Research on Prediction of Breast Cancer Type using Machine Learning

Dehui Kong *

School of statistics, Renmin University of China, Beijing, 100872, China

* Corresponding author: kongdehui@bjeaedu.com

Abstract. The most typical cancer type among women worldwide is breast cancer. In 2020 alone, it afflicts about 0.68 million people and 6.9% of all cancer cases. How to categorize tumors as benign (non-cancerous) or malignant (cancerous) is one of the main obstacles to its diagnosis. This study helps to make an accurate and reliable diagnosis based on the initial data of the tumor, such as smoothness, texture, area using machine learning models. This study uses five machine learning models, Logistic Regression (RF), Random Forest (RF), Support Vector Machine (SVM), K-nearest Neighbor (KNN), Naive Bayes Classifier (NBC) and three modelling systems, feature selection-ML and principal component analysis (PCA)-ML system to make predictions of the type of the tumor of Wisconsin Breast Cancer Dataset. Model performance are assessed by three performance evaluation which are accuracy, precision, recall. The results of full model show that random forest has the highest prediction accuracy of 98.25% out of the sample and 100% in the sample, and SVM's sigmoid-based kernel model has the lowest prediction accuracy of 83.33% outside and 85.27% inside the sample. The results of the feature selection model based on RF and LR shows that the RF with only 13 variables has the highest prediction accuracy 98.25% out-of-sample and 100% in-sample. Among all the PCA--ML models, PCA--NBC has the highest prediction accuracy of 97.33% out-of-sample. Nevertheless, PCA-RF has the highest prediction accuracy of 100% in-sample.

Keywords: Breast cancer, classification, machine learning.

1. Introduction

A Cancer Journal for Clinicians, the journal of the American Cancer Society forecasts the number of new cases and fatalities from cancer in the United States in that year. By 2023, there will be 1,958,310 new instances of cancer and 609,820 cancer-related deaths in the US. Among other things, mortality from female breast cancer has also plummeted, dropping 43% since 1989, with an average 5-year survival rate of 90% [1].

Early diagnosis and treatment are the key to reduce the mortality rate of breast cancer. Unlike other malignant tumors, female breast cancer can be detected early through self-examination and early screening, thus increasing the cure rate of breast cancer and reducing the death caused by breast cancer.

Cancer typically begins when cells in the breast start to proliferate uncontrollably. The breast tissue's expandable cells cause fast cell division, which results in lumps. These lumps, which can be felt as lumps in the breast region or seen on an X-ray, are also known as tumors. Tumors can be categorized as benign or malignant. Malignant tumors destroy body tissues by spreading aberrant cells throughout them. In contrast, benign tumors can develop anywhere in the body, but they do not spread or metastasis to other areas of the body. Breast cancer in the breast can be triggered by malignant tumors [2]. Therefore, in order to greatly increase survival rates and establish a probability of recovery, treatment should be taken into account for a proper tumor diagnosis.

The main obstacle to its discovery is determining whether a tumor is malignant (cancerous) or benign (non-cancerous). Benign tumors and malignant tumors often differ in smoothness, texture, area, circumference, and radius, and traditional treatment modalities involve imaging, such as ultrasound and mammography [3], which can show further differences in the mass. These tests can then be used to determine whether the mass is benign or malignant. However, some tumors cannot be clearly identified as benign or malignant by these non-invasive examinations, the final point of difference between benign and malignant is still through aspiration to see the pathological results.

Despite the importance of these strategies, it lacks satisfaction in diagnostic performance [4]. Physician interpretation of mammography may vary because of the limitations of false-positive results [5] and false-negative results on mammograms [6].

There are many machine learning classification models that can be used to improve the accuracy of diagnostic results, such as K-nearest Neighbor, Random Forest, Support Vector Machine, Naive Bayes, Logistic Regression, Decision Trees, C5.0, Adaboost, Bagging, Artificial neural network. There have been many studies using these models to analyze the accuracy of diagnosis based on Wisconsin Diagnostic Breast Cancer (WDBC) dataset. SVM algorithm has higher sensitivity, specificity, and lower misclassification error than other ML algorithms, allowing it to diagnose breast cancer as triple negative and non-triple negative with greater accuracy [7]. SVM has the highest prediction accuracy than another remaining algorithm such as LR, LDA, QDA, RF [8]. Raman spectroscopy and machine learning methods allow for quick investigation of breast cancer subtypes, revealing differences in intracellular composition and molecular structure [9]. The accuracy of the combination of LASSO+LR is almost 99%, which is better than the LR classification without any feature selection [10]. With the help of artificial intelligence tools, many applications can detect breast cancer even if the human eye cannot see the obvious tumor. To aid radiologists in identifying medical images, several deep learning tools and methodologies have been created [11]. AdaBoost obtained the highest accuracy of breast cancer classification based on the WDBC dataset [12].

Although the use of machine learning techniques applied to the further diagnosis of breast tumor cells is of great interest, combining PCA and feature selection with other machine learning algorithm are still worthy to study further. The cancer dataset is usually high-dimension with many features, it's necessary that cancer dataset is first dimensionally reduced or feature selection and then classified based on machine learning algorithms. The accuracy of the training model can be increased while minimizing the sophistication by using feature selection or feature extraction via PCA.

In this research, experiments were conducted to classify breast cancers as benign and malignant using nine machine learning algorithm including Logistic Regression (LR), Principal Component Analysis (PCA)- LR, Random Forest (RF), PCA-RF, K-Nearest Neighbor (KNN), PCA-KNN, Support Vector Machine (SVM), Naive Bayes Classifier (NBC), PCA-NBC based on the Wisconsin Diagnostic Breast Cancer (WDBC) database. The purpose of the study is to evaluate the accuracy, precision, and recall of breast cancer classification.

2. Materials and Methods

2.1. Breast Cancer Dataset

Wisconsin Diagnostic Breast Cancer (WDBC) dataset are obtained from Kaggle. There are 569 observations and 32 variables in the dataset. The first two variables are ID and diagnosis which are integer-value and categorical variable respectively. The remaining 30 variables are numerical values. The response variable in the study is diagnosis which is categorized into two value malignant with 212 observations and benign with 357 observations. There are 10 features related to tumor which are radius, texture, perimeter, area, smoothness, compactness, concavity, concave, symmetry and fractal dimensions. Each feature has three statistics: mean, standard error and worst. To sum up, there are 30 explanatory variables in the dataset.

2.1.1. Descriptive statistics

The summary statistics of WDBC dataset is shown in the Table 1.

Table 1. Descriptive statistics of dataset

Features	Range		
	mean	standard error	worst
Radius	6.981 - 28.11	0.112 - 2.873	7.93 - 36.04
Texture	9.71 - 39.28	0.360 - 4.885	12.02 - 49.54
Perimeter	43.79 - 188.5	0.757 - 21.98	50.41 - 251.2
Area	143.5 - 2501	6.802 - 542.2	185.2 - 4254
Smoothness	0.053 - 0.163	0.0017 - 0.031	0.0712 - 0.223
Compactness	0.019 - 0.345	0.0023 - 0.135	0.027 - 1.058
Concavity	0 - 0.427	0 - 0.396	0 - 1.252
Concave points	0 - 0.201	0 - 0.0528	0 - 0.291
Symmetry	0.106 - 0.304	0.0079 - 0.079	0.157 - 0.664
Fractal dimensions	0.050 - 0.097	0.00089 - 0.030	0.055 - 0.208

2.1.2. Graphical analysis

As shown from the Figure 1, the radius, area, texture, perimeter of malignant cells is larger than cells of benign. Malignant cancer cells are larger in size than benign cancer cells. The area occupied by malignant cancer cells is greater than that of benign cancer cells.

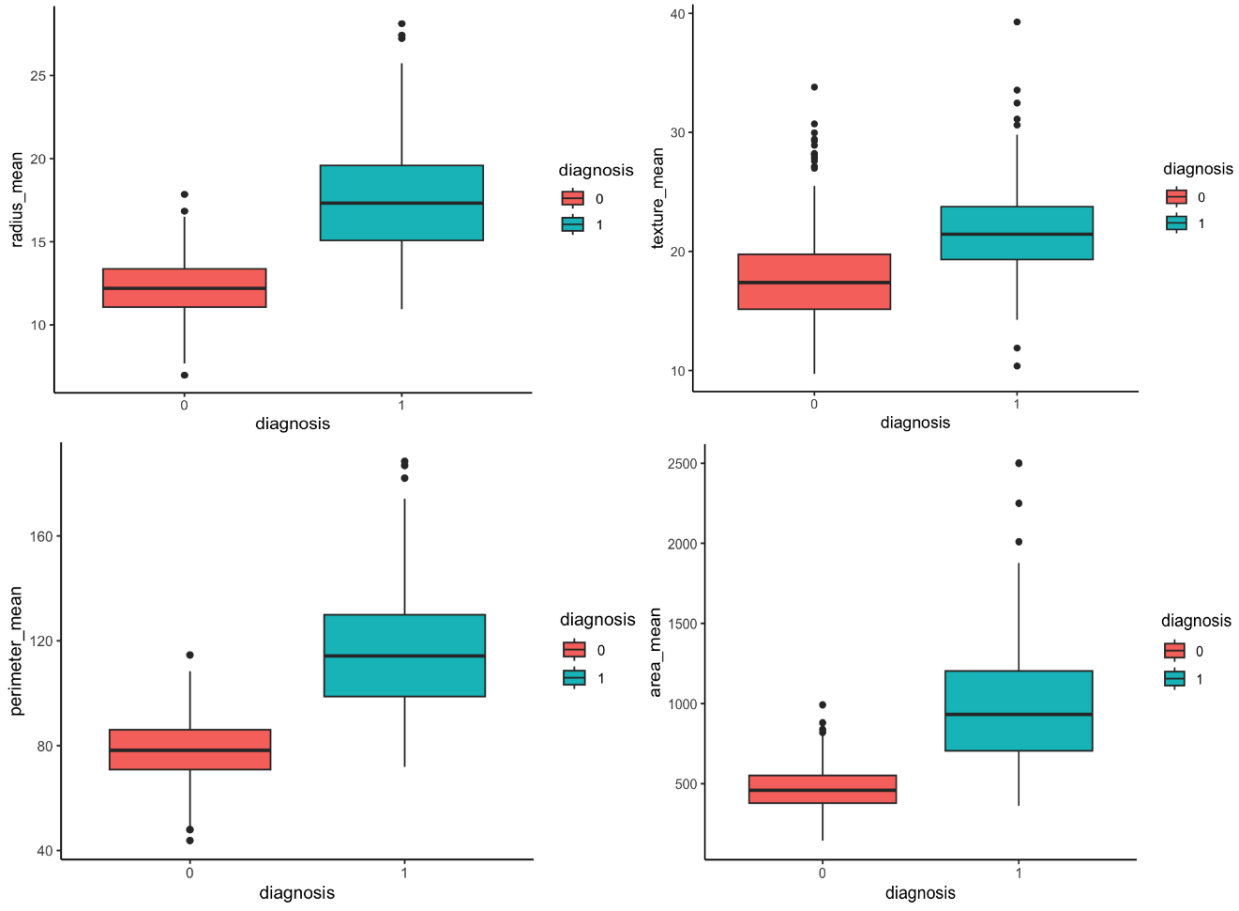

Figure 1. Boxplots of diagnosis VS features

Figure 2 shows the correlation between the 30 variables, there shows some strong correlation between these variables with dark blue areas. There must be multicollinearity between variables. When there are precise correlations or significant correlations between the explanatory variables in a regression model, the model estimates become distorted or difficult to estimate effectively. This is called the multicollinearity. It is particularly important to perform PCA to reduce the dimensionality of data that contains repetitive information before modeling.

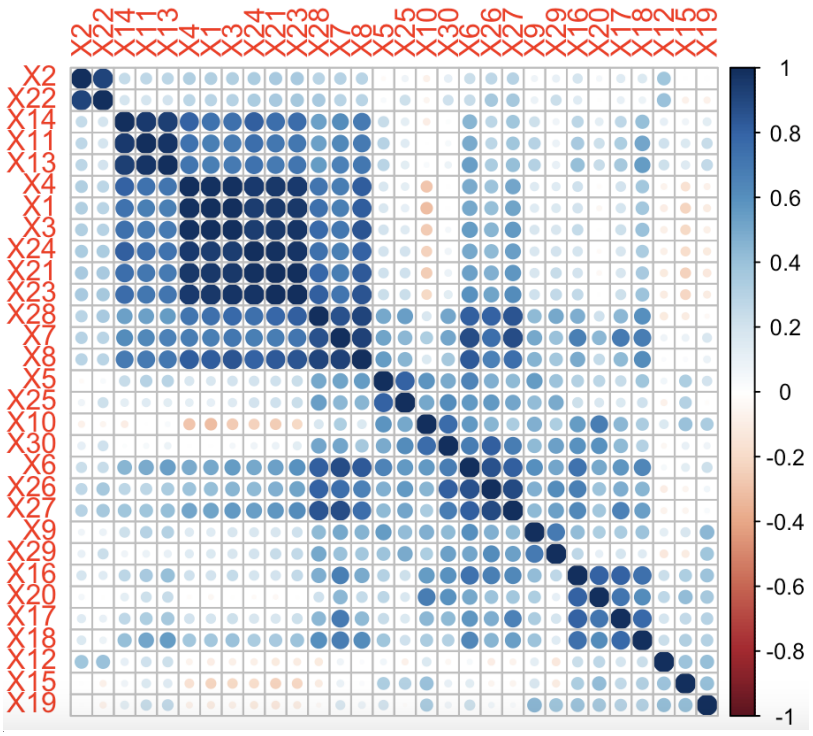

Figure 2. Correlation between variables

2.2. Performance Evaluation

Accuracy, sensitivity, specificity, recall are common evaluation methods for classification. And all these measurements can be calculated from the Confusion Matrix. The Confusion Matrix is defined in Table 2.

Table 2. Confusion Matrix

	Test outcome positive	Test outcome negtive
Actual condition positive	True postive(TP)	False negative(FN)
Actual condition negtive	False positive(FP)	True negative(TN)

The performance measurements and indicators are defined as followed.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+FP+TN}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{FN+TP}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{3}$$

In this study, the four performance measurements are calculated to assess the performance of the classifier models.

3. Results and Discussion

The dataset has one variable “ID” which consists of a string of unique integer number indicated unique observation and one response variable “diagnosis” which is categorized into two values Malignant and Benign. In this study, “ID” has been removed in the modelling part. Generally, as shown in the Fig.3, there are three modelling systems. The general idea of the first modelling system is that all the variables are introduced into the classification models LR, RF, KNN, SVM, NBC. The second modelling system is that feature selection is performed before modelling, the feature selection

is carried out dependent of the classification method. For instance, LR uses step forward, step backward and both direction, three selection methods; however, RF uses the cross validation to choose the best number of variables and uses variable importance such as mean decrease accuracy to select most important variables into model. The basic idea of the third modelling system is that there is multicollinearity between variables as mentioned in the EDA part, principal component analysis is carried out to remove the duplicate variables (closely related variables) and create new uncorrelated component and that these new components maintain as much information as possible about the dataset. Further step after PCA is that use different classification model to classify the tumor into benign and malignant. In order to choose the most effective and best classification method, performance of each model is compared based on accuracy, precision, recall.

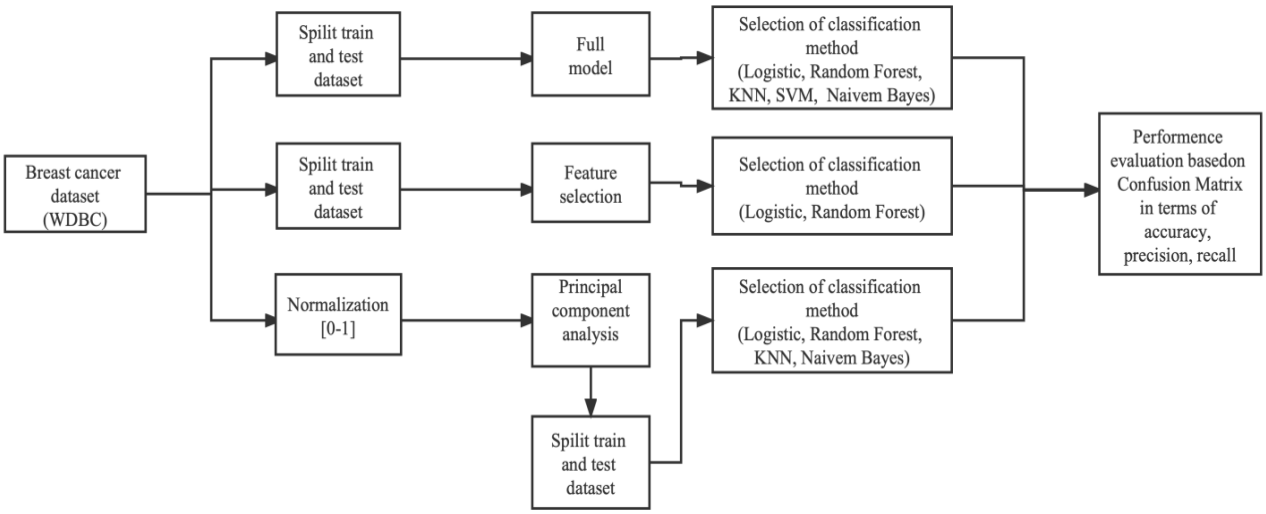

Figure 3. Process of the Classification System

The study is carried out with a ratio of training to testing of 80%:20%. After training the model on the train dataset, predictions on both the test and train datasets are obtained. The Confusion Matrix are also calculated to compute the accuracy, precision and recall. Table 3 summarizes the results of performance evaluation for classification models.

Figure 4 and Figure 5 shows the performance evaluations for classification models outside and inside of the model respectively. Apparently, all the models performs well that all the performance evaluations are greater than 75%. Furthermore, most of the measurements are greater than 90%, indicating all the classification models performs extremely well in and outside of the sample.

For accuracy measurement, outside the sample, RF and selective RF with 13 variables performs best of 98.25% accuracy and SVM with sigmoid kernel performs worst of 83.33% accuracy. Within the sample, LR, RF, selective LR with 11 variables, selective RF with 13 variables, RF--PCA (7) (7 components), RF--PCA (5) (5 components) have 100% accuracy, and SVM with sigmoid has the worst accuracy of 85.27%.

For precision measurement, outside the sample, RF, selective RF with 13 variables, LR, selective LR with 11 variables and NBC—PCA (7) has the 100% precision and SVM with sigmoid kernel has the worst precision of 75%. Within the sample, LR, RF, selective LR with 11 variables, selective RF with 13 variables, RF--PCA (7), RF--PCA (5), SVM with ploy kernel, SVM with linear kernel, SVM with radial kernel all the models have a precision of 100%. However, SVM with sigmoid kernel has the worst precision of 81.18%.

Table 3. Performance Evaluation for Classification Models

Modelling System	Model	outside sample(20%)			within sample(80%)		
		Accuracy(%)	Precision(%)	Recall(%)	Accuracy(%)	Precison(%)	Recall(%)
Full Model	LR	97.37	100	92.31	100	100	100
	RF	98.25	100	95.12	100	100	100
	SVM(linear)	96.49	92.68	97.44	99.12	100	97.69
	SVM(poly)	92.11	85.71	92.31	99.78	100	99.42
	SVM(radial)	90.35	80.43	94.87	99.12	100	97.69
	SVM(sigmoid)	83.33	75	76.92	85.27	81.18	79.77
	KNN	97.37	97.50	95.12	97.36	99.38	93.57
	NBC	90.35	90.91	85.11	95.51	93.21	91.52
Selection Model	LR(11)	97.37	100	92.31	100	100	100
	RF(13)	98.25	100	95.12	100	100	100
PCA + classification method	LR+PCA(7)	96.49	97.30	92.3	98.24	97.69	97.69
	LR+PCA(5)	96.49	97.3	92.31	98.24	98.25	97.11
	RF+PCA(7)	96.49	95.12	95.12	100	100	100
	RF+PCA(5)	96.49	95.12	95.12	100	100	100
	KNN+PCA(7)	92.98	94.74	85.71	96.70	99.36	91.76
	KNN+PCA(5)	91.23	97.06	78.57	95.60	99.34	88.82
	NBC+PCA(7)	97.34	100	92.68	92.09	90.91	87.72
	NBC+PCA(5)	95.61	97.37	90.24	91.87	92.95	84.80

For recall measurement, outside the sample, SVM with linear kernel has the highest recall of 97.44% among all the models. However, SVM with sigmoid kernel has the worst recall of 76.92% among all the models. Within the sample, RF, selective RF with 13 variables, LR, selective LR with 11 variables, RF--PCA (7), RF--PCA (5) have the highest recall of 100%. Still, SVM with sigmoid kernel has the worst recall of 79.77% among all the model.

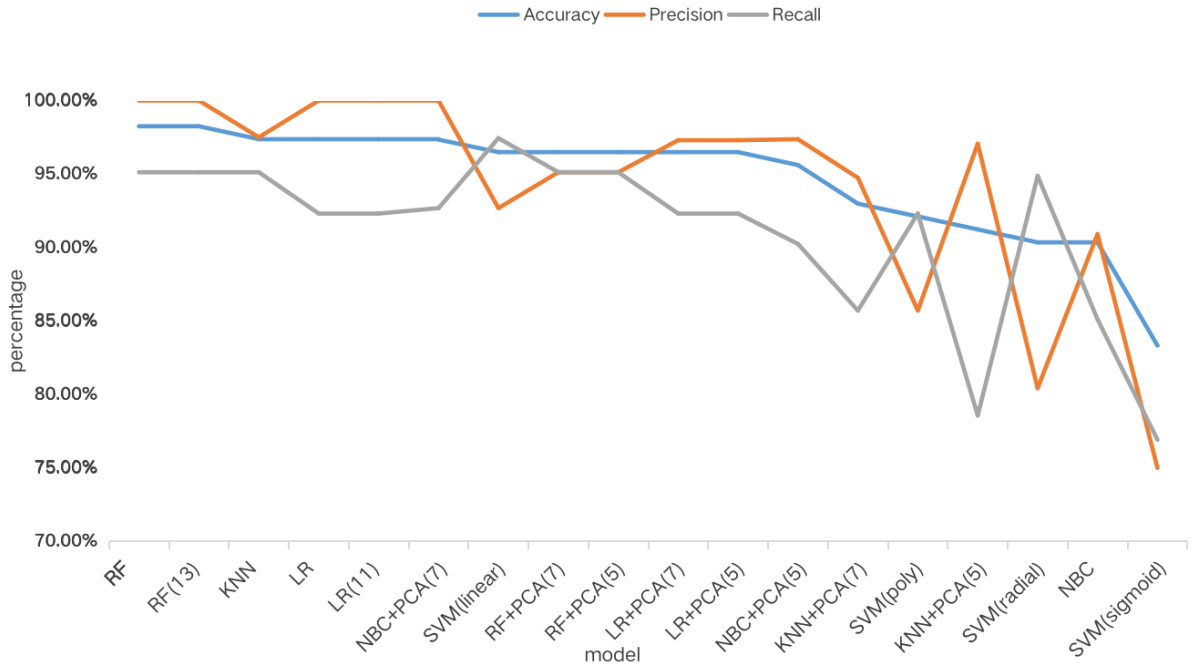

Figure 4. Comparison of performance evaluation on the testing dataset

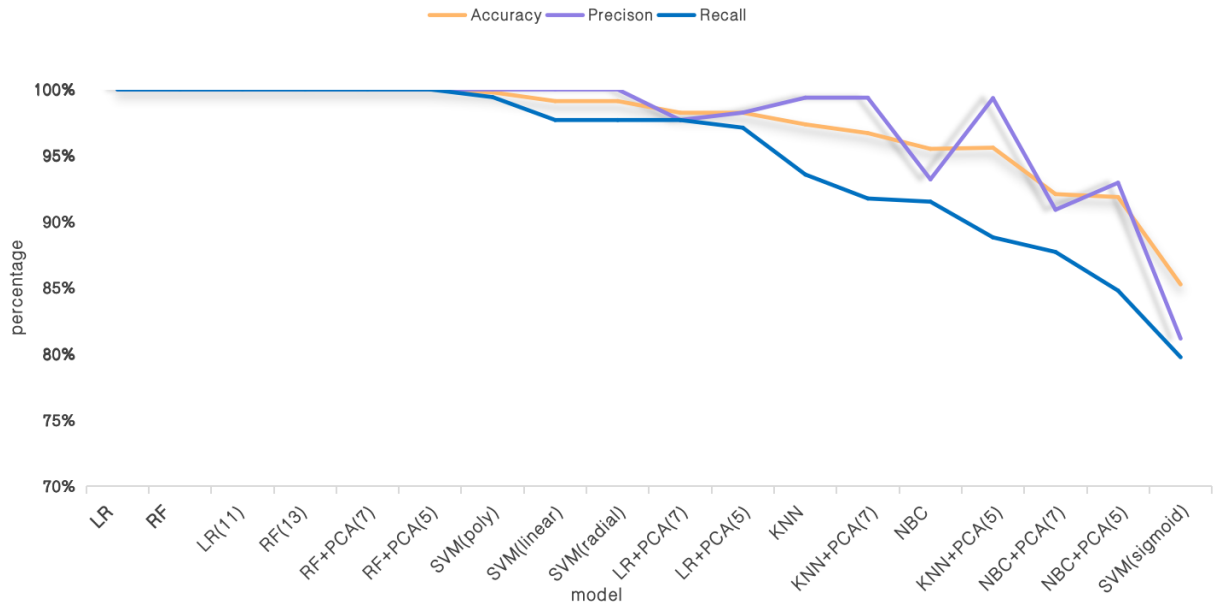

Figure 5. Comparison of performance evaluation on the training dataset

From the perspective of the modelling system's, conclusions might be a little bit different. For full model, RF has the highest accuracy of 98.25% and highest precision of 100% outside the sample. SVM with linear kernel has the highest recall of 97.44% outside the sample. Moreover, RF has the highest 100% on the three measurements within the sample. SVM with sigmoid kernel still performs the worst on all the measurements both outside and within the sample. For selective model, RF with 13 variables performs the best on all three measurements both within and outside of the sample. For PCA--Classification models, outside the sample, NBC--PCA (7) has the highest accuracy of 97.34% and precision of 100%, RF--PCA (5) and RF--PCA (7) has the highest recall of 95.12%. Within the sample, RF--PCA (5) and RF--PCA (7) performs the best on all three measurements of 100%.

4. Conclusion

This study used three modeling systems which consist of five common machine learning classification models to predict the type of the breast tumor which has two types benign and malignant based on the Wisconsin Diagnostic Breast Cancer (WDBC) database. In the pre-processing part of the study, the dataset is split into test and train dataset in the ratio of 80% to 20%. In the three-modelling system, classification models are trained in the train dataset, and then predictions are calculated based on the test dataset. Moreover, predictions on train dataset are also calculated for the purpose of further study and comparison. Confusion Matrix and three performance evaluations in terms of accuracy, precision, and recall are calculated for all the classification models to assess the performance of the models. In the full modelling system, eight classification models are carried out. Among all the models in the full model, RF has the highest accuracy of 98.25% and highest precision of 100% outside the sample. SVM with linear kernel has the highest recall of 97.44% outside the sample. Moreover, RF has the highest 100% on the three measurements within the sample. SVM with sigmoid kernel performs the worst on all the measurements both outside and within the sample. In the selective modelling system, two models are carried out. Among all the models in the selective model, RF with 13 variables performs the best on all three measurements both within and outside of the sample. In the PCA--Machine Learning models, NBC--PCA (7) has the highest accuracy of 97.34% and precision of 100%, RF--PCA (5) and RF--PCA (7) has the highest recall of 95.12% outside the sample. However, RF--PCA (5) and RF--PCA (7) performs the best on all three measurements of 100% within the sample. Among all the classification models, Random Forest "family" has the strongest performance both inside and outside of the sample. Moreover, RF with 13 variables has the same performance as the full model but lower model complexity as well as RF--PCA (7). According

to the results, PCA--Machine learning classification models all have high performance evaluation indicating that combine the principal component analysis and machine learning classification models to predict the type of the breast tumor has a great significance and worth more study in the future.

References

[1] Siegel R L, et al. Cancer statistics. CA Cancer J Clin, 2023, 73 (1): 17 - 48.

[2] Mashudi N A, Rossli S A, Ahmad N, Mohd Noor N. Breast Cancer Classification: Features Investigation using Machine Learning Approaches. International Journal of Integrated Engineering, 2021.

[3] Kalaf J M. Mammography: A history of success and scientific enthusiasm. Radiol Bras, 2014.

[4] Bellaachia A, Guven E. Predicting breast cancer survivability using data mining techniques. Age, 2016, 58 (13), 10 - 110.

[5] Saygılı A. Classification and Diagnostic Prediction of Breast Cancers via Different Classifiers. International Scientific and Vocational Studies Journal, 2018, 2 (2): 48 - 56.

[6] Amrane M, Oukid S, Gagaoua I, Ensarİ T. Breast cancer classification using machine learning, in 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 2018, 1 - 4.

[7] Anuradha R. Support Vector Machine Classifier for Prediction of Breast Malignancy Using Wisconsin Breast Cancer Dataset. Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN), 2022.

[8] Oyewola D, Hakimi D, Adeboye K, Shehu M D. Using Five Machine Learning for Breast Cancer Biopsy Predictions Based on Mammographic Diagnosis, 2017.

[9] Zhang L, et al. Raman spectroscopy and machine learning for the classification of breast cancers. Working paper, 2022.

[10] Nursabillilah M, et al. Comparison of microarray breast cancer classification using support vector machine and logistic regression with LASSO and boruta feature selection. Indonesian Journal of Electrical Engineering and Computer Science, 2022, 20 (2): 712 – 719.

[11] Ahmed A, et al. A Neutrosophic based C-Means Approach for Improving Breast Cancer Clustering Performance. Working paper, 2023.

[12] Mashudi N A, Rossli S A, Ahmad N, Mohd N. Breast Cancer Classification: Features Investigation using Machine Learning Approaches. International Journal of Integrated Engineering, 2021.