J. Sci. Technol. Environ. Inform. 09(02): 665-672 | Ahmed et al. (2020)
EISSN: 2409-7632, Journal home: www.journalbinet.com
Crossref: https://doi.org/10.18801/jstei.090220.67

JSTEI ONLINE JOURNAL
Journal of Science, Technology & Environment Informatics
Journal BiNET Scientific Publication

Analysis of Wisconsin Breast Cancer original dataset using data mining and machine learning algorithms for breast cancer prediction

Md. Toukir Ahmed, Md. Niaz Imtiaz and Animesh Karmakar

Dept. of Computer Science and Engineering, Pabna University of Science and Technology, Bangladesh

Article received: 03.07.2020; Revised: 22.07.2020; First published online: 30 July 2020

Article Information

Key Words:
Classification, Decision tree, MLP, WDBC, Naïve bayes and SVM

Access by Smart Phone

For any information:
toukirahmedreal@gmail.com

ABSTRACT

Breast cancer has become a concerning issue in recent years. The rate of women having breast cancer seemed to be increased significantly. The disease has become life-taking if it is not diagnosed at all and in many cases, separation of limbs is the only way to prevent it, if it is diagnosed at the last stage. As a result, a good predictor of this issue can be fruitful in successful diagnosis. The main focus of this paper is to perform different machine learning classification algorithms to correctly predict the target class and improve it by checking the effectiveness of particular attributes of original Wisconsin Breast Cancer dataset (WDBC) for breast cancer diagnosis prediction. After running classifiers on the dataset, the comparison was made among them to find the best performing algorithm and then effective attributes of dataset were analyzed to improve performance further. In this paper, we have used algorithms- Naïve Bayes, Support Vector Machine (SVM), Multilayer Perceptron (MLP), J48 and Random Forest. Here, for comparing the result, we have used performance metrics: Accuracy, Kappa statistic, precision, recall, F-measure, MCC, ROC area, PRC area. Based on the values of performance metrics, Naïve Bayes classifier gave the best result among the algorithms used. Moreover, we also tried to optimize our proposed model and made a comparison among state-of-the-art approaches proposed by different researchers, on the same dataset.

Citation: Ahmed, M. T., Imtiaz, M. N. and Karmakar, A. (2020). Analysis of Wisconsin Breast cancer original dataset using data mining and machine learning algorithms for breast cancer prediction. Journal of Science, Technology and Environment Informatics, 09(02), 665-672.
Crossref: https://doi.org/10.18801/jstei.090220.67

I. Introduction

The second leading disease causing of women's death is breast cancer and about 2.6% women die due to the disease when they are affected. This rate of death decreasing day by day because of finding latent causes of breast cancer, earlier diagnosis, increased awareness. The health and medical sector have more in need of data mining which helps medical practitioners to extract valuable information from large database very useful to take decision, improve health services, prediction about situation means

in which stage the disease is that helps to provide medication dose. For breast cancer data mining can act very effective avoidance, indication base medication, rectifying hospital data errors. Breast cancer is a very common and second leading cause of death innumerous developed and developing country. Cause of death in breast cancer is acute in US and Asia. Among Asian countries, breast cancer is very common in Pakistan, with approximately 90,000 new breast cancer disease diagnosed in Pakistan (Mushtaq et al., 2019). Data mining and machine learning strategies have been implied in many sectors of medicine health care systems (Hung et al., 2018). We can get a very useful result by using these technologies, by generating the impact of some cause of breast cancer by analyzing data and discovering secret patterns among data. Data mining are solving many problems and supporting medicine analyst and doctors. WEKA (Waikato Environment for Knowledge Analysis) (Hall et al., 2009) is a powerful tool as it contains supervised learning as well as unsupervised learning methods. It contains classification, clustering, Association, Mining, Feature selection, Data visualization etc (Hall et al., 2009). WEKA is a very useful data mining tool that helps researchers to implement classification, analyze data, visualize data, comparing classification techniques to get easily a better performing algorithm using various parameters. Breast cancer is a malignant and benign tumor, inside breast wherein cell divide and grow without control (Shah and Jivani, 2013). Scientists have tried the exact reason behind breast cancer, as there are few risk factors that increase the likelihood of a woman growing breast cancer. Age, genetic risk and family background, obesity, Gene variation, smoking, taking alcohol are such factors being considered for breast cancer. There exists two types of breast cancer treatments named: local and systematical. Local types of treatments are surgery, radiation. Chemotherapy and hormone therapy are treated systematically. Medical practitioners continue both types of treatment collaterally for checking which one impacting best.

There had been numerous research works done on Wisconsin Breast Cancer dataset for prediction of breast cancer. Shah and Jivani, (2013) have used three different classification strategies for predicting Breast Cancer, emphasizing mainly on the greatest accuracy and low computation time and got Naïve Bayes as the best performing algorithm with 95.9943% accuracy and 0.02 seconds computation time. The main setback with their approach is relatively lower accuracy. Bazazeh and Shubair (2016) have made a comparison among the three most popular classification techniques, namely Support Vector Machine (SVM), Random Forest (RF), Bayesian Networks (BN) and claimed that, on this dataset the accuracy of SVM is about 97%. There exists some drawbacks with this approach as SVM doesn't seem to have the property of retraining as when a new point is added to it, it has to train again with several key parameters. Amrane et al. (2018) explained that *k*NN gives the accuracy of 97% with this dataset. A setback is, kNN compares the Euclidean distance with respect to all training points which has a great toll on testing time. Nematzadeh et al. (2015) have stated, neural networks showed the accuracy of 98% on this dataset. Drawback with this approach is it is time consuming too. Priyanka et al. (2019) have used KE's algorithm and the accuracy assessed for the test data is 98.53% which is great but the drawback is the taken split ratio of dataset for testing performance was larger which usually made an overall satisfactory accuracy for this dataset. Singh and Thakral (2018) compared decision tree classifier (J48, Simple CART), Bayes classifier (Naïve Bayes, Bayesian Logistic Regression). Simple CART provided higher accuracy among their proposed classifiers. In their work, they only used two parameters for performance metrics which were accuracy and time complexity, but choosing performance metrics containing only two parameters may not be adequate and secure to compare the classifiers performance. Senturk and Kara, (2014) have used seven different algorithms and to apply data mining with the proposed algorithms Rapid Miner 5.0 data mining tool used. In this paper they diagnosed breast cancer. They have collected the data of the patients who are in trouble with breast cancer. According to accuracy rates, they have gotten SVM as the best performing algorithm.

In this paper, different data classification algorithms are used which is described in section II (b) and the used dataset which discussed in detail in section II (a) is Wisconsin Breast Cancer original dataset. First of all, a superior classification algorithm has found out among used classifiers compared with accuracy and other standard parameters, then the dataset is analyzed to extract the feature of attributes effectiveness and based on that the dataset is modified. After that, the best performing classifier is run on the modified dataset for getting the improved performance result.

II. Materials and Methods

Keeping in mind the shortcomings of earlier works, the experiment was carried out to have improved performance in less computational time. Thus the emergence of WEKA (Hall et al., 2009) tool was felt necessary as it efficiently runs the classification algorithms and gives output with respect to all performance metrics. The experiment was run in the last quarter of 2019 and as different split ratios and validations were tried, we had to re-run it multiple times.

Data Description

The data used in this study involving Breast Cancer data extracted from UCI Machine Learning Repository (Wolberg, 1992). Dr. WIlliam H. Wolberg (physician) University of Wisconsin Hospitals Madison, Wisconsin, USA has been collected the data since 1989 to 1991 (Wolberg, 1992). The dataset is available for everyone for research (Dua, 2017). There are 699 instances of 11 attributes with 19 missing values. The first attribute is for id number and it is unnecessary for research that's why we removed it from dataset. The number 10 attribute represent class value which have two value 2 and 4 where 2 represents benign and 4 represents malignant. Rest of attributes are ranged from 1 to 10. Pathologist assigned these numbers based on their characteristics. Large value represents greater chance of malignancy. The detailed information about WDBC is given below in (Table 01).

Table 01. Attributes information of WDBC

Attributes Name	Value
Clump-thickness	1 – 10
Uniformity of Cell Size	1 – 10
Uniformity of Cell Shape	1 – 10
Marginal Adhesion	1 – 10
Single Epithelial Cell Size	1 – 10
Bare nuclei	1 – 10
Bland chromatin	1 – 10
Normal Nucleoli	1 – 10
Mitoses	1 – 10
Class	2 for benign, 4 for malignant

Used Classifiers

Naïve Bayes: Naïve Bayes is a probabilistic statistical base classifier based upon Bayes' theorem which is strong supervised machine learning classification technique. It assumes that all the features are conditionally independent (Kim et al., 2017) which means the effect of an attribute value has no effect on other attribute value. Naïve Bayes is a very light weight classifier can be used to classify big dataset easily. It is very robust to ignore noise and irrelevant attributes. It is very easy to construct and no need of complicated iterative parameter estimation schemes (Wu, 2008).

SVM: Support vector machine (SVM) is considered as a supervised machine learning classification technique that is built, based on the concept of decision planes that define decision boundaries. This algorithm function by making "hyperplane" and categories the data based on class values, SVM algorithm performs margin maximization which means it tries to make maximum difference between classes (Parkin, 1998). SVM creates complex non-linear boundaries that are robust to over fitting and the major advantage is high classification accuracy.

J48: J48 classifier is a supervised machine learning classification method, simple decision tree algorithm for classification. It creates small binary tree. J48 is an extension of ID3. The feature of this algorithm is accounting for missing values, derivational rules, pruning for decision trees (Kaur and Chhabra, 2014). J48 is an open source algorithm of C4.5 in WEKA data mining tool. The function of J48 is to generate a threshold and divides the data into two groups which are bigger than threshold and which are same and lower than threshold.

Multilayer perceptron (MLP): MLP belongs to one of the classes of neural network and a branch of artificial intelligence (Gardner and Dorling, 1998). It is a kind of acyclic graph and comes under the

category of supervised feed forward networks. MLP has three or more layers of nodes such as input layer, hidden layer and output layer, each node is a neuron which used non-linear activation function except for the input node. It connects multiple layers in a directed graph to establish one way directed signal path through the nodes and each node has a non-linear activation function. It can generalize new unseen data easily and the core feature of multilayer perceptron is it don't make any prior assumption regarding data ordering.

Random Forest: Random forest is a supervised learning method that is a decision tree based algorithm. As the name suggest as forest the random forest classifier is an ensemble of decision trees where a random vector sample produce each classifier from input vector (Pal, 2005) and each tree cast a unit vote for the most popular class to classify an input vector, most of the time trained with a bagging method. The general idea of bagging method is that compose of the learning method increases the overall result. The Random Forest is less sensitive than other streamline machine learning classifiers to over fitting and to the quality of training samples (Belgiu and Drăguț, 2016).

Performance Metrics

Performance parameters are the most important metrics to compare among classifier methods to get the best classifier. We have applied 10 performance parameters (Wu, 2008) which are Accuracy, Sensitivity, Specificity, Kappa statistics, Precision, Recall, F-measure, Matthew's Correlation Coefficient (MCC), Receiver Operating Characteristic (ROC) Area and Precision-Recall Curves (PRC). These parameters are calculated from a confusion matrix which is situated in every step of classification. A general view of confusion matrix is illustrated in Table 02.

Table 02. Confusion Matrix

	Predicted YES	**Predicted NO**
Actual YES	TP	FN
Actual NO	FP	TN

TP represents the number of correctly classified positive instances.
FP represents the number of misclassified positive instances.
FN represents the number of misclassified negative instances.
TN represents the number of correctly classified negative instances.

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

Specificity = $\frac{TN}{TN+FP}$

Sensitivity = $\frac{TP}{TP+FN}$

Precision = $\frac{TP}{TP+FP}$

Recall = $\frac{TP}{TP+FN}$

F-measure = $\frac{2*precision*recall}{precision+recall}$

MCC = measure of quality of binary classification.
ROC Area = Most important parameter in WEKA which defines the classifiers performance in general.
PRC area = this is also an important parameter which more informative, a plot of precision and recall. This is very useful to get the difference between precision and recall.

Implementation

Here, we discuss about our proposed methodology which we apply step by step to get expected result. At first, we analyze Wisconsin Breast Cancer Dataset by using WEKA data mining tool 3.8 (Hall et al., 2009). The tool is very helpful to analyze and have various techniques embedded in it. We identify the

effectiveness of dataset attributes which attribute has the most impact on the result of accuracy with different parameters. After that we identify the performance of proposed classifiers and choose the better performing classifier. We selected best performing classifier based on not only accuracy but also others important parameters which are very useful for checking the performance of a classifier in all side. Current work is attempt to improve performance of previous work done on this dataset and ignoring ineffective attribute using proposed algorithms which make sure the lightweight environment to detect Breast Cancer correctly.

Description of Method Procedures

We implemented the proposed algorithms in WEKA. We used 10-fold cross validation method to investigate the performances of algorithms. For performance measurement, we used accuracy, sensitivity, specificity, precision, Recall, ROC, PRC area. We checked all these parameters to find out the better performing algorithm and came to a point that Naïve Bayes giving the best result among them. After that we analyze the dataset by removing each attribute on Naïve Bayes algorithm to find out the effectiveness of attribute and got to know which attribute is more or less effective and has impact on performance result parameters. At first, we removed each attribute at a time to get the less effective attribute, then we removed two attributes at a time and ran Naïve Bayes on this dataset to know the effect of these two attributes. But we got the best result by removing one attribute "Single Epithelial Cell" Size which has very less effect on result parameters and gets improve result of performance parameters.

III. Results and Discussion

Study of the algorithm's performance

In this section, we firstly demonstrate the obtained result of proposed classifier methods with different parameters. At first, we used total attributes of dataset with 699 instances. We used 10-fold cross validation on the training dataset of WDBC. The obtained results of different parameters are illustrated in Figure 01 and Table 03.

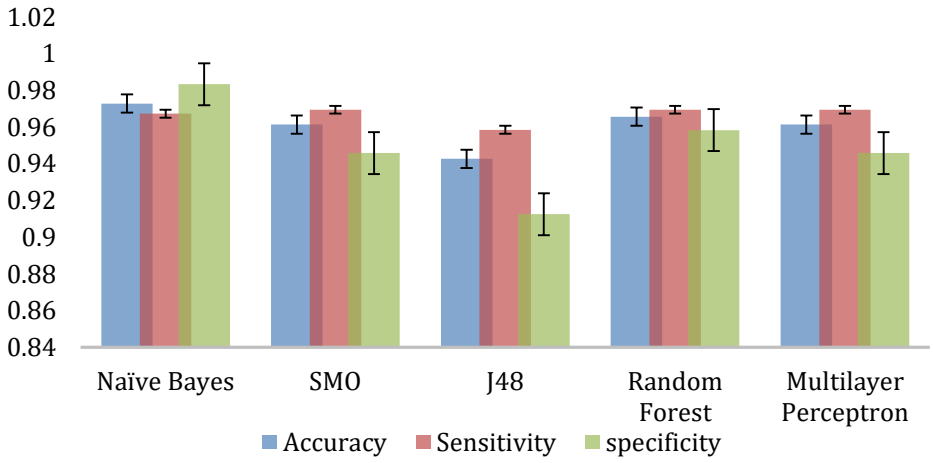

Figure 01. Comparison between classifiers based on accuracy, sensitivity and specificity.

Table 03. Performance results of used Classifiers for 10 attributes (Bold indicates the best result**)**

Classifier Name	Accuracy %	Kappa statistics	Precision	Recall	F-measure	MCC	ROC area	PRC area
Naïve Bayes	**97.2779**	**0.9403**	**0.974**	**0.973**	**0.973**	**0.941**	**0.992**	**0.992**
J48	94.2693	0.8727	0.943	0.943	0.943	0.873	0.965	0.956
Random forest	95.5616	0.9241	0.966	0.966	0.966	0.924	0.988	0.987
SMO	96.1318	0.9144	0.961	0.961	0.961	0.914	0.958	0.944
Multilayer Perceptron	96.1318	0.9144	0.961	0.961	0.961	0.914	0.987	0.986

Figure 01 shows that, Naïve Bayes classifier performs better in terms of Accuracy and Specificity. Besides, it is also noticeable that, Multilayer Perceptron gave highest Sensitivity. But with respect to other performances metrics, Naïve Bayes classifier shows it supremacy that is noticeable in Table 03.

Performance improvement study based on WDBC analyzing and modifying

As Naïve Bayes classifier worked best among our proposed classifiers, we tried to optimize result further. We tried to find the effectiveness of each feature and their effects on the performance.

Table 04. Performance results of Naïve Bayes after removing attribute "Single Epithelial Cell Size" from WDBC

Test Mode	Accuracy %	Kappa statistics	Precision	Recall	F-measure	MCC	ROC area	PRC area
5-fold cross validation	97.4212	0.9435	0.975	0.974	0.974	0.944	0.993	0.993
10-Fold cross validation	97.4212	0.9435	0.975	0.974	0.974	0.944	0.993	0.993
15-Fold cross validation	97.4212	0.9435	0.975	0.974	0.974	0.944	0.993	0.993
Split 66.6% train, remainder test	96.9957	0.935	0.971	0.970	0.970	0.936	0.993	0.993
Split 85.5% train, remainder test	99.0099	0.9788	0.990	0.990	0.990	0.979	0.997	0.997

So, we extracted the features from data by removing the attributes one by one and check out the performance to know the effectiveness of classifier then got to a point that "Single Epithelial Cell Size" have less impact in dataset have negative effect on accuracy. By removing this we get better accuracy with better results with other parameters also. Table 04 shows the details results in information after removing attribute "Single Epithelial Cell Size".

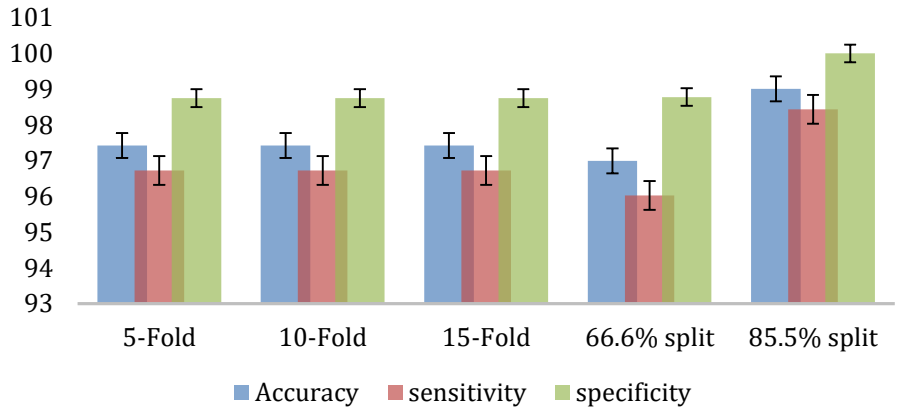

Figure 02. Accuracy, sensitivity, specificity of Naïve Bayes classifier for different test mode in modified dataset which modified by removing one attribute named as "Single Epithelial Cell Size".

After getting good performance as indicated in Table 04 we imposed different splitting and folding mechanism on the dataset and accessed results in Figure 02. It is worth mentioning that, in Table 04, for 85.5% split of train data, Naïve Bayes gave superior accuracy which is 99.0099% and also showed the best results for other parameters as well giving 100% specificity. So according to Table 04 and Figure 02, it can be observed that Naïve Bayes classifier with one attribute "Single Epithelial Cell Size" removed in WDBC dataset, improves the prediction performance in terms of all parameters.

To assess the performance of our proposed work, we have also included a comparison among state of the art approaches executed on WDBC dataset. The comparison is summarized in Table 05. It is noticeable from the comparison that, our proposed Naïve Bayes classifier with one attribute "Single Epithelial Cell Size" removed has outperformed the approaches mentioned here in terms of accuracy.

The nearest model having accuracy closer to our model is KE's algorithm proposed by Priyanka et al. (2019).

Table 05. Performance of different state of the art approaches. (Bold indicates the best result)

Authors	Classifier	Accuracy
Chintan Shah and Dr. Anjali G. Jivani (Shah and Jivani, 2013)	Naïve Bayes	95.99%
Dana Bazazeh and Raed shubair (Bazazeh and Shubair, 2016)	SVM	97%
Meriem Amrane Meriem Amrane ; Saliha Oukid ; Ikram Gagaoua and Tolga Ensarl (Amrane et al., 2018)	*K*NN	97%
Zahra Nematzadeh, Roliana Ibrahim and Ali Selamat (Nematzadeh et al. 2015)	NN	98%
G. Priyanka, V. Rohith Dr.Prasanta Kumar Sahoo, Dr.K.Eswaran (Priyanka et al., 2019)	KE's algorithm	98.53%
Our Proposed Method	Naïve Bayes ("Single Epithelial Cell Size" removed)	99.01%

So, accessing the results above it is justified to mention that, Naïve Bayes classifier with "Single Epithelial Cell Size" removed in WDBC dataset performs best in terms of performance metrics than other classifiers we imposed. Besides, it is also examined that attribute named "Single Epithelial Cell Size" has less significance on the experimental result as its removal showed better performance. But one thing worth noticing that, above 99% accuracy in training data is a indication of higher bias leading to over fitting. So, in future it will be a daunting task to remove over fitting from dataset.

IV. Conclusion

We have used the best five classification algorithms and come to a point that Naïve Bayes is superior to others compared with standard parameters. After that analysis of the dataset to extract the features of attributes to know the effectiveness of different attributes and we have checked out the performance and get a result that is better than prior result and the algorithm performing better. We removed attribute named as 'Single Epithelial Cell Size' and come to better accuracy and overall performance. But leading to such an impressive amount of accuracy, it made us think if the dataset was become oversaturated or not. So, in the future, we will try to find out the oversaturation of the reformed dataset and run a Principal Component Analysis (PCA) on the dataset.

References

[1]. Amrane, M., Oukid, S., Gagaoua, I. and Ensarl, T. (2018). Breast cancer classification using machine learning. Proceedings Article published at the 2018 Electric Electronics, Computer Science, Biomedical Engineering's' Meeting (EBBT) https://doi.org/10.1109/EBBT.2018.8391453

[2]. Bazazeh, D. and Shubair, R. (2016). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), 1-4. https://doi.org/10.1109/ICEDSA.2016.7818560

[3]. Belgiu, M. and Dragut, L. (2016). Random forest in remote sensing: A review of applications and future directions. Journal of Photogrammetry and Remote Sensing, 114, 24-31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

[4]. Dua, D. A. (2017). Breast Cancer Wisconsin (Diagnostic) Data Set. Retrieved from {UCI} Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)

[5]. Gardner, M. W. and Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric Environment, 32 (14), 2627-2636. https://doi.org/10.1016/S1352-2310(97)00447-0

[6]. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, H. I. (2009). The WEKA data mining software: an update. SIGKDD Explorations Newsletter. 11, 1 (June 2009), 10–18. https://doi.org/10.1145/1656274.1656278

[7]. Hung, P. D., Hanh, T. D. and Diep, V. T. (2018). Breast Cancer Prediction Using Spark MLlib and ML Packages. Paper presented at the Proceedings of the 2018 5th International Conference on Bioinformatics Research and Applications, Hong Kong, Hong Kong. https://doi.org/10.1145/3309129.3309133

[8]. Kaur, G. and Chhabra, C. (2014). Improved J48 Classification Algorithm for the Prediction of Diabetes. International Journals of Computer Applications, 98 (22), 13-17. https://doi.org/10.5120/17314-7433

[9]. Kim, T., Chung, B. D. and Lee, J. S. (2017). Incorporating receiver operating characteristics into naive Bayes for unbalanced data classification. Computing, 99(3), 203-218. https://doi.org/10.1007/s00607-016-0483-z

[10]. Mushtaq, Z., Yaqub, A., Hassan, A. and Su, S. F. (2019). Performance Analysis of Supervised Classifiers Using PCA Based Techniques on Breast Cancer. International Conference on Engineering and Emerging Technologies (ICEET) Lahore, Pakistan, 2019, pp. 1-6. https://doi.org/10.1109/CEET1.2019.8711868

[11]. Nematzadeh, Z., Ibrahim, R. and Selamat, A. (2015). Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques. Paper presented at the 2015 10th Asian Control Conference (ASCC). https://doi.org/10.1109/ASCC.2015.7244654

[12]. Pal, M. (2005). Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 26 (1), 217-222. https://doi.org/10.1080/01431160412331269698

[13]. Parkin, D. M. (1998). Epidemiology of cancer: global patterns and trends. Toxicology Letters, 102-103, 227-234. https://doi.org/10.1016/S0378-4274(98)00311-7

[14]. Priyanka, G., Rohith, V., Sahoo, P. K. and Eswaran, K. (2019). Breast Cancer Prediction System using KE Sieve Algorithm. International Journal of Scientific & Engineering Research, 10(1), 19-21.

[15]. Senturk, Z. R. and Kara, R. (2014). Breast Cancer diagnosis via data mining: Performance analysis of seven different algorithms. Computer Science and Engineering: An International Journal, 4 (1), 35-46. https://doi.org/10.5121/cseij.2014.4104

[16]. Shah, C. and Jivani, A. G. (2013). Comparison of data mining classification algorithms for breast cancer prediction. Paper presented at the 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). 1-4. https://doi.org/10.1109/ICCCNT.2013.6726477

[17]. Singh, S. N. and Thakral, S. (2018). Using Data Mining Tools for Breast Cancer Prediction and Analysis. Paper presented at the 2018 4th International Conference on Computing Communication and Automation (ICCCA). https://doi.org/10.1109/CCAA.2018.8777713

[18]. Wolberg, W. H. (1992). Breast cancer Wisconsin (diagnostic) data set [uci machine learning repository]. Retrieved from https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)

[19]. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., Geoffrey J. McLachlan, G. J.. Ng, A., Liu. B., Philip S. Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J. and Steinberg, D. (2008). The Top Ten Algorithms in Data Mining. Knowledge and Information Systems, 14, 1-37. https://doi.org/10.1007/s10115-007-0114-2