**IMPACT: International Journal of Research in Applied,
Natural and Social Sciences (IMPACT: IJRANSS)
ISSN (P): 2347–4580; ISSN (E): 2321–8851
Vol. 8, Issue 3, Mar 2020, 7–16
© Impact Journals**

BREAST CANCER DIAGNOSIS

Supriya. U, Divya. R M, Janani. P & Nivetha. R

*Research Scholar, Department of Computer Science and Engineering, Bannari Amman Institute of Technology,
Sathyamangalam, Erode, TamilNadu, India*

Received: **18 Mar 2020** **Accepted:** **23 Mar 2020** **Published:** **31 Mar 2020**

ABSTRACT

After lung cancer, Breast Cancer is the most intricate disease diagnosed among 8% of women. Breast Cancer is characterized by the mutation of genes, constant pain, changes in the size, color (redness), skin texture of breasts. The detection and treatment of breast cancer involves screening at an early stage by which the mortality can be reduced by a quarter. Data Mining techniques are used to predict the probability of a person who are affected by Breast Cancer using certain Data Mining Techniques. The techniques include KNN Classification, Support Vector Machine Algorithm, and Confusion Matrix Algorithm. The project is implemented using Python programming and deploys the analytical results as a comparision plot. The data-set has been fetched from the UCI repository, and it contains different attributes or factors such as patient_id, unif_cell-size, unif_cell-shape, marg_adhesion and bare_nuclei. In this project, the results are deployed as a comparision chart, which provides better visualization. Various techniques have been used to understand the behavior patterns and to predict the percentage of occurrence of breast cancer. We infer from our study that KNN is well suited algorithm for diagnosis with a accuracy of 97%.

KEYWORDS: *K-Nearest Neighbors, Support Vector Machine, Confusion Matrix, Feature Extraction, Principle Component Analysis, Algorithm Comparision and Optimization*

INTRODUCTION

Breast cancer is the main reason that causes thedeath of women. It is the second vicious cancer after lung cancer. According to the statistics provided by World Cancer Research Fund in 2019, it is estimated that over 2 million new cases were recorded out of which 626,679 deaths were approximated. Of all cancers, breast cancer constitutes of about 24.2% of cancer among Women. If any symptoms diagnosed, people are advised to consult doctor immediately and if needed, they are referred to Oncologist. The Oncologist diagnoses breast cancer through medical history, physical examination and also by checking the symptoms such as swelling or hardening of any lymph nodes in the armpit. The prior diagnosis can enhance the prediction and survival rate, so that the patients can be intimated to take the clinical treatment at the right time. The research is to be carried out for the proper diagnosis of Breast Cancer and prediction of occurrence rate among patients. The analysis is being carried out in this area by applying various machine learning and data mining techniques on many different datasets for Breast Cancer. The analysis result shows that different algorithms give different accuracy rates on working with different datasets. On working with the datasets from UCI repository, it is observed that KNN algorithm provides the greater accuracy and high performance rate than any other machine learning algorithms. The usage of this work provides a better solution than a medical suggestion from the specialized doctors, subject to the availability of medical logs of patients. We have supervised machine learning techniques such as support vector machine, K-Nearest Neighbors and confusion matrix which utilize the given input attributes to predict the probability of occurrence of breast cancer and identify the optimized algorithm among them.

Machine learning is a technique to derive insights from examples and experiences, while not being expressly programmed. Rather than writing code, we can feed knowledge to generic algorithmic program, and it builds required logic with the help of information fed. Machine leaning is a field of Artificial Intelligence (AI) that is derived out of computing. By applying AI, we have a potential to make higher and intelligent machines. There are three kinds of machine learning algorithms such as supervised Learning, unsupervised Learning and Reinforcement Learning. Machine learning is associated with data science that additionally focuses on prediction-making through the utilization of data sets. Within the domain of analytics, machine learning is a technique that devises advanced models and algorithms and lends themselves to prediction. In business use, this is often called prognostic analytics. By victimization, these two prospective Breast Cancer Diagnosis technique is meant to predict the probability of occurrence of breast cancer and to identify the best classification algorithm that can provide greater performance and accuracy for a given particular problem statement.

KNN stores all available attributes and classifies new attributes based on a similarity measure. The classification parameters and the number of neighbours i.e number of data points to be enclosed are provided as input. SVM helps in identifying an optimal separating hyperplane which maximizes the margin between different classes of the training data. The classification attributes are grouped as training set, test set and are fitted into SVC classifier to produce a hyperplane and to predict the accuracy score. The confusion matrix is a table, often used to provide an overview of the performance of a classifier on a set of test data for which the true and false values are known. It produces true and false value percentage based on comparison with actual and predicted value. It allows the visualization of the performance of an algorithm.

LITERATURE REVIEW

Different deliberate machine learning methods have been used to diagnose the breast cancer. Among those methods, few related works are discussed.

Amrane M et al. [1] have used six supervised classification techniques for the classification of breast cancer disease namely, K-Nearest Neighbors, Naive Bayes, Support vector machine, Random Forest, Desicion Tree and Linear Regression and evaluated the dataset for senstivity, specificity and total accuracy. The prediction on performance shows that Support Vector Machine shows the utmost performance accuracy rate of about 97.07%. However, Naive Bayes and Random Forest achieve the second utmost performance accuracy rate.

Bazazeh D et al. [2] work compares three of the most popular machine learning techniques commonly used for breast cancer detection and diagnosis, namely Support Vector Machine, Random Forest and Bayesian Networks. The Wisconsin original breast cancer data set was used as a training set to evaluate and compare the performance of the three machine learning classifiers in terms of key parameters such as accuracy, recall, precision and area of Receiver Operating Characteristic curve (ROC).

Divik Jain et al. [3] proposed healthcare applications still do not fit into many of the existing techniques as there are drawbacks in contexts such as non-availability of treatment, doctors and other resources and hence many people lose their lives. They have identified the potentials of Machine Learning (ML) usage for the development of robust healthcare system. It is believed that inherent capabilities of Machine Learning such as learning from experience, independency from human intervention etc., can play a vital role. Hence it concludes that how Machine Learning has been applied in various healthcare systems.

Islam M M et al. [4] presented a novel modality for the prediction of breast cancer and introduced with the Support Vector Machine and K-Nearest Neighbors which are supervised machine learning techniques for breast cancer detection by training its attributes. The proposed system uses 10-fold cross validation to get an accurate outcome. The performance of the proposed system is appraised considering accuracy, sensitivity, specificity, false discovery rate, false omission rate and Matthews correlation coefficient. The approach provides better results both for training and testing.

Jasmine Awatramani et al. [5] presented an overview to evolve the machine learning techniques in cancer disease by applying learning algorithms on breast cancer Wisconsin data - Linear Regression, Random Forest, Multi-layer Perceptron and Desicion Trees (DT). The result outcome shows that Multilayer perception performs better than other techniques.

Jini Ret al. [6] proposed a mammogram image retrieval technique using pattern similarity scheme. Comparing previous and current mammogram images associated with pathologic conditions are used to diagnose the real stage of breast cancer by doctors. In this work, the retrieval process is divided into four distinct parts that are feature extraction, KNN classification, pattern instantiation and computation of pattern similarity.

Lokanayaki K et al. [7] work highlights the prediction of unknown primary tumors in the dataset. The multiclass classifier with Random forest is used for classification of multicast dataset as it gives much higher accuracy than binary classifiers. Synthetic Minority Oversampling Technique (SMOTE) method for this imbalanced dataset with Randomize technique is applied during preprocessing for reducing the biasness among classes.

Maity M G et al. [8] study developed a machine learning prototype for the prognosis of dementia using rule-based forward chaining method. The results showed an accuracy value of 100%, which suggested that the prognosis has complied with the expert rules.

Pritom A I et al. [9] put a special emphasis on the Convolutional Neural Network (CNN) method for breast image classification. Along with the CNN method, they have also described the involvement of the conventional Neural Network (NN), Logic Based classifiers such as the Random Forest (RF) algorithm, Support Vector Machines (SVM), Bayesian methods, and a few of the semi-supervised and unsupervised methods which have been used for breast image classification.

Turgut S and et al [10] made an attempt to apply SSL through Multi-Modal Curriculum Learning (MMCL) strategy over medical images. Through this, medical images can be categorized into normal and abnormal images. Experimental results demonstrate good accuracy for classification.

Prior Work

Different breast cancer diagnosis and detection techniques have been created. As of late, a few investigations have proposed that different dataset provide best performance accuracy for different algorithms. Amrane M et. al. have developed a diagnosis system that employs six different data mining classification algorithms which provided the best performance algorithm based on its accuracy score but an optimization algorithm and a graphical comparision was lacked.

Bazazeh D et al proposed a system that compared Support Vector Machine, Random Forest and Bayesian Networks throughout the key-parameters namely acuuracy, recall and precision with Receiver Operating Characteristic Curve. The system utilized only three specific algorithms whereas there are many classification algorithms which provide better accuracy than those. Habib Dhahri et al proposed a system based on genetic programming technique which employed a single data mining algorithm to select the best features and perfect parameter values of the machine learning classifiers. It only classifies the stages of breast cancer as benign or malignant but does not provide any other accurate information such as accuracy and precision.

Ashwaq Qasem et al discussed a radio-pathalogical correlation system that used two technical methods such as mammography and histopathology. Manual detection and grading are tedious and so employed two techniques such as CADe and CADx that uses Machine Learning. CADe is a rejection model based on SVM algorithm that reduces the False Positive of the output. It also helps the medical experts in providing second opinion through precise detection.

PROPOSED METHOD

Diagnostic Analysis System Architecture

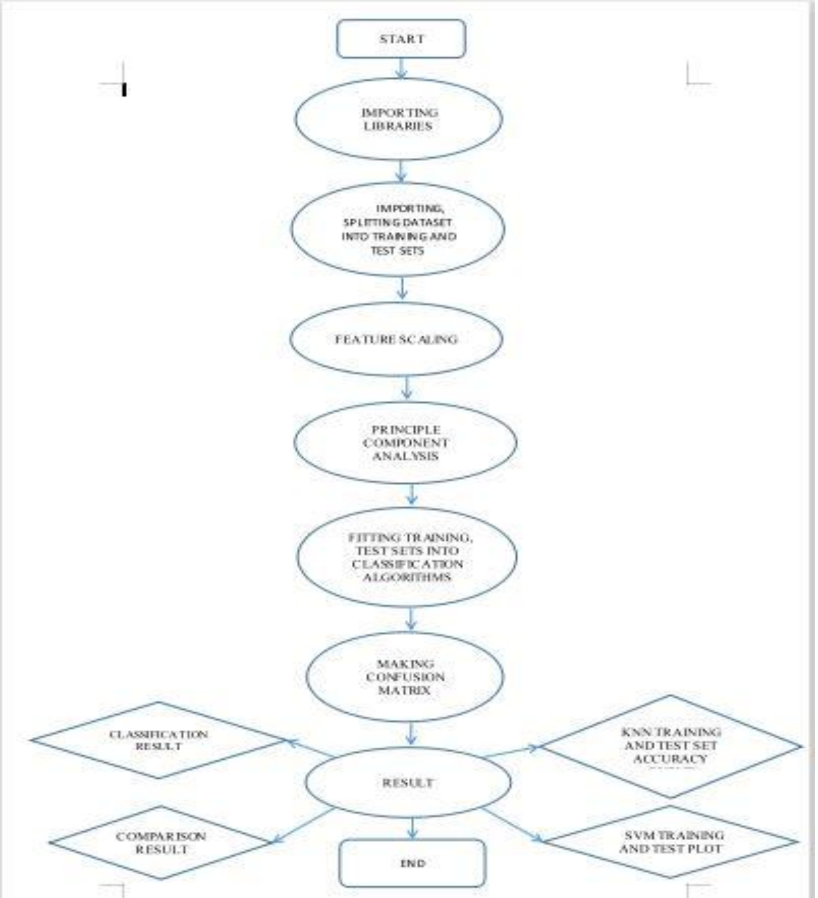
START
IMPORTING LIBRARIES
IMPORTING, SPLITTING DATASET INTO TRAINING AND TEST SETS
FEATURE SCALING
PRINCIPLE COMPONENT ANALYSIS
FITTING TRAINING, TEST SETS INTO CLASSIFICATION ALGORITHMS
MAKING CONFUSION MATRIX
CLASSIFICATION RESULT
KNN TRAINING AND TEST SET ACCURACY
RESULT
COMPARISON RESULT
END
SVM TRAINING AND TEST PLOT

Figure 1: Architecture.

In this system, the dataset has been imported and classified into training and test datasets. Similarly, the necessary libraries have been imported. Later, feature scaling and principle component analysis have done. Feature Scaling is done to pre-process the data i.e normalise the data within a particular range and it helps in speeding up of calculations in the algorithm. Principle Component Analysis is done to identify inter-relation between attributes of a dataset.

Machine Learning

Machine learning is a collection of methods that can automatically identify patterns in data, and then use those patterns to predict future outcomes, or to perform other types of decision making below certain conditions. Machine learning introduces various algorithms, those enable machines to understand the current situations and on the basis of that machines can take appropriate decisions. Machine learning works independently and takes decision at its own. The two main types of machine learning are supervised learning and unsupervised learning.

Supervised Learning: In supervised learning, the input and its corresponding output is already known. This is called supervised learning because it learns from training data set and creates model from it and when this model applied on new data set, it gives predicted results. Decision Tree, naive Bayes etc are the examples of supervised learning.

Unsupervised Learning: Unsupervised learning is where we have only input data and no corresponding output variable. The main job of unsupervised learning is to build up class labels automatically. The relationship between the data can be found using unsupervised learning algorithms to discover whether the data can characterize to form a group. This group is known as clusters. Unsupervised learning can be also described as cluster analyses".

Feature Scaling

The data mining algorithms that use Euclidean Distance measures are sensitive to Magnitudes. Thus feature scaling helps in weighing all the attributes equally.The attributes in the real world datasets varies highly in units, magnitudes and range. So, normalization plays an essential role. It reduces the irrelevance and misleading of attributes in a dataset. In a dataset, if a feature is compared in big scale to other features, then Euclidean distance is measured as a big scale in these algorithms. Then, the feature becomes highly dominating and has to be normalized. Initially, we observe the dataset is in an inconvenient form. The first feature to be noticed in a dataset is that its magnitude and range varies or differs a lot. For example, one feature measures in hundreds while the other measures in lakhs. In feature scaling, the Standard Scaler is used as a scaling algorithm. The Standard Scaler assumes that the dataset follows a normal distribution.

Principle Component Analysis

In principal component analysis, the relationship between the attributes is identified by finding a list of principal axes in the data. The principle component is a linear combination of normalized predictors of a dataset and it helps in feature reduction. It has both magnitude and direction. The principle axes are used to describe the dataset and its behavior is easy for visualization. It preserves the essential features with more variation by removing the non-essential features with lesser variation.

Algorithm 1

K- Nearest Neighbors

The K- Nearest Neighbors classification algorithm works on the fact that similar things tends to exist in close proximity. KNN stores all available attributes and classifies new attributes based on a similarity measure. It is widely useful for non-linear attribute dataset since it does not provide any of the assumptions about data in this algorithm. KNN algorithm, a type of supervised Machine Learning algorithm is used for both classification and regression problems.

Based on feature similarity property, the KNN algorithm assigns the value for new data points. The K-factor plays an essential role in identifying the data-point values. For accurate selection of k-factor, the algorithm has to be executed several times with different K values.The K -factor that reduces the number of error occurrence during predictions has to be selected for accurate calculations and measurements.

The classification parameters and the number of neighbours i.e number of data points to be enclosed are provided as input. The predicted results are displayed as its accuracy score. KNN algorithm provides result in terms of plots and classification error. Plots or graphs are simple, yet effective way of representing the results. It provides quick understanding and draws better attention compared to other representation. It plays a major role in depicting the actual comparison based on user defined attributes. Classification Error depends on the number of incorrectly classified samples and it is evaluated by the use of formula. It helps in increasing the accuracy rate.

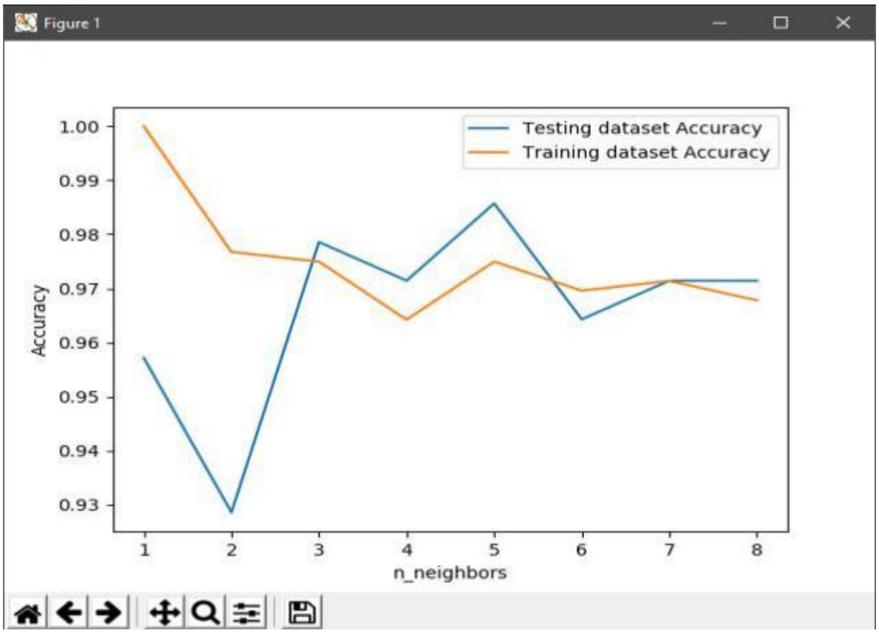

Figure 2: KNN Training and Test Set Accuracy Curve.

Algorithm 2

Support Vector Machine

Support Vector Machine, a supervised machine learning algorithm that is used for both classification and regression problems. Mostly, it is used in classification problems. In SVM algorithm, we plot each data item as a data point to a high-dimensional space with the value of each feature as particular coordinate value. An SVM model is basically a representation of different classes in a hyper plane and helps in identifying an optimal separating hyper plane which maximizes the margin between different classes of the training data. The classification attributes are grouped as training set, test set and are fitted into SVC classifier to produce a hyper plane and to predict the accuracy score.

Hyper plane is a line in a dimensional vector space which divides the space into two disconnected parts and helps to distinguish the different classes of data attributes. Support vectors, the data points whose position and orientation helps in building a hyper plane. Further, these points helps in maximizing the margin of the classifier.

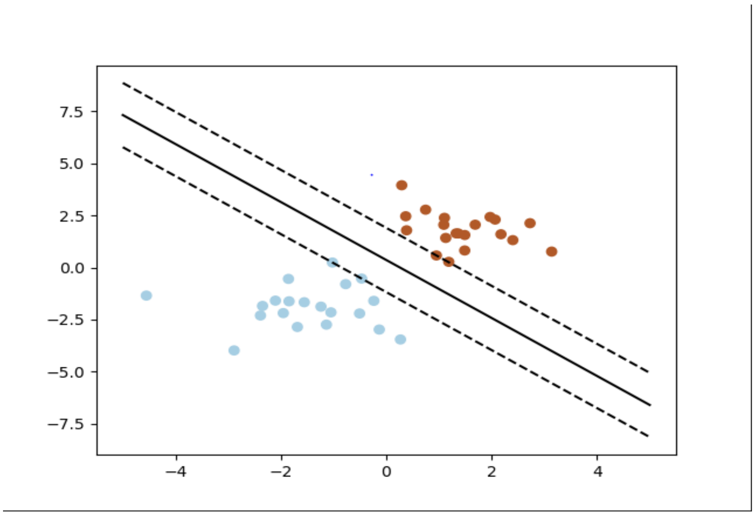

Figure 3: Hyper Plane.

Algorithm 3

Confusion Matrix

A table, often used to provide an overview of the performance of a classifier on a set of test data for which the true and false values are known. It produces true and false value percentage based on comparison with actual and predicted value. It allows the visualization of the performance of an algorithm.

It provides a summary of prediction results of a classification problem and provides the result in a table format. In the table, the number of correct and incorrect predictions is summarized with count values and is listed down with each class.

EXPERIMENTAL RESULTS

Our experimental results showed that KNN outperformed SVM. Moreover, there is a very marginal difference between the results obtained by SVM and KNN.

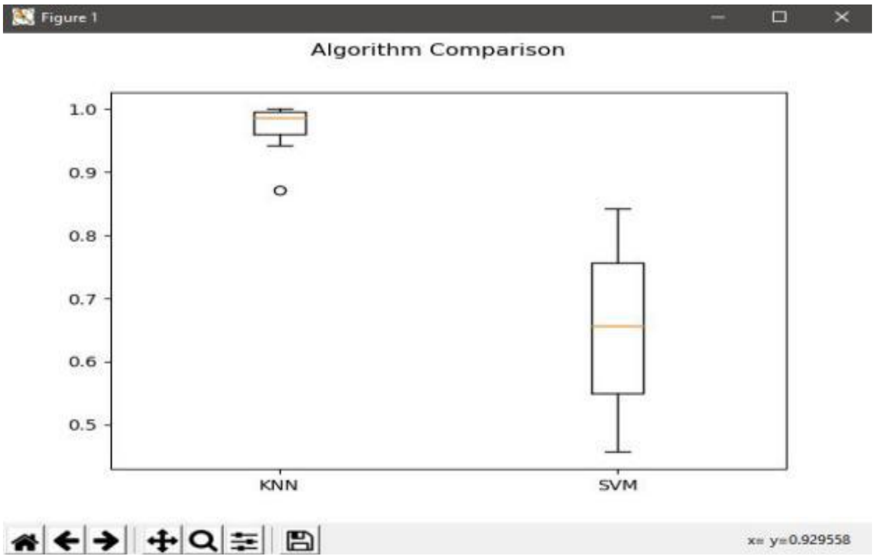

Figure 4: Hyper Plane.

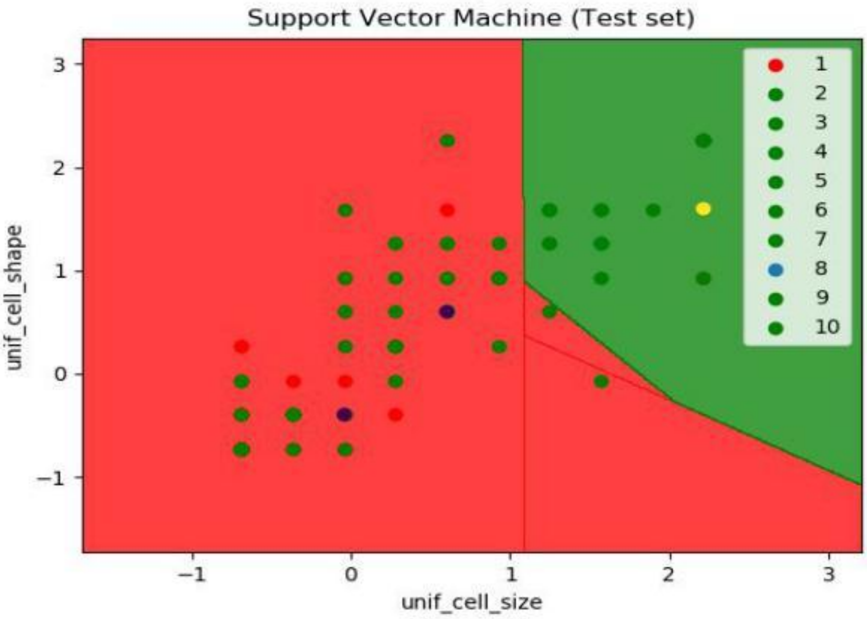

Figure 5: SVM Test Plot.

Accuracy

The accuracy obtained for KNN is 97.143% whereas for SVM, it is 96.753%. On comparing both the classification algorithm with the help of optimization algorithm, it is found that KNN has higher performance rate than SVM.

CONCLUSIONS

It is essential to build an intelligent and a trusted system that predicts breast cancer occurrence accurately based on the attributes such as unif_cell-size, unif_cell-shape, marg_adhesion and bare_nuclei and some domain knowledge of experts in the field. This project aims to predict the disease on the basis of the symptoms. The project is designed in such a way that the system takes symptoms as attributes from a data set or from the user as input and produces output i.e. predict the occurrence percentile of breast cancer. Average prediction accuracy probability of 97% is obtained. History about the user’s disease can be kept as a log and recommendations can be implemented for medications.

REFERENCES

1. *Amrane M, Oukid S, Gagaoua I,Ensari T, "Breast Cancer Classification Using Machine Learning Int. Conf. on Electric Electronics", Computer Science Biomedical Engineerings' Meeting, April 18–19, 2018.*
2. *Bazazeh D, Shubair R, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis", 2016 5th International Conference on Electronic Devices Systems and Applications, pp. 1-4, 2016.*
3. *Divik Jain, Brijesh Kadecha, Sailesh Iyer, "A Comparative Study of Machine Learning Techniques in Healthcare", Computing for Sustainable Global Development (INDIACom) 2019 6th International Conference on, pp. 455-460, 2019.*
4. *Islam M M, Iqbal H, Haque M R, Hasan M K, "Prediction of breast cancer using support vector machine and K-Nearest neighbors", 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), pp. 226-229, 2017.*

5. *Jasmine Awatramani, Nitasha Hasteer, "Early Stage Detection of Malignant Cells: A Step Towards Better Life", Computing Communication and Intelligent Systems (ICCCIS) 2019 International Conference on, pp. 262-267, 2019.*

6. *Jini R. Marsilin, G. Wiselin Jiji, "An efficient cbir approach for diagnosing the stages of breast cancer using knn classifier", Bonfring International Journal of Advances in Image Processing, vol. 2, no. 1, pp. 1, 2012.*

7. *Lokanayaki K and A. Malathi, "Exploring on Various Prediction Model in Data Mining Techniques for Disease Diagnosis", International Journal of Computer Applications,2013.*

8. *Maity M G, Das S, "Machine learning for improved diagnosis and prognosis in healthcare", 2017 IEEE Aerospace Conference, pp. 1-9, 2017.*

9. *Pritom A I, MunshiM A R, Sabab S A, Shihab S, "Predicting breast cancer recurrence using effective classification and feature selection technique", 2016 19th International Conference on Computer and Information Technology (ICCIT), pp. 310-314, 2016.*

10. *Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), 570-577.*

11. *Holmes, M. D., Chen, W. Y., Feskanich, D., Kroenke, C. H., & Colditz, G. A. (2005). Physical activity and survival after breast cancer diagnosis. Jama, 293(20), 2479-2486.*

12. *Holmes, M. D., Chen, W. Y., Feskanich, D., Kroenke, C. H., & Colditz, G. A. (2005). Physical activity and survival after breast cancer diagnosis. Jama, 293(20), 2479-2486.*

13. *Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), 570-577.*

14. *Turgut S,Dagtekin M,Ensari T, "Microarray Breast Cancer Data Classification Using Machine Learning Methods", Int. Conf. on Electric Electronics Computer Science Biomedical Engineerings' Meeting, April 18–19, 2018.*