Comparative analysis of PCA and KPCA on paddy growth stages classification

Hendra Halim[1], Sani M. Isa[2] (*Author*)
Master of Information Technology
Bina Nusantara University
Jakarta, Indonesia
hendrahalim92@gmail.com[1], sani.m.isa@binus.ac.id[2]

Sidik Mulyono (*Author*)
Agency For The Assessment And Application Of Technology (BPPT)
Jakarta, Indonesia
sidik.mulyono@bppt.go.id

***Abstract*— Hyperspectral image is capable to distinguish paddy growth stages with classification methods. Hyperspectral has disadvantages. One of the disadvantages is hyperspectral image has high dimensionality that can cause curse of dimensionality. In this paper, PCA and Kernel PCA are used to reduce the dimension of hyperspectral data. The objective in this research is to analyze the effect of using dimension reduction techniques on hyperspectral data on paddy growth stages classification. The result will show the effect of dimension reduction techniques whether it is capable to improve the classification accuracy and execution time.**

***Keywords*— PCA; Kernel PCA; Hyperspectral; Growth stages; Classification;**

I. INTRODUCTION

Indonesia is one of Asia countries that has rice as staple food. Lacking of rice production can impact to Indonesia food security directly [1]. According to Statistics Indonesia in 2012, Indonesia had produced rice about 65.740.946 ton but it was insufficient to fulfil the needs of rice. Import became the only way to fulfil the needs of food. But in rice import calculation, there was problems because of lack of accurate information about rice productivity in Indonesia. The lack of accurate information was because of variety of planting time and using conventional method to calculate rice production [2]. Field officer of district agricultural bureau usually carries out harvest area prediction by eyes-estimation around some sampling areas of paddy field, without measuring the real harvested area. It is time consuming and laborious to gather and compile the data from district level to national level. At the national level, the gathered data is lack of validity because the method is used to estimate the harvest area at district level is neither precise nor accurate [3]. Accurate harvest area calculation can be predicted based on the paddy growth stage at that time [2].

Nowadays, remote sensing and hyperspectral image can solve paddy growth stage determination problem [2]. According to Sidik Mulyono et al., hyperspectral remote sensing is able to extract spectral information that uniquely characterizes and identifies the chemical, moisture, and physical properties of the constituent parts of an input object, scene region, or an agricultural product [4]. Hyperspectral image is an image that contains information from various electromagnetic spectrum that is saved in stack of layers. Each of layers have electromagnetic spectrum range that is called spectral band [2]. Some of airborne and space remote sensing tools that have over than 200 bands, have been developed and used. Because of the hyperspectral image have narrow bands, the gathered data must be compared spectral object reflectance characteristic from spectral library or field observation. Field observation gjves additional information for image preprocessing and processing from remote sensing, especially airborne remote sensing when is used on satellite or multispectral or hypersectral airborne sensor [5]. But hyperspectral has disadvantages. One of the main disadvantages is often because of lack of sample in training data. It is because the expensive cost needed to conduct data gathering in the field [4].

Paddy growth stages can be determined by classification methods. Classification of paddy growth stages need a right classifier model to have a high accuracy [2]. Some of well-known classification methods are Support Vector Machine (SVM) and Naïve Bayes. The decision to use SVM and Naïve Bayes in this research is because SVM and Naïve Bayes are popular classifier in hyperspectral image classification. SVM and Naïve Bayes are used as classifier in classification of contaminants from wheat using near-infrared hyperspectral imaging [6]. SVM and Naïve Bayes also used in classification of Local Climate Zone Based on Multiple Earth Observation Data [7]. SVM also used in hyperspectral data classification to distinguish the land cover [8]. Because of the popularity, SVM and Naïve Bayes are used in this research.

According to Senthilnath et al., hyperspectral image can distinguish paddy growth stages based on paddy spectral reflectance [9]. Maspiyanti et al. state that the dimensions of hyperspectral data are the features that can be used in classification to determine the paddy growth stage. But because of having too many features, it can cause curse of dimensionality that the more features do not always have maximum accuracy with a chance to remove the ineffective features to have optimum accuracy [2].

Dimension reduction is one of techniques that can solve the curse of dimensionality. Dimension reduction is a process of reducing variable numbers at certain condition according to Han et al. [10]. Fang et al. mentioned that one of the main methods in dimension reduction is feature extraction [11]. Feature extraction is a process to construct a low-dimensional

978-1-5090-0931-2/16/$31.00 ©2016 IEEE

representation that reduces redundancy because of high dimensionality data according to Subasi and Gursoy in 2010 [12]. Cunningham in 2008 state that Principal Component Analysis (PCA) is one of the well-known methods for unsupervised feature extraction [13]. According to L.J.P. van der Maaten in 2007, PCA is a dimension reduction techniques that linearly transforms a high dimensional data into a low dimensional data. PCA can handle a nonlinear data with a kernel function that is called Kernel PCA (KPCA). Kernel PCA (KPCA) is a reformulation of traditional linear PCA in a high dimensional space that is constructed using a kernel function [14].

In this paper, PCA and KPCA are used on Ultra genjah growth stages classification. This research will analyze the comparison of the classification accuracy and execution time between Ultra genjah data without dimension reduction and dimension reduction Ultra genjah data. The result of this research will show the dimension reduction techniques can improve the classification accuracy and/or the execution time on paddy growth classification.

II. DIMENSION REDUCTION TECHNIQUE

Principal Component Analysis (PCA) is a well-known method for dimension reduction. PCA linearly transforms a high dimensional input vector into a low-dimensional one whose components are uncorrelated [15]. PCA construct a low-dimensional representation of the data that describes as much of the variance in the data as possible. This is done by finding a linear basis of reducted dimensionality for the data, in which the amount of the variance in the data is maximal [14]. PCA has been successful in classification with Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) to detect insect-damage wheat kernels using near-infrared hyperspectral imaging. The classification accuracy reached between 85% until 100% with only 2 PC where the first PC (PC 1) variability is 94% and the second PC (PC 2) is 5%. PCA also has been successful in detection of fusarium damaged kernels in Canada Western Red Spring wheat using visible/near-infrared hyperspectral imaging. PCA reduced the data dimension to 10 features/PC and with LDA as classifier, the accuracy reach 92%.

Kernel PCA (KPCA) is a reformulation of traditional linear PCA in a high dimensional space that is constructed using a kernel function. KPCA computes the principal eigenvectors of the kernel matrix, rather than those of the covariance matrix. The reformulation of traditional PCA is straightforward, since a kernel matrix is similar to the inproduct of the data points in the high-dimensional space that is constructed using the kernel function. The application of PCA in kernel space provides KPCA the property of constructing nonlinear mappings. Since KPCA is a kernel-based method, the mapping performed by KPCA highly relies on the choice of the kernel function. Possible choices for kernel function are Linear kernel, Polynomial kernel, and Gaussian kernel [14]. KPCA has been successful to improve the classification accuracy in face recognition [16]. KPCA also provide a better result than PCA and Independent Component Analysis (ICA) in classification of hyperspectral data over urban areas based on extended morphological profile with partial reconstruction [17].

III. METHODOLOGY

The objective of this research is to analyze the impact of implementation PCA and KPCA on paddy growth stages classification. The processes of this research are consist of 3 steps i.e. dimension reduction process, classification process, and analysis process. The processes can be seen in Figure 1.

The data are distinguished into 2 types i.e. data with 826 features and dimension reduction data. The dimension reduction data are the output of dimension reduction process. Besides the data, the output of the dimension reduction process is the execution time of dimension reduction. The dimension reduction data are consist of data that are dimension reduced using Principal Component Analysis (PCA), Kernel PCA (KPCA) Gaussian kernel, KPCA Linear kernel, and KPCA Polynomial kernel. Each of the dimension reduction data are consist of data that have 1 feature/Principal Components (PC) until 10 PC. Because of the value of execution time in dimension reduction process is not definite, the dimension reduction process for each dataset is repeated 20 times to earn an average value of execution time. The output of the dimension reduction process are 40 datasets of dimension reduction data and the average value of execution time. As the input of the classification process, there are 40 datasets of dimension reduction data and 1 data with 826 features that are used in this research.

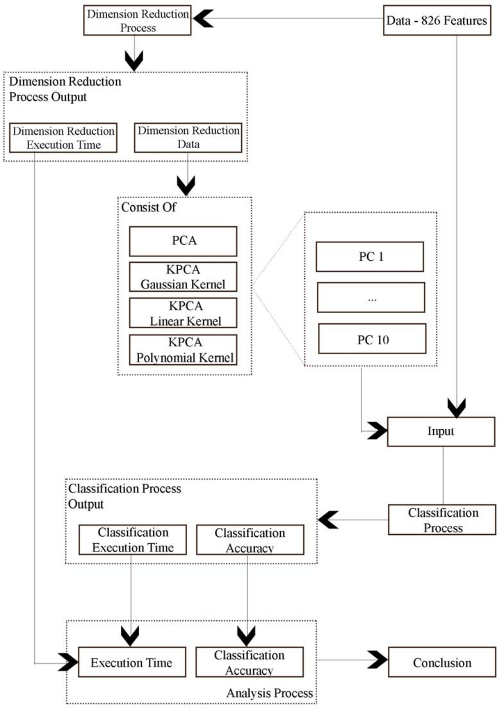

Figure 1. Methodology Flow Process

Each of the dataset is consist of 6 classes/growth stages. These dataset are used as an input in the classification process. The classification process use SVM Linear and Naïve Bayes as classifier. The result of this process is evaluated using 10-folds cross validation method. The outputs of the classification process are the classification accuracy and the execution time of classification process. Like the execution time of the dimension reduction process, the execution time in the classification process is not definite either. The classification process is looped 100.000 times to earn an average value of the execution time.

In the analysis process, there are 3 metrics that are analyzed. The metrics are the classification accuracy, the total execution time, and f-measure. The total execution time is sum of the execution time of dimension reduction process and execution time of classification process. In other words, the execution time of the data without dimension reduction process is only from the classification process. The output of analysis process is the conclusion of the analysis of the result.

This research use Ultra Genjah data that were gathered by Agency For The Assessment And Application Of Technology/Badan Pengkajian Dan Penerapan Teknologi (BPPT) in May 2012. The data were result of BPPT's research that had objective to create paddy spectral library from field observation in Subang. There were two variety of paddy that were used in the experiment i.e. Ultra genjah and Ciherang. The experiment was divided into 3 main fields, 2 fields for Ultra Genjah and 1 field for Ciherang, where each field was given different treatment. The differences of the treatment were in the doses of Nitrogen for fertilization that are 0, 45, 90, 135 kg N/ha, and time for distributing the Nitrogen that were 2 times and 3 times each season. Each measurement of spectral data, which were given 3 times Nitrogen distribution, was taken 5 times to get an average value that could represent the spectral data itself. The spectral data were gathered using spectrometer, ILT900 and other additional tools like a sensor, Spektralon white reference, SpectrlLight software, netbook, spectroradiometer sensor holder, netbook holder, digital camera, and Global Positioning System (GPS). The gathered data were through pre-processing steps i.e. selection, average, smoothing and cropping. The selection and average process were done by taking measurement of the field that was given 3 times Nitrogen distribution each season 5 times. The result of the measurement was averaged to get a representative value. Then the smoothing process was done using Savitzky-Golay method. The cropping process was done based on the capability of ILT900 that works effectively with wavelength between 250-950 nm. The result of the cropping process showed that the wavelength than 400 nm and the wavelength greater than 900 nm were not good enough so the wavelength were cropped between 400-900 nm. [18]

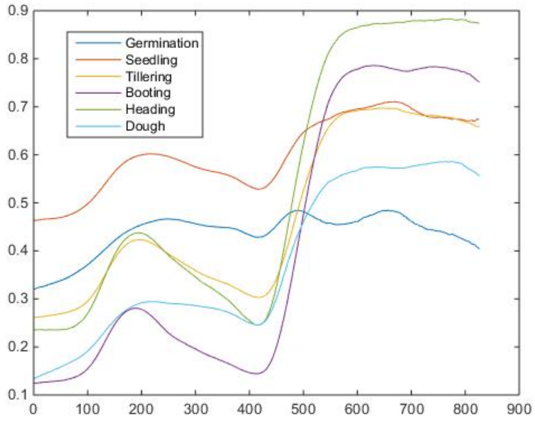

Figure 2. Ultra Genjah data plot for each growth stage

IV. EXPERIMENT AND RESULT

There are two software that are used in this research. They are Rapid Miner and MATLAB. Rapid Miner is used for classification and MATLAB is used for dimension reduction. The dimension reduction algorithms in MATLAB are using the toolbox for dimension reduction that were created by L.J.P. van der Maaten [14].

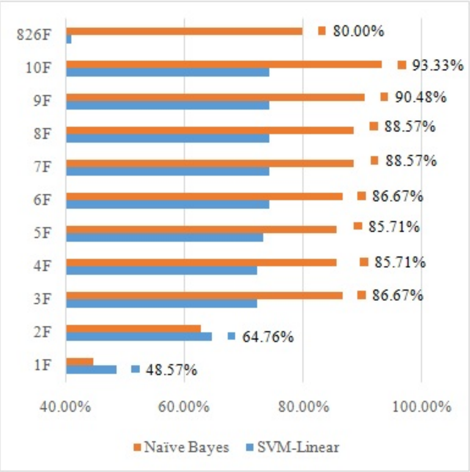

Figure 3. Comparison of classification accuracy among the data without dimension reduction and dimension reduction data using PCA.

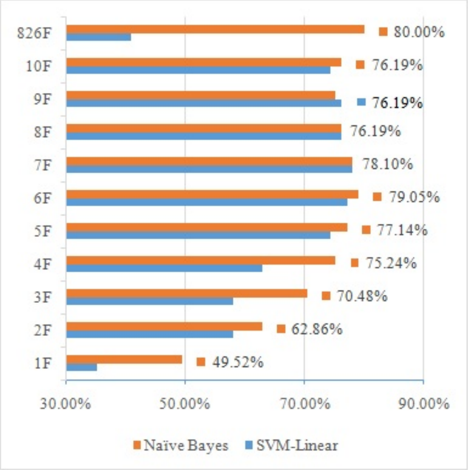

Figure 4. Comparison of classification accuracy among the data without dimension reduction and dimension reduction data using KPCA Gaussian kernel.

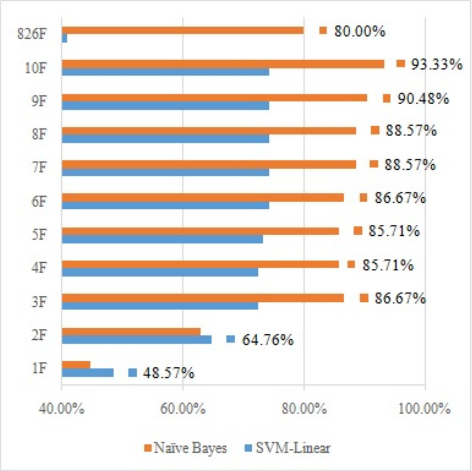

Figure 5. Comparison of classification accuracy among the data without dimension reduction and dimension reduction data using KPCA Linear kernel.

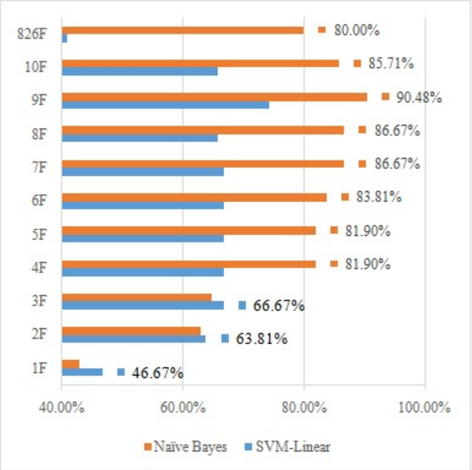

Figure 6. Comparison of classification accuracy among the data without

Figure 3 shows the comparison of classification accuracy among the data without dimension reduction and dimension reduction data from PC 1 until PC 10 with PCA. The classification accuracy from PC 3 until PC 10 are better than the classification accuracy with data with 826 features. The classification accuracy using Naïve Bayes as classifier are higher than using SVM Linear except for PC 1 and PC 2. PCA can improve the classification accuracy over 80%. The highest classification accuracy is at PC 10. From this result, PCA is capable to improve the classification accuracy with reduce the data dimension from 826 features to 3 features with higher classification accuracy. Figure 4 shows the comparison of classification accuracy among the data without the dimension reduction and dimension reduction data from PC 1 until PC 10 with KPCA Gaussian kernel. KPCA with Gaussian kernel is not effective to increase the classification accuracy. The classification accuracy using Naïve Bayes as classifier are higher than using SVM Linear except for PC 7, PC 8, and PC 9. Even though KPCA Gaussian kernel is not improve the classification accuracy but KPCA Gaussian kernel can reduce the data dimension from 826 features to 6 features with decrease the classification accuracy to 79.05%. It shows that the dimension reduction with KPCA Gaussian kernel decreases the classification accuracy, but it may affect much in the execution time. Figure 5 shows the comparison of classification accuracy among the data without the dimension reduction and the data from PC 1 until PC 10 with KPCA Linear kernel. KPCA with Linear kernel has the same result with the PCA. There is no difference in classification accuracy between PCA and KPCA with linear kernel. The highest accuracy is at PC 10. Besides improve the classification accuracy, KPCA Linear kernel is also capable to reduce the data dimension from 826 features to 3 features with higher classification accuracy. In Figure 6 shows the comparison of classification accuracy among the data without dimension reduction and the dimension reduction data from PC 1 until PC 10 with KPCA Polynomial kernel. The classification accuracy using Naïve Bayes as classifier are higher than using SVM Linear except for PC 1, PC 2, and PC3. KPCA Polynomial kernel can improve the classification accuracy to the best at PC 9. KPCA Polynomial kernel is capable to reduce the data dimension from 826 features to 4 features with higher classification accuracy.

From the result, the classification accuracy with the data without dimension reduction and dimension reduction data are higher using Naïve Bayes as classifier than using SVM linear as classifier. The highest classification accuracy is at PC 10 that is using PCA or KPCA with linear kernel data and Naïve Bayes as classifier. Classification PC 10 for PCA and KPCA with Linear Kernel also produce f-measure value over 85% for each class that means the classifier can distinguish those classes very well. The detail of precision is showed in table 1. Besides KPCA Gaussian kernel, the other dimension reduction techniques are capable to reduce the data dimension with higher classification accuracy.

Table 1. Confusion Matrix for Classification of 10 features of PCA data using Naïve Bayes

True Prediction	Germination	Seedling	Tillering	Booting	Heading	Dough	Class Precision
Germination	12	1	0	0	0	0	92.31%
Seedling	3	13	0	0	0	0	81.25%
Tillering	0	0	14	1	0	0	93.33%
Booting	0	0	1	30	0	0	96.77%
Heading	0	0	0	0	14	1	93.33%
Dough	0	0	0	0	0	15	100.00%
Class Recall	80.00%	92.86%	93.33%	96.77%	100.00%	93.75%	
F-Measure	85.72%	86.67%	93.33%	96.77%	96.55%	96.77%	

Figure 7 shows the comparison of execution time of dimension reduction data with PCA from PC 1 until PC 10. It shows the execution time of Naïve Bayes as classifier is faster than SVM Linear. The Execution time of SVM Linear as classifier is ±1.7 until 2.1 times slower than the execution time of Naïve Bayes as classifier. Figure 8 shows the comparison of execution time of dimension reduction data with KPCA Gaussian kernel from PC 1 until PC 10. It shows the execution time of Naïve Bayes as classifier is faster than SVM Linear. The Execution time of SVM Linear as classifier is ±1.5 times slower than the execution time of Naïve Bayes as classifier. Figure 9 shows the comparison of execution time of dimension reduction data with KPCA Linear kernel from PC 1 until PC 10. It shows the execution time of Naïve Bayes as classifier is faster than SVM Linear. The execution time of SVM Linear as classifier is ±1.5 until 2 times slower than the execution time of Naïve Bayes as classifier. Figure 10 shows the comparison of execution time of dimension reduction data with KPCA Polynomial kernel from PC 1 until PC 10. It shows the execution time of Naïve Bayes as classifier is faster than SVM Linear. The execution time of SVM Linear as classifier is ±1.3 until 1.8 times slower than the execution time of Naïve Bayes as classifier.

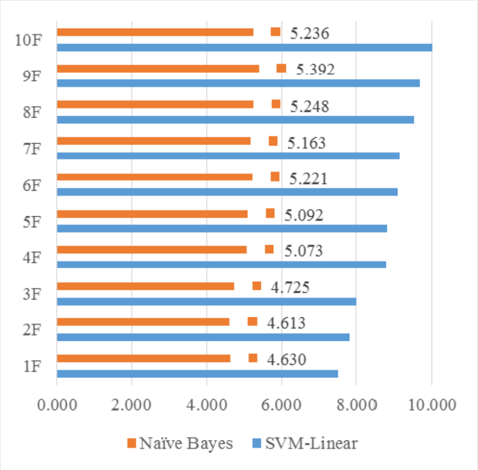

Figure 7. Time Comparison among PCA data from PC 1 until PC 10 in millisecond.

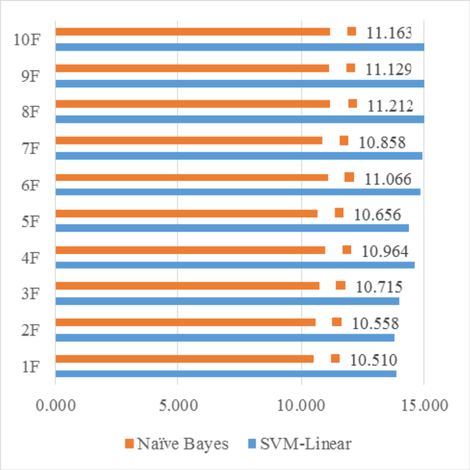

Figure 8. Time Comparison among KPCA Gaussian kernel data from PC 1 until PC 10 in millisecond.

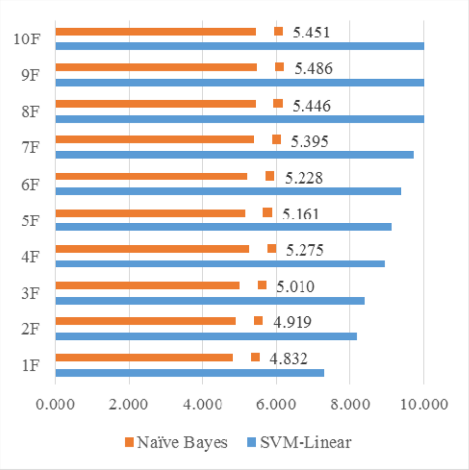

Figure 9. Time Comparison among KPCA Linear kernel data from PC 1 until PC 10 in millisecond.

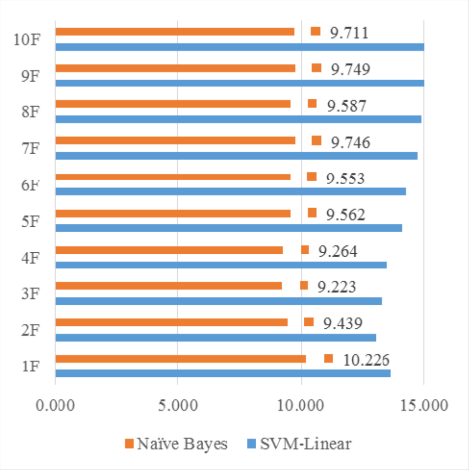

Figure 10. Time Comparison among KPCA Polynomial kernel data from PC 1 until PC 10 in millisecond.

From the result, the classification using Naïve Bayes as classifier have faster execution time than using SVM Linear with PCA or KPCA with every kernel method. It also shows that the dimension reduction techniques are capable to reduce the execution time.

The execution time for the data without dimension reduction for both classfiers are 364.06 ms for SVM and 65.88 ms for Naïve Bayes. Table 2 shows the comparison of execution time of data without dimension reduction and execution time from data using dimension reduction with highest accuracy in millisecond. From those table, the execution time of the dimension reduction data are faster than the data without dimension reduction. The execution time of data with 826 features using SVM Linear as classifier are around 24 until 40 times slower than the execution time of the dimension reduction data. The execution time of data with 826 features using Naïve Bayes as classifier are around 6 until 12 times slower than the execution time of the dimension reduction data.

Table 2. Time Comparison between data without dimension reduction and data with dimension reduction with highest accuracy.

	SVM	Naïve Bayes
Without Dimension Reduction	364.06 ms	65.88 ms
PCA	9.10 ms	5.24 ms
KPCA (Gaussian)	14.97 ms	11.07 ms
KPCA (Linear)	9.40 ms	5.45 ms
KPCA (Polynomial)	13.27 ms	9.75 ms

V. CONCLUSION AND FUTURE WORK

The objective of this research is to analyze the effect of dimension reduction techniques on paddy growth stages classification. The classification accuracy with PCA or KPCA data are higher using Naïve Bayes as the classifier than SVM linear as classifier. The highest classification accuracy is 93.33% which is using PCA or KPCA with linear kernel data and Naïve Bayes as the classifier. In PCA data and KPCA with linear kernel with Naïve Bayes as classifier, PC 10 have the same classification accuracy that is 93.33% but there is process time difference. PC 10 with PCA data process time is 3.85% faster than PC 10 with KPCA linear kernel process time. In PC10, besides have the highest accuracy, the f-measure values are higher than 85% for each class that means the classifier with the reducted dimension data can distinguish the classes very well. There are some consideration in classification accuracy and process time, even though the classification accuracy are not the highest. In PCA data, PC 3 classification accuracy is 86.67% which is lower 7.14% than PC10 but the process time is 9.76% faster than PC 10 process time. In KPCA with linear kernel and KPCA with polynomial kernel also have 86.67% classification accuracy at PC 3 and PC 6 for linear kernel, and PC 7 and PC 8 for Polynomial kernel but the process time is not faster than PC 3 in PCA data. Out of all PC that have classification accuracy greater than 80%, there is no PC that is faster than PC 3 in PCA data with 86.67% classification accuracy. From the result, it shows that PCA and KPCA are capable to improve the execution time and/or the classification accuracy.

PCA and KPCA are unsupervised dimension reduction techniques. The Future work should focus with supervised dimension reduction techniques. Using supervised dimension reduction, there is a possibility to have higher classification accuracy.

ACKNOWLEDGMENT

This research is supported by BPPT and BINUS University. Thanks to Sidik Mulyono from BPPT that allow us to use the spectral data. Thanks to BINUS University for supporting the students to publish their research. Thanks to Sani M. Isa for being my research supervisor.

REFERENCES

[1] T. Takayama, A. Uchida, H. Sekine, K. Fukuhara, K. Yoshida, O. Kashimu, S. Muljono, D. Arief, M. Evri, and M. Sadly, "VALIDATION OF BIPLS FOR IMPROVING YIELD ESTIMATION OF RICE PADDY FROM HYPERSPECTRAL DATA IN WEST JAVA , INDONESIA Mitsubishi Research Institute , Inc ., 2-10-3 , Nagata-cho , Chiyoda-ku , Tokyo 100-8141 , JAPAN Japan Space Systems , JAPAN Agency for the A," pp. 6581–6584, 2012.

[2] F. Maspiyanti, M. I. Fanany, and A. M. Arymurthy, "Klasifikasi Fase Pertumbuhan Padi Berdasarkan Citra Hiperspektral dengan Modifikasi Logika Fuzzy (Paddy Growth Stages Classification based on Hyperspectral Image using Modified Fuzzy Logic)," *J. Penginderaan Jauh*, vol. 10, pp. 41–48, 2013.

[3] S. Mulyono, M. I. Fanany, and T. Basaruddin, "A paddy growth stages classification using MODIS remote sensing images with balanced branches support vector machines A Paddy Growth Stages Classification Using MODIS Remote Sensing Images with Balanced Branches Support Vector Machines," no. September 2015, 2012.

[4] S. Mulyono, M. I. Fanany, and T. Basaruddin, "Genetic Algorithm Based New Sequence of Principal Component Regression (GA-NSPCR) For Feature Selection And Yield Prediction Using Hyperspectral Remote Sensing Data," *Int. Geosci. Remote Sens. Symp.*, pp. 4198–4201, 2012.

[5] F. Yamazaki, K. Hara, and W. Liu, "Urban land-cover classification based on airborne hyperspectral data and field observation," p. 92440P, 2014.

[6] L. Ravikanth, C. B. Singh, D. S. Jayas, and N. D. G. White, "Classification of contaminants from wheat using near-infrared hyperspectral imaging," *Biosyst. Eng.*, vol. 135, no. November, pp. 73–86, 2015.

[7] B. Bechtel and C. Daneke, "Classification of local climate zones based on multiple earth observation data," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 5, no. 4, pp. 1191–1202, 2012.

[8] C. N. Ozturk and G. Bilgin, "A comparative study on manifold learning of hyperspectral data for land cover classification," vol. 9443, no. Icgip 2014, p. 94431L, 2015.

[9] J. Senthilnath, S. N. Omkar, V. Mani, N. Karnwal, and S. P. B., "Crop Stage Classification of Hyperspectral Data Using Unsupervised Techniques," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, pp. 1–7, 2012.

[10] J. Han, M. Kamber, and P. Jian, *Data Mining : Concepts and Techniques Third Edition*. Morgan Kaufmann, 2011.

[11] L. Fang, H. Zhao, P. Wang, M. Yu, J. Yan, W. Cheng, and P. Chen, "Feature selection method based on mutual information and class separability for dimension reduction in multidimensional time series for clinical data," *Biomed. Signal Process. Control*, vol. 21, pp. 82–89, 2015.

[12] A. Subasi and M. I. Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8659–8666, 2010.

[13] P. Cunningham, *Machine Learning Techniques for Multimedia Chapter 4 - Dimension Reduction*, no. 11. Springer Berlin Heidelberg, 2008.

[14] L. J. P. van der Maaten, "An Introduction to Dimensionality Reduction Using Matlab," *Tech. Rep. MICC*, 2007.

[15] L. J. Cao, K. S. Chua, W. K. Chong, H. P. Lee, and Q. M. Gu, "A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine," *Neurocomputing*, vol. 55, no. 1–2, pp. 321–336, 2003.

[16] Q. Wang, "Kernel Principal Component Analysis and its Applications in Face Recognition and Active Shape Models," 2012.

[17] W. Liao, R. Bellens, A. Pizurica, W. Philips, and Y. Pi, "Classification of Hyperspectral Data Over Urban Areas Using Directional Morphological Profiles and Semi-Supervised Feature Extraction," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 5, no. 4, pp. 1177–1190, 2012.

[18] M. E. Dr. Ir. Sidik Mulyono, *Teknologi Hiperspektral Untuk Pemetaan Sentra Produksi Pertanian*. 2012.