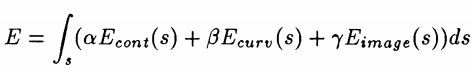
1. Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993, July). Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization* (Vol. 1905, pp. 861-870). International Society for Optics and Photonics.

Instead of using a full biopsy, Fine needle aspirations (FNAs) enable to examine a small amount of tissue from the tumor. For the diagnosis, they used image processing and machine learning techniques. First step was the finalizing boundary of cell nucleus by an active contour model, which is known as ‘snake’. The initial step is to draw approximate boundary which is rough initial outline. After the initial step, it finishes drawing boundaries. The interactive process took 2 ~ 5 minutes.

They use following formula to generate boundaries.



To determine the boundaries, they use 3 elements, which are continuity, curvature, Image. The computation will be conducted till all points settle into a local minimum of energy function. They extracted features from the image, which are radius, perimeter, area, compactness, smoothness, concavity, concave points, symmetry, fractal dimension, texture. Finally, the classification procedure used is a variant on the Multisurface Method (MSM)10' known as MSM-Tree 23. This method uses a linear programming model to iteratively place a series of separating planes in the feature space of the examples. In this classifier, the predicted correctness was 97%.For the linear classification, they use 2 factors, sensitivity, specificity to identify. They have applied this machine learning techniques to Wisconsin Hospital to determine and diagnose breast cancer.

1. Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1995). Image analysis and machine learning applied to breast cancer diagnosis and prognosis. *Analytical and Quantitative cytology and histology*, *17*(2), 77-87.
2. Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1994). Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. *Cancer letters*, *77*(2-3), 163-171.
3. Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1997). Computerized diagnosis of breast fine‐needle aspirates. *The Breast Journal*, *3*(2), 77-80.

The main aim of this research paper is to control the precision of computer-based image analysis in identifying breast fine-needle aspirates (FNA). Depending on the separation accuracy achieved, the procedure is recursively repeated on the two regions generated by each plane until satisfactory parting is realized, that is, each of the final regions contains mostly points of one group.

Using machine learning approaches, an algorithm was developed on the basis of an initial series of 569 patients (357 benign and 212 malignant) with established diagnoses. Tenfold cross-validation of these data led us to believe that 97.5% accuracy could be achieved.

In this article, the author reported the potential computer diagnosis of 192 successive FNAs. The mathematical programming on which the data analysis is based has been described in detail elsewhere. Briefly, image dispensation produces a file consisting of one 30-dimensional point for each sample. We then wish to construct a surface that splits the benign points from the malignant points in this 30-dimensional space.

Correct visual diagnoses can effortlessly be made for FNA samples at the excesses of malignancy and benignity. Diagnostic difficulties arise for equivocal samples. One of the attributes of Xcyt is the rendering of a value that estimations the likelihood of malignancy.

All computer misclassified FNAs were correctly diagnosed visually based on contextual features. The author proposed a computer-based system as a diagnostic assistant rather than as a stand-alone system

1. Wolberg, W. H., Street, W. N., Heisey, D. M., & Mangasarian, O. L. (1995). Computer-derived nuclear features distinguish malignant from benign breast cytology. *Human Pathology*, *26*(7), 792-796.

This research paper describes the use of computer-based diagnostic methods to describe nuclear size, shape, and texture structures. These features are then used to differentiate amongst benign and malignant breast cytology. The difference in accuracy is largely caused by the subjectivity that is inherent in visual interpretation. Visually assessed size, shape, and consistency features that differentiate benign from malignant cells are now measured by computers. Nuclear feature analysis is better achieved on cytological FNA measures than on the more commonly used histological tissue samples. FNA cells are preserved intact, whereas histological processing cuts cells at various planes. A stepwise logistic regression selection process selected a model consisting of the variables standard error of the radius, worst radius, worst texture, and worst concave point. Overall , Classification built on digital feature analysis is vigorous; similar classification accuracy was obtained by logistic regression and inductive machine learning classification. Generally, simpler classifiers achieve better on new data than do more complex ones. So, it has been minimized by not only the number of extrication planes but also the amount of structures used in building the planes. A categorising procedure is formed from the nine remaining parts, and the correctness of the classifier is tested on the 10th part. The 10th part is then repaid, and the process is recurring until all parts have been verified. The capability to properly categorise samples as benign or malignant on the basis of these features was determined by inductive machine learning and logistic regression. Cross-validation was used to test the validity of the predicted diagnosis.

1. Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M., & Razavi, A. R. (2013). Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform, 4(124), 3.

In this research paper, the author worked with a combination of three machine learning methodologies that will be employed for prediction of breast cancer repetition. With the application of data mining methods, authors established models to predict the repetition of breast cancer by analyzing data collected from ICBC (Iranian Center for Breast Cancer) registry. In the research paper author demonstrated that data mining could be used as an important tool in identifying patterns in breast cancer cases, which can be used for analysis, prediction, and treatment purposes.

The EM (Expectation Maximization) algorithm is a technique for efficient estimation from incomplete data. For any unfinished dataset, there is subsidiary indication about the probable values of the unnoticed values. This evidence, when joint with some conventions, includes a predictive probability distribution for the missing values that should be averaged in the statistical investigation. To the most important finding, these results shown that SVM are the best classifier forecaster with the test dataset, tracked by ANN and DT thereafter. Moreover, studies should be directed for improvement in the presentation of these classification techniques by using more variables and selecting for a longer follow-up length. At the same time, there are some limitations in the current study. There were many cases misplaced in the follow-up and there were records with missing values that were omitted inappropriately. Some important variables such as S-phase fraction and DNA index were not encompassed in the study because of their inaccessibility which may have diminished the presentation of the models and also and there were some degree of missingness in our data. The SVM classification model predicts breast cancer recurrence with least error rate and highest precision. The predicted accuracy of the DT model is the lowest of all. The results are attained using 10-fold cross-validation for computing the impartial prediction exactness of each model.

1. Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, 1064-1069.

In this article, the aim is to evaluate efficiency and effectiveness of those algorithms in terms of accuracy, sensitivity, specificity and precision. In order to evaluate classifiers, they created 4 criteria for the performance, such as ‘Time to build a model’, ‘Correctly classified instances’, ‘Incorrectly classified instances’, ‘Accuracy(%)’ . For the calculating simulation errors, they made 5 criteria ‘KS’, ‘MAE’, ‘RMSE’, ‘RAE(%)’, ‘RRSE(%)’.

They used 4 classification algorithm which are ***C4.5, SVM, NB, and k-NN***. According to the report “Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability,” written by V. Chaurasia and S. Pal, the experimental result shows that SVM-RBF kernel is more accurate than other classifiers. SVM proves to be the most accurate classifier with accuracy of 96.99%.

Decision tree classifier scored 69.23% which was conducted by V. Chaurasia and S. Pal.

Aside from the previous two experiments, their experimental results show that *SVM* achieves the highest accuracy (97.13%) with *the lowest error rate (0.02%)* unlike C4.5, Naïve Bayes and k-NN that have an accuracy that varies between 95.12 % and 95.28 % and an error rate that varies between 0.03 and 0.06.

In summary, **SVM** was able to show its power in terms of effectiveness and efficiency based on accuracy and recall. Compared to a good amount of research on Breast-cancer-Wisconsin found in literature that compare classification accuracies of data mining algorithms, our experimental results make the highest value of accuracy (97.28 %) in classifying breast cancer dataset.

1. Gayathri, B. M., Sumathi, C. P., & Santhanam, T. (2013). Breast cancer diagnosis using machine learning algorithms-a survey. International Journal of Distributed and Parallel Systems, 4(3), 105.

At the beginning of this article, they provided relevant references and reports that tried to obtain the highest accuracy using various algorithms and methods, such as Back propagation algorithm used for training Multilayer Perceptron(MLP), Genetic algorithm and feature selection, Radial basis function(RBF), Probabilistic Neural Network(PNN), and so on.

Tuba kiyan(2004) : 96.18% for Radial Basis Function (RBF), 97% PNN, 98.8% for GRNN and 95.74% for MLP.

David B.fogel : used back propagation method with multilayer perceptron. The first experiment

consisted of five trials with 9-2-1 Multi Layer Perceptron. The second experiment consisted of 9-9-1 Multi Layer Perceptron. The accuracy recorded 97.5% and 98.2% respectively.

- SVM

To increase SVM accuracy, other authors tried several things, such as image processing.

1. Y.Iraneus Anna Rejani and Dr.S.Thamarai selvi

They used mammograms to detect breast cancer. Filtered image was used for contrast stretching. This method can be summarized as the initial step based on gray level information of image enhancement. For each tumor region extract, morphological features were extracted to categorize the breast tumor and finally SVM classifiers were used for classification.

2. Z.Qinli : modified kernel functions to improve the performance of SVM

- RVM

RVM is applied for detecting optical cancer, ovarian cancer, etc.

The author found the fact that RVM is applied for detecting optical cancer, ovarian cancer etc by going through various articles. They finally chose RVM for the better accuracy to detect breast cancer.

1. Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic.

In this research paper the authors made the use of expert systems and machine learning techniques in medical diagnosis. Using automatic diagnostic systems, the likely errors medical experts made in the process of diagnosis can be removed, and the medical data can be inspected quickly and in more fine tune level. The planned method can be explained into a 2-stage method. In the stage one, the learning algorithm RF was trained and tested on the training set and validation set in order to choose the best feature. A backward elimination approach was used to check the inner involvement of each feature to the classifier through one-by-one eliminating feature and associating classification precision earlier and later removing the feature. In order to significantly improve predictive accuracy for breast cancer diagnosis and prognosis the author worked on developing a classifier model by uniting the random forest classifier besides feature selection method. The main reason of the validation is to determine whether the removed feature is actually noisy feature or not. Validation can be achieved by comparing the classification precision before and after disregarding the feature. Also, one of the most important key attributes is that there are four variable importance measures have been executed using the RF software code. Two measures are calculated that is based on the GINI index of node impurity and classification accuracy of OOB data, are usually used. At the end it can be concluded that the proposed system can be very supportive to the physicians for them as a second view for their final verdict. By using such an efficient tool, they can make very accurate decisions and helpful in making quality measures that is beneficial for the society. The calculations done at the end of the comparative study made is a verdict that the algorithms that has been used are superior than that of previously implemented algorithms as they showed more accuracy than before. The results indicate that the proposed method is a reliable diagnostic tool for breast cancer diagnosis and prognosis.

1. Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016). Machine learning models in breast cancer survival prediction. Technology and Health Care, 24(1), 31-42.

In this , the author worked on the creation of a system that will detect the more reliable in predicting the breast cancer survival prediction. The model was based on the combination of various rules and different machine learning techniques. The main objective of the paper is to study rule-based classification method with an additional machine learning method that can be equipped for the prediction of various Breast cancer survival. This was based upon the exploitation of various patterns and relationships among the large datasets in order to forecast the outcome of disease with the help of the historical cases that are captured in the datasets. Ensemble learning model was used to improve the classification. For the k-fold cross validation technique the data was partitioned into k -subsets and with each iteration k-1 was used for training. In this process, it was repeated for k times for training and testing. The author made use of WEKA software. The algorithms were evaluated on the basis of sensitivity, accuracy, specificity, ROC curve and precision. The set of the models were implemented, and a comparative summary was generated. One of the most important features that was that higher accuracy which doubled from the last models. The outcome of assessment shows that the presentation of the TRF model was the best. Consequently, TRF model can be anticipated for breast cancer survival. This method is fast to in application and can be used for feature collection (finding efficient risk factors) alone. In their projected methods, three approaches specifically Bagging, Boosting, and Random subspace were employed to detect heart valve disorders. The database used in this method included 91 risk factors. A radial basis function (RBF) network is a kind of non-natural neural network that the activation functions are the combination of radial based models used for the breast cancer prediction . A linear grouping of radial basis functions of the inputs and neuron parameters are employed as the output of the network.