Machine Learning based System for Prediction of Breast Cancer Severity

Sara LAGHMATI
Departement of computer science
Faculty of science (FS)
Chouaib Doukkali University (UCD)
El Jadida, Morocco
saralaghmati@gmail.com

Amal TMIRI
LaROSERI Laboratory
Faculty of science (FS)
Chouaib Doukkali University (UCD)
El Jadida, Morocco
b_tmiri@yahoo.fr

Bouchaib CHERRADI
STICE Team, CRMEF-El Jadida
LaROSERI laboratory, FS
Chouaib Doukkali University (UCD)
El Jadida, Morocco
cherradi1@hotmail.com

***Abstract*—Breast cancer is one of the most common diseases and the leading cause of death to mostly females all over the world. Early detection can provide higher treatment efficiency and better healing chances. Even though mammography screening is handy in diagnosing breast cancer at an early stage, Computer-Aided Diagnosis (CAD) systems can help to reduce the cancer death-rate. Radiologists, physicians, and doctors, in general, make use of these CAD systems to diagnose, detect, analyze and make decisions whether the patient is benign or malignant. The present paper presents some data mining techniques used in the diagnosis of cancer such as Artificial Neuron Network (ANN), K-Nearest Neighbors (KNN), Binary Support Vector Machine (Binary SVM), and Decision Tree (DT). Within this framework, the database utilized is the Mammographic Mass dataset. This database contains data of probabilistic breast cancer patients and the advanced results by experts in the field. The paper adopts a confusion matrix for binary prediction as a method of data analysis. The present paper provides a comparison between the different Computer-Aided diagnosis systems techniques regarding accuracy, specificity, and sensitivity amidst many other criteria to find the most accurate alternative among ANN, KNN, Binary SVM, and DT.**

Keywords—Breast cancer, Computer-aided Diagnosis, Artificial Neuron Network, K-Nearest Neighbors, Binary Support Vector Machine, Decision Tree, Database.

I. INTRODUCTION

Breast cancer is the most commonly diagnosed cancer and the leading cause of cancer death of women around the globe [1]. In the US alone, they will be about 268600 newly diagnosed female breast cancer cases and 41760 deaths for 2019[2]. Early detection is crucial; it can reduce the breast cancer mortality rate [3]. To diagnose breast cancer at its different stages; various imaging techniques could be used such as mammography, magnetic resonance imaging (MRI), positron emission tomography (PET), Computed tomography (CT), and single-photon emission computed tomography (SPECT) [4].

The quest for early detection has set a mandatory use of machine learning techniques and recently CAD systems have become ubiquitous part of routine clinical work especially for breast cancer detection on mammograms at many screening sites and hospitals [5,6]. Since 1980, and through the years, CAD systems survived face to many challenges to develop a certain degree of dependency and become more accurate in terms of increasing diagnosis rate and decreasing incorrectly diagnosed cases [7,8]. The classification of cancer based on CAD systems may use as an input statistics attributes from data set which regroups different kind of features such as dynamic features, textural features, and morphological features [9].

Machine learning (ML) techniques, such as Artificial Neural Network, K-Nearest Neighbours, and Decision Tree, are widely used to predict and pre-diagnose frequent diseases like cancers, hepatitis, and heart diseases [10, 11, 12].

In this paper, four of the ML techniques; Artificial Neural Network, K-Nearest Neighbours, Decision Tree, and SVM are analysed and applied on a data set of mammography patients to predict the severity of breast cancer. The contribution of this paper is to provide a benchmarking on the four models and gives the best systems based on many performance evaluations measures such as accuracy, sensitivity, specificity, and others.

The rest of this paper is organized as follows: In section II, the paper presents an overview of some relevant related work in the field of cancer diseases especially breast cancer. In section III, the specifications and characteristics of the dataset utilized and the ML algorithms used are presented. Section IV gives details of the implementation and the results. Finally, section V provides a conclusion to this paper.

II. RELATED WORK

Over the last few years, the majority of scientists have embraced and developed new various ML techniques that aim to predict cancer accurately and effectively [5]. Many efforts have been spent on breast cancer diseases to develop and test ML techniques on different datasets [13].

This section presents some of the research work related to this field with the type of data exploited, the predictive models used, and the performance measurement.

Uma Ojha et al. [14] have compared various classifier algorithms on WPBC dataset. Their results demonstrate the classification algorithms; C5.0 and SVM have shown 81% accuracy in classifying the recurrence of the disease.

Vikas Chaurasia et al. [15] used three different data mining algorithms Naïve Bayes, RBF Network and J48 to develop prediction models using Wisconsin dataset (683 breast cancer cases) they also used 10-fold cross-validation methods to compare the performance of the three prediction models. The results indicated that the Naïve Bayes is the best predictor with 97.36% accuracy.

Mohammed H. Tafish et al. [16] proposed a model to help in resolving the difficulty of determining the degree of risk for the disease and to get best practices, abatement time and expense with the objective of advancing wellbeing, based on data collected from hospitals in the Gaza Strip. The model is applying classification techniques such as SVM, ANN and KNN on the collected breast cancer data, which in turn

978-1-7281-2625-8/19/$31.00 ©2019 IEEE

predicts the severity of breast cancer. After evaluation and testing using the mentioned classification techniques on the breast cancer dataset, they obtained an accuracy of 77%.

Abdel-Zaher, A. M. et al. [17], developed a CAD scheme for detection of breast cancer using deep belief network unsupervised path followed by back propagation supervised path. The construction is based on back-propagation neural network with learning function of Liebenberg Marquardt and weights initialized from the deep belief network path (DBN-NN). They tested their technique on the Wisconsin Breast Cancer Dataset (WBCD). The classifier complex gives an accuracy of 99.68%.

III. MATERIALS AND METHODS

A. Mammographic mass dataset

The data used in this study are Mammographic Mass dataset. The dataset has 6 Attributes in total, this data set can be used to predict the severity (benign or malignant) of a mammographic mass lesion from Breast Imaging-Reporting and Data System (BI-RADS) attributes and the patient's age. It contains attributes together with the severity field for 516 benign and 445 malignant masses that have been identified on full field digital mammograms collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006 [18,19]. The dataset can be used as an indication of a CAD system efficiency compared to radiologists. The data contain the following variables/features.

1. BI-RADS assessment: Breast Imaging-Reporting and Data System 1 to 5 (ordinal, non-predictive!)
2. Age: patient's age in years (integer).
3. Shape: mass shape differs from round=1 to oval=2 to lobular=3 or irregular=4, (nominal).
4. Margin: mass margin and it represent the nature of tissues surrounding the tumour: circumscribed=1, micro lobulated=2, obscured=3, ill-defined=4, spiculated=5 (nominal).
5. Density: mass density varies from almost all fatty tissue to extremely dense tissue with very little fat high=1 iso=2 low=3 fat-containing=4 (ordinal).
6. Severity: the nature of the tumor benign=0 or malignant=1 (binominal, goal field!).

In [18] we find more details on main dataset features and their values range.

B. Machine learning algorithms for prediction

1) Artifitial Neural Network (ANN)

An artificial neural network is a model that its design is inspired by, not necessarily identical to, the functioning and the organization of biological neurons [20,21]. An artificial neural network is based on connections between many different processing elements each corresponding to a single neuron (node), each neuron takes many input signals then based on an internal weighting gives a single output that acts as input to another neuron. The architecture is divided into different layers: The input layer receives the inputs, the output layer produces the final output, and between them one or more hidden layers [22]. Fig. 1 shows a diagram of an artificial neuron model.

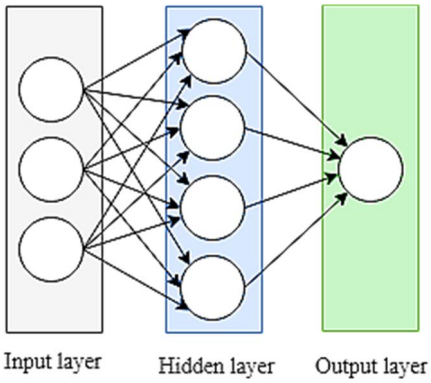

Fig. 1. Diagram of an artificial neuron model with 1 hidden layer.

2) K-Nearest Neighbours (KNN)

The K-Nearest Neighbours search technique and KNN-based algorithms are widely used as benchmark learning rules. The basic idea is to identify the class of a compound based on the class of the k most similar compounds, where similarity is defined by calculating the Euclidean distances (Eq. (1)) between the descriptor vectors [23].

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(P_i - q_i)^2} \quad (1)$$

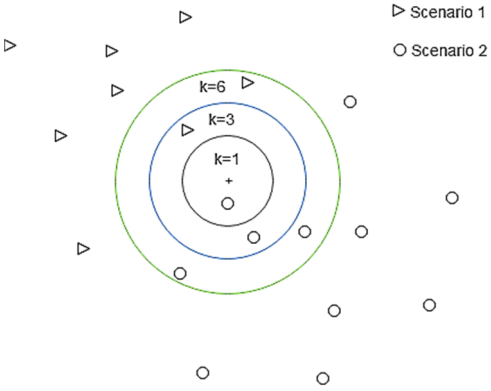

Fig. 2. Classification based on KNN.

3) Decision tree (DT)

A decision tree is a special type of tree that aim is to classify input data into one of several classes. Each internal node of a decision tree contains a *decision rule*, and each leaf node contains a class label [24]. Here is a simple decision tree with 4 decision nodes and 5 leaf nodes.

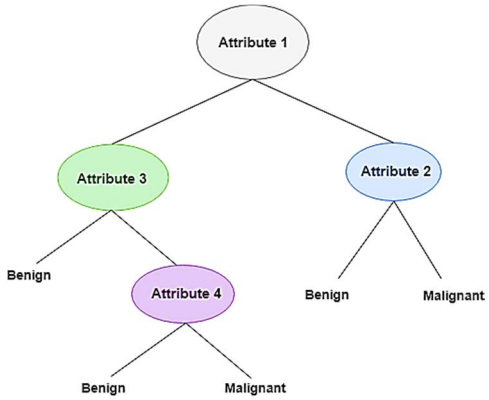

Fig. 3. A two level decision tree sample.

4) Binary SVM

A support vector machine (SVM) in binary classification is a concept in statistics and computer science for a set of related supervised learning methods that analyses data and recognizes patterns, used for classification and regression analysis [25]. SVM is used in the cases of data with two different scenarios. it separates points of one class from the others by the best hyper lane (where the margin is the largest). The following figure illustrates a classification by SVM where triangle indicates data points of type1 and circles indicate data points of type 2.

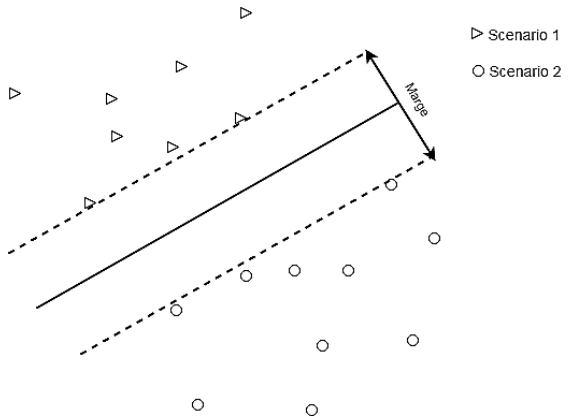

Fig. 4. SVM sample for separable data.

C. Performance evaluation methods

1) Confusion matrix for binary prediction

In the machine learning field and more specifically the statistical classification problem, a confusion matrix is a particular table layout that grants visualization of the performance of an algorithm, typically a supervised learning one. Each column represents the instances in an actual class while each row of the matrix represents the instances in a predicted class (or vice versa) [26]. It is a table with two rows and two columns that reports the number of false positives FP, false negatives FN, true positives TP, and true negatives TN with:

True positive (TP): Sick patients correctly identified as sick

False positive (FP): Healthy patients incorrectly identified as sick

True negative (TN): Healthy patients correctly identified as healthy

False negative (FN): Sick patients incorrectly identified as healthy

TABLE I. CONFUSION MATRIX FOR BINARY PREDICTION

Predicted output/labels (Ground truth)	**Algorithm outputs/labels**	
	Patient without disease (Negative/0)	***Patient with disease (Positive/1)***
Patient without disease (False/0)	TN	FP
Patient with disease (True/1)	FN	TP

2) Performance evaluation measures

- Sensitivity

Sensitivity relates to the test's ability to correctly detect ill patients who do have the condition [27].

$$\text{Sensivity} = \frac{\text{TN}}{\text{FN} + \text{TN}} \tag{2}$$

- Specificity

Specificity refers to the test's ability to correctly reject healthy patients without a condition [27].

$$\text{Specificity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

- Accuracy

The accuracy represents the proportion of true results, both true positives and true negatives, among the total number of cases examined [28].

$$\text{Accuracy} = \frac{\text{TP} + \text{FN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{4}$$

- Matthews's correlation coefficient MCC

MCC Measure the quality of binary can be used even if the classes are of very different sizes unlike accuracy [28].

$$\text{MCC} = \frac{\text{TP} * \text{TN} - \text{FP} * \text{FN}}{\sqrt{(\text{TP} + \text{FP}) * (\text{TP} + \text{FN}) * (\text{TN} + \text{FP}) * (\text{TN} + \text{FN})}} \tag{5}$$

- The Jaccard Similarity coefficient

The Jaccard Similarity coefficient is used to algorithm to work out the similarity between two things [28].

$$\text{Jaccard} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \tag{6}$$

- Dice similarity

The Dice coefficient, just like Jaccard, is a statistic used to gauge the similarity of two samples but in a more semi metrical way [29].

$$\text{Dice} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}} \tag{7}$$

IV. RESULTS AND DISCUSSION

A. Training and algorithms best configuration

The dataset was divided into training, validation and testing sets with a ratio of 70%, 15%, and 15% respectively. The procedure has been repeated 30 times. TABLE II. III. IV. AND V. regroup different parameters used for the training of the algorithms.

TABLE II. PROPOSED ANN ARCHITECTURE AND TRAINING PARAMETERS

Dataset	Training dataset: 70% Validation dataset: 15% Testing dataset: 15%
Layers number	1
Weights and bias	Randomlyinitialized
Number of neurons/layers	Input: 5 Hidden: 20 Output: 1
Learning rule	Levenberg–Marquardt
Learning rate	0.001

TABLE III. PROPOSED KNN ARCHITECTURE AND TRAINING PARAMETERS

Kmax	21
Learn rate	1
Cross validation Losskloss	0.1844
Resubstitution Loss rLoss	0.1738
Evaluation of the generated model quality E	0.4869
N learn	10

TABLE IV. PROPOSED DT ARCHITECTURE AND TRAINING PARAMETERS

Optimization of model DT	
Total elapsed time:	26.1539 seconds
Observed objective function value	0.14458
Function evaluation time	0.051076
Best estimated feasible point	MinLeafSize: 30
Estimated objective function value	0.1445
Estimated function evaluation time	0.060423

TABLE V. PROPOSED BSVM ARCHITECTURE AND TRAINING PARAMETERS

Optimization of model BSVM	
Total elapsed time:	218.067 seconds
Observed objective function value	0.15835
Function evaluation time	1.1889
Best estimated feasible point	BoxConstraint:2.5976 KernelScale: 0.57175
Estimated objective function value	0.15817
Estimated function evaluation time	0.56694

As the training goes on, Fig. 5 describes how the loss decreases with function evaluation.

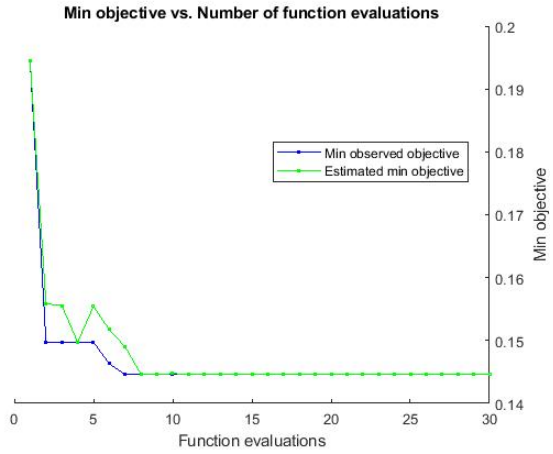

Fig. 5. Min objective vs Number of functions evaluations for DT.

After taking the best combination between KemelScale and BoxConstraint values to find the objective function model for SVM the loss decreases with more evaluation as it is shown in Fig. 6.

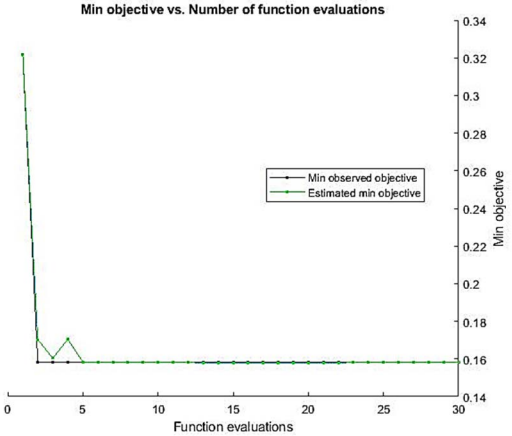

Fig. 6. Minobjective vs Number of functions evaluations for BSVM.

B. *Comparison of the testing results*

Among the four algorithms applied on the 249 patients of the testing data, Binary SVM gives, the highest rate of TN and FP indicates 133 healthy people correctly identified as healthy, 49 healthy people incorrectly identified as sick, respectively, also the least number of FN only 2 people ill incorrectly identified as healthy. On the other hand, ANN provides the highest number of TP or 99 patients correctly identified as sick and the lower number of FP 15 people healthy incorrectly identified as ill patients. As for KNN, it provides the second-highest TP number and the highest FN number equal to 32 people ill incorrectly identified as healthy. The decision tree gives the second-highest TN number and second-lower FN number. TABLE VI represents the confusion matrix values for each algorithm.

TABLE VI. CONFUSION MATRIX OF BREAST CANCER DISEASES PREDICTION USING ML ALGORITHMS TESTED ON 249 PATIENTS FROM MAMMOGRAPHIC MASS DATASET.

	TN	**FN**	**TP**	**FP**
ANN	112	23	99	15
KNN	114	32	82	21
DT	128	7	76	38
Binary SVM	133	2	65	49

The results show that Binary SVM has the best specificity and that ANN is better than the other models in terms of accuracy, sensitivity, MCC, Jaccard, and Dice. In other words, BSVM is the best model with the highest probability to correctly reject healthy patients. ANN is the most efficient model to correctly detect ill patients, it is the best system in providing the proportion of true results among the number of testing patients, and it is the most likely to differentiate between healthy and ill patients. TABLE VII. regroups the values of the different statistic parameters (Accuracy, Sensitivity, Specificity, MCC, Jaccard, and Dice) of the four models.

TABLE VII. EVALUATION PERFORMANCE MEASURE OF ML ALGORITHMS TESTED ON 249 PATIENTS FROM MAMMOGRAPHIC MASS DATASET.

	ANN	**KNN**	**DT**	**Binary SVM**
Accuracy	0.84	0.78	0.81	0.79
Sensitivity	0.86	0.79	0.66	0.57
Specificity	0.82	0.78	0.94	0.98
MCC	0.69	0.57	0.64	0.62
Jaccard	0.72	0.60	0.62	0.56
Dice	0.83	0.75	0.77	0.71

C. *Discussion*

In recent years, many studies have tackled the problem of breast cancer detection and classification based on CAD systems. Many ML models have been developed, however improving accuracy seems to be always an opened topic for research.

The present study represents a benchmarking of several ML algorithms to classify breast cancer patients as sick or healthy. The results indicate that ANN is a more effective model in indicating breast cancer severity then SVM, KNN or DT, and it shows an accuracy of 84%. On the contrary, the results provided in section II, which claims that SVM is better than ANN and KNN. The difference in the results can be due to the use of different datasets or different parameters.

V. CONCLUSION AND PERSPECTIVES

This paper presents an overview of some Machine learning techniques; ANN, KNN, Binary SVM, and decision tree. The four CAD techniques applied to the mammographic mass dataset; groups together 5 statistical attributes as input and a single binary output. The data analysis and classification are based on the confusion matrix to extract the number of FN, TP, TN and FP. The results show that ANN is more reliable to assist in decision making regarding breast cancer severity.

CAD systems are far from reaching their peak and are always open to further research to develop and optimize the algorithms to achieve the best combination between the accuracy, the specificity, the sensitivity, and the other performance measures.

ACKNOWLEDGMENTS

This study is achieved by the support of National Centre for scientific and technical research (CNRST), Morocco within the Research Excellence Program.

REFERENCES

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." CA Cancer J Clin. (2018). https://doi.org/10.3322/caac.21492

[2] Rebecca L. Siegel, Kimberly D. Miller, Ahmedin Jemal DVM "Cancer statistics, 2019"(2019). https://doi.org/10.3322/caac.21551

[3] Oeffinger KC, Fontham ETH, Etzioni R, et al. Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. *JAMA*. 2015;314(15):1599–1614. doi:10.1001/jama.2015.12783

[4] Jafari SH, Saadatpour Z, Salmaninejad A, et al. Breast cancer diagnosis: Imaging techniques and biochemical markers. J Cell Physiol. 2018;233:5200–5213. https://doi.org/10.1002/jcp.26379

[5] Kumar, Ajay and Sushil, Rama and Tiwari, Arvind, "Cancer Survival Analysis Using Machine Learning" (2019). Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur - India, 2019. Available at SSRN: http://dx.doi.org/10.2139/ssrn.3354469

[6] Timothy W. Freer, Michael J. Ulissey, "Screening Mammography with Computer-aided Detection: Prospective Study of 12,860 Patients in a Community Breast Center" (2001). https://doi.org/10.1148/radiol.2203001282

[7] Leichter, I., Fields, S., Nirel, R. et al. Eur Radiol, "Improved mammographic interpretation of masses using computer-aided diagnosis" (2000) 10: 377. https://doi.org/10.1007/s003300050059

[8] Doi, K. (2007). "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential." Computerized Medical Imaging and Graphics. https://doi.org/10.1016/j.compmedimag.2007.02.002

[9] Fusco, R., Sansone, M., Filice, S. et al. J. Med. Biol. Eng. (2016) 36: 449. https://doi.org/10.1007/s40846-016-0163-7

[10] El Houby, E. M. F. (2018). "A survey on applying machine learning techniques for management of diseases." Journal of Applied Biomedicine. https://doi.org/10.1016/j.jab.2018.01.002

[11] Oumaima Terrada, Bouchaib Cherradi, Abdelhadi Raihani, Omar Bouattane, " A fuzzy medical diagnostic support system for cardiovascular diseases diagnosis using risk factors", in 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), 2018, p. 1‑6. DOI: 10.1109/ICECOCS.2018.8610649

[12] Oumaima Terrada, Bouchaib Cherradi, Abdelhadi Raihani, Omar Bouattane, "Classification and Prediction of atherosclerosis diseases using machine learning algorithms", in 2019 5th International Conference on Optimization and Applications (ICOA), 2019, p. 1‑5. DOI: 10.1109/ICOA.2019.8727688

[13] Jafari-Marandi R, Davarzani S, Gharibdousti MS, Smith BK, "An Optimum ANN-based Breast Cancer Diagnosis: Bridging Gaps between ANayN Learning and Decision-making Goals", Applied Soft Computing Journal (2018). https://doi.org/10.1016/j.asoc.2018.07.060

[14] Uma Ojha ; Savita Goel(2017) "A study on prediction of breast cancer recurrence using data mining techniques" 2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence. DOI: 10.1109/CONFLUENCE.2017.7943207

[15] V Chaurasia, S Pal, BB Tiwari, (2018) "Prediction of benign and malignant breast cancer using data mining techniques". https://doi.org/10.1177/1748301818756225

[16] M. H. Tafish and A. M. El-Halees, "Breast Cancer Severity Degree Predication Using Data Mining Techniques in the Gaza Strip," *2018 International Conference on Promising Electronic Technologies (ICPET)*, Deir El-Balah, 2018, pp. 124-128. doi: 10.1109/ICPET.2018.00029

[17] Ahmed M.Abdel-Zaher, Ayman M.Eldeib, (2016) "Breast cancer classification using deep belief networks" https://doi.org/10.1016/j.eswa.2015.10.015

[18] https://www.kaggle.com/overratedgman/mammographic-mass-data-set/home

[19] https://sci2s.ugr.es/keel/dataset.php?cod=86

[20] Noppakorn Klintong, Pakpachong Vadanasin, Natcha Thawesaengskulthai,(2011). "Decision Support System using Artificial Neural Network for Managing Product Innovation" International Journal of the Computer, the Internet and Management Vol. 19. No.3.

[21] TimHilla, LeoreyMarquezb, MarcusO'Connorc, WilliamRemus, "Artificial neural network models for forecasting and decision making"(1994). https://doi.org/10.1016/0169-2070(94)90045-0

[22] Amato, F., López, A., Peña-Méndez, E. M., Vaňhara, P., Hampl, A., & Havel, J. (2013). Artificial neural networks in medical diagnosis. Journal of Applied Biomedicine. https://doi.org/10.2478/v10136-012-0031-x

[23] Asikainen, A., Kolehmainen, M., Ruuskanen, J., & Tuppurainen, K. (2006)."Structure-based classification of active and inactive estrogenic compounds by decision tree, LVQ and kNN" https://doi.org/10.1016/j.chemosphere.2005.04.115

[24] Quinlan, J.R. Mach Learn (1986) 1: 81. https://doi.org/10.1007/BF00116251

[25] Hetal Bhavsar, Amit Ganatra "Support Vector Machine Classification using Mahalanobis Distance Function " (2015) International Journal of Scientific & Engineering Research, Volume 6, Issue 1, January-2015.

[26] Ting K.M. (2017) Confusion Matrix. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA.

[27] Fawcett, T. & Flach, P. Mach Learn (2005) 58: 33. https://doi.org/10.1007/s10994-005-5256-4

[28] Powers, David & , Ailab. (2011). "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation". J. Mach. Learn. Technol. 2. 2229-3981. 10.9735/2229-3981.

[29] David Gutman1, Noel C. F. Codella, Emre Celebi , Brian Helba , Michael Marchetti , Nabin Mishra, Allan Halpern, (2016) "Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC)".