Advanced feature selection methodology for cancer datasets to improve accuracy of classification

Dr.D.Venkatesh[1], Dr.L.Venkateswara Reddy[2]

[1,2]Professor

[1]Department of CSE, [2]Department of IT

[1]Anantha Lakshmi Institute of Technology and Sciences, Anantapur, Andhra Pradesh

[2]Sree Vidyanikethan Engineering College, Tirupati, Andhra Pradesh

[1]dvvenkatesh@yahoo.co.in, [2]lakkireddy.v@gmail.com

Abstract

In cancer disease, there are four stages in which first two stages are called as early stage and the last two stages are known as last stages. As we know the cancer is a non-curable disease if we diagnose at the last stages and it is also not an easy task to diagnose the cancer at early stages also. For the early stage diagnosis, we are can apply machine learning techniques. Machine Learning is one of the trending domains in the computer science discipline and also machine learning algorithms are so effective for analyzing the biological datasets. By the help of machine learning methodologies, we can analyze the cancerous tumors of a human body with high accuracy. In this type of prediction type of model, we can have some issues related to overfitting model. In which, we are having less priority features to be removed. There will be complexity in computational access while eliminating low priority features and also some high priority features. Hence, in this research work we are focusing on few numbers of high priority features and predicting the cancer more accurately.

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

1. Introduction

The usage of machine learning and data science technology, in medical area proves to be prolific and a more assistance in decision-making procedures, for reducing death rate due to breast cancer[1-4]. 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 one in eight women (nearly 11%) are having breast cancer. In the 2019, a roughly 269,600 new cases of invasive breast cancer are likely to be diagnosed, inclusive of 63,390 new cases of non-invasive breast cancer. Approximately 40,760 women are estimated to die in 2019 from breast cancer. However, since 1988 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 has been explained in six parts. 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 [5-10]. 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.

2. Related Work

Performed an experimental result of feature selection for breast cancer datasets. The scholar included 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 [10-15]. 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 [15-21].

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 naïve Bayes, ANN, SVM, & decision tree. 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. methodologies, we can analyze the cancerous tumors of a human body with high accuracy. In this type of prediction type of model, we can have some issues related to overfitting model. In which, we are having less priority features to be removed. There will be complexity in computational access while eliminating low priority features and also some high priority features.

Zheng, B. et al. (2013) refines each suitable element data to help the treatment of bosom malignant growth illness. Information Mining procedure is utilized to remove the tumor highlights from the bosom and analyze it. K-mean and SVM strategy are utilized to find the shrouded example of the tumor. Gathering K-SVM limits the calculation time alongside keeping up the precision in the conclusion procedure. Pritom, A. et al. (2016) proposed a respectable way to deal with improve the precision of the model for the events of bosom disease utilizing information mining techniques. The dataset gathered from UCI AI store have 35 properties in which unmistakable ML Algorithm, C4.5 Decision Tree, Naïve Bayes & SVM, have been utilized. Highlight determination calculation used to extemporize the exactness by considering the upper positioned fields in the datasets. Bayes and Decision Tree gives better results after component determination method.

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 breast cancer. Osareh, A. et al. (2010) gathering SVM, K-NN and Probabilistic Neural Networks classifiers with Signal-to-Noise Ratio Feature Ranking, Sequential Forward Selection-based component determination and PCA highlight extraction so as to separate carcinogenic and non-dangerous tumors of bosom disease. Comprehensively the exceptional exactness for bosom malignancy determination is brought to 96.33% by utilizing SVM-RBF classifier for dataset having highlight 25 through 98.80 precision by utilizing SVM-RBF classifier for dataset having 11 highlights. Malik, A. et al. (2015) utilizes Extreme Learning Machine (ELM) that created an exactness of 93% for bosom malignant growth location [22-25].

3. Methodology

In this below figure 1 flowchart we have represented the feature selection & elimination for breast cancer datasets using machine learning algorithm.

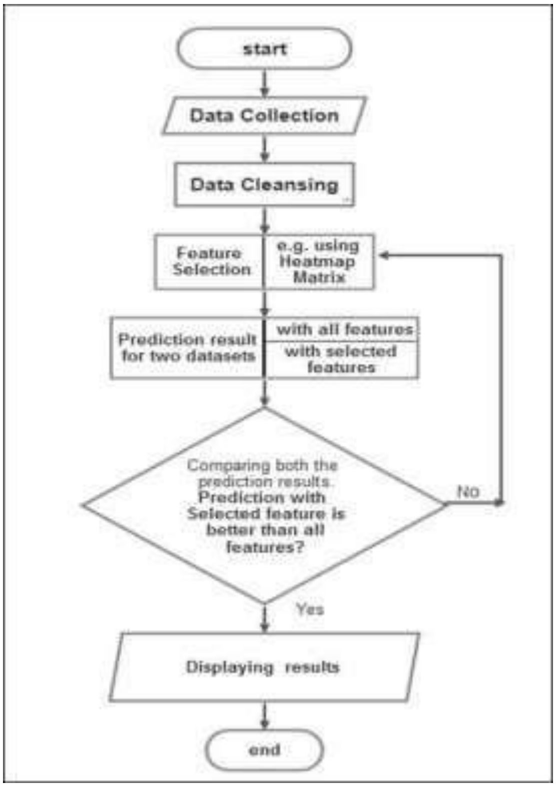

start
Data Collection
Data Cleansing
Feature Selection
e.g. using Heatmap Matrix
Prediction result for two datasets
with all features
with selected features
Comparing both the prediction results. Prediction with Selected feature is better than all features?
No
Yes
Displaying results
end

Fig. 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 extracted from 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.

List of Features:

Id, diagnosis,
Mean of Radius,
Mean of texture,
Mean of perimeter,
Mean of Area,
Mean of
smoothness,
Mean of
compactness,
Mean of concavity,
Concave_mean of points,
mean of symmetry,
Mean of fractal dimension,
se of radius,
se of texture,
se of perimeter,
se of area,
se of smoothness,
se of compactness,
se of concavity,
se of concave
se of points,
se of symmetry,
se of fractal dimension,

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 categorized 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 different ways. Thusly, most encouraging path is to supplant the missing qualities in the dataset by mean estimation of that information section.

3.3. Feature Selection

Various tools are available for distinguished visualization of features available in the dataset. One of the major visualization tools is heatmap 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, recall, precison & F1-score is produced.

4. Experiment

The experimental setup is comprised of python 3.x, windows operating system 64-bit with 2 GB NVIDIA Geforce graphics card, Jupyter notebook. The dataset is analyzed and then uploaded in Jupyter 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 heatmap 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 heatmap matrix of all 30 features with their correlation value.

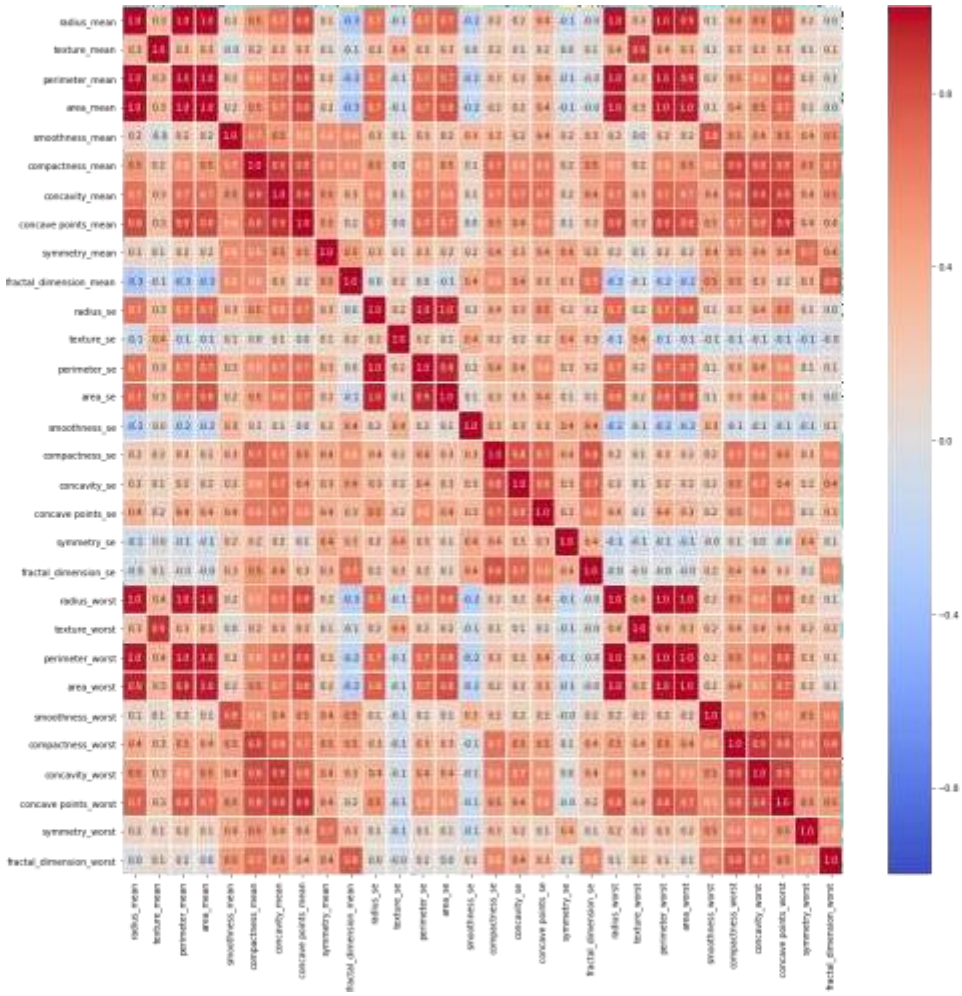


Fig. 2. Heatmap matrix for correlation of features

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

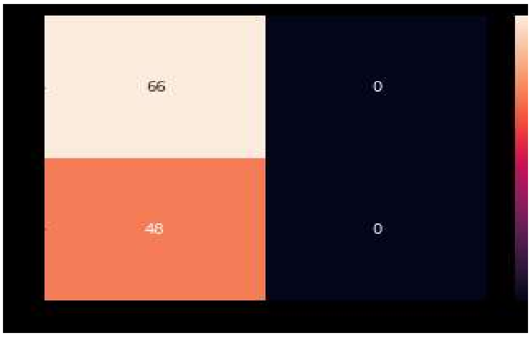
Feature status	mean set	standard error (se) set	worst set
Extracted Features	Mean ofCompactness, mean of concavity, mean of concave points,	Se of Concavity, se of compactness,	Worst of concavity, worst of compactness, worst of concave points

	Smoothness-mean, perimeter-mean, area-mean,		
Eliminate Feature	mean of radius, mean of texture, mean of symmetry, mean of fractal-dimension	se of radius, se of texture, se of perimeter, se of area, se of smoothness, se of concave points, se of symmetry, se of fractal dimension	Worst in Radius, worst in texture, worst in perimeter, worst in area, worst in smoothness, worst in symmetry, worst in fractal dimension

5. Result & Discussion

Data is analyzed for prediction using following selected 11 features and all 30 features too. Compactness mean, concavity mean, concave point mean, smoothness mean, perimeter mean, area mean, concavity se, compactness se, 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.



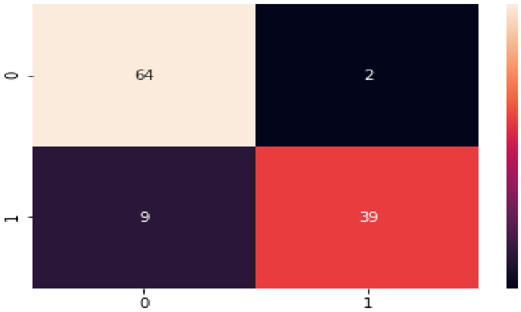


Fig. 2. (a) Confusion matrix for 30 features Fig. 3. (b) Confusion matrix for 11 features

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	
	M	0.00	0.00	0.00	48	57%
Avg/Total		0.34	0.58	0.42	114	
11	B	0.88	0.97	0.92	66	
	M	0.95	0.81	0.88	48	90%
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.

6. Conclusion & Future work

It is concluded from this research work that best selection of features has been crucial 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.

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

[3] American Cancer Society: A Report, “Cancer Fact & Figures 2019”, Atlanta USA

[4] 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

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

[6] 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

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

[8] 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.

[9] 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.

[10] Ummadi Janardhan Reddy, Pandluri Dhanalakshmi, Pallela Dileep Kumar Reddy Image Segmentation Technique Using SVM Classifier for Detection of Medical Disorders Ingénierie des Systèmes d'Information, Vol. 24, No. 2, pp. 173-176, April 2019, https://doi.org/10.18280/isi.240207 (Scopus) ISSN: 1633-1311

[11] G. Ramu, P. Dileep Kumar Reddy, Appawala Jayanthi "A Survey of Precision Medicine Strategy Using Cognitive Computing" International Journal of Machine Learning and Computing, Vol. 8, No. 6, December 2018 DOI: 10.18178/IJMLC2018.8.6.741 (Scopus) (UGC Approved) Journal No: 48748, pp 530 to 535

[12] Ramu, G. A secure cloud framework to share EHRs using modified CP-ABE and the attribute bloom filter. Educ Inf Technol 23, 2213–2233 (2018). DOI https://doi.org/10.1007/s10639-018-9713-7

[13] Ramu, G., Reddy, B.E., Jayanthi, A. et al. Fine-grained access control of EHRs in cloud using CP-ABE with user revocation. Health Technol. 9, 487–496 (2019). http://dx.doi.org/10.1007/s12553-019-00304-9

[14] P. Dileep Kumar Reddy, R. Praveen Sam, C. Shoba Bindu "Optimal Blowfish Algorithm based Technique for Data Security in Cloud"Int. J. Business Intelligence and Data Mining, ISSN online 1743-8195, ISSN print 1743-8187, Vol. 11, No. 2, 2016.Pp.171–189.DOI: 10.1504/IJBIDM.2016.10001484. (Inder Science)(UGC Approved). Journal No: 16481

[15] J. Somasekar a, , G. Ramesh , Gandikota Ramu, P. Dileep Kumar Reddy, B. Eswara Reddy e, Ching-Hao Lai, "A dataset for automatic contrast enhancement of microscopic malaria infected blood RGB images", Data in brief, Elsevier, https://doi.org/10.1016/j.dib.2019.104643,2352-3409/2019

[16] 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.

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

[18] 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,

[19] 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.

[20] 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.

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

[22] 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

[23] 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.

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

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

[26] Lakshmi Prasanna, K., & Ashwini, S. (2019). Automatic breast cancer detection and classification using deep learning techniques. Test Engineering and Management, 81(11-12), 5505-5510. Retrieved from www.scopus.com

[27] Kumar, A., Sushil, R., & Tiwari, A. K. (2019). Feature extraction and elimination using machine learning algorithm for breast cancer biological datasets. International Journal of Advanced Science and Technology, 28(20), 425-435. Retrieved from www.scopus.com