WWW.IJRPR.COM

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 -A Survey

[1]VanitaParmar R.,[2] Prof. SaketSwarndeep J.

[1]Student, Post Graduate scholar,Post Graduate Department, L.J.Institute of Engineering and Technology, L J University
[2] Assistant Professor, L.J.Institute of Engineering and Technology, L J University

ABSTRACT

Breast cancer is one of the diseases that make a high number of deaths every year.Here we use cclassification and machine learning methods classify data into different categories for predicting breast cancer.The main purpose of this study is to use machine learning algorithms like: Support Vector Machine (SVM), Logistic Regression, Random Forest and k Nearest Neighbours (k-NN),Naïve Bayes to predict the breast cancer with better accuracy and precision, accuracy, sensitivity and specificity.

Keywords: SVM, Logistic Regression ,Decision Tree, Random Forest, KNN, Naïve Bayes.

INTRODUCTION

Breast cancer is starts in the breast. It can start in one or both breasts. Cancer starts when cells begin to grow out [42]of control. Breast cancer occurs almost entirely in women, but men can get breast cancer, too[31].

TYPES:

Firstly, to go with types of Breast Cancer we need to know about different parts of the breast from where the breast cancer starts[42]. Breast cancers can start in different parts of the breast[47].

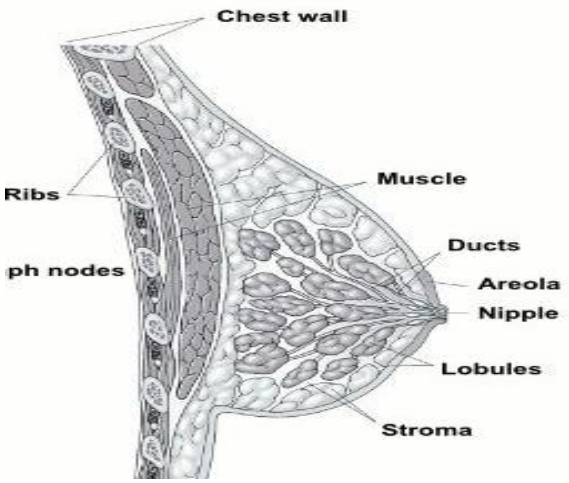

Fig1: Parts of Breast From where Breast cancer starts[31]

- Lobules are the glands that make breast milk. Cancers that start here are called lobular cancers.
- Ducts are small canals that come out from the lobules and carry the milk to the nipple. Cancers that start here are called ductal cancers.
- The nipple is the opening in the skin of the breast where the ducts come together and turn into larger ducts so the milk can leave the

breast[51]. The nipple is surrounded by slightly darker thicker skin called the areola. Cancer that start here are called Paget disease of the breast.

- The fat and connective tissue (stroma) surround the ducts and lobules and help keep them in place. Cancer that start here are called phyllodestumer[31].

Symptoms of Breast Cancer[41]

Symptoms of breast cancer include a lump in the[41] breast, bloody discharge from the nipple and changes in the shape or texture of the nipple or breast[30][45].

Reduce Breast Cancer Risk

Researchers continue to look for medicines that might help lower breast cancer risk, especially women who are at high risk using Estrogen blocking, Tamoxifen, raloxifene used for many years to prevent Breast Cancer[30].

Many of the researchers are also provide many usefulsolutions to predict Breast cancer using machine learning algorithms[52] for early diagnosis[30].

Risk Factors

- Family History: Women whose mother or sister had breast cancer carry a higher risk of developing this disease[35].
- Breast lumps- Women who have some type breast lumps have a higher risk of breast cancer later on[35].
- Dense breast tissue -Women with thicker breast tissue[35] are more likely to develop breast cancer.
- Age[10]: Therisk of breast cancer increases as women get older.
- Diet and lifestyle choices[34]: Women who smoke, eat high fat diet, drink alcohol are more at risk of developing breast cancer[35].
- Radiation Exposure: Frequent exposure to X-Rays and CT scans may raise a women's chance of developing breast cancer[35].
- Obesity: Women which are overweighed are also have a risk of developing breast cancer[35].
- Estrogen exposure –Womenwho[35] start menstruating earlier have a higher risk of developing breast cancer[35].
- Due to pregnancy over older age or miscarriages can also lead the breast cancer.

PREVIOUS WORK

The increase in health problems especially breast cancer has instigated many researchers to make further developments in finding the most reliable and efficient diagnostic system[33].

In [2],shows that the how to classify whether the breast cancer is benign or malignant and predict the recurrence and non-recurrence of malignant cases after a certain period[33]. This include machine learning techniques such as Support Vector Machine, Logistic Regression, KNN and Naive Bayes[33]. These techniques are coded in MATLAB using UCI machine learning depository[33] to observed the results with high accuracy[2]. In this study, each algorithm gives different outcomedepending on the dataset and the parameter selection[33]. It shows that for overall methodology, KNN technique gives the best results. Naive Bayes and logistic regression have also performed well in diagnosis of breast cancer ,SVM is a strong technique for predictive analysis[33] are used for recurrence/non-recurrence prediction of breast cancer[2]. This paper shows that ,SVM provides 74% and Naïve Bayes provides 81.33% accuracy.The limitation of this study is this that, it is only applicable when the there are binary variables are usedi.e. there are more than 2 classes. To solve this problem scientists have come up with multiclass SVM[2][33].

In [6],the main objective of this research paper is to predict and diagnosis breast cancer, using machine-learning algorithms, and find out the most effective with respect to confusion matrix, accuracy and precision[34]. It compare all algorithms to get higher accuracy for predict breast cancer.A ssupport vector machine has demonstrated its efficiency in breast cancer prediction and diagnosis and achieves the best performance in terms of accuracy and precision[34]. The experimental result shows that SVM[53] provides higher accuracy[7] (97.2%) compared to other algorithms[6].Limitation this work, it is apply same algorithms and methods on other databases[34] will return different result and adding new parameters on larger data sets with more disease classes not obtain higher accuracy[8][34].

In [10],In this study, they used Principal Component Analysis (PCA) and Support Vector Machine (SVM)[9] to develop[38] BC risk assessment and early diagnosis model[38] that is capable of accurately establishing BC at the early stage[39].PCA was used to extract features at the first preprocessing[10] and the features were further reduced after the second preprocessing[39]. The multi pre-processed data were assessed for breast cancer's risk and diagnosis using SVM[39]. The result, it was concluded that model has the potential to multi pre-process breast cancer data and classify patients into likely and unlikely categories, based on risk factors, and classify cancer cases into malignant and benign[38].It provides , another pre- processing technique used to dual pre-process breast cancer data for feature selection or feature extraction, prior to risk assessment and diagnosis[11][48].computational techniques evaluation were not devoid of missing data, noise and redundant data. Thus, feature selection was not done in all cases[10][48].

In [12],shows every technique has different accuracy rate and it varies for different situations, tools and datasets being used[36]. Main focus of this study is to comparatively analyze different existing Machine Learning and Data Mining techniques[12] in order to find out the most appropriate method that will support the large dataset with good accuracy of prediction[36].The algorithms used are Machine Learning Techniques, Ensemble Techniques[13] and Deep Learning Techniques which take large number of data, analyze that data and on the basis of that train model make a prediction about future[50].The main purpose of this research is to review different machine learning and data mining algorithms that helped people for the prediction of breast cancer[12][50]. Focus is to find out the most accurate and suitable algorithm for breast cancer prediction[12][50].

In [16],shows that benefits and risks of breast multi-imaging modalities, segmentation schemes, feature extraction, classification of breast abnormalities through state-of-the-art deep learning approaches[36]. This study shows the use of computer-aided-diagnosis, deep learning techniques, medical image analysis, lesions classification, segmentation[49]. The CAD system is separating benign and malignant lesions with higher accuracy[16][49].Also suggested that the mammographic image is the most effective[14] and reliable tool used for early breast lesions prognoses[49] andwith the advancement of DL approaches the process of breast abnormalities segmentation and classification is improved which truly assisted radiologists and researchers[49] schemes by analyzing medical multi-image modalities[49].

METHODOLOGY

Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data[32]. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do 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 learning focuses on prediction.based on known properties learned from the training data[43].

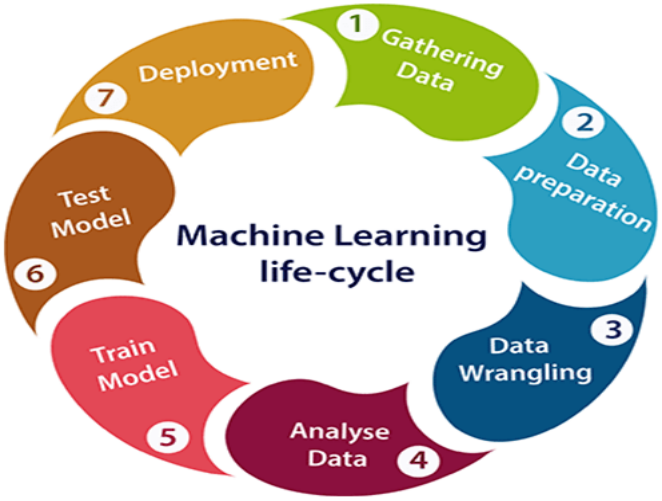

Fig: 2 Machine Learning Life cycle[30][37]

Machine learning has given the computer systems the abilities to automatically learn without being explicitly programmed[37]. So, it can be described using the life cycle of machine learning. Machine learning life cycle is a cyclic process to build an efficient machine learning project. The main purpose of the life cycle is to find a solution to the problem or project[27][37].

We obtained the breast cancer dataset from UCI repository[33]. Our methodology involves use of classification techniques like Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Logistic Regression[33], Decision Tree, Naïve Bayes[44].

Support Vector Machine(SVM)

Support vector machine is a very strong and sophisticated machine learning algorithm especially when it comes to predictive analysis[2]. A binary classifier is built by the SVM algorithm [13]. This binary classifier is constructed using a hyper plane where it is a line in more than 3-dimensions.

Logistic Regression

Logistic regression can be binomial or multinomial. The outcome is usually coded as "0" or "1". It is used to predict the odds, the odds are defined as the probability that a particular outcome is a case divided by the probability that it is a non case. Logistic regression technique uses sigmoid function to carry out the classification[2]. Logistic Regression uses a simple equation which shows the linear relation between the independent variables.

Random Forest

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems[27]. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

KNN

K-Nearest Neighbor is Supervised Learning technique.K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories[28].K-NN algorithm stores all the available data and classifies a new data point based on the similarity[7]. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification[28], it performs an action on the dataset.

Naïve Bayes

Naïve Bayes Classifier is one of the simple and most effective Classification algorithms[5] which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object[1].

PROPOSED SYSTEM

Breast Cancer is the most common type of cancer World Wide and is the leading cause of death among women. The most effective way to reduce breast cancer deaths is by detecting it earlier[33] based on Concept of accuracy and precision[19].

- As far as research is done so far, small large data set is used in SVM with KNN algorithm which provide use of complex data with multiple classes and provide higher accuracy.
- Accuracy according to initial machine learning techniques.
- Sensitive to noise and outliers, so a small number of such data can cause of higher error rate.
- Provide the use of same algorithm with different database with same result

There are four stages in proposed system are shown below,

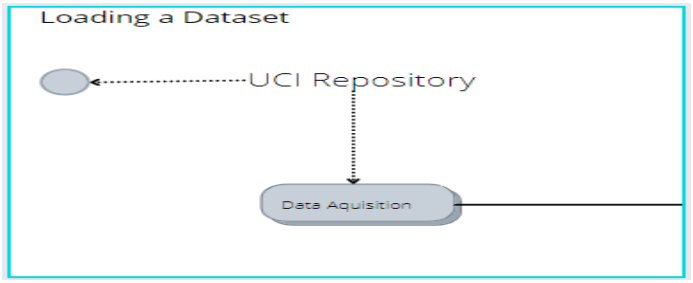

Fig :3 (a) Loading Dataset[29]

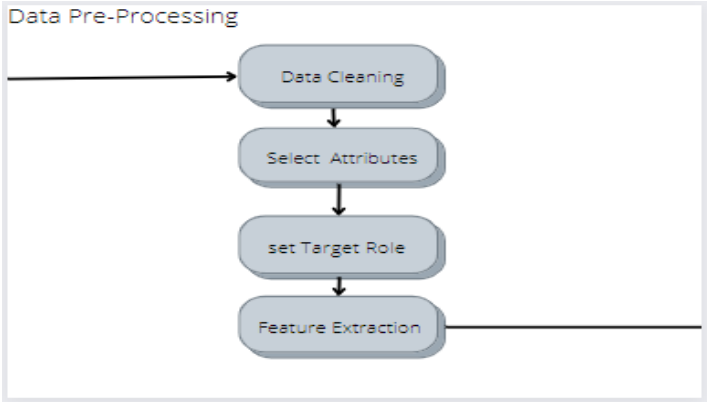

Fig: 4 (b) Data Pre-Processing[29]

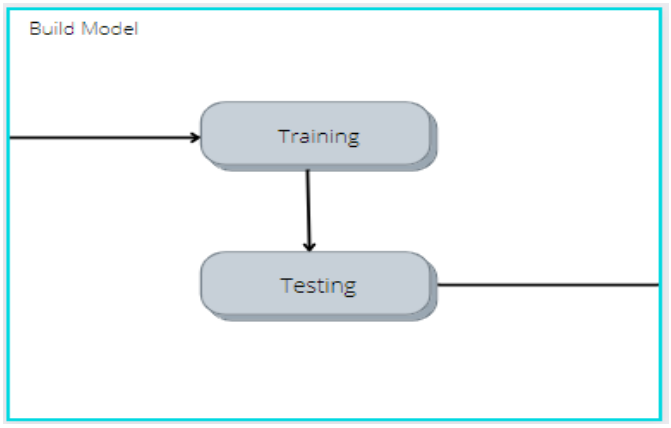

Fig: 5 (c) Build Model[29]

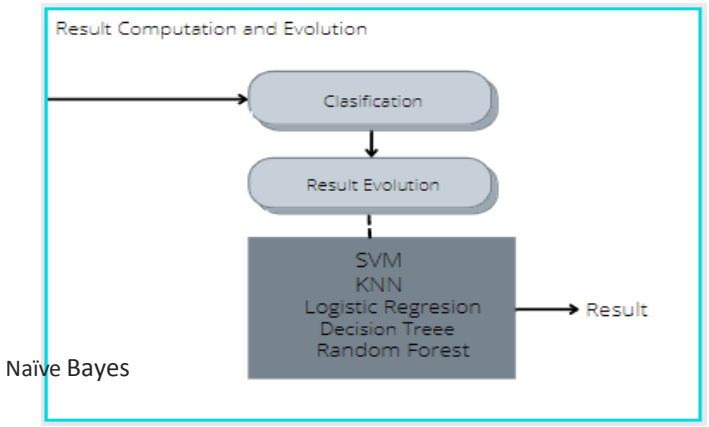

Fig: 6 (d) Result Computation and Evaluation[29]

Stage 1: Loading the Dataset.
The dataset is taken as UCI repository.

Stage 2: Data Pre-Processing.
Here pre-processing, which contains four steps : data cleaning, select attributes, set target Role and features extraction[22][34]. To create machine learning algorithms, that can predict breast cancer, prepared data is used to build[34]ML algorithms.

Stage 3: Build the 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.

Stage 4: Prediction and Result Computation and Evolution.
The Breast Cancer disease will be predicted on the basis of its absence or presence. After testing the models we compare the obtained results to select the algorithm that provides the high accuracy[23] and identify the most predictive algorithm for the detection of breast[34] to provide Early diagnosis.

Stage 5: Result
The Prediction and Accuracy will be given based on algorithms.

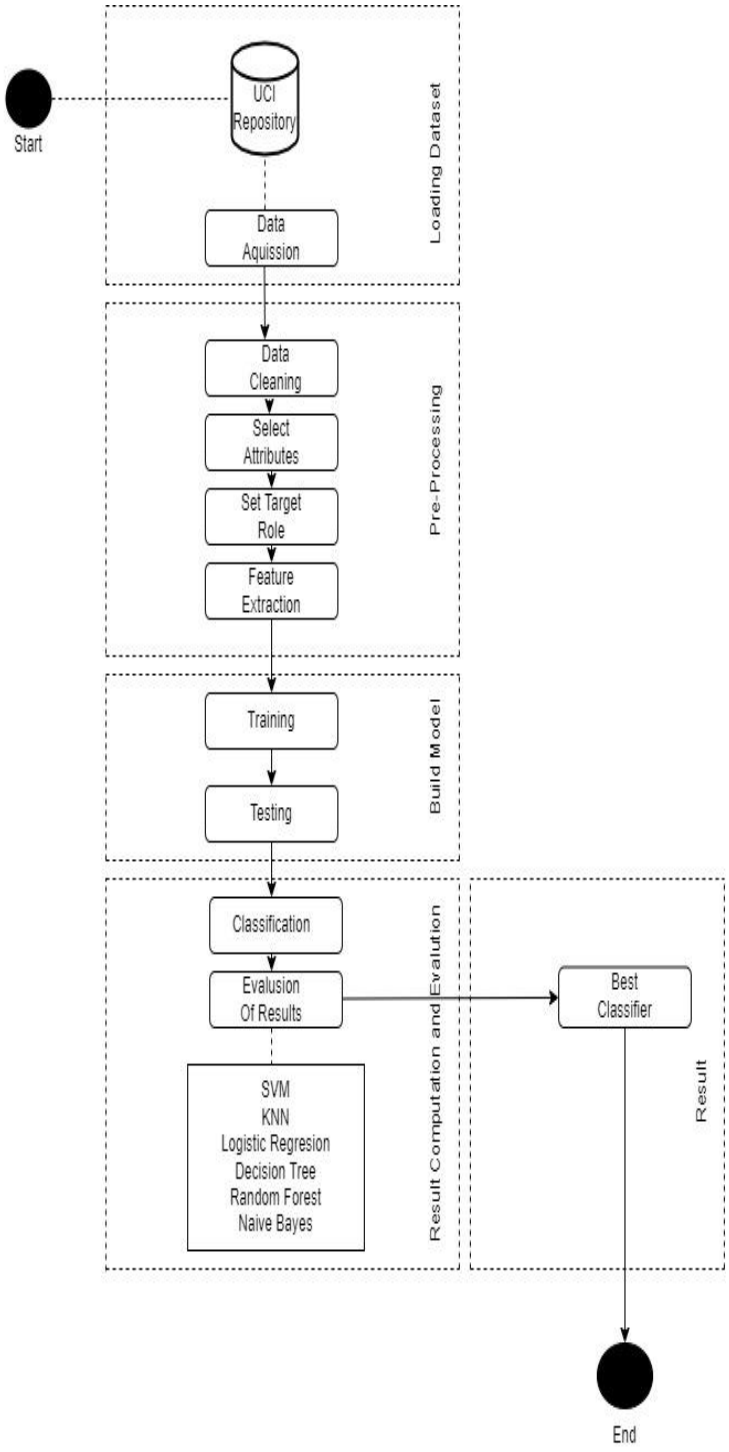

Fig: 7 Proposed System[29

]

CONCLUSION

The proposed system will definitely help in improving accuracy level with use of KNN algorithm by increasing its accuracy and by reducing the unusual data and error.The Naïve Bayes algorithm is used to improve accuracy when using public datasets with huge content. The existing systems are focused on using large data sets[24] with improved k-means algorithm only, whereas the proposed system is working with combined stacking approach which is quite advantageous for improving the accuracy[25]. The Euclidean distance method that compute the distance between every pair of all object. The SVM also Used for provide higher accuracy.

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] MandeepRana, PoojaChandorkar, Alishiba Dsouza , NikahatKazi ,"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] HibaAsri, HajarMousannif, 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, KawtarAarika ,EL Habib Benlahmar, RachidaAitAbdelouhahide, Olivier Debauche," Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis", ELSEVIER,Volume. 191,pp.487-492,Sep 2021.

[7] Ch. Shravya, K. Pravalika, ShaikSubhani, "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] NareshKhuriwal,NidhiMishra,"Breast Cancer Diagnosis Using Deep Learning Algorithm",IEEE,Oct 2018.

[9] AmraneMeriemEt.Al."Breast cancer classification using machine learning",IEEE,Apr2018.

[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. RezwanulHaque, Hasib Iqbal, Md. MunirulHasan, MahmudulHasan& Muhammad NomaniKabir,"Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques", Springer Nature Singapore,pp.1-14, Sep 2020.

[12] Noreen Fatima ,Sha Hong AND HaroonAhmed,"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 PuneetKumar,"Prediction of Breast Cancer Using Machine Learning", BanthamScience,Volume 13 , Issue 5 ,pp.901 - 908, 2020.

[15] TawseefAyoub 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,MohammedNassef,AmrBadr,"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] ShuaiLiu,HanLi,QichenZheng,LuYang,MeiyuDuan,XinFeng,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 PuneetKumar,"Prediction of Breast Cancer Using Machine Learning", BanthamScience,Volume 13 , Issue 5 ,pp.901 - 908, 2020.

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

[22] ZexianHuang,DaqiChen,"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, HarshwardhanKharpate, YogeshNarekar, SakshiGulhane, TanviVirulkar,"Breast Cancer Detection and Prediction using Machine Learning",IEEE,Sep 2021.

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

[25] BegümErkal, TülinErçelebiAyyıldız,"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] app.diagrams.net

[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] BoluwaliA.Akinnuwesi,BabafemiO.Macaulay,Benjamin S.Aribisala,"Breast Cancer risk assessment and earlydiagnosis using principle component analysis andsupport vector machine techniques",informatics in Medicine unlocked,2020.

[49] Tariq mahmood,jianqiangli,yanpei,faheemakhtar,AzharImran,khalilurRehman."A Brief Survey on Breast CancerDiagnostic with Deep Learning Schemes Using Multi-Image Modalities",IEEE Access,2020.

[50] Noreen Fatima,LILiu,ShaHong,Haroon Ahmed.
"Prediction of Breast Cancer,Comparative Review of Machine LearningTechniques, and Their Analysis",IEEE Access,2020.

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

[52] R.Preetha,S.Vinilajinny."Breast-cancer predictionstrategies and experimental processing using DEFS algorithm",Materials Today:Proceedings,2021.

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