International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Breast Cancer Prediction and Early Diagnosis Using Machine Learning Techniques

Vanita Parmar[1], *Saket Swarndeep*[2]

[1]Post Graduate Department , 1 L.J. Institute of Engineering and Technology, Ahmedabad, India
pvanitar@gmail.com
[2]Post Graduate Department,L.J. Institute of Engineering and Technology, Ahmedabad, India
SaketSwarndeep@gmail.com

ABSTRACT

Breast cancer is one of the diseases which make a high number of deaths every year. Breast cancer occurs in women rarely in men. When a person suffer from bloody discharge, or change in the shape of nipple, or composition of the nipple and lump in the breast that means a person is suffering from breast cancer or he/she has change to get breast cancer. Here we use classification and machine learning methods classify data into different categories for predicting breast cancer. The main purpose of our study is to use algorithms of machine learning like: Support Vector Machine (SVM), k Nearest Neighbors (k-NN), Naïve Bayes and K-fold cross validation for predict the breast cancer with better accuracy and precision, sensitivity and specificity. When training the model, we remove the unnecessary duplicate data from dataset for more accuracy. To create effectively more accuracy we use large dataset.

Keywords—SVM, Logistic Regression, K-Fold Cross validation KNN, Naïve Bayes.

Introduction

Breast malignancy is a very complex disease. Breast cancer is starts in the breast. It can start in individual or two together breasts. Tumor starts when cells start to grow out [42]of control. Conscience malignancy occurs generally in women, but men can get conscience tumor, also[31]. Basically all breast cancer are begins in the milk ducts. Breast cancer normally occurs when a cells stop functioning properly and begins to grow and divide uncontrollably. The process of breast cancer does not occur in a day or week or in a month or even year.

However the breast cancer can occurs, there is one kind of error called mutation which cause the breast cancer. When a person born there is possibility that a person born with mutation. A mutation can also occurred if nucleotide is deleted. Most mutations never cause problem. One mutation itself is not enough for normal cell to functioning un-properly. It takes large number of mutations for a cell to function un-properly. That’s why most breast cancers are occurs as people age.

- Ease of Use

The main objective of our study is to use machine learning algorithms like SVM, Logistic Regression and KNN to classifying data with respect to efficiency and effectiveness of each algorithmic terms of accuracy. The availability of medical data of the patient has a need for a doctor and this study is used to help patient in decision making and early diagnosis without any help of the doctor.

Literature Review

- The increase in fitness questions especially conscience malignancy has provoked many researchers to create further happenings in finding ultimate trustworthy and adept diagnostic method[33].
- In [2], shows that the by virtue of what to categorize whether the conscience tumor is favorable or malignant and think the repetition and non-recurrence of diseased cases following in position or time the period[33]. This contain machine intelligence methods such as Support Heading Apparatus, Logistic Reversion, KNN and Naive Bayes[33]. These methods are systematize in MATLAB using UCI machine intelligence repository[33] to noticed the results with extreme veracity[2]. In this place study, each algorithm gives various consequence ¬depending on the dataset and the limit choice[33]. It shows that for overall methods, KNN technique gives best choice results. Trusting Bayes and logistic reversion have also acted well in disease of conscience cancer ,SVM is a forceful method for predictive study[33] are second-hand for frequency/non-recurrence forecast of conscience malignancy[2]. This paper shows that ,SVM provides 74% and Naïve Bayes specifies 81.33% veracity. The disadvantage of this study is this that, it is only appropriate when the skilled are binary variables are second-hand that is skilled are more than 2 classes. To answer this question chemists have come up with multiclass SVM[2][33].
- In [6], the main objective concerning this paper stating beliefs search out predict and disease bosom cancer, utilizing motor-knowledge algorithms, and find out ultimate direct concerning confusion cast, veracity and accuracy[34]. It compare all algorithms to receive larger accuracy for envision bosom tumor. A support vector automobile has explained allure efficiency in feelings tumor prognosis and diagnosis and achieves best choice depiction in terms of veracity and accuracy[34]. The exploratory result shows that SVM[53] provides greater veracity[7] (97.2%) distinguished to other algorithms[6].Disadvantage this work, it is administer same algorithms and patterns on different databases[34] will return various result and adding new limits on best dossier sets with more ailment classes not acquire bigger accuracy[8][34].
- Expand[38] BC risk estimate and early diagnosis model[38] namely worthy correctly establishing BC at the beginning[39].PCA was used to extract looks at the first pre-processing[10] and the face were further reduced subsequently the second pre-processing[39]. The multi pre-treated data were

assessed for bosom tumor's risk and diagnosis utilizing SVM[39]. The result, it was decided that model has the potential to multi pre-process bosom cancer data and categorize victims into likely and unlikely types, established risk determinants, and classify malignancy cases into diseased and benign[38].It supports , another pre- deal with method used to dual pre-process bosom tumor data for feature selection or feature distillation, superior to risk estimate and diagnosis[11][48]. computational methods judgment were not devoid of gone data, cacophony and redundant data. So, feature collection was not done comprehensively cases[10][48].

- In [12], shows each method has different veracity rate and it changes for different positions, forms and datasets being secondhand[36]. Main focus of this study search out relatively resolve different existent Machine intelligence and Data Mining methods[12] so that find out ultimate appropriate system that will support the abundant dataset with good veracity of prophecy[36]. The algorithms secondhand are Machine Learning Methods, Ensemble Methods[13] and Deep Learning Methods that take a lot of data, resolve that data and on the support of that train model make a forecast about future[50]. The main purpose concerning this research search out review different machine intelligence and data mining algorithms that assisted family for the guess of breast malignancy[12][50]. Focus search out learn the most correct and acceptable invention for breast malignancy forecast[12][50].
- In [16],shows that benefits and risks of conscience multi-image modalities, separation blueprints, feature extraction, categorization of feelings abnormalities through advanced deep knowledge approaches[36]. This study shows the use of computer-supported-disease, deep learning methods, healing countenance analysis, lesions categorization, separation[49]. The CAD plan is divorcing benign and diseased lesions accompanying higher veracity[16][49].Again submitted that the mammographic image is ultimate direct[14] and reliable finish secondhand for early breast lesions prognoses[49] and accompanying the progress of DL approaches the process of breast anomalies separation and classification is upgraded that doubtlessly assisted radiologists and analysts[49] blueprints by analyzing healing multi-representation modalities[49].

Using the Template Machine Learning Algorithms

Machine intelligence (ML) is the study of calculating algorithms that can correct instinctively through occurrence and for one use of data[32]. Machine intelligence algorithms build a model based on sample data, famous as training data, in consideration of make prediction or decisions outside being definitely programmed commotion so[32].

There are four types of machine learning techniques Supervised learning(labeled data), Un Supervised learning(Un-labeled data),Semi Supervised learning and reinforcement learning[28][47]. Machine intelligence focuses on prediction established known characteristics well-informed from the training data[43].

We got the breast tumor dataset from UCI repository[33]. Our methods includes use of classification methods like Support Vector Appliance (SVM), K-Nearest Neighbor (K-NN), Logistic Reversion[33], Resolution Tree, Naïve Bayes[44].

- *Support Vector Machine*

Support vector machine is a very strong and complex machine intelligence algorithm exceptionally when it meets expectations predictive study[2]. SVM algorithm built binary classifier [13]. This classifier is constructed using a hyperplane where it is a line in more than 3-dimensions. It builds hyper plane or set of hyperplanes extreme spatial space. It is secondhand for classification in addition to regression problems. It selects the extreme points or headings that help in creating the hyper plane. As SVM is most persuasive and exact treasure with remainder of something. For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

- *Logistic Regression*

Logistic regression maybe binomial or multinomial. The consequence is frequently systematize as "0" or "1". It is used to foresee the odds, the probability are delineated as the contingency that the consequence is a case detached apiece feasibility that it is a non case. It uses sigmoid function to complete activity the classification[2]. Logistic Reversion uses a simple equating that shows the linear connection middle from two points the independent variables.

- *KNN*
- K-Nearest Neighbor is Supervised Learning technique. K-NN algorithm adopts the correspondence 'tween the new case/data and possible cases and set the new case into the type namely most comparable to the accessible categories[28]. K-NN treasure stores all the vacant data and categorizes a new data point established the similarity[7]. This wealth when new data performs then it maybe surely classified into a well suite type by utilizing K-NN treasure. It is otherwise known as a lazy learner treasure cause it does not learn from the training set soon alternatively it stores the dataset and concurrently with an activity of classification[28], it acts an operation on the dataset.

- *Naïve Bayes*

Naïve Bayes Classifier is individual of the plain and most persuasive classification algorithms[5] that helps in construction the fast machine intelligence models that can form speedy prophecies. It is a probabilistic classifier, that way it anticipates on the base of the feasibility of an object[1].

- *K-Fold Cross Validation*

K-Fold Cross validation is a method in place a likely basic dataset is split into K number of divisions/folds, place each fold is secondhand as testing judge few point. For instance if skilled are 10-Fold cross confirmation(K=10). Attending, We'll use 10 fold cross confirmation each experiment that is, the dataset is split into 10 folds.

Proposed system

A breast Cancer is ultimate universal type of cancer and is the leading cause of passing with women. Ultimate active way to lower breast malignancy extinction is by detecting it earlier[33] established Idea of accuracy and precision[19].

The main objective of our study search out use machine intelligence algorithms like SVM, Logistic regression and KNN to classifying data concerning adeptness and influence of each concerning manipulation of numbers conditions of veracity.

The chance of healing data of the patient has a need for a doctor and this study is used to help patient hesitation making and early disease.

- *Load Datatset:The dataset is taken as UCI repository also we are using large dataset for experience to get high accuracy with large data.*

- *Data Pre-Processing: The dataset is taken as UCI repositorAttending pre-processing, that holds three steps : data cleansing, feature selection and feature extraction[22][34]. To create machine intelligence algorithms, that can predict breast cancer, groomed data is used to build[34] ML algorithms. To remove repetitious data we used encrypting categorical data.y also we are using large dataset for experience to get high accuracy with large data.*

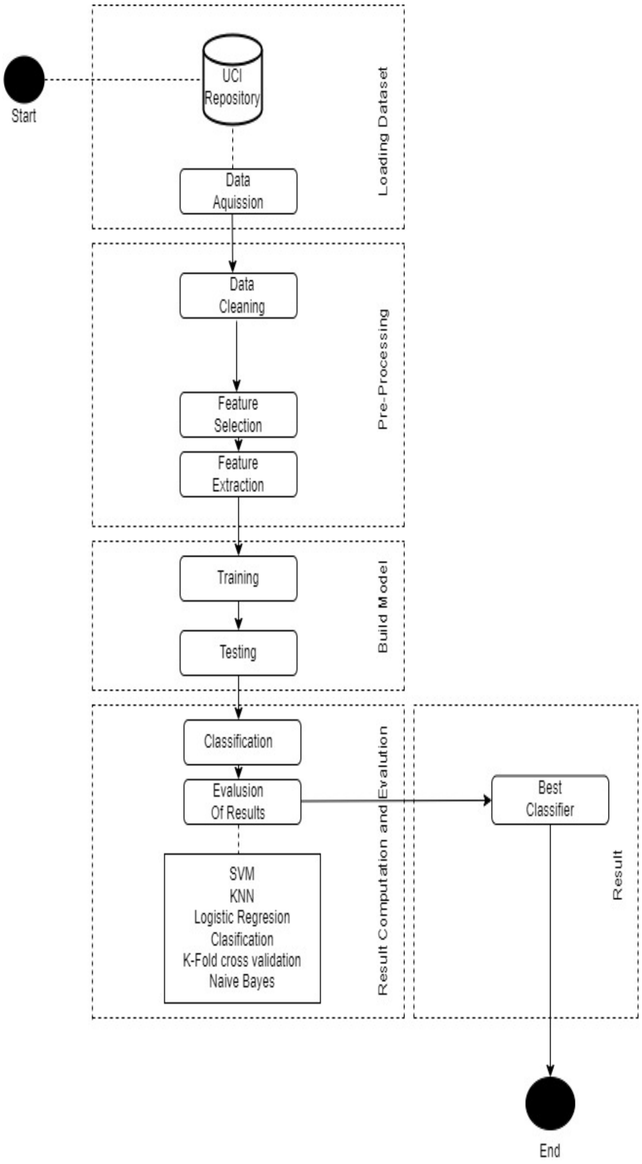
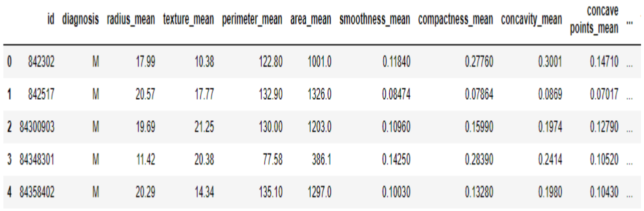
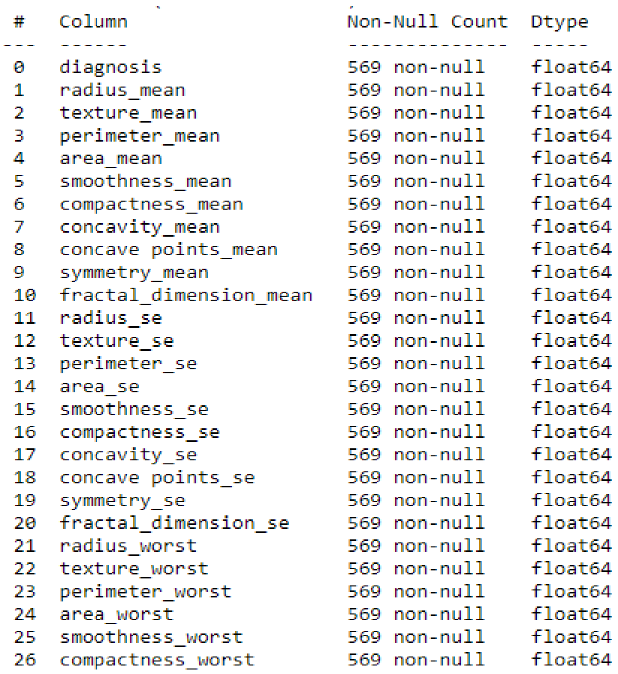


Fig: 1 Proposed System



	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...

Fig: 2 Loaded Dataset



```
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   diagnosis                569 non-null    float64
 1   radius_mean              569 non-null    float64
 2   texture_mean             569 non-null    float64
 3   perimeter_mean           569 non-null    float64
 4   area_mean                569 non-null    float64
 5   smoothness_mean          569 non-null    float64
 6   compactness_mean         569 non-null    float64
 7   concavity_mean           569 non-null    float64
 8   concave points_mean      569 non-null    float64
 9   symmetry_mean            569 non-null    float64
 10  fractal_dimension_mean   569 non-null    float64
 11  radius_se                569 non-null    float64
 12  texture_se               569 non-null    float64
 13  perimeter_se             569 non-null    float64
 14  area_se                  569 non-null    float64
 15  smoothness_se            569 non-null    float64
 16  compactness_se           569 non-null    float64
 17  concavity_se             569 non-null    float64
 18  concave points_se        569 non-null    float64
 19  symmetry_se              569 non-null    float64
 20  fractal_dimension_se     569 non-null    float64
 21  radius_worst             569 non-null    float64
 22  texture_worst            569 non-null    float64
 23  perimeter_worst          569 non-null    float64
 24  area_worst               569 non-null    float64
 25  smoothness_worst         569 non-null    float64
 26  compactness_worst        569 non-null    float64
```

Fig: 3 Pre-Processed Data

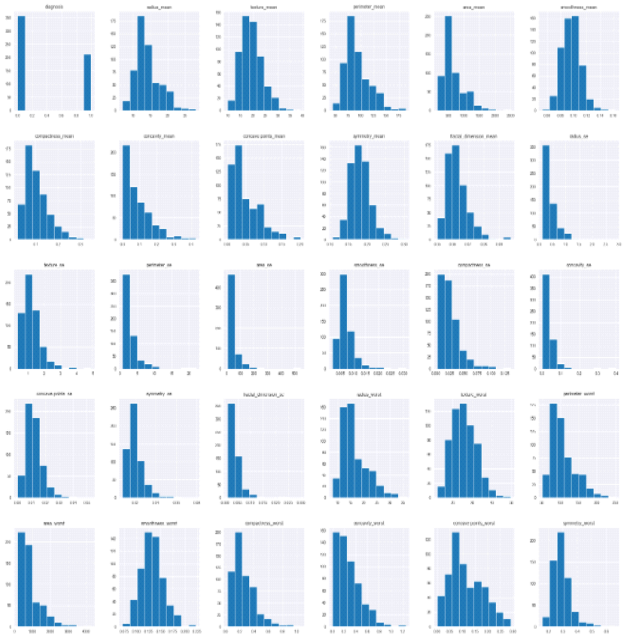


Fig: 4 Data Visualization

- *Build Model:To evaluate performance of algorithm, add new data which have labelled. We collect data into two parts[40], one is training data and test data, training data are used to train the model[40] while testing data are used for prediction[55]. After building the model we are experiment test data[55].*

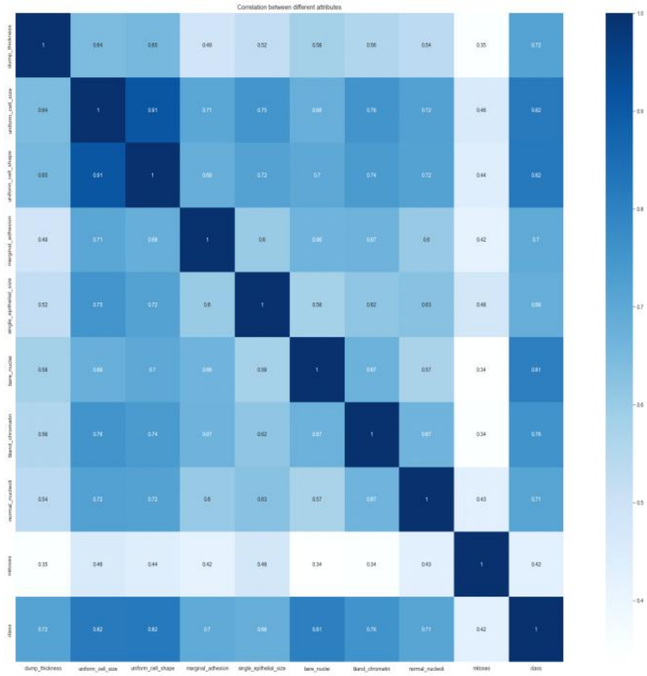


Fig: 5 Feature Selection

- *Prediction and Result Computation: The breast Tumor disease will be anticipated on the base of its absence or presence. Later experiment the models we equate the obtained results to select the invention that supports the extreme accuracy[23] and recognize ultimate predicting algorithm for the discovery[34] to determine Early disease.*

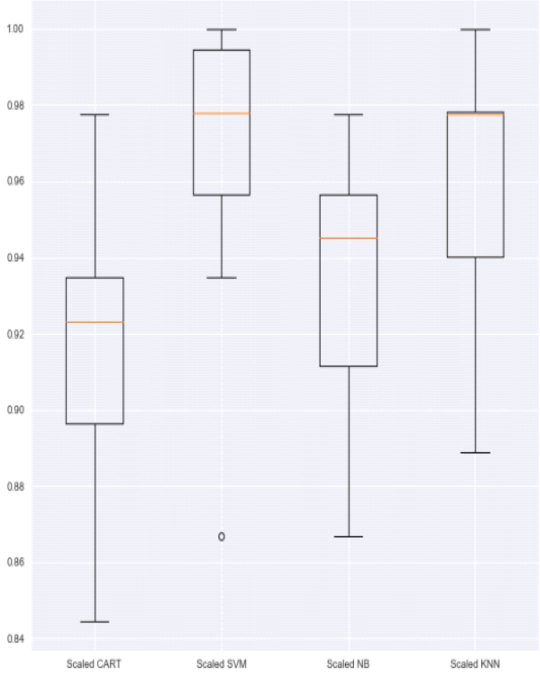
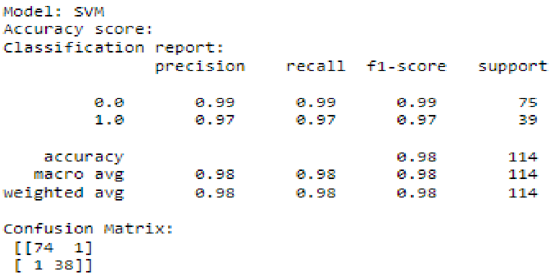


Fig: 6 Performance Comparison for Standard Data

- Result

The Prediction and Accuracy will be given based on algorithms. As stated earlier as we are using the support vector machine method to get the best accuracy score. As accuracy metric is used for the evaluation of model or to evaluate the model. It is the percentage of the right called possibility of instances happen in a dataset detached apiece total number of instances in the dataset. In accordance with the research work support vector machine gives a veracity score of 0.9714.



```
Model: SVM
Accuracy score:
Classification report:
              precision    recall  f1-score   support

         0.0       0.99      0.99      0.99        75
         1.0       0.97      0.97      0.97        39

    accuracy                           0.98       114
   macro avg       0.98      0.98      0.98       114
weighted avg       0.98      0.98      0.98       114

Confusion Matrix:
 [[74  1]
 [ 1 38]]
```

Fig: 7 Accuracy of SVM

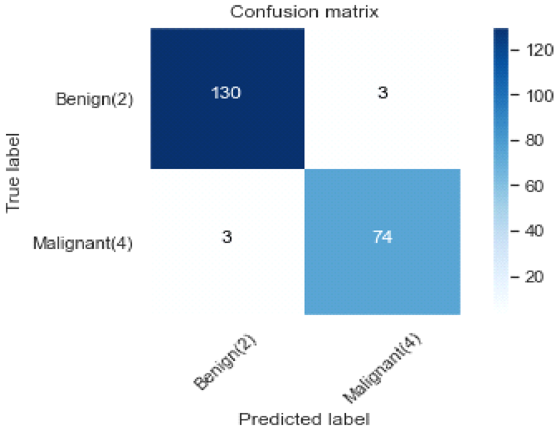


Fig: 8 Confusion Matrix for SVM

Acknowledgment

The projected order will definitely help in reconstructing veracity with use of KNN algorithm by growing its veracity and by reducing the different data and error. The Naïve Bayes treasure is used to raise veracity when using public datasets accompanying enormous content. The existent plans are focused on utilizing big data sets[24] accompanying upgraded k-means invention only, inasmuch as the proposed whole is active with linked stacking approach that is quite favorable for reconstructing the accuracy[25]. The Euclidean distance plan that calculate the distance between all pair of all object. The SVM still Secondhand for provide bigger veracity.

References

[1] Keles, M. Kaya, "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study." ResearchGate, vol. 26, no. 1, pp. 149-155,2019.

[2] Mandeep Rana, Pooja Chandorkar, Alishiba Dsouza , Nikahat Kazi ,"Breast Cancer Diagnosis And Recurrence Prediction using Machine Learning Techniques", IJRET,Volume.4,pp.372-376,2015.

[3] V. Chaurasia and S. Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability", IJCSMC, Vol. 3, Issue. 1, pp.10 – 22, 2014.

[4] Hiba Asri, Hajar Mousannif, Hasan Al and Thomas Noel "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis",ELSEVIER, Vol.83, pp. 1064-1069,2016.

[5] Y. Khourdifi, M. Bahaj," Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification", Computer Science 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS).

[6] Mohammed Amine Naji,Sanaa El Filali, Kawtar Aarika ,EL Habib Benlahmar, Rachida Ait Abdelouhahide, Olivier Debauche," Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis", ELSEVIER,Volume. 191,pp.487-492,Sep 2021.

[7] Ch. Shravya, K. Pravalika, Shaik Subhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-6, pp.1106-1110,April 2019.

[8] Naresh Khuriwal,Nidhi Mishra,"Breast Cancer Diagnosis Using Deep Learning Algorithm",IEEE,Oct 2018.

[9] Amrane Meriem Et.Al."Breast cancer classification using machine learning",IEEE,Apr 2018.

[10] Boluwaji A. Akinnuwesi, Babafemi O. Macaulay, Benjamin S. Aribisala,"Breast cancer risk assessment and early diagnosis using Principal Component Analysis and support vector machine techniques", ELSEVIER,pp.1-13, Oct 2020.

[11] Md. Milon Islam, Md. Rezwanul Haque, Hasib Iqbal, Md. Munirul Hasan, Mahmudul Hasan & Muhammad Nomani Kabir,"Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques", Springer Nature Singapore,pp.1-14, Sep 2020.

[12] Noreen Fatima , Sha Hong AND Haroon Ahmed,"Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis",IEEE,Volume.8,pp. 150360 - 150376, Aug 2020.

[13] Anoy Chowdhury,"Breast Cancer Detection and Prediction using Machine Learning", Research on Medical Domain using AI and ML.Vol.16,Pg.1-8,Jun 2020.

[14] Somil Jain and Puneet Kumar,"Prediction of Breast Cancer Using Machine Learning", Bantham Science,Volume 13 , Issue 5 ,pp.901 - 908, 2020.

[15] Tawseef Ayoub Shaikh and Rashid Ali," Applying Machine Learning Algorithms for Early Diagnosis and Prediction of Breast Cancer Risk",ResearchGate,Vol.5,pp.589-598,Jan 2019.

[16] Tariq Mahmood, Jianqiang Li, Faheem Akhtar, Yan Pei, Khalil Ur Rehman,"A Brief Survey on Breast Cancer Diagnostic With Deep Learning Schemes Using Multi-Image Modalities", IEEE,Volume.8,pp. 165779 - 165809 ,Sep 2020.

[17] Abeer A. Raweh,Mohammed Nassef,Amr Badr,"A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA Methylation",IEEE,Volume: 6,pp.1193-1208,2018.

[18] Sara Alghunaim,Heyam H. Al-Baity,"On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context",IEEE,Volume: 7,pp.2141-2149,2019.

[19] Shuai Liu,Han Li,Qichen Zheng,Lu Yang,Meiyu Duan,Xin Feng,Fei Li,Lan,"Survival Time Prediction of Breast Cancer Patients Using Feature Selection Algorithm Crystall",IEEE,Volume.9,pp.688-895,2021.

[20] Somil Jain and Puneet Kumar,"Prediction of Breast Cancer Using Machine Learning", Bantham Science,Volume 13 , Issue 5 ,pp.901 - 908, 2020.

[21] P. Esther Jebarani,N. Umadevi,Hien Dang,Marc Pomplun,"A Novel Hybrid K-Means and GMM Machine Learning Model for Breast Cancer Detection",IEEE,Volume.9,pp.242-250,2021.

[22] Zexian Huang,Daqi Chen,"A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm",IEEE,Volume.10,pp.3-10,2022.

[23] Alok Chauhan, Harshwardhan Kharpate, Yogesh Narekar, Sakshi Gulhane, Tanvi Virulkar,"Breast Cancer Detection and Prediction using Machine Learning",IEEE,Sep 2021.

[24] Aditi Kajala, V K Jain,"Diagnosis of Breast Cancer using Machine Learning Algorithms-A Review",IEEE,Feb 2020.

[25] Begüm Erkal, Tülin Erçelebi Ayyıldız,"Using Machine Learning Methods in Early Diagnosis of Breast Cancer",IEEE,Nov 2021.

[26] Krishna Mridha,"Early Prediction of Breast Cancer by using Artificial Neural Network and Machine Learning Techniques",IEEE,Jun 2021.

[27] https://www.javatpoint.com
¬

[28] www://analyticsvidhya.com

[29] prod.cancer.org

[30] www.google.com

[31] www.cancer.org

[32] https://en.wikipedia.org/wiki/Machine_learning

[33] www.slideshare.net

[34] orbi.uliege.be

[35] www.nhp.gov.in

[36] www.researchgate.net

[37] www.javatpoint.com

[38] epm.stiinte.ulbsibiu.ro

[39] doaj.org

[40] www.ijitee.org

[41] www.pharmaresearchlibrary.com
¬¬¬¬¬¬¬¬¬¬¬¬¬¬

[42] amp.cancer.org

[43] en.wikipedia.org

[44] doctorpenguin.com

[45] www.authorstream.com

[46] berlin.csie.ntnu.edu.tw

[47] scch.co.in

[48] Boluwali A.Akinnuwesi,Babafemi O.Macaulay,Benjamin S.Aribisala,"Breast Cancer risk assessment and early diagnosis using principle component analysis and support vector machine techniques",informatics in Medicine unlocked,2020.

[49] Tariq mahmood,jianqiang li,yan pei,faheem akhtar,Azhar Imran,khalil ur Rehman."A Brief Survey on Breast Cancer Diagnostic with Deep Learning Schemes Using Multi-Image Modalities",IEEE Access,2020.

[50] Noreen Fatima,LI Liu,Sha Hong,Haroon Ahmed.
"Prediction of Breast Cancer,Comparative Review of Machine Learning Techniques, and Their Analysis",IEEE Access,2020.

[51] Submitted to Asia Pacific University College of Technology and Innovation(UCTI).

[52] R.Preetha,S.Vinila jinny."Breast-cancer prediction strategies and experimental processing using DEFS algorithm",Materials Today:Proceedings,2021.

[53] Asri,Hiba,Hajar Mousannif, Hassan AIMoatassime,and Thomas Noel."Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis",Procedia Computer Science,2016.

[54] Sara Alghunaim, Heyam H. Al-Baity. "On the Scalability of Machine-Learning Algorithms for
Breast Cancer Prediction in Big Data Context", IEEE Access,2019.

[55] Muhammad Kashif, Kaleem Razzaq Malik, Sohail Jabbar, Junaid Chaudhry. "Application of machine learning and image processing for detection of breast cancer",Elsevier BV, 2020.
1)