BREAST CANCER DETECTION USING MACHINE LEARNING

Vanlalhmangaihsanga, P.C. Vanlalbeiseia, Laledenthara
CSE Student, CSE Student, CSE Student.
, Computer Science and Engineering, School of Engineering and Technology,
Mizoram University(MZU), Tanhril, Aizawl, 796009, Mizoram, India.

Abstract : In the contemporary world Breast Cancer has become the second leading cause of death among women. Early detection, assessment and followed by appropriate treatment of the cancer can reduce the deadly risk. Technology such as data mining and machine learning can substantially improve the diagnosis accuracy and reduce errors that can be made by medical professionals. This paper emphasizes on the algorithm on which machine can learn to accurately detect breast cancer. Machine Learning algorithm that are composed in this paper are Support Vector Machine, Logistic Regression, K-Nearest Neighbour Classifier, Decision Tree and Random Forest. The Algorithm will be tested accordingly with the World Breast Cancer (WBC) datasets. Benchmarking the aforementioned algorithm and picking the most efficient amongst the algorithm is the aim of this paper.

Keywords: Breast Cancer, Support Vector Machine, K-Nearest Neighbour Classifier, Logistic Regression, Decision Tree, Random Forest.

I. Introduction:

Breast Cancer is one of the most common type of cancer in the world. Although it can occur in man, it is much more common in woman and has become the second leading cause of death amongst women. With roughly 1 in 8 women developing breast cancer in their lifetime, the odds are high that nearly everyone is affected by this disease in some way. Early detection and appropriate treatment of the cancer is the most effective way to reduce the death among women due to breast cancer. This can be done through performing of different kind of diagnostic test like Magnetic resonance imaging (MRI), mammogram, ultrasound and biopsy.

Cancer is disease in which uncontrolled growth of abnormal cells which can divide and invade nearby tissues. Cancer cells can spread to all other parts of the body through blood and lymph systems if it is not controlled. Breast Cancer is a cancer that forms in the cells of the breast. Breast cancer most often begins with cells in the milk-producing ducts (invasive ductal carcinoma). Breast cancer may also begin in the glandular tissue called lobules (invasive lobular carcinoma) or in other cells or tissue within the breast. Researchers have identified hormonal, lifestyle and environmental factors that may increase your risk of breast cancer. But it's not clear why some people who have no risk factors develop cancer, yet other people with risk factors never do. It's likely that breast cancer is caused by a complex interaction of genetic makeup and environment. Also Doctors estimate about 5-10 percentage of breast cancer are linked to gene mutations passed through generations of family.

Breast Cancer Diagnosis is done by identifying the tumor. Tumors can be *benign* (noncancerous) or *malignant* (cancerous). Benign tumors tend to grow slowly and do not spread. Malignant tumors can grow rapidly, invade and destroy nearby normal tissues, and spread throughout the body. Unfortunately not all medical experts are able to distinguish between between benign and malignant tumor cells. Thus, we need a reliable diagnostic system that runs on an efficient algorithm that could predict the recurrence and non-recurrence of breast cancer from the patient with malignant cancer after the patient has undergone treatment. The available data in manual diagnosis is noisy and raw which increases the cost of management of data. Hence there is a need for proper parameter and feature selection provided to the algorithm of the machine for minimizing the error rate and cost of the diagnosis.

In this paper, SVM (SMO) model using linear and Gaussian kernels for separable and non-separable data are proposed respectively. We have implemented K- nearest neighbour where similarity criteria are Euclidean distance and Manhattan distance. Naïve Bayes and logistic regression are also implemented. Regularization parameter lambda is implemented in logistic regression for solving the problem Logistic regression and Logistic regression with regularization parameter, kNN – Euclidean and Manhattan measures, Naive Bayes. All these algorithms perform in a different manner, thereby giving different values of correct classification and prediction on both the datasets.

II. Motivation:

Globally stating, Breast Cancer is the most common cancer disease present in women. There is an estimate of 19.3 million new cancer cases and almost 10.0 million cancer deaths occurred in 2020. Breast cancer has surpassed lung cancer as the most commonly diagnosed cancer, with an estimated 2.3 million new cases (11.7%). The identification and development of Breast Cancer got interesting. The UCI Wisconsin Machine Learning Repository Breast Cancer Dataset drew attention as a sample set as substantial patients with variable attributes are present.

III. Literature Review:

The increasing health problem with consistent mortality rate concerning breast cancer has instigated many researchers to develop a more efficient and reliable diagnostic system.

Tarigopulla V.S Sriram [19] using the Parkinson disease dataset obtained from UCI depository dataset proposed SVM, kNN and Naïve Bayes techniques and made conclusions based on different values of accuracy.

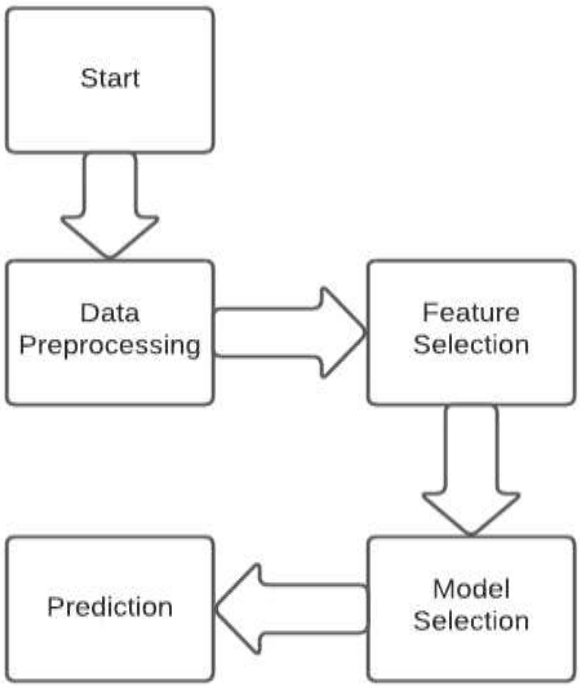
Also a work by S. Kharya [20] states that though the structure of artificial neural network is difficult to understand it has been the most widely used predictive technique in medical prediction. The limitations and advantages among the various machine learning techniques such as Decision trees, Naïve Bayes, neural network and SVM has been listed out in the paper.

From studying and clear observation the works of these previous researchers, we see that all the algorithms like SVM, kNN, Decision tree, etc. perform in a different manner, hence giving different values of correct classification and prediction on given datasets. And we can thus prepare aforementioned ML techniques to be implemented in breast cancer detection. Resulting in it being used in this paper. A list of a few literature studies associated with this subject is given below.

Table 1. Breast cancer detection research and accuracy using different machine-learning algorithm.

Paper Title	**Datasets**	**Algorithms**	**Results**
Breast Cancer Detection using Machine Learning[1]	WBC Mammogram	NN, SVM, KNN, DNN	ANN Accuracy:94.3% SVM Accuracy:85-91% KNN Accuracy:70-80% DNN Accuracy:96.3%
A study on prediction of breast cancer recurrence using data mining techniques [5]	WPBC	KNN, SVM, NB, c5.0 Classification. K-means, EM, PAM and Fuzzy c-means	Classification yields higher Accuracy than clustering. SVM and C5.0 Accuracy: 81%
Analysis of Breast Cancer Detection Using Different Machine Learning Techniques[6]	WBC Breast-Cancer	J48, SMO	J48 Accuracy: 75.52% SMO Accuracy: 96.99% After applying preprocessing techniques J48 Accuracy: 98.20% SMO Accuracy: 99.56%
Research Article on breast cancer detection: an application of machine learning algorithms on the Wisconsin diagnostic dataset[7].	WDBC	GRU-SVM, MLP, KNN, Softmax Regression, SVM	Accuracy for all used ML yielded test accuracy of above 90% with MLP standing out with 99.04%.
Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for breast Cancer Detection[8].	WDBC	Back-Propagation MSM-T Fuzzy Genetic GRNN Fuzzy + KNN Hybrid SVM	All ML yielded a test accuracy above 90% with Fyzzy+KNN and Hybrid SVM yielding 99,14% and 99.51%
Breast Cancer Daignosis by Different Machine Learning Methods using Blood Analysis Data[9].	Breast Cancer Data	ANN, ELM, KNN, SVM	ANN Accuracy: 79.43% ELM Accuracy: 80% KNN Accuracy: 77.5% SVM Accuracy: 73.5%

IV. PROPOSED METHODOLOGY:



We start by pre-processing the data using discretise filter and resampling it and then making it undergo k-fold cross validation, then features are selected to be trained or tested and a model is then used to evaluate aforementioned model's accuracy and other attributes. After all of these prediction can then be made.

4.1.Machine Learning Algorithms Used:

4.1.1. k-Nearest Neighbour (kNN):

k-Nearest Neighbour is one of the simpler algorithms, it is a supervised learning algorithm that is used to solve problems mainly categorized in classification and regression areas. With its ease of use, it comes with a few drawbacks of its own; Accuracy depending on quality of data, Sensitivity to large scale data and slow prediction, its computation being stored indefinitely makes it demanding and as such requires high memory which makes it highly demanding.

A generic kNN is often used for most average datasets to classify the means of a given cluster set. In this paper we are going to try to use it for the main purpose of generating predictions by training it for pattern recognition. kNN is a non-parametric method used for classification and regression. Both cases undergo training in a feature space, as kNN is instanced based learning[10]. In classification, output is given by selecting most votes done by neighbouring clusters of a given k, where k will have some arbitrary value or a fixed value based on the purpose. This method of extracting output from input data after transformation is known as Feature extraction.

A predominant procedure of how kNN Algorithm works is given below:

1)Assigning value to k, or giving it an arbitrary value.

2)Calculation is done, on both testing and training datasets.

3)Classification based on data received.

4)Finalizing results received and processing it as Output.

4.1.2 Random Forest Algorithm:

Another popular ML algorithm that falls under the supervised learning category, used in both Classification and Regression Problems. It is one of the more advanced ensemble learning, flexible algorithms[11]. It makes use of multiple decision trees to form a family of trees for classification methods[12], the output of this algorithm heavily relies on this simple yet multiplex and compounded method. And as such its robustness is seen when used in a large database. However, as good as it is in classification, its efficiency and advantage in this field falls off when used in regression problems.

A simple method of how Random Forest Algorithm works is given below:

1)From a given dataset, we specify random data points(k).

2)Constructing decision tress upon data points based on association as per needed.

3)Selecting N number for decision trees to be assembled.

4)Gathering particulars and details from N trees to forego prediction of new data, i.e assigning random data points and repeating construction of decision trees.

5)Taking prediction of each decision tree and reviewing the category and assigning it as a new data point, based on the majority vote of neighbouring trees.

4.1.3. Decision Tree:

In Decision Tree Algorithm, we have two nodes, the Decision Node and Leaf Node[13]. Decision Node being used for making decisions and erecting more multiple branches while Leaf Nodes are the output of aforementioned Decision Nodes. This Algorithm shows easy to understand output, a tree like structure resulting in effective consortium and interrelation among datasets and can be easily understood so it is usually associated with human thinking as it is quite easy to understand. Splitting, Pruning etc. are some

keywords users should familiarize with. As a result, they are quite easy to prepare and require less data cleaning. In the context of breast cancer detection, the nodes are usually required to be classified or de-label as Benign or Malignant.

Below is a given few steps involved in the working of this Algorithm:

1)Selecting a root node (S).

2)Laying out data set to probe the best attribute.

3)Splitting S into subsets and generating tree nodes, which contain the best attributes.

4)Repeating the process until the node cannot be further classified or split, resulting in the best outcome.

4.1.4. Support Vector Machine(SVM):

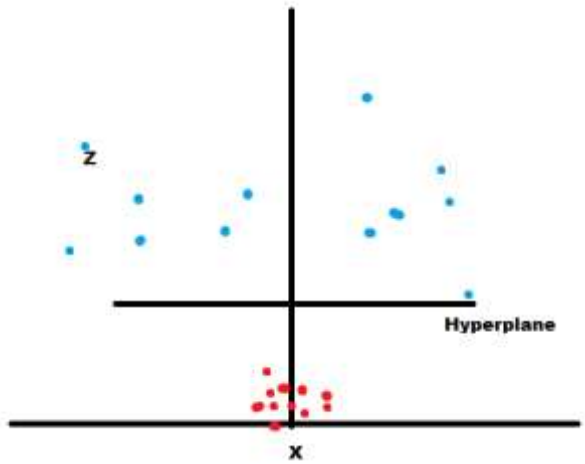
SVM is one of the most sought-after and favoured Supervised Learning Algorithm. Primarily used in Classification problems, end target and objective of it being to create a decision boundary that can set apart and isolate n-dimensional space into classes, putting them into contemporary data points in the correct category, this process being labelled 'hyperplane'[14]. Advantages it has over others are , but not limited to, memory efficiency, high dimensional space. Boasting its applicability on both Linearly separable and non-sparable data.

"*Generalized dot product or Kernel tricks are way of calculating dot product of two vectors to check how much they make an effect on each other. According to Cover's Theorem, the chances of linearly non-separable data sets have higher chances in higher dimensions*" 16]. So, we have linear and non-linear data types;

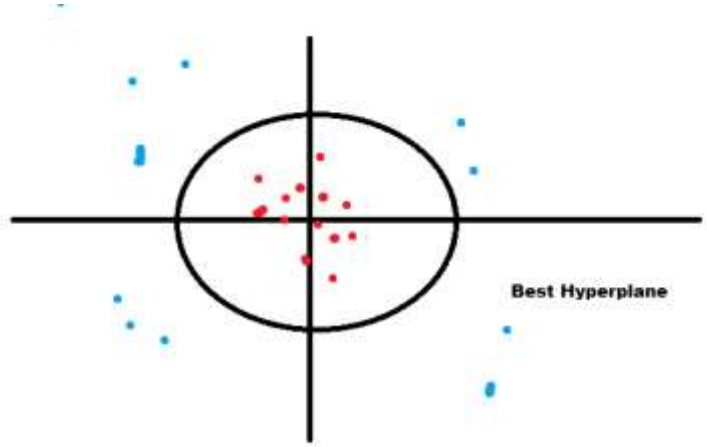
In Linear SVM, for simplicity's sake and for example purposes, let's say we have 2 different data sets that are laid out on a 2D plane and we want a classifier that can classify them each separately, a **Line** can be drawn across or in-between these datasets who have specific coordinates on the 2D plane. SVM assists in finding the most favourable line or decision boundary that can classify them apart. This line is coined, hyperplane. The points closest to the hyperplane is then used as a reference to draw another line parallel to the hyperplane, this is known as a support vector and the distance between them is termed margin. SVM tries to maximize this margin and the measurement with the highest value of margin is called Optimal Hyperplane[17].

In Non-Linear SVM, to separate data points, an additional dimension is required, i.e. a third dimension.

By adding a third dimension, we can have a much clearer picture.



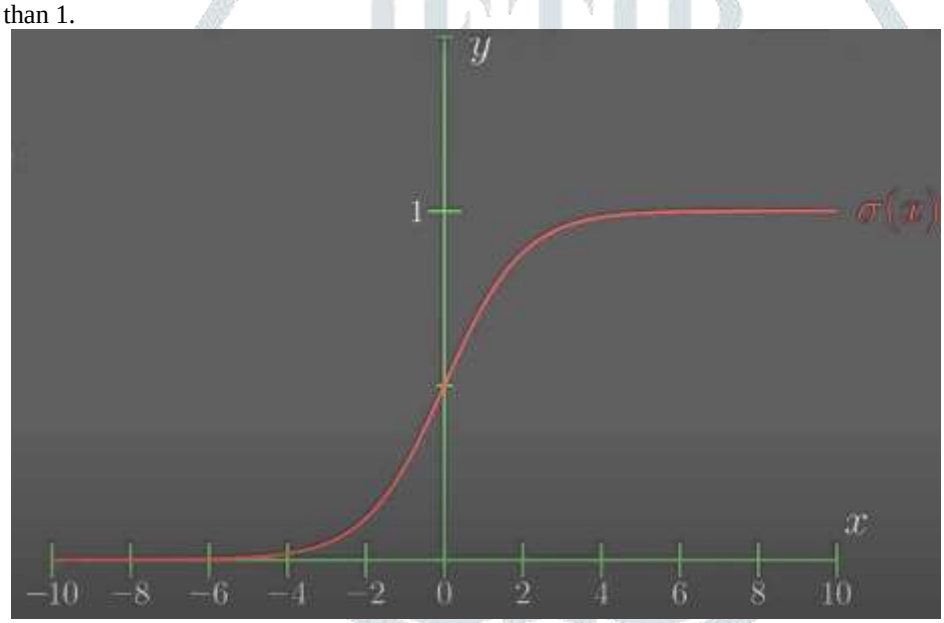
SVM divides the dataset into classes, and since it is in a 3D space, now, imagine the Hyperplane as a line in the above 3D graph as viewed from above, if we revert it back it into 2D space, then the Hyperplane will be given as:



4.1.5. Logistic Regression:

This Algorithm is used for predicting categorical dependent variable by making use of an acquired set of independent variables. And as such, it gives output as a categorical or discrete value. It can either be Yes/No, 1/0, True/False but it is known for giving probabilistic values ranging from 0 to 1. What is interesting about this algorithm is that, unlike the name implies, it is used to classify samples even though it uses the concept of predictive modelling as regression,

Since the value of the output given by logistic regression falls in line between 0 and 1, it forms an S-shape in a graphical representation. This is called a sigmoid function. Here for any value of x, Sigmoid function will always give an output no less than 0 and no more than 1.



V. RESULT AND DISCUSSION OF PROPOSED METHODOLOGY

The work was implemented on a i5 processor with 2.30 Ghz speed, 8 GB RAM, 1 TB of external storage and all tests and experiments on the classifiers and algorithms explained and laid out thus far in this paper have been conducted using libraries from JUPYTER notebook, Python 3, version 6.1.5 and using scikit learning machine. In this study, we have segregated and split apart our dataset in a 70:30 ratio. 70% for training and 30% for testing. We have used JUPYTER as it has a reputable collection of machine learning algorithms for pre-processing, clustering, classification and regression to choose from to be used on the dataset we have acquire.

In our work, we have applied k-fold cross validation test to estimate the skill of the model on a new data or unseen data set after training it on our own given dataset. It is a procedure used to evaluate proposed machine learning models on a custom data sample. It is simple to understand and its results are generally less biased and gives a more favourable and positive side estimate on the model. Course of action is as given below:

1)Dataset gets shuffled randomly.
2)Dataset is split into ‘k’ number of groups.
3)Each group gets taken out separately as test data set.
4)Remaining group is categorized and taken as training data set.
5)Takes a model and starts training it on the training data set.
6)After skimming through the training data set, model is evaluated on the test set.
7)Output or Evaluation score is then stored and the model is discarded.
8)Skill of the model is then summarized using the Model Evaluation scores.

5.1.Preprocessing Phase:

Using discretized filter, data is discretized, missing values removed from the dataset and then using resample filter on instances and then using a k-fold cross validation, the experiments are then carried out.

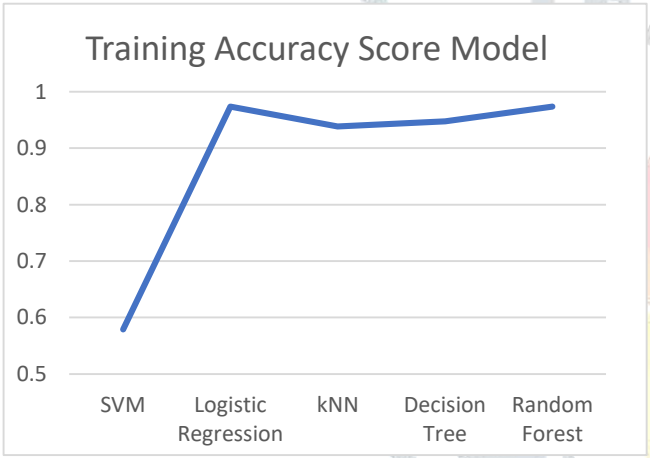
After pre-processing and preparation methods are applied, we try to evaluate and analyse data and figure out the output to categorize them in terms of accuracy, sensitivity, specificity and precision.

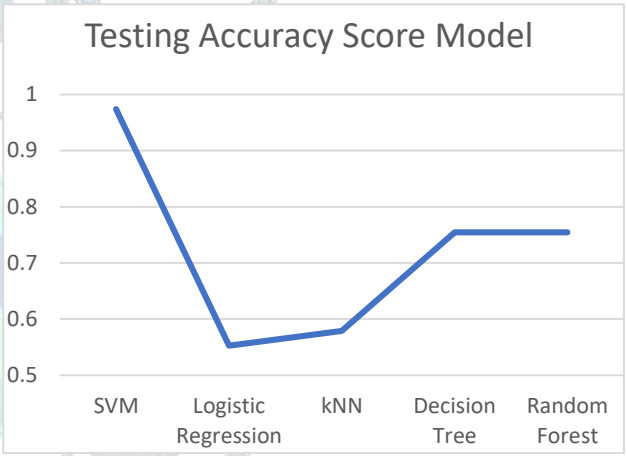
The experiments consisted of three main steps which were data gathering, data pre-processing and performance evaluation. The dataset used in this experiment was acquired and gathered from Breast Cancer Wisconsin Dataset, obtained from the work of a collaboration of The University of Wisconsin hospital, Madison between Dr. William H. Wolberg of the department of Surgery and Human Oncology and Prof. Olvi L. Mangasarian of the Computer Sciences Department[18]. The data has 699 instances with 10 attributes plus the class attributes. The dataset has been used in various machine learning algorithms to evaluate their attributes on diagnosis and prognosis of Breast Cancer.

5.2.EFFICIENCY:

Effectiveness of all classifiers is evaluated, each based in terms of accuracy, time taken to build the model and classification error.

Algorithm	Accuracy	Sensitivity(%)	Specificity(%)
k-Nearest Neighbour	0.9385	92.8	98.8
Random Forest	0.9736	93.8	97.7
Decision Tree	0.9476	92.5	94.3
Support Vector Machine	0.5789	95.7	98.6
Logistic Regression	0.9736	95.4	99.1





VI. CONCLUSION AND FUTURE WORK:

We can now make a conclusion based on the results we have taken in. Referring to Table 2. and Figure 1 and Figure 2. We can see that k-NN and Logistic Regression have insultingly low accuracy in the training process, unlike other models that we have built. Whereas the accuracy obtained by SVM is 96%, making it better than both Decision Tree and Random Forest classifier, which both have an accuracy of 75%. SVM also performed far better in terms of classification error, i.e, correctly classified instances and incorrectly classified instances.

After training the model and evaluating efficiency of the algorithms on the training dataset we can now deploy it on the testing data set. Surprisingly Logistic Regression and Random Forest classifier came up top with both boasting an accuracy of 97% while SVM felled off to 57%, Decision tree and kNN had an accuracy in the range of 93% and 94% respectively. Based on just accuracy value Logistic Regression and Random Forest classifier are on the same level. But adding the fact of efficiency, Sensitivity and Specificity we can conclude that **Random Forest Classifier** performs far superior.

VII. FUTURE WORK:

Observation and Inspection of the result denotes that use of multidimensional data on different classification models and feature selections can provide more robust and promising tools for use in this domain. For better performance of the proposed and used ML languages, more research should be implemented and prioritized to predict more variables and increase accuracy and precision.

VIII. REFERENCES:

[1] R.Chtihrakkannan, P.Kavitha, T. Mangayarkarsi, R.Karthikeyan: Breast Cancer Detection using Machine Learning; IJITEE ISSN:2278-3075, Volume 8, Issue 11, September 2019.

[2] "Latest Global Cancer Data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018", International Agency for Research on Cancer. World Health Organization 12 September 2018.

[3].M.M.Mehdy, E.E.Sharir and P.Y.Ng, "Artificial Neural Networks in Image processing for Earlier Detection of Breast Cancer" Hindawi, Computational and Mathematical Methods in Medicine, Volume 2017, Article ID 2610628.
[4]. Moh'd Rasoul A Al-Hadidi, MohammedY. Al-Gawagzeh. "Solving Mammography Problems of Breast Cancer Detection Using Artificial Neural Networks and Image Processing Techniques". Indian Journal of Science and Technology, Vol.5, No.4 (April 2012) ISSN:094-6846.
[5]. Jha U., Goel, S.: A study on prediction of breast cancer recurrence using data mining techniques. In: 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence, IEEE, pp. 527–530, 2017
[6].Analysis of Breast cancer Detection Using Different Machine Learning Techniques, Siham A. Mohammed, Sadeq Darrab, Salah A. Noaman, Gunter Saake. Conference Paper, 11 July 2020, CCIS, Volume 1234.
[7]. Abien Fred M. Agarap, Research Article on breast cancer detection: an application of machine learning algorithms on the Wisconsin diagnostic dataset, ICMLSC '18: Proceedings of the 2nd INternational Conference on Machine Learningand Soft Computing
[8]. IJRTE, ISSN: 2277-3878, Volume-8, issue-2s3, July 2019: Breast Cancer Detection using Machine Learning Way, Sri Hari Nallamala, Pragnyaban Mishra, Suvarna Vani Koneru.
[9]. International Journal of Intelligent Systems and Applications in Engineering, ISSN:2147-6799, Breast Cancer Diagnosos by Different Machine LEarning Methods using lood Analysis Data; Muhammet Fatih Aslan, Yunus Celik, Kadir Sabanci, Afik Durdu.
[10]. Detection of Cancer in Lung with kNN Classification Using Genetic Algorithm; P. Bhuvaneswari, Dr A Brintha Therese/ Procedia Materials Science10(2015) 433-440.
[11]. Random Forest for Breast Cancer Prediction; T.L. Octaviani, Z.Rustam. Department of Mathematics, Faculty of Mathematics and Natural Sciences(FMIPA), Universitas Indonesia. AIP Conference Proceedings 2160, 020050(2019)
[12]. Random Forest Algorithm, Javapoint. https://www.javatpoint.com/machine-learning-random-forest-algorithm.
[13]. Diagnosis of Breast Cancer using Decision Tree Models and SVM; Puneet Yadav, Rajat Barshney, Vishan Kumar Gupta. International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395-0056. Volume 05 Issue 03. March 2018.
[14]. SVM Algorithm, javapoint: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm.
[15]. Early detection of Breast Cancer using SVM Classifier Technique; Y. Iraeneus Anna Rejani. Dr. S Thamarai Selvi. International Journal on Computer Science and Engineereing Vol. 1(3) 2009, 127-130.
[16]: AnalyticSteps; How Does Support Vector machine (SVM) Algorithm work in Machine Learning?; Rohit Dwivedi.
[17]: Support Vector Machine Alogrithm; javapoint.
[18]. Machine Learning for Cancer Diagnosis and Prognosis; Dr. Wolberg
[19]. Tarigoppula V.S Sriram, M. Venkateswara Rao, G V Satya Narayana, DSVGK Kaladhar, T Pandu Ranga Vital - Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms, International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 3, September 2013.
[20]. S.Kharya, D. Dubey, and S. Soni - Predictive Machine Learning Techniques for Breast Cancer Detection, IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (6), 2013, 1023-1028.