Final Report: Breast Cancer Wisconsin

Ziyuan Shen

April 27, 2019

1 Problem Description

Over the past few decades, machine learning techniques have been increasingly applied to disease studies [1]. The advanced image technology provides increasing clinical image information. The development of single-cell technology greatly facilitates the availability of multi-omics data [2, 3]. Accordingly, delicately designed data mining and machine learning techniques are desired to explore the biological conclusions underlying the large datasets. Relevant studies can aid disease diagnosis and precision medicine, as well as disease early detection. Cancer classification is one of the top promising topics [4]. By far, various biological datasets [5, 6] have been used for cancer classification studies. Improving the classification performance is undoubtedly meaningful as human beings' overall health conditions will be improved by these efforts.

Cancer classification historically requires prior biological knowledge. However, most machine learning engineers do not have a biological background. Therefore, the general classification performance is worth improving. Prediction models with high accuracy are aimed to aid oncologists with diagnosis and prognosis.

The goal is to create a good predictor with samples from known classes, and possibly identify hidden cancer subtypes, without any biological priors. Most existing works develop models based on specific cancer types [5, 7, 8]. Thus, a more general and systematic procedure remains open-ended. Although biological prior is not require for performing classification, different estimators will yield different performances when tested on a certain dataset. The task is to explore the data structure, choose appropriate algorithms, compare different algorithms and analyze the result.

2 Data Description

2.1 Background

Breast Cancer Wisconsin is a classic cancer dataset for classification and has been explored by many machine learning researchers for testing algorithms [8–13]. The dataset is composed of digital image information of breast cancer cell nuclei. The cells are labeled as malignant and benign.Therefore, cancer subtypes have the potential to be classified with only cells data, without any prior biological knowledge. The dataset contains two classes, thus should serve as a good example for conducting classification algorithms. According to the original paper [9] generating this dataset, discrimination between the two classes is hard by one single characteristic. By integrating all features, the two masses are classifiable.

2.2 Summary Statistics

The data contains 569 samples in total with two classes: malignant cells and benign cells. The original image provides 10 characteristics for each sample: radius, texture, perimeter, area, smoothness,

compactness, concavity, concave points, symmetry and fracture. The mean value, standard error, and mean of the three largest values are calculated from the original data, respectively, for each data sample. Therefore, 30 features are generated for each sample. Statistics of the dataset are summarized in Table 1. Statistics of the first ten features are summarized in the form of boxplots in Figure 1.

Table 1: Summary Statistics of Breast Cancer Datasets.

Statistics \ **Classes**	**Malignant**	**Benign**
#Characteristic	10	
#Features	30	
#Total Samples	569	
#Samples	212	357
#Proportion	37.3%	62.7%

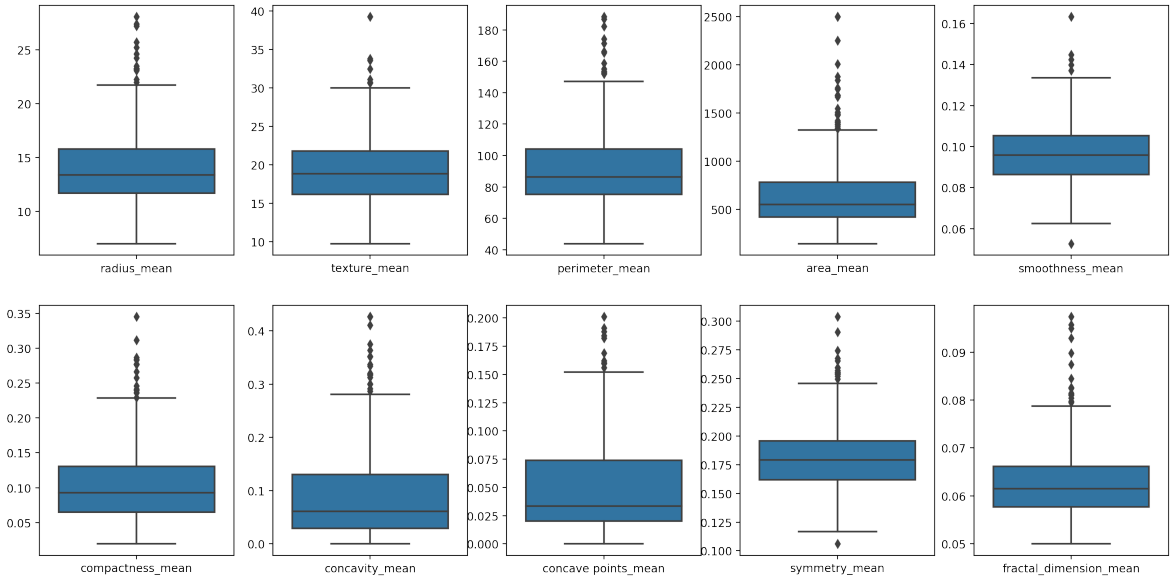

Figure 1: Summary statistics of the first ten features.

To explore the correlation between these samples, the correlation heat map of the first 10 features, which are the mean values of the 10 characteristics, is plotted in Figure 2. By observing the heat map, radius, perimeter and area are highly correlated. We can also draw this conclusion from the definition of these three characteristics. Therefore, we can choose any of these three features when performing classification. Similarly, compactness mean, concavity mean and concave points mean are highly related.

2.3 Data Visualization

2.3.1 Pairwise Scatter Plot

By dropping features with high correlation, six features are extracted for pairwise scatter plot as depicted in Figure 3. From the graph, statistics of the two classes overlap each other but should be separable if treated with appropriate classifier. The diagonal squares show the histograms of the six features.

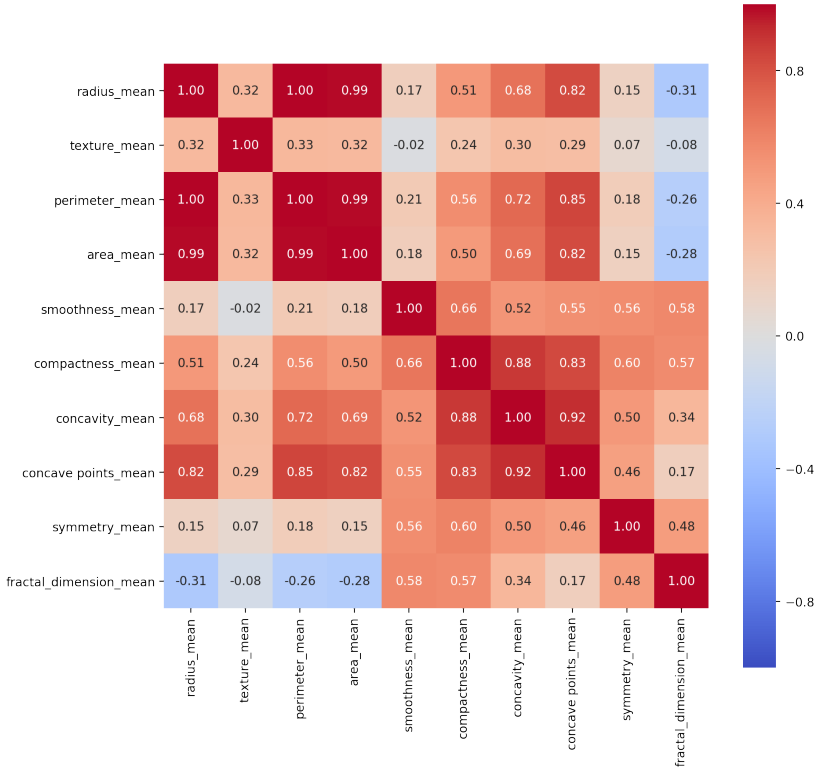

Figure 2: Correlation heatmap.

2.3.2 t-SNE

According to Figure 1, different features of this dataset have entirely different magnitudes. Therefore, some preprocessing might be necessary before visualizing the data. In order to conduct data visualization, the data is firstly reprocessed with z-scoring so that the processed data has zero mean and standard error. T-SNE [14] visualization is shown in Figure 4. A 3D plot for three components is provided as well as 2D plots for every two components.

2.3.3 Principle Component Analysis

As the number of features of this dataset is high, Principle Component Analysis (PCA) [15] is appropriate for visualizing the data. According to PCA, the first three components only contains 72.6% of the variance, which is large enough to represent the main structure of the data. PCA data visualization is shown in Figure 5.

2.4 Source

The source data can be downloaded from UCI machine learning repository [16] or kaggle website [17].

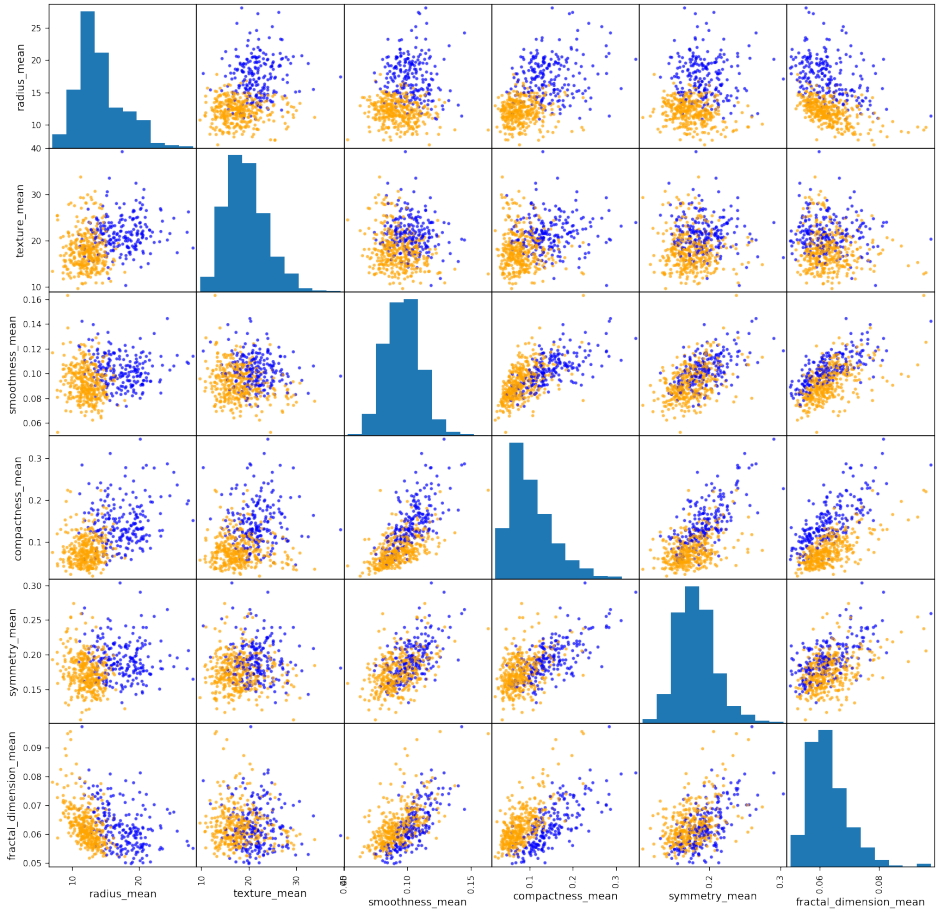

Figure 3: Pairwise scatter plot of six features.

3 Approach/Methods

3.1 Preprocessing

3.1.1 Normalization

As the data features have entirely different scales, preprocessing is important for obtaining accurate prediction performance. For instance, if KNN with $l2$ norm is applied, the calculation of distances will be dominated by the feature of the largest scale, preventing the estimator to learn from other features. As illustrated by the histograms in Figure 3, the data are very close to Gaussian distribution. Therefore, we can believe that the data are normal distributed. Then, z-Scoring may be an ideal normalization approach. z-Scoring scales data to zero mean and unit variance, as in Eq. (1):

$$z = \frac{x - u}{s}, \tag{1}$$

where u is the mean of the training samples and s is the standard deviation of the training samples.

3.1.2 Dimension Reduction

As the dataset has high dimensions, an approach for dimension reduction will be important to data visualization and the prediction performance.

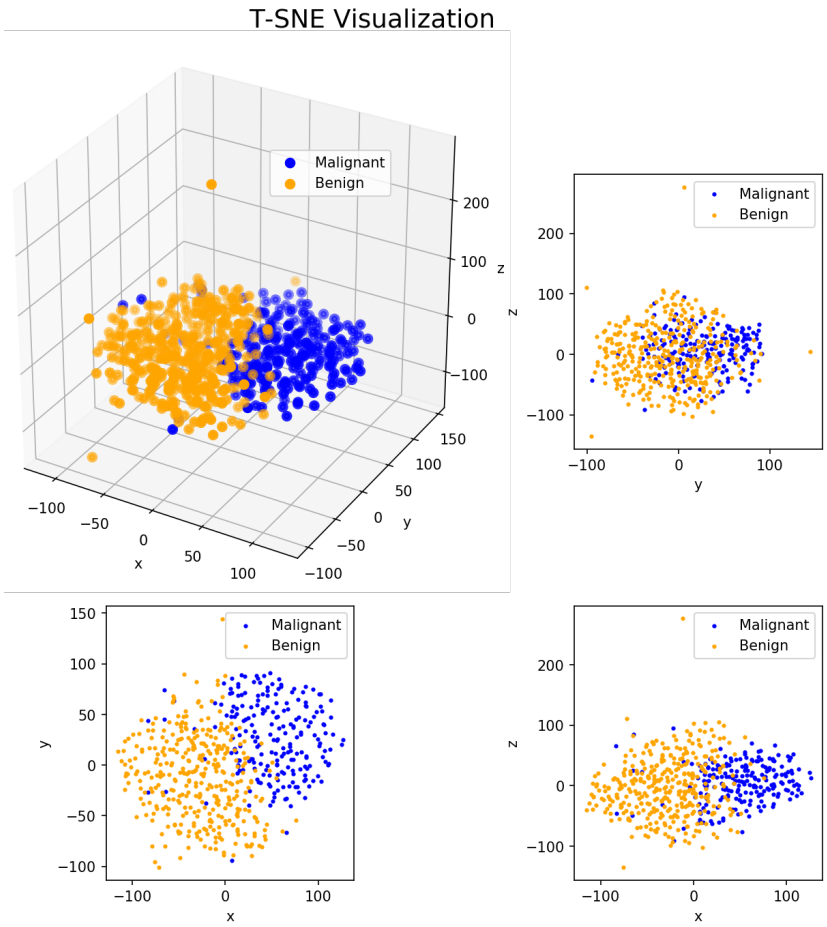

Figure 4: T-SNE visualization.

Principle Component Analysis (feature transformation): As we believe that the data is Gaussian distributed and z-scoring is applied to the dataset as part of preprocessing, Principle Component Analysis (PCA) [15] is a useful tool for dimension reduction. PCA decomposes a high dimensional dataset into uncorrelated and orthogonal components, which explain the maximum variance of the data. The maximum number of components is $\min(p, n-1)$, where p is the number of dimensions of the original data and n is the number of samples. Usually, less than maximum number of components are selected to achieve the goal of dimension reduction. The transformation is in the form:

$$\boldsymbol{y} = \mathbf{A}^T \boldsymbol{x}, \tag{2}$$

where $\boldsymbol{x}$ is the input vector, $\boldsymbol{y}$ is the output vector, and $\mathbf{A}$ is the transformation matrix. PCA requires that $E\{\boldsymbol{x}\} = 0$, thus $E\{\boldsymbol{y}\} = 0$. As $\boldsymbol{y}$ should have uncorrelated components, the correlation matrix should be a diagonal matrix:

$$R_y = E\{\boldsymbol{y}\boldsymbol{y}^T\} = E\{\mathbf{A}^T\boldsymbol{x}\boldsymbol{x}^T\mathbf{A}\} = \mathbf{A}^T E\{\boldsymbol{x}\boldsymbol{x}^T\}\mathbf{A} = \mathbf{A}^T R_x \mathbf{A}. \tag{3}$$

By eigendecomposition, $\mathbf{A}$ is chosen to have orthonormal eigenvectors of R_x. Visualization of the first three components of PCA, which contain 72.6% of the data variance, has already been shown in Section 2.3.3 (Figure 5). As the dataset has 30 dimensions, the maximum number of dimensions is 30. The variance each component explains as long as the cumulative variance is plotted in Figure 6. The first 15 components contain 98.6% of the variance, which convey almost all the information of the dataset. Therefore, by PCA, the data dimension can be reduced from 30 to 15.

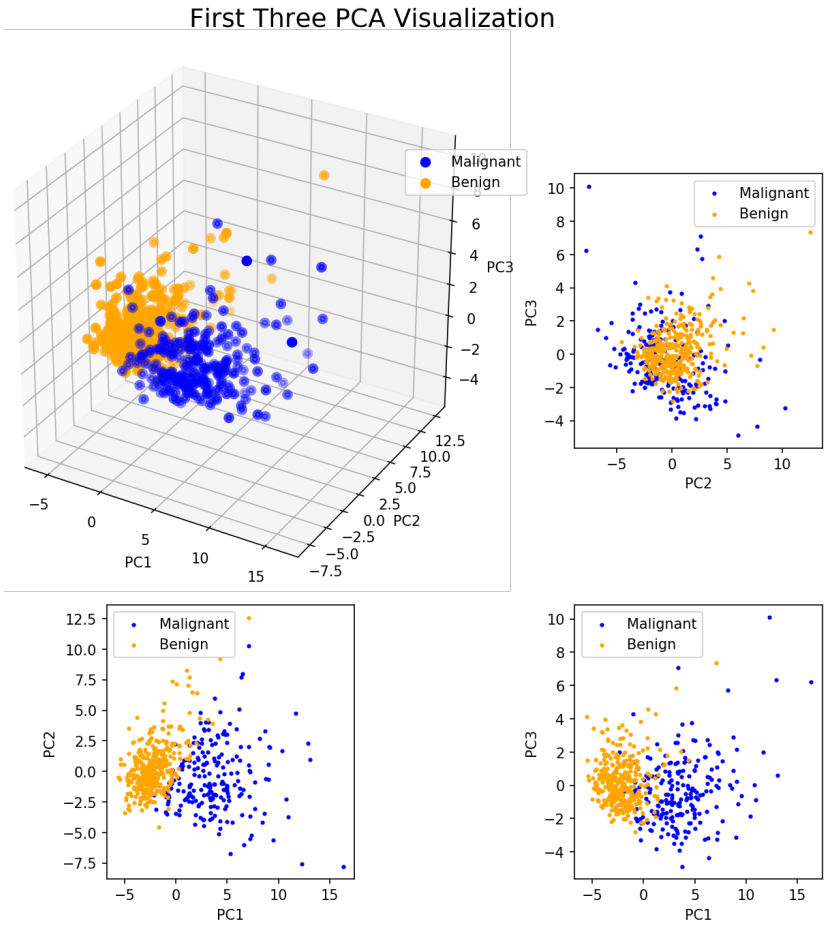

Figure 5: First three principle components visualization.

Feature Elimination: From the data correlation as shown in Figure 2, we can tell that some features (radius, perimeter, area and compactness, concavity, concave points) with high correlation are very likely to provide redundant information. Therefore, we can choose to directly eliminate some features to save computational complexity. If we choose radius out of radius, perimeter and area, and choose compactness out of compactness, concavity and concave points, then $6 \times 3 = 18$ features will be maintained ultimately.

Trade-offs: Some anticipated trade-offs are summarized in Table 2.

Table 2: Trade-off Analysis of Dimension Reduction.

Trade-offs	**Fewer Features**	**More Features**
Interpretability	High	Low
Overfitting	Less likely	Likely
Prediction Accuracy	Low	High

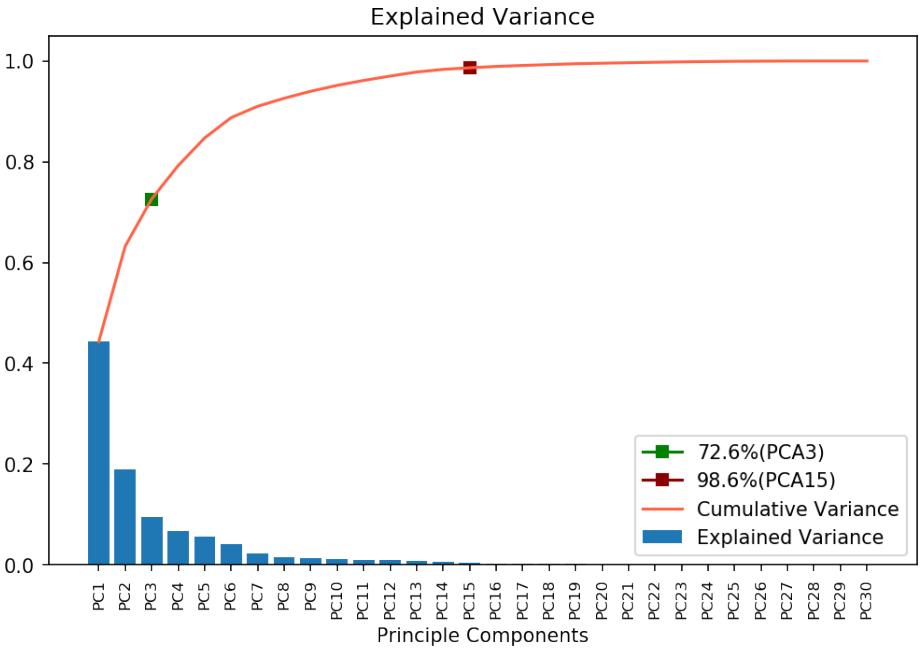

Figure 6: Explained variance for all PCA components.

3.2 Classifiers

Defined as a classification problem with two classes, the dataset can be tested by the following algorithms:

1) **Gaussian Bayes Classifier**[18]

For the case of two classes, decision statistic of Bayes classifier is defined by the likelihood ratio. To ease computation, we usually take ln likelihood. As the dataset has relatively balanced classes, we can assume same priors. The discriminant function is then stated in Eq. (4), if we assume symmetric costs.

$$g(x) = g_1(x) - g_0(x) = \ln \frac{p(x|\omega_1)}{p(x|\omega_0)} \tag{4}$$

Based on the belief in our data, we assume multivariate normal distribution:

$$p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right]. \tag{5}$$

Hence,

$$g_i(\mathbf{x}) = \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^T \mathbf{x} + w_{i0}, \tag{6}$$

where $\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}$, $\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$ and $w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i|$. z-Scoring is not necessary for Gaussian Bayes classifier.

2) **Linear Discriminant Analysis [19]**

The LDA (linear discriminant analysis) method tries to project the original high dimensional data onto one dimension such that the distance between means is large and the sample variance is small after projecting. Ultimately, LDA solves:

$$\underset{\mathbf{w}}{\operatorname{argmax}} \frac{|\widetilde{\mathbf{m}}_1 - \widetilde{\mathbf{m}}_0|}{\tilde{s}_1^2 + \tilde{s}_0^2}, \tag{7}$$

where $\widetilde{\mathbf{m}}_i = \mathbf{w}^T\mathbf{m}_i$, $\tilde{s}_i^2 = \sum_{n=1}^{N_i}(\mathbf{w}^T\mathbf{x}_{in} - \mathbf{w}^T\mathbf{m}_i)^2$.

3) **Logistic Regression [20]**
Logistic regression aims to model the posterior probabilities of all classes via linear functions of all features. More specifically, logistic regression assumes that the log odds of the model is a linear function of $\mathbf{x}$ such that:

$$\ln\frac{p(H_1|\mathbf{x})}{p(H_0|\mathbf{x})} = \beta_0 + \boldsymbol{\beta}^T\mathbf{x} = \mathrm{L}(\mathbf{x}). \tag{8}$$

In order to maximize the posterior likelihood, the problem is equivalent to finding $\boldsymbol{\beta}$ such that:

$$\boldsymbol{\beta} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \sum_{i=1}^{n} \left\{ y_i(\boldsymbol{\beta}^T\mathbf{x}_i) - \ln(1 + e^{\boldsymbol{\beta}^T\mathbf{x}_i}) \right\}. \tag{9}$$

4) **Support Vector Machine [21]**
A SVM (support vector machine) classifier generates a hyperplane for separating two classes. In general, SVM solves the following optimization problem:

$$\begin{aligned} &\underset{w,b}{\operatorname{argmin}} \frac{1}{2}w^Tw + C\sum_{i=1}^{n}\xi_i \\ \text{subject} \quad \text{to} \quad &y_i(w^Tx_i + b) \geq 1 - \xi_i, \\ &\xi_i \geq 0, \quad i = 1, 2, \ldots, n. \end{aligned} \tag{10}$$

5) **K-Nearest Neighbors [22]**
A K-Nearest Neighbors classifier adopts information around each query point for classification. A query point is assigned the data class which dominates its neighborhood. The decision statistic is computed as follows:

$$\lambda(x) = \frac{1}{k}\sum_{i=1}^{k}\mathbb{I}(y_i = 1), \tag{11}$$

where $\mathbb{I}$ is the indicator function that holds:

$$\mathbb{I}(y_i = 1) = \begin{cases} 1, \text{if} \quad y_i = 1, \\ 0, \text{if} \quad y_i = 0. \end{cases} \tag{12}$$

$y_i, i = 1, \ldots, k$ are labels of the k nearest neighbors of the query point x. Here we employ l_2 norm for computing the distance:

$$D(x, y) = ||x - y||_2 = \left[\sum_{d=1}^{D}|x_d - y_d|^2\right]^{\frac{1}{2}}, \tag{13}$$

where D is the number of dimensions of the data.

Remark. *For the implementation of the above methods in this report, LDA and Bayes classifier are coded by myself and the others (z-scoring, PCA, t-SNE, KNN, logistic regression and SVM) are leveraged from sci-kit learn python packages [23].*

4 Results

As the number of samples is limited in the given dataset, cross-validation [24] will be necessary to evaluate performance of the predictor. When evaluating a predictor, accuracy is certainly of high priority. Apart from accuracy, computational complexity, model interpretability and potential of overfitting should also be taken into consideration.

4.1 Algorithmic Pipeline

The pipeline for creating the classifier is depicted in Figure 7. The original data is firstly assigned to one of M folds. Here we choose $M = 5$. Then z-scoring and dimension reduction are applied for preprocessing. Finally one classifier is chosen to train the classifier and cross validation performance is evaluated. The cross validation performance can then be used to tune classifier parameters to achieve the optimum choice. It should be noted that z-scoring is applied separately to the training and testing data sets to ensure no prior knowledge of the testing data set. 15 PCA components are firstly computed from the training data, then testing data is projected onto these components before running each classifier.

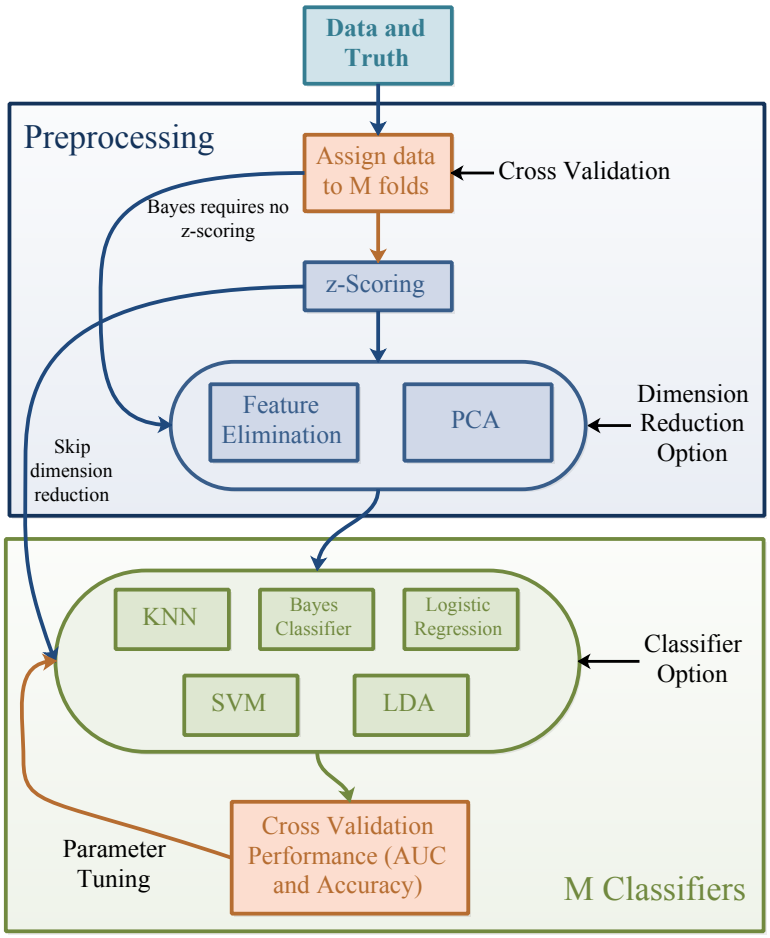

Figure 7: Classification pipeline.

4.2 ROC Results

The ROC results as well as corresponding AUC values of all specified classifiers by using full data dimensions, feature elimination and PCA are shown in Figure 8. As SVM produces no decision statistic, it is not included.

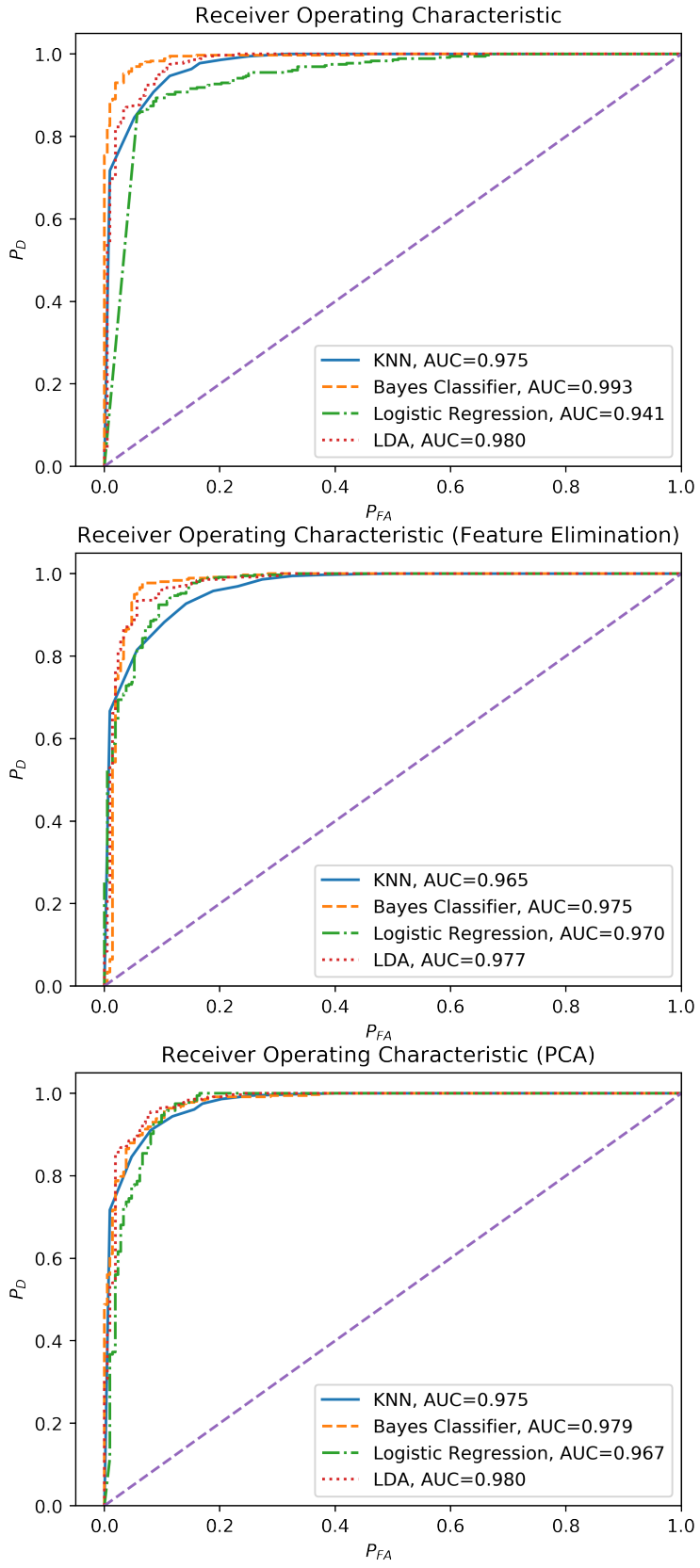

Figure 8: Comparison of ROC curves.

4.3 Visualization of Decision Statistics

Accordingly, decision statistics of these four classifiers are visualized in Figure 9. Note that no dimension reduction is employed here. By cross validation, each of the 569 samples gets a decision value. All of the data points are visualized by T-SNE with two components. Malignant and benign samples are represented by different scatter shapes (i.e. top triangles and squares) and different colors reflect the decision statistic values.

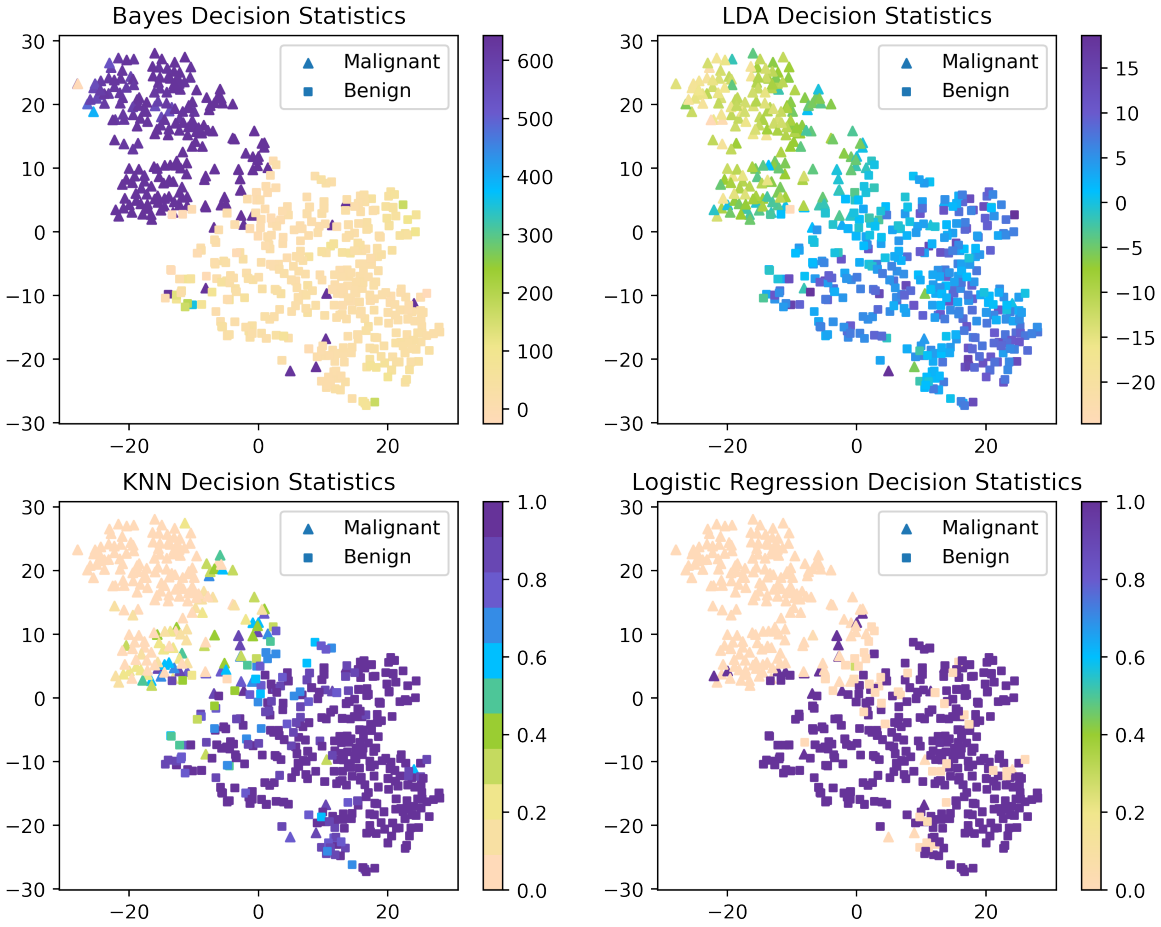

Figure 9: Comparison of decision statistics.

4.4 Classification Accuracy

Classification accuracy of all the classifiers are illustrated in Figure 10. Accuracy of the winner for each kind of preprocessing is explicitly displayed in the figure.

5 Conclusion

From the ROC results, it is obvious that Bayes classifier performs best in general. This result is not surprising as Bayes classifier assumes Gaussian distribution and the data is indeed multivariate normal distributed from our prior knowledge. In the meantime, KNN, LDA and logistic regression also provide similarly good performance. This is because that our data inherently has good linear separability, which can be observed from our data visualization.

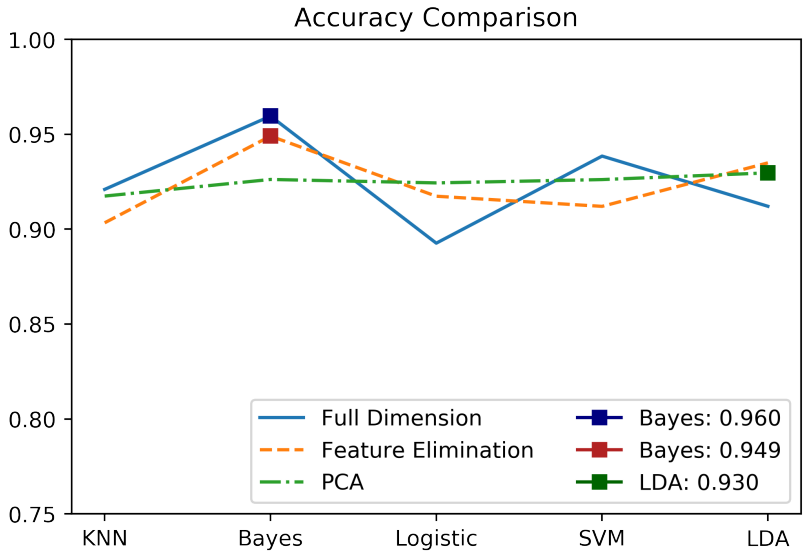

Figure 10: Classification accuracy comparison.

From the decision statistics, we can also observe that Bayes classifier provides best performance. By setting appropriate decision boundaries, all the classifiers have the potential to provide high accuracy for this dataset. Bayes classifier and logistic regression provide relatively robust decision statistics, as only one color dominates each class's data points, while LDA and KNN provide decision statistics that are dependent on the distance between the query point and the decision boundary. In conclusion, if we only care about prediction results, Bayes classifier and logistic regression are good choices. If we also care about prediction accuracy, KNN and LDA are more suitable choices.

For KNN, SVM and Bayes classifier, dimension reduction (feature elimination and PCA) reduces classifier accuracy, possibly because part of the data information is lost by applying dimension reduction. However, logistic regression's prediction accuracy increases after dimension reduction because feature selection and PCA may help relieve overfitting. LDA's performance also increases after preprocessing as feature elimination and PCA help reduce data correlation. As we evaluate our model by cross validation, the better the performance is, the less the classifier is overfitted.

References

[1] P. Sajda, "Machine learning for detection and diagnosis of disease," *Annu. Rev. Biomed. Eng.*, vol. 8, pp. 537–565, 2006.

[2] M. Ibrahim and K. Chakrabarty, "Cyber-physical digital-microfluidic biochips: Bridging the gap between microfluidics and microbiology," *Proceedings of the IEEE*, 2017.

[3] M. Ibrahim, K. Chakrabarty, and U. Schlichtmann, "Synthesis of a cyberphysical hybrid microfluidic platform for single-cell analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2018.

[4] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer informatics*, vol. 2, p. 117693510600200030, 2006.

[5] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *science*, vol. 286, no. 5439, pp. 531–537, 1999.

[6] X. Huang and W. Pan, "Linear regression and two-class classification with gene expression data," *Bioinformatics*, vol. 19, no. 16, pp. 2072–2078, 2003.

[7] S. Pineda, F. X. Real, M. Kogevinas, A. Carrato, S. J. Chanock, N. Malats, and K. Van Steen, "Integration analysis of three omics data using penalized regression methods: an application to bladder cancer," *PLoS genetics*, vol. 11, no. 12, p. e1005689, 2015.

[8] A. F. M. Agarap, "On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset," in *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing*. ACM, 2018, pp. 5–9.

[9] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology." *Proceedings of the national academy of sciences*, vol. 87, no. 23, pp. 9193–9196, 1990.

[10] W. H. Wolberg, O. L. Mangasarian, and R. Setiono, "Pattern recognition via linear programming: Theory and application to medical diagnosis," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 1989.

[11] K. P. Bennett, "Decision tree construction via linear programming," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 1992.

[12] K. P. Bennett and O. L. Mangasarian, "Robust linear programming discrimination of two linearly inseparable sets," *Optimization methods and software*, vol. 1, no. 1, pp. 23–34, 1992.

[13] K. Liu, G. Kang, N. Zhang, and B. Hou, "Breast cancer classification based on fully-connected layer first convolutional neural networks," *IEEE Access*, vol. 6, pp. 23 722–23 732, 2018.

[14] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[15] J. Lever, M. Krzywinski, and N. Altman, "Points of significance: Principal component analysis," 2017.

[16] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[17] "Breast Cancer Wisconsin (Diagnostic) Data Set: Predict whether the cancer is benign or malignant," https://www.kaggle.com/uciml/breast-cancer-wisconsin-data, 2019.

[18] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*. Springer Science & Business Media, 2013, vol. 31.

[19] H. Trevor, T. Robert, and F. JH, "The elements of statistical learning: data mining, inference, and prediction," pp. 106–119, 2009.

[20] ——, "The elements of statistical learning: data mining, inference, and prediction," pp. 119–127, 2009.

[21] ——, "The elements of statistical learning: data mining, inference, and prediction," pp. 417–438, 2009.

[22] ——, "The elements of statistical learning: data mining, inference, and prediction," pp. 463–471, 2009.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] H. Trevor, T. Robert, and F. JH, "The elements of statistical learning: data mining, inference, and prediction," pp. 241–247, 2009.