INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 2 - V7I2-1247)
Available online at: https://www.ijariit.com

Predictive classification of breast cancer using machine learning

E. S. Dharani
dharanisenthil0512@gmail.com
A.V.C College of Engineering, Mayiladuthurai, Tamil Nadu

S. Ishwarya
ishuselvam2000@gmail.com
A.V.C College of Engineering, Mayiladuthurai, Tamil Nadu

Dr. R. Kanimozhi
kanimozhiit@avccengg.net
A.V.C College of Engineering, Mayiladuthurai, Tamil Nadu

ABSTRACT

Breast cancer is a disease in which cells in the breast grow out of control. There are different kinds of breast cancer. The kind of breast cancer depends on which cells in the breast turn into cancer. Breast cancer can begin in different parts of the breast. A breast is made up of three main parts: lobules, ducts, and connective tissue. The lobules are the glands that produce milk. The ducts are tubes that carry milk to the nipple. The connective tissue (which consists of fibrous and fatty tissue) surrounds and holds everything together. Most breast cancers begin in the ducts or lobules. Breast cancer can spread outside the breast through blood vessels and lymph vessels. When breast cancer spreads to other parts of the body, it is said to have metastasized. Advances in screening and treatment for breast cancer have improved survival rates dramatically since 1989. According to the American Cancer Society (ACS), there are more than 3.1 million breast cancer survivors in the United States. The chance of any woman dying from breast cancer is around 1 in 38 (2.6%). The ACS estimate that 268,600 women will receive a diagnosis of invasive breast cancer and 62,930 people will receive a diagnosis of noninvasive cancer in 2019. In the same year, the ACS report that 41,760 women will die as a result of breast cancer. However, due to advances in treatment, death rates from breast cancer have been decreasing since 1989. However, The required facility for diagnosing cancer accurately and at the earliest stage using the results of the biopsy is not available to all general hospitals. Identifying and diagnosing cancer at the earliest stage is crucial as the possibility of cancer spreading increases. Therefore, A computerized system that identifies cancer at the earliest stage with minimal time with the greatest accuracy and which reduces cancer recurrence and mortality has to be developed. This paper concentrates and summarises the different machine learning algorithms which may be implied in cancer diagnosis to improve the accuracy of the diagnosis and identification.

Keywords: *Random Forest (RF), Conditional Inference Tree (CT)*

1. INTRODUCTION

Breast cancer is cancer that develops from breast tissue.[7] Signs of breast cancer may include a lump in the breast, a change in breast shape, dimpling of the skin, fluid coming from the nipple, a newly inverted nipple, or a red or scaly patch of skin.[1] In those with distant spread of the disease, there may be bone pain, swollen lymph nodes, shortness of breath, or yellow skin.[8]

About 5–10% of cases are the result of a genetic predisposition inherited from a person's parents,[1] including BRCA1 and BRCA2 among others.[1] Breast cancer most commonly develops in cells from the lining of milk ducts and the lobules that supply these ducts with milk.[1] Cancers developing from the ducts are known as ductal carcinomas, while those developing from lobules are known as lobular carcinomas.[1] There are more than 18 other subtypes of breast cancer.[2] Some, such as ductal carcinoma in situ, develop from pre-invasive lesions.[2] The diagnosis of breast cancer is confirmed by taking a biopsy of the concerning tissue.[1] Once the diagnosis is made, further tests are done to determine if the cancer has spread beyond the breast and which treatments are most likely to be effective.[1]

The balance of benefits versus harms of breast cancer screening is controversial. A 2013 Cochrane review found that it was unclear if mammographic screening does more harm than good, in that a large proportion of women who test positive turn out not to have the disease.[9] A 2009 review for the US Preventive Services Task Force found evidence of benefit in those 40 to 70 years of age,[10] and the organization recommends screening every two years in women 50 to 74 years of age.[11] The medications tamoxifen or raloxifene may be used in an effort to prevent breast cancer in those who are at high risk of developing it.[2] Surgical removal of both breasts is another preventive measure in some high risk women.[2] In those who have been diagnosed with cancer, a number of treatments may be used, including surgery, radiation therapy, chemotherapy, hormonal therapy, and targeted therapy.[1] Types of surgery vary from breast-conserving surgery to mastectomy.[12][13] Breast reconstruction may take place at the time of surgery or at a later date.[13] In those in whom the cancer has spread to other parts of the body, treatments are mostly aimed at improving quality of life and comfort.[13]

2. TYPES OF BREAST CANCER

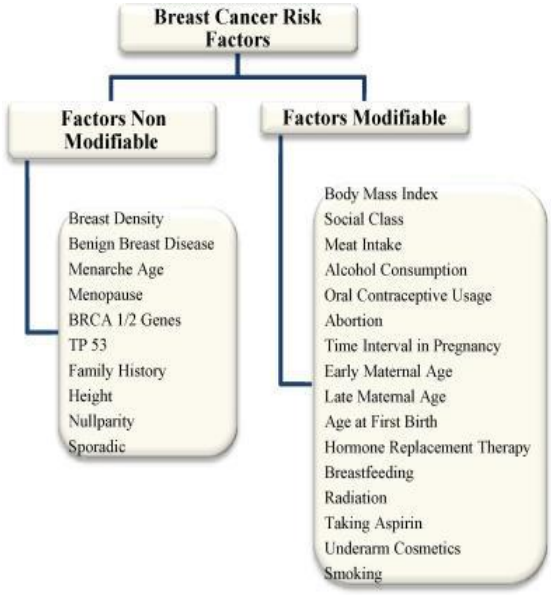
1. Benign Breast Cancer (Non-Invasive) [3]: It is vastly referred to as in situ carcinoma. As the name indicates this disease remains entirely in its place of origin (insitu) and doesn't spread to nearby tissue regions. Ductal carcinoma in situ cancer type is grown usually inside the milk duct. This is developed in both men and women.
2. Malignant Breast Cancer (Invasive) [3]: This type has an ability to spread to the
3. surrounding tissues and is a threat to life. Invasive ductal cancer is a widely occurring type of invasive BC.
4. Other types of Breast Cancer [3]: This is the least common type of breast cancer which includes invasive lobular BC(developed in the milk producing lobules), tubular BC, inflammatory BC, papillary BC and medullary BC.

3. RISK FACTORS

Something that affects the individual to acquire some disease, such as breast cancer is defined as a potential risk factor. There have been numerous situations where ladies have breast malignancy without evident symptoms or risk factors.. It can be classified based on non-preventable and preventable [4]

Non-Preventable Risk Factors:

1. Gender: In this case women are more prone to the risk. The disease is likely to occur in women at a chance about 100 times higher compared to men.
2. Age: Age also plays a vital role in growing cancer. Women within the age of 55 or older than that are more prone to the risk of Breast Cancer.
3. Genetic risk factors: Due to solid hereditary characteristics 5% of malignant growth are seen as acquired to a person. There are two autosomal predominant qualities, BRCA1 and BRCA2 that represent most instances of familial cancer disease. 65% to 85% of ladies with destructive BRCA transformation have danger of creating malignancy.
4. Family History: On the off chance that the woman''s mom, sister, father or youngster has been determined to have ovarian malignant growth, at that point the danger of developing the infection increases(twice). Regardless of whether the relative was analyzed before the age of 50 the hazard increments.



Preventable Risk Factors:

Alcoholic consumption, smoking and being overweight has high risk of recurrence. Other hazard factors include High BMI after menopause, Lack of activity, Radiation Therapy on the body (before age 30), Hormonal use – postmenopausal ,Weight increase after menopause, Africa-,American race has higher chances of BC,High bone thickness, Late pregnancy at an older age

4. ALGORITHMS

Machine learning algorithms are described as learning a target function (f) that best maps input variables (X) to an output variable (Y): Y = f(X)

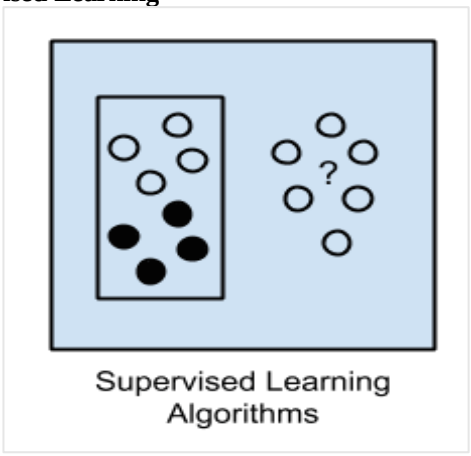
This is a general learning task where we would like to make predictions in the future (Y) given new examples of input variables (X). We don't know what the function (f) looks like or its form. If we did, we would use it directly and we would not need to learn it from data using machine learning algorithms.The most common type of machine learning is to learn the mapping Y = f(X) to make predictions of Y for new X. This is called predictive modeling or predictive analytics and our goal is to make the most accurate predictions possible.

There are different ways an algorithm can model a problem based on its interaction with the experience or environment or whatever we want to call the input data. It is popular in machine learning and artificial intelligence textbooks to first consider the learning styles that an algorithm can adopt.

There are only a few main learning styles or learning models that an algorithm can have and we'll go through them here with a few examples of algorithms and problem types that they suit.This taxonomy or way of organizing machine learning algorithms is useful because it forces you to think about the roles of the input data and the model preparation process and select one that is the most appropriate for your problem in order to get the best result.

4.1 Three different learning styles in machine learning algorithms

1. Supervised Learning



Input data is called training data and has a known label or result such as spam/not-spam or a stock price at a time.

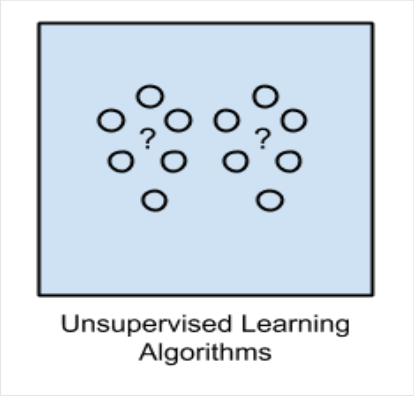
A model is prepared through a training process in which it is required to make predictions and is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy on the training data.Example problems are classification and regression.Example algorithms include: Logistic Regression and the Back Propagation Neural Network.

2. Unsupervised Learning

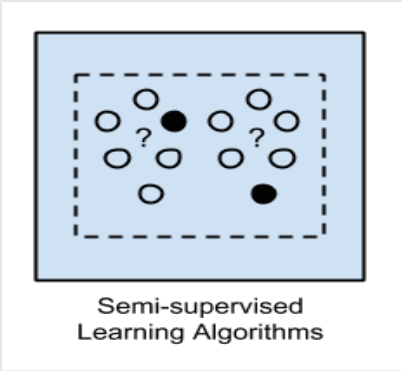
Input data is not labeled and does not have a known result.A model is prepared by deducing structures present in the input data. This may be to extract general rules. It may be through a

mathematical process to systematically reduce redundancy, or it may be to organize data by similarity.

Example problems are clustering, dimensionality reduction and association rule learning.Example algorithms include: the Apriori algorithm and K-Means.



3. Semi-Supervised Learning



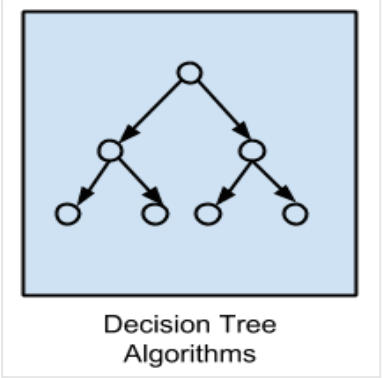
Input data is a mixture of labeled and unlabelled examples. There is a desired prediction problem but the model must learn the structures to organize the data as well as make predictions.

Example problems are classification and regression. Example algorithms are extensions to other flexible methods that make assumptions about how to model the unlabeled data.

Decision Tree Algorithms

When crunching data to model business decisions, you are most typically using supervised and unsupervised learning methods.

A hot topic at the moment is semi-supervised learning methods in areas such as image classification where there are large datasets with very few labeled examples.



Decision tree methods construct a model of decisions made based on actual values of attributes in the data.Decisions fork in tree structures until a prediction decision is made for a given record. Decision trees are trained on data for classification and regression problems. Decision trees are often fast and accurate and a big favorite in machine learning.

The most popular decision tree algorithms are:

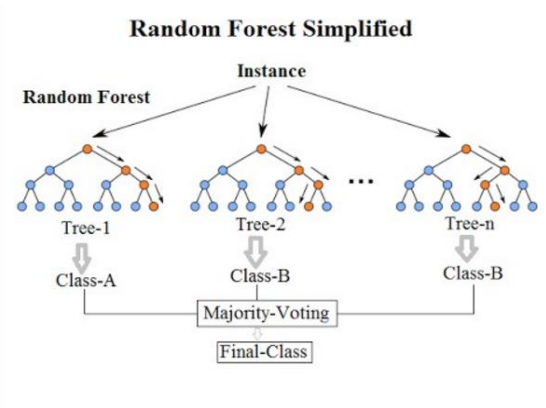
- Classification and Regression Tree (CART)
- Iterative Dichotomiser 3 (ID3)
- C4.5 and C5.0 (different versions of a powerful approach)
- Chi-squared Automatic Interaction Detection (CHAID)
- Decision Stump
- Conditional Decision Trees

5. RANDOM FOREST

Random Forest is one of the most popular and most powerful machine learning algorithms. It is a type of ensemble machine learning algorithm called Bootstrap Aggregation or bagging.

The bootstrap is a powerful statistical method for estimating a quantity from a data sample. Such as a mean. You take lots of samples of your data, calculate the mean, then average all of your mean values to give you a better estimation of the true mean value.

In bagging, the same approach is used, but instead for estimating entire statistical models, most commonly decision trees. Multiple samples of your training data are taken then models are constructed for each data sample. When you need to make a prediction for new data, each model makes a prediction and the predictions are averaged to give a better estimate of the true output value.



Random forest is a tweak on this approach where decision trees are created so that rather than selecting optimal split points, suboptimal splits are made by introducing randomness.

The models created for each sample of the data are therefore more different than they otherwise would be, but still accurate in their unique and different ways. Combining their predictions results in a better estimate of the true underlying output value.
If you get good results with an algorithm with high variance (like decision trees), you can often get better results by bagging that algorithm.

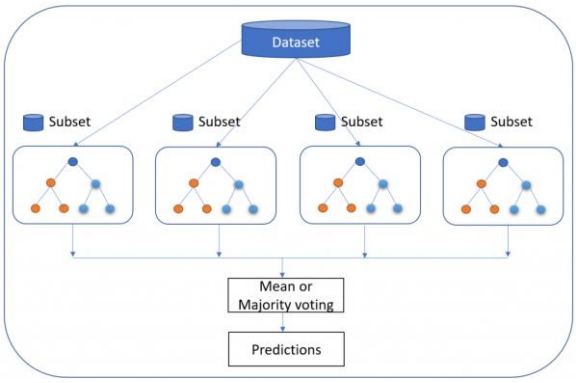
5.1 Random Forest

Random Forest is perhaps the most popular classification algorithm, capable of both classification and regression. It can accurately classify large volumes of data.

The name "Random Forest" is derived from the fact that the algorithm is a combination of decision trees. Each tree depends

on the values of a random vector sampled independently with the same distribution for all trees in the "forest." Each one is grown to the largest extent possible.

Predictive analytics algorithms try to achieve the lowest error possible by either using "boosting" (a technique which adjusts the weight of an observation based on the last classification) or "bagging" (which creates subsets of data from training samples, chosen randomly with replacement). Random Forest uses bagging. If you have a lot of sample data, instead of training with all of them, you can take a subset and train on that, and take another subset and train on that (overlap is allowed). All of this can be done in parallel. Multiple samples are taken from your data to create an average.



While individual trees might be "weak learners," the principle of Random Forest is that together they can comprise a single "strong learner."

The popularity of the Random Forest model is explained by its various advantages:

- Accurate and efficient when running on large databases
- Multiple trees reduce the variance and bias of a smaller set or single tree
- Resistant to overfitting
- Can handle thousands of input variables without variable deletion
- Can estimate what variables are important in classification
- Provides effective methods for estimating missing data
- Maintains accuracy when a large proportion of the data is missing

5.2 Conditional Interference Trees

Conditional inference trees (ctrees) embed tree-structured regression models into a well-defined theory of conditional inference procedures. They use a significance test procedure to select variables instead of selecting the variable that maximizes any information measure. In addition, ctree is applicable to all types of regression issues, including nominal, ordinal, numeric, censored, and multivariate response variables, as well as arbitrary measurement scales of covariates. A flexible and extensible computational tool in the "partykit" package of R is suitable for fitting and visualizing ctrees [54,55].

6. LITERATURE REVIEWS

S No.	Title	Author	Year of publishing	Publishing details
1	A Pancreatic Cancer Detection Support Tool Using Mass Spectrometry Data and Support Vector Machines	Emmanuel Briones, Angelyn Lao Geoffrey A. Solano	April 2020	International Conference on Artificial Intelligence and Signal Processing (AISP) **Web link:** https://ieeexplore.ieee.org/document/9073503
2	An Analysis On Breast Disease Prediction Using Machine Learning Approaches	F. M. Javed Mehedi Shamrat, Md. Abu Raihan, A.K.M. Sazzadur Rahman, Imran Mahmud, Rozina Akter	February 2020	International Journal Of Scientific & Technology Research **Web link:** https://www.academia.edu/download/62205809/An-Analysis-On-Breast-Disease-Prediction-Using-Machine-Learning-Approaches_220200226-113416-1oeghbr.pdf
3	A Breakup Machine Learning Approach for Breast Cancer Prediction	Sabari Vishnu Jayanthan Jaikrishnan, Orawan Chantarakasemchit, Phayung Meesad	December 2019	International Conference on Information Technology and Electrical Engineering (ICITEE) **Web link:** https://ieeexplore.ieee.org/abstract/document/8929977/authors
4	Applying Support Vector Machine to Electronic Health Records for Cancer Classification	Xudong Zhang, Jiehao Xiao, Feng Gu	June 2019	Spring Simulation Conference (SpringSim) **Web link:** https://ieeexplore.ieee.org/document/8732906
5	Support Vector Machine - Recursive Feature Elimination (SVM - RFE) for	Amazona Adorada, Ratih Permatasari, Panji Wisnu Wirawan, Adi Wibowo,	January 2019	International Conference on Informatics and Computational Sciences (ICICoS) **Web link:**

	Selection of MicroRNA Expression Features of Breast Cancer	Adi Sujiwo		https://ieeexplore.ieee.org/document/8621708
6	Breast cancer classification using machine learning	Meriem Amrane, Saliha Oukid, Ikram Gagaoua, Tolga Ensarİ	June 2018	Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) **Web link:** https://ieeexplore.ieee.org/abstract/document/8391453
7	Prediction of breast cancer using support vector machine and K-Nearest neighbors	Md. Milon Islam, Hasib Iqbal, Md. Rezwanul Haque, Md. Kamrul Hasan.	February 2018	IEEE Region 10 Humanitarian Technology Conference (R10-HTC) **Web link:** https://ieeexplore.ieee.org/abstract/document/8288944/
8	Revealing determinant factors for early breast cancer recurrence by decision tree	Jimin guo, Bengamin C.Fung, Farkhund Iqbal, Peter J.Kuppen, Rob A.Tollenaar, Wilma E.Mesker, Jean-Jacques Lebrun.	December 2017	Information Systems Frontiers **Web link:** https://dl.acm.org/doi/10.1007/s10796-017-9764-0
9	Investigating the effect of correlation based feature selection on breast cancer diagnosis using artificial neural network and support vector machines	Reem Alyami, Jinan Alhajjaj, Batool Alnajrani, Ilham Elaalami, Abdullah Alqahtani, Nahier Aldhafferi, Taoreed O. Owolabi, Sunday O. Olatunji	April 2017	International Conference on Informatics, Health & Technology (ICIHT) **Web link:** https://ieeexplore.ieee.org/document/7899011
10	Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review	Pedro Henriques Abreu, Miriam Seoane Santos, Miguel Henriques Abreu, Bruno Andrade, Daniel Castro Silva	October 2016	ACM (Association for Computing Machinery) Computing Surveys **Web link:** https://dl.acm.org/doi/abs/10.1145/2988544

7. CONCLUSION

In our paper, cancer at the side of ML was introduced and studied in addition. an associate degree in-depth literature survey was performed on varied ML ways used for cancer detection. The findings of those researchers recommend that SVM is the most recommended technique used for cancer detection applications. RF and CT were used either alone or combined with another technique to improve the performance.It was found that RF and CT has the very best accuracy of 99.8% .Therefore, developing a computerised cancer diagnosis can facilitate to scale back the quantity of your time to diagnose the cancer at the earliest stage with the best accuracy and reduce cancer repetition and mortality. This paper summarizes the survey on varied machine learning algorithms and ways that square measure is used to boost the accuracy of predicting cancer at the earliest stage.

8. REFERENCES

[1] India against cancer 2019, "Breast Cancer", National Institute of Cancer Prevention and Research, viewed 12 November 2019,

[2] World Cancer Research Fund 2018, "Breast Cancer", American Institute of Cancer Research, viewed 15 November 2019, .

[3] American Cancer Society 2019, "What is Breast Cancer" American Cancer Society, viewed 16 November 2019, .

[4] McElroy, J A.Newcomb, P A.Trentham-Dietz, Titus-Ernstoff, L Hampton, J M. Egan, K M.,"BreastCancer Risk Associated With Electromagnetic field Exposure From Computer Work Ascertained From Occupational HistoryData",17 th conference, ISEE, September,2005.

[5] AbienFred M.Agarap, "On Breast Cancer Detection:An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset',International Conference on Machine Learning, February 2–4, 2018, Phu Quoc Island, Viet Nam.

[6] H.Asri, Mousannif, Hajar, Al Moatassime, Hassan, Noël, Thomas, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", Procedia Computer Science,2016, vol. 83, pp.1064-1069.

[7] David A. Omondiagbe , Shanmugam Veeramani

,Amandeep S. Sidhu ,"Machine Learning Classification Techniques for Breast Cancer Diagnosis", IOP Conference Series: Materials and Science,2019.

[8] M . Kumari, V. Singh,"Breast Cancer Prediction Systems", Procedia Computer Science, 2018, vol.132,pp. 371-376. Puja Gupta et al. / Procedia Computer Science 171 (2020) 593–601 601 Author name / Procedia Computer Science 00 (2019) 000–000

[9] K.Kourou, T.Exarchos , " Machine Learning Application in Cancer Prognosis and Diagnosis",Computational and Structural Biotechnology Journal,2014.

[10] J. Guo, M.Fung, F. Iqbal, Kuppen, R. Tollenaar,J. Lebran," Revealing Early Determinants of Occurance of Breast Cancer", Information Systems Frontiers,2017 issue 6, pp.1233-1241.

[11] Karabatak M, Ince MC," An expert system for detection of breast cancer based on association rules and neural network", Expert systems with Applications",2009 March,vol 36(2),pp.3465-3469.

[12] Kharya S, Soni S.," Weighted naive bayes classifier: A predictive model for breast cancer detection", International Journal of Computer Applications. 2016 Jan,vol.133(9), pp.32-37.

[13] M. R. Mohebian, H. R. Marateb, M. Mansourian, M. A. Mañanas, and F. Mokarian, "A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning,"Computer Structural Biotechnology,2017,vol. 15, pp. 75– 85.

[14] B.M.Gayathri.,C.P.Sumathi and T.Santhanam ," Breast Cancer Diagnosis Using Machine Learning Algorithms –A Survey ",International Journal of Distributed and Parallel Systems (IJDPS) Vol.4, No.3, May 2013.

[15] PayamZarbakhsh,Abdul Jalil Addeh," Breast cancer tumor type recognition using graph feature selection technique and radial basis function neural network with optimal structure",

[27] ". arXiv preprint arXiv:1412.6980. 2014 Dec 22.

Journal of Cancer Research and Therapeutics,2018,vol.14,pp.625-33.

[16] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.

[17] Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M.Blondel, P. Prettenhofer, R. Weiss,V. Dubourg, J. Vanderplas, A. Passos, D. Cour-napeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011.,"Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12 (2011)",pp. 2825–2830.

[18] F.Bunea,"Honest Variable Selection in Linear and Logistic Regression models" International Journal of Statistics2 (2008).

[19] Octaviani TL, Rustam Z,"Random forest for breast cancer prediction",AIP Conference Proceedings, AIP Publishing Nov 4 ,2019 ,vol. 2168.

[20] LeCun Y, Bengio Y, Hinton," Deep learning. nature.", May,2015,issue 521(7553),pp.436-44.

[21] Medjahed SA, Saadi TA, Benyettou," A. Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules",International Journal of Computer Applications. 2013 Jan 1,vol.62(1).

[22] Yi L, Yi W," Decision Tree Model in the Diagnosis of Breast Cancer" ,International Conference on Computer Technology, Electronics and Communication (ICCTEC) , IEEE. Dec 19 ,2017 ,pp. 176-179 .

[23] Mert A, Kilic N, Akan A.,"Breast cancer classification by using support vector machines with reduced dimension.",ELMAR, ,IEEE,2011 Sep 14 ,pp.37-40.

[24] Gou J, Du L, Zhang Y, Xiong T ,"A new distance-weighted k-nearest neighbor classifier", J.ournal of Information of Computer Science,June 2012 ,vol.9(6),pp.1429-36.

[25] UCHealth 2015, "How Accurate are mammograms?", UCHealth viewed 16 November 2019,

[26] Kingma D P, Ba J, "Adam: A method for stochastic optimization

