Wisconsin Breast Cancer Dataset Analysis

Abstract Breast cancer is the second leading cause of death among women in the United States [1]. A prediction of breast cancer in an early stage can improve survivor rates. It is difficult for doctors to achieve an accurate diagnosis of disease before its treatment based on complicated tests. It is necessary to create breast cancer prediction tools to differentiate malignant tumors from benign tumors. The purpose of the paper is to provide the reader with insight on Wisconsin Breast Cancer Dataset (WBCD). The analysis involved building and comparing the integrated approaches. In this paper, five algorithms of machine learning including Decision Tree (DT), Random Forest (RF), Gradient Boosting and Ada Boosting, Support Vector Machine (SVM) and Artificial Neural Network (ANN) will be compared for creating classification models. Prior to the classification, exploratory data analysis methods will be conducted for finding relationships between dependent attributes and the target variable. Principal Component Analysis (PCA), Random Forest (RF), and Wrapper methods will be employed to extracting features. The paper will show the results of the five methods on Wisconsin Breast Cancer Dataset. The paper uses metrics such as accuracy, sensitivity, and specificity and AUC scores to compare the test results. The experimental results have shown that the Neural Net Work method with feature selection have produced the optimal performance with an accuracy of 97%.

1. Introduction

Biomedical engineering is a collection of theories of engineering and cutting-edge technologies for the first stage of diagnostic tool and information systems to fill the gaps between engineering and medicine. Physicians and doctors use their experiences and knowledge to diagnose a disease [2]. It is evident that physicians are challenged by analyzing certain diseases because new results outcast the old ones; novel treatment techniques and medicines are introduced each day, and occurrence of rare and unique diseases bring experienced physicians and newcomers at the same level. Therefore, it is important to analyze the ability of computer-aided models to promote decision accuracy by improving the natural capacities of physicians.

Breast cancer is the most common disease among females. According to recent data, breast cancer is the second-leading cause of death in women [1]. Women in developed nations have a much higher occurrence rate than those in developing nations. The primitive detection is the crucial aspect in its diagnosis, and machine- learning and soft computing help its simplification [3].

Machine-learning can produce models to make predictions, find unusual patterns, and trends by using the training set, while soft computing is the mixture of methodologies, which in synergism improves the flexibility of information processing for dealing with real-life ambiguous situations [4]. In medical diagnosis, the majority is classificatory problems, and the approach a computer paradigm follows is that it learns significant features of a dataset so that the user could classify the input data into one or more classes based on the historical data [5]. Consequently, artificial neural networks and fuzzy logic are extensively used for classification in intelligent systems. The neural network consists of numerous neurons and connections, which operate in parallel to solve complex problems while the wrapper methods involve building multiple models on different sets of features, looking for the optimal one with different kinds of search such as greedy, random, genetic algorithms. Feature subset selection ranks features and could be a method to handle high dimensionality by removing irrelevant and redundant features. Therefore, these techniques can promote the learning performance like predictive accuracy and improve the transparency of learned results for the classification algorithms. In this paper, the integrated approach is based on using dimensionality reduction and feature ranking at the first step for reducing the number of attributes to deal with. In the second stage, five machine learning methods, Decision Tree, Random Forest, Gradient Boosting and Ada Boosting, Support Vector Machine, are used to perform classification of the clinical instances into benign and malignant. The entire approach is developed to increase learning performance and reduce computational complexity due to high dimensionality and multicollinearity.

2 Literature Survey

The literature review significantly helped form the analysis which the paper conducted. A lot of research that has already been performed for the diagnosis of breast cancer indicates a direction to the methods the paper used. There are many techniques used for classification and prediction of breast cancer, and the primary goal of these methods is to assign patients to either benign or malignant group. Jhajharia's work is one of them, which introduced the use of a multivariate statistical approach has been combined with an artificial intelligence-based learning technique to implement a prediction model. Principal components analysis pre-processes the data by extracting the most relevant features for training, and artificial neural networks learn the patterns in the data and classify new instances [6]. Aragones, Ruiz, Jimenez, Perez, and Conejo proposed a combination of neural network and decision trees model for the prognosis of breast cancer while Shieu-Ming et al. employed an artificial neural network and multivariate adaptive regression splines [7, 8]. Integrated and hybrid techniques have also been widely used to conduct a classification of breast cancer. S¸ahan's work introduced a new mixture approach, which was based on a fuzzy-artificial immune system and a k-NN algorithm [9]. A new neural pattern recognition used two approaches of fuzzy systems, and an evolutionary algorithm was introduced. [10]. A method which is a collection of fuzzy systems and ACO algorithm produced better accuracy in diagnosing breast cancer [11]. Statistical methods including PCA, PLS linear regression analysis, data mining methods mixed with rough sets and probabilistic neural network have also been employed for classification. An integrated approach to linear discriminant analysis and principal component analysis as along with ANN and ANFIS adopts a modular approach by us in small and individual neural networks to achieve a more specific and accurate artificial neural network for attaining better accuracy [12].

The review of the past literature gives a detailed explanation of some approaches, which makes it difficult to choose an efficient and effective algorithm. It is important to conduct a comparative study of the current methodologies and methods. There are many researchers, such as Delen, Glenn Walker, and Amit Kadam, have been comparing three data mining paradigms for estimating breast cancer [13]. In another comparative study on Wisconsin breast cancer data, four fuzzy generation methods are examined and compared [14]. Ubyelihas also studied Wisconsin breast cancer data employing multilayer perceptron neural networks, combined neural networks, probabilistic neural networks, recurrent neural networks, and support vector machines [15]. R.R.Janghel compared a hybrid intelligent system using SANE (Symbolic Adaptive Neuro-Evolution) with evolutionary neural networks [16].These past works have provided a strong foundation for performing important research work in the future.

3 Methodology

The basic idea is to analyze the Wisconsin Breast Cancer Dataset (WBCD) and to classify problems of breast cancer. The dataset is from the Kaggle website [17], which contains 569 instances for 31 attributes, in which, 30 attributes are numerical dependent variables and 1 class variable that classifies between benign and malignant. The first step performed exploratory data analysis which includes correlation matrices which can check two independent variables correlation, violin plots which can find the distribution of the two groups including B=Benign and M=malignant, swarm plots which can show variance more clearly and box plots which can provide statistics of the two groups. Then Features selection methods, such as with tree-based with RF are Principal Component Analysis (PCA), and Wrapper methods, were used to select the feature and rank the features. The 16 features are ranked according to their importance by tree-based feature selection methods. Five features, including 'concave points_1ean’, 'radius_worst', 'texture_worst’, 'peri1eter_worst’, and 'concave points_worst’, were selected by Wrapper methods with parameters as activation = 'logistic', solver = 'lbfgs', alpha = 0.0001, max_iter=1000, layer_sizes=(10,), random_state=rand_st, cross_validate, cv=5. The second task is to apply soft computing techniques over the reduced dataset and not reduced data set. The classification will be done using five methods, including Decision Tree, Random Forest, Gradient Boosting and Ada Boosting, Support Vector Machine, and Neural Network, with different parameters. During the implement of the five methods in training and testing, cross validation methods are used. The results are compared and are shown in Figure 1.

3.1 Wisconsin Breast Cancer Database

The University of Wisconsin Hospital made the data in November 1995 based on fine-needle aspiration test. The creators are Dr. WilliamH. Wolberg who is in General Surgery Dep W. Nick Street, Computer Sciences Dept, and Olvi L. Mangasarian, Computer Sciences Dept at the University of Wisconsin-Madison Hospitals resulted in the Wisconsin Breast Cancer Diagnosis (WBCD) database, which is presently used and analyzed by research scientists. There is no missing attribute value. Class distribution is 357(63) benign and 212(37%) malignant and the number of instances is 569. The number of attributes is 32 (ID, diagnosis, 30 real-valued input features). Attribute information includes ID number, Diagnosis (M = malignant, B = benign). Ten real-valued features are computed for each cell nucleus: a) radius (mean of distances from center to points on the perimeter) b) texture (standard deviation of grayscale values) c) perimeter d) area e) smoothness (local variation in radius lengths) f) compactness (perimeter^2 / area - 1.0) g) concavity (severity of concave portions of the contour) h) concave points (number of concave portions of the contour) i) symmetry j) fractal dimension ("coastline approximation" – 1). The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius; field 13 is Radius SE, field 23 is Worst Radius. All feature values are recoded with four significant digits.

3.2 Visualization

Visualization can help understand the data and the results produced by the machine learning methods.

Correlation matrix:

In order to find the relationships between the independent variables, the correlation matrix was produced. Two R packages, ggplot2 and PerformanceAnalytics, were used to create the correlation matrices shown below. Figure 2 shows the correlation matrix of 30 independent variables. It shows that there 21 pairs of variables have strongly correlated.

Figure 2 correlation matrix of 30 variables

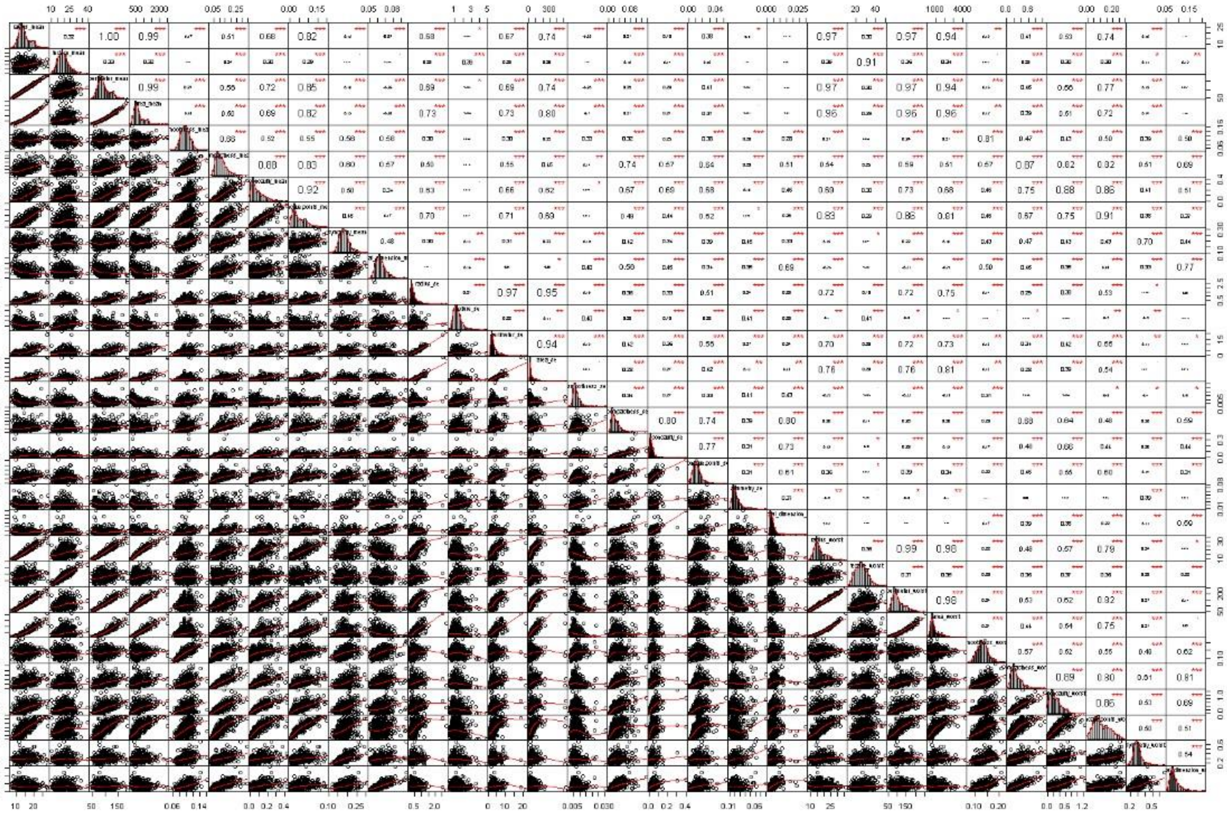


Figure 3 below shows a matrix with 10 variables from the first 10 variables with “mean” was produced to get more detailed information. Variables radius_mean and perimeter_mean has coefficient 1.00 statistically significant. Radius_mean and area_mean, area_mean, and perimeter_mean are strongly correlated with both coefficient 0.99. Also, there exists multicollinearity because variables concavity_mean and concave_point_mean correlated with multiple variables.

Figure 3 correlation matrix with 10 variables of mean from column 1-10 independent variables

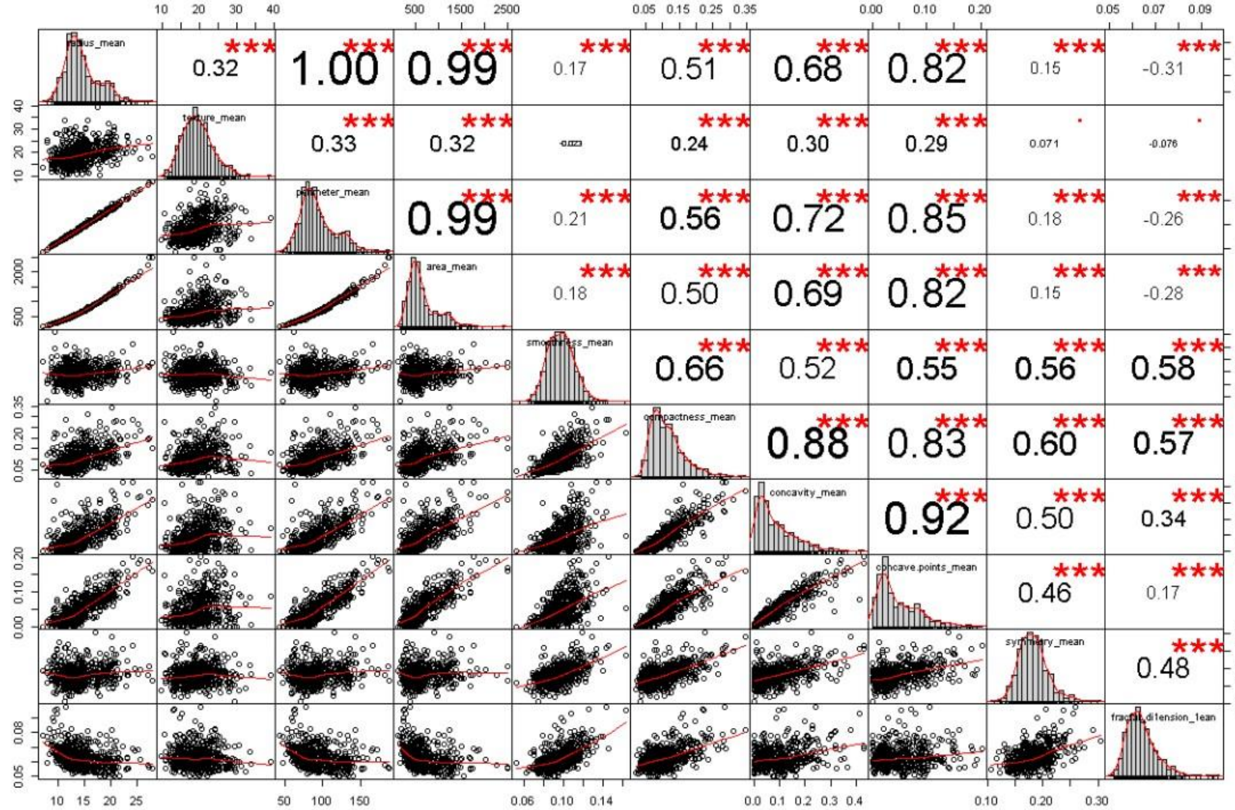


Figure 4 below shows that shows a matrix with 10 variables from the second group 10 variables with “se” (standard error). There are three pairs of variables strongly correlated. There exist one variable correlated concave_point_mean three other variables.

Figure 4 another correlation matrix with 10 variables of standard error(se) from 10-20

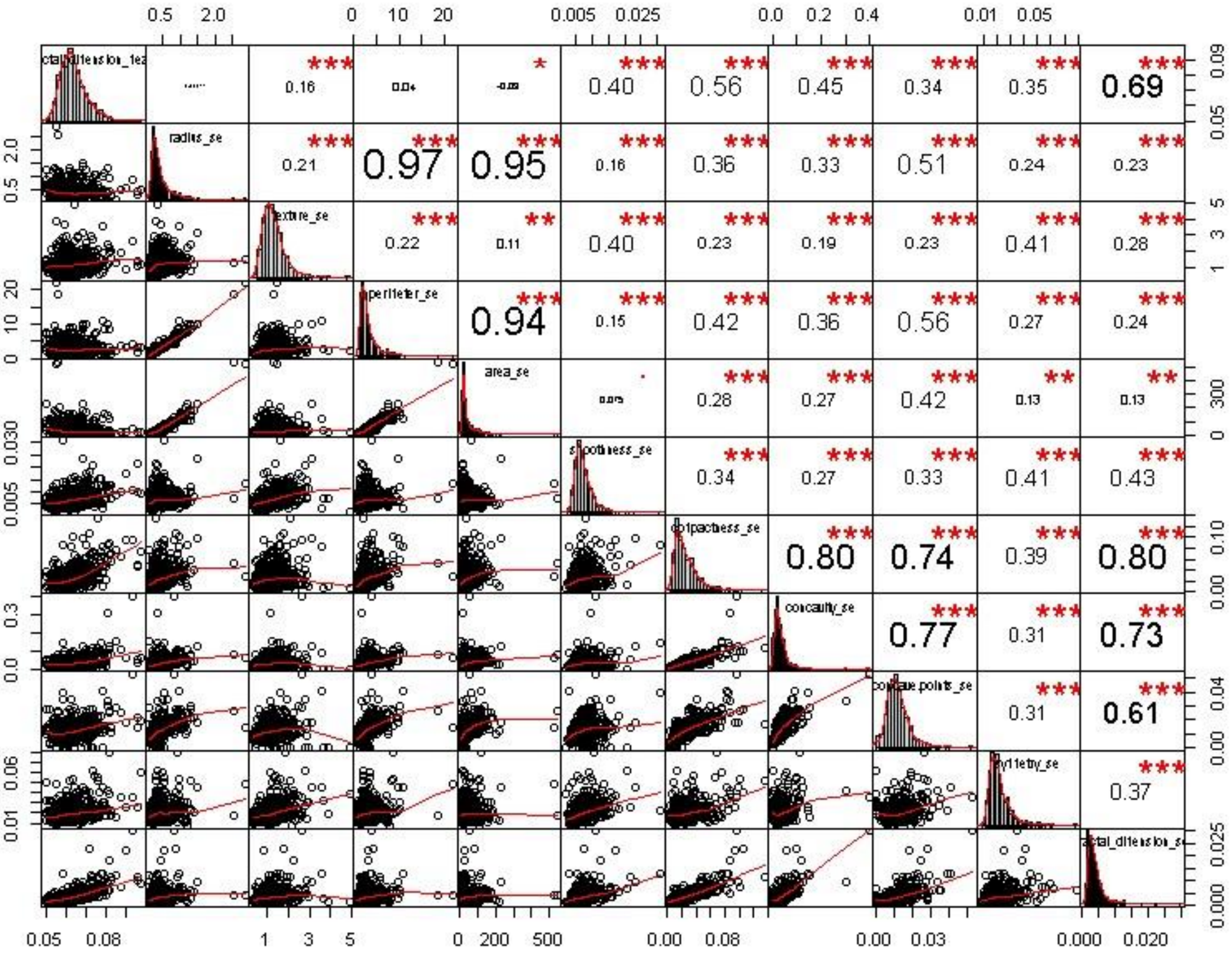
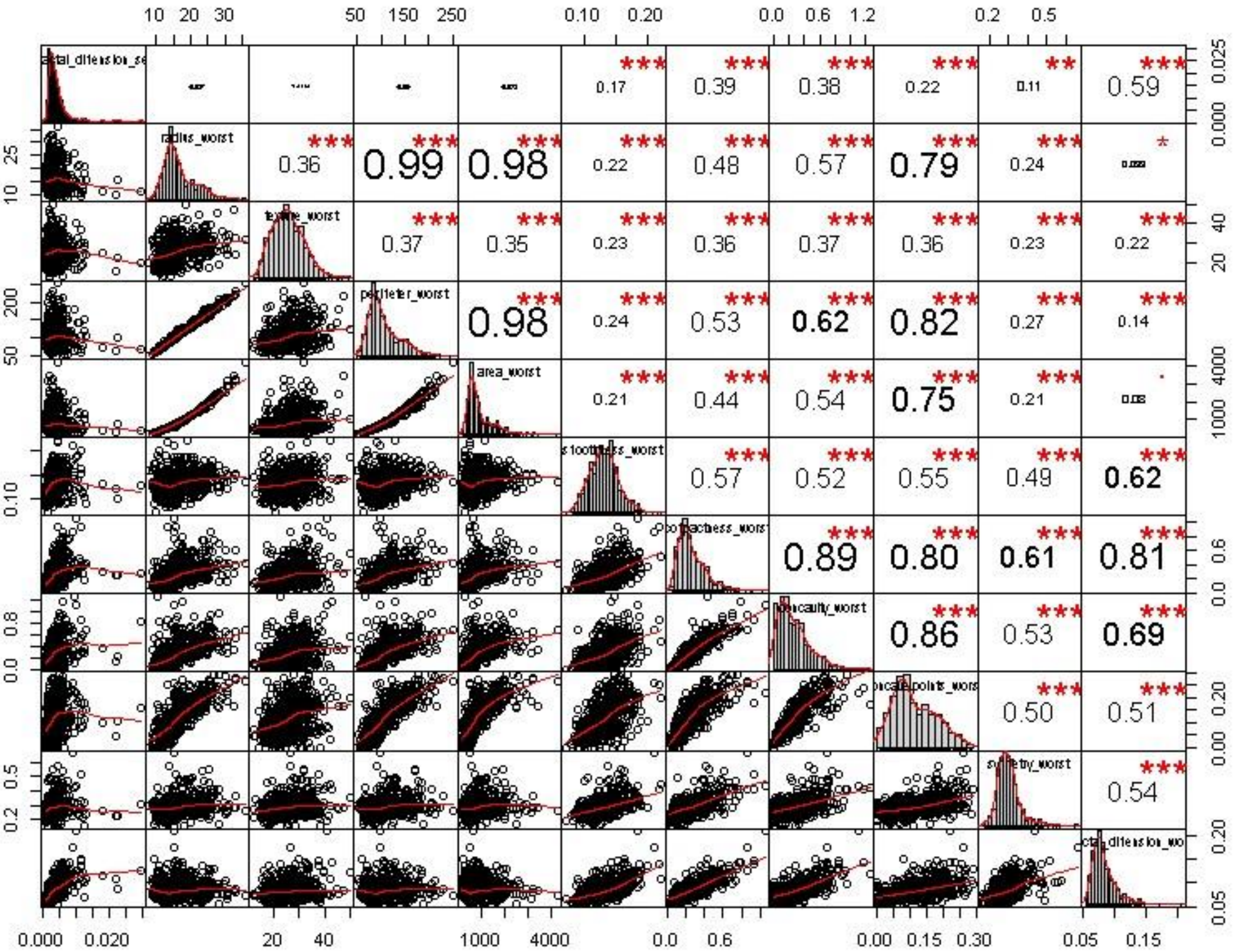


Figure 5 below shows that shows a matrix with 10 variables from the last group 10 variables with "worst" or largest (mean of the three largest values) There are five pairs of variables are strongly correlated with coefficients from 0.89 to 0.99. And one variable cancave_point_worst strongly correlated five variables. Another variable compactness_worst strongly correlated other three variables with coefficient 0.80 to 0.89.

Figure 5 correlation matrix with 10 variables of worst from 20 -30



Swarm plots

Figure 6 shows the variance of the two groups benign and malignant of the variables with mean. The blue dots represent the malignant, and the orange dots represent the benign. There are three variables, radius_mean, perimeter_mean, and area_mean of the two groups split more clearly than the others.

Figure 6 swarm plot of 10 variables with mean

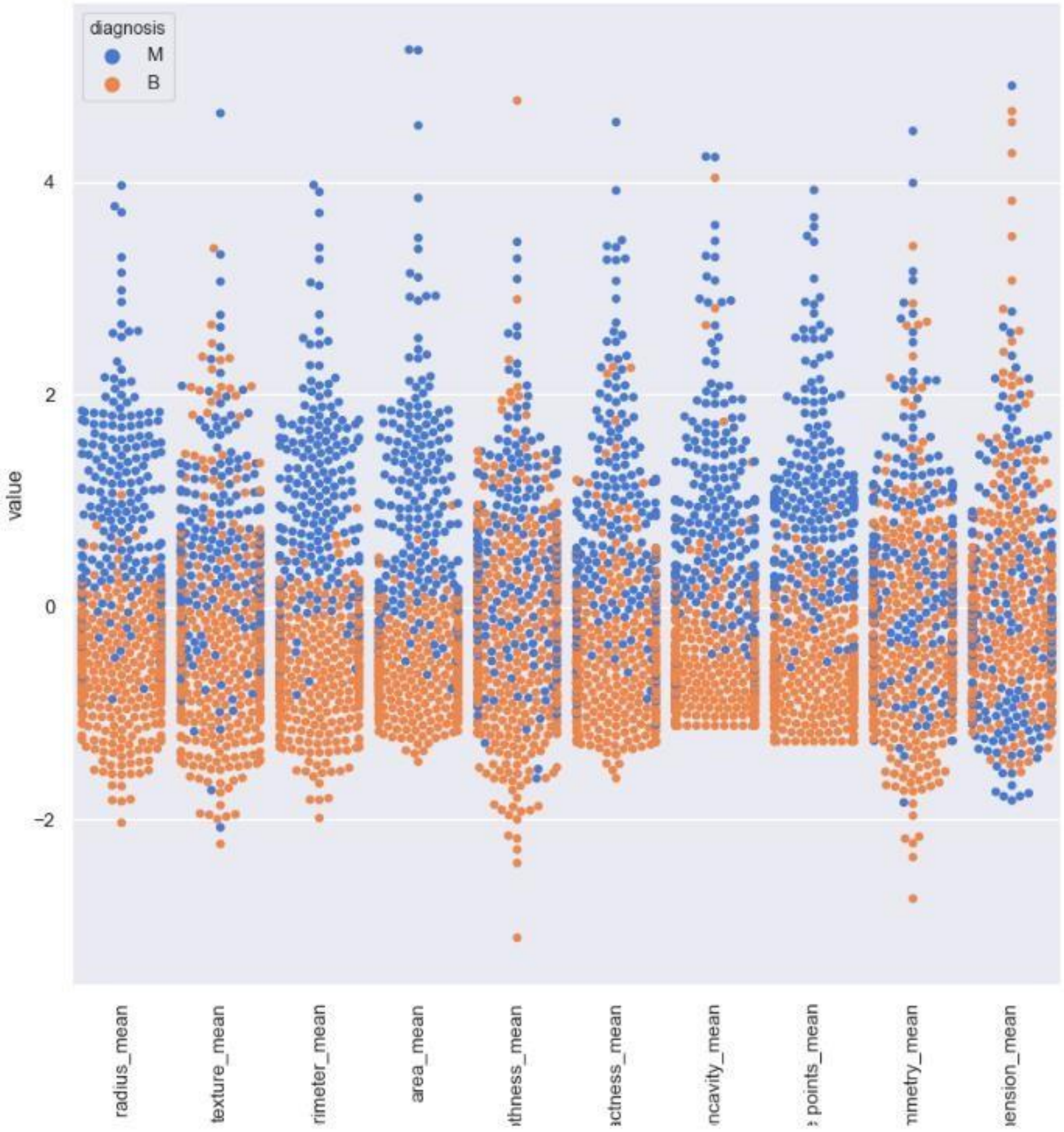


Figure 7 shows the distribution of 10 variables of standard error. The cancave_point_se variable presented a similar situation as the mean warn plot above, such as the area_mean clearly split by the two groups.

Figure 7 swarm plot of 10 variables of standard error

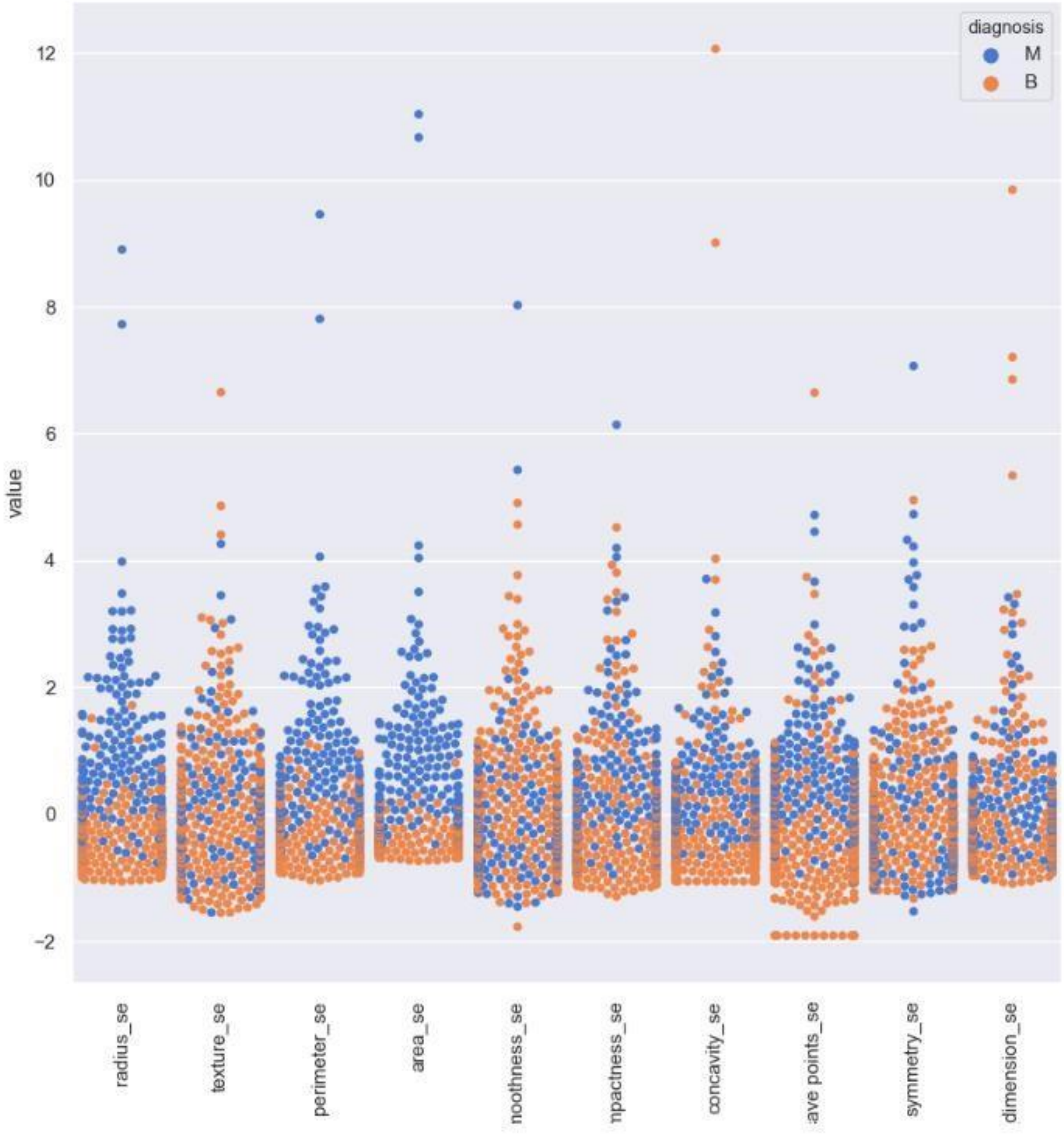
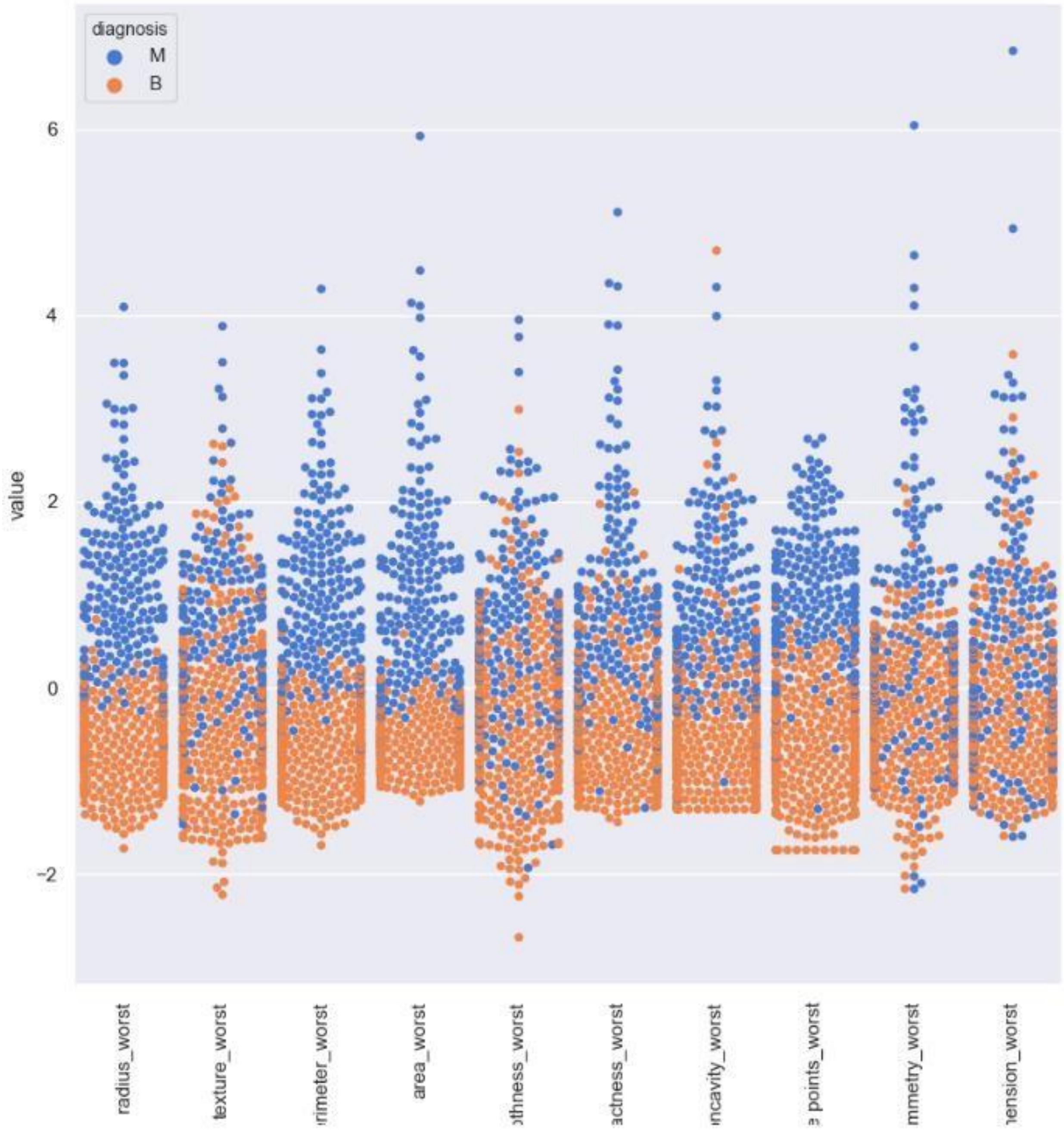


Figure 8 shows the distribution of 10 variables of worst and four variables, including radius_worst, perimeter_worst and area_worst, and concave_point_worst split very clearly by the two groups. Compared to the standard error group, the concave_point_worst changed the distribution from the not splitin standard error above in Figure 7 to split into the worst groups in Figure 8.

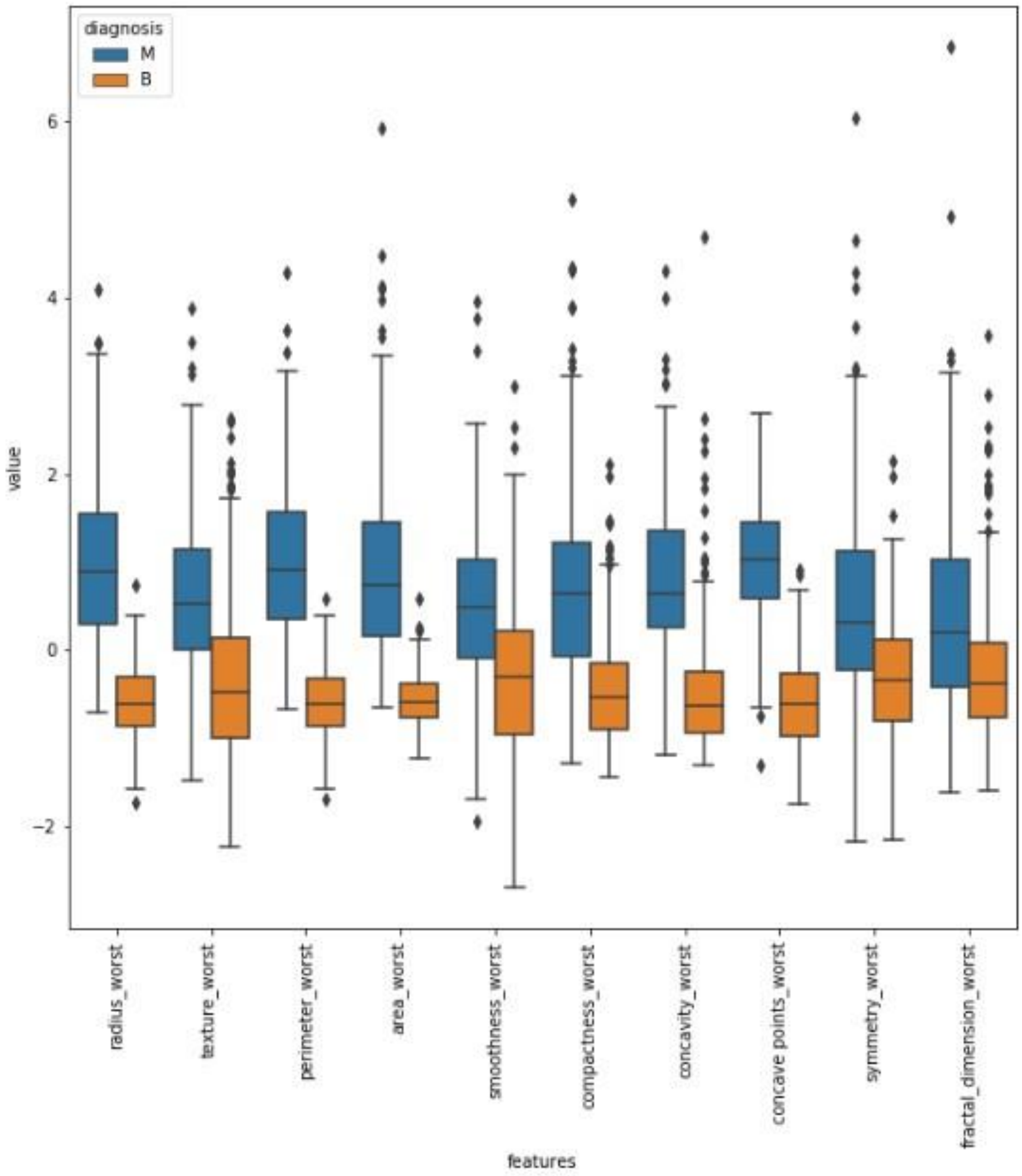
Figure 8 swarm plot of 10 variables of worst



Box plots

The box plot shows the majority of the data of the two groups are split vary clearly, the blue box and the orange box of each variable are separated.

Figure 9 box plot of the 10 variables 8of worst



Violin plots

Figure 10 shows the distribution of the 10 variables of the mean. The radius_worst, perimeter_worst and area_worst, and concave_point_worst variables were clearly split by the two groups.

Figure

10 violin plot of the variables of mean

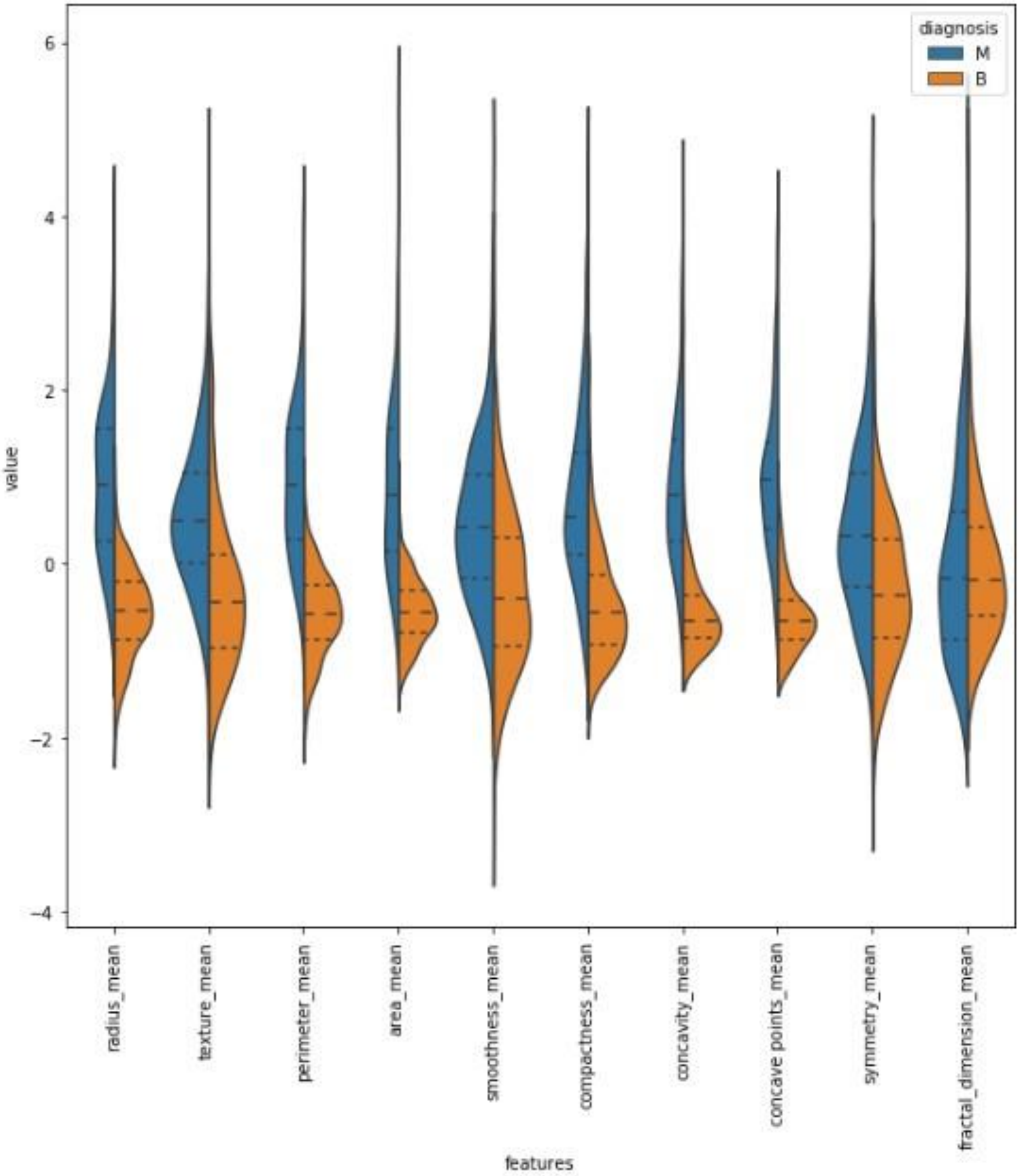


Figure 11 shows the distribution of 10 variables of standard error. The shapes of the violins are much less split than the mean groups.

11 violin plot of the 10 variables of standard error("se")

Figure

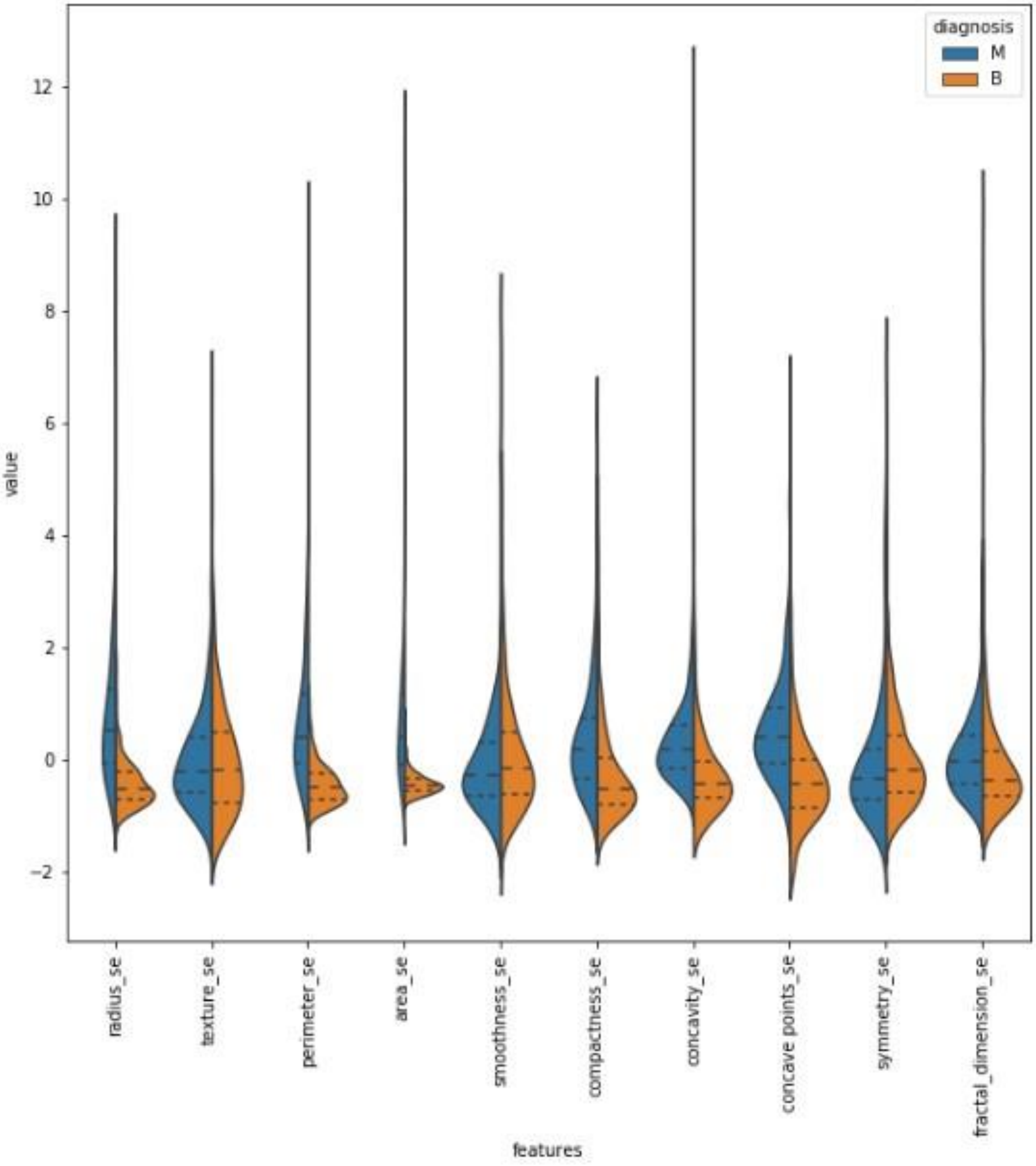
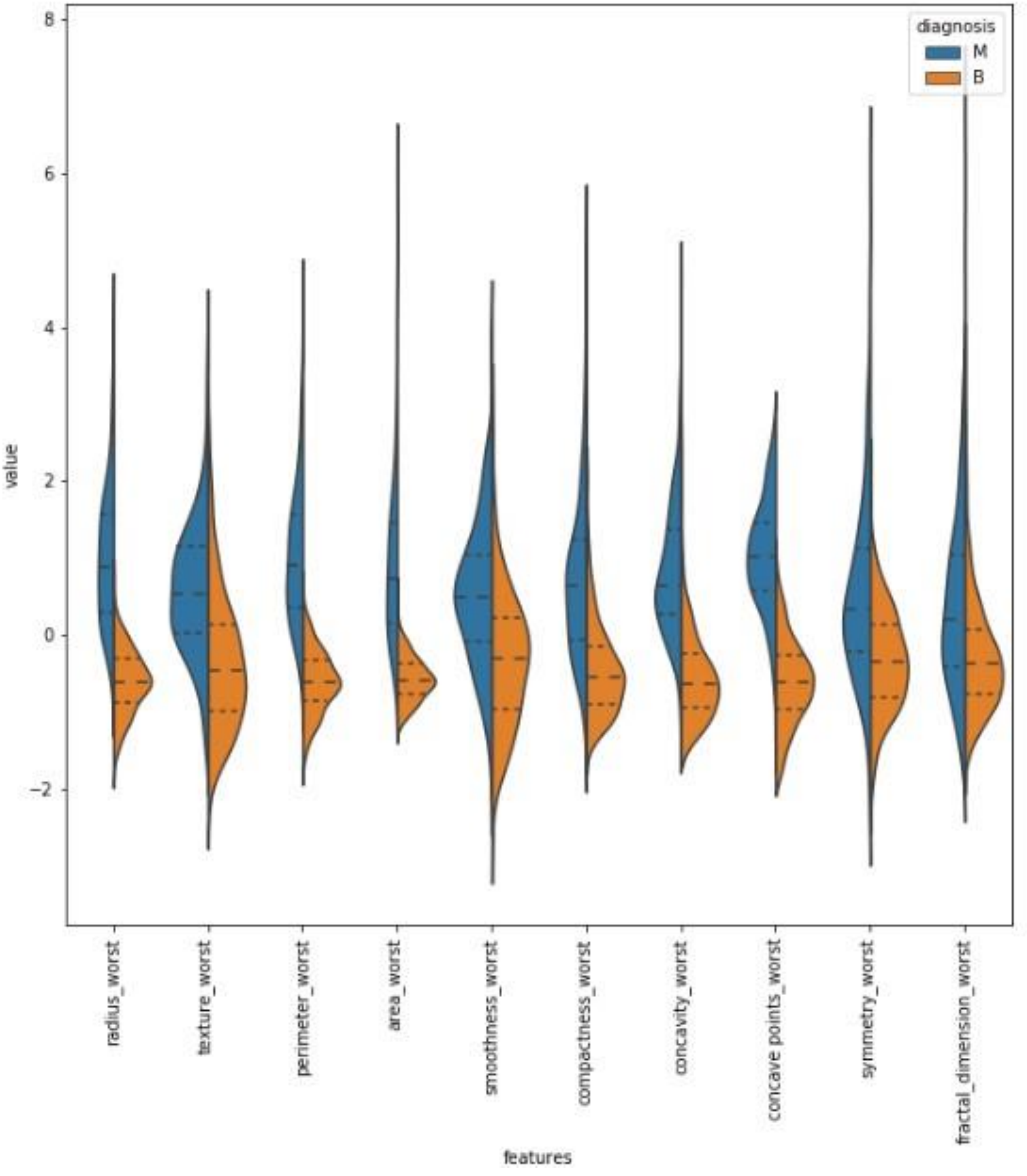


Figure 12 shows the distribution of the variables of worst and four variables, including radius_worst, perimeter_worst, and area_worst, and concave_point_worst split very clearly.

12 violin plot of 10 variables of worst

Figure

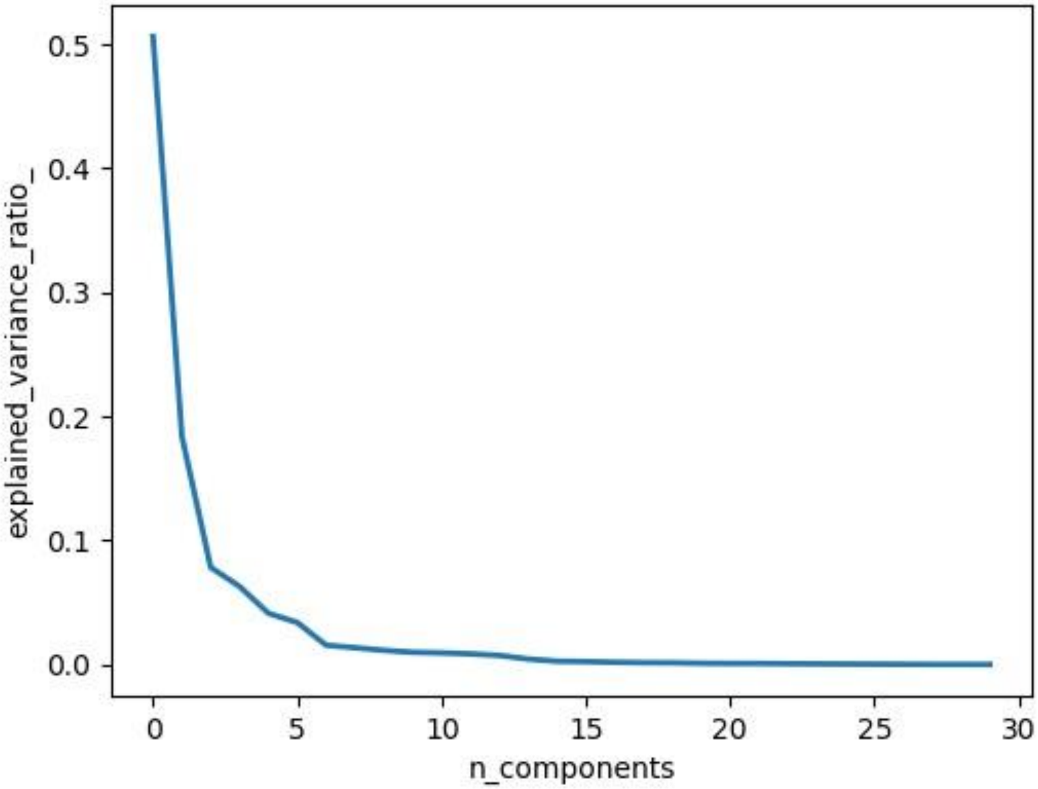


3.3 Tree-based feature selection and random forest classification

PCA

Dimensionality reduction is an important aspect of machine learning because it projects a highdimensional data into a lower dimensional space by keeping the inherent structure of the data. PCA was performed for features extraction, and the output was shown in Figure 13. Explained variance ratio is the percentage of variance being explained by each of the selected components. The curve in Figure 13 shows that Explained variance ratio reduces fast after the first three components. These means three components can be used to fulfill dimensionality reduction.

Figure 13 PCA feature extraction output



Rand Forest Classifier for feature ranking

A random forest is a meta estimator that fits a number of decision tree classifiers on various subsamples of the dataset and uses averaging to improve the predictive accuracy and control overfitting. The sub-sample size is always the same as the original input sample size, but the samples are drawn with a replacement if bootstrap=True (default).

Figure 14 feature ranking of 16 variables by tree- based method random forest classifier

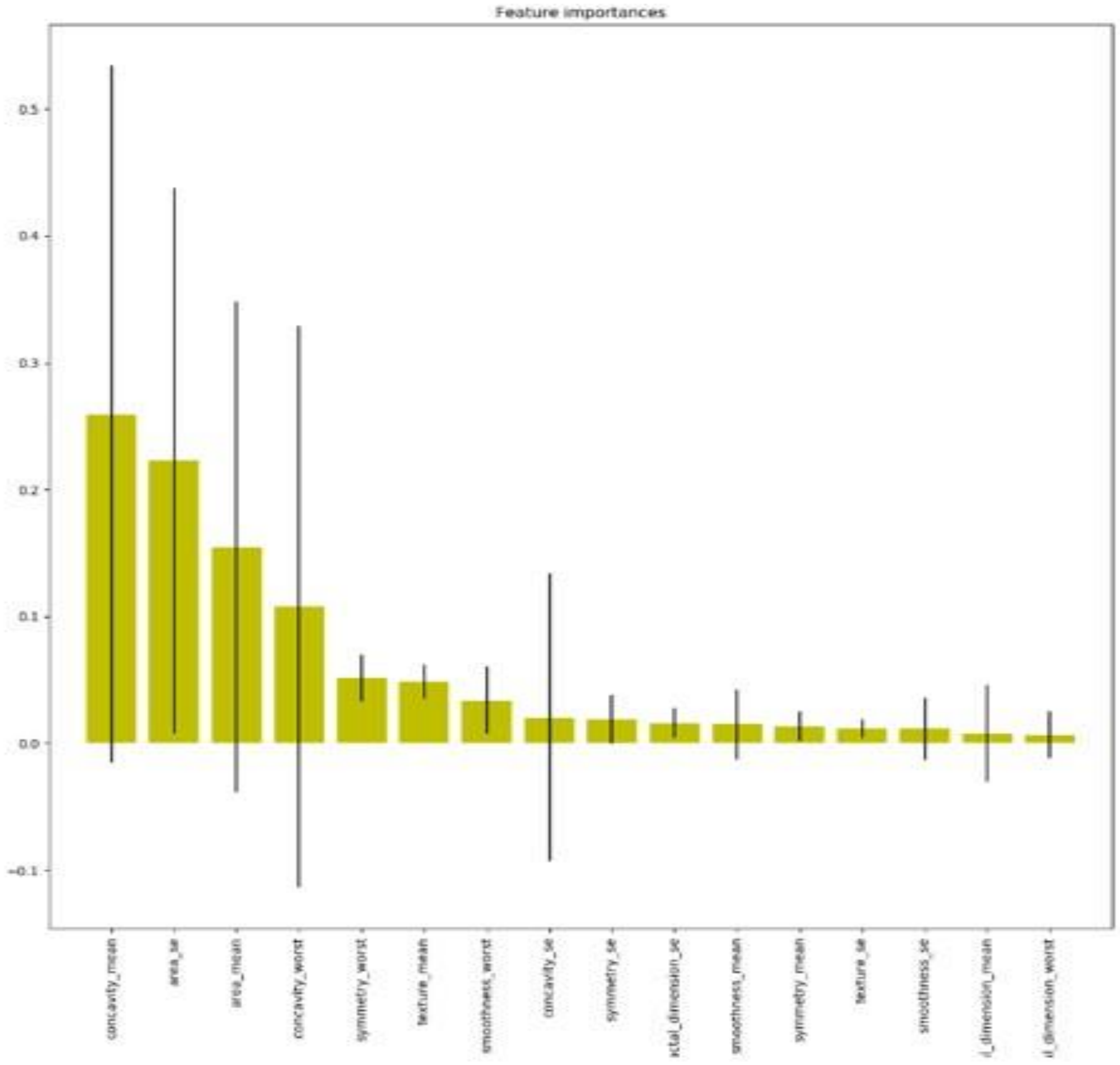


Figure 14 shows that after 5 best features the importance of features decreases. These 5 features will be used for RF classification.

Wrapper method for feature selection

Wrapper methods selected five features including 'concave points_mean’, 'radius_worst', 'texture_worst’, 'perimeter_worst’, 'concave points_worst’. These selected features are matched with the results of the visualization methods, including swarm plots, violin plots, and box plots plots. Only one feature was selected in strongly correlated pairs of variables. These means the Wrapper method worked well and provided a correct feature selection.

3.4 Classification

Five machine learning methods including Decision Tree (DT), Random Forest (RF), Gradient Boosting and Ada Boosting, Support Vector Machine (SVM) and Neural Network are performed on the WBCD with different parameters.

4 Results

The best results of each method are presented in Table 1. Below. The results show that RF and Neural Network with feature selection produced the best results in ACC and AUC scores with 0.97 and 0.99, but Neural Network with feature selection run time is about 4 times short than that of the RF. Compared the Neural Network with feature selection and without feature selection, the one with feature selection produced a better result than the other. the ACC scores increased from 0.91 to 0.97, and the AUC scores increased from 0.95 to 0.99. The Gradient boosting and Ada boosting produced slightly lower ACC and AUC scores, which is 0.96 and 0.99. The Decision tree method did not perform well.

Table 1 Results

	Decision Tree	random forest	Gradient Boosting	Ada Boosting	MLPClassifier with Feature Selection	MLPClassifier without Feature Selection	SVM
ACC	0.935	0.97	0.96	0.96	0.97	0.91	0.95
AUC	0.932	0.99	0.99	0.99	0.99	0.95	0.98
Run Time		1.58	0.708	1.045	0.38	1.68	0.12
	entropy, train/test: 0.65/0.35	Tree number N=100, cross validation: cv=5,	depth=3		activation = 'logistic', solver = 'lbfgs', alpha = 0.0001, max_iter=1000, layer_sizes=(10,), random_state=rand_st, cross_validate, cv=5	activation = 'logistic', solver = 'lbfgs', alpha = 0.0001, max_iter=1000, layer_sizes=(10,), random_state=rand_st, cross_validate, cv=5	

Figure 15 shows that the sensitivity is 0.96, that specificity is 0.98, that accuracy is 0.97, and that AUC is 0.99.

Figure 15 Confusion matrix by Neural Network with feature selection

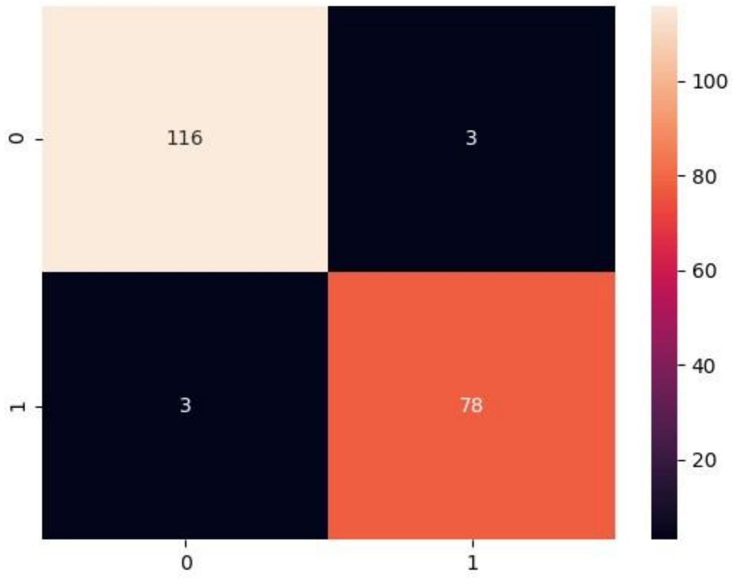


Figure 16 confusion matrix by RF with five features

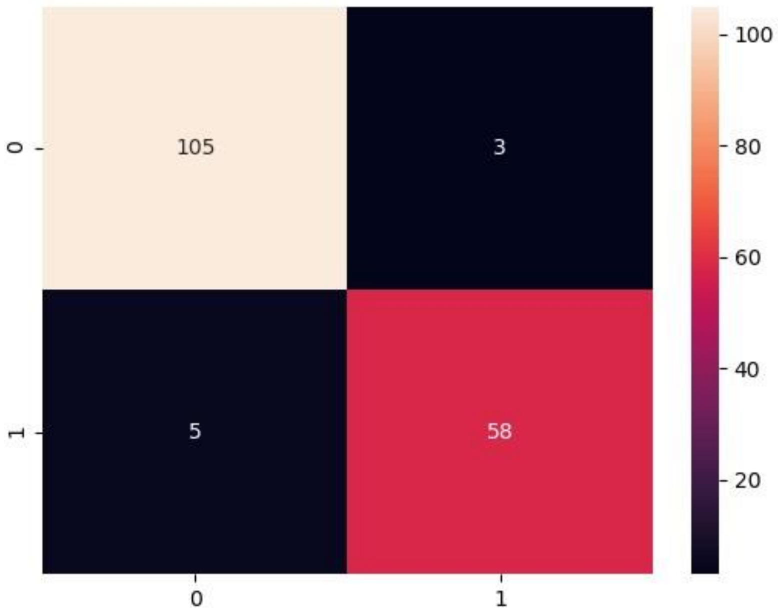


Figure 16 shows the result of the best five features, including area_mean, area_se, texture_mean, concavity_worst, and concavity_mean by Univariate feature selection and random forest classification. The sensitivity is 0.95, specificity is 0.96, accuracy is 0.95, and AUC is 0.99.

5. Discuss

The Wrapper methods selected the five features which are shown in the swarm plots, violin plots, and box lots very clearly. The Rand Forest and Gradient boosting, and Neural Network performed well on this data sets. But the running times are different, the Neural Network (NN) with feature selection is much short than methods which can produce the dame level of scores. The SVM run time is the shortest, but it did not produce higher scores. Decision tree methods did not perform well. From the results of the NN with feature selection and without feature selection, the feature selection one not only produced the better scores but also run much fast. This implies that when dealing with the data set with a lot of correlated variables, methods with feature selection may be more efficient and accurate. The correlation matrix of 30 variables gives a picture of how the dependent variables are correlated. The visualization plots give an ideal about how the variables of two groups, benign and malignant, are distributed, and the results matched with that of the machining learning methods.

6. Conclusion and future work

The visualization tools can help know the data well. The Neural Network with feature selection produced the accuracy 97% and AUC the score 99% with short run time. The 100% accuracy is the ideal model to pursue. In the future work, some probability methods such as Bayesian Network and some fuzzy methods will be tried to improve accuracy.

References

1. Nichols, H. (2017, February 23). "The top 10 leading causes of death in the United States." Medical News Today. Retrieved from https://www.medicalnewstoday.com/articles/282929.php.
2. © Springer Nature Singapore Pte Ltd. 2019133
N. K. Verma and A. K. Ghosh (eds.), Computational Intelligence: Theories, Applications and Future Directions—Volume I, Advances in Intelligent Systems and Computing 798, https://doi.org/10.1007/978-981-13-1132-1_11
3. West, D., Mangiameli, P., Rampal, R., West, V.: Ensemble strategies for a medical diagnostic decision support system: a breast cancer diagnosis application. Eur. J. OperationalRes.162(2), 532–551 (2005)
4. Yardimci, A.: Soft computing in medicine. Appl. Soft Comput. 9(3), 1029–1043 (2009)
5. Kala, R., Janghel, R.R., Tiwari, R., Shukla, A.: Diagnosis of breast cancer by modular evolutionary neural networks. Int. J. Biomed. Eng. Technol. 7(2), 194–211 (2011)
6. S. Jhajharia, H. K. Varshney, S. Verma and R. Kumar, "A neural network based breast cancer prognosis model with PCA processed features," *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI),* Jaipur, 2016, pp. 1896-1901. doi: 10.1109/ICACCI.2016.7732327
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7732327&isnumber=77320137
7. Shieu-Ming, C.,Tian-Shyug, L., Shao, Y. E. ,Chen, I. -F.:Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. Expert Syst. Appl.27(1), 133–142 (2004)
8. José,M.,Jerez-Aragonés,J.M.,Gómez-Ruiz,J.A.,Ramos-Jiménez,G.,Muñoz-Pérez,J.,AlbaConejo, E.: A combined neural network and decision trees model for prognosis of breast cancer relapse. Artif. Tell. Med. 27(1), 45–63 (2003)
9. S¸ahan, S., Polat, K., Kodaz, H., Güne¸s, S.: A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. Computers in Biology and Medicine 37(3), 415–423 (2007)
10. Pena-Reyes, C.A., Sipper, M.: A fuzzy-genetic approach to breast cancer diagnosis. Artif.Intell. Med. 17(2), 131–155 (1999)
11. Einipour, A.: A fuzzy-aco method for detect breast cancer. Global J. Health Sci. 3(2), 195 (2011)
12. Janghel, R.R., Shukla, A., Tiwari, R.: Hybrid computing based intelligent system for breastcancer diagnosis. Int. J. Biomed. Eng. Technol. 10(1), 1–18 (2012)
13. Delen,D.,Walker,G.,Kadam,A.:Predictingbreastcancersurvivability:acomparisonofthree data mining methods. Artif. Intell. Med. 34(2), 113–127 (2005)
14. Jain, R., Abraham, A.: A comparative study of fuzzy classification methods on breast cancer data. Australasian Phys. Eng. Sci. Med. 27(4), 213–218 (2004)

15. Übeyli, E.D.: Implementing automated diagnostic systems for breast cancer detection. Expert Syst. Appl. 33(4), 1054–1062 (2007
16. Janghel, R.R. Shukla, A., Tiwari, R., Kala, R.: Breast cancer diagnostic system using symbioticadaptive neuro-evolution (sane). In: Soft Computing and Pattern Recognition (SoCPaR), 2010 International Conference of, pp. 326–329. IEEE, 2010
17. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/home