco, USA
16. S. G. Jacob and R. G. Ramani, “Discovery of Knowledge Patterns in Clinical Data through Data Mining Algorithms: Multiclass Categorization of Breast Tissue Data,” Int. J. Comput. Appl. Vol., vol. 32, no. 7, pp. 46–53, 2011.
17. V. Chaurasia, S. Pal, and B. B. Tiwari, “Prediction of benign and malignant breast cancer using data mining techniques,” J. Algorithms Comput. Technol., vol. 12, no. 2, pp. 119–126, 2018.
18. Y. Li and Z. Chen, “Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction,” vol. 7, no. 4, 2018, pp. 212–216.
19. “Current scenario of Breast Cancer In Rural India |Scientific India Magazine.”
20. “R- Documentation - Classification with PLS dimension Reduction and Linear Discriminant Analysispls.”
21. "Scientific India - The Changing Face of Breast Cancer - http://www.scind.org/1099/Health/the-changing-face-of-breast-cancer-in-india.html", 2018
22. Statistics Solutions, “Binary Logistic Regression - Statistics Solutions.” 2018.
23. “Tanagra - Data Mining and Data Science Tutorials, http://data-mining-tutorials.blogspot.com/” .
24. “Trends of Breast Cancer in India, http://www.breastcancerindia.net/statistics/trends.html.” .
25. Y. Ahuja and S. Kumar Yadav, “Multiclass Classification and Support Vector Machine,” Glob. J. Comput. Sci. Technol. Interdiscip., vol. 12, no. 11, pp. 14–19, 2012.