Multi-class Texture Analysis
Image Classification

Chao Wang

July 26, 2020

Contents

1 Introduction

One tricky thing about colorectal cancer (CRC), one of the most prevalent cancer types, is that tumour architecture varies within the period of tumour progression and is also related to patient prognosis. [1, 2]

Although it is still indispensable to manually identify and classify different tissues in clinical routine, recent years has witnessed an eruption of automatic image processing and classification regarding texture analysis, which can greatly boost efficiency of clinical trial and save doctors' time for something more valuable.

Research development on this topic has evolved quickly from considering only two categories of tissues (tumour and stroma) [3] to classifying up to eight types of texture images almost perfectly [4]. As for the paper about eight-class texture analysis, there is still room for potential improvement, even though it has already achieved a reasonably good and impressive progress.

In this report, I proposed two main proposals with respect to a possible refinement of the paper's work. On one hand, the paper adopted serveral distinct sets of texture descriptors mainly in a gesture to reduce the dimensionality and meanwhile maintain as much information such as variability and directions as possible. To name a few, histogram features such as mean, standard deviation and central moments represent the overall characteristics of images, whereas local binary patterns (LBP) [5], a concept within computer vision, is referred to as the summary of images' local patterns including the consideration of different directions. As a result, I incorporated one more set of features, Local Phase Quantization (LPQ) [6], which should render the performance even better.

On the other hand, the paper only applied four machine learning algorithms, signifying that I could try different methods and utilize emsemble in the end to yield more robust outcomes. I came up with two paths: directly analyze data based on the gray-scale features or exploit the feature sets by unsupervised and supervised learning algorithms appropriately.

2 Material and Methods

2.1 Dataset

There were two types of dataset available, original images (250,000px) and compressed ones (4,096px). Given the time and hardware limit, I decided to choose the latter one first to get a basic idea of whether the improvement is feasible. Time permitting, the original data can be involved to produce a more professional research project. Before analyzing, the dataset was shuffled to ensure randomness.

The eight categories are: (**a**) tumour epithelium, (**b**) simple stroma, (**c**) complex stroma, (**d**) immune cell, (**e**) debris, (**f**) mucosal gland, (**g**) adipose tissue, (**h**) background. Each has 625 images.

2.2 Method I: Gray-scale Analysis

2.2.1 Data Preparation

Preprocessing First, I looked at the variances of each feature to see whether there were columns with near zero variation, which then should be removed as they rarely provide any information. The following is the histogram.

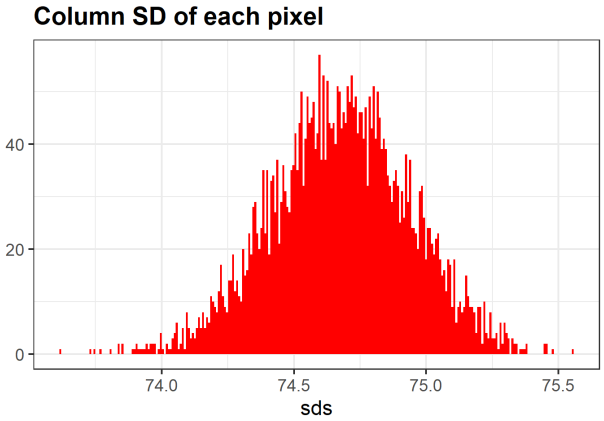


Figure 1: Feature Variance

As is shown in the graph, all the features have relatively large variability, meaning they can all be potentially informative.

Principal Component Analysis It seems particularly plausible to apply PCA in this case where every feature just accounts for a tiny pixel in that extremely scantly will any of them have a dominant effect on the target variable. Additionally, it is also necessary because running over multiple thousand of columns is such a burden for common computers.

I scaled each feature to remove any column effect before using PCA and then plotted how the variance proportion changes over the number of columns.

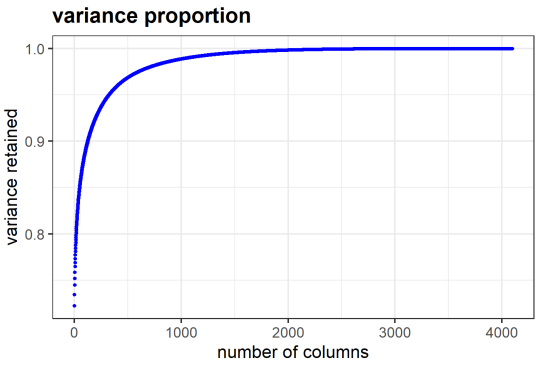


Figure 2: Variance Proportion

We can conclude from the plot that around 100 and 1000 columns are still needed in order to retain 90% and 99% of variability. Due to the hardware limitation, I made the cutoff at the point where 90% of variance was remained.

To obtain a more comprehensive understanding of the data, we can take a closer look at the first few principal components (PC).

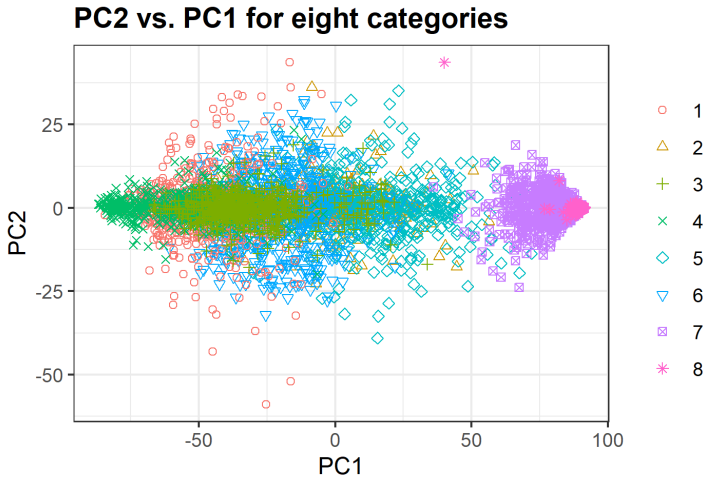
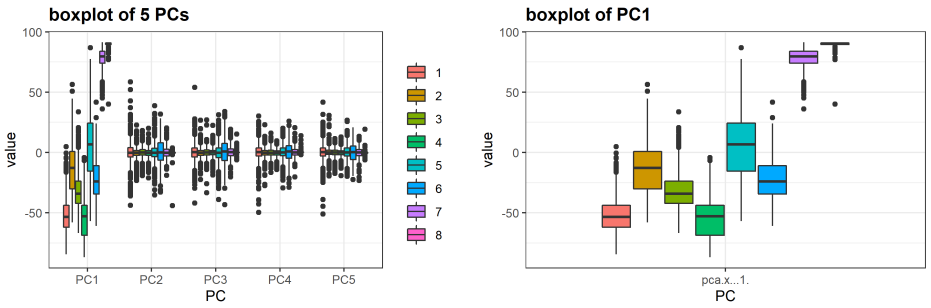


Figure 3: PC1 vs. PC2

The scatterplot indicates that type 7 and 8 are easy to be separated while the others entangle with each other. This is more conspicuous when using boxplot.



(a) 5 PCs

(b) 1 PC

Figure 4: Boxplots

On the left, when plotting based on first 5 PCs, we can clearly see that only the first PC make a difference as it accounts for the most variance (over 70%). Hence, we only need to study it further by plotting it individually. On the right, the relationship among each category is confirmed anew. Although the first six categories, to some extent, overlapp with each other, it still looks promising.

2.2.2 Model Fitting

Multinomial Logistic Regression The first model I used was multinomial logistic regression.

Table 1: Multinomial Logistic Regression

	Sensitivity	Specificity	Precision	F1
Class: 1	0.317	0.902	0.317	0.317
Class: 2	0.238	0.912	0.278	0.256
Class: 3	0.286	0.889	0.269	0.277
Class: 4	0.460	0.912	0.426	0.443
Class: 5	0.540	0.898	0.430	0.479
Class: 6	0.206	0.918	0.265	0.232
Class: 7	0.778	0.991	0.925	0.845
Class: 8	1	0.982	0.887	0.940

The overall accuracy and kappa are 47.8% and 40.3%, respectively. As is demonstrated, specificity is high for all classes, whereas sensitivity denotes an obvious gap between class 7-8 and the rest, which was signified in the previous boxplots. And in generall, the results are not satisfactory.

Linear Discriminant Analysis (LDA) Next I went through a series of discriminant analysis models, starting with LDA. Generally speaking, discriminant analysis is more suitable when there are multiple classes to be categorized.

Table 2: Linear Discriminant Analysis

	Sensitivity	Specificity	Precision	F1
Class: 1	0.063	0.873	0.067	0.065
Class: 2	0.127	0.927	0.200	0.155
Class: 3	0.143	0.902	0.173	0.157
Class: 4	0.143	0.900	0.170	0.155
Class: 5	0.143	0.878	0.143	0.143
Class: 6	0.127	0.893	0.145	0.136
Class: 7	0.413	0.887	0.342	0.374
Class: 8	0.841	0.882	0.505	0.631

The performance of LDA model is very poor where the accuracy is only 25%. And almost all metrics displayed are less than those of logistic regression. This might be because many features are not normally distributed which violated the assumption of LDA.

Quadratic Discriminant Analysis (QDA) QDA is more flexible than LDA in that it does not assume normal distribution for all features.

Table 3: Quadratic Discriminant Analysis

	Sensitivity	Specificity	Precision	F1
Class: 1	0.302	0.977	0.655	0.413
Class: 2	0.254	0.932	0.348	0.294
Class: 3	0.397	0.887	0.333	0.362
Class: 4	0.302	0.964	0.543	0.388
Class: 5	0.270	0.893	0.266	0.268
Class: 6	0.825	0.773	0.342	0.484
Class: 7	0.603	0.989	0.884	0.717
Class: 8	0.937	0.998	0.983	0.959

The overall accuracy was 48.6%.

Mixture Discriminant Analysis (MDA) For MDA, each class is assumed to be a Gaussian mixture of subclasses. Equality of covariance matrix, among classes, is still assumed.

Table 4: Mixture Discriminant Analysis

	Sensitivity	Specificity	Precision	F1
Class: 1	0.476	0.932	0.500	0.488
Class: 2	0.302	0.907	0.317	0.309
Class: 3	0.270	0.878	0.239	0.254
Class: 4	0.476	0.934	0.508	0.492
Class: 5	0.476	0.925	0.476	0.476
Class: 6	0.397	0.914	0.397	0.397
Class: 7	0.651	0.995	0.953	0.774
Class: 8	1	0.950	0.741	0.851

Accuracy is 50.6%.

Flexible Discriminant Analysis (FDA) FDA uses non-linear combinations of predictors such as splines and is useful to model multivariate non-normality or non-linear relationships among variables within each group.

Table 5: Flexible Discriminant Analysis

	Sensitivity	Specificity	Precision	F1
Class: 1	0.365	0.896	0.333	0.348
Class: 2	0.190	0.932	0.286	0.229
Class: 3	0.302	0.873	0.253	0.275
Class: 4	0.413	0.914	0.406	0.409
Class: 5	0.619	0.909	0.494	0.549
Class: 6	0.206	0.925	0.283	0.239
Class: 7	0.524	0.993	0.917	0.667
Class: 8	1	0.932	0.677	0.808

The accuracy was 45.2%.

Regularized Discriminant Analysis (RDA) RDA builds a classification rule by regularizing the group covariance matrices allowing a more robust model against multicollinearity in the data. It can be viewed as a trade-off between LDA and QDA where it shrinks the separated covariances of QDA toward a common covariance as in LDA.

Table 6: Regularized Discriminant Analysis

	Sensitivity	Specificity	Precision	F1
Class: 1	0.206	0.959	0.419	0.277
Class: 2	0.333	0.948	0.477	0.393
Class: 3	0.571	0.853	0.356	0.439
Class: 4	0.587	0.887	0.425	0.493
Class: 5	0.556	0.921	0.500	0.526
Class: 6	0.429	0.961	0.614	0.505
Class: 7	0.873	0.993	0.948	0.909
Class: 8	0.984	0.984	0.899	0.939

The accuracy is 56.8%.

There is certainly no denying that the performances of all discriminant analysis models are abysmal, which could be attributed to the limitations of the models as they have strong assumptions. Actually, I kind of knew it would end up this way, thus indirectly proving why powerful algorithms such as support vector machine are so prevalent.

Support Vector Machine (SVM) SVM is actually the winning algorithm in the paper. Therefore, I chose to try it here as an ending method for gray-scale analysis.

Table 7: Linear SVM

	Sensitivity	Specificity	Precision	F1
Class: 1	0.397	0.891	0.342	0.368
Class: 2	0.333	0.868	0.266	0.296
Class: 3	0.286	0.891	0.273	0.279
Class: 4	0.444	0.925	0.459	0.452
Class: 5	0.444	0.950	0.560	0.496
Class: 6	0.222	0.925	0.298	0.255
Class: 7	0.714	0.995	0.957	0.818
Class: 8	1	0.959	0.778	0.875

Table 8: Radial Basis SVM

	Sensitivity	Specificity	Precision	F1
Class: 1	0.270	0.937	0.378	0.315
Class: 2	0.492	0.868	0.348	0.408
Class: 3	0.302	0.934	0.396	0.342
Class: 4	0.286	0.959	0.500	0.364
Class: 5	0.159	0.959	0.357	0.220
Class: 6	0.683	0.805	0.333	0.448
Class: 7	0.762	0.982	0.857	0.807
Class: 8	0.984	0.975	0.849	0.912

They score 48% and 49% in terms of accuracy, respectively, which is far from expectation. This also implies that gray-scale analysis really is not suitable for texture analysis. One particular reason is that the observations (images) in texture analysis are more complicated than those in digit classification, so dimensionality is rather difficult to be reduced.

The following is a summary table of the performances, with the best highlighted in red and the worst in green.

Table 9: Summary of Algorithm Performance

Algorithm	Accuracy	Kappa
MLR	47.8%	40.3%
LDA	25%	20.4%
QDA	48.6%	44.1%
MDA	50.6%	45.9%
FDA	45.2%	38.3%
RDA	56.8%	50.4%
Linear SVM	48%	42.7%
Radial Basis SVM	49%	42.9%

2.3 Method II: Histogram-Feature Analysis

2.3.1 Motivation

Texture analysis has complicated images that usually require taking all pixel information into account, thus rendering direct gray-scale analysis inefficient. Hence, histogram-feature analysis is called for, which can be viewed as a special case of principal component analysis but is based more on the image context itself and thus more efficient.

2.3.2 Overview

Feature Set Instead of training on and analyzing pixel values of images, the histogram-feature analysis is interested in extracting specific image features via distinct approaches.

- **Lower-Order Histogram Features**
  Lower-order statistics can be used to describe texture [7]. In particular, this set contains the mean, variance, skewness, kurtosis, and the 5th central moment of the histogram.
- **Higher-Order Histogram Features**
  The higher-order feature set consists of the central moments from 2nd to 11th. It is invariant to changes in the average gray-scale intensity of the input image because it does not contain the mean.
- **Local Binary Pattern Histogram Fourier Features (LBP-HF)**
  The LBP-HF operator considers the probability of occurence of all possible binary patterns that arise from a neighbourhood of predefined shape and size, and applies Fourier transformation to prevent the negative effect of image rotations. [5]

- **Local Phase Quantization (LPQ)**
  LPQ, a texture descriptor that is robust to image blurring, utilizes decorrelated phase information computed locally in a window for every image position. [6, 8]

Algorithm I adopted seven classification strategies: 1) Multinomial Logistic Regression, 2) Linear SVM, 3) Radial Basis SVM, 4) Polynomial SVM, 5) Random Forest, 6) XGBoost, 7) Ensemble.

- **Multinomial Logistic Regression (MLR)**
  I used multinomial logistic regression as a baseline method so as to get a basic idea of how the features extracted performs.

- **Support Vector Machine (SVM)**
  I employed support vector machine with one-vs-one class decisions [9] in terms of linear SVM, radial basis SVM and polynomial SVM, with radial basis SVM being the winning method in the paper. The features were automatically standardized before training to maintain equal mean and variance.

- **Random Forest**
  Random forest is also considered a fast and robust technique in which it is particularly effective on skewed data [10]. Although it is not the case here, it is still worth trying given its quickness and robustness.

- **XGBoost**
  Tree boosting is a highly effective and widely used machine learning method which can achieve state-of-the-art results on many machine learning challenges. [11]

- **Ensemble**
  The core idea of ensemble is to carry out a majority vote system for prediction, incorporating all results of the algorithms mentioned above. Theoretically, it is guaranteed to yield a better outcome.

Experiment I planned to implement the experiment in two stages after having all data processed and prepared.

- **Pure Feature Set**
  Firstly, I conducted 4 x 7 = 28 supervised image classification experiments to test the performance of each combinantion of one of four feature sets and one of seven algorithms. As a result, I managed to group by and order feature sets as well as algorithms based on the accuracy.

- **Combined Feature Set**
  Next, pursuant to the results obtained from the analysis of pure feature sets, I could form grouped feature sets in the form of best two, three, four and all five, thereafter ending up with the most satisfactory result.

2.3.3 Model Fitting

Pure Feature Trial During the first step, I trained seven models one by one using one out of five pure feature sets. Consequently, the best result of each set was displayed. It is not difficult to notice that all of the tables are yielded by the ensemble method, which is theoretically tenable.

Table 10: Ensemble of Lower-Order

	Sensitivity	Specificity	Precision	F1
Class: 1	0.649	0.966	0.712	0.679
Class: 2	0.691	0.958	0.723	0.707
Class: 3	0.667	0.936	0.588	0.625
Class: 4	0.738	0.972	0.800	0.768
Class: 5	0.776	0.957	0.703	0.738
Class: 6	0.866	0.979	0.866	0.866
Class: 7	0.938	0.998	0.984	0.960
Class: 8	1	0.995	0.968	0.984

Table 11: Ensemble of Higher-Order

	Sensitivity	Specificity	Precision	F1
Class: 1	0.702	0.914	0.513	0.593
Class: 2	0.603	0.956	0.683	0.641
Class: 3	0.583	0.898	0.438	0.500
Class: 4	0.354	0.975	0.676	0.465
Class: 5	0.603	0.968	0.714	0.654
Class: 6	0.851	0.961	0.770	0.809
Class: 7	0.875	0.991	0.933	0.903
Class: 8	1	0.991	0.938	0.968

Table 12: Ensemble of LBP-HF

	Sensitivity	Specificity	Precision	F1
Class: 1	0.719	0.944	0.621	0.667
Class: 2	0.765	0.940	0.667	0.712
Class: 3	0.383	0.939	0.460	0.418
Class: 4	0.769	0.977	0.833	0.800
Class: 5	0.638	0.980	0.804	0.712
Class: 6	0.791	0.947	0.697	0.741
Class: 7	0.938	0.993	0.952	0.945
Class: 8	0.984	0.998	0.984	0.984

Table 13: Ensemble of LPQ

	Sensitivity	Specificity	Precision	F1
Class: 1	0.614	0.939	0.565	0.588
Class: 2	0.706	0.958	0.727	0.716
Class: 3	0.583	0.907	0.461	0.515
Class: 4	0.815	0.989	0.914	0.862
Class: 5	0.552	0.973	0.727	0.627
Class: 6	0.776	0.956	0.732	0.754
Class: 7	0.922	0.998	0.983	0.952
Class: 8	1	0.995	0.968	0.984

After a close look into the tables listed, it is somewhat valid to conclude that the overall performance of histogram-feature analysis has been improved remarkably, compared to that of gray-scale analysis, even with just one set of features. Furthermore, a stricter and more sensible selection could be made after comparing the results within the feature sets by a summary table.

Table 14: Summary of Pure Feature Set Performance

Accuracy	**MLR**	**Lin SVM**	**rb SVM**	**Poly SVM**	**RF**	**XGBoost**	**Ensemble**
Lower	69.4%	70.6%	78%	58.4%	77.6%	78.4%	79.2%
Higher	65.2%	63.8%	64.4%	46.4%	68%	66.6%	69.6%
LBP-HF	72.2%	71.8%	68.6%	68.6%	73.8%	74.2%	75.2%
LPQ	70.6%	71.8%	69.8%	72.4%	75%	79%	79.8%

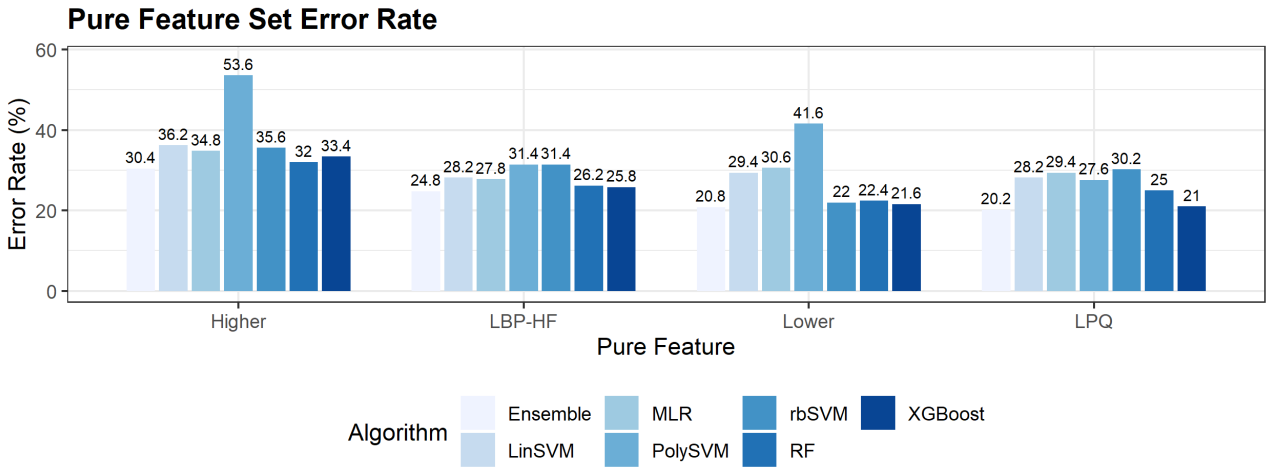


Figure 5: Pure Feature Set Error Rate

The rb SVM is highlighted in green as it is the winning algorithm in the paper, so it is treated as a baseline here. The best results are all accentuated in red, which undoubtedly belong to the ensemble method. In addition, yellow represents the best performance by a single algorithm where three models (XGBoost) out of four are newly introduced into my research, implying that there is genuinely great room for improvement with regard to the original paper.

Meanwhile, another crucial implication is that LPQ becomes the most representative feature set among all, even beating the previous champion, the lower-order feature set. To further elaborate, I applied PCA to LPQ features as its generic version had 256 columns which led to overfitting during training. As a consequence, a better result, after trial and error, was achieved when I kept only 71 columns retaining 80% of variability. In the remainder I refer to the 71-column LPQ features as LPQ for simplicity. Then, I explored and established the combined feature set analysis.

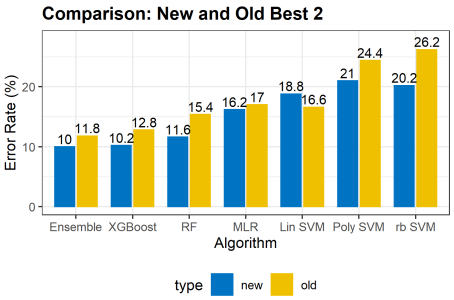
Combined Feature Trial At this stage, I computed the accuracies of the best two, three, four and all five combined models.

- Best 2: LPQ + Lower-Order
  According to the pure feature set analysis, the best 2 is LPQ feature and lower-order histogram feature. After undergoing PCA over 256-column LPQ feature set, the best 2 has 76 variables only but an accuracy up to 90%, with XGBoost scoring 89.8%.

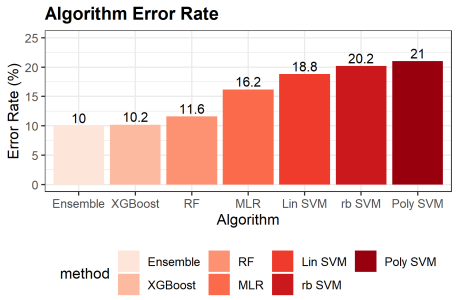
Table 15: Performance Summary of Best 2

Algorithm	Accuracy	Kappa
MLR	83.8%	81.5%
Linear SVM	81.2%	78.5%
rb SVM	79.8%	76.9%
Poly SVM	79%	76%
Random Forest	88.4%	86%
XGBoost	89.8%	88.3%
Ensemble	90%	88.6%

More importantly, it is proved that the original best 2 (lower-order + LBP-HF) is indeed eclipsed on the left barplot where its lowest error rate is 1.8% higher than the new best 2. The right barchart shows directly the error rates of different algorithms of new best 2, with lowest on the left and highest on the right.



(a) Error Rate: Best 2 New vs. Old



(b) Algorithm Performance of Best 2

To be more specific, if we look at different evaluation metrics, we could spot how much progress the model has already achieved.

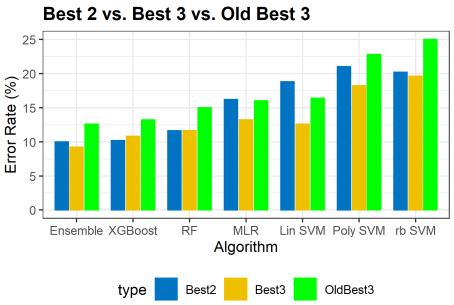
Table 16: Ensemble of Best 2

	Sensitivity	Specificity	Precision	F1
Class: 1	0.877	0.977	0.833	0.855
Class: 2	0.897	0.968	0.813	0.853
Class: 3	0.750	0.970	0.776	0.763
Class: 4	0.831	0.995	0.964	0.893
Class: 5	0.914	0.993	0.946	0.930
Class: 6	0.955	0.986	0.914	0.934
Class: 7	0.969	1	1	0.984
Class: 8	1	0.995	0.968	0.984

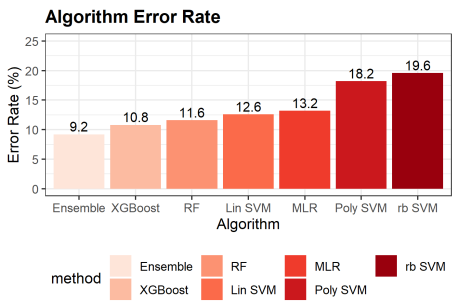
- Best 3: 70% LPQ + Lower-Order + Higher-Order
  Quite surprisingly, the third best feature set, LBP-HF, is absent here, which is supposed to be a part. Instead, the higher-order histogram feature set becomes the champion. In the mean time, a minor modification has been made to LPQ in that 70% of variance is adopted as 80% showed a sign of overfitting which gave rise to a decline in accuracy.

Table 17: Performance Summary of Best 3

Algorithm	**Accuracy**	**Kappa**
MLR	86.8%	83.3%
Linear SVM	87.4%	84.1%
rb SVM	80.4%	76.6%
Poly SVM	81.8%	76.9%
Random Forest	88.4%	84.8%
XGBoost	89.4%	87.3%
Ensemble	90.8%	88.6%



(a) Best 3 vs. Best 2 vs. Old Best 3



(b) Algorithm Performance of Best 3

Table 18: Ensemble of Best 3

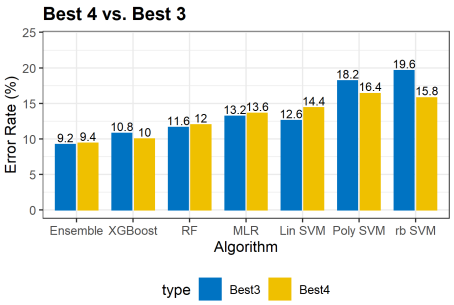
	Sensitivity	Specificity	Precision	F1
Class: 1	0.895	0.975	0.823	0.857
Class: 2	0.926	0.972	0.840	0.881
Class: 3	0.783	0.970	0.783	0.783
Class: 4	0.815	0.995	0.964	0.883
Class: 5	0.948	0.998	0.982	0.965
Class: 6	0.940	0.991	0.940	0.940
Class: 7	0.953	1	1	0.976
Class: 8	1	0.993	0.953	0.976

From the above, the first conclusion is that new models both surpass the performance of the original combined best 3 model, which consolidates this report's goal and assertation anew. Second, when taking the previous metric table of best 2 into account, we could notice a slight increase in the overall accuracy of the new best 3 with a smaller variance across all algorithms.

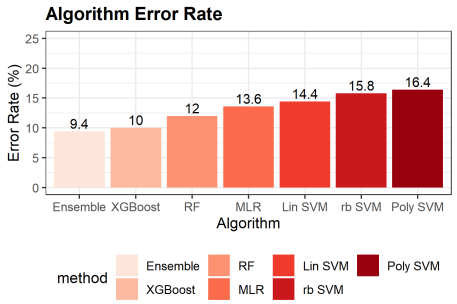
- All 4: 70% LPQ + Lower-Order + Higher-Order + LBP-HF
  The same as the case of best 3, 70% LPQ is employed to avoid overfitting.

Table 19: Performance Summary of All 4

Algorithm	Accuracy	Kappa
MLR	86.4%	83.1%
Linear SVM	85.6%	82.5%
rb SVM	84.2%	80.9%
Poly SVM	83.6%	79%
Random Forest	88%	84.5%
XGBoost	90%	87.4%
Ensemble	90.6%	88.3%



(a) Error Rate: Best 4 vs. Best 3



(b) Algorithm Performance of Best 4

Table 20: Ensemble of All 4

	Sensitivity	Specificity	Precision	F1
Class: 1	0.912	0.980	0.852	0.881
Class: 2	0.853	0.970	0.817	0.835
Class: 3	0.800	0.968	0.774	0.787
Class: 4	0.877	0.993	0.950	0.912
Class: 5	0.931	0.991	0.931	0.931
Class: 6	0.925	0.993	0.954	0.939
Class: 7	0.953	1	1	0.976
Class: 8	1	0.998	0.984	0.992

Unexpectedly, the last model is not the best, although being the seemingly most comprehensive one. Nonetheless, it is by far the most stable across all different methods concluded from the barplot on the left. Likewise, as for the summary table, the sensitivity, in particular, has all transcended 80% for the first time, highlighting the stableness.

One possible explanation of the decrease of highest accuracy could be the innate property of LPQ and LBP-HF. They both belong to a category called "dense" descriptors which extracts characteristics pixel by pixel over an input image [12]. Thus, the functional imcompatibility is a likely culprit where their positive predicting power has overlapped, whereas their bias has added up.

3 Discussion

3.1 Major Findings

In this report I investigated texture analysis for discriminating multi-class colorectal cancer images. I sorted out several key points.

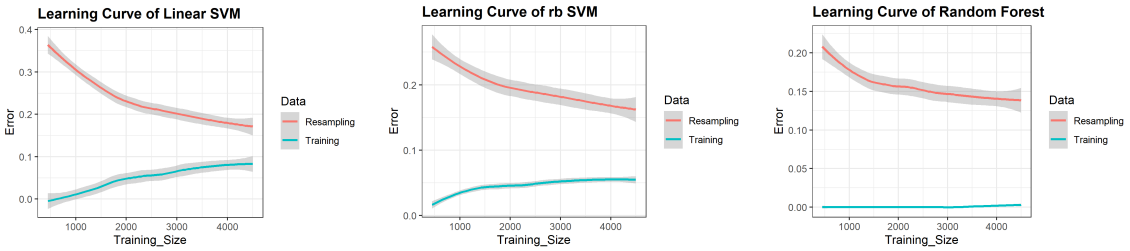
- Histogram-feature analysis is far more accurate than gray-scale analysis when images have complex patterns and backgrounds, as the former is essentially a content-based and thus more efficient way to reduce dimensionality.
- With regard to the single feature set, the LPQ feature set with PCA is the most informative descriptor in this case, followed by lower-order histogram, LBP-HF and higher-order histogram. This finding indicates that the first aim of this report is achieved because the newly introduced LPQ actually outperforms the original winner - the lower-order histogram, which is the reason why I did not include the other three "worse" feature sets mentioned in the paper.
- As for the overachingly best combined set, the best 3 (70% LPQ + lower-order + higher-order) tops the list. Again, its superior is proved by comparison.
- When it comes to algorithms, ensemble method dominates the test, comprising usually the multinomial logistic regression, random forest and XGBoost, XGBoost being the best individual. This point completes the second mission of this report - to seek improvement in the aspect of algorithms.
- To sum up, my study managed to produce a reasonably good and similar results as the paper, using only compressed images which by all means contained less information. Nonetheless, the idea of extracting different feature sets is highly generalizable that it could be applied to any kind of gray-scale images, no matter how large the pixel number, thus reducing dimensionality.
- The following is a confusion matrix of the best performance in a gesture to show the progress more directly. In conclusion, only the complex stroma (3) is somewhat harder to distinguish, other parts, especially the tumour, can be separated quite precisely.

Table 21: Best Prediction vs. Reference

	1	2	3	4	5	6	7	8
1	51	1	6	3	0	2	0	0
2	2	62	6	2	3	0	0	0
3	2	2	46	6	1	1	0	0
4	1	0	1	53	0	0	0	0
5	0	2	0	0	54	0	0	0
6	1	1	1	1	0	64	0	0
7	0	0	0	0	0	0	62	0
8	0	0	0	0	0	0	2	61

3.2 Limitations

The biggest limitation of this report lies in the lack of data, which is likely to lead to overfitting.



(a) Linear SVM (b) rb SVM (c) Random Forest

Figure 9: Learning Curve

Displayed here are three typical learning curve plots. All of them demonstrate a trend of high variance and thus overfitting, which could be alleviated by enlarging the sample size. However, a larger data set is not available in this specific case. Other ideas have also been explored such as decreasing the number of predictors and using regularization, but the influence is minor, partly attributed to the limitation of hardware condition.

3.3 Outlook

Machine learning in image classification and texture analysis is constantly gaining popularity and momentum nowadays. Though reasonably good resutls have been achieved in this case of colorectal cancer histology, there are still some perspectives to be studied.

First, a new feature set could be involved. Weber Local Descriptor (WLD) is a robust local image descriptor which consists of two components: differential excitation and orientation. The differential excitation component is a function of the ratio between two terms: on is the relative intensity differences of a current pixel against its neighbours; the other is the intensity of the current pixel. The orientation component is the gradient orientation of the current pixel. [12]

Second, novel algorithms also deserve a closer look. To name a few, deep learning algorithms have a great potential. In particular, the CNN-XGBoost model was built specifically for image classification. [13]

Generally speaking, the ideas and approaches of texture analysis could be applied to various kinds of complicated image classification problems, regardless of how large the images are. As a result, it is predictable that more and more fields can utilize the techniques in this report to do interesting and meaningful jobs.

All details and results are stored in Github for further open discussion:
https://github.com/Kennethws/ntu-report.git

References

[1] Egeblad M, Nakasone ES, Werb Z. Tumors as organs: complex tissues that interface with the entire organism. Developmental cell. 2010 Jun 15;18(6):884-901.

[2] Huijbers A, Tollenaar RA, v Pelt GW, Zeestraten EC, Dutton S, McConkey CC, Domingo E, Smit VT, Midgley R, Warren BF, Johnstone EC. The proportion of tumor-stroma as a strong prognosticator for stage II and III colon cancer patients: validation in the VICTOR trial. Annals of oncology. 2013 Jan 1;24(1):179-85.

[3] Bianconi, F., *Álvarez-Larrán*, A. & Fernández, A. Discrimination between tumour epithelium and stroma via perception-based features. Neurocomputing. 2015 Apr 22;154:119-26.

[4] Kather JN, Weis CA, Bianconi F, Melchers SM, Schad LR, Gaiser T, Marx A, Zöllner FG. Multi-class texture analysis in colorectal cancer histology. Scientific reports. 2016 Jun 16;6:27988.

[5] Ahonen T, Matas J, He C, Pietikäinen M. Rotation invariant image description with local binary pattern histogram fourier features. InScandinavian conference on image analysis 2009 Jun 15 (pp. 61-70). Springer, Berlin, Heidelberg.

[6] Ojansivu V, Heikkilä J. Blur insensitive texture classification using local phase quantization. InInternational conference on image and signal processing 2008 Jul 1 (pp. 236-243). Springer, Berlin, Heidelberg.

[7] Beyerer J, Le*ó*n FP, Frese C. Machine vision: Automated visual inspection: Theory, practice and applications. Springer; 2015 Oct 1.

[8] Heikkila J, Ojansivu V, Rahtu E. Improved blur insensitivity for decorrelated local phase quantization. In2010 20th International Conference on Pattern Recognition 2010 Aug 23 (pp. 818-821). IEEE.

[9] Allwein EL, Schapire RE, Singer Y. Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of machine learning research. 2000;1(Dec):113-41.

[10] Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. RUSBoost: Improving classification performance when training data is skewed. In2008 19th International Conference on Pattern Recognition 2008 Dec 8 (pp. 1-4). IEEE.

[11] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794).

[12] Chen J, Shan S, He C, Zhao G, Pietikainen M, Chen X, Gao W. WLD: A robust local image descriptor. IEEE transactions on pattern analysis and machine intelligence. 2009 Aug 18;32(9):1705-20.

[13] Ren X, Guo H, Li S, Wang S, Li J. A novel image classification method with CNN-XGBoost model. InInternational Workshop on Digital Watermarking 2017 Aug 23 (pp. 378-390). Springer, Cham.