Classification of Colon Cancer Data

Ting Sheng Tan
S1975313
s1975313@ed.ac.uk

Shuai Pan
S1925253
s1925253@ed.ac.uk

Jialu Wang
S1972355
s1972355@ed.ac.uk

Jiale Peng
S1928189
s1928189@ed.ac.uk

Abstract

Gene expression significantly aids in the development of efficient cancer diagnosis and classification platforms. In this report, we examine one set of gene expression data measured across sets of tumor(s) and normal clinical samples: it consists of 2,000 genes, measured in 62 epithelial colon samples. We aim to classify the samples as cancerous or not, in which the separation of tissue type is measured using individual gene expression levels. We use several feature selection methods to reduce the dimension and select useful features. We also utilize a few classification methods to assess the classification power of completing expression profiles. Leave one out cross-validation (LOOCV) is performed on training set, and three robust classifiers are selected to classify test set samples as cancerous or not. Lastly, we demonstrate a success rate of at least 69% in tumor versus normal classification, using sets of selected genes.

1 Introduction

Gene expression data can be useful in understanding of cancer. Normal cells can evolve into malignant cancer cells through a series of mutations in genes that control the cell cycle, apoptosis, and genome integrity. As determination of cancer type and stage is often crucial to the assignment of appropriate treatment[1], a central goal of the analysis of gene expression data is to classify samples as cancerous or not and predict the whether the sample is cancerous or not given its gene expression data. The colon data consists of expression level measurements of two thousands of genes, measured in 62 samples. We divided the colon data into two groups based on its label, one containing tumor samples, and the other containing normal tissue samples. We perform exploratory data analysis to get familiar with the data. This efficiently aid in extracting relevant biological information that prepare us for data mining.

The shape of colon data is quite challenging, as the dimensionality of the feature space is very high compared to the number of cases. So it is important to find small sets of genes that are sufficiently informative to distinguish between cells of different types for classification. We present four feature selection methods based on gene expression data to select informative genes. Some of these features are not relevant to the distinction between normal and tumor and introduce noise in the classification process, it is called data contamination. To realistically assess the performance of such methods one needs to address the issue of sample contamination. In the data, the authors[2] observed that the normal colon biopsy also included smooth muscle tissue from the colon walls. As a result, smooth muscle-related genes showed high expression levels in the normal samples compared to the tumor samples. This artifact, if consistent, could contribute to success in classification. But whether delete these data or not do not have any effects for the classification result [3], so we still include these data in our analysis.

We utilize several classification approaches on training set and use LOOCV and Bayesian optimization to tune the hyper-parameters. Then we evaluate the effect of gene selection on the classification methods and select three robust classifiers. Finally, we apply these classifiers to test set to show the classifiers' generalization performance.

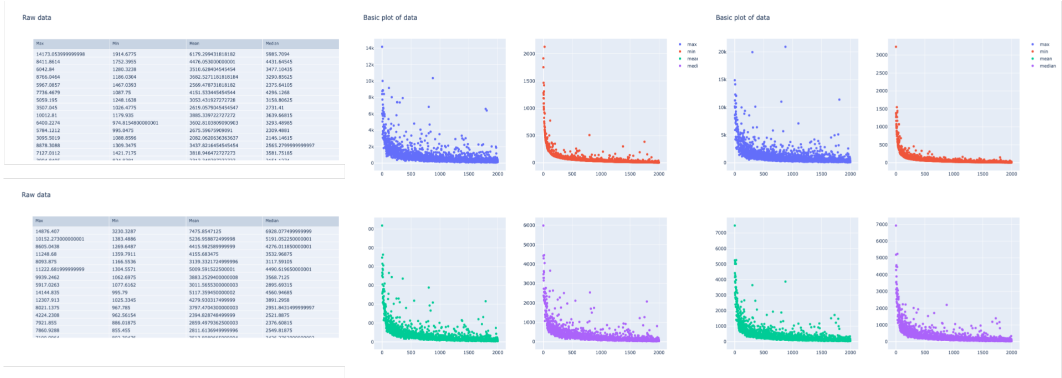
2 Descriptive Analysis

2.1 Descriptions of the Dataset

This dataset is a collection of expression measurements from colon biopsy samples reported by [2]. The dataset that were collected from colon-cancer patients. The "tumor" biopsies were collected from tumors, and the "normal" biopsies were collected from healthy parts of the colons of the same patients. The final assignments of the status of biopsy samples were made by pathological examination. Gene expression levels in these 62 samples were measured using high density oligonucleotide arrays. Of the genes represented in these arrays, 2000 genes were selected based on the confidence in the measured expression levels.

2.2 Basic Acknowledgement of the Data

Through *describe*(), we get a basic understanding of the data of different categories (normal tissue, tumor tissue), and extract the maximum, minimum, average, and median of each feature. *plotly* library is used to make them into the following graphics (Figure 1).



Raw data

Max	Min	Mean	Median
14173.053999999998	1914.6775	6179.299431818182	5985.7094
8411.8614	1752.3955	4476.053000000001	4431.64545
6042.84	1280.3238	3510.628404545454	3477.10435
8766.0464	1186.0304	3682.5271181818184	3290.85625
5967.0857	1467.0393	2569.478731818182	2375.64105
7736.4679	1087.75	4151.533445454544	4296.1268
5059.195	1248.1638	3053.431927272728	3158.80625
3507.045	1026.4775	2619.0579045454547	2731.41
10012.81	1179.935	3885.339722727272	3639.66815
6400.2274	974.8154800000001	3602.8103809090903	3293.48985
5784.1212	995.0475	2675.59675909091	2309.4881
3095.5019	1088.8596	2082.0620636363637	2146.14615
8878.3088	1309.3475	3437.8216454545454	2565.2799999999997
7127.0112	1421.7175	3818.946472727273	3581.75185

Raw data

Max	Min	Mean	Median
14876.407	3230.3287	7475.8547125	6928.077499999999
10152.273000000001	1383.4886	5236.958872499998	5191.052250000001
8605.0438	1269.6487	4415.982589999999	4276.011850000001
11248.68	1359.7911	4155.683475	3532.96875
8093.875	1166.5536	3139.3321724999996	3117.59105
11222.681999999999	1304.5571	5009.591522500001	4490.619650000001
9939.2462	1062.6975	3883.2529400000008	3568.7125
5917.0263	1077.6162	3011.5655300000003	2895.69315
14144.835	995.79	5117.359450000002	4560.94685
12307.913	1025.3345	4279.930317499999	3891.2958
8021.1375	967.785	3797.4704300000003	2951.8431499999997
4224.2308	962.56154	2394.828748499999	2521.8875
7921.855	886.01875	2859.4979362500003	2376.60815
7860.9288	855.455	2811.6136949999996	2549.81875

Figure 1: The upper table and the first scatter plot group are normal tissues. The table below and the second scatter plot group are tumor tissues.

It can be concluded from the figure that the interval size of each feature is generally consistent under different categories. For example, the feature data in the front is generally large, while the feature data in the back is generally small.

2.3 Data Distribution

In Figure 2, we only plot the data distribution from random 20 features, and we can see that whether it is tumor or normal gene data. It generally shows a right-biased distribution.

Therefore, for the genetic data under different labels, the skewness is measured. Here skewness measurement is measured using sample skewness and Galton's measure of skewness. We count the total amount whose result is greater than 0 and summarize it. The statistical results are showed in Table 1.

SS_Normal	*SS_Tumor*	*GMS_Normal*	*GMS_Tumor*
1872	2000	1241	1643

Table 1: Skewness summary

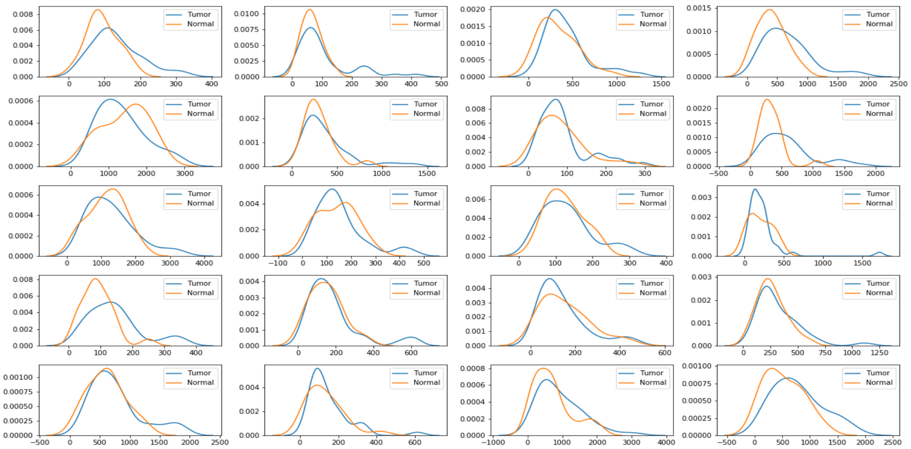


Figure 2: Kde plot about the random 20 features.

When using different labels to measure the sample skewness, it can be found that in the calculation of normal genetic data, a total of 1827 features have a right-skewed distribution. In tumor gene data, all of them are right-skewed. Right-skewed distribution means that there is a particularly large outlier, which causes its distribution to be right-skewed. In addition, under Galton's measure of skewness, all the genetic data, whether normal or tumor, has more than half of the data showing a right-skewed distribution. So next we explore the outlier situation in the data.

2.4 Outlier Detection

We use the absolute median deviation (AMD) to simply find the outlier. The calculated AMD of each column is compared. If it is found that the value of a certain data in a process similar to the "standardization" above is greater than the AMD, it is considered that there is an outlier in this column of data. Finally, the column data of the outlier will be returned. The following box diagram is obtained (Figure 3).

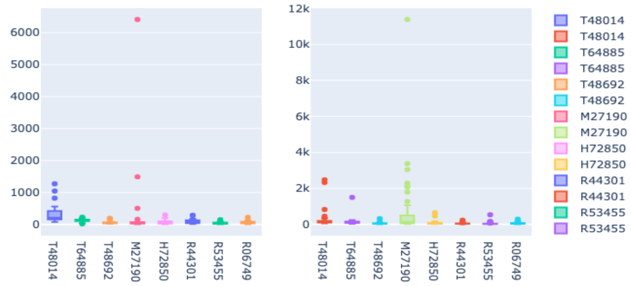


Figure 3: Outlier found by AMD. In the left figure genes are from normal samples, in the right figure, the genes are from tumor samples.

In the left Figure 3 the genes are normal label, comparing with the T48014 and M27190, which the outlier can easily find in the figure, other genetic data's outliers are not as obvious as them. In the right Figure 3, the gene with tumor label, the outlier of these genetic data is relatively obvious. It is initially believed that these data may interfere with the final result.

However, in the subsequent analysis, we will not artificially remove the outlier, as it is known from the literature [2] that these 2000 genes have been selected from 6,500 genes with high discrimination.

2.5 Correlation Analysis

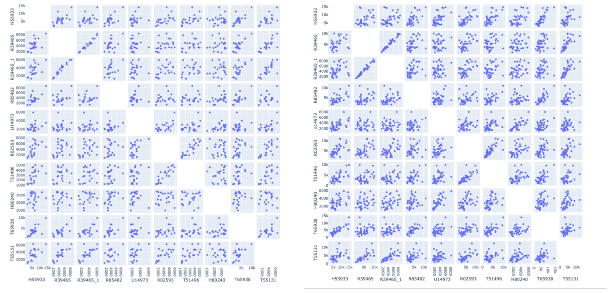


Figure 4: First ten features' scatter plot.

Only the first ten variables with different labels are selected (Figure 4). The normal label is on the left, and the tumor label is on the right. The scatter plot is drawn between the two in order to look at the variables relevance. It can be seen that the correlation of the scatter plots between different labels may have the same relationships. For example, in the first and second features where the label is normal, it is roughly a relationship, and the features where the label is tumor are the same.

The Figure 5 is the statistics of each variable whose correlation with other variables is greater than 0.8 (not repeated measurement). It is almost a solid triangle, so it shows that most of the variables are actually related. This is also conducive to the following data if it involves dimension reduction, providing a good evidence.

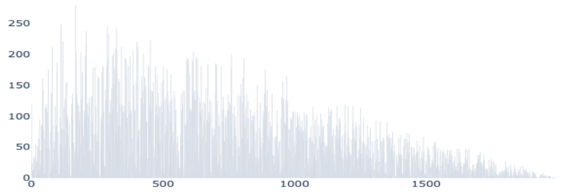


Figure 5: Summarize the total number of correlated data.

3 Feature Selection

In this section, we implement four feature selection methods that include F-test, MFA score, MFA+ score and TNoM score. Each methods explore the data in different perspective: F-test focuses on feature deviation; MFA score focuses on feature separability between classes and within classes; MFA+ score focus on reducing redundancy based on MFA; and TNoM score focuses on feature performance. The detail of each feature selection method can be seen below.

3.1 F-Test

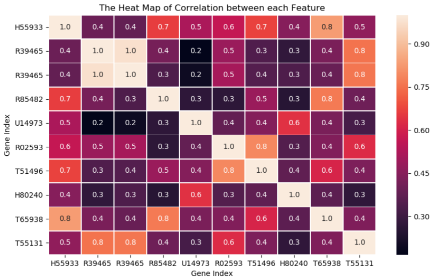
F-test is a statistical test in which the test statistics has F distribution under the null hypothesis and F statistics can be seen as kind of the extension of the T statistics. Here, we use One-Way ANOVA method to do the feature selection. This is a technique that can be used to analysis the differences between group means among a sample. The One-Way ANOVA tests the null hypothesis that 2 or more features have the same population mean. It is often used in the analysis of data and can help to select the significant feature based on P-value. ANOVA is a robust technique which assumes all sample of a data to be distributed in general, having equal variance and independent. We test the statistical significance of each feature by comparing their F test statistic to the F distribution and get the p-value. The smaller p-value means the more significant of that gene and we will choose the top significant genes as our features.

3.2 MFA Score

The Marginal Fisher analysis (MFA Score) is used to keep the local and global geometric properties of sample data. It doesn't need to take the data distribution into account and so can be applied to any data. The MFA Score use the formation of the neighbors between-class and within-class [4]. The bigger score represent the gene can gathers the same class samples and separates the different class samples better and so the gene is more significant. The details can been seen in article. We will choose the genes with bigger MFA Score as our featrues.

3.3 MFA+ Score

MFA+ Score consider both MFA Score and the redundancy of each gene. From figure 6, we can find there are many gene pairs selected by MFA Score are highly correlated (e.g. correlations are greater than 0.7). Assuming that two highly correlated genes have the same contribution to classification performance and exclude one redundancy gene will not have side-effect to the total procedure. The advantage of excluding redundancy genes is that, with only small number of features, the new genes added into feature set with lower redundancy may produce better classification accuracy [4]. Here, we use Pearson correlation coefficient as our evaluation method. To get the MFA+ Score, we first get the feature set in descending order of MFA Score, then exclude all low score genes with correlation of any high score genes greater than threshold σ. After that we get the feature by MFA+ Score method, we can choose the genes with higher MFA Score and lower redundancy as our features.



	H55933	R39465	R39465	R85482	U14973	R02593	T51496	H80240	T65938	T55131
H55933	1.0	0.4	0.4	0.7	0.5	0.6	0.7	0.4	0.8	0.5
R39465	0.4	1.0	1.0	0.4	0.2	0.5	0.3	0.3	0.4	0.8
R39465	0.4	1.0	1.0	0.3	0.2	0.5	0.4	0.3	0.4	0.8
R85482	0.7	0.4	0.3	1.0	0.3	0.3	0.5	0.3	0.8	0.4
U14973	0.5	0.2	0.2	0.3	1.0	0.4	0.4	0.6	0.4	0.3
R02593	0.6	0.5	0.5	0.3	0.4	1.0	0.8	0.3	0.4	0.6
T51496	0.7	0.3	0.4	0.5	0.4	0.8	1.0	0.4	0.6	0.4
H80240	0.4	0.3	0.3	0.3	0.6	0.3	0.4	1.0	0.4	0.3
T65938	0.8	0.4	0.4	0.8	0.4	0.4	0.6	0.4	1.0	0.4
T55131	0.5	0.8	0.8	0.4	0.3	0.6	0.4	0.3	0.4	1.0

Figure 6: Heatmap of pearson correlation between feature pairs with top MFA Score.

3.4 TNoM Score

The *threshold number of misclassification* which is also called TNoM Score is a measurement to compare the performance of each gene individually by weak learning algorithm. The intuition is that an informative gene should be able to separate different classes (normal and tumor) by a threshold value[5]. The TNoM Score of each gene is simply find the minimum number of misclassification of a decision stump by searching over all possible threshold and direction. The goal of TNoM Score is to find the gene with higher quality decisions and higher expression level. We will choose the genes with smaller TNoM Score as our features.

4 Classification Methods

The colon cancer dataset has a small number of samples. Besides, the dataset is unbalanced, in which there are 40 tumor tissue samples and 22 normal tissue samples. Although the number of samples is very small, we split the dataset with a 80/20 ratio. There are 49 samples in the train set and 13 samples in the test set. The test set is held out from the model training process, and is only used once in the end for evaluating the classifier generalisation performance. In this project, five types of classifiers will be trained: Linear Support Vector Machine (SVM), Gaussian Naive Bayes classifier, KNN classifier, Logistic Regression Classifier, and Multi-layer Perceptron classifier.

4.1 Baseline Classification

We use the simplest classifier as baseline to classify samples as the most frequent class in the training set. By LOOCV evaluating the baseline validation performance, the accuracy and log-loss of this classifier is 0.6531 and 11.983 respectively. We also used another classifier which make uniformly random predictions. This classifier yield a lower accuracy score on LOOCV. The accuracy and log-loss of this classifier is 0.3469 and 22.556 respectively.

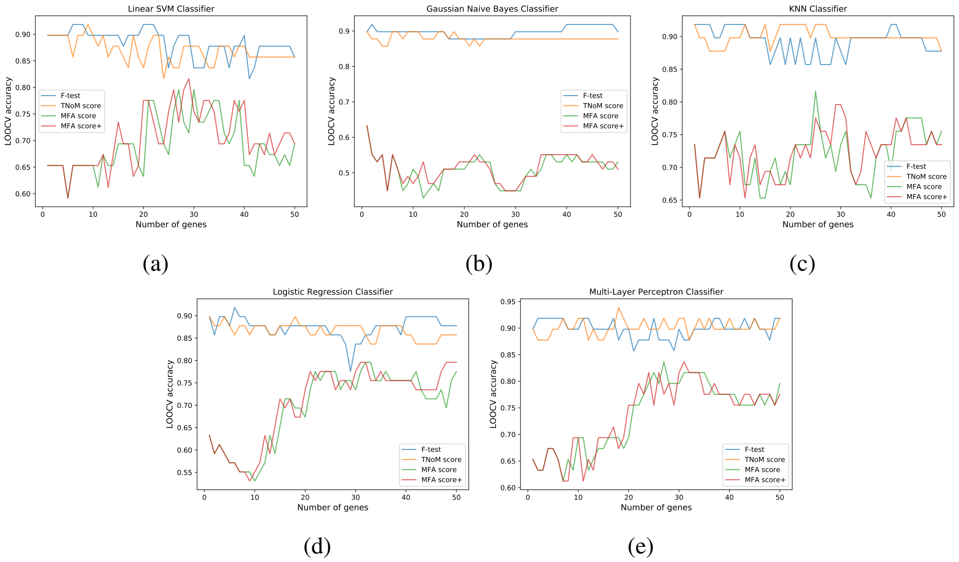
4.2 Classifier Training

The four types of feature selection methods (F-test, TNoM, MFA, MFA+) discussed in section 3 are used to select the most predictive genes when training our models, in order to avoid overfitting. When training each type of classifier, all the four feature selection methods are being considered separately. In addition, each feature selection method is used 50 times on each type of classifier. To be more precise, the number of most predictive genes selected by each method will be increment by one in each iteration until the top 50 predictive genes are all being considered.

Due to the small amount of train samples, LOOCV is used to evaluate the classifier validation performance. We also use Bayesian optimisation to find the optimum hyper-parameters for each classifier in each combination of feature selection method and number of genes selected.

4.3 Classifier Comparison

Figure 7 illustrates the accuracy of all the combination of five classifiers and four feature selection method. It shows that MFA score and MFA score+ have poor performance for all the classifiers in LOOCV evaluation compare to F-test and TNoM. So we do not use these two feature selection methods.



(a) (b) (c)

(d) (e)

Figure 7: The LOOCV accuracy for each classifier with four feature selection method based on different number of gene features.

We select six combinations of feature selection method and classifier with their optimal hyper-parameters which have high accuracy. The summary of these combination is in Table 2, which contain their accuracy, corresponding genes and hyper-parameters. In Table 2, F-test method shows good performance in these combinations. But the highest accuracy for the selected colon data is classified by Multi-layer Perceptron(TNoM) with 18 gene features which is 0.9388. However, for other combinations, the highest accuracy is 0.9184.

We also notice that M63391 is selected in all the combinations which have high accuracy. Especially in KNN(F-test) and KNN(TNoM), they only use M63391 that can achieve the high accuracy up to 0.9138. So it shows that M63391 contains most useful information and play a important role in the classification.

We need to select three more robust classifiers from the six combinations. KNN is sensitive to irrelative attributes, as these attributes will effect the distance between features. In the raw colon data, there are 2000 features, and KNN is sensitive to feature selection. So it just use one feature. It is not robust. The highest accuracy of KNN in LOOCV is less than Multi-layer Perceptron(TNoM). So we will abandon KNN(F-test) and KNN(TNoM). In the high dimensional colon dataset, Naive Bayes assumes each gene feature to be independent with each other. This normally is not a realistic

Classifier (Method)	Selected Genes (Count)	LOOCV Accuracy	Optimum Hyperparameters
Linear SVM (F-test)	M63391, J02854, M76378, M76378, T92451, M76378 (6)	0.9184	C = 446.2, gamma = 8.734
Gaussian Naive Bayes (F-test)	M63391, J02854 (2)	0.9184	var_smoothing = 4.462×10^{-7}
KNN (F-test)	M63391 (1)	0.9184	n_neighbors = 15, leaf_size = 13, p = 1
KNN (TNoM)	M63391 (1)	0.9184	n_neighbors = 15, leaf_size = 13, p = 1
Logistic regression (F-test)	M63391, J02854, M76378, M76378, T92451, M76378 (6)	0.9184	C = 87.35
Multi-layer perceptron (TNoM)	M63391, T71025, M76378, M76378, R87126, J02854, M26383, M76378, T92451, U25138, U19969, T60155, X86693, M36634, J05032, T60778, R78934, H64489 (18)	0.9388	hidden_layer_size = 16, alpha = 0.9998

Table 2: Summary of six combinations and corresponding optimum hyperparameters.

assumption in gene dataset. And the accuracy of Naive Bayes does not exceed Multi-layer Perceptron, so we abandon it too. Hence, we keep Linear SVM(F-test), Logistic regression(F-test) and Multi-layer Perceptron(TNoM), and apply them to test data to see the classifiers' generalization performance.

5 Results

We use the selected classifiers from previous section to test the previously held-out test set and the test results are in Table 3.

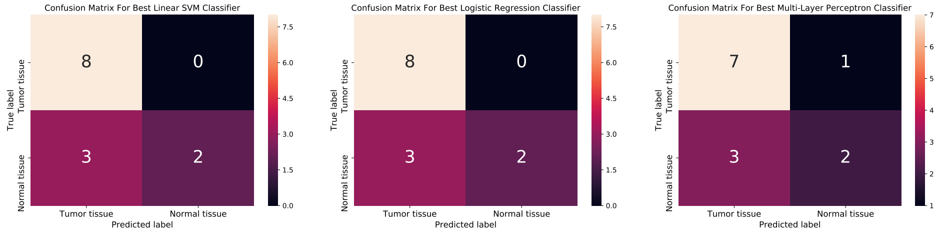
As shown in Table 3, both linear SVM and logistic regression classifiers have the highest test classification accuracy of 76.92%, while multi-layer perceptron classifier only has a test classification accuracy of 69.23%. On the other hand, linear SVM model has the lowest test log-loss of 0.5776 when compared with both logistic regression and multi-layer perceptron models. Linear SVM classifier performs the best in both test classification accuracy and test log-loss. We think that this might be explained by the use of slack variables and soft margin in SVM that allows linear SVM classifier to generalise better.

Apart from that, precision, recall, and F1 scores are also calculated for the models. Precision, recall, and F1 scores are important evaluation metrics for our classifiers because these metrics are related with the true positive rate, which is the proportion of actual tumor tissue samples that are correctly identified. In fact, it is more vital to correctly classify tumor tissue samples than to correctly classify normal tissue samples, because in medical use cases, misidentifying a tumor tissue sample as a normal tissue sample can bring deadly consequences. Among the three models, both linear SVM and logistic regression classifiers have the same precision, recall, and F1 scores of 0.7273, 1.000, and 0.8421 respectively. They perform better than the multi-layer perceptron classifier which only has precision, recall, and F1 scores of 0.7000, 0.8750, and 0.7778 respectively.

Classifier (Method)	Accuracy	Log-Loss	Precision	Recall	F1
Linear SVM (F-test)	0.7692	0.5776	0.7273	1.000	0.8421
Logistic regression (F-test)	0.7692	0.7375	0.7273	1.000	0.8421
Multi-layer perceptron (TNoM)	0.6923	0.6305	0.7000	0.8750	0.7778

Table 3: Summary of test results for the three robust classifier

The precision, recall, and F1 scores can again be verified in the classifiers' confusion matrices, as shown in Figure 8. In general, we can see that all three models are able to correctly identify almost all tumor tissue samples correctly, except one tumor tissue sample that has been misclassified by the multi-layer perceptron classifier as a normal tissue. However, all the three models do not perform well in identifying normal tissue. Out of the five normal tissue samples in test set, all the models have misidentified three normal tissue samples as tumor tissues. We think that this might be explained by the unbalanced nature of the colon cancer dataset, in which we have more tumor tissue samples than normal tissue samples.



(a) (b) (c)

Figure 8: Confusion matrix: (a) Linear SVM(F-test); (b) Logistic regression(F-test); (c) Multi-layer perceptron(TNoM).

6 Conclusion

Given colon data, which has 62 samples but with 2000 features, we tried four feature selection method and tried five classifiers. Each of this combination has corresponding number of features and hyper-parameters, we use LOOCV and Bayesian optimization to find the optimal hyper-parameters. In this process we found F-test and TNoM have good performance in the LOOCV evaluation, but MFA and MFA score+ have bad accuracy in all the five classifiers. We selected six classifier-method combinations based on each classifier-method's LOOCV accuracy. In the corresponding gene selection of the six classifier-method combinations, we found that M63391 has been selected in all the combinations. It shows that this gene is highly correlated with colon cancer, which is useful in colon cancer diagnostics. Then, we select three robust classifiers from the six classifier-method combinations, and test on the test set. The results show that all the three classifiers have achieved at least 69% test accuracy. Besides, the results also show that linear SVM classifier has the best generalization performance. It has high accuracy, precision, recall and F1 scores, and the lowest log-loss.

Future improvement: Since both MFA and MFA score+ have bad performance in LOOCV, we can tune the hyper-parameter *k* for MFA and MFA score+.[5]

References

[1] Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.

[2] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, 1999.

[3] Amir Ben-Dor, Ron Shamir, and Zohar Yakhini. Clustering gene expression patterns. *Journal of computational biology*, 6(3-4):281–297, 1999.

[4] Jiangeng Li, Lei Su, and Zenan Pang. A filter feature selection method based on mfa score and redundancy excluding and it's application to tumor gene expression data analysis. *Interdisciplinary Sciences: Computational Life Sciences*, 7(4):391–396, 2015.

[5] Amir Ben-Dor, Laurakay Bruhn, Nir Friedman, Iftach Nachman, Michèl Schummer, and Zohar Yakhini. Tissue classification with gene expression profiles. In *Proceedings of the fourth annual international conference on Computational molecular biology*, pages 54–64, 2000.