Data Set

Breast Cancer Wisconsin (Diagnostic) -- Predict whether the cancer is benign or malignant

Rationale and objectives of the study

1) Goal

The goal of this project is to predict whether the breast cancer is benign or malignant.

2) Public health concern for women worldwide

Breast cancer is the most common cancer among women and one of the major causes of death among women worldwide. One in eight women in the United States will be diagnosed with breast cancer in her lifetime. Worldwide, it is estimated that 124 out of 100,000 women are diagnosed with breast cancer, and approximately 23 out of the 124 women will die of this disease. On average, every 2 minutes a woman is diagnosed with breast cancer and 1 woman will die of breast cancer every 13 minutes. Public health data indicated that the global burden of breast cancer in women, measured by incidence, mortality, and economic costs, is substantial and on the increase.

3) Early detection helps recovery

When detected in its early stages, there is a 30% chance that the cancer can be treated effectively. Death rates from breast cancer have been declining since about 1990, in part due to better screening and early detection, increased awareness, and continually improving treatment options. A woman can get a better chance of complete regaining from the cancer if diagnosed at prior stage, thus firing the need of development of efficient diagnosis techniques like Information and Communication Technologies (ICT).

4) Detection dilemma

However, late detection of advanced-stage tumors makes the treatment more difficult. Currently, the most used techniques to detect breast cancer in early stages are: mammography (63% to 97% correctness), FNA (Fine Needle Aspiration) with visual interpretation (65% to 98% correctness) and surgical biopsy (approximately 100% correctness). Therefore, mammography and FNA with visual interpretation correctness varies widely, and the surgical biopsy, although reliable, is invasive and costly.

What is more, normally in traditional cases, double-reading (same mammograms are read by two radiologists individually) has been encouraged in decreasing the percentage of overlooked cancers and it is at present the supreme technique incorporated in most of the screening programs instead of the fact that it earns in surplus workload and costs.

Therefore, a platform for the computer-aided detection/diagnosis systems is established for backing up a single radiologist reading mammograms providing sustenance to her/his decisions.

5) Previous research work

Some papers were published during the last 30 years trying to achieve the best performance for the computational interpretation of FNA samples.

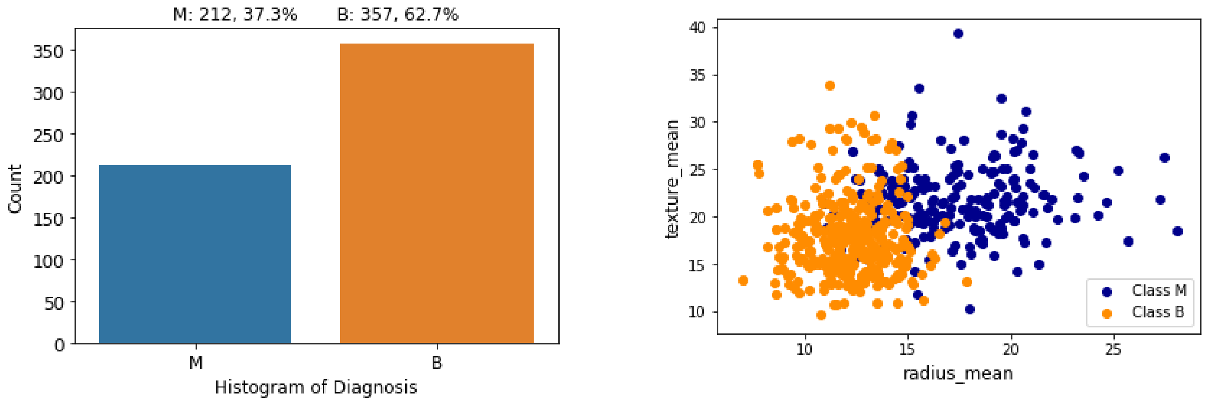
- Omar,et al [7] applied three machine-learning algorithms (Support Vector Machine, K-nearest neighbors, and Decision tree) to do classifier training. The outcomes of their study have revealed that quadradic support vector machine grants the largest accuracy of (98.1%) with lowest false discovery rates. Their experiments were carried out using Matlab.
- Vikas, et al [8] applied three popular data mining algorithms (Naïve Bayes, RBF Network, J48) to develop the prediction models. They used 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models. The results indicated that the Naïve Bayes performed best which received 97.36% accuracy as a predictor on the holdout sample, RBF Network came out to be the second with 96.77% accuracy, J48 came out third with 93.41% accuracy.
- Ioannis, et al [11] used probabilistic and generalized regression neural classifiers to deal with the breast cancer diagnosis and prognosis using the Wisconsin Diagnostic Breast Cancer (WDBC) as well as the Wisconsin Prognostic Breast Cancer (WPBC) data sets. The accuracy of the neural classifiers reaches nearly 98% for the diagnosis and 92% for the prognosis problem.

- Borges, et al. [12] used two well-known machine learning techniques to test their training models: Bayesian Networks and J48. The best accuracy in their paper was achieved by the Bayesian Networks algorithm, 97.80% of accuracy in its best configuration, while J48 had 96.5% of accuracy.

More previous research work could be searched in Google Scholar or Springer, archived paper could be downloaded via Duke Library.

Data Description

Wisconsin Diagnostic Breast Cancer (WDBC) dataset were utilized in this project. The dataset is publically available [1], [2], and was created by Dr. William H. Wolberg, physician at the University of Wisconsin Hospital at Madison, Wisconsin, USA. The dataset consists of **569** instances (357 benign – 212 malignant), where each one represents FNA test measurements for one diagnosis case. For this dataset each instance has **32** attributes, where the first two attributes correspond to a unique identification number and the diagnosis (benign / malignant). The rest 30 features are computations for ten real-valued features, along with their mean, standard error and the mean of the three largest values for each cell nucleus respectively. The attribute information of the features is summarized in Table 1, and diagnosis information is show in Fig.1, while more details of feature information examples could be found in Fig. 2.

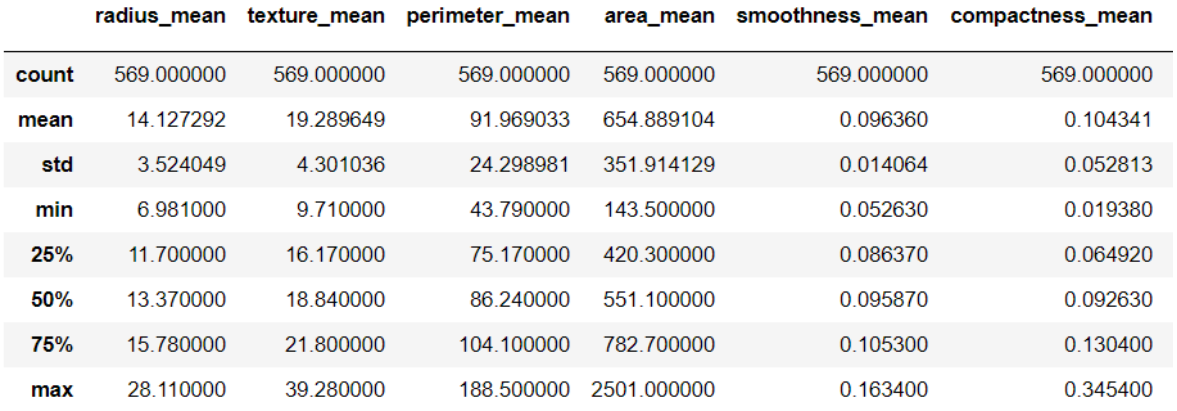


(a)

(b)

Fig. 1(a) Histogram of Diagnosis Distribution, (b) Scatter plot of radius_mean and texture_mean

Here to get a general view of the dataset, Fig. 1(a) shows Diagnosis distribution of dataset, and Fig.1(b) shows first two features plot classified by diagnosis value (M / B).



	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341
std	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813
min	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380
25%	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920
50%	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630
75%	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400
max	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400

Fig. 2 General numeric description of first six features

These are all real-valued features computed for each cell nucleus, from the graph and data visualization, we could observe that diagnosis value (M / B) are closed related to the shape or value of cell nucleus. Thus, we could apply machine learning algorithm to train an optimized model for the data set, then test our trained model within the dataset.

Table 1. Summary of Data Set

Class Distribution	Benign: 357, 62.7%		
	Malignant: 212, 37.3%		
Number of Instances	569		
Features		**Attribute**	**Type**
	1	radius (mean of distances from center to points on the perimeter)	Numeric
	2	texture (standard deviation of gray-scale values)	Numeric
	3	perimeter	Numeric
	4	area	Numeric
	5	smoothness (local variation in radius lengths)	Numeric
	6	compactness (perimeter^2 / area - 1.0)	Numeric
	7	concavity (severity of concave portions of the contour)	Numeric
	8	concave points (number of concave portions of the contour)	Numeric
	9	symmetry	Numeric
	10	fractal dimension ("coastline approximation" - 1)	Numeric
	11	ID number	Numeric
	12	Diagnosis	Nominal
Missing Values	None		

Anticipated Technical Approach/Methods

Original dataset before being publically available, there are missing points. However, this dataset has been revised, so it could be considered 'noise-free' and has none missing feature values.

Step 1: Data Pre-processing

First step of the dataset analysis would be pre-processing the data. A lot of factors affect the success of Machine Learning (ML) on a given task. The representation and quality of the instance data is first and foremost. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult [13].

Though it was claimed there were no missing values, we need to be meticulous when we are dealing with data. The pre-processing will focus on managing the missing attributes, the unbalanced data (outlier values) and the number of attributes used to train the classifier. Fig. 3 simply shows a schematic diagram to identify outliers. Here we would remove rows which contain value outside of 3 standard deviations.

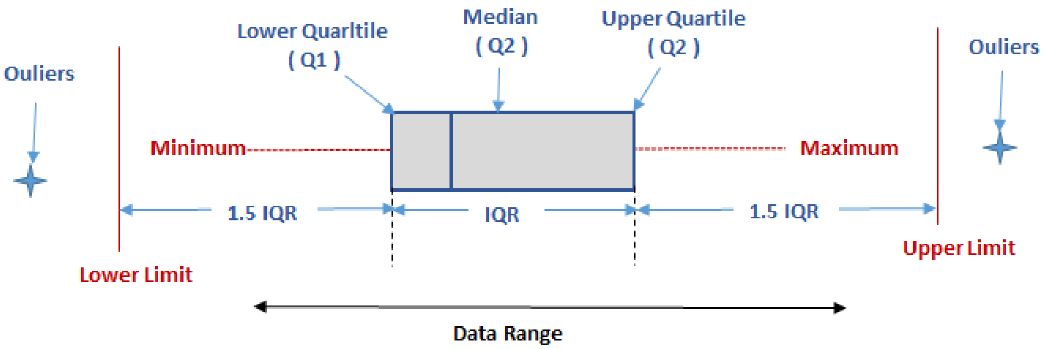


Fig. 3 Outlier Definition

Step 2: Data Visualization

Second step would be data visualization. As we have 32 features, it would be complex and time-costing to train such multi-features model. We could use PCA and t-SNE for Data Visualization.

Principal Components Analysis (PCA):

PCA is a technique for reducing the number of dimensions in a dataset whilst retaining most information. It is a method of spectral clustering. Spectral clustering is related to kernel principal components, a non-linear version of linear principal components. Standard linear principal components (PCA) are obtained from the eigenvectors of the covariance matrix, and give directions in which the data have maximal variance [14]. It is using the correlation between some dimensions and tries to provide a minimum number of variables that keeps the maximum amount of variation or information about how the original data is distributed.

One of the most important applications of PCA is for speeding up machine learning algorithms. It does not do this using guesswork but using hard mathematics and it uses something known as the eigenvalues and eigenvectors of the data-matrix. These eigenvectors of the covariance matrix have the property that they point along the major directions of variation in the data. These are the directions of maximum variation in a dataset.

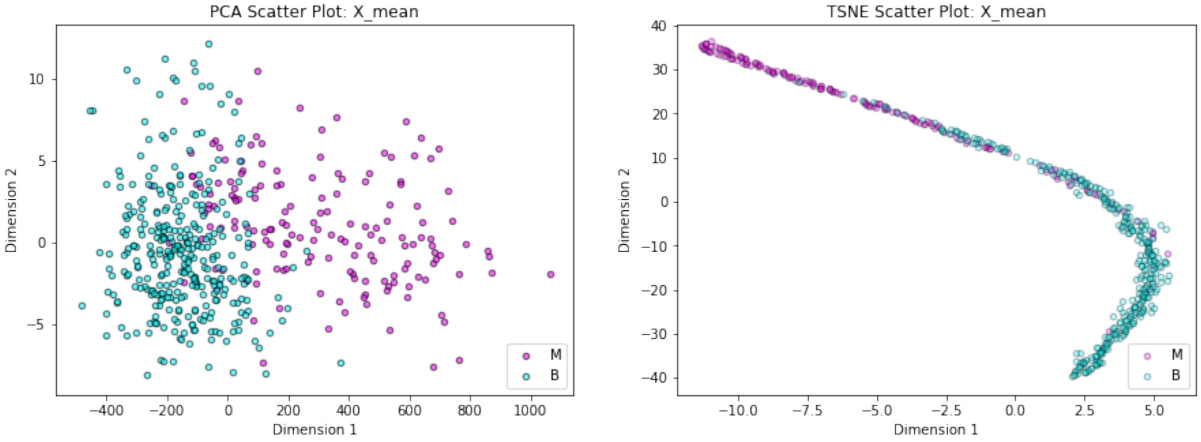
As there are three fields (mean, standard error, worst value) of the ten features, to explore the data, the data set is separated into three parts, X_mean, X_se, X_worst. X represent the whole feature data set. It might be possible that within one field, we could develop a good model.

T-Distributed Stochastic Neighbouring Entities (t-SNE):

t-Distributed Stochastic Neighbor Embedding (t-SNE) is another technique for dimensionality reduction and is particularly well suited for the visualization of high-dimensional datasets. Contrary to PCA it is not a mathematical technique but a probablistic one. t-Distributed stochastic neighbor embedding (t-SNE) minimizes the divergence between two distributions: a distribution that measures pairwise similarities of the input objects and a distribution that measures pairwise similarities of the corresponding low-dimensional points in the embedding [15].

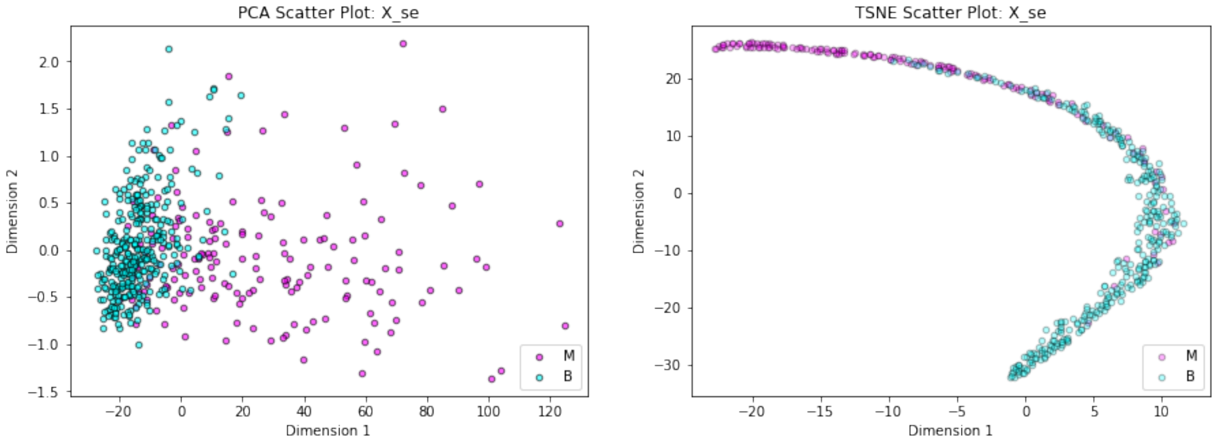
Before Standardization:

Fig. 4, Fig. 5, Fig. 6, Fig. 7 respectively shows the PCA scatter plot and TSNE scatter plot of corresponding dataset.



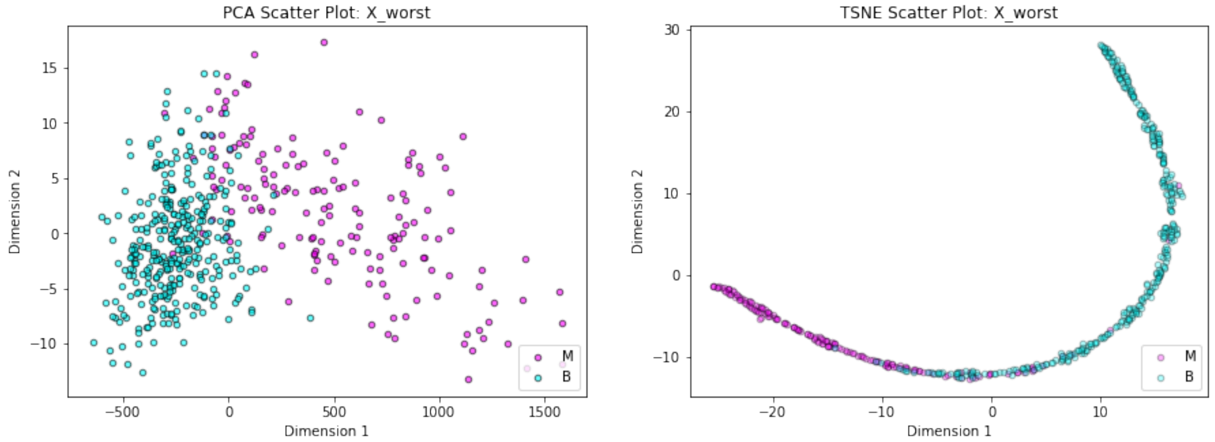
(a) (b)

Fig. 4 (a) PCA scatter Plot of Dataset (X_mean, y), (b) TSNE scatter Plot of Dataset (X_mean, y)



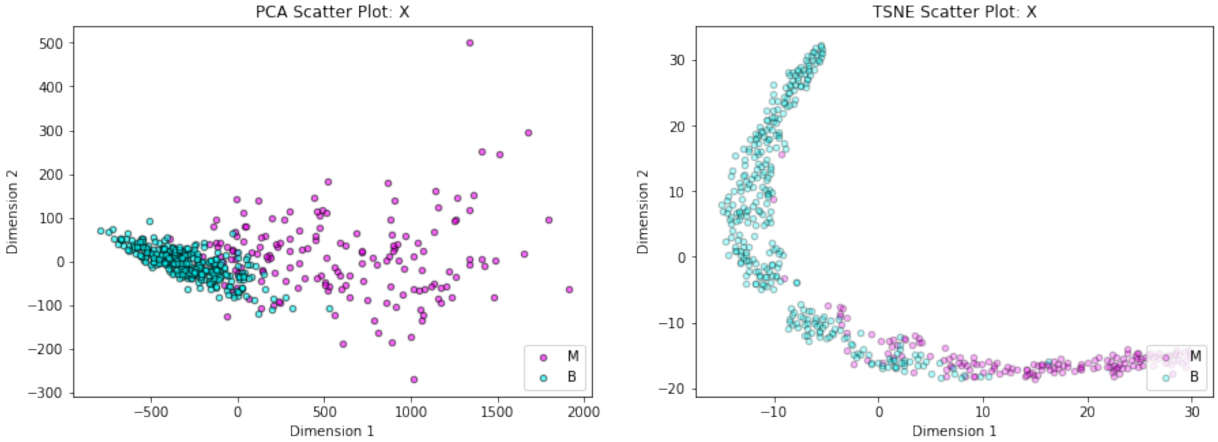
(a) (b)

Fig. 5 (a) PCA scatter Plot of Dataset (X_se, y), (b) TSNE scatter Plot of Dataset (X_se, y)



(a) (b)

Fig. 6 (a) PCA scatter Plot of Dataset (X_worst, y), (b) TSNE scatter Plot of Dataset (X_worst, y)

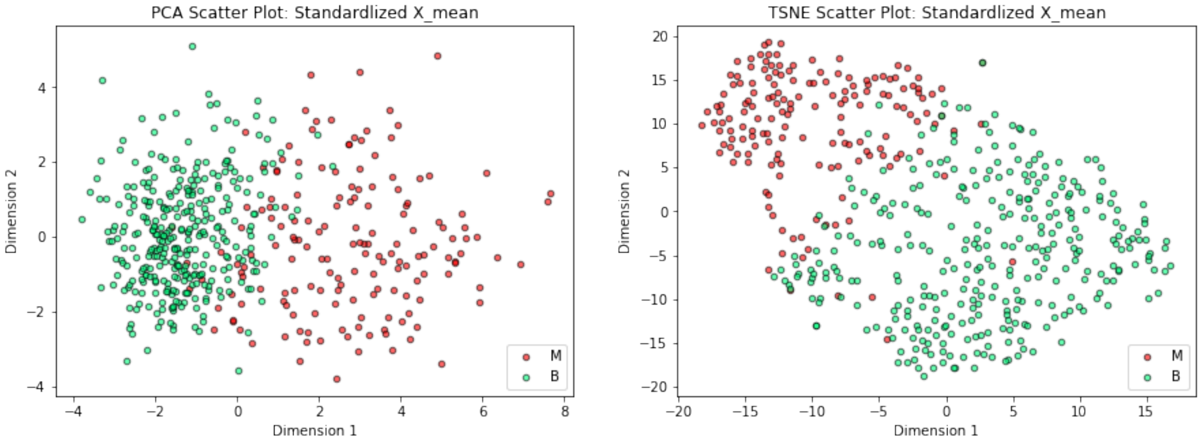


(a) (b)

Fig. 7 (a) PCA scatter Plot of Dataset (X, y), (b) TSNE scatter Plot of Dataset (X, y)

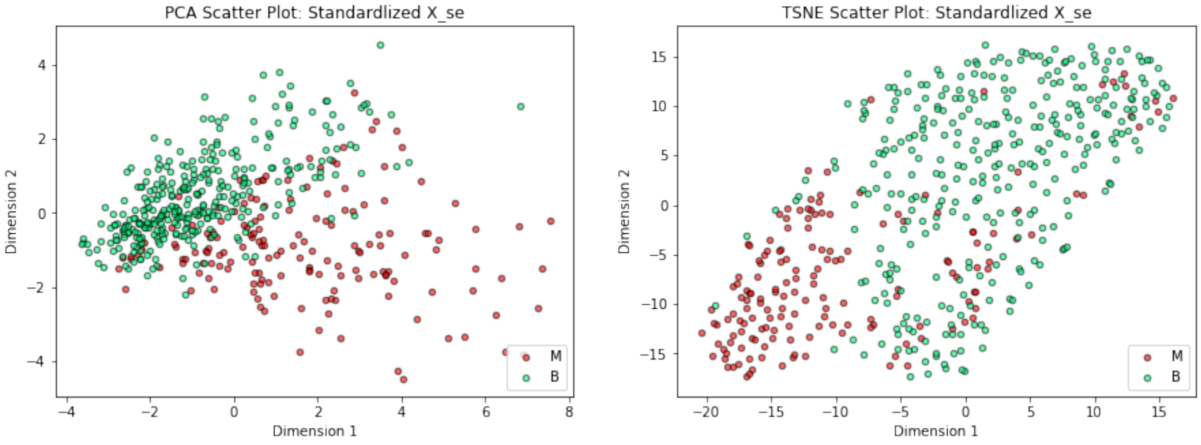
After Standardization:

Fig. 8, Fig. 9, Fig. 10, Fig. 11 respectively shows the PCA scatter plot and TSNE scatter plot of corresponding dataset.



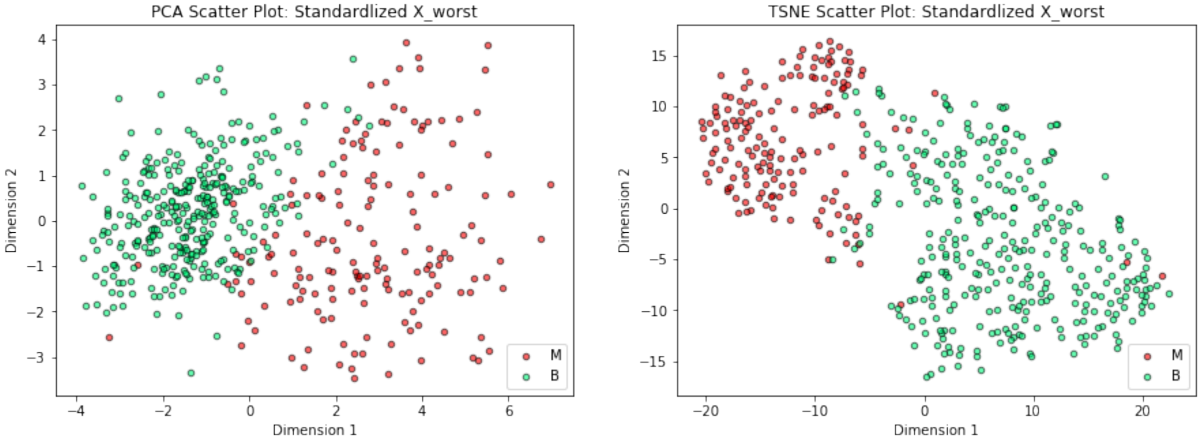
(a) (b)

Fig. 8 (a) PCA scatter Plot of Dataset (X_mean_std, y), (b) TSNE scatter Plot of Dataset (X_mean_std, y)



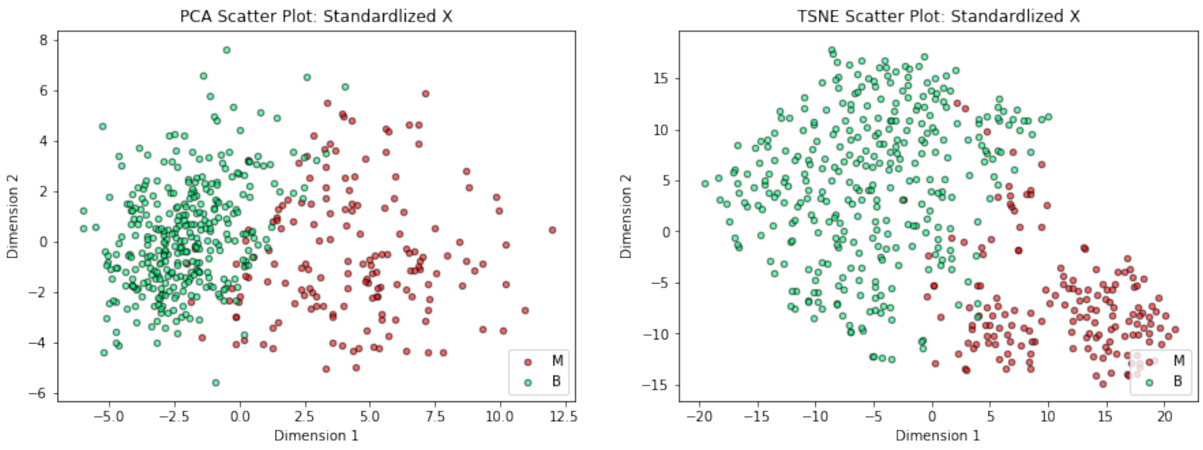
(a) (b)

Fig. 9 (a) PCA scatter Plot of Dataset (X_se, y), (b) TSNE scatter Plot of Dataset (X_se, y)



(a) (b)

Fig. 10 (a) PCA scatter Plot of Dataset (X_worst, y), (b) TSNE scatter Plot of Dataset (X_worst, y)



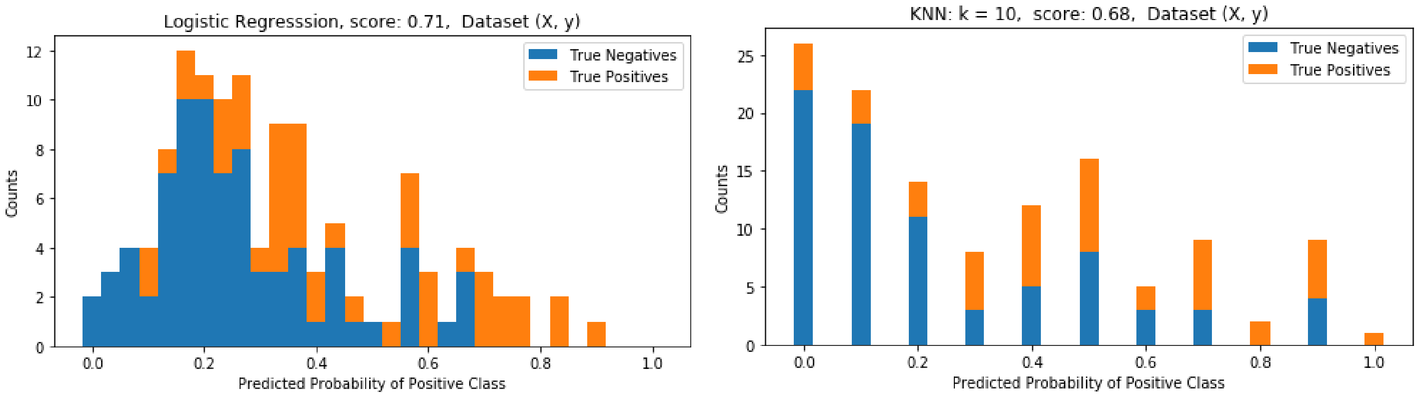
(a) (b)

Fig. 11 (a) PCA scatter Plot of Dataset (X_std, y), (b) TSNE scatter Plot of Dataset (X_std, y)

Step 3: Model Training

Third step would be using different data mining algorithms to develop prediction models. We could apply the algorithm we learn in class like KNN, or other applicable algorithms like (Naïve Bayes, RBF Network, J48, Support Vector Machine, and Decision tree) to solve the diagnosis problem. As the dataset is publically available and the same, previous work is replicable. The main reason for the project is to learn, thus I think reproduce previous work should be meaningful.

Here I simply shown Logistic Regression and KNN(k = 10) to show the Predicted Probability of Dataset (X, y) in Fig. 12. More work will be done and analyzed in the future.



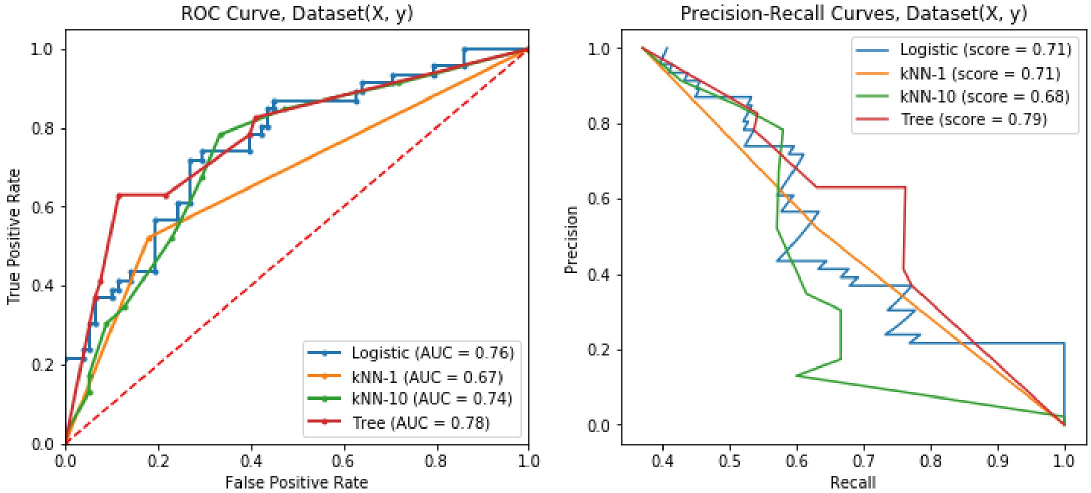
(a) (b)

Fig. 12 (a) Predicted Probability of Dataset (X, y) using Logistic Regression, (b) Predicted Probability of Dataset (X, y) using KNN, k=10

Step 4: Model Testing

Fourth step would be model performance measuring. We have many influencing factors to define how good a training model is. However, for simplicity, here we use accuracy to be the metric for machine learning algorithms (precision, recall, F1 Score, ROC Curve, etc would be fine). Furthermore, we could apply classification accuracy, sensitivity, specificity, positive and negative predictor values and confusion matrix to evaluate different models performance.

Here I simply shown four different machine learning performances of ROC curves and Precision-Recall curves in Fig. 13. More work will be done and analyzed in the future.

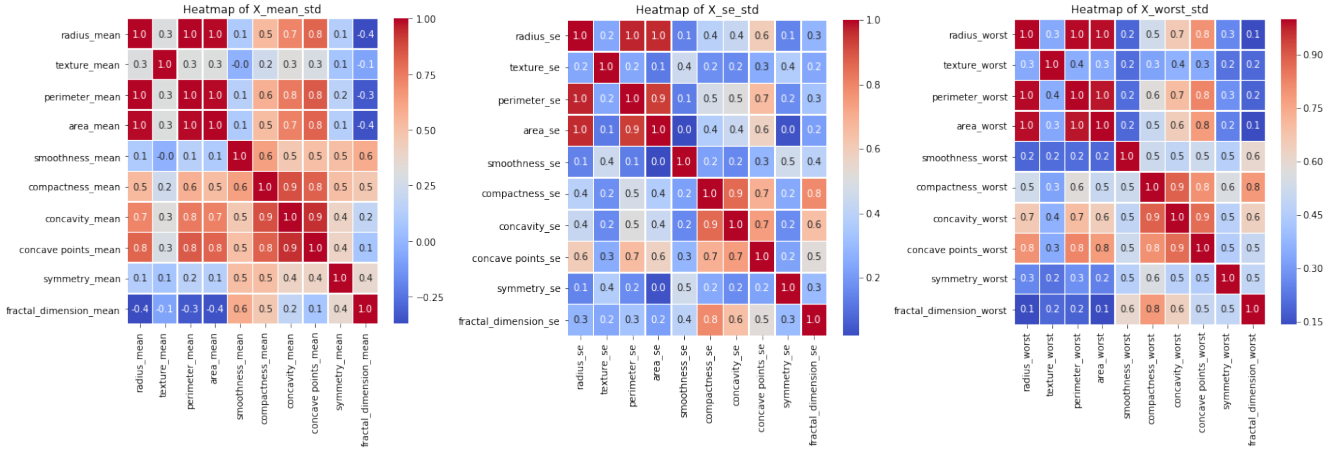


(a) (b)

Fig. 13 (a) ROC Curve of four algorithms in Dataset (X,y) , (b) Precision-Recall Curves of four algorithms in Dataset (X,y)

Future Plan:

From Fig.14 We know that different features have different level of importance for the prediction. In the future, I will try to pick most relevant features according to the correlation of three fields to train and test models. It might be helpful to reduce work it only some features are sufficient for a good model.



(a) (b) (c)

Fig. 14 (a) Heatmap of Dataset (X_mean_std,y) , (b) Heatmap of Dataset (X_mean_std,y), (c) Heatmap of Dataset (X_worst_std,y)

Reference:

[1] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/discussion/62297
[2] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
[3] https://www.nationalbreastcancer.org/
[4] Shaikh, Tawseef Ayoub, and Rashid Ali. "Applying Machine Learning Algorithms for Early Diagnosis and Prediction of Breast Cancer Risk." *Proceedings of 2nd International Conference on Communication, Computing and Networking*. Springer, Singapore, 2019.
[5] Henriksen, Emilie L., et al. "The efficacy of using computer-aided detection (CAD) for detection of breast cancer in mammography screening: a systematic review." Acta Radiologica 60.1 (2019): 13-18.
[6] Sri, M. Navya, et al. "A Comparative Analysis of Breast Cancer Data Set Using Different Classification Methods." Smart Intelligent Computing and Applications. Springer, Singapore, 2019. 175-181.
[7] Obaid, Omar Ibrahim, et al. "Evaluating the Performance of Machine Learning Techniques in the Classification of Wisconsin Breast Cancer." International Journal of Engineering & Technology 7.4.36 (2018): 160-166.
[8] Chaurasia, Vikas, Saurabh Pal, and B. B. Tiwari. "Prediction of benign and malignant breast cancer using data mining techniques." Journal of Algorithms & Computational Technology 12.2 (2018): 119-126.
[9] Bhattacherjee, Aindrila, et al. "Classification approach for breast cancer detection using back propagation neural network: a study." Biomedical image analysis and mining techniques for improved health outcomes. IGI Global, 2016. 210-221.
[10] Coughlin, Steven S., and Donatus U. Ekwueme. "Breast cancer as a global health concern." Cancer epidemiology 33.5 (2009): 315-318.
[11] Anagnostopoulos, Ioannis, et al. "The Wisconsin Breast Cancer Problem: Diagnosis and DFS time prognosis using probabilistic and generalised regression neural classifiers." *Oncology Reports, Special Issue Computational Analysis and Decision Support Systems in Oncology* (2005).
[12] Borges, Lucas Rodrigues. "Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection." *Group* 1.369 (1989).
[13] Kotsiantis, S.B., Kanellopoulos, D. and Pintelas, P.E., 2006. Data preprocessing for supervised leaning. International Journal of Computer Science, 1(2), pp.111-117.
[14] Friedman, J., Hastie, T. and Tibshirani, R., 2001. *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics.
[15] [15] Maaten, L.V.D. and Hinton, G., 2008. Visualizing data using t-SNE. Journal of machine learning research, 9(Nov), pp.2579-2605.