# Colorectal Cancer Tissue Classification Based on Machine Learning

*Completed Research Paper*

**Min-Jen Tsai[1], Imam Yuadi[2], and Yu-Han Tao[1]**

## Abstract

*For digital pathology, automatic recognition of different tissue types in histological images is important for diagnostic assistance and healthcare. Since histological images generally contain more than one tissue type, multi-class texture analysis plays a critical role to solve this problem. This study examines the important statistical features including Gray Level Co-occurrence Matrix (GLCM), Discrete Wavelet Transform (DWT), Spatial filters, Wiener filter, Gabor filters, Haralick features, fractal filters, and local binary pattern (LBP) for colorectal cancer tissue identification by using support vector machine (SVM) and decision fusion of feature selection. The average experimental results achieve high identification rate which is significantly superior to the existing known methods. In summary, the proposed method based on machine learning outperforms the techniques described in the literatures and achieve high classification accuracy rate at 93.17% for eight classes and 96.02% for ten classes which demonstrate promising applications for cancer tissue classification of histological images*

.

**Keywords:** Classification, Decision Fusion, Support Vector Machine, Machine Learning

## Introduction

The structures of human tumors comprise several tissue types that are able to be distinguished by histopathological evaluation of Hematoxylin and Eosin (H&E) stained tissue sections (Kather et al. 2016). During tumor progression associated with patient prognosis, colorectal cancer (CRC) is one type of cancer that is frequently evaluated clinically (A. Huijber et al. 2013). It is imperative to identify multiclass tumors based on tissue types of CRC histological images.

Among previous works, several methods for histological texture analysis have been studied. Kather et.al (2016) identified histological images of human colorectal cancer including eight different types of tissue by using texture descriptors and several classifiers. They divided ten original tissue images into patches before identification based on classes. Mattfeldt, et.al (2013) studied the correlation between epithelial cells and lumina from low grade to high grade prostatic cancer progression in terms of the Gleason score. They implemented multiclass pattern recognition constructed by spatial statistics, as contrasting to the usual method of binary pattern recognition. Huang and Lee (2009) examined variations of intensity and texture complexity for histological grading of prostate tissues

using two feature extraction methods based on fractal dimension. Several feature filter sets were applied and also used by the sequential floating forward selection method to optimize the classification results. Furthermore, Signolle et.al (2009) segmented histopathology slides to identify various types of ovarian carcinoma stroma using wavelet-domain hidden Markov tree model and a pairwise classifiers design and selection. Additionally, Yang et.al (2009) applied a grid-enabled decision support system for performing automatic analysis of breast tissue microarray images. Four different types of filter banks were applied to classify several major subtypes of breast cancer using k-nearest neighbor (*kNN*) and tree were integrated into a Bayesian framework.

The objective of this paper is to obtain the best performance for CRC tissue identification for histopathology images where we investigate different filter sets, expand feature filter combination, and theoretically diffusing the feature selection among feature space. Furthermore, the aim of the research is to acquire the best decision results.

As a consequence, this paper is organized as following: Section 2, contains the description about the theoretical background of feature extraction in different statistical approach and classification for CRC. Section 3 exhibits the justification of the proposed method and other approaches. Section 4 concludes the paper and recommends to the future works.

## Literature Review

The machine learning approach to the automatic separation of tissue types in histological images can be achieved by the method of segmented individual cells based on cell morphology and then classified into different categories such as tumor cells, stromal cells and immune cells. In medical image analysis, texture-based methods are very useful for classifying tissue types. Texture refers to the specific nature of the internal structure of the image area, for example coarse versus fine or oriented versus scattered randomly (Bianconi F. et al 2015). Usually, this method first extracts the texture feature

Generally, in classifying CRC images consisting of tumors and stroma is a difficulty obtained by researchers because the tumor part is heterogeneous. Similarly, various types of studies use their own drawing datasets whose classification performance is always different from previous studies. All published methods always show two general limitations when classifying tissue types in CRC images on a regular basis: first, they only consider two categories of tissues, namely tumors and stroma, which makes this approach unsuitable for more heterogeneous parts of the tumor. Second, all studies use their own image data sets whose classification performance cannot be compared (N. Linder et al. 2012). A number of studies have also investigated the development of automated methods for the assessment and classification of CRC networks. Most of their studies use benchmark data sets available for image classification problems such as handwriting recognition, facial recognition, universal computer vision problems and texture classification. In general, the dataset has no data available for classification of histopathological networks. While for computer-assisted diagnostic systems developed to classify the type of colorectal polyps using sequential image feature selection and classification using vector machine support. Processing pipes, including microscopic image segmentation, feature extraction, and classification, can also be used for automatic detection of cancer through an image (A.A. Nahid, M.A. Mehrabi, and Y. Kong 2018).

For image analysis in cancer patients can use the Convolutional Neural Network Technique, better known as CNN. CNN was first proposed by Fukushima (1980) which is also referred to as "Neocognotron". The main project of CNN is to find a stimulus pattern, where cancer can be tolerated with a limited amount. This "Negotron" model also functions as the first CNN model for biomedical signal analysis. Especially the CNN model was first introduced for breast image classification (Wu C.Y. et al. 1994). Jaffar classifies the mammogram-image dataset using the CNN model and obtains 93.35% and 93.00% of the area under curve (AUC) (Jaffar M.A. 2017). CNN for classification of mammogram images requires 2.5 and 10 map features to obtain an average accuracy of 71.40% (Qiu Y. et al 2016). For automatic mass positioning and image classification, obtain an accuracy of 85.00% (Ertosun M.G. and Rubin D.L. 2015). Classification of a set of mammogram images into a class of design and malignant cancers, in which they used a total of 560 ROI (Region of Interest) and characterized a set of

mammogram images in benign and malignant cancer images and obtained 96.70% accuracy. A set of mammogram images has been classified by Sahiner et al., And the ROC score achieved is 0.87 (Jiao Z et al. 2016).

## The Approach

A systematic framework for CRC classification has been developed and flowcharts are shown on Figure. 1 with the following procedure:
1) Feature extraction: All images are analyzed using ten different filter set features. We extract and convert the original data images into numerical forms so that the data has values that can be processed further. In extracting we use GLCM, DWT, Gabor, Gausian, Log, Unsharp, Wiener, fractal, Haralick, and LBP filters. These ten filters are explained in more detail in the Section below

2) Feature selection: The purpose of filtering is to eliminate filters that have low accuracy or choose the best filter to produce the highest classification accuracy. This study implements a feature selection algorithm with five fusion technique decisions to select all 306 features. Selection of this feature is done automatically to select features that most contribute to predictive variables or output on CRC classification. Furthermore, the irrelevant features in this study can reduce the accuracy of the model and make it maximal in determining CRC based on irrelevant features.

3) Classification: at the classification stage, we use SVM in classifying CRC textures. We chose this type of classification because it has the ability and good performance in classifying gray pixel images to find the best hyperplane that functions as a separator between classes in the input space. In comparison between extraction features for multi texture classification problems, SVM can be applied by combining extractor features to get the best results. In the classification stage, we first analyzed CRC histology in eight multi-class textures without overlapping images patches. These ten filters are explained in more detail in the Section below

In the first experiment, we use the selected datasets such as is shown in Fig. 2. It was conducted by Kather, et.al. (2016). Subsequently, it is also implemented in our approach by using 90% CRC images from a tissue type are randomly selected to train the SVM classifier, whereas at least another 10% images, randomly taken from the same tissue data sets, are tested during the identification step. In the second experiment, we present a dataset of 5,000 histological images of human colorectal cancer including eight types of tissue from Kather (2016). From the first 10 images per class shown in Figure. 3 has an average coloring intensity different from one network to another. It reflects the usual variability in routine histopathological slides. In addition to these images, we also extracted ten larger images with dimensions of 5000 x 5000 pixel sizes from different network areas than those applied to smaller images. In the second experiment, we included ten images application set that are shown in Fig. 3 and Fig. 4 in different types of tissue by dividing each image of 5000 pixel square as 10,000 images overlapped at 150x150 pixel sizes. To improve the results, we used 10,000 images of each tissue type and then implemented 5000 images as a training set and 1000 images as testing sets.

Since certain properties or patterns will be embedded in the CRC images, such action is similar to the operation of the active warden (2003). Those image features are categorized into ten different groups which will be briefly explained

### *The Spatial Feature*

GLCM features are the estimates of the second order probability density function of the pixels in the image where the overall spatial relationships are calculated. The other spatial features are DWT, Gaussian, Laplacian of Gaussian (LoG), Unsharp, Wiener, and Gabor (R.C. Gonzalez and R.E. Woods 2007). A 2D Gabor filter has two-dimensional filter which is Gaussian kernel function modulated by a complex sinusoidal plane wave and has several advantages such as invariance to illumination, rotation, scale and translation (L.R. Vega et al. 2013). Additionally, the co-occurrence matrix and texture features are the most popular second-order statistical features which are introduced by RM Haralick in 1973 (R. Haralick et al. 1973) also used in this study.

## *The fractal and LBP feature filters*

In this study, we extracted fractal based features by calculating the fractal dimension (A.F. Costa et al. 2012). These features are built on fractal dimension for gray-scale images which depict objects and structure boundary. LBP is a feature extractor that has an appropriate and powerful sub pattern-based texture descriptor. It characterizes the gray-scale invariant texture and combination between measuring texture from each neighborhood and the difference of the average gray level of those pixels based on binary numbers (T. Ojala et al. 2002).

## *Decision Fusion Approach*

The decision fusion model referring to the feature-selection and decision fusion technique is therefore explored. The floating search methods are implemented by the sequential selection procedures that are related to the plus $\boldsymbol{l}$ take-away $\boldsymbol{r}$ algorithms (P. Budil et al. 1994). Plus $\boldsymbol{l}$ minus $\boldsymbol{r}$ selection (LRS) starts from the empty set and repeatedly adds $\boldsymbol{l}$ features and removes $\boldsymbol{r}$ features when $\boldsymbol{l}$ is more than $\boldsymbol{r}$. Conversely, when $\boldsymbol{l}$ is less than $\boldsymbol{r}$, LRS starts from the full set and repetitively removes $\boldsymbol{r}$ features followed by $\boldsymbol{l}$ additions.

The plus **l** take-away **r** algorithms method can be described in an algorithmic way as following:

*Input:* $Y = \{y_j \mid j = 1, \dots, D\}$ //available measurements//

*Output:* $X_k = \{x_j \mid j = 1, \dots, k, x_j \in Y\}, k = 0,1, \dots, D$

*Initialization:* if $l > r$ then $k := 0$; $X_0 := \emptyset$; go to Step 1

else $k := D$; $X_D := Y$; go to Step 2

Step 1 (*Inclusion*)

repeat $m$ times

$x^+ := \arg\ max_{x \in Y - X_k} J(X_k + x)$

$X_{k+1} := X_k + x^+;\ k := k + 1;$ go to Step 2

Step 2 (*Exclusion*)

repeat r times

$x^- := \arg\ \max_{x \in X_k} J(X_k - x)$

$X_{k-1} := X_k - x^-;\ k := k - 1;$ go to Step 1

It can be implemented by using plus 2 minus 1 (P2M1) where ($l$ =2, $r$ =1), plus 3 minus 2 (P3M2) where ($l$ =3, $r$ =2), and plus 4 minus 3 (P4M3) where ($l$ =4, $r$ =3). Furthermore, to perform feature selection, Pudil ( 1994) proposed the SFFS and SBFS methods. The SFFS method is a modified plus-m-minus-r by one more mechanism in the minus step. The SFFS method can be described algorithmically in a similar way to the previous method as following:

*Input :* $Y = \{y_j \mid j = 1, \dots, D\}$ //available measurements//

*Output:* $X_k = \{x_j \mid j = 1, \dots, k, x_j \in Y\}, k = 0,1, \dots, D$

*Initialization:* $X_0 := \emptyset;\ K := 0$

(in practice one can begin with $k = 2$ by applying SFS twice)

*Termination:* Stop when $k$ equals the number of features required

Step 1 (*Inclusion*) $x^+ :=$ arg $\max_{x \in Y - X_k} J(X_k + x)$

$X_{k+1} := X_k + x^+; k := k + 1$

Step 2 (*Conditional Exclusion*)

$x^- := \arg \max_{x \in X_k} J(X_k - x)$

if $J(X_k - \{x^-\}) > J(X_{k-1})$ then

$X_{k-1} := X_k - x^-; k := k - 1$ ; go to Step 2

else go to Step 1

A challenge of feature selection integration or fusion represents the method of combining the above mentioned five different techniques of feature selection (P2M1, P3M2, P4M3, SFFS, and SBFS). The goal here is to gather the most useful features from all the selection methods, in such a way that the end-result is to achieve the maximum outcomes from each technique respectively and then making a fusion from each of them after aggregation.

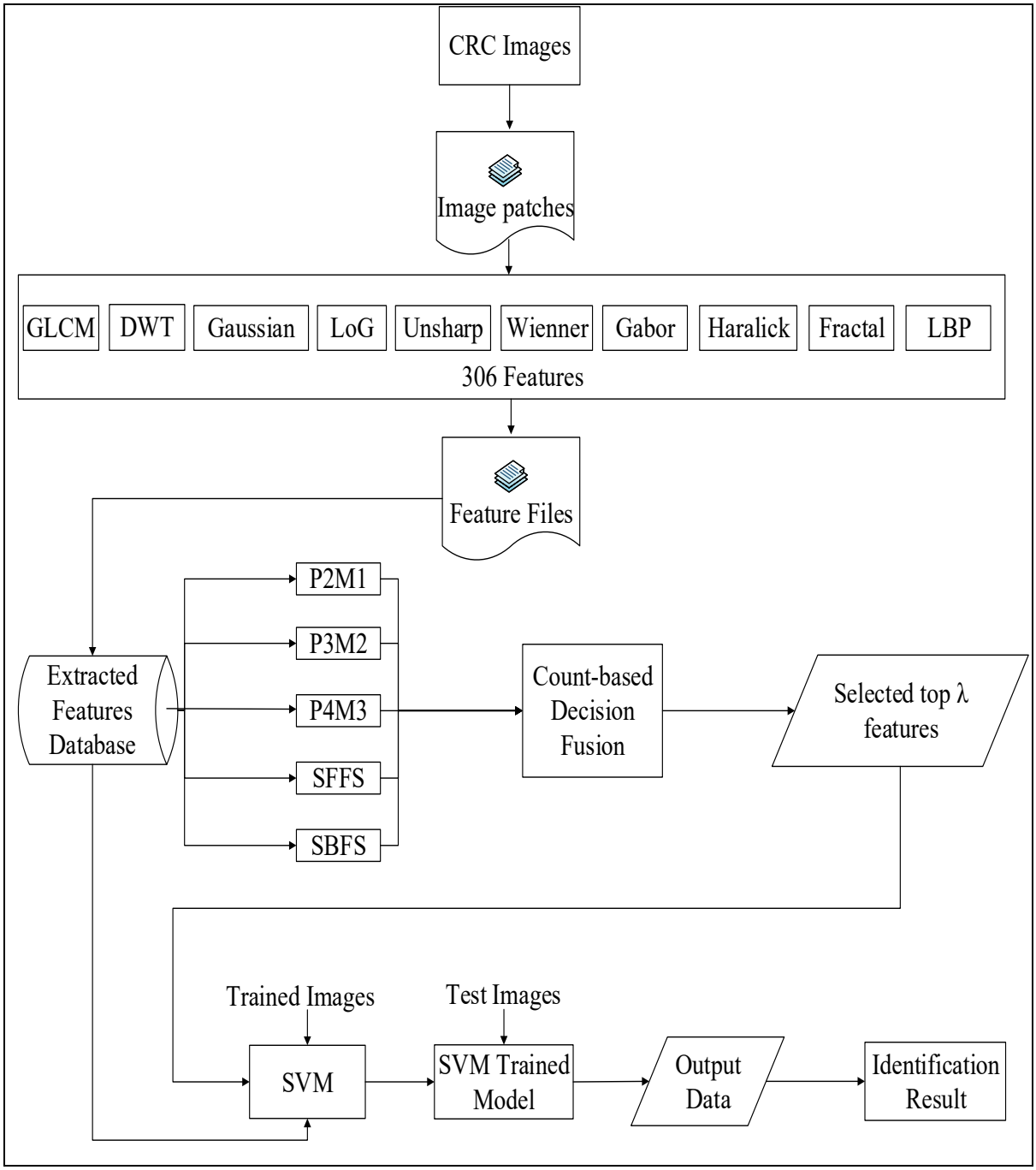


**Figure 1. Procedure of identifying CRC**

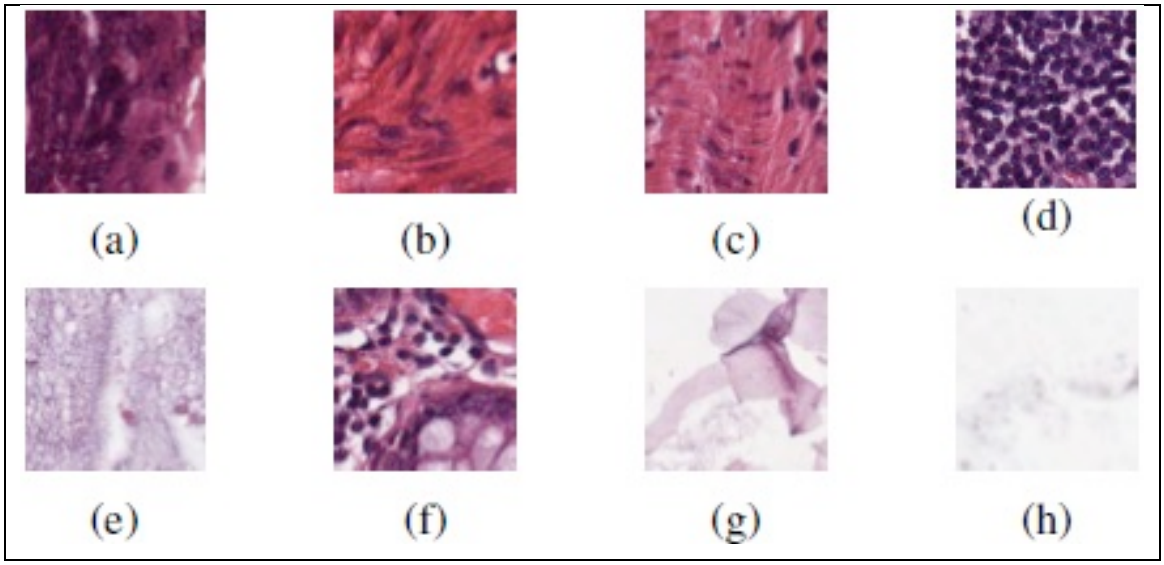


**Figure 1. Eight types of tissue image patch samples (a) tumour epithelium, (b) simple stroma, (c) complex stroma (stroma that contains single tumor cells and/or single immune cells), (d) immune cell conglomerates, (e) debris and mucus, (f) mucosal glands, (g) adipose tissue, (h) background.**

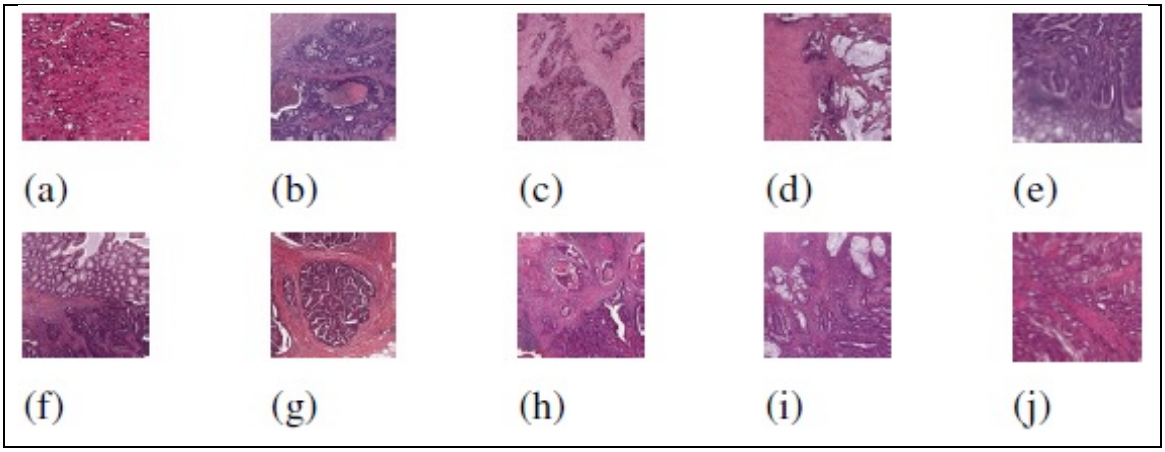


**Figure 2. Ten Types of original Tissue images.**

## Experiment Results

Extensive experiments have been conducted to verify the efficacy of our proposed method for histological tissue image identification. The experiments are performed in sequential steps and the results are tabulated for demonstration purpose.

### *Feature extraction*

In this study, there are total 306 statistical features which computation requirement increases fast while a large number of features as well as instances are processed. The Feature filters that are used in this experiment i.e., GLCM (22 features), DWT (12 features, Gaussian filter (21 features) , LoG filter (21 features) , Unsharp filter (21 features), Wiener filter (64 features), Gabor filter (48 features), Haralick filter (14 features), fractal filter (24 features), and LBP filter (59 features). Therefore, the computational complexity is a critical issue to be resolved in real applications. Therefore, feature selection conducted in the next step to alleviate the computation demands.

## *Decision Fusion for feature Selection*

The adaptive feature selection algorithm is implemented in this study in order to reduce the total evaluation time without the loss of accuracy while the most important λ features are selected. The number of chosen features is determined based on the accuracy rate for all 306 features.

*Deciding the most important λ features*
Five feature selection methods are adopted for the feature selection processes including: P2M1, P3M2, P4M3, SFFS and SBFS. The selection order during execution is recorded to choose the most important features and the feature value of λ= 222 is obtained after the experimental analysis.

*Determine the most λ effective features*
Since the highest accuracy rate can be achieved by using 222 features from above analysis, the counter-based decision fusion algorithm is used to decide the final top λ features from the recorded feature selection order. After the most important features have been decided, the tissue type identification can be finally investigated.

## *Classification*

By using confusion matrix, we analyze accuracy rate prediction based on each column of the matrix that represents the instances in a predicted class while each row represents the instances in an actual class. As shown in Table I, the average accuracy rate using 222 features for CRC classification for multiclass tissue separation can achieve 93.17% and using 306 features can obtain 92.90%. Alternatively, the average prediction results are shown in Table II can attain 95.76% for different tissue types when all 306 features are applied and achieved 96.02% when it used 222 features. Given these points, the average accuracy rate by using decision fusion of feature selection (222 features) is better than the using all features 306 features to identify histological images. It is clear that the proposed approach by using decision fusion with 222 data sets after feature selection is better than the other feature sets.

**Table 1. Accuracy Prediction Rates For Eight Classes**

| Filters | Number of features | Accuracy (%) |
|---|---|---|
| GLCM | 22 | 86.27 |
| DWT | 12 | 74.06 |
| Gaussian | 21 | 85.44 |
| LoG | 21 | 66.25 |
| Unsharp | 21 | 81.21 |
| Wiener | 64 | 82.60 |
| Gabor | 48 | 85.73 |
| Haralick | 14 | 81.60 |
| Fractal | 24 | 88.13 |
| LBP | 59 | 77.58 |
| **Selected features** | **222** | **93.17** |
| All features | 306 | 92.90 |

**Table 2. Accuracy Prediction Rates For Ten Classes**

| Filters | Number of features | Accuracy (%) |
|---|---|---|
| GLCM | 22 | 68.42 |
| DWT | 12 | 55.37 |
| Gaussian | 21 | 62.93 |
| LoG | 21 | 41.24 |
| Unsharp | 21 | 56.84 |
| Wiener | 64 | 71.86 |
| Gabor | 48 | 71.43 |
| Haralick | 14 | 56.03 |
| Fractal | 24 | 73.50 |
| LBP | 59 | 73.87 |
| **Selected features** | **222** | **96.02** |
| All features | 306 | 95.76 |

The previous studies are compared to our approach tabulated in Table III. Kather et.al (2016) implemented the same dataset to classify 8 classes could achieve 87.4% accuracy rate. Signol (2010) identified ovarian cancer histology to analyze tumor epithelium in the *wavelet-based* where they reported 71.5% accuracy rate. In addition, Yang et. al. (2009) used four different filter sets to extract imaged breast tissue microarrays for identifying breast cancer. They achieved the average accuracy rate of identification was 89% for three tissue types. On the other hand, deep learning could achieve accuracy rate up to 90% to identify breast cancer (F.A. Spanhol et al. 2016). It demonstrates that our identification results achieve higher accuracy rates (93.17% for eight classes and 96.02% for ten classes) than previous results. It demonstrates that our proposed method is superior to the previous studies and the technique can effectively identify the CRC based histological images.

In summary, from above analyses, the superior accuracy rates justify the effectiveness of our proposed method for eight and ten classes in identifying CRC tissue images by using feature selection with decision fusion. It is highly promising that the proposed technique can be a universal tool of cancer histology where further researches will be undertook to prove its ubiquity.

**Table 3. Prediction Results among Different Hispatological Images**

| Research | Filters | Research object | Classifier | Claimed accuracy rate (%) |
|---|---|---|---|---|
| J.N. Kather et al. (2016) | LBP, Gabor, GLCM | CRC for eight classes | SVM | 87.4 |
| Signolle N. (2010) | Wavelet | Tumor epithelium | Hidden Markov tree | 71.5 |
| Yang L. et al. (2009) | Four different filter sets | Breast cancer | kNN, Bayesian, C4.5 decision tree, and SVM | 89 |
| F.A. Spanhol et al. (2016) | | Breast Cancer | CNN | 90 |
| Our Approach | Decision fusion of Ten different filter sets | CRC for eight classes | SVM | 93.17 |

## Discussion

1) Automatic recognition is an essential part in the digital pathology to analyze different tissue types. To achieve the successful studies, multi-disciplinary experts are needed to collaborate among researches such as medical experts, pathologists, computer vision experts, etc.
2) How to identify the region where different tissue types locate is still a critical issue in digital histopathology in practice. In additional, limited data sources are available online and this study used the dataset from (J.N. Kather et al. 2016), but more medical data are needed to validate its capability in general use.
3) Currently, machine learning based approach is investigated but the preliminary results are not as good as feature based techniques. The reason may be due to the limited data sets available publicly to train the system. Further data collection may help to effectively enlighten the system for better classification accuracy

## Conclusion

This paper presents different tissue types in histological images classification for CRC detection using machine learning (SVM) based decision fusion. It demonstrates that our identification results achieve higher accuracy rates (93.17% for eight classes and 96.02% for ten classes) than previous results. It demonstrates that our proposed method is superior to the previous studies and the technique can effectively identify the CRC histological images. The results also confirm that human solid tumors with the complex structures can be distinguished based on the texture of tissue types. For future works, the textures of tissue morphology detection will be studied, and the information from biological visual fields will also be analyzed.

## References

Avcibas, I., Memon, N., dan Sankur, B. 2003. "Steganalysis using image quality metrics", *IEEE Transactions on Image Processing,* vol. 12, pp. 221-119.

Barker, J., Hoogi, A., Depeursinge, A. & Rubin, D. L. 2015. Automated Classification of Brain Tumor Type in Whole-Slide Digital Pathology Images Using Local Representative Tiles. *Med Imag Anal* 30, 60–71

Bianconi, F. & Fernández, 2014. A. An appendix to 'texture databases-A comprehensive survey'. *Pattern Recognit Lett* 45, 33–38

Bianconi, F., Álvarez-Larrán, A., & Fernández, A. 2015. Discrimination between tumour epithelium and stroma via perception-based features. *Neurocomputing* 154, 119–126

Costa AF.,G. Humpire-Mamani, A.J.M. Traina. 2012. An efficient algorithm for fractal analysis of textures. *SIBGRAPI Conference on Graphics, Patterns and Images, August, Ouro Preto*, pp. 39-46

Gonzales, R.C.,Woods, R.E. 2007 Digital Image Processing, 3rd, Edition, Prentice Hall.

Haralick, R., Shanmugam,K., Dinstein, I.. 1973 "Textural Features for Image Classification", *IEEE Transaction on Systems, Man, Cybernetics SMC*-3(6). pp. 610-621.

Huang, P.-W and Lee, C.-H. 2009. *Automatic classification for pathological prostate images based on fractal analysis*. IEEE T Med Imaging 28, pp. 1037–1050.

Huijbers, A. et al. 2013. The proportion of tumor-stroma as a strong prognosticator for stage II and III colon cancer patients: validation in the VICTOR trial. *Ann Oncol* 24, pp. 179–85.

Kather, J.N., Wei, C. A. Bianconi, F. Melchers, S. M. Schad, L. R. Gaiser, T. Marx, A., and Zöllner, F. G. 2016. "Multi-class texture analysis in colorectal cancer histology*", Scientific Reports* 6, Article number: 27988.

Linder, N. et al. 2012. Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. *Diagn Pathol* 7, 22

M.A., Jaffar. 2017. Deep Learning based Computer Aided Diagnosis System for Breast Mammograms. *International Journal of Advanced Computer Science and Applications*.

M.G., Ertosun, D.L.. 2015. Rubin. Probabilistic visual search for masses within mammography images using deep learning. *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2015*. pp. 1310–1315

Mattfeldt, T. Grahovac, P. and Luck, S. 2013. *Multiclass Pattern Recognition of the Gleason Score of Prostatic Carcinomas Using Methods of Spatial Statistics*. Image Anal Stereol 32, pp. 155–165.

Nahid, A. A., Mehrabi, M. A., & Kong, Y. 2018. Histopathological Breast Cancer Image Classification by Deep Neural Network Techniques Guided by Local Clustering. *BioMed research international*

Ojala, T., Pietikäinen, M., Mäenpää. T. 2002. Multiresolution gray-scale and rotation invariant texture classification with LBP, *IEEE Trans. Pattern Analysis & Machine Intelligence*. 24 (7), pp. 971-987

Pudil, P., Ferry, F.J., Novovicova, J and Kittler,J,.1994. Floating search methods for feature selection with no monotonic criterion functions. *IEEE*, 1051-465U9

Pudil, P., Novovicova, J and Kittler, J 1994. "Floating search methods in feature selection," *Pattern Recognition Letters*, vol.15, pp. 1119–1125, 1994

Qiu Y., Wang Y., Yan S., et al. 2016. An initial investigation on developing a new method to predict short-term breast cancer risk based on deep learning technology. *Proceedings of the Medical Imaging 2016: Computer-Aided Diagnosis; March 2016; San Diego, California, USA. SPIE. Digital Library*

Signolle, N., Revenu, M., Plancoulaine, B.,and Herlin, P. 2010. *Wavelet-based multiscale texture segmentation: Application to stromal compartment characterization on virtual slides*. Signal Process 90, pp. 2412–2422.

Spanhol, F.A., Oliviera, L.S., Petitjean, C and Heutte, L. 2016. Breast Cancer Hispathological Image Classification using Convolution Neural Networks, International Join Conference on Neural Networks, IEEE, 24-29 July, pp. 2560-2567

Vega, L.R., Rey, H..2013. A Rapid Introduction to Adaptive Filtering, Springer.

Wu C. Y., Lo S., Freedman M. T., Hasegawa A., Zuurbier R. A., Mun S. K. 1994. Classification of microcalcifications in radiographs of pathological specimen for the diagnosis of breast cancer. *Proc. SPIE*. 2167:630–641

Yang, L.2009.Virtual microscopy and grid-enabled decision support for large-scale analysis of imaged pathology specimens. *IEEE T Med Imaging* 13, pp. 636–644.

Z., Jiao, X. Gao, Y., Wang, J., Li. 2016. A deep feature based framework for breast masses classification. *Neurocomputing*.197:221–231