A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications

Chen Le, Ruolin Liao,Ying Ji

06/06/2021

Abstract

Breast cancer, one of the diseases that represent a large number of incidence and mortality in the world, is the most common cancer in women both in developed and underdeveloped countries. The most important problem for cancer treatment is early prediction. Early detection by doctors based on clinical features can greatly increase the probability of successful treatment. The purpose of this paper is to compare the accuracy of different machine learning algorithms, combining with two different dimension deduction methods, in defining whether a tumor is benign or malignant in the diagnosis of breast cancer, to obtain a more effective classification model. Five machine learning algorithms were evaluated in this report, including Logistic Regression, K Nearest Neighbor (KNN), Naive Bayes, Random Forest, and Support Vector Machine (SVM). The simulation of the algorithms is done on the Wisconsin breast cancer dataset available in the UCI machine learning repository.

1 Introduction

In the past 30 years, machine learning has developed into a multi-field interdisciplinary subject, involving probability theory, statistics, approximation theory, convex analysis, computational complexity theory, and other subjects. It has been widely used in medicine, finance, and defense. Machine learning theory is mainly to design and analyze some algorithms that allow computers to automatically "learn". The machine learning algorithm is a kind of algorithm that automatically analyzes and obtains rules from data, and uses the rules to predict unknown data. The death rate from breast cancer has declined gradually since 1990. The decline in the death rate is due in part to earlier medical testing methods, better diagnostic techniques, and improved treatment. Breast cancer detection and diagnosis in the early stage can have a great effect on treatment protocols since the type of treatment procedures that need to be followed will be varying according to the type of cancer and its stage. A tumor is often a condition that misleads people into believing that they have cancer, but not all tumors are cancerous. Most breast cancer patients' early symptoms are not obvious, it is easy to be ignored and did not seek medical treatment in time. Therefore, high-risk groups should pay attention to the screening of breast cancer. Breast cancer may occur in two different categories like malignant and benign tumors, and benign tumors grow slowly or do not grow at all, while malignant tumors are cancerous. Accurate detection and classification are the most important processes in breast cancer treatment since treatment in the early stage can save the lives of many patients.

In our project, we aimed to make a comparative analysis using data visualization, classification, and machine learning applications for breast cancer detection and diagnosis on five machine learning techniques. We started with data cleaning and exploratory data analysis and data visualization, and then reduced the date dimension in two different ways: high correlation filter and principal component analysis. Since this is an unbalanced class data set, we split our data into a training data set and testing test data randomly after balancing the data set. Machine learning algorithms were applied based on the training data set and tested upon the testing data set. Finally, we finished off with an analysis of the model results and compared the strength of the model with accuracy measures. The rest of this report is structured as follows. Part 2 is the background of our project and our data set. Part 3 gives a theoretical presentation of the methods and machine learning algorithms we have used in this report. Part 4 gives the data visualization and the analysis of some important variables. The result of the comparison is represented in part 5 and finally a discussion in part 6.

2 Background

Breast cancer is often called the "pink killer", which is the most common cancer in a woman's lifetime. About one in eight women will develop breast cancer, mainly in women over 50 years of age. According to the latest data from the 2018 International Agency for Research on Cancer (IARC) survey, the incidence of breast cancer in women worldwide is 24.2%, ranking the first among women's cancers, 52.9% of which occur in developing countries. When breast cancer is

detected at an early stage, a cure is possible, but undetected at the early stage can lead to a fatal stage, even death. Breast cancer is the phenomenon of uncontrolled proliferation of mammary epithelial cells under the action of a variety of carcinogenic factors. The early stage of the disease is often manifested as a breast mass, nipple discharge, axillary lymph node enlargement, and other symptoms, and the late stage can be caused by distant metastasis of cancer cells, multiple organ lesions, which directly threaten the life of the patient. The etiology of breast cancer is not clear, so far scientists have not found the exact cause of breast cancer, but have found a lot of high-risk factors related to the incidence of breast cancer. For example, genetic factors are a high-risk factor for breast cancer. A history of breast cancer among first-degree relatives (such as parents, children, and siblings) is two to three times the risk of developing breast cancer in the general population. Some genetic mutations also increase the risk of breast cancer. In addition, certain physical factors, such as childhood exposure to chest radiation, are also risk factors for breast cancer. As the risk factors for breast cancer accumulate, the risk increases. In addition to the above risk factors, there are some lifestyle and the incidence of breast cancer have a certain relationship, such as age, race, and genes, even exercise level, alcohol consumption, and lifestyle habits. Most of them, though, do not directly cause breast cancer. Prompt detection and treatment of breast cancer patients can help reduce the risk of mortality and complications and slow the progression of the disease. The survival of patients with breast tumors varies greatly. Staging is the most important independent prognosis factor to determine tumor recurrence and patient survival, and the 5-year survival rate of different staging patients is different. Whether the tumor can be completely removed is another important factor affecting the prognosis. In the past 20 years, thymoma diagnosis and treatment have improved significantly. The overall recurrence rate continues to decline, and the survival rate of patients, including thymus cancer, has shown an upward trend.

Benign tumors are non-dangerous tumors, they have well-defined contours. They develop slowly in the organ where they appeared without producing metastatic cases. Benign tumors are composed of cells that resemble normal cells of the breast tissue. Malignant tumors are dangerous tumors because they spread to other organs of the body and can produce metastatic cases. Cancer cells of malignant tumors have several abnormalities compared with normal cells in shape, size, and contours where cells lose their original characteristics. In this project, we used the Wisconsin Breast Cancer dataset was obtained from the UCI Machine Learning Repository. The dataset contains information of 357 benign and 212 malignant tumors, with the radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension of each cell nucleus. The mean, standard error, and largest value of the above-mentioned features were computed and it results in 30 features. These features were extracted from digital images of fine-needle aspirates of breast lumps that describe the core features of the current image.

3 Methodology

3.1 Upsampling

Upsampling in data analysis are techniques used to adjust the class distribution of a data set (i.e. the ratio between the different classes/categories represented). These terms are used both in statistical sampling, survey design methodology, and machine learning. It can be defined as adding more copies of the minority class, which is a good choice when we don't have a ton of data to work with. Upsampling always is done after splitting into test and train data sets.

3.2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a multivariate statistical method to investigate the correlation between multiple variables. A set of variables that may be correlated are converted into a set of linearly uncorrelated variables by orthogonal transformation, and the converted variables are called principal components. It studies how to reveal the internal structure of multiple variables through a few principal components, that is, to derive a few principal components from the

original variables, so that they retain the information of the original variables as much as possible, and are not correlated with each other. Note that each principal component is orthogonal, which can eliminate the interaction between original data components.

3.3 High Correlation Filter

High correlation filtering assumes that when two columns of data have similar trends, the information contained in them will also be displayed. In this way, the machine learning model can be satisfied by using one of the similar columns. The similarity between numerical columns can be expressed by calculating the correlation coefficient, and the correlation coefficient between noun columns can be expressed by calculating the Pearson Car-square value.

3.4 Balanced Accuracy

In real life, sometimes the data sets are uneven, such as the prediction of rare diseases. Assuming that the sample ratio of the positive category is hundreds of times that of the negative category, an exaggerated assessment may be given if the performance of the model is measured with conventional accuracy, which can be shown as high accuracy of both the positive category and the negative category in the training set. However, when the model is predicated on the test set, only the positive category has high accuracy. The negative category has low accuracy. The balanced accuracy in binary and multi-class classification problems to deal with imbalanced data sets. Balanced accuracy is based on sensitivity and specificity, can be computed from the confusion matrix as follows

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{FP + TN}$$

Where TP is the positive rate; FP is the false positive rate; TN is the true negative rate; FN is the false negative rate. Balanced accuracy is simply the arithmetic mean of the two:

$$Balanced \quad accuracy = \frac{Sensitivity + Specificity}{2}$$

3.5 Logistic Regression

This is a technique that can be used for traditional statistics as well as machine learning, which has time complexity $O(nd)$.Logistics regression is similar to linear regression except , logistic regression predicts something *true* or *false*, instead of predicting something continuous like size. Instead of fitting a line to the data, logistic regression fits an s-shaped logistic function, the curve goes from 0 to 1, which tells us the probability of y based on x. Also, it's usually used for classification. If the probability of y is great than 50%, then we classify it as 1.

3.6 K Nearest Neighbor (KNN)

KNN is a supervised learning technique that means the label of the data is identified before making predictions. Starting with a dataset with known categories. Then cluster that data by PCA or else. Adding a new cell, with an unknown category. Then we classify the new cell by looking at the nearest neighbors. The KNN has a time complexity of $O(knd)$. By the nature of the KNN method, we need to fit the data multiple times and select the best k. k represents a numerical value for the nearest neighbors. Predictions are made based on the Euclidean distance to k-nearest neighbors. Low values for K can be noisy and subject to the effects of outliers. Large values for K smooth over things, but we do not want k to be so large that a category with only a few samples in it will always be outvoted by other categories.

3.7 Naive Bayes

The Naive Bayes method is a classification method based on the Bayes theorem and independent assumption of characteristic conditions.6 Naive Bayes, originated from classical mathematics theory, has a solid mathematical foundation and stable classification efficiency. At the same time, the Naive Bayes model requires few parameters to estimate and is not sensitive to missing data, and the algorithm is relatively simple. In theory, the Naive Bayes model has the smallest error rate compared with other classification methods. However, in fact, this is not always the case, because the Naive Bayes model assumes that the attributes are independent of each other, which is often not valid in practical application, which has a certain impact on the correct classification of the Naive Bayes model.

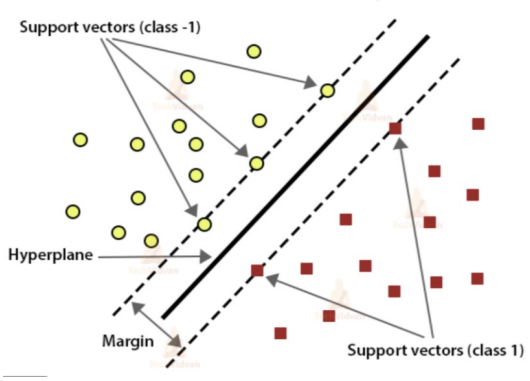
3.8 Random Forest

There are N cases in the training set. Then, samples of these N cases are taken at random with replacement. If there are m variables, a number m is specified such that at each node, m variables are selected at random out of N. The best split on these m is used to split the node. The value of m is held constant while we grow the forest. Each tree is grown to the largest extent possible.

When a new input is entered into the system, it is run down all of the trees. Finally, it predicts by taking the average of the output from various trees. Random forest classifier solves overfitting which is one of the biggest problems in machine learning. If there are enough trees in the forest, the classifier won't overfit the model.

3.9 Support Vector Machine (SVM)

Support vector machines (SVM) are supervised learning models that can be used for prediction as well as for the classification of linear and nonlinear data. Training complexity of nonlinear SVM is generally between $O(n^2)$ and $O(n^3)$ with n the amount of training instances. The principle of support vector machine algorithm is to use nonlinear mapping to transform the original learning data into larger dimensions. The objective SVM is to find an optimal hyper-plane in our n-dimensions that classifies the data points. The linear and RBF kernel are usually be used to optimize. N dimension diversifies based on the feature numbers. The plot of comparing two features is shown as follows.

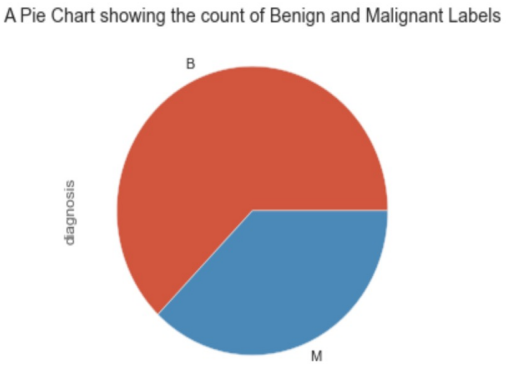


4 Visualization and Analysis

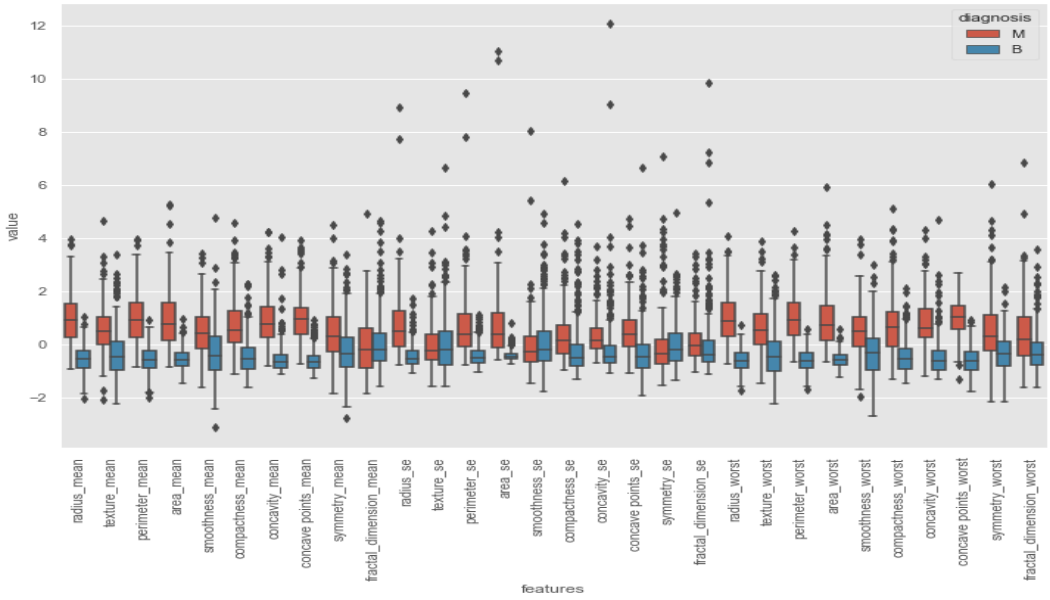
There are 33 columns in our data, but not all of them carry the useful information of classification. Before we doing the data visualization and machine learning algorithms. Since the first column, ID number, has no information on the tumor classification and the last column, Unnamed 33, is a whole column of missing value, so we removed them and 31 columns left. After removing the

useless columns and checking the miss values, finally there are 569 observations of 31 variables in our data. Note that these variables are in a large different scale, so we standardized them to make coefficients more comparable. Next we did the data visualization to analysis the relationship between the variables.

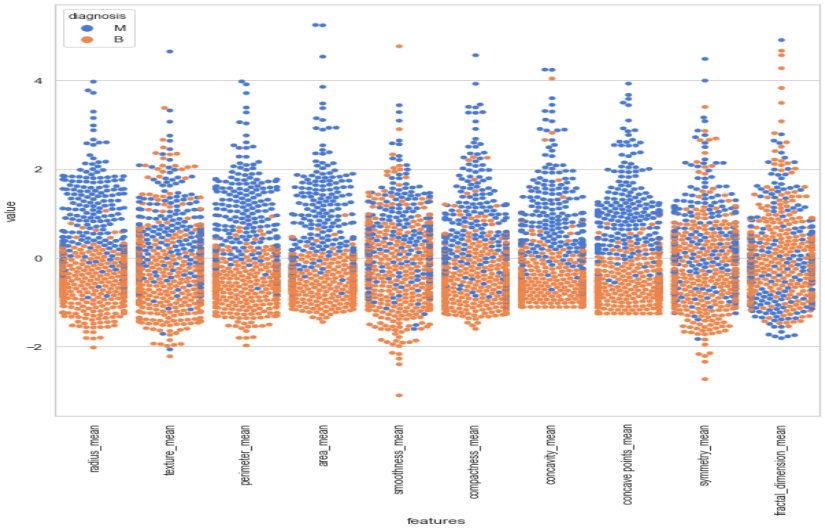
Then we made the pie chart to see the proportion of two types of tumors: 62.74% of the observations are benign and 32.76% are malignant, which means that it is an unbalanced class data, so we need to balanced our data set before applying the machine learning algorithms. Here we used the upsampling, which increases the sample of malignant tumors to the number that same as the benign tumors in the training dataset, also we used the balanced accuracy on the test dataset to remove the bias of unbalanced class.



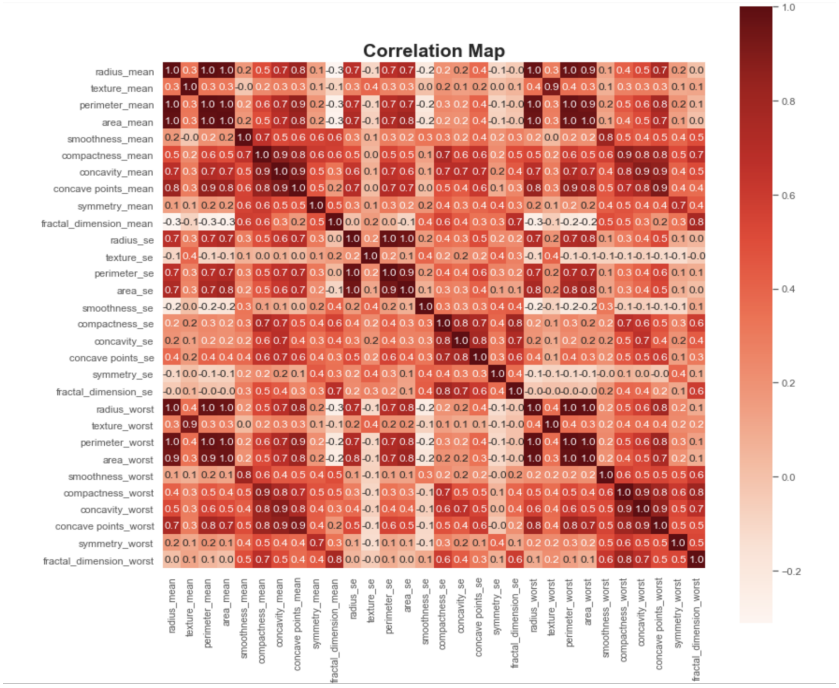
We draw a boxplot of all the variables based on the type of diagnosis below. We can see that there are many outliers in this data, where malignant tumors have more outliers than benign tumors. This makes sense, because malignant tumor cells vary in size and shape, and are usually larger and faster growing than their source cells, with a significantly thinner nucleo-cytoplasmic ratio than normal. The nuclear morphology is different and may appear megakaryocytic, dikaryotic, or multinuclear phenomena. Since the malignant tumor cells behave abnormally and uncontrollably, so it is more likely to have outliers. Even though the outliers carry some information of the classification, we still removed it because we have some methods are sensitive with the outliers such as the PCA. Besides these, we can see that the distributions of two different types with the same variable are different.



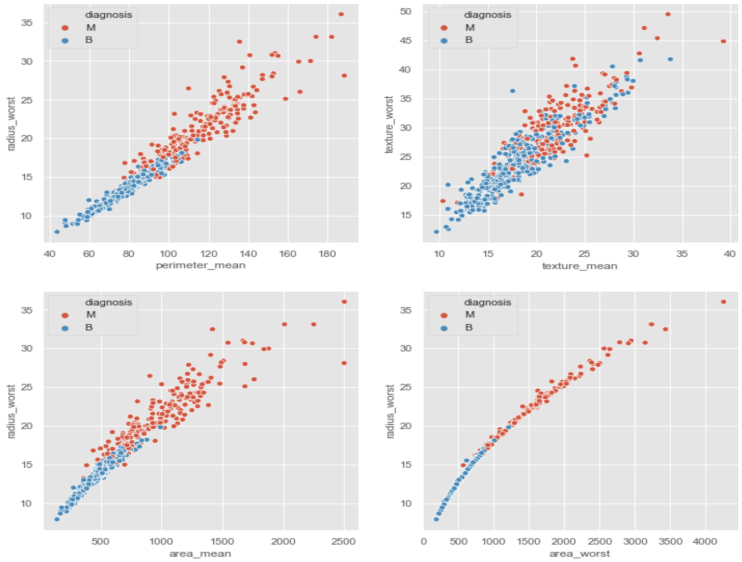
We generated the swarm plot using the subset of data which is the computed mean value of all the features since we believed the mean values are more representative than the standard deviation and the largest value of the features in this case. We can see some of the points have piled up against the edges of each column. The points of some variables are mixed, however, some others have been distinctly divided. For example, radius, perimeter, area, compactness, and concavity are significantly different based on the diagnosis. Rather others are not. This has corresponded to the distribution shown in the boxplot. We can observed that there exist large difference in the variables including *radius_mean*, *perimeter_mean*, *area_mean*, *compactness_mean* and *concavity_mean*. We claimed that variables with distinct differences will have a more significant effect on our classification. In general, malignant tumors are larger, smoother, more compact, more concave, and more symmetric.



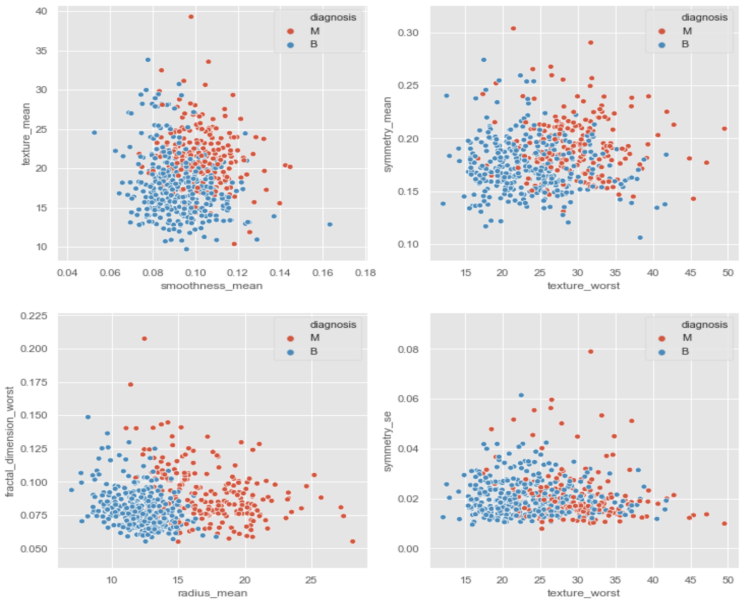
Additionally, we investigated the correlation between our variables. We used a heat map to describe the correlations between each predictor variable. The deeper the color, the higher the correlation. We discussed some correlation specifically later.



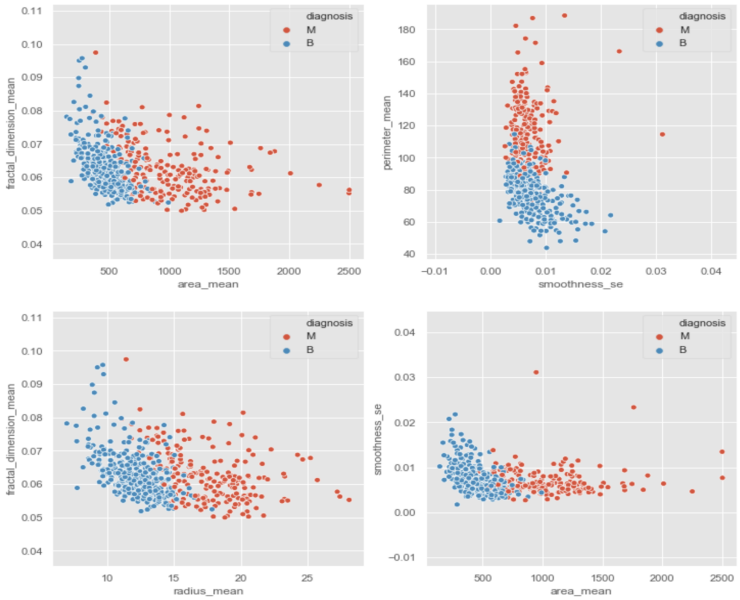
From the heat mat, we can see there are some highly correlated feature. We first drew four scatter plots where the correlation between features is positive and high. The plot on the right bottom shows the highest correlation between variables, while the plot on the top-right displays the lowest relationship among two variables. Overall, all four plots show a high positive correlation. Observations became increasingly correlated as nodes became closer, which shows multicollinearity. Since the logistic regression, KNN and Naive Bayes are sensitive to multicollinearity, which may decrease the performance of these models, so we did the dimension deduction to avoid over-fitting.



Next, we drew four scatter plots of the correlation between features is weak and irregular, which means there was no correlation between features and did not affect each other.



Lastly, we drew the scatter plot of correlation between features is negative, where radius mean and fractal dimension mean had the highest negative correlation, and smoothness standard error and area mean had the lowest negative correlation .



5 Result and Conclusion

Finally, we used five machine learning algorithms to do the classification and then compared with the balanced accuracy by using the proposed dimension deduction methods. We compare three different dimensions of the data set, which are denoted as the original data set, the PCA data set and the correlation filter data set. Since the more the number of features, the harder it gets to visualize the training set and then work on it, also in order to avoid overfitting, we did dimension deduction. To reduce the number of variables, we used two major methods, PCA and high correlation filter. For PCA method, we find the first 6 principal components are enough because they occupy 91% of the total variance. Each principal component is orthogonal, which can eliminate the interaction between original data components. For the high correlation filter method, we removed variables that have correlation coefficients above 0.8 based on common sense. A pair of variables with a high correlation increases the multicollinearity of the data set because they have the similar contribution, so it is necessary to delete them in this way.

Deduction Type	L.R	KNN	N.B	R.F	SVM(linear)	SVM(rbf)
Original	0.9705	0.9705	0.9414	0.9522	0.9473	0.9473
PCA	0.9815	0.9667	0.9277	0.9747	0.9815	0.9743
Correlation	0.9670	0.9605	0.9270	0.9737	0.9870	0.9737

Table 1: The Balanced Accuracy of Each Algorithm

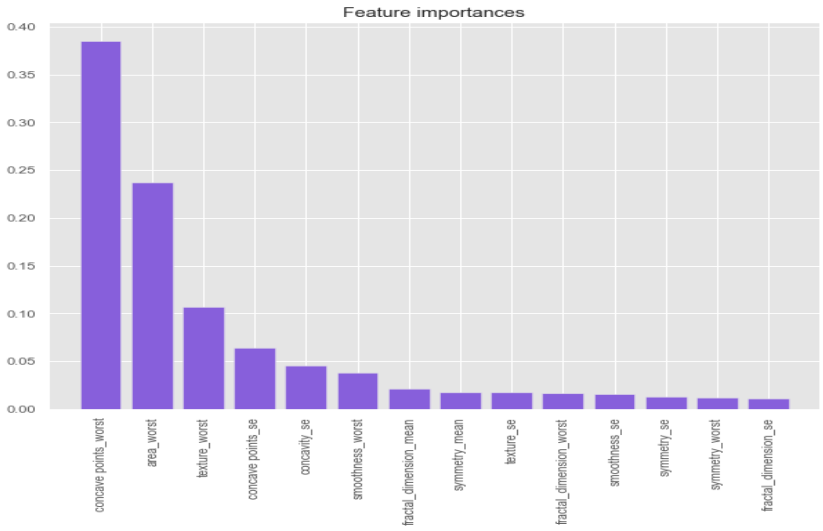
From the table above, we summarized the balanced accuracy of each machine learning algorithm with different dimension deductions by using the K-fold cross-validation test method with K = 10. In the appendix, we have all the results. Since the scoring times do not have a large difference, so we mainly compared the accuracy. We can see that all the balanced accuracy of the original data is higher than 0.94, which means there exists overfitting. And the SVM in the original data does not perform as well as in the other two data, that is because the SVM algorithm is sensitive to multicollinearity. When we used PCA and correlation filter to reduce the dimension, we can see that all the balanced accuracy of the SVM has improved a lot. The Naive Bayes algorithm has the worst results because it assumes that features or predictors are independent which is not true in this case. The results of PCA data and the correlation filter data are similar except the logistic regression, where the PCA data of Logistic Regression has

a higher balanced accuracy. So using PCA to reduce the dimension gives Logistic Regression a better performance. The reason may be the correlation filter method had removed half of the correlation so the data may lose some important information, while the first six principal components can explain 91% of the original data and contain more information for classification. All in all, the balanced accuracy of each algorithm is high. And the highest one is when we use the SVM with kernel linear in the correlation filter data. The second one, we can choose to use Logistic Regression or SVM with linear kernel in PCA data. Also, the SVM algorithm and the Logistic Regression are better algorithms to be used in this data.

6 Discussion

Some of algorithms might be improved. For example, the feature importance plays an important role in dimension reduction and feature selection, which can improve the efficiency and effectiveness. For example, in random forest method, even in a highly interpretive decision tree model, if the tree is too large, it is difficult for us to explain the results it makes. Random forests are usually made up of hundreds of trees, making them more difficult to explain. Fortunately, it is more important that we find those characteristics to assist us in interpreting the model. More importantly, unimportant features can be removed to reduce noise. It is more easy to understand than the result of dimension reduction by using PCA.

From below plot, the scores suggest that the model found the three important features concave points_worst,area_worst, texture_ worst with feature importance over 0.1 among 14 variables, which make a huge contribution to classification.



Removing the noisy features will help with memory, computational cost and the accuracy of our model. Also, it may help avoid the overfitting. In our analysis, we analyzed the accuracy without removing the variables that the feature importance are less than 0.1, which may have some adverse effects on our accuracy. For another example, for Logistic Regression, we may use stepwise regression removes and adds terms to the model for the purpose of identifying a useful subset of the terms. This might help us to find the significant variables during the classification. Minitab stops when all variables not in the model have p-values that are greater than the specified alpha-to-enter value and when all variables in the model have p-values that are less than or equal to the specified alpha-to-remove value.

In this report, we simply compared five different machine learning classification algorithms on three different dimensions of the data. Since too many variables contained in the model are not convenient for people to collect and record. Also, it has a higher computation complexity. To

reduce the number of variables, we used PCA and high correlation filter. However, there are certain advantages or limitations of both these methods of course. For example, in the high correlation filter method, we decided to remove variables that have correlation coefficients above 0.8 based on common sense. But the limitation is 0.8, which may not be the best benchmark of high correlation in this case. The choice of the value is kind of subjective.

Moreover, a commonly used rule says that a data point is an outlier if it is more than 1.5IQR above the third quartile or below the first quartile. we used this method to identify outliers and replacing them by upper bound and lower bound of our data. However, outliers also carry some useful information of the classification. And when we did the dimension deduction, some outliers also can be removed. So there may be some negative effect on our model after removing these points.

The incidence rate of breast cancer is increasing year by year. The screening of breast cancer can detect diseases as soon as possible. If breast cancer is detected early and timely taken measures for treatment, it can play a better control role. Early breast cancer can be treated by surgery, which has a relatively good effect and can effectively inhibit the metastasis and spread of cancer cells. Finding an efficient and accurate way to test could help doctors save patients' lives. In future for getting more than 99% accuracy, the optimization techniques are planned to use.

7 Reference

Ak, Muhammet Fatih. *A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications.* Multidisciplinary Digital Publishing Institute, 2020.

Saoud, Hajar and Ghadi, Abderrahim and Ghailani, Mohamed and Abdelhakim, Boudhir Anouar. *Application of data mining classification algorithms for breast cancer diagnosis.* Proceedings of the 3rd International Conference on Smart City Applications, 2018.

Abderrahim Ghadi *Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data.* inbook,2018

Habib Dhahri , Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, and Mohammed Faisal Nagi *Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms.* Volume 2019

Kalaiyarasi, M and Dhanasekar, R and Ram, S Sakthiya and Vaishnavi, P. *Classification of Benign or Malignant Tumor Using Machine Learning.* IOP Conference Series: Materials Science and Engineering, 2020

8 Appendix

	Model	Fitting time	Scoring time	Accuracy	Balanced accuracy	Precision	Recall	F1_score
4	Support Vector Machine_linear	0.017430	0.006370	0.975590	0.974288	0.976352	0.975616	0.975578
5	Support Vector Machine_radial	0.021640	0.005950	0.973835	0.974288	0.974677	0.973830	0.973820
0	Logistic Regression	0.011073	0.006284	0.970327	0.970531	0.971668	0.970259	0.970298
3	K-Nearest Neighbors	0.002522	0.006655	0.959891	0.970531	0.961244	0.959914	0.959860
1	Random forests	0.026830	0.005807	0.966848	0.952174	0.967390	0.966810	0.966840
2	Naive Bayes	0.004364	0.006120	0.938990	0.941363	0.941817	0.938978	0.938890

Comparison table of original data

	Model	Fitting time	Scoring time	Accuracy	Balanced accuracy	Precision	Recall	F1_score
0	Logistic Regression	0.007008	0.004582	0.965024	0.981494	0.965904	0.961399	0.964743
4	Support Vector Machine_linear	0.023603	0.006150	0.960628	0.981494	0.961438	0.956734	0.960454
1	Random forests	0.007873	0.003441	0.955894	0.974288	0.954843	0.951148	0.955693
5	Support Vector Machine_radial	0.016941	0.008094	0.958406	0.974288	0.964857	0.947401	0.957769
3	K-Nearest Neighbors	0.000860	0.004369	0.962609	0.966667	0.967283	0.954132	0.962034
2	Naive Bayes	0.001506	0.004422	0.925024	0.927692	0.926514	0.913377	0.924550

Comparison table of PCA data

	Model	Fitting time	Scoring time	Accuracy	Balanced accuracy	Precision	Recall	F1_score
4	Support Vector Machine_linear	0.017650	0.006234	0.975125	0.987013	0.976252	0.975185	0.975106
1	Random forests	0.023820	0.006035	0.973371	0.973684	0.974418	0.973399	0.973338
5	Support Vector Machine_radial	0.024054	0.006353	0.969799	0.973684	0.970528	0.969828	0.969784
0	Logistic Regression	0.009056	0.010207	0.955608	0.967006	0.957324	0.955665	0.955562
3	K-Nearest Neighbors	0.002827	0.010411	0.948622	0.960470	0.951014	0.948768	0.948544
2	Naive Bayes	0.004363	0.010374	0.930827	0.926992	0.933127	0.930727	0.930722

Comparison table of high correlation filter data