BBRC
Bioscience Biotechnology Research Communications

Dimentionality Reduction Using Principal Compound Analysis in Supervised Machine Learning Techniques

G Nirmala[1], S Prabu[1], A Azhagu Jaisudhan Pazhani[2] and S Vairaprakash[3]

[1]Department of ECE, Mahendra Institute of Technology, Namakkal, India
[2](Sr.Gr.), Department of ECE, Ramco Institute of Technology, Rajapalayam, India
[3]Department of ECE, Ramco institute of Technology, Rajapalayam, India

ABSTRACT

Breast cancer is the most common diseases among women's world-wide. The survival rate of the women may increase early diagnosis of the disease. Researchers helping the physicians for analyzing and predicting the breast cancer as early as possible using various technologies. This research explores the feature reduction property of Principal Compound Analysis (PCA) on breast cancer decisions support system from wisconsin breast cancer dataset which are analyzed in both two dimensional and 3 dimensional components. The data are reduced to 4 features using chi-square method and evaluated the accuracy of classifiers such as K-Nearest Neighbor (K-NN), Linear Regression(LR), Support Vector Machine (SVM), Random Forest(RF), Decision Tree (DT),Gaussian Naïve Bayes(GNB) and Artificial Neural Network (ANN). This is validated with 10 fold cross-validations. These classifiers are evaluated ,in which the ANN method provides high accuracy of 97.00% and also yields better selectivity and sensitivity rates rather than other machine learning algorithms.

KEY WORDS: BREAST CANCER, PCA, MACHINE LEARNING ALGORITHMS, ROC AND ANN

INTRODUCTION

Breast cancer is the massive disease which causes death among women around the globe. We need effective methodology and algorithms to identify the masses, calcifications and architectural distortion. Also we need of detecting whether that is benign or malignant. In recent days the affected victims are slightly increased to 30% (Siegel et al. 2018). Early detection leads the survival rate of the women. Mammogram is one of the low cost effective tool and diagnosis and accuracy plays an important role in diagnosis. In the last few decades, the development of machine learning techniques is too vast for detection of breast cancer and classification (Yue et al. 2018). The feature extraction plays a major role in enhancing the features and diagnosing the abnormality presence and to determine whether it is benign or malignant. Since having many difficulties in diagnosing in analyzing the breast cancer researchers focusing on many parameters and algorithms. The analysis of diseases by utilizing medical imaging is popular in the field of medical (Al-Hajj et al. 2003). The intelligent healthcare always supports the physicians to achieve meaningful benchmarks (Yue et al. 2018).

Literature Survey: Ahmad et al., analyzed the performance of the classifiers decision tree (C4.5), SVM, and ANN. They used the Iranian center dataset for breast cancer. They identified SVM was the best classifier. Liu et al applied predictive model based on Decision Table (DT) to predict the survivability of breast cancer patient. The C5 technique has the 86.52 % accuracy for predicting the patients survival rate and bagging algorithm is applied to deal with data imbalance problem. This model increase the prediction performance on breast cancer. Chaurasia and Pal compare the performance metrics of Simple Classification and Regression Tree (CART), Decision Tree (DT) (J48), RBF neural networks, SVM-RBF kernel, Naïve

ARTICLE INFORMATION

**Corresponding Author: g.s.nila@gmail.com*
Received 15th Oct 2020 Accepted after revision 31st Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

Clarivate Analytics   crossref

NAAS Journal Score 2020 (4.31)
A Society of Science and Nature Publication,
Bhopal India 2020. All rights reserved.
Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.13/50

Bayes to identify the best method for classification of breast cancer datasets.

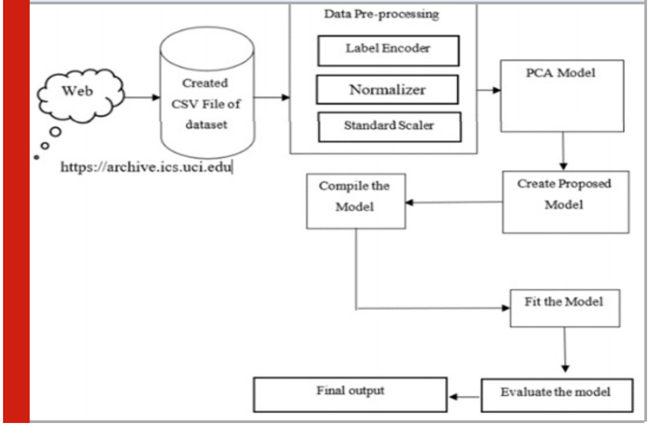
The result shows that SVM-RBF kernel has higher accuracy of 96.84 % in the Wisconsin Breast Cancer (original) datasets. The DL based method have become integral part of the doctors and pathologists in clinical practices. Some examples are breast cancer detection and classification, lung cancer detection, Alzheimer and Brain tumors detection. In 2004, two ML methods such as ANN and DT were compared with statistical method of linear regression to predict the survival of breast cancer in large dataset of more than 200,000 cases and demonstrate that ML methods could be a promising classification for practicle use. The result shows that DT method has higher accuracy of 93.6 % with ANN achieves 91.2 % and both have higher performance than linear regression achieving 89.2 % accuracy.

Doctors requires regular breast screening for the early detection of breast cancer. Early detection of disease helps to identify the breast cancer before patient felt the symptoms. This includes observation of tissues in breast-part for any abnormal lumps. In this stage, a Fine-Needle Aspiration (FNA) biopsy method is required if any mass or lump is appeared on the breast while screening (Sannasi Chakravarthy et al., 2019). The SVM process on the training dataset and each data tuple is allocated with a class label in training dataset. In the training examples, each data represented as a point in an n-dimensional space, where n denotes the number of features. The SVM maps new data to the nearest class. This process provides the huge gap splitting and dividing them by representing the data in the various categories and this gap is known as hyperplane. The hyperplane separates the data and the optimal hyperplane provides the big gap for classifying the data.

The decision tree applies the tree structure to visualize the data and denotes in consequences and sequences. The “root node” is the topmost node of the tree and the internal nodes present a test on the attributes. The outcome of the test is called “branch”. The leaf nodes are the nodes without further branching and this denotes the class label of all prior decisions (Kowalczyk 2017). The random forest is based on the generating trees. The random forest is the simple algorithm that applies only two parameters namely the number of trees in the forest and the number of variables in the random subset. This algorithm creates the different trees based on the original data and the best split predictor is used to prunes the trees at each nodes. The trees predictions are aggregating to predict the new data (Liaw & Wiener 2002).

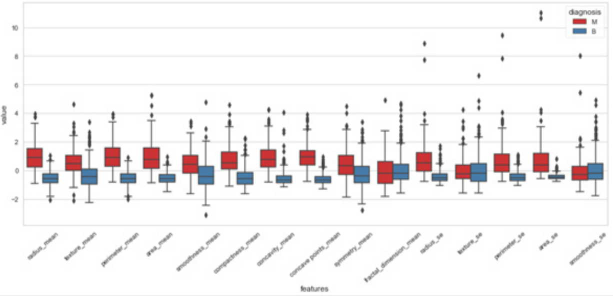
Machine Learning Approach: Machine learning can employ in various fields to classify patterns or to build prediction models. It can be broadly divided into two types: (1) Supervised learning and (2) Un-supervised learning based on the used data and their availability. The machine learning demand is increasing day by day based on the service requirements. But this field has higher barriers and frequently needs expert knowledge. Typical machine learning approach pipeline is as follows (1) preprocessing (2) feature selection, and (3) classification. In the proposed method, the selection of methods, parameters in the process of preprocessing and classification stage is automatically detected. By using various machine learning Algorithms like K-NN, LR, SVM, RF,DT, GNB and ANN,can effectively determine the problems and can the solve.

Figure 1: Machine Learning Pipeline



3.1. Data Exploration: Classification generally used to optimize problem. Wisconsin breast cancer dataset was used (Mangasarian et al. 1990; Wolberg et al. 1990. The data set includes recording collected from the biopsies of patients in various hospitals of wisconsin. Grouped the data points chronologically in a way which original medical cases are reported. The datasets includes 699 samples or instances that are characterized by nine features or attributes. The some more datasets are available for analyzing the breast cancer. They are(1) Wisconsin Prognostic Breast-Cancer Chemotherapy (WPBCC) and (2)Wisconsin Diagnostic Breast-Cancer (WDBC) (Sharma et al. 2017). The larger number of ML algorithms are utilized to analyze the data set. The three main features in diagnosis are preprocessing, feature selection and classification.

Figure 2: Benign and Malignant classification in the dataset

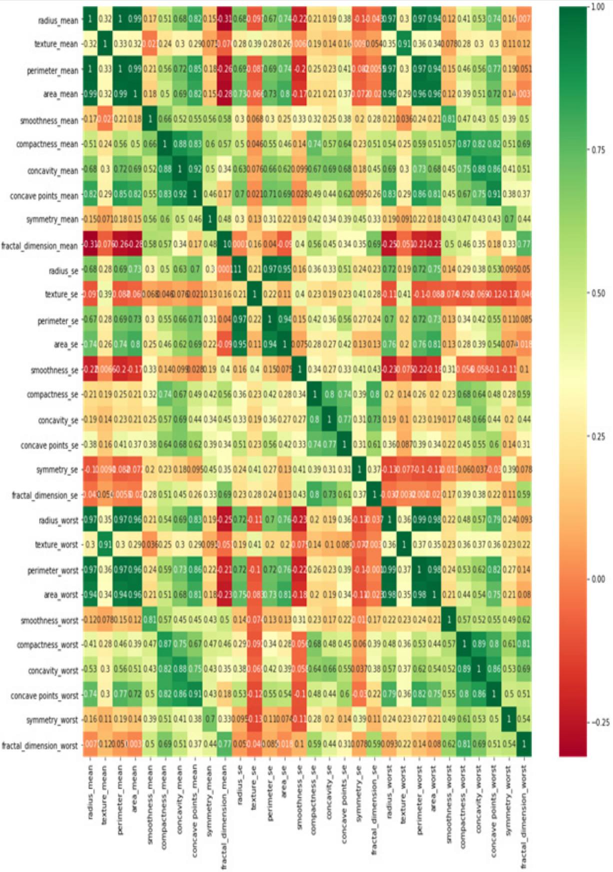


3.2. Preprocessing: Data pre-processing utilized for filtering the data in an usable format. Noise is a common factor with the real time dataset.so we need to preprocess the database. Data pre-processing transforms the dataset into usable format by using standardization approach to preprocess the UCI dataset(Parameshachari et al. 2020). The raw data is preprocessed to scale the feature and using standard scalar modules. Statistical Analysis is distributed in the graph. The errors are calculated here as outliers. In order to calculate the outliers the following steps taken.

1. Calculate first quartile (Q1) (25%)
2. Find IQR (inter quartile range) = Q3-Q1
3. Compute Q1 - 1.5IQR and Q3 + 1.5IQR

3.3. Correlation Matrix: Correlation based Feature Selection (CFS) is an fitter based approach used to select the features and feature attributes weight at the intrinsic property of data (David et al. 2019). Generally, the feature attributes are largely correlated with one another. The features that largely correlates gives redundant details and are excluded by CFS. Similarly, the features that largely interrelate with the labels of class are selected and retained (Kowalczyk 2017). The correlation matrix provides the heat-map for the input features.

Figure 3: Heat Map for the correlated value of 30 attributes

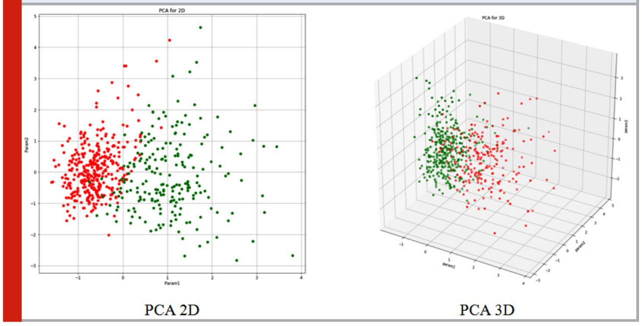


The greater the correlation value between two features, then it is observed as the related and inclusion of one feature is more sufficient. Hence, the feature selection technique provides the reduced attributes such as Radius_Mean, Perimeter_Mean, Concavity_Mean ,Area_ Mean and concave Points_Mean shown in figure.3. The selected features from the heat map can be used for the classification and regression analysis. The benign and malignant parts are explained in correlation and Non-correlation features. We can see several of them are highly correlated between each other.

3.4. Principal Compound Analysis: PCA is a statically method used for data analysis. Basically it is a dimension reduction technique which includes the relate features. The main function of PCA is to detect the patterns in dataset and find similarity and differences between each individual attributes. Variance of breast cancer dataset can be determined as. Feature extraction plays a vital role while processing data. It helps to distinguish the benign and malignant (Salembier & Garrido 2000) which helps to create an better predictive approach.

It includes various benefits for applying the method of feature selection: (a) very faster and effective in training the ML algorithms, (b) decreases the complexity of model and easy for interpretation (c) improves the model accuracy with chosen subset (d) Decreases the problem of over fitting (Dhahri et al. 2019). Before learning, we are reducing the data set for maintain the most significant features along with randomized Single Value Decomposition, Low variance, Univarience and recursive features (Shlens 2005). We are using min-max scaler for preparing data.Here we are using 70 % of data for training and 30% of data for testing.

Figure 4: Principal compound Analysis (a) 2D PCA (b) 3D PCA



PCA 2D

PCA 3D

The Chi-Square method basically checks the independency and tests the relationship among different variables. The Null hypothesis in Chi-Square test means no relationship exists among the different variables in population which are independent of two variables, that is, it checks whether there is a considerable correlation between them. According to these correlations, it ranks the features according to their importance. In this part, we used this method to select the 4 best features shown in Table.1.The Chi-Square statistic calculation is quite intuitive and straightforward:

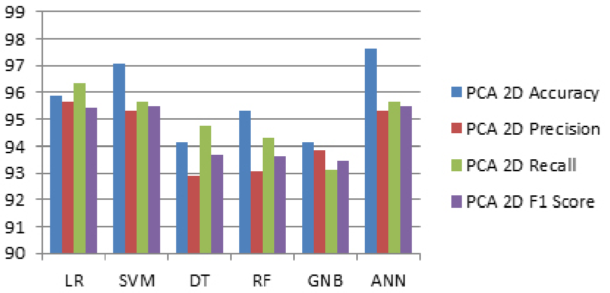
$$\chi_c^2 = \sum \frac{(Oi-Ei)2}{Ei} \quad (1)$$

Where
Oi, is the observed frequency counts in the cell
Ei is the expected frequency if no relationship existed among the variables

Figure 5: Confusion Matrix

Predicated values	Actual Values	
	Positive	Negative
Positive	True Positive	False Positive
Negative	False negative	True negative

Figure 6: Comparision Matrices of PCA 2D



RESULTS AND DISCUSSION

Calculation Matrices: Generally various parameter metrics are used to evaluate a model in machine learning. We will use the following common metrics to evaluate our models' performances: i) accuracy; ii) precision; iii) recall or sensitivity; iv) F1 score; v) Receiver Operating Characteristic (ROC) curve; with 10 fold cross validation.

$$Accuraccy = \frac{TP+TN}{FP+FN+TP+TN} \qquad \text{------ (2)}$$

$$Recall = \frac{TP}{TP+FN} \qquad \text{------ (3)}$$

$$Precision = \frac{TP}{TP+FP} \qquad \text{------ (4)}$$

$$F1\ score = 2\ X\ \frac{(Precision\ X\ Recall)}{(Precision+Recall)} \qquad \text{------ (5)}$$

Confusion Matrix: Confusion matrices are estimated in classifier to predict benign and malignant analysis for all the Machine Learning algorithms. The confusion matrix forms a two-by-two dimension shown in figure 5.

Table 2. Comparison Table of PCA 2D, PCA 3D and 4 Best features

Algorithm	PCA 2D				PCA 3D				4 Best features			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
Linear Regression	95.90	95.64	96.31	95.45	97.07	96.31	97.3	96.80	92.98	93.35	91.03	92.03
KNN	94.73	95.31	95.67	95.49	95.90	94.33	93.97	94.15	94.74	94.34	93.91	94.15
SVM	97.08	95.31	95.67	95.49	95.91	96.60	96.97	96.97	94.15	94.72	92.38	93.36
Decision Tree	94.15	92.90	94.73	93.67	94.15	92.90	94.73	93.67	90.06	92.90	94.73	93.67
Random Forest	95.32	93.06	94.33	93.62	94.15	94.31	95.62	95.32	92.98	93.36	91.03	92.03
Gaussian Naïve Bayes	94.15	93.86	93.13	93.48	94.15	93.53	93.53	93.53	93.58	93.39	92.28	92.79
ANN	97.66	95.31	95.67	95.49	95.90	97.08	97.81	97.43	95.62	94.68	94.42	95.01

Figure 7: Comparision Matrices of PCA 3D

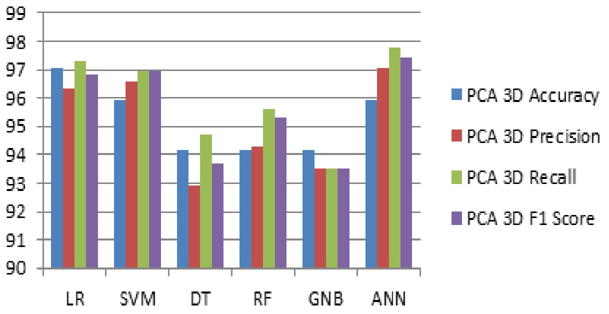


Figure 8: Comparision Matrices of 4 Best Features

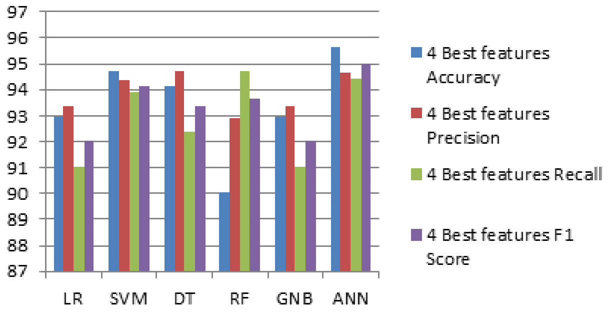


Figure 9: (a),(c)Confusion Matrix for ANN with 1 hidden layer for PCA 2D and PCA 3D (b),(d) Confusion Matrix for ANN with 2 hidden layer for PCA 2D and PCA 3D

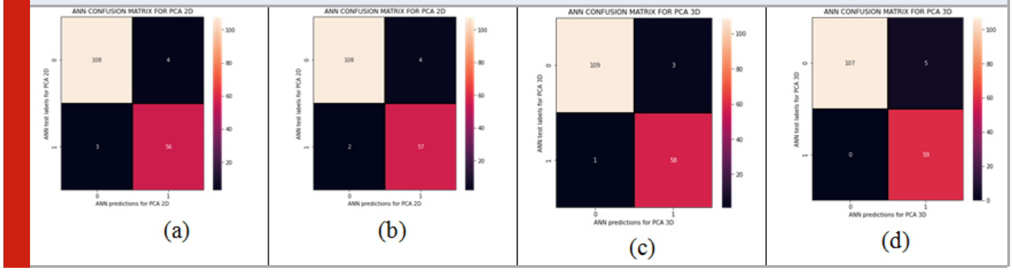
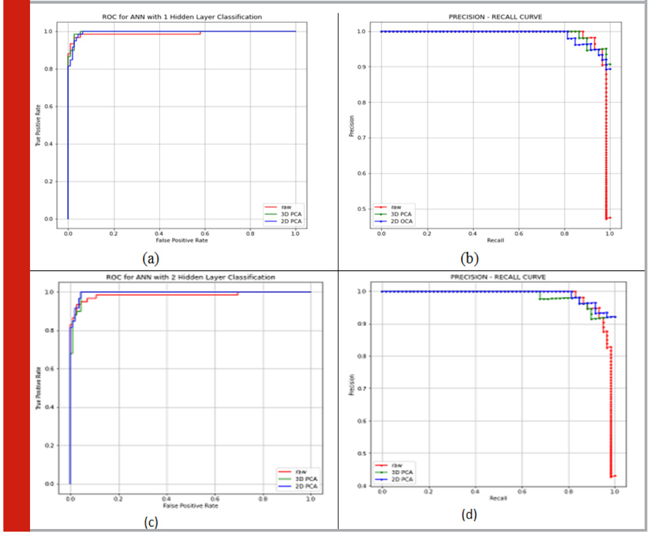


Figure 10: ROC and Precision –Recall curves. (a) ROC for ANN 1hidden layer classification (b)Precision-Recall curve for ANN 1hidden layer classification (c) ROC for ANN 2hidden layer classification (d) Precision-Recall curve for ANN 2hidden layer classification



From the Table.2, we have seen that all the Machine Learning Algorithms are working effectively on both the 2D, 3D and selection features. Some algorithms performs well and yields good accuracy in PCA 3D than the PCA 2D such as Linear Regression (97.07%), KNN(95.90%). Few algorithms performs well on PCA 2D and provide good accuracy like SVM, RF and ANN. The DT provides same values in both 2D and 3D PCA. The 4 Best features out performs well and yields very close results to the PCA 2D and 3D, shows need to reform it. Out of all the ML algorithms the ANN out performs well, yields very Accuracy (97.66%) for PCA 2D, Precision (97.08%), Recall (97.81%) and F1 score (97.43) for PCA 3D. Rather than accuracy, quite interesting is that, PCA 3D provides very good precision, recall and F1 score in all the ML algorithms which shows out performs well than the PCA 2D with 10 fold cross validation.

The ROC curves is the effective technique used to compute the performance of classifier for different training and testing separations with True Positive (TP) and False Positive(FP). It is parameterized by the probabilities of threshold values. The TP rate describes the fraction of positive cases which are classified correctly by model. So, it provides trade-off among ANN 1 hidden layer and 2 hidden layer precision and Recall ROC curve which shown in fig10.

CONLUSION

Various ML techniques can be utilized for the breast cancer prediction. The difficulty is to develop an computationally efficient and accurate medical data classifiers. Each algorithm performs in a various ways depending on the datasets and selection of parameters. In the proposed ML Algorithms ANN achieves high accuracy rate in PCA 2D and Precision, Recall and F1 score are very good at PCA 3D. In our future work, we would like to choose an optimal feature that plays an major role in improving the classifier performance .Further, this can also be implemented on cloud platform for ease of utilization.

REFERENCES

"UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set." [Online]. Available:

A. Kowalczyk, Support Vector Machines Succinctly. 2017.

A. Liaw, and M. Wiener, "Classification and regression by randomForest," R News, vol. 2, no. 3, pp. 18–22, 2002.

AC. Tan, D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification",

Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison SJ, Clarke MF, Prospective identification of tumorigenic breast cancer cells, NCBI, 2003 May 27;100(11):6890.,DOI:10.1073/pnas.0530291100.

American Cancer Society. 2018. Global Cancer: Facts & Figures, 4th edition http://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/global-cancer-facts-and-figures/global-cancer-facts-and-figures-4th-edition.pdf.Appl. Bioinform, Vol. 2, pp. 75-83, 2003.

Chaurasia V and Pal S. Data mining techniques: to predict and resolve breast cancer survivability. Int J Comput Sci Mobile Comput 2014; 3: 10–22.

David A. Omondiagbe 1 , Shanmugam Veeramani 1*, Amandeep S. Sidhu, Machine Learning Classification Techniques for Breast Cancer Diagnosis, IOP Conf. Series: Materials Science and Engineering 495 (2019)

012033 doi:10.1088/1757-899X/495/1/012033

Delen, D.; Walker, G.; Kadam, A. Predicting breast cancer survivability: A comparison of three data mining methods. Artif. Intell. Med. 2005, 34, 113–127.

Habib Dhahri ,Eslam Al Maghayreh,Awais Mahmood,Wail Elkilani,and Mohammed Faisal NagiAutomated Breast Cancer Diagnosis Based on Machine Learning Algorithms, Journal of Healthcare Engineering,Volume 2019, Article ID 4253641,https://doi.org/10.1155/2019/4253641 https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29. [Accessed: 29-Dec-2015].

International Conference on Advances in Computing, Communication Control and Networking (ICACCCN2018) ISBN: 978-1-5386-4119-4/18/$31.00 ©2018 IEEE 98 Breast Cancer Diagnosis Using Deep Learning Algorithm

J. R. Quinlan, “Induction of decision trees,” Machine Learning, vol.1, no. 1, pp. 81–106, 1986.

J. Shlens, ”A tutorial on principal component analysis”, Systems Neurobiology Laboratory Salk Institute for Biological Studies, 2005.

JF McCarthy, M.K., PE Hoffman, ”Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management”, Ann N Y Acad Sci, Vol.62, pp. 10201259, 2004

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical image analysis, 42, 60-88.

Liu, Y-Q, Wang, C, Zhang, L. Decision tree based predictive models for breast cancer survivability on imbalanced data. In: 3rd international conference on bioinformatics and biomedical engineering, 11-13 June 2009, Beijing, China, 2009.

Mangasarian, O.L.; Setiono, R.; Wolberg, W.H. Pattern recognition via linear programming: Theory and application to medical diagnosis. In Large-Scale Numerical Optimization; SIAM: Philadelphia, PA, USA, 1990;pp. 22–31.

P. Salembier and L. Garrido, “Binary partition tree as an efficient representation for image processing, segmentation,and information retrieval,” IEEE Transactions on Image Processing, vol. 9, no. 4, pp. 561–576, 2000.\

Parameshachari, B.D., Panduranga, H.T. and liberata Ullo, S., 2020, September. Analysis and Computation of Encryption Technique to Enhance Security of Medical Images. In IOP Conference Series: Materials Science and Engineering (Vol. 925, No. 1, p. 012028). IOP Publishing.

R.L. Siegel, K.D. Miller, and A. Jemal, “Cancer statistics, 2018”,CA: A Cancer Journal for Clinicians 68 (1), 7–30 (2018).

Sharma, A.; Kulshrestha, S.; Daniel, S. Machine learning approaches for breast cancer diagnosis and prognosis. In Proceedings of the International Conference on Soft Computing and Its Engineering Applications, Changa, India, 1–2 December 2017.

Wenbin Yue , Zidong Wang,, Hongwei Chen , Annette Payne and Xiaohui Liu, “Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis”, Designs 2018, 2(2), 13; https://doi.org/10.3390/designs2020013

Wolberg, W.H.; Mangasarian, O.L. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proc. Natl. Acad. Sci. USA 1990, 87, 9193–9196.