Prediction Accuracy and Sensitivity Analysis on Different Models for Breast Cancer Wisconsin Dataset

Shiyu Liu
shiyu_liu@brown.edu
Data 1030 | Fall 2019
GitHub URL: github.com/shiyuliu1/data1030_project.git

1 Introduction

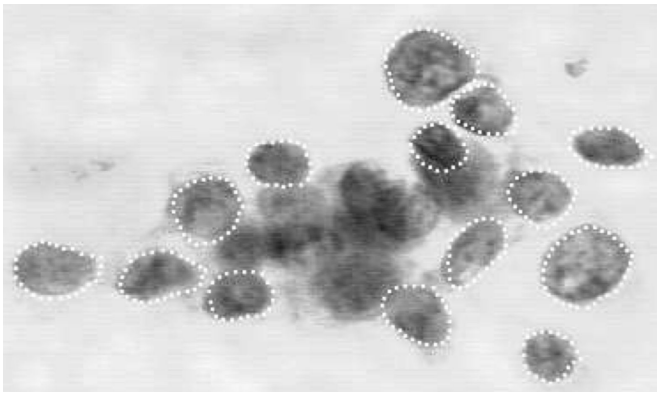

Figure 1. Cells obtained through breast fine needle aspirate. The boundaries of the cells were outlined.

Forecasting breast cancer can significantly increase the survival rate of patients, and classifying the sample tumor cells as malignant or benign is one of the best and most direct ways to make accurate predictions. Breast Cancer Wisconsin from UCI Machine Learning Repository was chosen as the dataset to implement machine models and predict diagnosis result on practical cases. The target variable in this dataset is the final diagnosis whether the tumor is malignant or benign. There are 30 real-valued columns as independent variables with three properties, mean, standard error and worst cases, for 10 cytological features. Such feature values are processed and computed based on the image of the tissue cells, which were obtained using Fine needle aspirations (FNAs) (Street 1993). Figure 1 shows the cells that were obtained through fine needle aspirate, and the boundaries of the cells were outlined by the contour model called "snake" (Street 1993).

Previous studies have focused on how logistic regression and inductive machine learning such as decision trees could achieve the classification of the diagnosis (Mangasarian 1995). In this study, LASSO regression, support vector machine, random forest, gradient boosting algorithms and Gaussian naive bayes classifier, were applied to this classification problem. Gradient boosting algorithm include XGBoost and CatBoost. After one of the models is selected, GridsearchCV was used for tuning the model hyperparameters, and the percentage of correctly classified observations was recorded as the accuracy. In addition to comparing the accuracy score, an averaged confusion matrix for each method was generated. This estimates the models' misclassification property across the two classes.

2 Exploratory Data Analysis

2.1 Target Variable

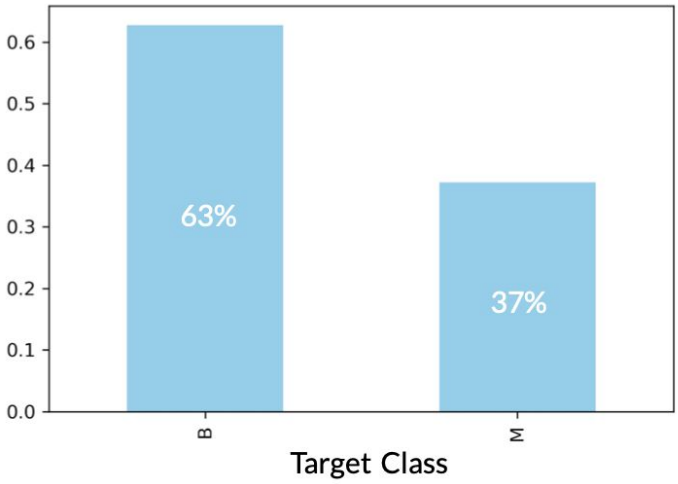

Figure2: The dataset is slightly imbalanced. 63% of the observations are in class 0 and the rest 27% are in class 1. Baseline model accuracy is 0.63.

Among the 569 cell observations, 357 of them are benign cells and the other 212 are malignant cells. Figure 2 shows the percentage of each class. This is regarded as a slightly imbalanced dataset and stratified train test split should be applied for training and testing later. The target values were then labeled as 0 for benign and 1 for malignant.

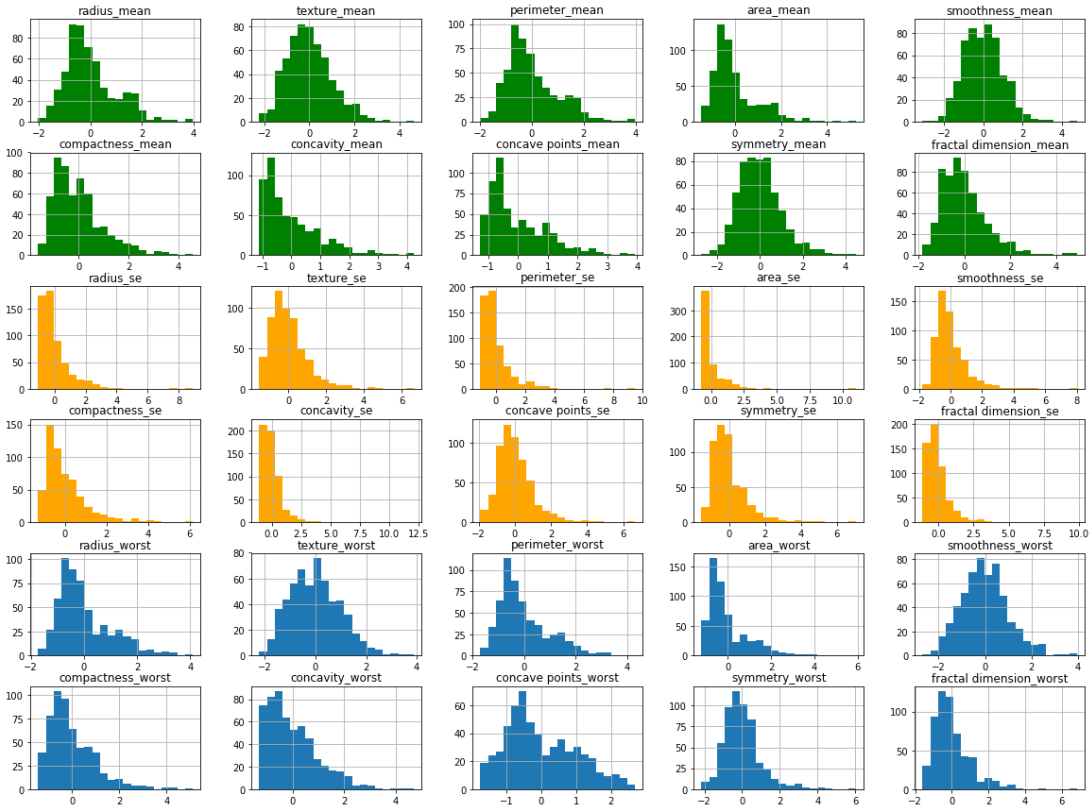

Figure 3: The histogram of each feature measurements. The top two rows show the mean of feature measurements; the middle two rows show the standard error of feature measurements; the bottom two rows show the worst cases of feature measurements.

2.2 Features

The independent variables are all numerical. As shown in figure 3, it can be observed from the histogram of each feature that there is no apparent outlier. Since there is no upper or lower bound set for the feature values, standard scaling shall be applied to such numerical variables. The data preprocessing process used standard scaling inside the machine learning pipeline.

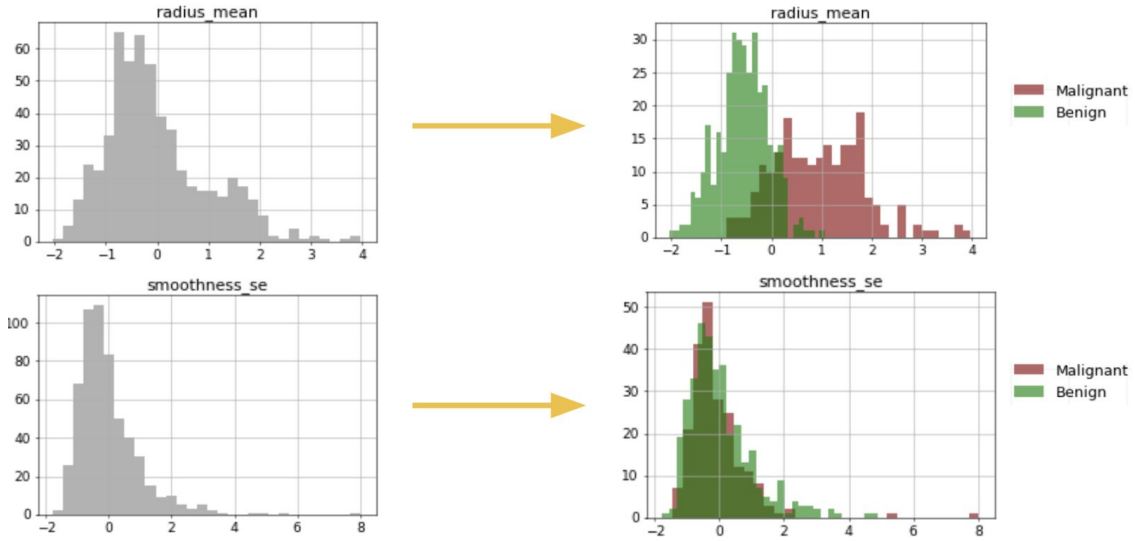

Figure 4: A demonstration that different features can have different ability in classifying the target variable. Radius mean is much more correlated with final diagnosis than smoothness standard error.

Among the 30 features, some features are more correlated to the target variable, while others hardly provide information predicting the target variable. Figure 4 selected two representative features to show the two features can have huge correlation differences with the target variable. A better method relationship between features and target variable is to quantitatively describe their pairwise dependency. In this exploratory process, correlation, F score and the mutual information score were calculated between features and target. Figure 5 describe the three score tendency for the features after ordering them based on their correlation coefficients with diagnosis. It can be observed that both mutual information and F score have a tendency of decreasing. Generally, the first few features are more informative than the last few features.

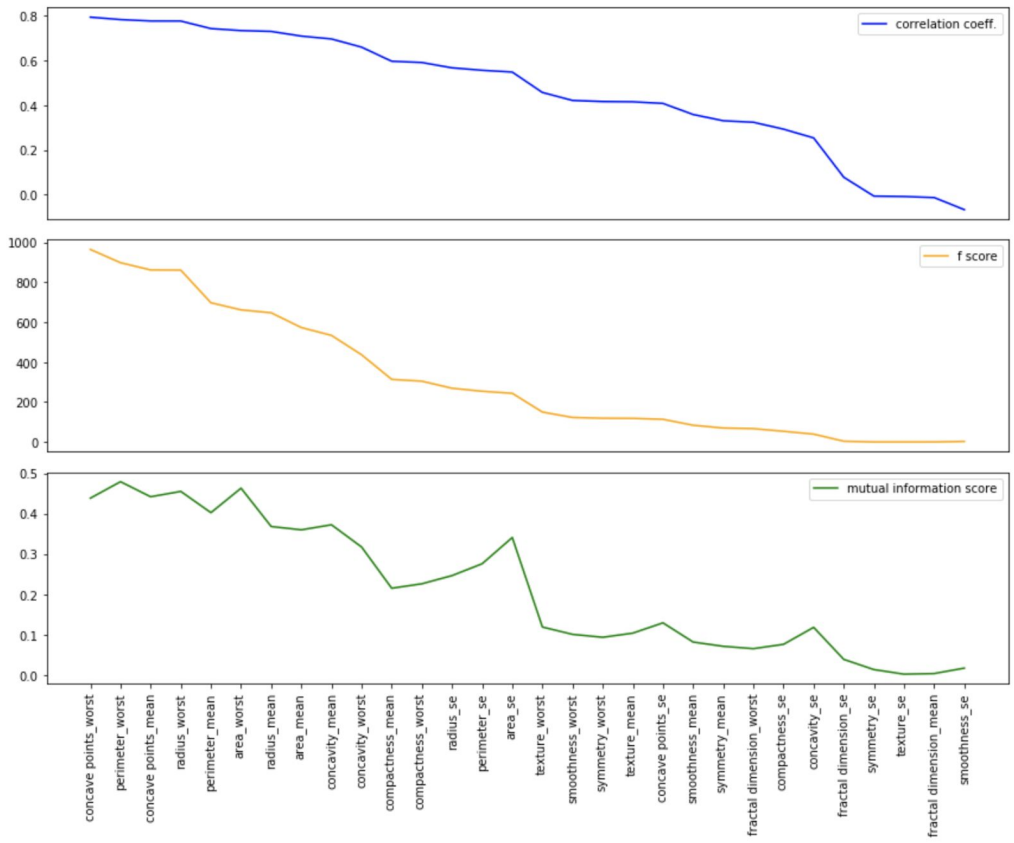

Figure 5: Correlation coefficients, mutual information and F score have similar tendency for the ordered features, although there are some fluctuations in the middle for mutual information score.

3 Methods

3.1 Pipeline

A machine learning pipeline was first developed to facilitate the workflows when applying different machine learning algorithms. Before using the pipeline, the dataset was stratified into train/cv data and test data by label. Specifically, 80 percent of the data was for training and the other 20 percent data was for testing. This aimed to solve the problem that this dataset is slightly imbalanced. The pipeline consists of two parts, the preprocessor based on column transformer and the classifier model. As the input of the pipeline, the train/cv data will be splitted into train data and cross validation data to find the best parameters. A five-fold design was implemented. To prevent data leakage, the standard scaling in the preprocessor was fit and transformed based on train data only. Then the cv and test data was transformed based on this scaler. The best parameter combination was selected based on the highest average cross validation score. The prediction score can be calculated by predicting the test data through the best estimator returned by the pipeline. For the evaluation metric, the accuracy score was selected because the percentage of correct classification is intuitive to evaluate the classification models.

3.2 Models and Parameter Tuning

Table 1. Hyperparameters Tuned and Ranges

Methods	Parameter 1	Parameter 1 Range	Parameter 2	Parameter 2 Range	Parameter 3	Parameter 3 Range
Lasso Logistic Regression	Alpha	logspace(-2,2,20)	N/A	N/A	N/A	N/A
Random Forest	Max Depth	2,3,4,5,6,8,10,12	Max Feature	1,3,5,7,9,11,13, 18,20,25,30	N/A	N/A
Support Vector Machine	Gamma	logspace(-5,0,20)	C	logspace(-3,4,20)	N/A	N/A
XGBoost	n_estimators	100, 200, 600, 1000	Learning Rate	0.01, 0.05, 0.1, 0.2, 0.3, 0.5	Max Depth	2, 3, 5, 6, 8, 10
CatBoost	Iterations	600	Learning Rate	0.05, 0.1	Depth	5, 6
Naïve Bayes Classifier	N/A	N/A	N/A	N/A	N/A	N/A

The pipeline and data were fitted by six mahcine learning algorithms, logistic regression, support vector machine, random forest, XGBoost, CatBoost and Gaussian naive bayes classifier. The hyperparameters tuned and their range are shown in Table 1. The parameters range were based on the empirical tests for the specific model so that the parameters of the best estimator are not at the edge of their corresponding ranges.

Limited combinations were put inside the pipeline because the running time is also a factor. For each algorithm, same ten different random states were used to conduct the train-test split to test the algorithm's performance stability and avoid the uncertainties when splitting data due to the random state. For each random state, the pipeline and its grid search function returned the best test score as well as the model parameters. Aftering obtaining the ten best test scores, we used the mean of the ten scores to estimate the overall performance of the model.

3.3 Algorithm Stability(uncertainty), Sensitivity and Specificity

To investigate whether the algorithms provide stable performance, standard deviations for the ten test scores were calculated to estimate the level of score fluctuations. In addition to the overall test score, an overall confusion matrix is obtained by summing up the corresponding categories of the confusion matrix for each model based on the random state. The confusion matrix is then normalized by row to obtain detection rate, miss rate, false alarm rate and true negative rate. While detection rate represents the sensitivity on detecting cancer cells of a model, true negative rate represents the specificity that the model can successfully classify patients not having cancer cells.

4 Results

4.1 Algorithm Performance

Table 2. Mean and Standard Deviation of Model Test Scores & Comparison to baseline

Methods	Mean Test Score	Test Score Standard Deviation	No. std. above Baseline
Lasso Logistic Regression	0.966	0.017	19.3
Random Forest	0.952	0.027	11.6
Support Vector Machine	0.972	0.016	20.9
XGBoost	0.961	0.022	14.7
CatBoost	0.970	0.014	23.7
Naïve Bayes Classifier	0.934	0.014	21.1

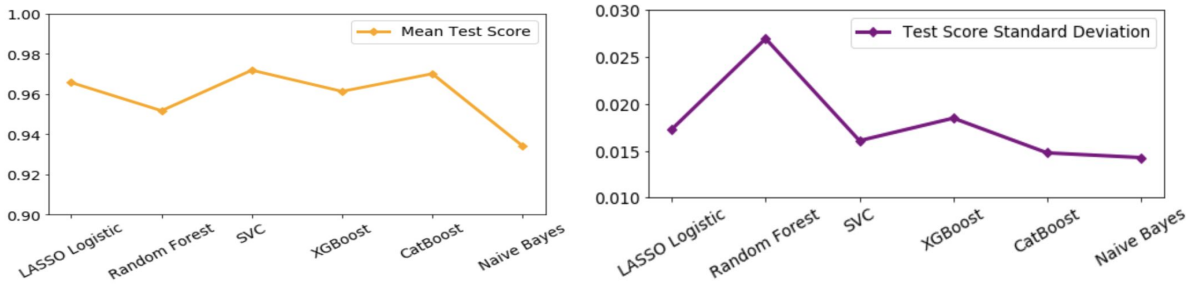

Figure 6. The left panel shows the mean test score comparison across the six models. The right panel shows the test score standard deviation comparison.

Table 2 shows the mean and standard deviation of the test scores for each model. Figure 6 compares the two quantities across the methods. These results show that the support vector machine classifier has the best mean score and the lowest standard deviation. This indicates SVC algorithm performs best and second most stable in classifying the cancer cell types for this dataset. For random forest algorithms, the two gradient-boosting-based algorithms perform much better than random forest classifier. Such improvement reveals that correcting the errors of weak learners can effectively reduce the bias during training for this dataset. Comparing to XGBoost, CatBoost algorithm is slightly higher in accuracy and lower in standard deviation, resulting in the largest number of standard deviations above baseline. Logistic regression also performs well. Moreover, although having a relatively low standard deviation, the prediction accuracy is relatively low compared to other methods.

4.2 Sensitivity and Specificity

Table 3. Summary of Confusion Matrix

Methods	Detection Rate	False Alarm Rate	Miss Rate	True Negative Rate
Lasso Logistic Regression	0.938	0.018	0.062	0.982
Random Forest	0.92	0.029	0.081	0.971
Support Vector Machine	0.948	0.014	0.052	0.986
XGBoost	0.933	0.022	0.067	0.978
CatBoost	0.940	0.013	0.060	0.988
Naïve Bayes Classifier	0.889	0.039	0.112	0.961

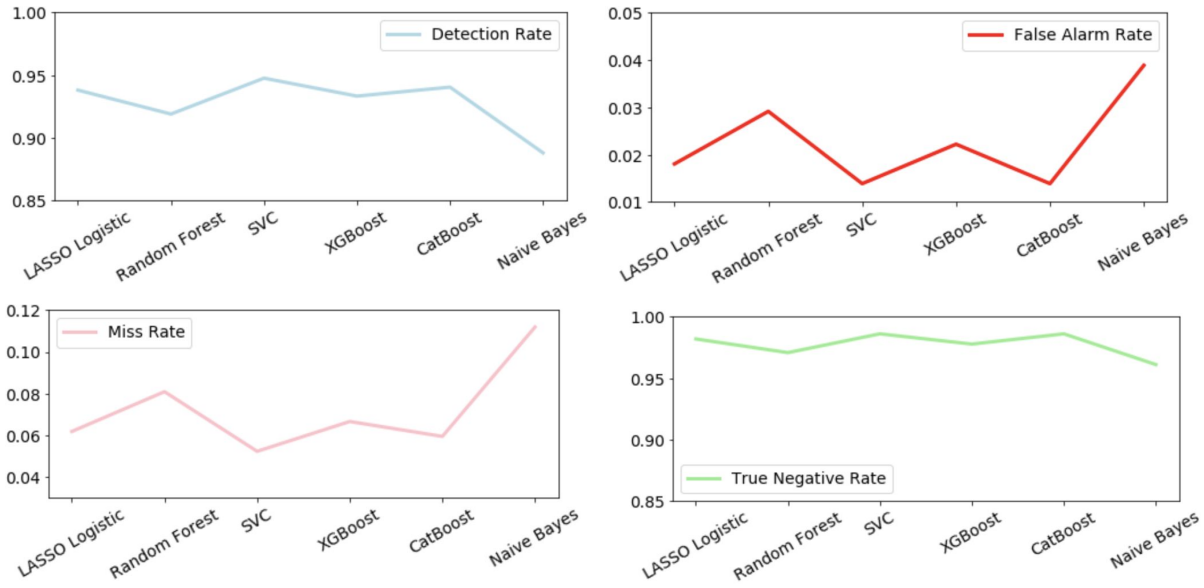

Figure 7. The values of the four categories in confusion matrix are extracted and compared across models. Being consistent with the precision accuracy, SVC has the highest sensitivity and specificity, and the lowest misclassification rate.

Table 3 shows the confusion matrix values for each matrix and each model. The comparisons of detection rate, false alarm rate, miss rate and true negative rate between models is shown in Figure 7. Generally, a model with higher accuracy is higher in detection rate and true negative rate, lower in false alarm rate and miss rate. Similar to test accuracy, SVC is the highest in detection rate and true negative rate, indicating its highest sensitivity and specificity.

4.3 Local Feature Importance

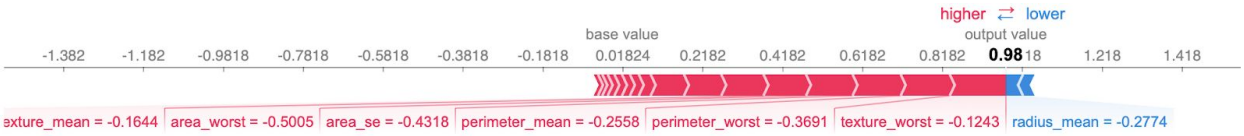

Figure 8. Force plot for a single point based on SVC model. The red bars length shows that the most important three features are texture worst, perimeter worst and perimeter mean.

Since the support vector machine classifier has the best performance on this dataset, the point with index 20 was selected to conduct the local feature importance investigation using SVC model. Figure 8 shows the force plot for classifying this point to class 0. SVC model classified this point right, and the most important three features are texture worst, perimeter worst and perimeter mean.

4.4 Practical Case Discussion

When classifying tumor cells as malignant or benign in practical cases given the tissue cell cytological features, the result of this study would suggest using the support vector machine classifier due to its high accuracy and low variance. Another important practical concern is that SVC has the lowest miss rate.

We hope to make the miss rate as low as possible because delaying the treatment might result in disastrous health issues.

5 Outlook

In this study, all the machine learning algorithms perform well for the Breast Cancer Wisconsin dataset. The misclassified points are likely to be outliers, which are cells having small features values but actually in malignant group, and vice versa. More investigation could have done on investigating which features can be removed but remaining similar accuracy performance. Enlightened by the feature engineering part in EDA, we can reasonably delete the last few features which have nearly no mutual information and dependency with the target variable. Ruling our irrelevant features and using the reduced dataset with less features can significantly accelerate the training and cross validation process. It also simplify the feature data preparation process when generating them from the computer vision system.

In addition to the machine learning algorithms used in this study, neural networks might have similar or better performance. Fogel et al. obtained high correctness using evolutionary artificial neural networks with two and nine hidden nodes (Fogel 1996). A EANN with pareto-differential evolution (PDE) algorithm performed slightly better with low standard deviation (Abbass 2002).

Future studies might also benefit from investigating the cell property that have been misclassified by most of the methods. Such cancer cells have special property that their type cannot be determined by the current cytological features. It is possible that there are other features that are necessary to be collected so that the types can be correctly classified using the new features information. To prevent offering unnecessary treatments for healthy patients and delaying treatments for patients with malignant tumors, other cancer screening techniques need to be combined with this biopsy exams to ensure accurate results.

References:

Abbass, Hussein A. An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artificial Intelligence in Medicine*, 25. 2002.

Fogel, David B., et al. “Evolving Neural Networks for Detecting Breast Cancer.” *Cancer Letters, vol. 96, no. 1*, 1995, pp. 49–53.

Mangasarian, O.L., W.N. Street and W.H. Wolberg. *Breast cancer diagnosis and prognosis via linear programming*. Operations Research, 43(4), pages 570-577, July-August 1995.

Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. *IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905*, pages 861-870, San Jose, CA, 1993.