Colorectal Cancer Classification using PCA and Fisherface Feature Extraction Data from Pathology Microscopic Image

Fajri Rakhmat Umbara, Adiyasa Nurfalah, The Houw Liong
Graduate school of Telkom Institute of Technology, St. Telekomunikasi num. 1, Bandung, Indonesia

Keywords:

Colorectal Cancer
Feature Extraction
Fisherface
PCA
Random Tree Algorithm

ABSTRACT

Colorectal cancer is the one of variant cancer which can kill people on this earth. World Health Organization, from their website wrote about 608,000 people can get killed every year because of it. The variant of colorectal cancer such as lymphoma and carcinoma strikes colon from the inside and outside. Lymphoma can be found in white corpuscle and attack colon through lymphocytes, whereas carcinoma can attack the outer layer of colon. Early detection is needed to decrease the number of death because of this cancer.

The study about colorectal cancer is to classified lymphoma, carcinoma, and normal colon. It is doing by using 198 pathology microscopic images data from Hasan Sadikin Hospital in Bandung, Indonesia. Feature extraction using PCA and Fisherface and each generate 2, 5, 10, 50, 100 features. The study compared these two methods and using WEKA to testing the accuracy.

Using 10 folds cross-validation and 3 different classifier in WEKA such as Random Tree, Multi Layer Perceptron, and Naïve Bayes, Fisherface has capability for classified colorectal cancer around 84% - 100% for accuracy. It came from almost all features. Difference result is much visible in PCA. From this result, Fisherface is better than PCA for feature extraction.

Corresponding Author:

Fajri Rakhmat Umbara,
Department of Informatics Graduate School
Telkom Institute of Technology,
Jalan Telekomunikasi Number 1, Bandung, Indonesia.
Email: fajri.umbara@gmail.com

1. INTRODUCTION

Cancer is the uncontrolled growth and spread of cells. It can affect almost any part of the body. The growths often invade surrounding tissue and can metastasize to distant sites (WHO Cancer Fact Sheet). According to data from the Department of Health Republic of Indonesia in 2009, colon cancer is a type of cancer that was ranked the 3rd most suffered by both men and women, with an age range 40-59 years (Dharmais Hospital, National Cancer Center). Similarly, the World Health Organization said that colon cancer has donated 608,000 deaths in 2008 with an estimated it will increase in 2030. Unhealthy lifestyles such as frequent exposure to preservatives and dyes in foods, less exercise, obesity, and tobacco smoke polluted environment are major factor in colon cancer.

Colon cancer (colorectal cancer) is the growth of cancer cells that are in the large intestine (colon) or the area above the anus (rectum). In the medical world often come across several types of colon cancer such as sarcoma, carcinoma and lymphoma. The three types of cancer have different characteristics so it can be a sign that a person may be suffering colon cancer. However, in reality, the majority of patients with colon cancer check the condition when the cancer is already in an advanced stage.

Even though cancer is a deadly disease, it still curable by medical action, such as operation, radiotherapy, or chemotherapy which is suitable regarding the type of cancer owned by the patient while it has been detected earlier. There is a traditional way in detecting cancer tissue, called pathology collation. Pathology collection is processed by observing cell samples under the microscope to determine whether cancer symptom is exist by looking at their abnormality (comparing the observed cell and the healthy one). Therefore, that kind of method has a close relationship to the observation quality conducted by doctor which is possible to produce failure that can influence the diagnostic result.

This research aims to find the best features extraction methods for classifying cancer which is useful for Pathology Division in any hospital to diagnose colorectal cancer in fast and accurate way. The data used in simulation are microscopic images of colorectal gland obtained from Hasan Sadikin Hospital Bandung.

Feature extraction of microscopic images has previously done before data are being classified. This feature extraction process intends to discover some prominent features from a set of data, so that the classification process will be held fast and precise. This research uses Principal Component Analysis (PCA) and Fisherface as its extraction feature methods.

PCA is a statistical method that is usually used in order to execute the extraction feature process. The purpose of this method is to generate some principle components from a set of data by implementing some statistical formulas, such as covariance or singular value decomposition [1][2]. Meanwhile, Fisherface is a method that combines PCA and LDA which is first used in face recognition [3][4]. After the feature extraction process is done, classification process is taking place by using WEKA Classifier Tools that rae Random Tree, Multi Layer Perceptron, and Naïve Bayes. These methods used to testing both feature extraction and calculate the accuracy.

2. RESEARCH METHOD

PCA and Fisherface are the methods for finding a useful information from images [5]. Data from Hasan Sadikin Hospital are images from Pathology Microscopic which divided into 3 classes :

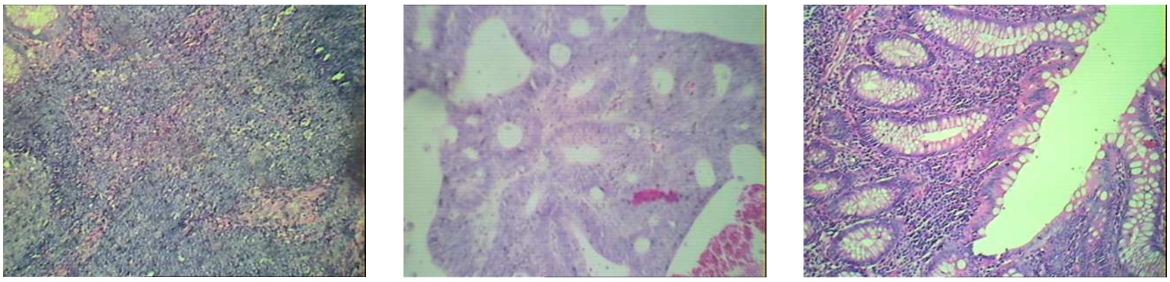

Figure 1. Colon Cancer Images Examples, from left to right : Lymphoma, Carcinoma, Normal Colon

Every images has 66 data and balanced class. Tested using 10 folds cross-validation and 3 different WEKA classifier that are Random Tree, Multi Layer Perceptron, and Naïve Bayes. The flow of research are briefly explained from the diagram below :

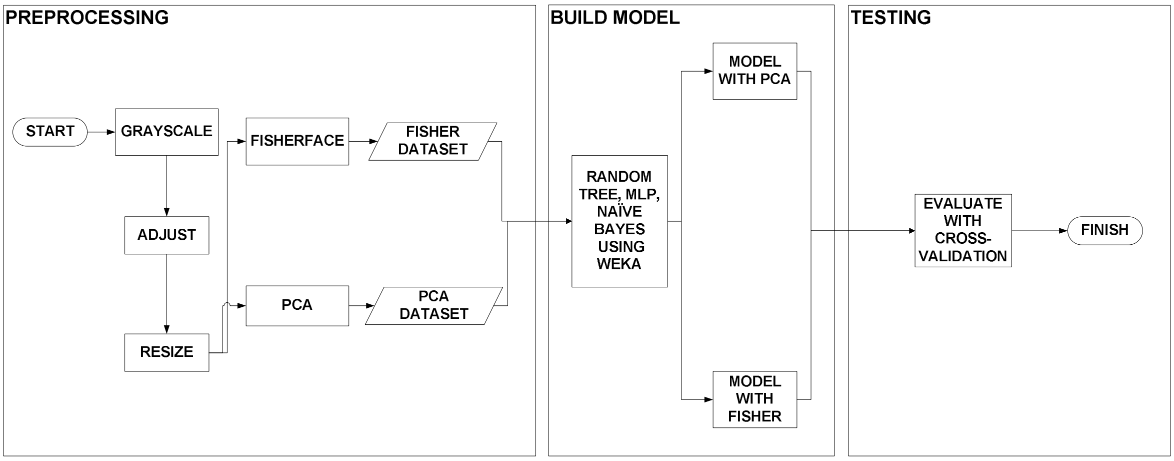

Figure 2. Flow Diagram of Classification Images

PCA and Fisherface can be calculated for grayscale images, so in this research, all of images transformed into a grayscale color. Furthermore, every image has a different size of pixels, so after grayscale, the images must be adjust at the center of observation and resized it into same resolution (255 x 255 pixels).

Using PCA and Fisherface, all images calculated into numerical value using MATLAB. It has 2, 5, 10, 50, and 100 features from all images. PCA [1][2] and Fisherface [3][4] calculated by the following algorithm :

Let an image X (x, y) is two-dimensional array M * N. N number of images in the database are represented as a matrix Γ = {X1, X2, X3, ..., XN}, where each Xi is a vector sized M * N. Matrix Γ which is the input method of Fisherface and PCA also.

PCA process:

1. Calculated the average matrix using formula :

$$\mu = \frac{1}{N}\sum_{k=1}^{N} x_k$$

N is the number of images and x is image vector.

2. Calculated covariant matrix using formula :

$$C = \sum_{k=1}^{N} (x_k - \mu)(x_k - \mu)^T$$

3. Calculated Eigenvalue and Eigenvector from images using formula :

$$CU_n = \lambda_n U_n$$

U is Eigenvector and λ is Eigenvalue.

4. Sorting Eigenvector descending from the largest Eigenvalue.

5. Choose the Eigenvector, the chosen Eigenvector called the Eigenface.

6. Get transformation matrix by projection data into Eigenvector

$$Y = U^T * (X - u)$$

Fisherface process:

1. PCA Process

2. Calculate Within-Class Scatter Matrix (Sw)

$$Sw = \sum_{i=1}^{L}\sum_{k=1}^{N} (Y_k - \mu_i)(Y_k - \mu_i)^T$$

Where L = number of class

3. Calculate Between-Class Scatter Matrix (Sb)

$$Sb = \sum_{i=1}^{L} (\mu_i - \mu)(\mu_i - \mu)^T$$

4. Calculate Eigenvalue and Eigenvector of Sw and Sb

5. Sort Eigenvector based on its best Eigenvalue

6. Get some Eigenvector according to its Eigenvalue

7. Projection PCA transformation matrix into Fisherface Eigenvector (last Eigenvector)

$Y_{fisherface} = U_{fisherface}^T * Y$

From that features, the classification model build using WEKA such as Random Tree Algorithm, Multi Layer Perceptron, and Naïve Bayes. Each classifier using default parameter from WEKA itself.

3. RESULTS AND ANALYSIS

PCA and Fisherface give the result with matrix of feature in 2, 5, 10, 50, and 100 features. For each features used to classify Lymphoma, Carcinoma, and Normal Colon.

Table 1. Ilustration of Projection Matrix which is Result from PCA Process dan Fisherface Process

Num of Data	*Feature 1*	*Feature 2*	*Feature n*	*Class*
1	PCA/FIS$_{1,1}$	PCA/FIS$_{1,2}$	PCA/FIS$_{1,n}$	*Lymphoma*
2	PCA/FIS$_{2,1}$	PCA/FIS$_{2,2}$	PCA/FIS$_{2,n}$	*Carcinoma*
m	PCA/FIS$_{m,1}$	PCA/FIS$_{m,2}$	PCA/FIS$_{m,n}$	*Normal*

The result of implementation using different classifier in WEKA and tested using 10 folds of cross-validation. We choose Random Tree, Multi Layer Perceptron, dan Naïve Bayes to perform this. We set every classifier parameter by default value from WEKA classifier. Testing result are shown in tables below.

Table 2. Testing Result Using Random Tree in WEKA

Feature Extration	**Num of Feature**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Accuracy**
PCA	2	0.395	0.389	0.391	0.542	38.8889
	5	0.536	0.535	0.536	0.652	53.5354
	10	0.505	0.51	0.507	0.633	51.0101
	50	0.412	0.414	0.412	0.561	41.4141
	100	0.479	0.48	0.479	0.61	47.9798
Fisherface	2	1	1	1	1	100
	5	1	1	1	1	100
	10	1	1	1	1	100
	50	0.857	0.854	0.854	0.89	85.3535
	100	0.846	0.843	0.844	0.883	84.3434

Table 3. Testing Result Using Multi Layer Perceptron in WEKA

Feature Extration	**Num of Feature**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Accuracy**
PCA	2	0.458	0.475	0.461	0.662	47.4747
	5	0.531	0.535	0.533	0.678	53.5354
	10	0.515	0.52	0.515	0.681	52.0202
	50	0.576	0.576	0.57	0.756	57.5758
	100	0.446	0.439	0.434	0.637	43.9394
Fisherface	2	1	1	1	1	100
	5	1	1	1	1	100
	10	1	1	1	1	100
	50	1	1	1	1	100
	100	1	1	1	1	100

Table 4. Testing Result Using Naïve Bayes in WEKA

Feature Extration	**Num of Feature**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Accuracy**
PCA	2	0.523	0.53	0.518	0.689	53.0303
	5	0.554	0.566	0.551	0.762	56.5657
	10	0.55	0.551	0.527	0.774	55.0505
	50	0.488	0.505	0.482	0.75	50.5051
	100	0.616	0.586	0.59	0.757	58.5859
Fisherface	2	1	1	1	1	100
	5	1	1	1	1	100
	10	1	1	1	1	100
	50	1	1	1	1	100
	100	1	1	1	1	100

The result of this research means Fisherface as a methods for feature extraction can divide image from colorectal cancer data better than PCA. It can be seen from the value of precision, recall, and f-measure. The ROC area shows that the model which build from classifiers can be used to classify data more accurately.

Cross-validation used to test all of data so that the quality of model which build with classifiers can be tested. The model produced by Fisherface has a better quality than PCA, so it can cause the number of accuracy from the prediction is good too.

4. CONCLUSION

From the testing shown Fisherface outperform PCA both using 3 different classifier in WEKA for each features. Constant and higher number of precision, recall, and f-measure prove that data which produced with Fisherface is better than PCA. Its clear, because in Belheumer et.al. [1] also shown Fisherface outperform PCA in case Face Recognition. Using 10 folds of cross-validation for each classifiers was produced a good result. The number of accuracy is around 84.3434 % - 100% for each classifier which using data from Fisherface. From the number of ROC area, the quality of model is also good, so that the model can be used for predicted unclassified data.

Many interesting issues in cancer field in the future especially in cancer classification using microscopic pathology image data. The immediate improvement of the presented method is to implement others feature extraction method that based on image processing such as Gabor Wavelet Transform and Edge Detection.

ACKNOWLEDGEMENTS

Authors wish to thank to Prof. The Houw Liong for guidance and feedback during this research work. Authors also thank to the staff of Hasan Sadikin Hospital especially dr. Anton and dr. Nita in pathology division for useful discussions on the subject of cancer.

REFERENCES

[1] M. Turk, A. Pentland, “Eigenfaces for Recognition”, Journal of Cognitive Neuroscience, Vol. 3, No. 1, pp 71-86, 1991.

[2] D. Stathis and D. Myronidis, “Principal Component Analysis of Precipitation In Thessaly Region (Central Greece)”, Global NEST Journal, Vol 11 no 4, pp 467-476, 2009.

[3] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, “Eigenfaces vs.Fisherfaces: Recognition using class specific linear projection”, IEEE Transactions Pattern Analysis Machine Intelligent, Vol. 19, pp 711-720, 1997.

[4] N. Adiyasa, “Pengenalan Individu Melalui Citra Wajah Menggunakan Metode Fisherface dan Jaringan Syaraf Tiruan Backpropagation”, Institut Teknologi Telkom, Bandung, 2011.

[5] M. Sharkas, M.A. Elenien, “ Eigenface vs Fisherface vs ICA for Face Recognation ; a Comparative Study”, Signal Processing 2008, ICSP 2008, 9th International Conference I, *2008.*

[6] Y. Zhao and Y. Zhang, “Comparison of Decision Tree Methods for Finding Active Objects”, Advance of Space research, 2007.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, “The WEKA Data Mining Software : an Update”, ACM SIGKDD Explorations Newsletter, Vol 11 issue 1, 2009.

[8] J. Yang, J.Y. Yang, “Why Can LDA be Performed in PCA transformed Space?”, Elsevier, The Journal of Pattern Recognition Society,2003, Pattern Recognition 36 (2003) 563 – 566.

[9] M. O’Neil and I. Damjanov, “Histopathology of Colorectal Cancer after Neoadjuvant Chemoradiation Therapy”, The Open Pathology Journal, 2009.

BIBLIOGRAPHY OF AUTHORS

	Fajri Rakhmat Umbara was born in Palembang, Indonesia, March 13 1988. He already finished his Bachelor Degree at Telkom Institute of Technology since 2012. Now, he continues his study at Master Degree of Informatics Data Mining in Telkom Institute of Technology. He already has several local publications like writing books and local research. This is the first International publication for him. He wants to be a lecturer after he finished his study from current University. For more information about him, please contact with his email at fajri.umbara@gmail.com or his LinkedIn at www.linkedin.com/in/fajriumbara.

	Adiyasa Nurfalah was born in Subang, West java, Indonesia, March 27 1988. He got his Diploma Degree from Politeknik Negeri Bandung in 2008, then he continued his study, Bachelor Degree at Telkom Institute of Technology and finished it in 2011. Now he still studies in the same campus at Master Degree of Informatics Data Mining while work at his own Campus as a Web Programmer. For more information about him, please contact with his email at adiyasa.nurfalah@gmail.com or add his Facebook account at www.facebook.com/adiyasan.
	Prof. The Houw Liong, he is a Professor in Department of Informatics, Graduate School, Telkom Institute of Technology since 2009. He was finished his doctor degree, Ph D of Physics in University of Kentucky at 1968. Besides being a lecturer, he is a researcher too. He began his career as a lecturer and researcher in Bandung Institute of Technology at 1963. He also a lecturer in other famous University in Bandung, such as ST Inten and Harapan Bangsa Institute of Technology and to be a researcher at UPTHB BPPT and LAPAN. He has many publications from National and International Level since the beginning of his career. For more information about him, please contact with his email at thehl007@gmail.com or visit his LinkedIn at www.linkedin.com/in/houwliong.