*Intelligent Systems and Computer Technology*
*D.J. Hemanth et al. (Eds.)*


*doi:10.3233/APC200173*

# Evaluation of Ensemble Machines in Breast Cancer Prediction

LeenaNesamani S[a,1] , NirmalaSugirthaRajiniS[b]
[a]*Research Scholar, Dept of Computer Application, Dr. M.G.R. Educational and Research Institute, India*
[b]*Professor, Dept of Computer Application, Dr. M.G.R. Educational and Research Institute, India*

**Abstract.**Breast cancer is one of the most deadly diseases encountered among women for which the cause is not clearly defined yet. Early diagnosis may help the physicians in the treatment of this deadly disease which could turn out fatal otherwise. Machine Learning techniques are employed in the process of detecting breast cancer with greater accuracy. Individual classifiers employed in this process, predicted the disease with less accuracy when compared with ensemble models. Ensemble methods employ a group of classifiers to individually classify the data. It then combines the result of the individual classifiers using weighted voting of their predictions. Ensemble machines perform better than individual models and show improved levels in the accuracy of the prediction system. This paper examines and evaluates different ensemble machines that are used in the prediction of breast cancer and tries to identify the combinations that prove to be better than the existing ones.

**Keywords.**Breast Cancer, benign, malignant, ensemble, classifier, Machine Learning.

## 1. Introduction

An ensemble classifier is a collection of a group of machine learning classifiers whose individual results are combined to get the final result of the classification process. The prime discovery in the research of constructing good ensembles is that the ensembles often show up with greater accuracy than the individual classifiers that are used in constructing them [1].

Algorithms like Support Vector Machine, Decision tree, Naïve Bayes, K-Nearest Neighbor, Perceptron, Logistic Regression etc were used to predict breast cancer in the past. These are individual classifiers whose performances are limited when compared with ensemble machines. In this paper, various ensemble models are examined analyzed to highlight their effectiveness over the prediction of breast cancer.

[1]LeenaNesamani S, Deptof Computer Application, Dr. M.G.R. Educational and Research, Chennai. Email id: leena.nesa@gmail.com

## 2. Performance of Ensemble Models

The model proposed by Maduri et al.[2], uses four machine learning approaches that make use of the Wisconsin Breast Cancer dataset. A total of twenty attributes were extracted by the PCA for the analysis.

The model used machine learning techniques such as Support Vector Machine, Logistic Regression, Decision Tree and K –Nearest Neighbor to classify the data individually. Soft voting technique was used to combine the results of the individual classifiers. Sequential Least Squares Programming Method (SLSQP) was used to assign the weights to each of the classifiers. Soft voting is a method that is used to combine similar machine learning techniques using the majority voting system. This system tries to predict the class labels for each of the input pattern based on the weight that is assigned to the classifiers.

**Table 1.** Performance of ML Techniques

| Classification | Ensemble | SVM | K-NN | DT | LR |
|---|---|---|---|---|---|
| Accuracy | **97.88** | 93.98 | 90.12 | 92.15 | 89.12 |
| F-score | **95** | 93 | 91 | 92 | 90 |
| $R^2$ | **0.9** | 0.7 | 0.7 | 0.8 | 0.6 |
| 10-Fold | **97.2** | 94.71 | 89.70 | 94.2 | 90.8 |

Accuracy, R square, F score, and 10 fold cross-validation were used to evaluate the performance of the system. The evaluation exhibited an increased performance for the ensemble system when measured against the individual classifiers. Also, the accuracy of prediction outperformed the equal weights method, and a weight of value 1 is applied to all the classifiers that make up the system.

The ensemble system designed by Quang et.al.[3], analyses both supervised and unsupervised models in the process of classifying breast cancer emloying the Wisconsin Breast Cancer dataset. The data is split in the ratio of 70:30 for training and testing. The data is prepared for processing by means of several pre-processing steps which include missing value check, checking of class imbalance, normalization checking, correlation checking and splitting up of training/testing set.

Classifiers like K- Nearest Neighbor, Support Vector Machines, Perceptron, AdaBoost, XGBoost[6], Gradient Descent, Extremely Randomised Trees (ERT), Logistic Regression were employed to perform the prediction of breast cancer. The results of the individual classifiers were combined through Ensemble Voting Classification. All the classifiers are assigned equal weights to make sure that all the classifiers have an equal preference to participate in the voting process. The training and the testing data are divided into 10 folds for validation.. Cross-Validation is

performed on the training and testing data and the average of the out-of-sample errors is obtained.

$$CV(\lambda) = \frac{1}{K}\sum_{K}^{k=1} E_k(\lambda)$$

The class label y can be predicted using the majority voting for each of the classifier Cj:

$$\hat{y} = mode(C_1(x), C_2(x), ..., C_m(x))$$

All these four models showed an accuracy above 98%. Out of these top four classifiers namely the Logistic Regression, Ensemble-Voting Classifier, and SVM Tuning showed higher values in accuracy, ROC-AUC, and F1 score.

The ensemble approach designed by Pragya et.al,[4] is hinged on Genetic Algorithm (GA). The motivation here was to assign the weights automatically. Eight classifiers were used for the initial classification process which include Decision Tree, Neural Network, Linear Model, AdaBoost, Naïve bays, Random Forest, SVM and SVM-Poly. Top three classifiers were identified based on their accuracy score. AdaBoost, SVM and Random Forest are the three classifiers that showed high accuracy levels. The predicted values of these three classifiers were used to train the weighted average ensemble model. The weighted average ensemble method built on GA showed higher accuracy score when compared to classical weighted average method. The weights were calculated manually in the classical method and the weights were optimized using the GA algorithm in the proposed method.

**Table 2.** Performance of three Nature Inspired Algorithms

| Model | MER | MWL | ER | Sens | Spec | Pre | Recall | TPR | FPR | F | Youden | TP | FP | TN | FN | Acc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PSO | .01 | .008 | .02 | .98 | .97 | .93 | .98 | .98 | .02 | .95 | .96 | 82 | 6 | 261 | 1 | 98 |
| DE | .01 | .01 | .02 | .97 | .97 | .93 | .97 | .97 | .02 | .95 | .95 | 81 | 6 | 261 | 2 | 97.71 |
| GA | .009 | .007 | .009 | .98 | .99 | .97 | .98 | .98 | .007 | .98 | .98 | 82 | 2 | 265 | 1 | 99.14 |

Out of the three algorithms which were examined, GA outperforms the other two. Performance measures like sensitivity, specificity, precision, recall, F1 score, Accuracy, and Youden were computed. An overall performance of 99.14% of accuracy was exhibited by the proposed system.

The ensemble model designed by Sheau-Ling et.al.[5], employed individual classifiers that include Neural fuzzy (NF) classifier, the Quadratic Classifier (QC), and K-nearest neighbor (KNN) classifier to create an ensemble model for the classifying breast cancer as either benign type or malignant type. The features that are necessary for the process of classification is retrieved using Information Gain (IG) algorithm. This technique uses the concept of Shannon entropy.

Information Gain is given by:

$$IG\left(\frac{x}{y}\right) = H(X) - H\left(\frac{X}{Y}\right)$$

Where, H(X) represents the entropy attribute of X and H(X/Y) represents the entropy attribute of the variable X after observing another variable Y. Three different classifiers were selected for the classification which includes Neural Fuzzy (NF) classifier, K Nearest Neighbor (KNN) classifier, and the Quadratic Classifier (QC). Three different ensemble models NFE, KNNE, QCE are created from the individual classifiers. And finally a fourth model which is an ensemble of the above three ensemble models was developed. The results of the ensemble models are combined through Majority Voting technique. It was observed that this model exhibited increased accuracy for the combined ensemble mode which consists of NF classifier , KNN classifier and QC classifier, at 97.14% which was also higher than the individual classifiers ensembles that were constructed for the study.

The model designed by Moloud et.al.[7], used a nested ensemble model for the detection of benign and malignant breast tumors. The nested ensemble makes use of the stacking and voting technique. Each nested ensemble is made up of a set of classifiers and meta-classifiers. The meta-classifiers are a combination of multiple classification algorithms. The model used the Wisconsin Diagnostic Breast Cancer (WDBC) dataset.

It follows a two-layer nested ensemble model, and four such models were created. Two of the ensembles have two meta-classifiers and the other two have three meta-classifiers. The first two models were named as SV-BayesNet2-MetaClassifier; SV-Na¨ıve Bayes-2-MetaClassifier; and the other two models were named as SVBayesNet-3-MetaClassifier and SV-Na¨ıveBayes-3-MetaClassifier. The result showed that the ensemble model that is made up of two nested layers performed better than the single classifiers and most of the work done before. The meta-classifiers obtained an accuracy of 98.07% for K = 10.

## 3. Result and Discussion

A new ensemble model that combines the above classifiers such as SVM, Ensemble - Voting Classifier, Logistics Regression, and SVM Tuning, GA based weighted average, NF, KNN and QC, SV-Na¨ıve Bayes-3-MetaClassifiers, Random Forest or by considering only the top three classifiers could be chosen to form a model, where the results of the individual classifiers could be combined through Majority Voting will be able to produce better results than all the other ensemble models considered for the evaluation.

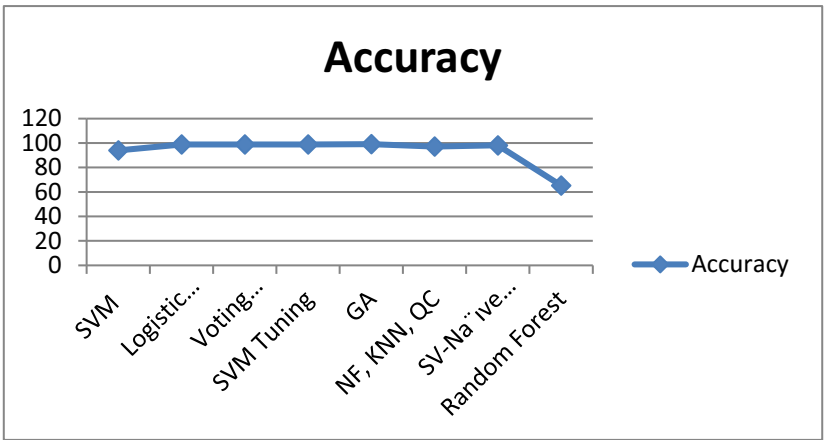


**Figure 1.** Performance of Ensemble Models

## 4. Conclusion

Several individual classifiers were employed in the prediction of breast cancer. And the top classifiers that showed higher accuracy in the prediction were selected to create an ensemble model where the results of the individual classifiers were combined to get the final prediction using either majority voting technique or simple average or bagging techniques. The final predictions of the ensemble models proved to be better than the predictions of the individual classifiers which conclude that the ensemble methods are better predictors in breast cancer and could be applied to any type of classification problem.

## References

[1] V.D.Ambeth Kumar, V.D.Ashok Kumar, S.Malathi and P.Jagaeedesh,Intruder Identification using Footprint Recognition with PCA and SVM Classifiers, International Journal of Advanced Materials Research, Vols.1345, PP 984-985, 2014.

[2] Quang H. Nguyen, TrangT.T.Do, Yijing Wang, Sin SweeHeng, Kelly Chen, Conceicao Edwin Philip, Misha Singh, Hung N. Pham, Binh P. Nguyen, Matthew C. H. Chua, Binh P. Nguyen , Matthew C. H. Chua, Breast Cancer Prediction using Feature Selection and Ensemble Voting, 2019 International Conference on System Science and Engineering (ICSSE)

[3] PragyaChauhan, Amit Swami, Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach, IEEE – 43488, 9th ICCCNT 2018, IISC

[4] Sheau-Ling Hsieh, Sung-Huai Hsieh, Po-Hsun Cheng, Chi-Huang Chen, Kai-Ping Hsu, I-Shun Lee, Zhenyu Wang, FeipeiLaiDesign, Ensemble Machine Learning Model for Breast Cancer Diagnosis, J Med Syst (2012) pp:2841–2847,#Springer Science+Business Media, LLC 2011

[5] Tsymbal, A., Pechenizkiy, M., and Cunningham, P., Diversity in search strategies for ensemble feature selection, Inform. Fusion pp: 83–98, 2005.

[6] A. Vignesh, T. YokeshSelvan, Ganesh Krishnan, Arjun N. Sasikumar, V. D. Ambeth Kumar, “Efficient Student Profession Prediction Using XGBoost Algorithm. Lecture Notes on Data Engineering and Communications Technologies, Volume 35, pp 140-148, 2020.