Bilge International Journal of Science and Technology Research

Web : http://dergipark.gov.tr/bilgesci - **E-mail:** kutbilgescience@gmail.com

Received: 11.11.2019
Accepted: 26.12.2019
DOI: 10.30516/bilgesci.645067

ISSN: 2651-401X
e-ISSN: 2651-4028
3 (Special Issue), 9-20, 2019

An Empirical Comparison of Machine Learning Algorithms for Predicting Breast Cancer

Hamit Taner Ünal[1], Fatih Başçiftçi[2]*

Abstract: According to recent statistics, breast cancer is one of the most prevalent cancers among women in the world. It represents the majority of new cancer cases and cancer-related deaths. Early diagnosis is very important, as it becomes fatal unless detected and treated in early stages. With the latest advances in artificial intelligence and machine learning (ML), there is a great potential to diagnose breast cancer by using structured data. In this paper, we conduct an empirical comparison of 10 popular machine learning models for the prediction of breast cancer. We used well known Wisconsin Breast Cancer Dataset (WBCD) to train the models and employed advanced accuracy metrics for comparison. Experimental results show that all models demonstrate superior accuracy, while Support Vector Machines (SVM) had slightly better performance than other methods. Logistic Regression, K-Nearest Neighbors and Neural Networks also proved to be strong classifiers for predicting breast cancer.

Keywords: Breast Cancer, Artificial Intelligence, Machine Learning, Medical Decision Support Systems

1. Introduction

Breast cancer is the second largest cause of cancer deaths among women (American Cancer Society, 2018). According to studies (Siegel & Jemal, 2015), there is an increase in the occurrence rate recently. Fortunately, breast cancer is also among the most curable cancer types if it is diagnosed in early stages (Akay, 2009). Early diagnosis followed by appropriate cancer treatment helps eliminate the deadly risk significantly. Furthermore, accurate classification of benign tumors can prevent patients undergoing unnecessary treatments.

Diagnosis of breast cancer and classification of patients into malignant or benign groups is the subject of recent research. Artificial Intelligence and Machine learning plays a vital role in a wide range of critical applications, such as image recognition, natural language processing, time series forecasting, regression and prediction. The use of an accurate machine learning algorithm for early detection could definitely save precious lives. Because of its unique advantages in critical features detection from complex breast cancer datasets, machine learning (ML) is widely recognized as the methodology of choice in breast cancer pattern classification and forecast modeling (Yue, Wang, Chen, Payne, & Liu, 2018). Previous studies had significant results for the classification of breast cancer by employing various Machine learning models (Agarap, 2018).

In this paper, we compare popular and trending machine learning methods effectively used in real-world classification problems. We utilized breast Cancer data available from the Wisconsin dataset from University of California at Irvine (UCI) machine learning repository with the aim of developing an accurate comparison.

The remainder of this paper is organized as follows. Firstly, a brief literature review has been conducted in Section 2. Then, basic information about ML models is documented in Section 3.

1 Selçuk University, Institute of Natural and Applied Sciences, Department of Information Technologies Engineering (PhD Student), Konya, Turkey
2 Selçuk University, Faculty of Technology, Department of Computer Engineering, Konya, Turkey
*Corresponding author (İletişim yazarı): basciftci@selcuk.edu.tr

Citation (Atıf): Ünal, H. T., Başçiftçi, F. (2019). An Empirical Comparison of Machine Learning Algorithms for Predicting Breast Cancer. Bilge International Journal of Science and Technology Research, 3(Special Issue): 9-20.

Next, the experimental setup and results with a thorough discussion is provided in Section 4 and Section 5 respectively. Finally, conclusions and future works are outlined in Section 6.

2. Literature Review

Numerous studies have been published and various classification techniques were developed for predicting breast cancer through Machine Learning. Most of the works evaluated their proposed models using the dataset taken from the UCI machine-learning repository.

Polat and Güneş (2007) conducted breast cancer diagnosis by using Least Square Support Vector Machines (LS-SVM). They evaluated the robustness of their model via k-fold cross validation with accuracy, sensitivity and specify metrics. They obtained classification accuracy of 98.53%. Akay (2009) employed an SVM-based method combined with feature selection. He obtained an accuracy of 99.52% with his proposed model containing five features. Sadhukan (2020) and Upadhyay compared KNN and SVM to predict breast cancer and analyzed digital image of a fine needle aspirate (FNA) of breast tissue with image processing to extract features of kernel of the cells. Sri et al (2019) proposed SVM and Neural Network models for tumor prognosis. By using WEKA tool, they applied 10-fold and 5-fold cross validation for high precision results. Kadam et al (2019) proposed feature ensemble learning based on Sparse Autoencoders and Softmax Regression for classification of Breast Cancer into benign or malignant. They obtained 98.60% true accuracy with their proposed model. Sethi (2018) compared evolutionary algorithms and machine learning predicting breast cancer in three different datasets. Jain et al (2018) presented a hybrid machine learning framework for the diagnosis of breast cancer and diabetes using feature selection and classification techniques. Their model identified significant risk factors related to both chronic disease datasets by applying different feature selection techniques and hybridization of Relief Feature Ranking with Principal Component Analysis (PCA) method. Rustam and Hartini (2019) proposed a new ML method based on kernel, which is a modification of KC-Means combining K-Means and Fuzzy C-Means algorithms together with kernel function. They applied C-Means algorithm on the centers of a fixed number of groups founded by K-Means, in order to improve the accuracy of classification. Rashed et al (2019) developed a novel network architecture with an inspiration from U-net structure to predict breast cancer in early stages. Omondiagbe (2019) investigated classification performance of Support Vector Machine (using radial basis kernel), Artificial Neural Networks and Naïve Bayes for breast cancer prediction, focusing on integrating machine learning techniques with feature selection and extraction methods. They proposed a hybrid approach by reducing the number of features with linear discriminant analysis (LDA).

3. Methodology

Literature on the subject of breast cancer prediction is mostly based on traditional methods. As a fairly new discipline, data science takes advantage of vast amounts of data at our disposal today and availability of advanced computational power. These factors make it possible the improvement of existing prediction methods and contribute to the development of new and better algorithms. In this section, we introduce several popular machine learning algorithms for prediction of breast cancer.

3.1. Dataset

The dataset used in this study is publicly available and was created by Dr. William H. Wolberg, physician at the University of Wisconsin Hospital at Madison, Wisconsin, USA. To create the dataset Dr. Wolberg used fluid samples, taken from patients with solid breast masses and a user-friendly graphical software called Xcyt, which is capable of perform the analysis of cytological features based on a digitized image of a fine needle aspiration (FNA) procedure of a breast mass.

The program describes the characteristics of the cell nuclei present in the image and uses a curve-fitting algorithm, to compute ten features from each one of the cells in the sample, than it calculates the mean value, extreme value and standard error of each feature for the image, returning a 30 real-valuated vector. Dataset consists of 569 instances of which 357 instances are benign and 212 are malignant cases. A Digitized Image of FNA is depicted in Figure 1.

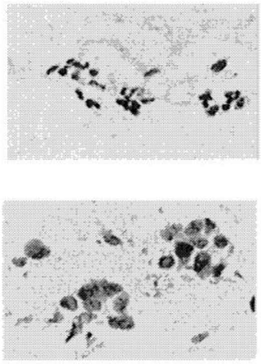


Figure 1. Digitized Image of FNA. Benign (top) and Malignant (bottom).

3.2. Feature selection and preprocessing

Feature Selection is the process of automatically or manually selecting features which contribute most to model prediction. Having irrelevant features in data can decrease the accuracy of the models and make the model learn based on irrelevant features. Feature selection is often confused with dimensionality reduction. The purpose of both methods is to simplify the data that feeds the algorithm. However, while the feature selection inputs the data to the model without altering it, the dimensionality reduction can process the data to obtain data of different structures and sizes.

An important feature of decision tree-based classifiers is that they can analyze the attributes in the dataset well and prioritize the data columns that yield the best results. Using this feature in the Extreme Gradient Boosting algorithm, we extracted the 10 most important features in the Wisconsin Breast Cancer Dataset (shown on Fig.2). This allowed us to simplify the model make better predictions in a reasonable amount of time.

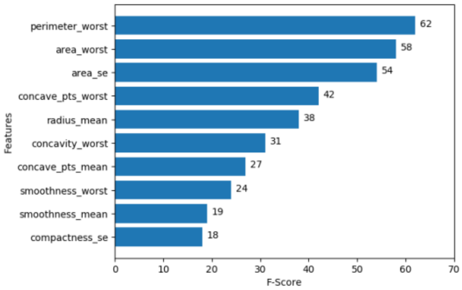


Figure 2. Feature Importance Table for Wisconsin Breast Cancer Dataset

3.3. ML Algorithms

Machine learning models have been employed in the last decade as serious competitors to classical statistical algorithms for prediction (Tokmak & Küçüksille). In machine learning, the ability of a model to predict categorical values based on a training dataset is called classification (ŞENOL & MUSAYEV). In this paper, we used Logistic Regression (Hosmer Jr, Lemeshow, & Sturdivant, 2013; Kleinbaum, Dietz, Gail, Klein, & Klein, 2002; Wright, 1995), K-Nearest Neighbors (Cover & Hart, 1967; Han, Kamber, & Pei, 2011; Ho, 1998), Support Vector Machines (Boser, Guyon, & Vapnik, 1992; Cortes & Vapnik, 1995; Vapnik, 1998), Classification and Regression Tress (Breiman, Friedman, Olshen, & Stone, 1984; Morgan & Sonquist, 1963; Timofeev, 2004), Naïve Bayes (Clark & Niblett, 1989; Frank, Hall, & Pfahringer, 2002; Zheng & Webb, 2000), Support Vector Machines (Boser et al., 1992; Cortes & Vapnik, 1995; Vapnik, 1998), ensemble techniques like Adaboost Classifier (Freund & Schapire, 1997; Li, Wang, & Sung, 2008), Gradient Boosting Classifier (Friedman, 2001, 2002), Random Forest (Breiman, 2001; Liaw & Wiener, 2002), Extreme Gradient Boosting (Chen & Guestrin, 2016; Chen, He, Benesty, Khotilovich, & Tang, 2015; Friedman, 2001) and finally Artificial Neural Networks with Multilayer Perceptron (Hecht-Nielsen, 1992; Mitchell, 1997; Rosenblatt, 1958; Tekin & Çan)

4. Experimental Setup

In order to evaluate the performance of given models, a series of experiments have been conducted. The models have been implemented with Python 3.7 programming language, using Keras framework and Tensorflow as a backend, running on Google Colaboratory Notebook. It is powered by single Tesla K80 GPU on free cloud service.

Initially, we load the dataset and make an exploratory data analysis to extract best features for determining major reasons of malignant tumors. Later, we modify data structure to prepare training on machine learning models. Finally, we apply selected algorithms and record results of their performance for prediction.

We used various advanced performance metrics to evaluate our model. The main score is based on the accuracy of the model, which basically shows the number of successful predictions. Precision (sensitivity) gives the percentage of the positive predictions that were correctly identified. Recall (specificity) which is another performance metric is the percentage of correctly predicted positive cases. F-measure metric is the harmonic average of the precision and recall. Calculation of those metrics is given in equation 1-4.

$$Accuracy = \frac{TN + FP}{FN + FP + TN + FP} \qquad (1)$$

$$F - Measure = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \qquad (2)$$

$$Precision = \frac{TP}{FP + TP} \qquad (3)$$

$$Recall = \frac{TP}{FN + TP} \qquad (4)$$

True Positive (TP): Benign samples classified as benign.
True Negative (TN): Malignant samples classified as malignant.
False Positive (FP): Benign samples classified as malignant.
False Negative (FN): Malignant samples classified as benign.

Confusion Matrix: The confusion matrix provides statistics about correct and incorrect predictions. It makes a comparison of expected values within the test set with the predicted values in the training set. The columns represent the predictions, while the rows indicate actual labels. The chart gives an idea about the performance of the classifier algorithm. Elements of a typical Confusion Matrix are given in Table 1.

Table 1. Elements of Confusion Matrix

Predicted Class	**Actual Class**		
		Positive (P)	Negative (N)
	True (T)	True Positive (TP)	True Negative (TN)
	False (F)	False Positive (FP)	False Negative (FN)

Cross-Validation (CV): Cross-validation is a technique used for making sure that our model is well trained, without using the test set. It consists in partition data into *k* portions of equal length. For each portion, we train the model on the remaining *k-1* parameters and evaluate it on partition *i*. The overall score is the mean of the *K* scores calculated.

There are two types of cross validation splits:
- Leave one out cross validation
- K-fold cross validation

Leave one out CV cycles over the dataset and removes one test group per iteration that will not be included in the training set but instead will be used to test the model's performance.

K-fold CV takes a K variable as an input, partition the dataset into K parts, cycles over the parts and for each cycle leaves the single portion out of training and use it as a test set.
By using k-fold validation we make sure that the model uses all the training data available for tuning the model, it can be computationally expensive but allows to train models even if little data is available. The main purpose of k-fold validation is to get an unbiased estimate of model generalization on new data.

Stratified k-Fold: In stratified k-fold, the aim is to include the same proportion of data labels for each test portion. For example, if the data has %75 Benign and %25 malignant samples, every fold contains the same percentage of benign and malignant samples.

Repeated k-fold: In repeated cross-validation, the cross-validation procedure is repeated n times, giving n random partitions of the original sample. Mean values of results for each n partition produce a single prediction.

5. Results and Discussion

The results were achieved using 10 fold cross-validation for each model, and are based on the average results obtained. Table 2 shows the complete set of results in a tabular format.

Cross-validation results were auspicious. The results indicate that Support Vector Machines gained top performance when compared to other nine models. Neural Networks, KNN and Logistic Regression Models have also demonstrated significant success when predicting breast cancer. Naïve Bayes and CART showed relatively poor performance but the prediction rates are at an acceptable level for our problem.

Table 2. Experiment Results

Algorithms	**Cross Validation Accuracy (Mean)**	**Std. (+/- %)**
Logistic Regression (LR)	0.9772	0.0207
K-Nearest Neighbors (KNN)	0.9701	0.0236
Decision Trees (CART)	0.9349	0.0197
Naïve Bayes (NB)	0.9332	0.0376
Support Vector Machines (SVM)	0.9789	0.0131
Adaboost Classifier (ADB)	0.9649	0.0157
Gradient Boosting Classifier (GRAD)	0.9612	0.0307
Random Forests (RF)	0.9613	0.0324
Extreme Gradient Boosting (XGB)	0.9630	0.0269
Neural Networks (MLP)	0.9772	0.0175

Comparison of ML Algorithms has been depicted in Figure 3. Detailed cross validation results for all algorithms have been tabulated at Table 3.

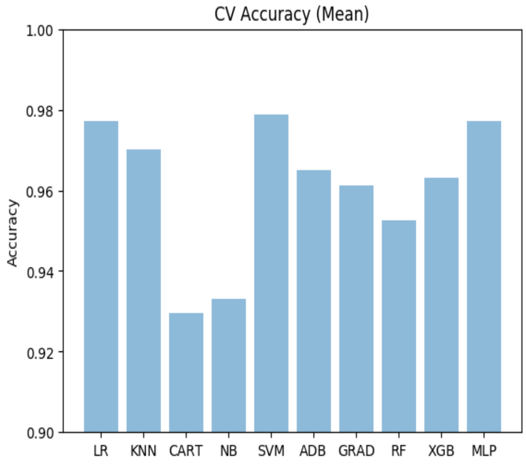


Figure 3. ML Comparison Chart

Table 3. Cross Validation Results for All Algorithms

LOGISTIC REGRESSION (LR)

Confusion Matrix			Precision	Recall	F1 Score	Acc.
35	1	B	0.9459	0.9722	0.9589	0.9483
2	20	M	0.9523	0.9090	0.9302	
36	0	B	1.0000	1.0000	1.0000	1.0000
0	22	M	1.0000	1.0000	1.0000	
36	0	B	0.9473	1.0000	0.9730	0.9649
2	19	M	1.0000	0.9048	0.9500	
36	0	B	1.0000	1.0000	1.0000	1.0000
0	22	M	1.0000	1.0000	1.0000	
35	1	B	0.9722	0.9722	0.9722	0.9649
1	10	M	0.9524	0.9524	0.9524	
36	0	B	1.0000	1.0000	1.0000	1.0000
0	22	M	1.0000	1.0000	1.0000	
34	2	B	0.9714	0.9444	0.9577	0.9474
1	20	M	0.9091	0.9524	0.9302	
34	1	B	1.0000	0.9714	0.9855	0.9821
0	21	M	0.9545	1.0000	0.9767	
35	0	B	1.0000	1.0000	1.0000	1.0000
0	21	M	1.0000	1.0000	1.0000	
35	0	B	0.9459	1.0000	0.9722	0.9643
2	19	M	1.0000	0.9048	0.9500	
Mean Accuracy :						**0.9772**
Std (+/-) % :						**0.0207**

K-NEAREST NEIGHBOR (KNN)

Confusion Matrix			Precision	Recall	F1 Score	Acc
35	1	B	0.9211	0.9722	0.9459	0.9310
3	19	M	0.9500	0.8636	0.9048	
36	0	B	0.9730	1.0000	0.9863	0.9828
1	21	M	1.0000	0.9545	0.9767	
35	1	B	0.9722	0.9722	0.9722	0.9649
1	20	M	0.9524	0.9524	0.9524	
36	0	B	0.9730	1.0000	0.9863	0.9825
1	20	M	1.0000	0.9524	0.9756	
36	0	B	0.9730	1.0000	0.9863	0.9825
1	20	M	1.0000	0.9524	0.9756	

36	0	B	1.0000	1.0000	1.0000	1.0000
0	21	M	1.0000	1.0000	1.0000	
36	0	B	0.9474	1.0000	0.9730	0.9649
2	19	M	1.0000	0.9048	0.9500	
35	0	B	1.0000	1.0000	1.0000	1.0000
0	21	M	1.0000	1.0000	1.0000	
35	0	B	0.9459	1.0000	0.9722	0.9643
2	19	M	1.0000	0.9048	0.9500	
34	1	B	0.9189	0.9714	0.9444	0.9286
3	18	M	0.9474	0.8571	0.9000	
Mean Accuracy :						**0.9701**
Std (+/-) % :						**0.0236**

DECISION TREES (CART)						
Confusion Matrix			Precision	Recall	F1 Score	Acc.
35	1	B	0.9459	0.9722	0.9589	0.9483
2	20	M	0.9524	0.9091	0.9302	
34	2	B	0.9444	0.9444	0.9444	0.9310
2	20	M	0.9091	0.9091	0.9091	
33	3	B	0.9429	0.9167	0.9296	0.9123
2	19	M	0.8636	0.9048	0.8837	
35	1	B	0.9459	0.9722	0.9589	0.9474
2	19	M	0.9500	0.9048	0.9268	
33	3	B	0.9706	0.9167	0.9429	0.9298
1	20	M	0.8696	0.9524	0.9091	
33	3	B	1.0000	0.9167	0.9565	0.9474
0	21	M	0.8750	1.0000	9333	
36	0	B	0.9474	1.0000	0.9730	0.9649
2	19	M	1.0000	0.9048	0.9500	
32	3	B	1.0000	0.9143	0.9552	0.9464
0	21	M	0.8750	1.0000	0.9333	
32	3	B	0.9697	0.9143	0.9412	0.9286
1	20	M	0.8696	0.9524	0.9091	
33	2	B	0.8919	0.9429	0.9167	0.8929
4	17	M	0.8947	0.8095	0.8500	
Mean Accuracy :						**0.9349**
Std (+/-) % :						**0.0197**

NAIVE BAYES CLASSIFIER (NB)						
Confusion Matrix			Precision	Recall	F1 Score	Acc
34	2	B	0.9444	0.9444	0.9444	0.9310
2	20	M	0.9091	0.9091	0.9091	
34	2	B	0.9444	0.9444	0.9444	0.9310
2	20	M	0.9091	0.9091	0.9091	
33	3	B	0.9429	0.9167	0.9296	0.9123
2	19	M	0.8636	0.9048	0.8837	
34	2	B	0.9714	0.9444	0.9577	0.9474
1	20	M	0.9091	0.9524	0.9302	
33	3	B	0.8684	0.9167	0.8919	0.8596
5	16	M	0.8421	0.7619	0.8000	
36	0	B	1.0000	1.0000	1.0000	1.0000
0	21	M	1.0000	1.0000	1.0000	
36	0	B	0.9474	1.0000	0.9730	0.9649
2	19	M	1.0000	0.9048	0.9500	
33	2	B	1.0000	0.9429	0.9706	0.9643
0	21	M	0.9130	1.0000	0.9545	
34	1	B	0.9189	0.9714	0.9444	0.9286
3	18	M	0.9474	0.8571	0.9000	
33	2	B	0.8919	0.9429	0.9167	0.8929
4	17	M	0.8947	0.8095	0.8500	
Mean Accuracy :						**0.9332**
Std (+/-) % :						**0.0376**

SUPPORT VECTOR MACHINES (SVM)						
Confusion Matrix			Precision	Recall	F1 Score	Acc.
35	1	B	0.9722	0.9722	0.9722	0.9655
1	21	M	0.9545	0.9545	0.9545	
36	0	B	1.0000	1.0000	1.0000	1.0000
0	22	M	1.0000	1.0000	1.0000	
35	1	B	0.9722	0.9722	0.9722	0.9649
1	20	M	0.9524	0.9524	0.9524	
36	0	B	0.9730	1.0000	0.9863	0.9825
1	20	M	1.0000	0.9524	0.9756	
35	1	B	0.9722	0.9722	0.9722	0.9649
1	20	M	0.9524	0.9524	0.9524	
36	0	B	1.0000	1.0000	1.0000	1.0000

0	21	M	1.0000	1.0000	1.0000	
36	0	B	0.9730	1.0000	0.9863	0.9825
1	20	M	1.0000	0.9524	0.9756	
34	1	B	1.0000	0.9714	0.9855	0.9821
0	21	M	0.9545	1.0000	0.9767	
35	0	B	0.9722	1.0000	0.9859	0.9821
1	20	M	1.0000	0.9524	0.9756	
35	0	B	0.9459	1.0000	0.9722	0.9643
2	19	M	1.0000	0.9048	0.9500	
Mean Accuracy :						**0.9789**
Std (+/-) % :						**0.0131**

ADABOOST CLASSIFIER (ADB)						
Confusion Matrix			Precision	Recall	F1 Score	Acc
35	1	B	0.9459	0.9722	0.9589	0.9483
2	20	M	0.9524	0.9091	0.9302	
35	1	B	0.9722	0.9722	0.9722	0.9655
1	21	M	0.9545	0.9545	0.9545	
35	1	B	0.9459	0.9722	0.9589	0.9474
2	19	M	0.9500	0.9048	0.9268	
36	0	B	0.9730	1.0000	0.9863	0.9825
1	20	M	1.0000	0.9524	0.9756	
35	1	B	0.9459	0.9722	0.9589	0.9474
2	19	M	0.9500	0.9048	0.9268	
35	1	B	1.0000	0.9722	0.9859	0.9825
0	21	M	0.9545	1.0000	0.9767	
35	1	B	1.0000	0.9722	0.9859	0.9825
0	21	M	0.9545	1.0000	0.9767	
34	1	B	1.0000	0.9714	0.9855	0.9821
0	21	M	0.9545	1.0000	0.9767	
35	0	B	0.9459	1.0000	0.9722	0.9643
2	19	M	1.0000	0.9048	0.9500	
34	1	B	0.9444	0.9714	0.9577	0.9464
2	19	M	0.9500	0.9048	0.9268	
Mean Accuracy :						**0.9649**
Std (+/-) % :						**0.0157**

GRADIENT BOOSTING CLASSIFIER (GRAD)						
Confusion Matrix			Precision	Recall	F1 Score	Acc.
35	1	B	0.9722	0.9722	0.9722	0.9655
1	21	M	0.9545	0.9545	0.9545	
35	1	B	0.9722	0.9722	0.9722	0.9655
1	21	M	0.9545	0.9545	0.9545	
36	0	B	0.9730	1.0000	0.9863	0.9825
1	20	M	1.0000	0.9524	0.9756	
35	1	B	0.9722	0.9722	0.9722	0.9649
1	20	M	0.9524	0.9524	0.9524	
35	1	B	0.9459	0.9722	0.9589	0.9474
2	19	M	0.9500	0.9048	0.9268	
36	0	B	1.0000	1.0000	1.0000	1.0000
0	21	M	1.0000	1.0000	1.0000	
35	1	B	0.9722	0.9722	0.9722	0.9649
1	20	M	0.9524	0.9524	0.9524	
35	0	B	1.0000	1.0000	1.0000	1.0000
0	21	M	1.0000	1.0000	1.0000	
35	0	B	0.8974	1.0000	0.9459	0.9286
4	17	M	1.0000	0.8095	0.8947	
33	2	B	0.8919	0.9429	0.9167	0.8929
4	17	M	0.8947	0.8095	0.8500	
Mean Accuracy :						**0.9612**
Std (+/-) % :						**0.0307**

RANDOM FOREST (RF)						
Confusion Matrix			Precision	Recall	F1 Score	Acc
35	1	B	0.9211	0.9722	0.9459	0.9310
3	19	M	0.9500	0.8636	0.9048	
36	0	B	0.9474	1.0000	0.9730	0.9655
2	20	M	1.0000	0.9091	0.9524	
36	0	B	0.9730	1.0000	0.9863	0.9825
1	20	M	1.0000	0.9524	0.9756	
36	0	B	1.0000	1.0000	1.0000	1.0000
0	21	M	1.0000	1.0000	1.0000	
35	1	B	0.9211	0.9722	0.9459	0.9298
3	18	M	0.9474	0.8571	0.9000	
36	0	B	1.0000	1.0000	1.0000	1.0000
0	21	M	1.0000	1.0000	1.0000	
36	0	B	0.9730	1.0000	0.9863	0.9825
1	20	M	1.0000	0.9524	0.9756	

33	2	B	1.0000	0.9429	0.9706	0.9643
0	21	M	0.9130	1.0000	0.9545	
35	0	B	0.9459	1.0000	0.9722	0.9643
2	19	M	1.0000	0.9048	0.9500	
33	2	B	0.8919	0.9429	0.9167	0.8929
4	17	M	0.8947	0.8095	0.8500	
Mean Accuracy :						**0.9613**
Std (+/-) % :						**0.0324**

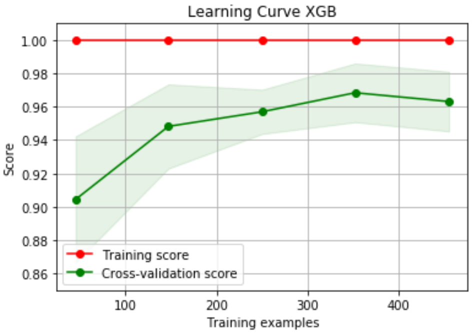
EXTREME GRADIENT BOOSTING (XGB)						
Confusion Matrix			Precision	Recall	F1 Score	Acc.
35	1	B	0.9722	0.9722	0.9722	0.9655
1	21	M	0.9545	0.9545	0.9545	
34	2	B	0.9714	0.9444	0.9577	0.9483
1	21	M	0.9130	0.9545	0.9333	
36	0	B	0.9730	1.0000	0.9863	0.9825
1	20	M	1.0000	0.9524	0.9756	
36	0	B	0.9730	1.0000	0.9863	0.9825
1	20	M	1.0000	0.9524	0.9756	
35	1	B	0.9459	0.9722	0.9589	0.9474
2	19	M	0.9500	0.9048	0.9268	
36	0	B	1.0000	1.0000	1.0000	1.0000
0	21	M	1.0000	1.0000	1.0000	
36	0	B	0.9730	1.0000	0.9863	0.9825
1	20	M	1.0000	0.9524	0.9756	
34	1	B	1.0000	0.9714	0.9855	0.9821
0	21	M	0.9545	1.0000	0.9767	
35	0	B	0.8974	1.0000	0.9459	0.9286
4	17	M	1.0000	0.8095	0.8947	
34	1	B	0.8947	0.9714	0.9315	0.9107
4	17	M	0.9444	0.8095	0.8718	
Mean Accuracy :						**0.9630**
Std (+/-) % :						**0.0269**

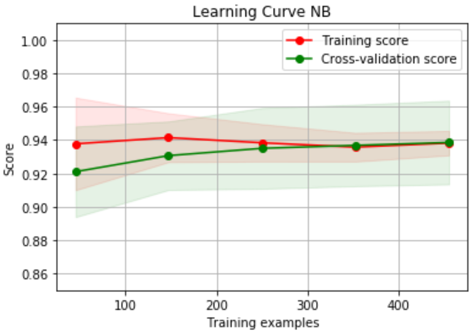
NEURAL NETWORK (MLP)						
Confusion Matrix			Precision	Recall	F1 Score	Acc
35	1	B	0.9459	0.9722	0.9589	0.9483
2	20	M	0.9524	0.9091	0.9302	
36	0	B	1.0000	1.0000	1.0000	1.0000
0	22	M	1.0000	1.0000	1.0000	
36	0	B	0.9474	1.0000	0.9730	0.9649
2	19	M	1.0000	0.9048	0.9500	
36	0	B	1.0000	1.0000	1.0000	1.0000
0	21	M	1.0000	1.0000	1.0000	
35	1	B	0.9722	0.9722	0.9722	0.9649
1	20	M	0.9524	0.9524	0.9524	
35	1	B	1.0000	0.9722	0.9859	0.9825
0	21	M	0.9545	1.0000	0.9767	
35	1	B	0.9722	0.9722	0.9722	0.9649
1	20	M	0.9524	0.9524	0.9524	
34	1	B	1.0000	0.9714	0.9855	0.9821
0	21	M	0.9545	1.0000	0.9767	
35	0	B	1.0000	1.0000	1.0000	1.0000
0	21	M	1.0000	1.0000	1.0000	
35	0	B	0.9459	1.0000	0.9722	0.9643
2	19	M	1.0000	0.9048	0.9500	
Mean Accuracy :						**0.9772**
Std (+/-) % :						**0.0175**

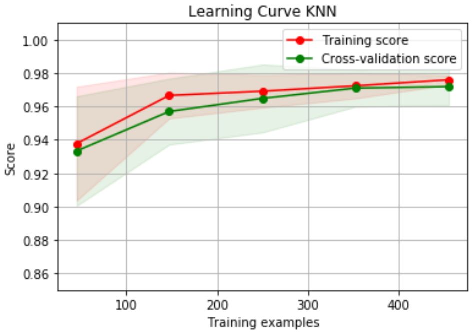
In order to evaluate overall performance of an algorithm, learning curves can be utilized to measure how many training sample is required to reach optimum performance. Learning curves depict the test and training score of a classifier for various lengths of training samples. It is a visual indicator of how the performance is increased by adding new training samples and whether the classifier is adversely affected from a variance error or a bias error. Learning curves for the algorithms are depicted on Figure 4.

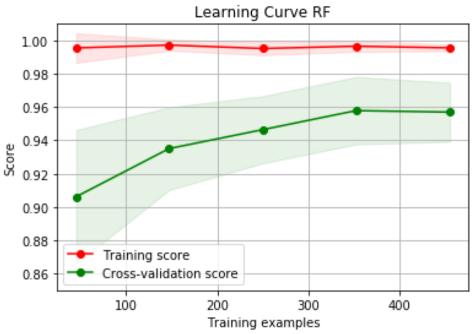
Gap between the two curves shown above, determines the interpretation of the models in the bias-variance landscape. A narrow distance shows low variance. On the contrary, a wide gap indicates that validation error will generally be higher. If we change the training set sizes, the pattern will likely to continue, and the value between training and validation errors will draw that distance between the two learning curves. If a model performs better on the training set and poor on the test set, overfitting will occur.

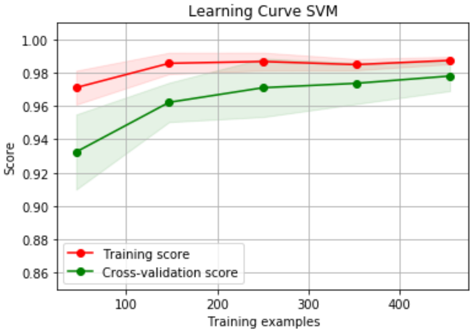
Learning curves of above models clearly demonstrate that KNN and Logistic Regression show significant performance with less training examples. Although, performing best among the other models, SVM needs more samples to reach best accuracy.

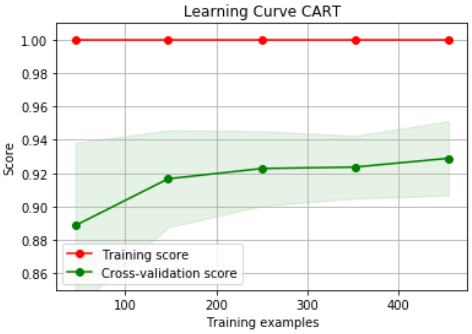


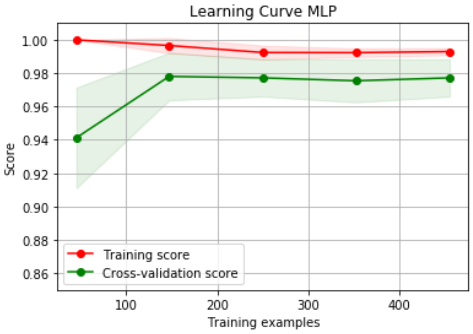












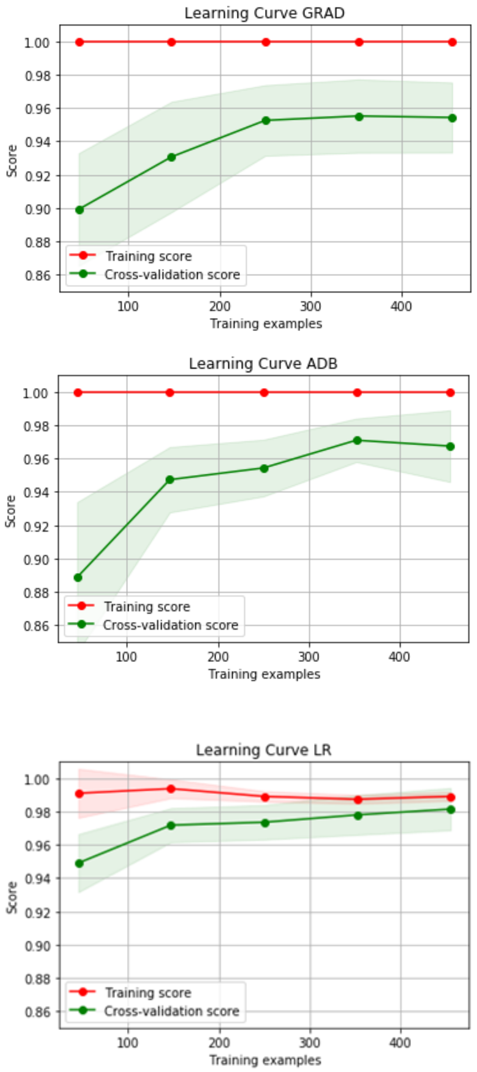


Figure 4. Learning Curves for ML Algorithms

6. Conclusions and Future Work

This paper presented a comparison between several classification techniques in machine learning (Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees (CART), Naïve Bayes, Support Vector Machines (SVM), Adaboost Classifier, Gradient Boosting Classifier, Random Forests, Extreme Gradient Boosting and Neural Networks), for the prediction of breast cancer. The selected techniques were applied to the well-known 'Wisconsin Breast Cancer dataset' as a part of UCI data repository. The results showed that all methods performed remarkable performance, while LR, SVM and MLP gained the best metrics.

We also have identified which features contribute to predict breast cancer by using tree based classifiers. Application of data visualization and data analytics techniques, together with the application of data science tools, allowed the understanding of feature's predictive relevance.

Our ongoing research efforts are geared toward applying advanced Machine Learning techniques to recent real world problems on various domains and increasing the prediction performance with fine-tuning of hyper parameters.

Code for the models can be found on Github repository: https://github.com/htanerunal/breast_cancer/blob/master/breast_cancer_ML_comparison.ipynb

Acknowledgment

We would like to express gratitude to our families for their patience and continuous support.

References

Agarap, A. F. M. (2018). *On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset.* Paper presented at the Proceedings of the 2nd International Conference on Machine Learning and Soft Computing.

Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications, 36*(2), 3240-3247.

American Cancer Society. (2018). "Cancer Facts & Figures 2018". *Atlanta, American Cancer Society.*

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). *A training algorithm for optimal margin classifiers.* Paper presented at the Proceedings of the fifth annual workshop on Computational learning theory.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Wadsworth Int. *Group, 37*(15), 237-251.

Chen, T., & Guestrin, C. (2016). *Xgboost: A scalable tree boosting system.* Paper presented at the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1-4.

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. *Machine learning, 3*(4), 261-283.

Cortes, C., & Vapnik, V. (1995). Soft margin classifiers. *Machine learning, 20*, 273-297.

Cover, T. M., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory, 13*(1), 21-27.

Frank, E., Hall, M., & Pfahringer, B. (2002). *Locally weighted naive bayes.* Paper presented at the Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences, 55*(1), 119-139.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis, 38*(4), 367-378.

Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems*, 83-124.

Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. In *Neural networks for perception* (pp. 65-93): Elsevier.

Ho, T. K. (1998). *Nearest neighbors in random subspaces.* Paper presented at the Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR).

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398): John Wiley & Sons.

Jain, D., & Singh, V. (2018). *Diagnosis of Breast Cancer and Diabetes using Hybrid Feature Selection Method.* Paper presented at the 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC).

Kadam, V. J., Jadhav, S. M., & Vijayakumar, K. (2019). Breast Cancer Diagnosis Using Feature Ensemble Learning Based on Stacked Sparse Autoencoders and Softmax Regression. *Journal of medical systems, 43*(8), 263.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression*: Springer.

Li, X., Wang, L., & Sung, E. (2008). AdaBoost with SVM-based component classifiers. *Engineering Applications of Artificial Intelligence, 21*(5), 785-795.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3), 18-22.

Mitchell, T. M. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill, 45*(37), 870-877.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association, 58*(302), 415-434.

Omondiagbe, D. A., Veeramani, S., & Sidhu, A. S. (2019). *Machine Learning Classification Techniques for Breast Cancer Diagnosis.* Paper presented at the IOP Conference Series: Materials Science and Engineering.

Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital signal processing, 17*(4), 694-701.

Rashed, E., & El Seoud, M. (2019). *Deep learning approach for breast cancer diagnosis.* Paper presented at the Proceedings of the 2019 8th International Conference on Software and Information Engineering.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review, 65*(6), 386.

Rustam, Z., & Hartini, S. (2019). *Classification of Breast Cancer using Fast Fuzzy Clustering based on Kernel.* Paper presented at the IOP Conference Series: Materials Science and Engineering.

Sadhukhan, S., Upadhyay, N., & Chakraborty, P. (2020). Breast Cancer Diagnosis Using Image Processing and Machine Learning. In *Emerging Technology in Modelling and Graphics* (pp. 113-127): Springer.

Sethi, A. (2018). *Analogizing of Evolutionary and Machine Learning Algorithms for Prognosis of Breast Cancer.* Paper presented at the 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO).

Siegel, R., & Jemal, A. (2015). Cancer facts & figures 2015. *American Cancer Society Cancer Facts & Figures*.

Sri, M. N., Sailaja, D., Priyanka, J. H., Chittineni, S., & RamaKrishnaMurthy, M. (2019). *Performance Evaluation of SVM and Neural Network Classification Methods for Diagnosis of Breast Cancer.* Paper presented at the International Conference on E-Business and Telecommunications.

ŞENOL, Ü., & MUSAYEV, Z. Estimating Wind Energy Potential by Artificial Neural Networks Method. *Bilge International Journal of Science and Technology Research, 1*(1), 23-31.

Tekin, S., & Çan, T. Yapay Sinir Ağları Yöntemi ile Ermenek Havzası'nın (Karaman) Kayma Türü Heyelan Duyarlılık Değerlendirmesi. *Bilge International Journal of Science and Technology Research, 3*(1), 21-28.

Timofeev, R. (2004). Classification and regression trees (CART) theory and applications. *Humboldt University, Berlin*.

Tokmak, M., & Küçüksille, E. U. Kötü Amaçlı Windows Çalıştırılabilir Dosyalarının Derin Öğrenme İle Tespiti. *Bilge International Journal of Science and Technology Research, 3*(1), 67-76.

Vapnik, V. (1998). Statistical Learning Theory Wiley-Interscience. *New York*.

Wright, R. E. (1995). Logistic regression.

Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. *Designs, 2*(2), 13.

Zheng, Z., & Webb, G. I. (2000). Lazy learning of Bayesian rules. *Machine learning, 41*(1), 53-84.