ANALYSIS ON EARLY PREDICTION OF BREAST CANCER USING DIFFERENT ALGORITHMS

[1] P.Laura Juliet M.C.A., M.Phil., Assistant Professor of Computer Applications, Vellalar College for Women(Autonomous),Erode, India.

[2] Dr. P.R.Tamilselvi., M.Sc., M.Phil., Ph.D., Assistant Professor of Computer Science, Government Arts & Science College, Komarapalayam, India.

Abstract

Cancer is a critical wellbeing trouble around the world. It is evaluated that more than 9 million cancer patients are expected to pass on in creating nations from various sorts of cancers by 2030. The frequency of various kinds of cancers is expanding because of an undesirable way of life. Breast Cancer (BC) is the second most habitually analyzed cancer and the fifth reason for cancer mortality around the world. Prediction assumes a critical job in diseases with related high death rates, since it has the ability to assist clinicians with defining every patient's anticipation, consequently permitting to customize the comparing medications. The data mining and factual learning methods were utilized to find steady and valuable examples in huge datasets. We have analyzed the exactness level of various data mining learning calculations, and the best model has executed for predicting disease. This paper examination Several data mining methodologies like a Support Vector Machine (SVM), Random Forest and Artificial Neural network (ANN) to discover the most accurate outcome.

Keywords: Classification, Prediction, Data mining, Breast cancer, Support Vector Machine, Random Forest, Artificial Neural Network.

1. Introduction

Cancer is the name given to the marvel of uncontrolled development of abnormal cells. BC is the name given to dangerous tumors that start in the breast, subsequently the name. In any case, numerous patients that have BC don't have genuine side effects, or may partner weariness and weight reduction (conceivable cancer side effects) to various different causes (stress, diverse eating regimen, less rest). The mammogram, a X-beam picture of the patient's breast, assumes a significant job in the early discovery of BC, detecting cancer much before any manifestations appear. Outer indications of BC may incorporate an irregularities in the breast, or general changes. At the point when a patient finds an oddity in the breast (by

means of self-assessment or in a medical checkup) or a mammogram uncovers it, the doubt of cancer shows up. A biopsy is then performed, and a pathologist inspects it to affirm the finding, while radiology can be utilized to distinguish inaccessible association in different organs by cancerous cells (metastases).Invasive BC can be separated by the beginning neighborhood of the tumor inside the breast, and the two most continuous are ductal and lobular. These names start in the names of the conduits, channels that convey the milk from the creating organs to the areola, and the lobules, the organs themselves.

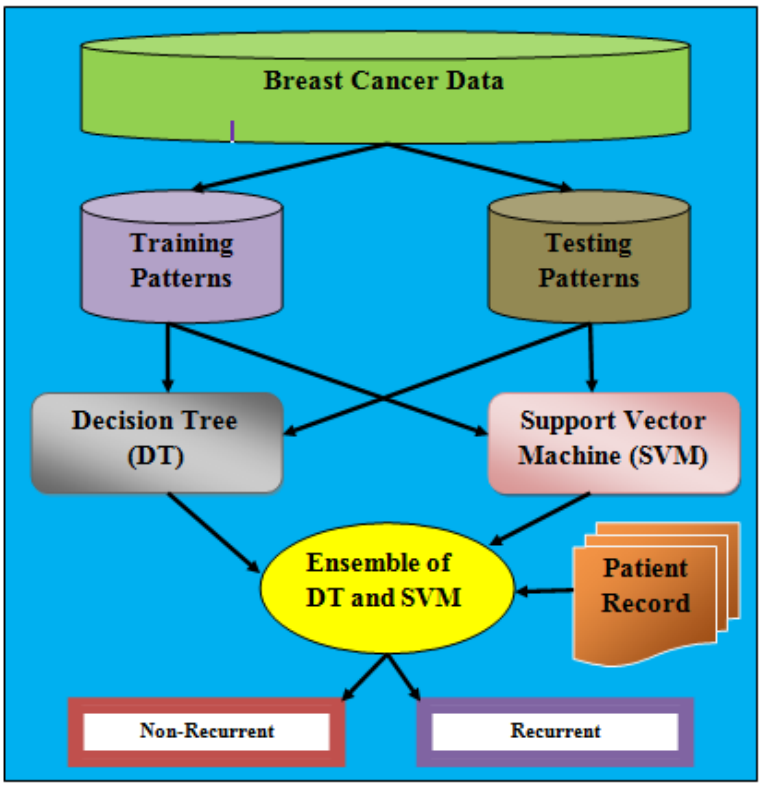

Figure 1: Ensemble Model for Breast cancer identification

It is the most widely recognized type of BC, representing around 80% of obtrusive BC. Intrusive lobular carcinomas start in the lobules, speaking to about 10% of obtrusive BC.BC subtypes are a method for classifying patients based on some significant highlights of the tumors. The factors used to recognize these subgroups are surveyed in a compound procedure called immunohistochemistry (IHC), and speak to the nearness or nonappearance of changed protein in the tumor (individually positive, +, and negative, −). Estrogen Receptors (ER) are receptors of the hormone Estrogen, while Progesterone Receptors (PR) are receptors of the hormone Progesterone. HER2 (human epidermal development factor receptor 2) is another significant protein, connected with the movement of BC tumors.BC is usually treated by one or a few blends of what has been referenced previously: medical procedure, radiation treatment, chemotherapy, and hormone treatment. The data mining and factual learning classification calculations were utilized to extricate information from the breast cancer dataset. The presentation of these calculations, specifically, Support Vector Machines (SVM), Random Forest, AdaBoost, Decision Trees, Artificial Neural Networks

(ANN), Rule Based Classifiers, Bagging, Boosting, Principal Component Analysis (PCA) and Bayesian Based classifiers were analyzed by utilizing classification accuracy, confusion matrix and stratified 10-crease cross approval strategies.

2. Literature Survey

Ahmed Iqbal Pritom Shahed Anzarus Sabab, Md. Ahadur Rahman Munshi, Shihabuzzaman Shihab et.al Proposed to Predicting Breast Cancer Recurrence utilizing effective Classification and Feature Selection technique. To give a respectable approach so as to improve the accuracy of those models. Utilizing appropriate property selection technique, any classification algorithm can be improved significantly. Attributes with less contribution in dataset frequently misleads the classification and results in poor prediction. In this proposed, to establish Support Vector Machine giving much better yield both when property selection. Territory under ROC curve analysis indicated brings about the support where Naïve Bayes and Decision Tree demonstrated much better improvement after feature selection technique. An efficient feature selection algorithm helped us to improve the accuracy of each model by reducing some lower positioned attributes. Not just the contributions of these attributes are less, yet their expansion likewise deceives the classification algorithms. **Tarek Gaber, Gehad Ismail, Ahmed Anter , Mona Soliman , Mona Ali , Noura Semary , Aboul Ella Hassanien, Vaclav Snasel et.al** Proposed the Thermogram Breast Cancer Prediction Approach based on Neutrosophic Sets and Fuzzy C-Means Algorithm. This approach consists of two fundamental stages: automatic segmentation and classification. Likewise, post-segmentation process was proposed to section breast parenchyma (for example return on initial capital investment) from thermogram pictures. For the classification, diverse portion functions of the Support Vector Machine (SVM) were utilized to classify breast parenchyma into ordinary or abnormal cases. The framework originally extracted ROI utilizing Neutrosophic Set, FFCM and morphological operators. It at that point utilized a few features (statistical, surface and vitality) with the SVM to detected typical and abnormal breast. Utilizing a benchmark database, the proposed framework was assessed through recall, accuracy, precision, and mistake rate demonstrating that the CAD framework achieving an excellent outcomes. Additionally, it was discovered that NS sets with F-FCM is an effective segmentation strategy for thrmogram pictures as NS enhanced warm picture and reduced the indeterminacy. **Valentina Giannini, Samanta Rosati, Cristina Castagneri, Laura Martincich , Daniele Regge, Gabriella Balestra et.al** Proposed the radiomics for pretreatment prediction of pathological response to neoadjuvant treatment utilizing magnetic resonance imaging: influence of feature selection. To extracted 27 3D surface features from the dynamic contrast enhanced-MRI, and to created four feature subsets utilizing diverse FS algorithms.To compare the performance of a Bayesian classifier in predicting pCR, when utilizing subset of features got from FS algorithms. FS is a procedure for dimensionality reduction of multivariate data, extensively utilized in the biomedical field for supporting data, sign and picture analysis. It consists in choosing a subset of features,

from the underlying arrangement of factors, ready to protect the first information content. By excluding from the dataset of unimportant or repetitive attributes, it is conceivable to increase classifier performance since certain factors could be source of clamor. Besides, uniquely in contrast to other dimensionality reduction algorithms which play out a change of the first factor space (for example PCA), FS selects a subset of the first parameters, protecting the factors meaning and facilitating the period of information understanding. From a clinical perspective, this is significant since a twofold bit of leeway for patients can be gotten: (an) an early modification of the treatment for those patients that are not likely reacting, (b) a reduction of toxicity because of unnecessary medications. **Youness Khourdifi Mohamed Bahaj et.al** Proposed to Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification. To anticipate breast cancer, which is the subsequent driving reason for death among ladies around the world, and with early discovery and counteractive action can significantly diminish the risk of death, utilizing a few machine-learning algorithms that are Random Forest, Naïve Bayes, Support Vector Machines SVM, and K-Nearest Neighbors K-NN, and picked the best. To concentrated on the utilization of classification systems in restorative science and bioinformatics. Classification is the most usually utilized information mining method and utilizations a lot of pre-arranged guides to build up a model to characterize the number of inhabitants in records. The principle goal of the classification system is to precisely foresee the objective class for each case in the information. To break down information from a breast cancer dataset utilizing a classification method in the field of restorative bioinformatics to precisely anticipate the class for each situation, utilizing the weka information mining apparatus and its utilization for classification. It initially orders the informational collection and then decides the best calculation for the finding and prediction of breast cancer sickness. Prediction starts with recognizing side effects in patients, at that point distinguishing sick patients from a huge numer of sick and sound patients.

3. Prediction Algorithms

Random Forest (RF)

Random forests are a blend of tree predictors to such an extent that each tree relies upon the estimations of a random vector tested autonomously and with a similar dissemination for all trees in the forest. The speculation mistake for forests combines a.s. as far as possible as the quantity of trees in the forest turns out to be enormous. The speculation blunder of a forest of tree classifiers relies upon the quality of the individual trees in the forest and the connection between's them. Utilizing a random determination of features to part every hub yields mistake rates that contrast positively with Adaboost .however are increasingly strong as for commotion. Interior appraisals screen blunder, quality, and relationship and these are utilized to demonstrate the reaction to expanding the quantity of features utilized in the parting. Inner evaluations are likewise used to quantify variable significance. These thoughts are likewise relevant to regression.Random forests or random choice forests are a group learning strategy for classification,

regression and different errands that works by building a large number of choice trees at training time and yielding the class that is the method of the classes (classification) or mean prediction (regression) of the individual trees. Random choice forests right for choice trees' propensity for overfitting to their training set.

Support Vector Machine (SVM)

Support Vector Machine is a supervised learning model which is related with learning calculations that dissect data and recognize patterns that is utilized for classification and regression analysis. The data set is isolated into two sets: training and test. To begin with, the example is isolated in two gatherings: masses and non masses. Next, each gathering is randomly separated into10 subsets, from which one subset is picked for training and the staying ones are utilized for test. This procedure is rehashed until the sum total of what subsets have been tested. The support vector machine (SVM) was utilized with outspread bit and standard parameters (C=1 and y=0.5). A classification task more often than not includes isolating data into training and testing sets. The objective of SVM is to deliver a model (in view of the training data) which predicts the objective estimations of the test data given just the test data characteristics. Since the extent of nonmasses chose in the division stage is around multiple times higher than the quantity of masses, higher weight was alloted to the training of masses. This implies in training, the punishment for a mass classification blunder is more noteworthy than for a non-mass. A decent harmony between these two files was accomplished utilizing weight 9 for the mass example and weight 1 for the non-mass sample. SVM is a paired classifier dependent on supervised learning which gives preferable execution over different classifiers. SVM arranges between two classes by building a hyper plane in high-dimensional feature space which can be utilized for classification.

Artificial Neural Networks (ANN)

An Artificial Neural Network (ANN) is a scientific model that attempts to reproduce the structure and functionalities of organic neural networks. Essential structure square of each artificial neural network is artificial neuron, that is, a straightforward scientific model. Such a model has three basic arrangements of principles: augmentation, summation and initiation. At the passage of artificial neuron the data sources are weighted what implies that each info worth is duplicated with individual weight. In the center segment of artificial neuron is entirety work that wholes every single weighted information and predisposition. At the exit of artificial neuron the aggregate of recently weighted sources of info and predisposition is passing trough initiation work that is additionally called exchange work. Artificial neural networks endeavor to streamline and emulate this mind conduct. They can be prepared in a supervised or unsupervised manner. In a supervised ANN, the network is prepared by giving coordinated info and yield data tests, with the aim of getting the ANN to give an ideal yield to a given information. A model is an email spam channel – the information training data could be the include of different words in the body of the email, and the yield training data would be a classification of whether the email was genuinely spam or not. In the event that

numerous instances of messages are gone through the neural network this enables the network to realize what info data makes it likely that an email is spam or not. Artificial neural networks are the demonstrating of the human mind with the least difficult definition and building squares are neurons.

4. Experimental Results

Accuracy Ratio

Random Forest	Support Vector Machine	Artificial Neural Network
67.2	57	69.5
69.7	59	69.9
70.8	62	69.5
72.6	66	70.9
75	69	72

Table 1: Comparison table of Accuracy Ratio

The comparison table of accuracy ratio explains the different values of random forest, support vector machine and artificial neural network. In each level the values are increasing. Random forest values are starts from 67.2 to 75, support vector machine values are starts from 57 to 69 and artificial neural network values are starts from 69.5 to 72.

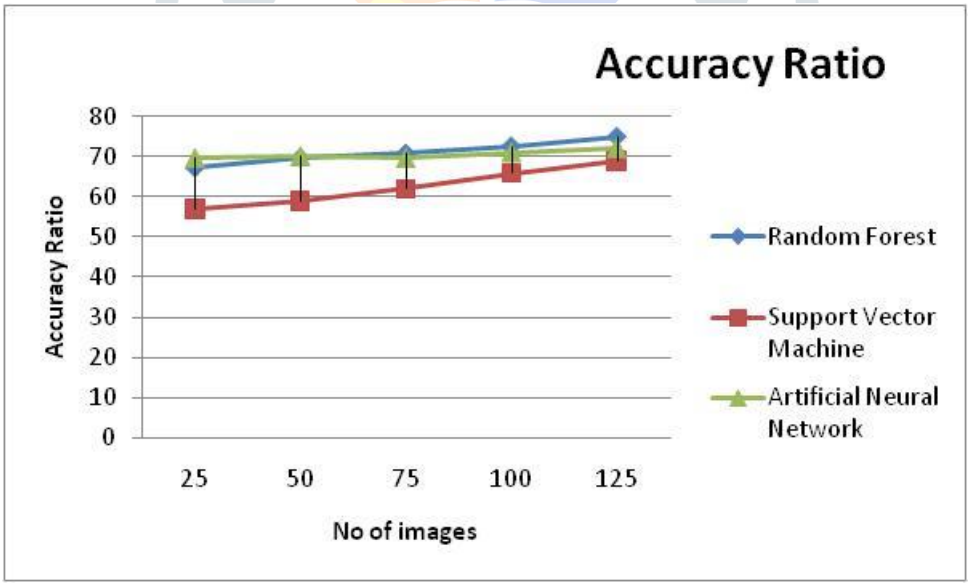

Chart 1: Comparison chart of Accuracy Ratio

The comparison chart of accuracy ratio shows the different values of random forest, support vector machine and artificial neural network. No of images in X axis and accuracy ratio in Y axis. Random forest values are starts from 67.2 to 75, support vector machine values are starts from 57 to 69 and artificial neural network values are starts from 69.5 to 72.

Intensity Ratio

Random Forest	Support Vector Machine	Artificial Neural Network
39	26.77	66
45	31.98	72
49	34.56	76.5
55	38.92	79.8
58	44.56	85

Table 2: Comparison table of Intensity Ratio

The comparison table of intensity ratio explains the different values of random forest, support vector machine and artificial neural network. In each level the values are increasing. Random forest values are starts from 39 to 58, support vector machine values are starts from 26.77 to 44.56 and artificial neural network values are starts from 66 to 85.

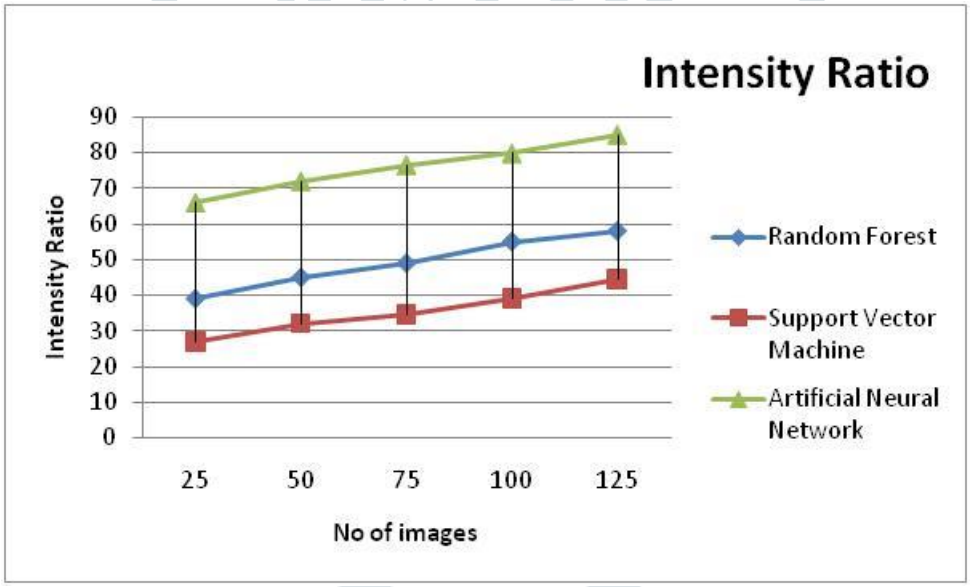

Chart 2: Comparison chart of Intensity Ratio

The comparison chart of intensity ratio shows the different values of random forest, support vector machine and artificial neural network. No of images in X axis and intensity ratio in Y axis. Random forest values are starts from 39 to 58, support vector machine values are starts from 26.77 to 44.56 and artificial neural network values are starts from 66 to 85.

Classification Ratio

Random Forest	Support Vector Machine	Artificial Neural Network
55	75	67
58.6	78.9	70.1
62.3	83.86	74.8

68.9	88.21	78.89
72	92.06	81

Table 3: Comparison table of Classification Ratio

The comparison table of classification ratio explains the different values of random forest, support vector machine and artificial neural network. In each level the values are increasing. Random forest values are starts from 55 to 72, support vector machine values are starts from 75 to 92.06 and artificial neural network values are starts from 67 to 81.

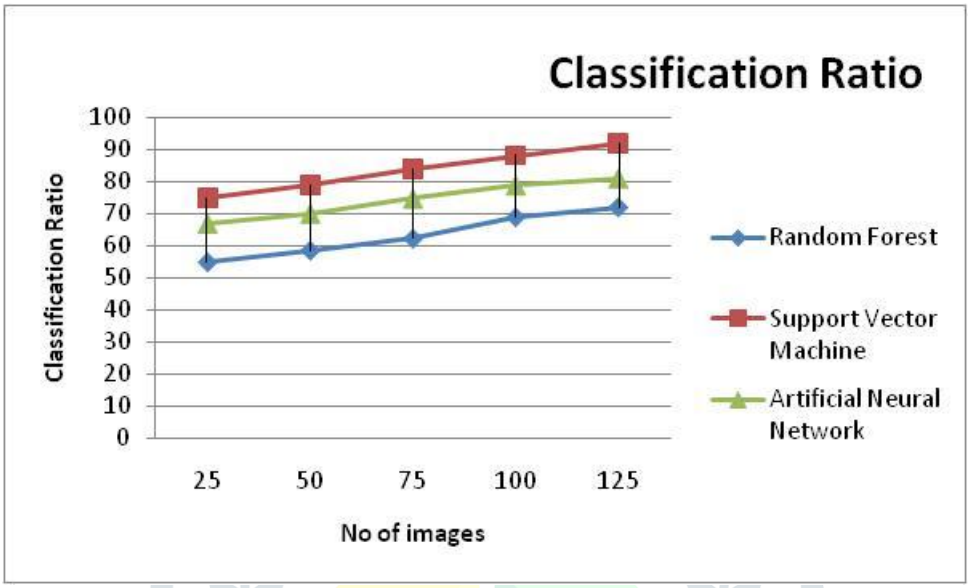

Chart 3: Comparison chart of Classification Ratio

The comparison chart of classification ratio shows the different values of random forest, support vector machine and artificial neural network. No of images in X axis and classification ratio in Y axis. Random forest values are starts from 55 to 72, support vector machine values are starts from 75 to 92.06 and artificial neural network values are starts from 67 to 81.

Detection Ratio

Random Forest	Support Vector Machine	Artificial Neural Network
13	22	7
18	27	15
21	35	20
24	44	22
30	51	26

Table 4: Comparison table of Detection Ratio

The comparison table of detection ratio explains the different values of random forest, support vector machine and artificial neural network. In each level the values are increasing. Random forest values are starts from 13 to 30, support vector machine values are starts from 22 to 51 and artificial neural network values are starts from 7 to 26.

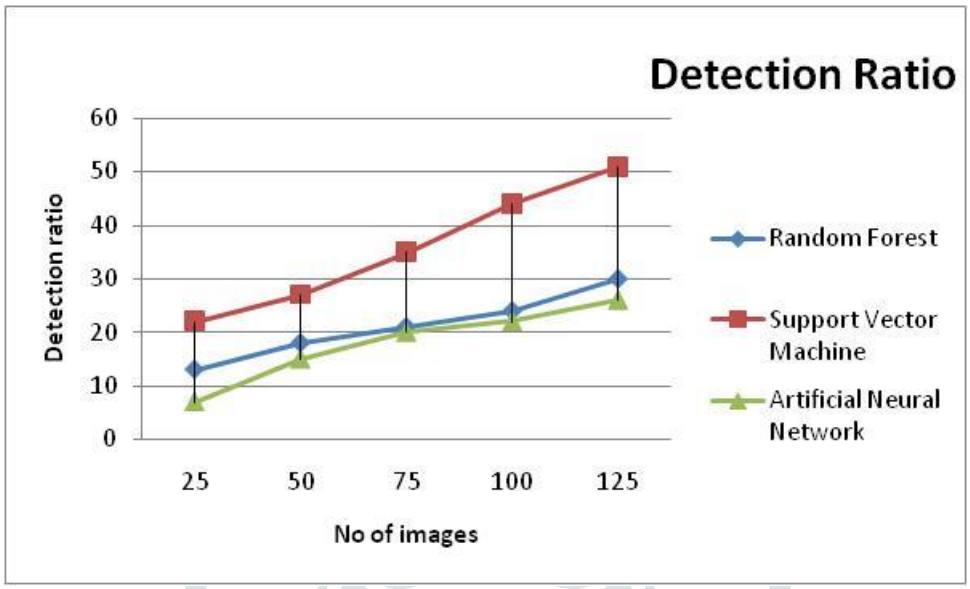

Chart 4: Comparison chart of Detection Ratio

The comparison chart of detection ratio shows the different values of random forest, support vector machine and artificial neural network. No of images in X axis and detection ratio in Y axis. Random forest values are starts from 13 to 30, support vector machine values are starts from 22 to 51 and artificial neural network values are starts from 7 to 26.

Conclusion

Breast cancer is one of the main sources of death in ladies. In this way, early detection is significant. Breast cancer detection done by less intrusive system gives digitized result which makes prediction simpler. There are numerous calculations accessible that can be utilized so as to prepare the dataset and get the precise outcomes. Data mining is the system of recover a pattern from huge data set regarding machine learning, data base and measurements. A data mining strategy such clustering, classification and association which is suitable for medical diagnosis. These different calculations for Support Vector Machine (SVM), Random Forest and Artificial Neural network (ANN) are examined and some are looked at dependent on their exhibition.

References:

[1] Ahmed Iqbal Pritom Shahed Anzarus Sabab, Md. Ahadur Rahman Munshi, Shihabuzzaman Shihab," Predicting Breast Cancer Recurrence using effective Classification and Feature Selection technique", 19th

International Conference on Computer and Information Technology, December 18-20, 2016, North South University, Dhaka, Bangladesh

[2] Tarek Gaber, Gehad Ismail, Ahmed Anter , Mona Soliman , Mona Ali , Noura Semary , Aboul Ella Hassanien, Vaclav Snasel," Thermogram Breast Cancer Prediction Approach based on Neutrosophic Sets and Fuzzy C-Means Algorithm", ©2015 IEEE

[3] Valentina Giannini, Samanta Rosati, Cristina Castagneri, Laura Martincich , Daniele Regge1, Gabriella Balestra," RADIOMICS FOR PRETREATMENT PREDICTION OF PATHOLOGICAL RESPONSE TO NEOADJUVANT THERAPY USING MAGNETIC RESONANCE IMAGING: INFLUENCE OF FEATURE SELEC", 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)

[4] Youness Khourdifi Mohamed Bahaj ," Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification", ©2018 IEEE

[5] Tudorica, K. Y. Oh, S. Y. Chui, et al., "Early prediction and evaluation of breast cancer response to neoadjuvant chemotherapy using quantitative DCE-MRI," Translational Oncology, vol. 9, no. 1, pp. 8-17, 2016.

[6] V. Giannini, S. Mazzetti, A. Marmo, et al., "A computeraided diagnosis (CAD) scheme for pretreatment prediction of pathological response to neoadjuvant therapy using dynamic contrast-enhanced MRI texture features," Br J Radiol, vol. 90, no. 1077, p. 20170269, 2017.

[7] S. Rosati, K. Meiburger, G. Balestra et al., "Carotid wall measurement and assessment based on pixel-based and local texture descriptors," Journal of Mechanics in Medicine and Biology, vol. 16, no. 1, p. 1640006, 2016.

[8] M. Terry, J. McDonald, H. Wu, S. Eng and R. Santella," Epigenetic biomarkers of breast cancer risk: across the breast cancer prevention continuum," Adv Exp Med Biol, vol. 882, pp. 33–68, 2016.

[9] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, et al., "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012," International journal of cancer, vol. 136, 2015

[10] J. Ma, Y. Qiao, G. Hu, Y. Huang, A. K. Sangaiah, C. Zhang, et al., "De-Anonymizing Social Networks With Random Forest Classifier," IEEE Access, vol. 6, pp. 10139-10150, 2018.

[11] Joshi, Jahanvi, Rinal Doshi, and Jigar Patel. "Diagnosis of breast cancer using clustering data mining approach." International Journal of Computer Applications 101.10 (2014).

[12] Chaurasia, Vikas, and Saurabh Pal. "A novel approach for breast cancer detection using data mining techniques." International Journal of Innovative Research in Computer and Communication Engineering 2.1 (2014): 2456-2465.

[13] R. Siegel, J. Ma, Z. Zou, and A. Jemal, “Cancer statistics, 2014,” CA: a cancer journal for clinicians, vol. 64, no. 1, pp. 9–29, 2014.

[14] M. Milosevic, D. Jankovic, and A. Peulic, “Comparative analysis of breast cancer detection in mammograms and thermograms,” Biomedical Engineering/Biomedizinische Technik, vol. 60, pp. 49–56, 2014

[15] Kylili, P. A. Fokaides, P. Christou, and S. A. Kalogirou, “Infrared thermography (irt) applications for building diagnostics: A review,” Applied Energy, vol. 134, pp. 531–549, 2014