Classification of Recurrence of Breast Cancer Cells Using Machine Learning

Ammara Zamir✉ and Rida Zahra

Department of Computer Science, University of Wah, Wah Cantt., Pakistan

ABSTRACT

Breast cancer is the most common cancer spreading in woman of developed and developing countries. Statistics shows the breast cancer cause so many deaths every year. Symptoms of breast cancer is lump and thickening of tissues of breast. There are many techniques including supervised and unsupervised learning methods used in medical science for prediction of breast cancer. Supervised learning methods are more popular and also used to find the type of cancer cells. They are also used for prediction of recurrence rate of cancer cells and the survival rate of woman diagnosed with breast cancer. This research study presents a comparison of machine learning (ML) classifiers: Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB) and Artificial Neural Network (ANN) with use of popular feature selection techniques including: Information Gain (IG), Gain Ratio (GR), Relief-F and Gini-index. The experimental results show that ANN outperforms all other classifiers with 99.6 % accuracy.

Keywords: *Neural Network, Recurrence cells, Breast Cancer, Random Forest*

1. INTRODUCTION

Breast cancer is the second major cause of deaths in world [1]. About one woman among 8 is diagnosed with breast cancer. Lumps or compact breast tissues is the symptom of breast cancer. The cause of breast cancer in woman can be due to family history, increased aged, obesity and excessive use of alcohol [2]. In developing countries, 70% deaths caused by breast cancer. Continuous research studies have been made to evaluate the techniques to predict the breast cancer [3].

Non-invasive breast cancer does not have ability to spread outside of the breast. It seldom shows by a breast lump. While invasive breast cancer is most common type of breast cancer. It has ability to spread outside of the breast [4]. Screening method is used to detect the type of cancers before they give rise to symptoms [5].

To predict the cancer disease, symptoms and its causes is more challenging. Arrival of new technologies in field of medical science lead towards the collection of cancer data for further experiments. Machine Learning (ML) techniques are widely used in field of medical science. ML is used to find the patterns and relations of cancer cells and use to find out the

✉ ammara.zameer@uow.edu.pk

type of cancer [6-8].

Usually, clinical and genomic data is integrated for experiments in earlier studies. The testing and validation process seems weak on integrated collection of dataset. Therefore, due to accuracy of ML methods, these techniques are used to predict the cancerous cells in hazard of recurrence [2]. The ML methods include following steps: collection of data, selection of data, training and testing of data and classification. ML methods have improved the classification accuracy of cancerous cells up to 15%-20% [9].

Many studies are carried out to predict the miRNA [10] which is the most prominent class of cancer. But, these studies based on gene expression lack in prediction accuracy due to the sensitivity issues. Multiple kernel learning method [9], Random Forest [11, 12], SVM [13], bagging with NN [14] are used to predict the susceptible cancer cells, balance the imbalance clinical data and predict the type of cancer cells. The performance of each algorithm depends on the configuration settings of different classifiers.

The aim of this research work is to find out the best suited machine learning algorithm for prediction of breast cancer and the effect of feature reduction/ selection methods on the performance of classifiers.

The rest of the paper is organized as follows: Section 2 reviews the earlier studies. Section 3 shares the methodology. Section 4 discusses the results. Section 5 presents conclusions.

2. RELATED WORK

Existing ML approaches related to prediction of breast cancer and data reduction techniques to improve the accuracy of classifiers are discussed in the following:

There are two types of ML techniques used in medical science field for prediction including supervised learning and un-supervised learning approaches. In supervised learning labeled data is used and in un-supervised learning approaches un labeled data is used as input for classifiers. Semi-supervised learning method is also used for prediction in which supervised and un-supervised approaches are combined.

Dataset is the basic component of this prediction process. It consists of different attributes and its values. Selection of good dataset with improved quality increases the accuracy of classifiers [6].

Feature selection approaches are also used for selection of important features from data set. When there is a large number of dataset features are ranked and reduced by feature selection approaches. By using this approach unnecessary features are eliminated from given dataset [15]. Noise reduction is the main advantage of feature selection algorithms. Filter and

wrapper are the main feature selection methods. Prediction accuracy of classifiers based on the configuration settings among them most common settings include (i) k-fold cross validation and (ii) hold-out settings. After pre-processing and feature selection methods machine learning classifiers are applied to classify the attributes from diverse data set.

ML techniques are widely used for patterning the progress of breast cancer [16]. ML techniques are used to find the breast cancer cells' recurrence, susceptibility and survival. Major ML techniques include Decision Tree (DT) and Artificial Neural Network (ANN). Support Vector Machine (SVM), DT, and Neural Network are commonly used for prediction. ANN uses hidden layers that process on the input to get the best classification results [17]. Whereas, DT [18] comes with the structure of nodes. Input is represented by nodes and result is determined on leaves. DT's are simple and quick to learn. SVM is most commonly technique used for prediction of breast cancer. SVM deals with input using hyper plan with high dimensionality. Hyper plan splits input into two classes. Outputs based on probability can be obtained using SVM [19].

These ML techniques are also used to find out the survival rate of woman diagnosed with breast cancer. ANN, SVM and SSL used on SEER dataset to predict the survival rate . Multicellular neural network for prediction of breast cancer is applied on MIAS [20] dataset and results are evaluated accuracy, sensitivity, and specificity. CAD is used to predict and diagnose the breast cancer using ultrasound dataset [21].

3. METHODOLOGY

In this section, different machine learning algorithms, feature selection techniques, selected dataset along with performance metrics are described. After data preprocessing, feature selection algorithms are applied to select the important feature set. Afterwards, ML techniques are applied and results of classifiers are evaluated using performance evaluation measures.

A. Dataset

Breast cancer dataset is available freely for research . This dataset has 286 instances and 10 attributes. Some attributes are linear and some are non-linear attributes. The dataset contains data of people age from 10-90 years. The number of continuous values is zero and discrete values is ten. However, nine values are missing in the dataset. A number of research publications have used this dataset [22] [23-25].

B. Feature Selection Methods

For feature selection the most popular techniques are used to reduce the instances from selected dataset. Following applied feature selection methods are:

a) Information Gain (IG)

Information Gain gives information about an attribute. IG measure the reduction in Entropy. IG is good where number of instances are small. IG shows biasness towards large number of attributes. IG is used along with machine learning techniques to improve the accuracy of classifiers [26].

b) Gain Ratio (GR)

GR is another feature selection technique which is extension of IG. GR is used to decrease the biasness of IG towards multivariant attributes. GR performs better on large number of instances of dataset [27].

c) Relief-F

Relief-F is a feature selection method used to apply on binary classification problem. Relief compute the score of each attribute and then rank and select the top scoring attributes. Relief-F is widely used with machine learning algorithms to increase the classification accuracy [28].

d) Gini-index

Gini coefficient also known as Gini-Index used to measure the inequality among frequency distribution. Gini-index is favorable for the large partitions and simple to implement [29]. Gini-index is used to evaluate the quality of each split on dataset of cancer [29].

C. ML Algorithms

Different machine learning approaches are applied to predict the recurrence rate of cancer cell. These ML classifiers include discriminative, generative, Ensemble learner, and Neural Networks. Following algorithms are applied on selected dataset.

a) Support Vector Machine (SVM)

SVM is supervised machine learning technique which divides the dataset into hyper-plane and predict the class of given dataset. SVM creates many hyper-planes and perform margin maximization to bridge the gap between different classes. SVM is widely used in medical field to predict the disease [2, 6].

b) Naïve Bayes (NB)

NB classifier is another popular approach in classification techniques. NB makes a decision using simple framework. Variables used in NB model are independent of each other. NB is the probability based classifier used to predict the class [2].

c) Random Forest (RF)

RF is an ensemble learner and recursive technique. RF randomly pick a sample from dataset and replace it with another sample. This step is repeated for every iteration until data is divided. Unnecessary data is eliminated and trees are repeated again. In the end, decision is

made on majority votes [2, 12].

d) k-Nearest Neighbor (kNN)

kNN is the classification technique and a lazy leaner. kNN do not assume on the basis of implicit data. Besides, kNN can be used for regression. In classification, each test point has k nearest training data points. Data points are divided into classes. Most occurring class assigned as test data [2, 30].

e) Artificial Neural Network (ANN)

ANN is machine learning classifier. ANN working is similar to Human brain. ANN use Activation function, hidden layers and iterations to predict the class. ANN has four layers in its architecture. Each Layer has weighted connector. ANN use backward propagation to predict the class [2, 31].

D. Performance Evaluation Metrics (PEM)

After applying the machine learning methods, the classifiers accuracy is evaluated using PEM [32] to check which algorithm is best to predict cancer recurrence shown in Eq. (1), Eq. 2, Eq. 3, Eq. 4.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F-Measure} = 2 * \text{Precision}\ \frac{\text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

4. EXPERIMENTAL RESULTS

The performance of classifiers with and without feature selection techniques on selected dataset of breast cancer is discussed in this section. Experiments are carried out on selected data set using standard tool Rapid Miner [33, 34]. Where default settings of all classifiers are used to evaluate the classification accuracy except ANN. While, for Artificial Neural Network, the following settings are used; iter=400, hidden layers=300, activation='tanh', learning rate ='adaptive'.

After preprocessing experiments are performed on the selected data set including all 10 attributes. ANN performed well than any other classifiers in terms of PEM evident from results shown in Table 1. ANN predicts the recurrence rate up to 89.2% when applied on

all attributes of data set shown in Table 1.

Table 1. Performance of machine learning classifiers on all attributes

Classifier	ACC	F-Measure	Precision	Recall
SVM	0.801	0.781	0.824	0.808
kNN	0.772	0.743	0.754	0.766
NB	0.749	0.734	0.732	0.738
RF	0.870	0.813	0.824	0.825
ANN	**0.892**	**0.836**	**0.848**	**0.846**

Afterwards, feature selection algorithms are applied to select the important features of dataset. Five top attributes of dataset are selected by feature selection techniques shown in Table 2. Feature selection techniques ranked the same attributes but with different values as shown in Table 2. While, IG have ranked higher than GR and Relief F feature selection methods. IG performs well on small number of attributes as compared to large number of attributes as in case of the selected data set IG performed better than other feature selection techniques shown in Table 2.

Table 2. Top ranked attributes by feature selection techniques

Attributes	IG	GR	Gini	Relief-F
dea-malia	0.077	0.050	0.046	0.033
inv-nodes	0.069	0.052	0.042	0.006
Tumor-size	0.057	0.019	0.026	0.028
Node-caps	0.053	0.073	0.033	0.031
irradiant	0.026	0.033	0.016	0.014

Feature selection is performed to find out the impact on ML classifiers. Experiments are performed on selected attributes of dataset. Where Random forest an ensemble leaner and Artificial Neural Network performed far better than any other classifiers to predict the recurrence rate of breast cancer cells.

Artificial Neural Network outperformed all other classifiers in terms of classification accuracy, F-measure, precision and recall evident from results shown in Table 3. ANN classified attributes accurately up to 99.6%.

Table 3. Performance of machine learning classifiers on top 5 attributes

Classifier	ACC	F-Measure	Precision	Recall
SVM	0.919	0.853	0.978	0.518
kNN	0.782	0.759	0.700	0.329
NB	0.767	0.566	0.580	0.553
RF	0.984	0.927	0.971	0.776
ANN	**0.996**	**0.986**	**0.987**	**0.906**

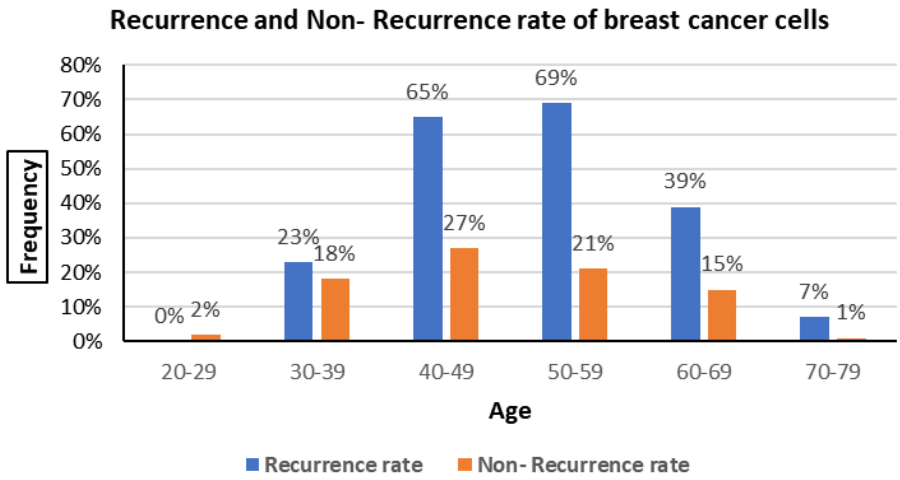

Fig. 1. Recurrence and Non- Recurrence rate of cancer cells among different age group

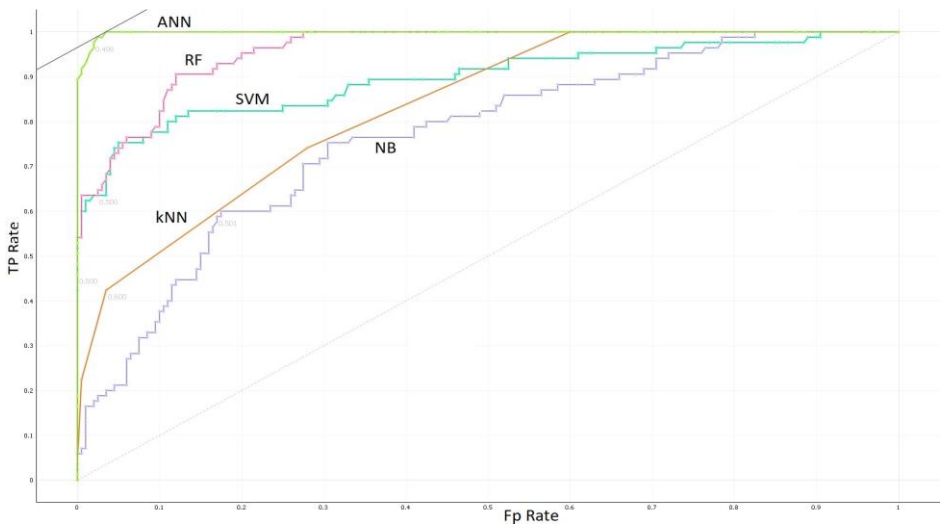

Fig. 2. ROC curve of ml classifiers

Artificial Neural Network outperformed all other classifiers in terms of classification accuracy, F-measure, precision and recall evident from results shown in Table 3. ANN classified attributes accurately up to 99.6%. The results show that feature selection techniques increased the accuracy of applied classifiers to predict the recurrence rate of cancer cells shown in Table 3.

Recurrence events are more prominent among age group of 50-59 years. And the recurrence rate is less among age group of 70-90 years. While recurrence rate of breast cancer cells is zero among age group of less than 30 years.

Recurrence and non-recurrence rate of cancer cell among different age of people is shown in Fig 1. Performance of all applied classifiers are measured by ROC curve shown in Fig 2. ROC shows the ability of classifiers up to which classifiers can classify the attributes correctly. ANN beats all classifiers by showing highest value of AUC= 99.6.

5. CONCLUSIONS

Breast cancer is the second most common type of cancer spreads among woman worldwide. The main focus of this study is to evaluate the importance of feature selection techniques and prediction accuracy of classifiers. Evident from experimental results that Information Gain (IG) remains best feature selection technique. ANN outperformed all other algorithms when applied on all attributes of dataset with 89.2% accuracy. While, when applied on the selected attributes of dataset ANN outperformed other classifies with 99.6% accuracy. So, Evident from the results ANN is most effective approach on selected dataset with feature selection technique IG to classify recurrence of breast cancer cells up to 99.6% accuracy.

REFFERENCES

[1] F. Bray and J. Ferlay, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA: a cancer journal for clinicians*,(2018), vol. 68, pp. 394-424.

[2] M. Amrane and S. Oukid, Breast cancer classification using machine learning, *in 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*,(2018), pp. 1-4.

[3] D. HanahanR.A. Weinberg, Hallmarks of cancer: the next generation, *cell*,(2011), vol. 144, pp. 646-674.

[4] J.C. GoochF. Schnabel, "Inflammatory Breast Cancer," in *Clinical Algorithms in General Surgery*, ed: Springer, 2019, pp. 105-108.

[5] M. Phillips and R.N. Cataneo, Prediction of breast cancer risk with volatile biomarkers in breath, *Breast cancer research and treatment*,(2018), vol. 170, pp. 343-350.

[6] K. Kourou and T.P. Exarchos, Machine learning applications in cancer prognosis and prediction, *Computational and structural biotechnology journal*,(2015), vol. 13, pp. 8-17.

[7] M.N.Q. Bhuiyan and M. Shamsujjoha, "Transfer Learning and Supervised Classifier Based Prediction Model for Breast Cancer," in *Big Data Analytics for Intelligent Healthcare Management*, ed: Elsevier, 2019, pp. 59-86.

[8] B.-J. KimS.-H. Kim, Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method, *Proceedings of the National Academy of Sciences*,(2018), vol. 115, pp. 1322-1327.

[9] D. Sun and A. Li, Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome, *Computer methods and programs in biomedicine*,(2018), vol. 161, pp. 45-53.

[10] X. Zhang and Q. Zou, Meta-path methods for prioritizing candidate disease miRNAs, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*,(2019), vol. 16, pp. 283-291.

[11] S. RajasegararA.S. Abdalrada, Breast Cancer Recurrence Prediction Using Random Forest Model, *in Recent Advances on Soft Computing and Data Mining: Proceedings of the Third International Conference on Soft Computing and Data Mining (SCDM 2018), Johor, Malaysia, February 06-07, 2018*,(2018), p. 318.

[12] X. Zhu and X. Du, Random forest based classification of alcohol dependence patients and healthy controls using resting state MRI, *Neuroscience letters*,(2018), vol. 676, pp. 27-33.

[13] H. Wang and B. Zheng, A support vector machine-based ensemble algorithm for breast cancer diagnosis, *European Journal of Operational Research*,(2018), vol. 267, pp. 687-699.

[14] I. Fakhruzi, An Artificial Neural Network with Bagging to Address Imbalance Datasets on Clinical Prediction, *Diabetes*,(2018), vol. 768, p. 2.
[15] G. ManikandanS. Abirami, "A Survey on Feature Selection and Extraction Techniques for High-Dimensional Microarray Datasets," in *Knowledge Computing and its Applications*, ed: Springer, 2018, pp. 311-333.
[16] S. Khan and N. Islam, A novel deep learning based framework for the detection and classification of breast cancer using transfer learning, *Pattern Recognition Letters*,(2019), vol. 125, pp. 1-6.
[17] T. Ayer and O. Alagoz, Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration, *Cancer*,(2010), vol. 116, pp. 3310-3321.
[18] S. Turgut and M. Dağtekin, Microarray breast cancer data classification using machine learning methods, *in 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*,(2018), pp. 1-3.
[19] L. Hussain and W. Aziz, Automated Breast Cancer Detection Using Machine Learning Techniques by Extracting Different Feature Extracting Strategies, *in 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*,(2018), pp. 327-331.
[20] L. Civcik and B. Yilmaz, Detection of microcalcification in digitized mammograms with multistable cellular neural networks using a new image enhancement method: automated lesion intensity enhancer (ALIE), *Turkish Journal of Electrical Engineering & Computer Sciences*,(2015), vol. 23, pp. 853-872.
[21] A. Jalalian and S.B. Mashohor, Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review, *Clinical imaging*,(2013), vol. 37, pp. 420-426.
[22] R.S. Michalski and I. Mozetic, The multi-purpose incremental learning system AQ15 and its testing application to three medical domains, *Proc AAAI 1986*,(1986), pp. 1,041-1,045.
[23] T. Niblett, Constructing decision trees in noisy domains, *in Proceedings of the 2nd European Conference on European Working Session on Learning*,(1987), pp. 67-78.
[24] M. TanL. Eshelman, "Using weighted networks to represent classification knowledge in noisy domains," in *Machine Learning Proceedings 1988*, ed: Elsevier, 1988, pp. 121-134.
[25] B. Cestnik, Assistant 86: A Knowledge-Elicitation Tool for Sophisticated Users, *Progress in Machine Learning*,(1987), vol. 62.
[26] R.B. Pereira and A. Plastino, Categorizing feature selection methods for multi-label classification, *Artificial Intelligence Review*,(2018), vol. 49, pp. 57-78.
[27] D.H. MazumderR. Veilumuthu, An enhanced feature selection filter for classification of microarray cancer data, *ETRI Journal*,(2019), vol. 41, pp. 358-370.
[28] K. YanH. Lu, Evaluating ensemble learning impact on gene selection for automated cancer diagnosis, *in International Workshop on Health Intelligence*,(2019), pp. 183-186.
[29] Y.-J. Tseng and C.-E. Huang, Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies, *International journal of medical informatics*,(2019), vol. 128, pp. 79-86.
[30] W. Yue and Z. Wang, Machine learning with applications in breast cancer diagnosis and prognosis, *Designs*,(2018), vol. 2, p. 13.
[31] Y. Xiao and J. Wu, A deep learning-based multi-model ensemble method for cancer prediction, *Computer methods and programs in biomedicine*,(2018), vol. 153, pp. 1-9.
[32] H.U. Khan, Mixed-sentiment classification of web forum posts using lexical and non-lexical features, *J Web Eng*,(2017), vol. 16, pp. 161-176.
[33] A. Viloria and J.R. López, Determinating Student Interactions in a Virtual Learning Environment Using Data Mining, *Procedia Computer Science*,(2019), vol. 155, pp. 587-592.
[34] B. McCullough and T. Mokfi, On the accuracy of linear regression routines in some data mining packages, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*,(2019), vol. 9, p. e1279.