Feature Extraction and Elimination using Machine Learning Algorithm for Breast Cancer Biological Datasets

Ajay Kumar[1], Rama Sushil[1], Arvind Kumar Tiwari[2] and Deepanshu S. Satwaliya[1]
[1]DIT University, Mussoorie Road, Dehradun 248001, India
[2]KNIT Sultanpur-Kadipur Road, Sultanpur 228118, U.P. India

Abstract- **Cancer is a non-curable disease if diagnosed at last stage. It is not easy to diagnose cancer at an early stage. Cancer stage is divided into four stages. First two stages are known as early stage and last two stages as last stage. In order to diagnose at an early stage, Machine Learning is applied. Machine Learning is a field of computer science and Machine Learning algorithm is effective in the operation on biological data. Machine Learning is the technology that can be used to predict the accuracy of cancerous tumor of a human body. Too many features might pose some issues of over fitting the model. In that case, less important features have to be eliminated. Elimination of less important features and including important features create less computational complexity. In this paper, it is marked that with the smaller number of effective features, cancer prediction can be more accurate.**

Keywords: Accuracy; Breast Cancer; Feature Selection; Heatmap Matrix; Machine Learning

I. INTRODUCTION

The use of data science and machine learning method, in medical area proves to be prolific and a great assistance in the decision-making procedures, for reducing death rate due to breast cancer. Current status of

Cancer cases can be visualized from below statistics. As per the report (A Report: Cancer Fact & Figures 2019) for the United States, approximately 1 in 8 women (nearly about 12%) will develop invasive breast cancer over the span of her life. In 2019, a roughly 268,600 new cases of invasive breast cancer are likely to be diagnosed, inclusive of 62,930 new cases of non-invasive breast cancer. Approximately 41,760 women are estimated to die in 2019 from breast cancer. However, since 1989 death rates have been continuously decreasing. The rate of breast cancer is slightly increased by 0.4% per year from 2006-2015, whereas the breast cancer death rate is declined by 1.8% per year from 2007-2016. In India also, a total of 1671 patients were diagnosed from 2007 – 2016 and over 5 years down the line, survival rate increases up to 88.3 % and disease-free survival is 85.7% (Naga, A. M. et al. 2019).

This research paper is divided into six sections. The introduction part is explained in the section one. The second section elaborates the related work which had been done by reputed researchers and scholars. The section three exhibits the methodology part where a proposed flowchart is represented and defined the terminologies used in the further part of the research paper. The section four has been designed for an experimental procedure to compare the number of features. Result & Discussion is narrated in the section five and the last section i.e. sixth section describes the future work and conclusion.

II. RELATED WORK

Dasgupta, S. et al. (2019**)** performed an experimental result of feature selection for breast cancer datasets. The scholar included Artificial Neural Network (ANN), Bayesian Network, Random Forest method and Decision tree to develop a model for cancer detection and accuracy. Later on, scholar also compared to find out the best algorithm for prediction of cancer type depending upon level of accuracy. A short conclusion is discussed for selection of features also. Prateek. (2019) performed an intensive experiment to select features in the breast cancer dataset to identify the least important features. The scholar discussed the various machine learning algorithm such as decision tree, k-nearest neighbor, logistic regression, neural networks, naïve Bayes, random forest, and support vector machine (SVM). In the conclusion he discussed that with thirty features, naïve Bayes, random forest and SVM yield the promising score of precision 0.94, while SVM shows a precision score of 0.95 with fifteen features.

Gupta, A. et al. (2018) elaborated the three types of feature selection using machine learning. They are filter methods, wrapper method, and embedded method. Further scholar listed the advantage and disadvantage of various machine learning algorithm such as Artificial neural network, naïve Bayes, support vector machine, and decision tree. Shi, P. et al. (2011) proposed a parameter-free classifier k-top scoring pair (k-TSP) algorithm ensembled with SVM classifier. Feature selection techniques is used to cancer micro array gene pairs for the outcome of cancer prediction. Further he compared with Fisher's discriminant criterion and found k-TSP+SVM outperforms in all datasets.

Agarap, A.F.M. (2019) compared six machine learning algorithms such as SVM, Linear Regression, MLP, KNN, Softmax Regression and SVM on WBCD dataset to find accuracy, sensitivity and specificity. MLP outperforms 99.04% accuracy out of all the applied ML algorithm.

Kourou, k. et al. (2015) studied varieties of machine learning algorithm including ANN, Bayesian Network, SVM and Decision Tree and applied to cancer dataset. Vanaja, S. et al. (2014) proposed a Feature Selection Algorithm that is used for forecasting the disease accurately. To maintain the accuracy the multiclass dataset should be in the original form without data reduction. The feature Selection algorithm is capable to choose the important features and also remove the non-important features that plays a vital role in sustaining the accuracy of the classification in ML.

Zheng, B. et al. (2013) refines each appropriate feature information to support the treatment of breast cancer disease. Data Mining technique is used to extract the tumor features from the breast and diagnose it. K-mean and SVM technique are used to discover the hidden pattern of the tumor. Ensemble K-SVM minimizes the computation time along with maintaining the accuracy in the diagnosis process. Pritom, A. et al. (2016) proposed a noble approach to improve the accuracy of the model for the occurrences of breast cancer using data mining methods. The dataset collected from UCI machine learning repository have 35 attributes in which distinct ML Algorithm, Naïve Bayes, C4.5 Decision Tree and Support Vector Machine, have been used. Feature selection algorithm used to improvise the accuracy by considering the upper ranked fields in the datasets. Naïve Bayes and Decision Tree provides better outcomes after feature selection procedure.

Asri, H. et al. (2016) & Akay, M.F. (2009) claimed that SVM algorithm is more efficient in forecasting the better decision about the diagnosis of breast cancer irrespective of C4.5, K-NN, NB. Out of these four algorithms, Accuracy measurement is outperformed by SVM algorithm only. It aims to the rightness in categorizing the data with efficiency and accuracy. Ojha, U. et al. (2017) brings the light on the performance of distinct classification and clustering algorithms on Wisconsin dataset for breast cancer. The experiments showed that the classification algorithm is better than clustering algorithm. The outcome proves that the C5.0 Decision Tree and SVM gives 81% accuracy while on the other hand fuzzy c-means gives 37%.

Dana, B. et al. (2016) depicts the difference in finding the accuracy for breast cancer detection using SVM, Random Forest (RF) and Bayesian Networks (BN). The dataset was used to calculate the performance of detecting the breast cancer in terms of precision, recall, and accuracy by using these three algorithms. The experimental results showed that SVM have the highest accuracy and precision while the RFs gives the highest probability of classifying the tumor rightly. Hussain, S. et al. (2015) uses the dataset from Surveillance, Epidemiology and End Results (SEER) mainly to forecast the Survivability of the women who have the breast cancer. Further he added that Principal Components are reduced to 5 variables from 14 variables (Delen D. et al. 2005) and finally, the outcomes are same that is it again captures 98% of the total variance which was also the outcome from 14 variables.

Gayathri, B. M. et al. (2016) justifies that Relevance Vector Machine (RVM) is much better than ML algorithms. RVM produce low computational cost in comparison of ML techniques that are used to diagnose the breast cancer. Osareh, A. et al. (2010) ensemble SVM, K-NN and Probabilistic Neural Networks classifiers with Signal-to-Noise Ratio Feature Ranking, Sequential Forward Selection-based feature selection and PCA feature extraction in order to differentiate cancerous and non-cancerous tumors of breast cancer. Inclusively the outstanding accuracy for breast cancer diagnosis is brought to 96.33% by using SVM-RBF classifier for dataset having feature 25 whereas 98.80 accuracy by using SVM-RBF classifier for dataset having 11 features. Malik, A. et al. (2015) uses Extreme Learning Machine (ELM) that produced an accuracy of 93% for breast cancer detection.

III. METHODOLOGY

Method used for feature extraction and elimination using machine learning algorithm for breast cancer dataset is diagrammatically shown in flowchart form in figure 1 below.

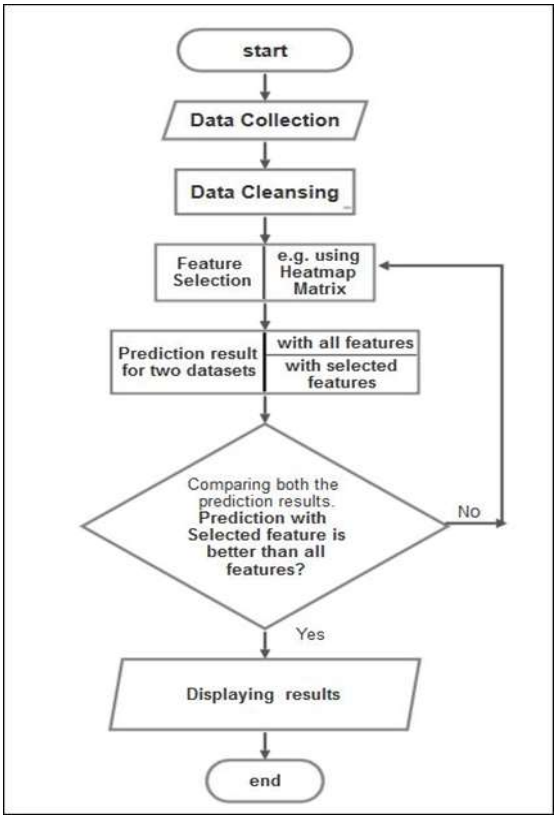

Figure 1 Flowchart for proposed methodology

All the steps used are described below in order.

3.1. Data Collection

Initially the dataset is collected from any resource repository. In this paper, data is collected from Wisconsin breast cancer dataset (WBCD) repository, i.e. available at Kaggle .com. The dataset is found in .csv format that can be seen easily through any software like MS-Excel, notepad. The type of dataset is mainly numerical whereas cancer type is denoted by a category M for malignin and B for benign. The dataset is built with 569 sample and 32 features. There are 32 features in the dataset which are listed below.

Id, diagnosis,		
radius mean, texture mean, perimeter mean, area mean, smoothness mean, compactness mean, concavity mean, concave points mean, symmetry mean, fractal_dimension_ mean,	radiuses, textures, perimeter's, areas, smoothness's, compactness's, concavity's, concave points_se, symmetrise, fractal_dimensi on_se,	radius worst, texture worst, perimeter worst, area worst, smoothness worst, compactness worst, concavity worst, concave points worst, symmetry worst, fractal_dimension _worst

Out of 32 features, first two features *id* and *diagnosis* are not considered for the experimental analysis in this paper as these are not a part of biological datasets. The datasets are categorised in three sets on the basis of *mean*, *standard error (se),* and *worst*, each having 10 features.

3.2. Data Cleansing

The next step is Data Cleansing where redundant data from the datasets is to be removed because it could provide unbiased prediction. It also includes the missing value in the dataset that could be replaced by various ways. One of the methods is to replace them by zero value but this may decrease the efficiency of the model. Therefore, most promising way is to replace the missing values in the dataset by mean value of that data column.

3.3. Feature Selection

Various tools are available for distinguished visualization of features available in the dataset. One of the major visualization tools is heat map matrix that represents a correlation between the features. Few of the feature might not have importance in order to analyze the prediction model. The feature (closer to zero) having co-efficient values should be extracted from the list of features. This step is called feature elimination. Data after elimination steps might give promising result.

3.4. Prediction of Cancer

With all features and selected features, comparing the results on basis of the parameters of confusion matrix such as accuracy level, precision, recall, and F1-score is produced.

IV. EXPERIMENT

The experimental setup is comprised of python 3.x, windows operating system 64-bit with 2 GB NVIDIA Before graphics card, Jupiter notebook. The dataset is analyzed and then uploaded in Jupiter notebook. During the data analysis, first two columns of *id* and *diagnosis* are to be dropped from the list of features as these two features might give unbiased prediction. Available data set is with 30 features only.

The heat map matrix, also known as Correlation matrix, is generated between all 30 features as shown below in figure. The correlation value ranges from -1 to 1. The value closer to 1 means features are highly correlated and inference says that the features are dependent on each other positively, while negative value which is closer to zero infer that features are independent to each other. The diagonal values are correlated with value 1. So, it is a perfect correlation. Figure 2 shows a heat map matrix of all 30 features with their correlation value.

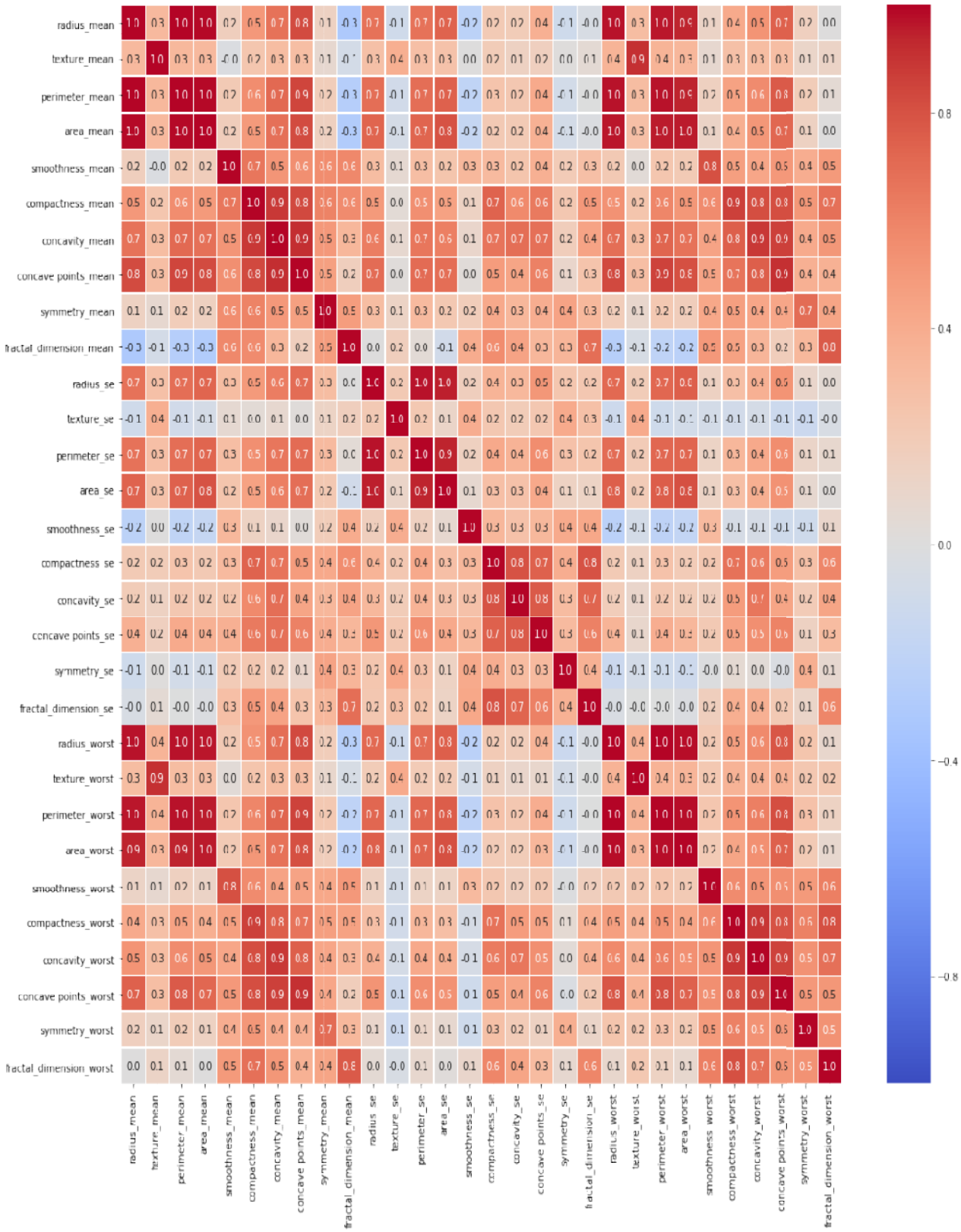

Figure 2 Heat map matrix for correlation of features

In this research work, only dependent features are considered on the basis of having correlation values greater than 0.5. Following table 1 shows the data in 3 sets on the basis of mean, standard error, & worst value. It also represents selected and eliminated features in the table.

Table 1 List of extracted features & eliminated feature

Feature status	mean set	standard error (se) set	worst set
Extracted Features	compactness mean, concavity mean, concave points mean, smoothness mean, perimeter mean, area mean,	concavities, compactness's,	concavity worst, compactness worst, concave points worst,
Eliminate Feature	radius mean, texture mean, symmetry mean, fractal _dimension_mean,	radiuses, textures, perimeters, areas, smoothness's, concave points_se, symmetry's, fractal_dimension_se,	radius worst, texture worst, perimeter worst, area worst, smoothness worst, symmetry worst, fractal_ dimension worst

V. RESULT & DISCUSSION

Data is analyzed for prediction using following selected 11 features and all 30 features too.
compactness mean, concavity mean, concave points mean, smoothness mean, perimeter mean, area mean, concavity's, compactness's, concavity worst, compactness worst, concave points worst.
For analysis, the dataset is split into training and testing part for 80% and 20% respectively. Support Vector Machine (Wang, H. et al. 2018) is applied to the dataset having 30 features & 11 features respectively. A confusion matrix is generated as shown in figure 3 below.

Figure 2 (a) Confusion matrix for 30 features

Figure 3 (b) Confusion matrix for 11 features

On the basis of confusion matrix, following table 2 shows the parameter measurement parameter such as accuracy, precision, f1-score, and support**.**

Table 2 Measurement Parameter

No. of Features	Diagnosis	Precision	Recall	F1-Score	Support	Accuracy
30	B	0.58	1.00	0.73	66	57%
	M	0.00	0.00	0.00	48	
Avg/Total		0.34	0.58	0.42	114	
11	B	0.88	0.97	0.92	66	90%
	M	0.95	0.81	0.88	48	
Avg/Total		0.91	0.90	0.90	114	

It is obvious from above results shown in table 2 that level of accuracy for breast cancer prediction is better for selected dependent features instead of taking all dependent and independent features.

VI. CONCLUSION & FUTURE WORK

It is concluded from this research work that best feature selection is an important part of data analysis for better prediction purpose with more accuracy. In future, we intend to work on to find better correlation value for better justification and selection of more dependent features for improving prediction accuracy.

VII. ACKNOWLEDGEMENTS

This research work is partially supported by a research group *Data Science & Machine Learning* at a research center in the institution. A sincere thanks to our head of the department Dr. Vishal Bharti, and Dean – School of Computing Dr. Debopam Acharya who insisted us in writing this research paper.

REFERENCES

[1] Agarap, A.F.M. 2019. "On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset", ICMLSC 2018, Phu Quoc Island, Vietnam.

[2] Akay, M.F. 2009. Support Vector Machines combined with feature selection for breast cancer diagnosis, ELSEVIER Expert Systems with Applications 36 (2009) 3240-3247, doi:10.1016/j.eswa.2008.01.009 American Cancer Society: A Report, "Cancer Fact & Figures 2019", Atlanta USA

[3] Asri, H. et al. 2016. "Using Machine Learning Algorithm for Breast Cancer Risk Prediction and Diagnosis", ELSEVIER 6th Intl Symposium. Frontier in Ambient and Mobile Systems (FAMS 2016), Procedia Computer Science 83 (2016) 1064 – 1069, doi: 10.1016/j.procs.2016.04.224

[4] Dana, B. et al. 2016. Comparative Study of machine learning algorithms for breast cancer detection and diagnosis, IEEE 978-1-5090-5306-3/16

[5] Dasgupta, S. et al. 2019. Feature Selection for Breast Cancer Detection using Machine Learning Algorithms, IJITEE, Vol 8, issue 9, ISSN 2278-3075, July 2019

[6] Delen, D. et al. 2005. Predicting breast cancer survivability: a comparison of three data mining methods, ELSEVIER Artificial Intelligence in Medicine, 34, 113-127

[7] Gayathri, B. M. et al. 2016. "Comparative study of Relevance Vector Machine with various machine learning techniques used for detecting breast cancer", IEEE Intl Conf. Computational Intelligence and Computing Research.

[8] Gupta, A. et al. 2018. Feature Selection from Biological Database for Breast Cancer Prediction and Detection using Machine Learning Classifier, J. of Artificial Intelligence, vol 11 issue 2, pp 55-64, Science Alert.

[9] Hussain, S. et al. 2015. "Reduction of variables for predicting breast cancer survivability using principal components analysis", IEEE 28th Intl Symp. Computer-Based Medical System, pp 131-134.

[10] Kourou, k. et al. 2015. Machine Learning application in cancer prognosis and prediction, ELSEVIER Computational and Structural Biotechnology Journal 13, p 8-17,

[11] Malik, A. et al. 2015. Extreme Learning machine-based approach for diagnosis and analysis of breast cancer, Taylor & Francis Journal of the Chinese Institute of Engineers,

[12] Naga, A. M. et al. 2019. Outcomes of breast cancer management from an urban specialist breast center in South India, Indian J. Medical and Paediatric Oncology, 40(5), p 102-108, 10.4103/ijmpo.ijmpo_206_17.

[13] Ojha, U. et al. 2017. "A Study on prediction of breast cancer recurrence using data mining techniques", IEEE 7th Intl Conf. cloud computing, data science & Engineering – Confluence, pp 527-530.

[14] Osareh, A. et al. 2010. "Machine Learning Techniques to Diagnose Breast Cancer", IEEE conf HIBIT Antalya, Turkey April 20-22, pp 114-120.

[15] Prateek. 2019. Breast Cancer Prediction: Important of Feature Selection, Springer Proceeding Advances in Computer Communication and Computational Sciences, (IC4S), series Advances in Intelligent Systems and Computing 924, pp 733-742, Singapore

[16] Pritom, A. et al. 2016. "Predicting Breast Cancer Recurrence using effective Classification and Feature Selection Technique", IEEE 19th Int. Conf. computer and information technology, North South University, Dhaka, Bangladesh, ISBM 978-1-5090-4089-6, p 310-314

[17] Shi, P. et al. 2011. Top Scoring pairs for feature selection in machine learning and applications to cancer outcome prediction, 12:375, BMC Bioinformatics.

[18] Vanaja, S. et al. 2014. Analysis of Feature Selection Algorithms on Classification: A Survey, Int. J Computer Application (0975-8887), vol 96 No.17.

[19] Wang, H. et al. (2018). "A Support Vector Machine-based ensemble algorithm for breast cancer diagnosis", ELSEVIER European Journal of Operational Research 267 (2018) 687-699, https://doi.org/10.1016/j.ejor.2017.12.001

[20] Zheng, B. et al. 2013. Breast Cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms, ELSEVIER Expert System with Applications, http://dx.doi.org/10.1016/j.eswa.2013.08.044