Breast Cancer Classification Using Ml Algorithms

Amira Khattab
Faculty of computers
and information
Menofya university

Sara A.Shehab
Faculty of computers
and artificial intelligence
Sadat university

Arabi Keshk
Faculty of computers
and information
Menofya university

Walid Attwa
Faculty of computers
and information
Menofya university

Abstract: *One of the most top diseases nowadays is breast cancer that causes death for many women over the world. Breast cancer is a heterogeneous disease defined by molecular types and subtypes. Artificial intelligence has an effect role in detecting and classification the breast cancer. In this work 13 classification method are used like Support Vector Machine, AdaBoost, MLP classifier and others. This work is evaluated using three keys accuracy, cross validation score and execution time. The results detect that Linear SVC Support Vector Machine achieved high accuracy (98.25%) and Random Forest and AdaBoost achieved high cross validation score (97.01%) when compared with other classification methods. Whereas Gaussian NB classifier achieved minimum execution time (0.01 seconds). A data set with 31 feature and 570 records are used for testing the algorithms. 20% of data set will be used in testing and 80% for training. The proposed work achieves high accuracy when compared with the previous works.*

Keywords: *Breast Cancer, Artificial Intelligence, Support Vector Machine, AdaBoost, MLP Classifier.*

1. Introduction

Breast Cancer has been raised in women with mortality rate according to many scholars mentions. In 2018, World Health Organization (WHO) detect that approximately 627,000 females have been died with this cancer and predict that this number will be maximized 2, 7 million in 2030[1]. The survival rate was very low according to the late discovery and complexity procedures of this disease. The risk of developing the cancer can be done when detecting the breast cancer earlier [2]. Creation of the abnormal cells in the body causes the cancer and can be spread speedy in the body. The latency in treatment and diagnosis causes dead. There are two types of breast cancer the first one is invasive and the second is non-invasive [3]. Machine Learning (ML) is a part of artificial intelligence. Machine Learning algorithms build a model based on sample data, known as training data, in order to make prediction or decisions without being explicitly programmed to do [4] so ML is used

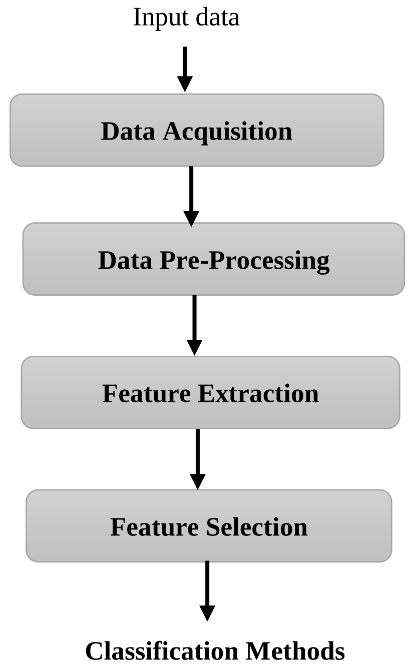
to give decision like human with high accuracy [5].ML includes 4 steps to employing the procedures (loading data, feature extraction, features selection and feature classification). Recently, that classification of tumors using machine learning has come into sharper focus [6-8]. Cancer is from uncontrolled and aberrant development of containers on account of consolidation of characteristic genetic and epigenetic defects. This unrestrained progress of the contributes to tumor incident. if the tumor starts to speedily metastasizes to added tools and wholes of the party as the cancer progresses ,the affection can then be incurable when found [9].Breast tumor generally affects wives (accompanying < 1%of cases moving nonwomen);roughly individual in eight mothers expand malignancy in their career[10].roughly 2.1 heap mothers are diagnosed accompanying bosom malignancy annually ,and ultimate harshly distressed are those between 40age or 70 age [11].Therefore , the early diseases of conscience malignancy is paramount to good forecast. Even though the syndromes concede possibility be weak in the inception, chances of endurance efficiently increase if detected early [12]. the miscellaneous protect forms used to investigate breast tumors contain fine teases aspiration plant structure (FNAC), ultrasound led surgical medical check-up and mammography. In dense feelings, the rate of malignancy discovery utilizing mammography is very poor and about 10 % to 30% of test cases go unfound [13], [14]. It is the main to recognize effective early malignancy diseases and situation to increase the survival rate of tumour inmates [15-18]. Machine learning is individual of ultimate well-known machines trained the models in an easily way and constitute predicting models for profitable in charge. machine Learning plays an important role in accompanying early diseases of breast malignancy and decides the type of the malignancy by analysing the tumour length. Machine learning forms are one of the approaches that lead to acquire favourable consequences with categorization and prophecy questions. Recently research in breast cancer keep benefit from ML methods used to recognize tumor and think the appearance or absence of tumors. Methods in machine learning can too be used to anticipate tumor virulence [19],[20]. To monitor and pronounce the diseases, normal forms second hand are very established the discovery of the presence of distinguishing signal facial characteristics by a human Spector. In the past ten of something, the growth of various computer-helped diseases(CAD)approaches has existed cued by many sufferers in exhaustive care unites needing perpetual listening. The mainly concerning qualities not quantities disease criteria are curve into more factual all features inclusive to surpass the issues of categorization in these methods [21],[22]. Three different data set are used in testing the multiple machines learning classifier, from the viewpoint of categorization veracity. Classifiers like subsequent littlest optimization and multilayer (MLP) perceptron were still contained in the study other than classifiers to a degree decision Trees, IBK and Naïve Bayesian. The detailed in above were proven on miscellaneous datasets that chiefly included the original data set (WBCD) Wisconsin Breast Cancer. Wisconsin Datasets are different in the way that the Wisconsin (WPBC) Prognostic breast cancer data and Wisconsin (WPBC) Diagnostic Breast Cancer data set, for the Diagnostic Breast Cancer data set the J48 and MLP classifiers were melded and helped by the feature selection order that shown a taller veracity rate than non-melded versions. For the Prognostic breast cancer data set, the mixture of all earlier ML algorithms shown a better veracity in verdict tumor than conventional plans, while the submitted SMO confirmed expected a more exact and better approach for Diagnostic Breast Cancer data set . Although, salaam et al. submitted a multi-classifier by calculation out that individual of the melded classifiers would support the best depiction foe some special dataset [23],[24]. The Model of the breast cancer accompanying Machine Learning algorithms is presented. The proposed model maybe secondhand for the detection of favorable and diseased malignancy containers. In the first step, the image of breast dossier is intoxicated, then the next step that takes place is Feature Extraction, and therefore that the last categorization model maybe prepared to load to full-fill task established above. Malignant cancer starts accompanying uneven container progress and can quickly pollute or pervade the encircling tissues, that form it a mortal condition; in another hand benign tumors are though-out non-malignant mostly non-deadly [20,25].

2. Related Work

Nowadays, the most important branch used in healthcare field is Artificial intelligence. It helps in improving the quality of care and providing safety. To classifying and identifying the tumours in the brain and breast the two branches of deep learning and machine learning is used [26]. In 2016, H.Asri proposed a classification for the breast cancer using four algorithms C4.5, K-NN,SVM and NB. The Wisconsin Breast Cancer data set is used. It consists of 699 instance and 11 attributes. The experimental result proved that SVM achieved high accuracy when compared with other three algorithms reached to 97.13%2[27]. In 2017, M.W.Hunang modified the SVM algorithm with boosting and bagging features and three functions. This paper used two data sets the first data set contains 117 attributes with 102294 instances and the second data set contains 699 instances and 11 attributes. The results proved that the accuracy is 96.85% and 95% [28]. In 2018, Y. K and M.Bahaj proposed four ML algorithms Naïve Bayes, KNN, SVM and Random Forest. The evaluation for these algorithms done using data set contains 699 instances with 30 attributes. The experimental results proved that SVM score high accuracy reached to 97.9% [29][30]. In 2020, R.Rawal use Wisconsin Prognostic Breast cancer data set with 32 attributes and 194 instances. A comparison between K-mean, EM, PAM, and fuzzy c-means with SVM and C5.0. the results approved that SVM achieved high accuracy reached to 97% [31]. In 2021, N.AI-Azzam compares the (SSL) semi supervised learning with (SL) supervised learning using different nine algorithms RBF Support vector machine, Logistic regression, Gaussian Naive Bayes, Decision Tree, Random Forest, Xgboost, Gradient Boosting, Linear Support vector machine and KNN. A data set with 30 attributes and 569 instances used for evaluation. The results detected that high accuracy for KNN (SL=98% &SSL= 97%) and logistics regression (SL=97% & SSL=98%) [32].

3. Classification Stages and Data Preparation

In this paper 4 stages will be covert to classify the breast cancer data set. The stages are Data acquisition, Data pre-processing, feature extraction and feature selection.



3.5 Data Acquisition

In this stage, firstly the data set is selected to train the model and secondly it is evaluated. There are many available data sets for breast cancer used for the classification. The data set used in this paper with 31 attributes and 570 instances. 20% of data set will be used in testing and 80% for training.

3.5 Data Pre-Processing

In the next stage the redundant and irrelevant data will be removed to get high performance from ML algorithms. Table 1 list the features of the data used in the paper. The duplicated data row will be removed. Find the cells in the data set which has no value and fill it with appropriate value. Converting all string data to numeric value as the ML algorithms can't deal with strings. The data set will be splitting into two parts training and testing, in this paper it will be 80% and 20% for training and testing respectively.

Table 1 Features of the data set

No	Features	No	Features
1	radius_mean	16	compactness_se
2	texture_mean	17	concavity_se
3	perimeter_mean	18	concave points_se
4	area_mean	19	symmetry_se
5	smoothness_mean	20	fractal_dimension_se
6	compactness_mean	21	radius_worst
7	concavity_mean	22	texture_worst
8	concave points_mean	23	perimeter_worst
9	symmetry_mean	24	area_worst
10	fractal_dimension_mean	25	smoothness_worst
11	radius_se	26	compactness_worst
12	texture_se	27	concavity_worst
13	perimeter_se	28	concave points_worst
14	area_se	29	symmetry_worst
15	smoothness_se	30	fractal_dimension_worst

3.5 Feature Extraction

In this stage, the features will be minimized to small new features with the same information as the old one. This step help in improving the accuracy of the ML algorithms and will reduce the risk of overfitting. It will also increase the training speeds.

3.5 Feature Selection

Selecting the features technique loads to effective classification, which rank the features from most important features to least important features and reduce the number of features. This reduction helps in statistical analysis e.g., increase training speed, improve accuracy and reduce the risk of overfitting. Several methos will be used as embedded method, wrapper method and filter method. The techniques involving removing and adding some features based on the performance of the model.

4. Experimental Result

The classification algorithms implemented using python under windows with Core i5. In this paper data set with 31 features and 570 records will be used. In this data set the code used 20% of this data for testing and 80% for training. The classification algorithms used in the implementation 13 algorithms. Gaussian NB, Dummy, K-Neighbours, Random Forest, Logistic Regression, Support Vector Machine, NuSVC Support Vector Machine, Linear SVC Support Vector Machine, SGD, MLP, AdaBoost, Gaussian Process, and Decision Tree classifiers. In the comparison three key comparisons are used: Accuracy, Cross Validation and Execution Time. The results detect that Linear SVC Support Vector Machine achieved high accuracy (98.25%) and Random Forest and AdaBoost achieved high cross validation score (97.01%) when compared with other classification methods. Whereas Gaussian NB classifier achieved minimum execution time (0.01 seconds). Table 2 List the comparison between 13 classifier algorithms with the three key parameters. Figures 1, 2 and 3 state the Accuracy, Cross validation and Execution time respectively.

Table 2 Difference Between 13 Classifier Algorithms

Classifier	Execution Time (seconds)	Cross Validation	Accuracy
1. Gaussian NB	0.015623	93.85%	97.36%
2. Dummy	0	62.74%	62.28%
3. K-Neighbours	0.078106	92.79%	95.61%
4. Random Forest	0.82739	97.01%	96.49%
5. Logistic Regression	0.14059	95.08%	96.49%
6. Support Vector Machine	0.046863	91.22%	94.74%
7. NuSVC Support Vector Machine	0.062486	91.22%	91.23%
8. Linear SVC Support Vector Machine	0.078094	91.22%	98.25%
9. SGD	0.031244	90.51%	94.74%
10. MLP	0.15621	91.39%	95.61%
11. AdaBoost	0.69518	97.01%	97.37%
12. Gaussian Process	14.242	95.43%	96.49%
13. Decision Tree	0.046864	92.26%	93.86%

5. Conclusion

In this paper 13 Machine Learning classification algorithms are used. A data set with 31 features and 570 instance are used for evaluation. 20% of data set will be used in testing and 80% for training. The data firstly acquisition, then it is pre-processing after that the two stages feature extraction and feature selection. The results conclude that Linear SVC Support Vector Machine achieved high accuracy (98.25%) and Random Forest and AdaBoost achieved high cross validation score (97.01%) when compared with other classification methods. Whereas Gaussian NB classifier achieved minimum execution time (0.01 seconds).

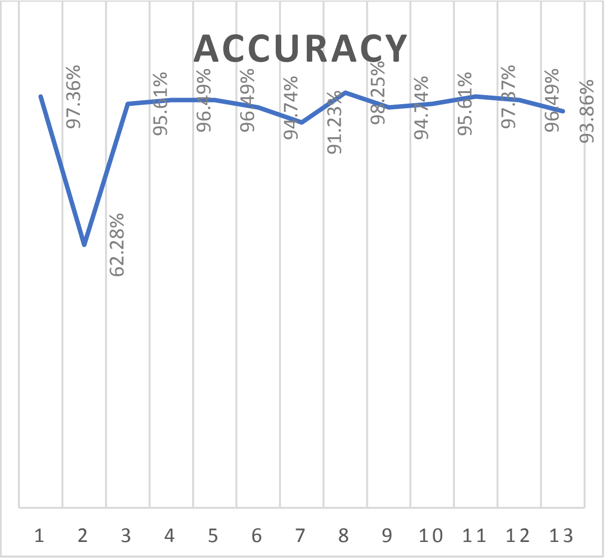


Figure 1. Accuracy for 31 classification method

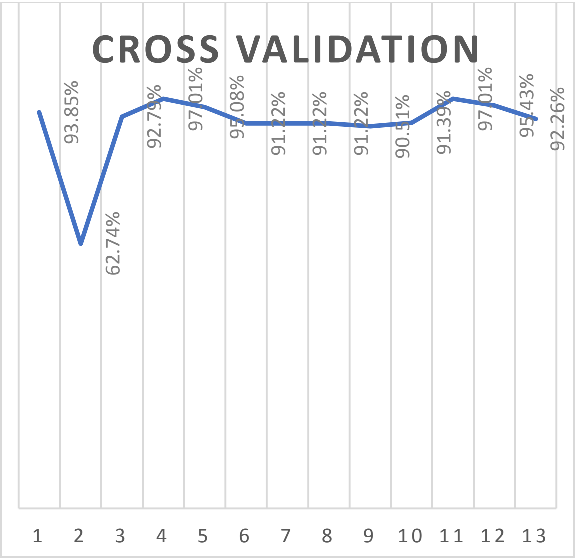


Figure 2. Cross Validation for 31 classification method

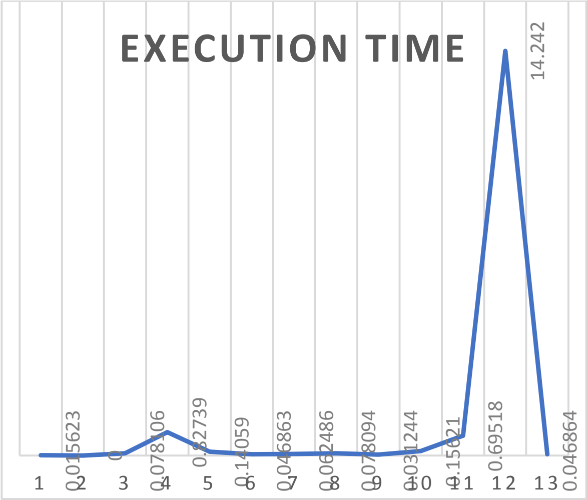


Figure 3. Execution time for 31 classification method

References

[1] “WHO | World Health Organization.” [Online]. Available: https://www.who.int/. [Accessed: 28-Jun-2021].

[2] D. Bardou, K. Zhang, and S. M. Ahmad, “Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks,” IEEE Access, vol. 6, pp. 24680–24693, 2018.

[3] Priyanka and K. Sanjeev, “A review paper on breast cancer detection using deep learning,” IOP Conf. Ser. Mater. Sci. Eng., vol. 1022, no. 1, 2021.

[4] G. Battineni, N. Chintalapudi, and F. Amenta, “Performance analysis of different machine learning algorithms in breast cancer predictions,” EAI Endorsed Trans. Pervasive Heal. Technol., vol. 6, no. 23, pp. 1–7, 2020.

[5] Op. P. Nave and M. Elbaz, “Artificial immune system features added to breast cancer clinical data for machine learning (ML) applications,” BioSystems, vol. 202, no. April, 2021.

[6] Mitra, M.; Mohadeseh, M.; Mahdieh, M.; Amin, “B. Machine learning models in breast cancer survival prediction”. Technol. Health Care, 24, 31–42, 2016.

[7] Tong, W.; Laith, R.S.; Jiawei, T.; Theodore, W.C.; Chandra,” M.S. Machine learning for diagnostic ultrasound of triple-negative breast cancer. Breast Cancer Res”. Treat, 173, 365–373, 2019.

[8] Riku, T.; Dmitrii, B.; Mikael, L.” Breast cancer outcome prediction with tumour tissue images and machine learning. Breast Cancer Res”, 177, 41–52, 2019.

[9] S Singh, A Deep, G Mohanta, *et al,* ”In next generation point-of-care biomedical sensors technologies for cancer diagnosis”, Springer, Singapore , pp. 253-278,2017.

[10]Mihaylov I, Nisheva M, Vassilev D, “Machine learning techniques for survival time prediction in breast cancer”. Lecture Notes in Computer Science Springer, Cham : 186–94, doi:10.1007/978-3-319-99344-7_17, 2018.

[11] H Benbrahim, H Hachimi, A, “AmineComparative study of machine learning algorithms using the breast cancer dataset Adv Intell Sys Comp”, 1103), pp. 83-91,2020.

[12]A Kamboj, P Tanay, A Sinha, et al.,”Breast cancer detection using supervised machine learning “,a comparative analysis Springer, Singapore , pp. 263-269,2021.

[13] RL Siegel, KD Miller, A. Jemal,” Cancer statistics, 2016 CA Cancer J Clin” , 66 (1) , pp. 7-30, 10.3322/caac.21332,2016.

[14] C Kaushal, A. Singla ,”Analysis of breast cancer for histological dataset based on different feature extraction and classification algorithms

“,Adv Intelligent Sys Comp, 1165 , pp. 821-833, 10.1007/978-981-15-5113-0_69 , 2021.

[15] JW. Uhr ,”Cancer diagnostics: one-stop shop”,Nature, 450 , pp. 1168-1169, 10.1038/4501168a, 2007.

[16] TR Geiger, DS. Peeper ,”Metastasis mechanisms,” Biochim Biophys Acta, 1796 , pp. 293-308, 10.1016/j.bbcan.2009.07.006, 2009.

[17] M Liberko, K Kolostova, V. Bobek,” Essentials of circulating tumor cells for clinical research and practice”, Critical Rev Oncol, 88 , pp. 338-356, 10.1016/j.critrevonc.2013.05.002,2013.

[18] K Wang, MQ He, FH Zhai, *et al., ”*A novel electrochemical biosensor based on polyadenine modified aptamer for label-free and ultrasensitive detection of human breast cancer cells”, Talanta, 166 , pp. 87-92, 10.1016/j.talanta.2017.01.052, 2017.

[19]I. Kononenko ,”Machine learning for medical diagnosis: history, state of the art and perspective”, Artif Intell Med, 23 , pp. 89-109, 10.1016/S0933-3657(01)00077-X, 2001.

[20] A. Al Bataineh ,”A comparative analysis of nonlinear machine learning algorithms for breast cancer detection”, Int J Mach Learn Comput, 9 , pp. 248-254, 10.18178/ijmlc.2019.9.3.794, 2019.

[21] NF Güler, ED Übeyli, I. Güler Recurrent neural networks employing Lyapunov exponents for EEG signals classification Expert Syst Appl, 29 , pp. 506-514, 10.1016/j.eswa.2005.04.011, 2005.

[22] A Osareh, B. Shadgar Proceedings of the 2010 5th international symposium on health informatics and bioinformatics, HIBIT , pp. 114-120, 10.1109/HIBIT.2010.5478895, 2010.

[23] Salama G, Abdelhalim MB, Zeid MA. Breast cancer diagnosis on three different datasets using multi-classifiers. 2012.

[24] A Ghasemzadeh, SS Azad, E. Esmaeili “Breast cancer detection based on Gabor-wavelet transform and machine learning methods “, Int J Mach Learn Cybern, 10 , pp. 1603-1612, 10.1007/S13042-018-0837-2, 2018.

[25] AA Nahid, Y. Kong ,”Involvement of machine learning for breast cancer image classification: a survey”, Comput Math Methods Med , Article 3781951, 2017.

[26] A. S. Assiri, S. Nazir, and S. A. Velastin, “Breast Tumor Classification Using an Ensemble Machine Learning Method,” J. Imaging, vol. 6, no. 6, p. 39, May 2020.

[27] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis,” Procedia Comput. Sci., vol. 83, no. Fams, pp. 1064–1069, 2016.

[28] M. W. Huang, C. W. Chen, W. C. Lin, S. W. Ke, and C. F. Tsai, “SVM and SVM ensembles in breast cancer prediction,” PLoS One, vol. 12, no. 1, pp. 1–14, 2017.

[29] Y. K. and M. Bahaj, “Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification,” in International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), pp. 1–5 , 2018.

[30] Y. K. and M. Bahaj, “Feature Selection with Fast Correlation-Based Filter for Breast Cancer Prediction and Classification Using Machine Learning Algorithms,” in 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), pp. 1–6 , 2018.

[31] R. Rawal, “BREAST CANCER PREDICTION USING MACHINE LEARNING,” J. Emerg. Technol. Innov. Res., vol. 7, no. 5, 2020.

[32] N. Al-Azzam and I. Shatnawi, “Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer,” Ann. Med. Surg., vol. 62, no. December 2020, pp. 53–64, 2021.