Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 192 (2021) 2742–2752

Procedia Computer Science

www.elsevier.com/locate/procedia

25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Machine Learning Classifiers on Breast Cancer Recurrences

Vincent Peter C. Magboo*, Ma. Sheila A. Magboo

*Department of Physical Sciences and Mathematics, College of Arts and Sciences, University of the Philippines Manila, Padre Faura St., Ermita, Manila, 1000, Philippines

Abstract

Breast cancer remains to be a leading cause of cancer-related deaths among women. Mortality is mainly attributed to metastasis and recurrence. Hence, early detection of breast cancer recurrence has become a real-world medical problem. Using data mining approaches, we compared four popular machine learning models (Logistic Regression, Naïve Bayes, K-Nearest Neighbors, and Support Vector Machines) on a high-dimensional but very small dataset, the Wisconsin Prognostic Breast Cancer Data Set for classifying breast cancer recurrences on four different configurations: a) only scaling applied, b) scaling with PCA, c) scaling with PCA and oversampling of minority class, and d) only with selected features (i.e. choose only one from each set of features that have high correlation with each other). Our results showed that Logistic Regression provided the best scores in almost all metrics (precision, recall, accuracy, f1 score (weighted), AUROC, AUPROC, and Cohen Kappa statistic in all four configurations, followed by Support Vector Machines, and then by K-Nearest Neighbors. Naive Bayes performed poorly especially in the scaling with PCA configuration, however, when we retained only one of many features that have high correlations with each other, Naïve Bayed performance improved. KNN improved its recall with oversampling while SVM's accuracy score has been fairly constant in all four configurations. Based on this study, the Logistic Regression model can serve as a potential model for predicting breast cancer recurrence that would enable clinicians to propose treatment options based on whether patient's features correspond to a good or bad prognosis (recurrence). This indicates the clinical utility of data mining methods for the early detection of breast cancer recurrence in post-surgical patients to save lives.

Keywords: breast cancer; data mining; machine learning, Naïve Vayes, Logistic Regression, K-Nearest Neighbors

* *Corresponding author email address: vcmagboo@up.edu.ph*

10.1016/j.procs.2021.09.044

1. Introduction

According to the World Health Organization, breast cancer is the most common malignancy among women affecting around 2.3 million women in 2020 and has produced around 685,000 deaths globally.[1] It affects women of any race with rising prevalence rates as the age group increases. It is also the leading cause of cancer-related deaths among women [2, 3]. Death from breast cancer is mainly associated to metastasis and recurrence (or relapse). Metastatic relapse can occur months to decades after initial diagnosis and treatment in breast cancer. [4, 5] Many researchers have considered early detection and prediction to be the best way to fight this highly invasive malignancy. Thus, predicting the recurrence of breast cancer has become a real-world medical problem and presents great challenges to the researchers using data mining approaches. [6, 7]. Accurate prediction of breast cancer behavior is very important as it aids clinicians in their decision-making process, enabling a more personalized treatment for patients leading to increased chances of recovery [8, 9]. Moreover, it is also associated with improved efficiency of healthcare resource allocation for these patients.[10] Understanding the underlying factors regarding early recurrence of breast cancer remains an important research priority not only among clinicians but to data scientists as well. Many research studies on cancer recurrence involve applying a variety of machine learning algorithms and statistical techniques which have improved breast cancer detection and prediction [7, 8, 11].

1.1. Related Works

Many studies reported in the literature have utilized several data mining techniques in breast cancer. In a comparative review study [12], several researches have been compiled and highlighted the use of machine learning, deep learning, and data mining techniques applied to predict breast cancer recurrence hoping to provide more appropriate information for beginning researchers in machine learning algorithms. Kumar et al. [9], compared 12 classification techniques on Wisconsin Breast Cancer data from UCI repository to predict malignant and benign breast cancer. In the study by Morales-Ortega et al. [11], several methods have been applied to detect breast cancer recurrence in post-surgical patients. The authors compared decision trees (DT), Naïve Bayes (NB) and Support Vector Machines (SVM), and then integrated the optimum results with Simple K-Means algorithm to generate significant improvements in precision. In [13], Aishwarja et al. compared K-Nearest Neighbors (KNN), SVM, NB and Random Forest in predicting breast cancer and its recurrence. The best performance was obtained by KNN in breast cancer prediction and SVM for breast cancer recurrence. Temesgen Abera Asfaw [14] analyzed the performance of DT, Logistic Regression (LR), NB and KNN for detecting breast cancer using the UCI Wisconsin breast cancer dataset. It showed LR with the best classification accuracy 96.93 %. In another study [15], authors applied LR in the detection of breast cancer. They have concluded that the use of a weighting factor β (which is a function of number of features and type of optimization techniques), to the existing logistic function provided significant improvement in accuracy, sensitivity and specificity. Lou et al. [10], compared the accuracy of forecasting models (Artificial Neural Networks (ANN), KNN, SVM, NB, and Cox Proportional Hazards Regression Models (COX)) to predict recurrence within 10 years after breast surgery. The study showed artificial neural networks with the highest prediction performance indices. In the study by Ak [16], data visualization and machine learning techniques including LR, KNN, SVM, NB, DT, random forest, and rotation forest were used to detect breast cancer. The highest classification accuracy (98.1 %) was obtained by LR and concluded further that machine learning techniques can impact cancer detection in the decision-making process. Borges [17], applied machine learning techniques in the Wisconsin Breast Cancer dataset to discriminate benign from malignant breast masses with Bayesian Networks obtaining the best accuracy of 97.80 %. Finally, Mohammed et al. [18], compared three machine learning models (DT, NB and Sequential Minimal Optimization (SMO)) on Wisconsin Breast Cancer dataset where it showed SMO having the highest classifier performance particularly after applying resample filter in the preprocessing phase.

2. Methodology

2.1. Dataset Description

In this study, the Wisconsin Prognostic Breast Cancer (WPBC) dataset is used. It is publicly available at the UCI Machine Learning Repository [19]. This dataset contains 198 instances with 34 attributes. The outcome variable is recurrence (151 nonrecur, 47 recur). The first 30 features in this dataset are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image, hence only tumor characteristics (such as area, perimeter, compactness, texture, concavity, concave points, size, lymph node involvement) are utilized. Generally, all these parameters are useful to classify cancer with those relatively large values pointing to a malignant condition. Ten real-valued features are computed for each cell nucleus. The mean, standard error, and "worst" or largest (mean of the three largest values) of these features are computed for each image, resulting in 30 features. Tumor size (diameter of the excised tumor in centimeters) and lymph node status (number of positive axillary lymph nodes observed at time of surgery) are the last two real-valued features.

2.2. Preprocessing and Exploratory Data Analysis

Preprocessing was performed to prepare the dataset for machine learning. The 4 records with missing values in the lymph nodes were removed, reducing the initial dataset from 198 (151 Nonrecur, 47 Recur) to 194 (148 Nonrecur, 46 Recur). The ‘id’ column was dropped since all identifiers should be removed prior to analysis. All columns were converted to numeric as this is required by the models to be used in this study. A pairplot of all features was created to see the distribution of values. From the pairplot, it is interesting to note in all features it is difficult to separate recurrence from non-recurrence as the values overlap. A sample pairplot containing only four features is shown in Fig. 1 below to illustrate this observation.

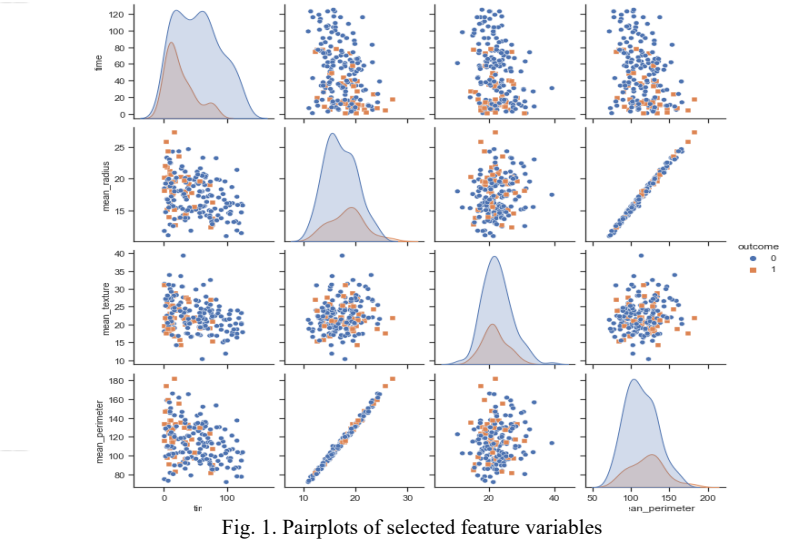

Fig. 1. Pairplots of selected feature variables

Feature scaling was applied in order to normalize the range of values of each feature. This will ensure that each feature contributes approximately proportionately to the final distance and also to comply with the requirements of algorithms to be used (ex. SVM, PCA, KNN).

2.3. The Model Configurations

We aim to predict whether cancer will recur or not, using Logistic Regression, Naïve Bayes, K-Nearest Neighbors, and Support Vector Machines. A total of 4 model configurations were built in order to see the effect of each

configuration on the model performance as seen in Fig. 2. The 4 model configurations are as follows: a) Only Scaling Applied, b) Scaling with Principal Component Analysis (PCA), c) Scaling with PCA and Oversampling of minority class, and d) Selected Features (i.e., choose only one from each set of features that have high correlation with each other).

PCA was applied in the 2nd and 3rd model configurations as a feature extraction technique in which a new set of fewer features but containing basically the same information as the original features is generated. PCA reduced the number of features while at the same time captured as much information as possible with high explained variance. In

Fig. 2. Model Configurations

this study we aim choose principal components that can cover a variance of 95%. This resulted in the selection of 13 principal components covering all 33 features. A scree plot of the PCA is shown in Fig. 3.

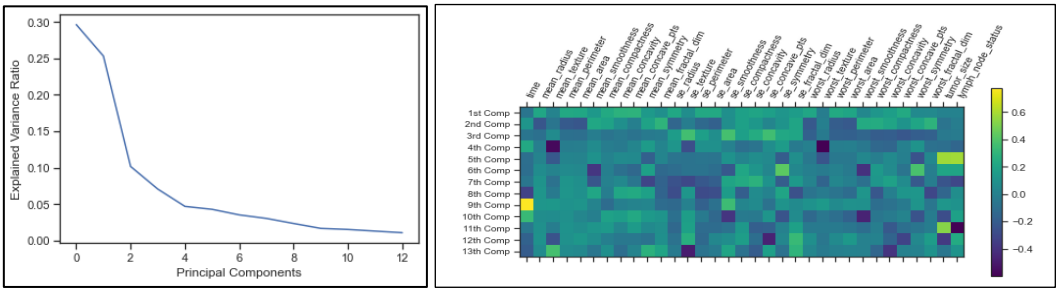

Fig. 3. Principal Component Analysis Result

The resulting principal components were used to create a new dataset consisting of 13 features and the same target variable 'outcome' (0: Nonrecur, 1: Recur) as shown in Fig. 4.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	outcome
0	-1.624414	-2.263547	0.442383	-1.755149	0.136050	0.358325	-1.362365	-0.301880	0.020172	0.499513	0.890479	0.407108	0.060793	0
1	6.219238	3.086070	-1.843870	3.524154	0.244660	0.050153	-0.896042	-1.228048	-1.116149	-0.326711	0.044955	0.407309	-0.622436	0
2	1.155083	-0.987948	-1.026530	3.095157	-1.076312	2.227237	0.565639	-0.489725	1.972983	0.344312	0.209328	0.071435	0.863631	0
3	2.996889	11.549788	-0.279260	1.576046	-1.420966	1.169724	-2.100026	-0.637150	0.948278	-1.103539	0.661119	0.383818	0.666323	0
4	1.369174	-1.856716	1.089515	2.643058	1.499046	-0.811371	1.918320	0.479474	-0.202146	-0.512728	0.377872	0.215685	0.345279	1

Fig. 4. New Dataset Generated after PCA

Random oversampling was used in combination PCA in the 3rd configuration in order to balance the distribution of the classes. As discussed above WPBC suffers from class imbalance as there is a 4:1 ratio between nonrecur vs recur. Random oversampling was used to duplicate examples in the minority class by randomly selecting examples from the minority class with replacement and adding them to the training dataset. 1/3 of the dataset was set aside for testing and remained unmodified. A total of 129 records (99 nonrecur, 30 recur) for training and 65 records (49 nonrecur, 16 recur) for testing were generated by the train-test split. All 4 configurations used this specification. For the 3rd configuration, 69 records were added by the random oversampling function to equalize the number of classes. Branco et al. reported that resampling appears to be an effective solution to an imbalanced data because it imposes non-uniform misclassification costs [17, 20].

For model configuration number of 4, a correlation matrix was constructed. Our interest is on eliminating the features that have high correlation with other features as having all of them in the model would hurt the performance of the model. We examined those features having at least 80% correlation with each other and selected only one feature as the group's representative. For example, mean_radius, mean_perimeter, mean_area, worst_area, worst_perimeter and worst_radius have correlation values more than 90%, and so we selected the mean_radius as the group's representative and removed the remaining features from the new feature list. The new feature list now contains 20 features: time, mean_radius, mean_texture, mean_smoothness, mean_concavity, mean_concave_pts, mean_symmetry, mean_fractal_dim, se_radius, se_texture, se_smoothness, se_concave_pts, se_symmetry, se_fractal_dim, worst_smoothness, worst_concavity, worst_concave_pts, worst_symmetry, tumor_size, and lymph_node_status.

As for the machine learning algorithms used, all 4 model configurations used the same machine learning algorithms: Logistic Regression, Naïve Bayes, K-Nearest Neighbors, and Support Vector Machines.

Logistic Regression is a classification algorithm that predicts a binary outcome based on a set of independent variables. It aims to maximize the posterior probabilities of classes. Logistic Regression uses the sigmoid function, a class of functions characterized by an S-shaped curve, to map probabilities to discrete classes. For example, if a

probability is over or under a certain threshold, it then falls into one or the other category. The sigmoid function generates outputs in the range of (0,1) [8, 16, 21].

Naïve Bayes is a supervised probabilistic classifier using the Conditional Probability (Bayes Theorem) based on the frequency and combination of the values on the dataset [8, 11, 14, 18, 21]. It can be used in binary as well as multiclass prediction problems. It is relatively fast compared to other machine learning algorithms but it assumes that each input variable is independent of the other variables which in real life is impossible as most of the predictors used have some form of correlation with each other. Another problem with Naïve Bayes happens when a categorical variable has a category that is not observed in the training dataset, the model will assign a zero probability and hence will not be able to make a prediction. To resolve this problem, a simple smoothing technique called Laplace estimation can be used [22, 23]. Naïve Bayes is typically applied to various supervised classification problems such as text classification, spam detection, sentiment analysis, among others [11, 16].

K-Nearest Neighbors (KNN) is a non-parametric supervised classification algorithm in which the k nearest neighbors of a point are determined by minimizing a similarity measure such as Euclidean distance. The unlabeled object is then classified either by majority voting (the predominant class in the neighborhood) or by a weighted majority, where a greater weight is given to points closer to the unlabeled object [14, 16, 21, 24]. The main drawback of KNN is when dealing with large databases as the algorithm needs to search for each instance's nearest neighbors through the entire dataset. Another problem is how to determine the optimal number of neighbors (k) and the most appropriate distance metric to use [8].

Support Vector Machines (SVM) is a supervised machine learning algorithm that can be used for both classification or regression problems. Each record is plotted as a point on an n-dimensional space where n is the number of features in the dataset. The goal is to look for the best hyperplane that can separate the classes well. This hyperplane should be able to maximize the distance between the point nearest it and the hyperplane. It is easy to visualize if we only have two features that can be mapped on the x-y plane. However most datasets have higher dimensions and as such, SVM maps a lower dimensional space to a higher dimensional space through a technique called the kernel trick. SVM likewise can handle multiple continuous and categorical variables [25, 26, 27].

2.4. Model Evaluation

A variety of performance metrics were used to evaluate the classification performance of each of the 4 machine learning algorithms over each of the 4 different configurations or a total of 16 different models. As the dataset exhibits class imbalance since random oversampling was applied only in one configuration, the other three configurations were evaluated using metrics recommended for datasets with imbalance problems. The test set consists of 65 samples of which 49 belong to the Nonrecur class while 16 belongs to the Recur class.

For datasets with class imbalance, the recommended performance evaluation metrics are Precision, Recall, F1-score, Area Under the Precision-Recall Curve and Cohen Kappa score [28 - 32]. Precision is a ratio that gives the number of correct predictions for a certain class. It is a measure of correctness as it measures how many are really positive out of the positive labeled samples. Recall is a measure of completeness. Recall measures how many of the positive class are labelled correctly. Note that Accuracy is not a recommended metric for imbalanced datasets as it places more weight on the majority class. Thus, it performs poorly in classifying rare classes. The F1-score is the harmonic mean or weighted average of precision and recall, hence, is also a good metric for imbalanced datasets. The Cohen Kappa statistic is a metric that compares observed accuracy with expected accuracy (random chance). It estimates how well the model can separate the instances into the right class. Landis and Koch considers 0-0.20 as slight agreement, 0.21-0.40 as fair agreement, 0.41-0.60 as moderate agreement, 0.61-0.80 as substantial agreement, and 0.81-1 indicates an almost perfect classifier [33]. The Precision-Recall Curve is a plot of the precision (y-axis) and recall (x-axis) for different probability thresholds. It focuses on the minority class making it an effective diagnostic

tool for imbalanced datasets. The ROC Curve is a plot that summarizes the performance of the binary classifier model on the positive class [28 – 32]. It plots the False Positive Rate in the x-axis vs the True Positive Rate in the y-axis

Of the four machine learning algorithms used in this study, the Logistic Regression performed the best in almost all performance metrics in the 4 different model configurations, followed by Support Vector Machines and K-Nearest Neighbors in that order as shown in Table 1. The best values are highlighted in red. Naive Bayes performed poorly especially in Model Configuration 2 (Scaling with PCA). However, when we retained only one of many features that have high correlations with each other as in Model Configuration 4, Naïve Bayes performance improved. One probable reason could be due to the fact that many of the features have high correlation with each other. However, Naïve Bayes assumes that they are conditionally independent and this affected its performance as a classifier. In general, if the features exhibit high correlation, this will have a negative affect on the Naïve Bayes assumption and thus would result to a negative performance of the Naïve Bayes classifier [14, 18, 21]. PCA does not remove the dependencies between features but just transforms them into a lower dimensional space hence did not help improve the performance of Naïve Bayes [34, 35]. Thus, the removal of highly correlated features in Model Configuration 4 somehow helped improve Naïve Bayes performance.

KNN improved its recall with PCA and oversampling (Model Configuration 3) while SVM's performance has been fairly constant in all four configurations. KNN classifies a datapoint based on the majority vote among its K-Nearest Neighbors [8, 14, 16]. If there is a huge class imbalance, more datapoints of the majority class will be present. Hence, the higher chance that the datapoint may be classified as belonging to the majority class as in the case for KNN in Model Configuration 1 and 4. Oversampling the minority class to balance the distribution helped improve the KNN classification [8, 21, 24].

SVM is effective for balanced dataset. It generally does not perform well on imbalanced datasets. SVM finds the hyperplane that can serve as decision boundary to split the classes. The margin should be maximized to separate the binary classes with adjustable bias-variance proportion. SVM tends to favor the majority class on imbalanced datasets although modifications can be made to incorporate costs proportional to class importance to adjust for the minority class. SVM is not prone to outliers as it considers only points closest the decision boundary (called support vectors) [25, 26]. In this study, oversampling slightly improved the F1-score of SVM in Model Configuration 3 compared to its F1 scores in the other model configurations that did not incorporate oversampling. Overall, SVM is the second-best classifier after Logistic Regression.

Logistic Regression is a linear classification model that learns the probability of a sample belonging to a class. It aims to look for the best decision boundary that can separate the classes. It models the posterior probability by learning the input to output mapping and minimizing the error [8, 16]. It works well even with the presence of correlated features. In this study Logistic Regression gave the most number of best scores in all four Model Configurations. In Model Configuration 1, 2 and 4 where class imbalanced issue is not addressed, we rely on the best metric for imbalanced datasets which include precision, recall, F1-score, AUPRC and Cohen Kappa statistic in which Logistic Regression is the clear winner. When the class imbalance issue was addressed via oversampling in Model Configuration 3, still Logistic Regression gave the best performance.

3. Conclusion

This study showed that Logistic Regression provided the best scores in almost all metrics (precision, recall, accuracy, f1 score (weighted), AUROC, AUPROC, and Cohen Kappa statistic in all four configurations, followed by Support Vector Machines, and then by K-Nearest Neighbors. Based on this study, the Logistic Regression model can serve as a potential model for predicting breast cancer recurrence that would enable clinicians to propose treatment options based on whether patient's features correspond to a good or bad prognosis (recurrence). This indicates the clinical utility of data mining methods for the early detection of breast cancer recurrence in post-surgical patients to save lives.

Scaling Only

			Predicted	
			Nonrecur	Recur
LR	Actual	Nonrecur	48	1
		Recur	12	4
	Precision: 0.80		**AUROC: 0.81**	
	Recall: 0.25		**AUPRC: 0.62**	
	F1-score: 0.76		**Cohen Kappa**	
	Accuracy: 0.80		**Statistic: 0.3**	

			Predicted	
			Nonrecur	Recur
NB	Actual	Nonrecur	34	15
		Recur	11	5
	Precision: 0.25		AUROC: 0.56	
	Recall: 0.31		AUPRC: 0.37	
	F1-score: 0.61		Cohen Kappa	
	Accuracy: 0.60		Statistic: 0.01	

			Predicted	
			Nonrecur	Recur
KNN	Actual	Nonrecur	46	3
		Recur	15	1
	Precision: 0.25		AUROC: 0.50	
	Recall: 0.06		AUPRC: 0.27	
	F1-score: 0.66		Cohen Kappa	
	Accuracy: 0.60		Statistic: 0.01	

			Predicted	
			Nonrecur	Recur
SVM	Actual	Nonrecur	48	1
		Recur	15	1
	Precision: 0.50		AUROC: 0.52	
	Recall: 0.06		AUPRC: 0.40	
	F1-score: 0.67		Cohen Kappa	
	Accuracy: 0.75		Statistic: 0.01	

Scaling with PCA

			Predicted	
			Nonrecur	Recur
LR	Actual	Nonrecur	48	1
		Recur	12	4
	Precision: 0.80		**AUROC: 0.75**	
	Recall: 0.25		**AUPRC: 0.62**	
	F1-score: 0.76		**Cohen Kappa**	
	Accuracy: 0.80		**Statistic: 0.3**	

			Predicted	
			Nonrecur	Recur
NB	Actual	Nonrecur	44	5
		Recur	15	1
	Precision: 0.17		AUROC: 0.60	
	Recall: 0.06		AUPRC: 0.23	
	F1-score: 0.64		Cohen Kappa	
	Accuracy: 0.69		Statistic: 0.05	

			Predicted	
			Nonrecur	Recur
KNN	Actual	Nonrecur	47	2
		Recur	15	1
	Precision: 0.33		AUROC: 0.54	
	Recall: 0.06		AUPRC: 0.31	
	F1-score: 0.66		Cohen Kappa	
	Accuracy: 0.74		Statistic: 0.03	

			Predicted	
			Nonrecur	Recur
SVM	Actual	Nonrecur	47	2
		Recur	14	2
	Precision: 0.50		AUROC: 0.64	
	Recall: 0.12		AUPRC: 0.42	
	F1-score: 0.69		Cohen Kappa	
	Accuracy: 0.75		Statistic: 0.11	

Scaling + PCA + Oversampling

			Predicted	
			Nonrecur	Recur
LR	Actual	Nonrecur	37	12
		Recur	6	10
	Precision: 0.45		**AUROC: 0.74**	
	Recall: 0.62		**AUPRC: 0.59**	
	F1-score: 0.74		**Cohen Kappa**	
	Accuracy: 0.72		**Statistic: 0.34**	

Selected Features

			Predicted	
			Nonrecur	Recur
LR	Actual	Nonrecur	43	4
		Recur	12	6
	Precision: 0.60		**AUROC: 0.78**	
	Recall: 0.33		**AUPRC: 0.56**	
	F1-score: 0.73		**Cohen Kappa**	
	Accuracy: 0.75		**Statistic: 0.29**	

			Predicted	
			Nonrecur	Recur
NB	Actual	Nonrecur	31	18
		Recur	8	8
	Precision: 0.31		AUROC: 0.60	
	Recall: 0.50		AUPRC: 0.47	
	F1-score: 0.62		Cohen Kappa	
	Accuracy: 0.60		Statistic: 0.11	

			Predicted	
			Nonrecur	Recur
KNN	Actual	Nonrecur	36	13
		Recur	8	8
	Precision: 0.38		AUROC: 0.58	
	Recall: 0.50		AUPRC: 0.47	
	F1-score: 0.69		Cohen Kappa	
	Accuracy: 0.68		Statistic: 0.21	

			Predicted	
			Nonrecur	Recur
SVM	Actual	Nonrecur	44	5
		Recur	12	4
	Precision: 0.44		AUROC: 0.57	
	Recall: 0.25		AUPRC: 0.44	
	F1-score: 0.71		Cohen Kappa	
	Accuracy: 0.74		Statistic: 0.17	

			Predicted	
			Nonrecur	Recur
NB	Actual	Nonrecur	35	12
		Recur	11	7
	Precision: 0.37		AUROC: 0.68	
	Recall: 0.39		AUPRC: 0.46	
	F1-score: 0.65		Cohen Kappa	
	Accuracy: 0.65		Statistic: 0.13	

			Predicted	
			Nonrecur	Recur
KNN	Actual	Nonrecur	42	5
		Recur	14	4
	Precision: 0.44		AUROC: 0.60	
	Recall: 0.22		AUPRC: 0.44	
	F1-score: 0.67		Cohen Kappa	
	Accuracy: 0.71		Statistic: 0.14	

			Predicted	
			Nonrecur	Recur
SVM	Actual	Nonrecur	46	1
		Recur	12	2
	Precision: 0.67		AUROC: 0.54	
	Recall: 0.11		AUPRC: 0.51	
	F1-score: 0.66		Cohen Kappa	
	Accuracy: 0.74		Statistic: 0.12	

Table 1. Confusion Matrices and Performance Metrics

4. References

[1] Breast cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer/ Accessed December 11, 2020.

[2] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *68*(6): 394–424. https://doi.org/10.3322/caac.21492

[3] Fitzmaurice, C., Akinyemiju, T. F., Al Lami, F. H., Alam, T., Alizadeh-Navaei, R., Allen, C., ... & Yonemoto, N. (2018) Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2016: a systematic analysis for the global burden of disease study. JAMA oncology, 4(11): 1553-1568. https://doi.org/10.1001/jamaoncol.2018.2706

[4] Rosa Mendoza, E. S., Moreno, E., & Caguioa, P. B. (2013) Predictors of early distant metastasis in women with breast cancer. Journal of cancer research and clinical oncology, 139(4): 645–652. https://doi.org/10.1007/s00432-012-1367-z

[5] Riggio, A.I., Varley, K.E. & Welm, A.L. (2021) The lingering mysteries of metastatic recurrence in breast cancer. Br J Cancer 124: 13–26. https://doi.org/10.1038/s41416-020-01161-4

[6] A. I. Pritom, M. A. R. Munshi, S. A. Sabab and S. Shihab. (2016) Predicting breast cancer recurrence using effective classification and feature selection technique. *2016 19th International Conference on Computer and Information Technology (ICCIT)*, pp. 310-314, doi: 10.1109/ICCITECHN.2016.7860215.

[7] Mosayebi A, Mojaradi B, Bonyadi Naeini A, Khodadad Hosseini SH (2020) Modeling and comparing data mining algorithms for prediction of recurrence of breast cancer. PLOS ONE 15(10): e0237658. https://doi.org/10.1371/journal.pone.0237658

[8] Pedro Henriques Abreu, Miriam Seoane Santos, Miguel Henriques Abreu, Bruno Andrade, and Daniel Castro Silva. (2016) Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review. ACM Comput. Surv. 49, 3, Article 52. 40 pages. DOI:https://doi.org/10.1145/2988544

[9] Kumar V., Mishra B.K., Mazzara M., Thanh D.N.H., Verma A. (2020) Prediction of Malignant and Benign Breast Cancer: A Data Mining Approach in Healthcare Applications. In: Borah S., Emilia Balas V., Polkowski Z. (eds) Advances in Data Science and Management. Lecture Notes on Data Engineering and Communications Technologies, vol 37. Springer, Singapore. https://doi.org/10.1007/978-981-15-0978-0_43

[10] Lou, S. J., Hou, M. F., Chang, H. T., Chiu, C. C., Lee, H. H., Yeh, S. J., & Shi, H. Y. (2020). Machine Learning Algorithms to Predict Recurrence within 10 Years after Breast Cancer Surgery: A Prospective Cohort Study. Cancers, 12(12), 3817. https://doi.org/10.3390/cancers12123817

[11] Roberto Cesar, M. O., German, L. B., Paola Patricia, A. C., Eugenia, A. R., Elisa Clementina, O. M., Jose, C. O., Marlon Alberto, P. M., Fabio Enrique, M. P., & Margarita, R. V. (2020). Method Based on Data Mining Techniques for Breast Cancer Recurrence Analysis. Advances in Swarm Intelligence: 11th International Conference, ICSI 2020, Belgrade, Serbia, July 14–20, 2020, Proceedings, 12145, 584–596. https://doi.org/10.1007/978-3-030-53956-6_54

[12] N. Fatima, L. Liu, S. Hong and H. Ahmed. (2020) Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. IEEE Access, vol. 8 pp. 150360-150376. doi: 10.1109/ACCESS.2020.3016715.

[13] Aishwarja A.I., Eva N.J., Mushtary S., Tasnim Z., Khan N.I., Islam M.N. (2021) Exploring the Machine Learning Algorithms to Find the Best Features for Predicting the Breast Cancer and Its Recurrence. In: Vasant P., Zelinka I., Weber GW. (eds) Intelligent Computing and Optimization. ICO 2020. Advances in Intelligent Systems and Computing, vol 1324. Springer, Cham. https://doi.org/10.1007/978-3-030-68154-8_48

[14] Temesgen Abera Asfaw. (2019) Comparative Analysis Of Classification Approaches For Breast Cancer. International Journal of Computer Engineering and Technology (IJCET) - Scope Database Indexed.Volume:10,Issue:4,Pages:10-16. http://www.iaeme.com/IJCET/issues.asp?JType=IJCET&VType=10&IType=4

[15] Laila Khairunnahar, Mohammad Abdul Hasib, Razib Hasan Bin Rezanur, Mohammad Rakibul Islam, Md Kamal Hosain. (2019) Classification of malignant and benign tissue with Logistic Regression. Informatics in Medicine Unlocked, Volume 16, 100189, ISSN 2352-9148. https://doi.org/10.1016/j.imu.2019.100189.

[16] Ak MF. (2020) A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications. Healthcare (Basel). 8(2):111. doi: 10.3390/healthcare8020111.

[17] Borges, L. (2015) Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection. Workshop de Visa~o Computacional.

[18] Mohammed, S. A., Darrab, S., Noaman, S. A., & Saake, G. (2020) Analysis of Breast Cancer Detection Using Different Machine Learning Techniques. Data Mining and Big Data: 5th International Conference, DMBD 2020, Belgrade, Serbia, July 14–20, 2020, Proceedings, 1234, 108–117. https://doi.org/10.1007/978-981-15-7205-0_10.

[19] Breast Cancer Wisconsin Data Set, available at: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Prognostic%29.

[20] Branco, P., Torgo, L., & Ribeiro, R. P. (2016) A survey of predictive modeling on imbalanced domains. ACM Computing Surveys (CSUR), 49(2), 1-50. https://doi.org/10.1145/2907070.

[21] Islam, M.M., Haque, M.R., Iqbal, H. *et al.* (2020) Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. *SN COMPUT. SCI.* **1,** 290. https://doi.org/10.1007/s42979-020-00305-w

[22] Kim, W., Kim, K. S., & Park, R. W. (2016). Nomogram of Naive Bayesian Model for Recurrence Prediction of Breast Cancer. Healthcare informatics research, 22(2), 89–94. https://doi.org/10.4258/hir.2016.22.2.89.

[23] Yang C, Yang J, Liu Y and Geng X. (2020). Cancer Risk Analysis Based on Improved Probabilistic Neural Network. Front. Comput. Neurosci. 14:58. doi: 10.3389/fncom.2020.00058.

[24] Desuky, A.S., Hussain, S. (2021) An Improved Hybrid Approach for Handling Class Imbalance Problem. Arab J Sci Eng 46, 3853–3864. https://doi.org/10.1007/s13369-021-05347-7

[25] Ahmad L.G., Eshlaghy A.T., Poorebrahimi A., Ebrahimi M., Razavi A.R. (2013). Using three machine learning techniques for predicting breast cancer recurrence. Journal of Health & Medical Informatics. 4(2): 124.

[26] Huang M-W, Chen C-W, Lin W-C, Ke S-W, Tsai C-F (2017) SVM and SVM Ensembles in Breast Cancer Prediction. PLoS ONE 12(1): e0161501. https://doi.org/10.1371/journal.pone.0161501

[27] K. M. M. Lopez and M. S. A. Magboo, "A Clinical Decision Support Tool to Detect Invasive Ductal Carcinoma in Histopathological Images Using Support Vector Machines, Naïve-Bayes, and K-Nearest Neighbor Classifiers," A. Tallón-Ballesteros and C.-H. Chen, Eds. Netherlands: IOS Press BV, 2020, pp. 46–53.

[28] Yang, P., Wu, W., Wu, C., Shih, Y., Hsieh, C. & Hsu, J. (2021). Breast cancer recurrence prediction with ensemble methods and cost-sensitive learning. Open Medicine, 16(1), 754-768. https://doi.org/10.1515/med-2021-0282

[29] David A. Omondiagbe *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **495** 012033

[30] Zahra Nematzadeh, Roliana Ibrahim and Ali Selamat, (2015). Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques, Proc. in 2015 10th Asian Control Conf. (ASCC), pp 1-6, IEEE.

[31] Magboo, M. S. A., & Coronel, A. D. (2019). 30-Day Hospital Readmission Prediction Model for Diabetic Patients within the 30-70 Age Group. Proceedings of the Academics World 130th International Conference, Madrid, Spain, 10th - 11th June, 2019, 1–8. https://www.worldresearchlibrary.org/up_proc/pdf/2968-15656902101-8.pdf

[32] Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019) The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, 91, 216-231. ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2019.02.023. (https://www.sciencedirect.com/science/article/pii/S0031320319300950)

[33] Landis, J., & Koch, G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1), 159-174. doi:10.2307/2529310

[34] Zain, Zuhaira Muhammad et al. (2020). Predicting breast cancer recurrence using principal component analysis as feature extraction: an unbiased comparative analysis. International Journal of Advances in Intelligent Informatics, [S.l.], v. 6, n. 3, p. 313-327. ISSN 2548-3161. Available at: http://ijain.org/index.php/IJAIN/article/view/462%7Cto_array%3A0

[35] Bian, K., Zhou, M., Hu, F., & Lai, W. (2020). RF-PCA: A New Solution for Rapid Identification of Breast Cancer Categorical Data Based on Attribute Selection and Feature Extraction. Frontiers in genetics, 11, 566057. https://doi.org/10.3389/fgene.2020.566057