Comparative Study of PCA and LDA for Rice Seeds Quality Inspection

Samson Damilola Fabiyi*, Hai Vu†, Christos Tachtatzis*, Paul Murray*, David Harle*,
Trung-Kien Dao†, Ivan Andonovic*, Jinchang Ren*, Stephen Marshall*
*Department of Electronic and Electrical Engineering
University of Strathclyde
Glasgow, United Kingdom
samson.fabiyi@strath.ac.uk
†International Research Institute MICA
Hanoi University of Science and Technology
Hanoi, Vietnam

***Abstract*—Contamination of rice seeds affects the crop quality, yield and price. Inspection of rice seeds for purity is a very important step for quality assessment. Promising results have been achieved using hyperspectral imaging (HSI) for classification of rice seeds. However, the relatively high number of spectral features in HSI data continues to pose problems during classification which necessitates the use of techniques like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) for dimensionality reduction and feature extraction. This paper presents a comparative study of LDA and PCA as dimensionality reduction techniques for classification of rice seeds using hyperspectral imaging. The results of LDA and PCA on spectral features extracted from hyperspectral images were used for classification using a Random Forest (RF) classifier. Classification results shows that LDA is a superior dimensionality reduction technique to PCA for quality inspection of rice seeds using hyperspectral imaging.**

Keywords—rice seed variety, hyperspectral imaging, PCA, LDA

I. INTRODUCTION

In the Agri-tech sector, inspection of rice seeds purity is a very important step for quality assessment. Professional inspectors operating at seedling propagation stations find it difficult to put in place and maintain pure rice seeds. Rice seeds become contaminated through the introduction of weeds and off-types which make them vulnerable to disease. Consequently, the quality and price of rice seeds can be affected. The rice seed screening exercise can be enhanced by the use of automatic inspection systems which require less time and effort when compared to visual inspection.

The application of hyperspectral imaging for inspecting rice seeds quality has produced very promising results [1, 2]. Hyperspectral images are characterised by the presence of hundreds or thousands of spectral channels such that each pixel in the spatial domain contains an entire spectrum of reflected light; the spectral range is determined by the system used for acquisition. This high number of spectral features is responsible for the problem of curse of dimensionality which poses a challenge for accurate classification of hyperspectral images. Hence, there is a need to remove redundant data and retain important information with very minimal loss through the process of dimensionality reduction and feature extraction. Examples of techniques capable of achieving this are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Application of these techniques reduces dimensions of hyperspectral imaging data and improves the classification results.

Table 1 presents a summary of existing dimensionality reduction techniques used by various authors as a pre-processing step when classifying rice seeds. It is noted from the related work reported, that no comparison of the effectiveness of these two techniques for inspection of rice seeds purity is studied and that PCA is the most commonly applied technique to deal with the problem of high dimensionality for rice seeds classification.

In this paper, firstly, a dataset of rice seeds containing 256 spectral features is prepared. Secondly, PCA and LDA are used to reduce the dimensions of spectra features of HSI data before a Random Forest classifier is trained on each dataset and used for rice seed species classification. The performance of the classifier is evaluated on a dataset containing twenty different rice seed species – significantly more than alternative approaches that have already been proposed in the literature. This paper aims to make a comparative study of the performance of PCA and LDA as dimensionality reduction techniques for classification of rice seeds.

II. METHODS

A. Data Acquisition and Description

Twenty rice seeds varieties were obtained from the National Centre of Protection of New Varieties and Goods of Plants (NCPNVGGP) in Vietnam. 96 rice seeds obtained for each of the 20 species (making a total of 1,920 rice seeds) were imaged using a Visible-range pushbroom HSI system consisting of a Specim V10E Imaging Spectrograph and Hamamatsu ORCA-05G CCD camera to form a rice seeds dataset. The spectral range of the system is: ~ [385 - 1000] nm. The 20 species considered in this study are: N54, VietHuong8, CTX30, TQ14, H229, NC2, NepThomHungYen, 9d, HS1, DTH155, BT6, NN4B, MyHuong88, BacThomSo7, TB13, TruongXuanHQ, BC15, ND9, KL25, and ThuanViet2.

B. Dimensionality Reduction Using PCA and LDA

Principal Component Analysis (PCA) is a common statistical analysis technique applied for hyperspectral data cubes. PCA is an unsupervised dimensionality reduction technique which does not use the labels provided in the dataset. For our application, PCA aims to reduce high

TABLE 1. A SURVEY ON DIMENSIONALITY REDUCTION TECHNIQUES (DRT) FOR RICE SEEDS CLASSIFICATION

Reference	Number of Varieties	DRT	Classifiers	Modality	Year
[1]	6	PCA	SVM, RF	HSI	2016
[2]	6	PCA	SVM, RF	HSI	2016
[3]	5	PCA	PNN	Electronic nose systems	2012
[4]	5	PCA, Segmented PCA	SAM	HSI	2010
[5]	2	PCA, LLE	SVM, KNN	Electronic nose and tongue	2015
[6]	3	PCA	BPNN	HSI	2014
[7]	5	PCA	BPNN, PLSDA, LS-SVM	Multi spectral imaging	2016
[8]	2	PCA	PNN	HSI	2011
[9]	6	PCA	SVM	Fluorescence technology	2017
[10]	9	PCA	ANOVA	RGB	2012

dimensionality using the spectral information in the rice seeds dataset. In the HSI context, PCA transforms the spectral data to linearly uncorrelated (orthogonal) components and subsequently ranks the components based on the percentage of variance attributed to each of the components in the dataset. In this study, the whole dataset was collected to evaluate variance rate using PCA on the spectral features (originally 256 in dimensions). We observed the percentage of variance explained by 1 to 256 principal components (PCs) and illustrate this in Figure 1. For instance, if 50 components which explains 99.999% of the total variance are used (instead of the original data for 256 wavelengths) for further tasks such as classification or discriminant analysis of the rice seed species, only a small loss of information occurs (~0.001%).

Linear Discriminant analysis (LDA) is another statistical analysis technique that can be applied to reduce the dimensionality of hyperspectral data cubes and extract features for classification. LDA is a supervised dimensionality reduction technique which makes use of labels alongside the features in the dataset. For our application, LDA also aims to reduce high dimensionality using the spectral information in the rice seeds dataset. LDA reduces dimensions of spectra features (initially 256) by maximizing the between-species variance and minimizing the within-species variance of the 90 species. In this way, LDA can maximize the separability of species in the rice seeds datasets [11]. Features which explain the highest variance among species are selected by LDA. LDA ranks features based on how much variation (among the species) they account for. This is different to PCA which looks for directions with the most variation in the dataset and ranks components based on how much variation they explain within a given dataset. Results (explained variances and number of features) obtained after applying LDA on spectral features are illustrated in Figure 2.

C. Purity Inspection Using Random Forest Classifier

Random Forest (RF) classification is an ensemble model which is constructed using many decision trees. The aggregate of the decisions reached by each of the trees is used for making predictions with new data. Apart from a Random Forest's ability to deal with the problem of overfitting (through the use of many trees) and handle large datasets swiftly and effectively, it offers attainment of comparable classification results and reduction of variance. In this work, 500 and 4:1, which gave the best classification result during preliminary analysis, are the number of decision trees and ratio of training to testing samples used respectively. Performances of this classifier with and without the use of both dimensionality reduction techniques are presented in the next section.

III. RESULTS AND ANALYSIS

To evaluate and compare performance of the RF classifier on outputs of LDA and outputs of PCA, three performance metrics are adopted. They are precision, recall and f1 score which are defined in (1) and (2).

$$\text{Precision (P)} = \frac{t_p}{t_p + f_p} \text{ and Recall (R)} = \frac{t_p}{t_p + f_n} \quad (1)$$

where t_p, f_p, t_n and f_n represent the number of true positive, number of false positive, number of true negative and number of false negative respectively.

$$\text{F1 score} = 2 * \frac{P*R}{P+R} \quad (2)$$

Where P is the precision and R is the recall.

We trained the RF classifier with the dataset before applying the two dimensionality reduction techniques. We observed that the RF classifier gave a below par performance when trained using the raw spectral features on full band with

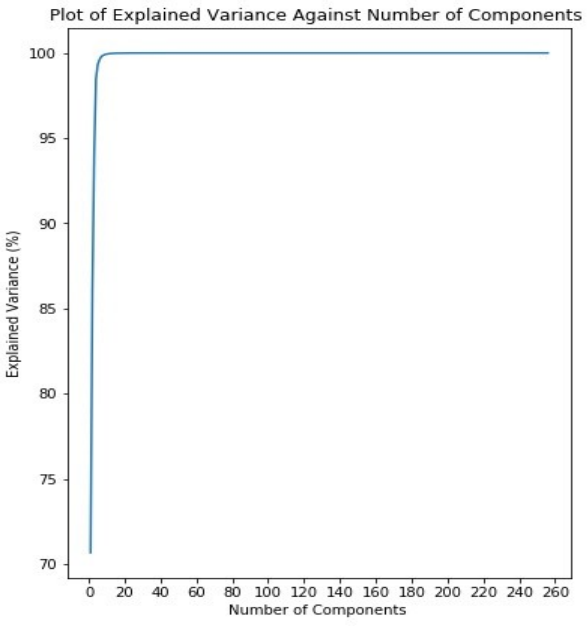

Figure 1 Result of PCA on spectra

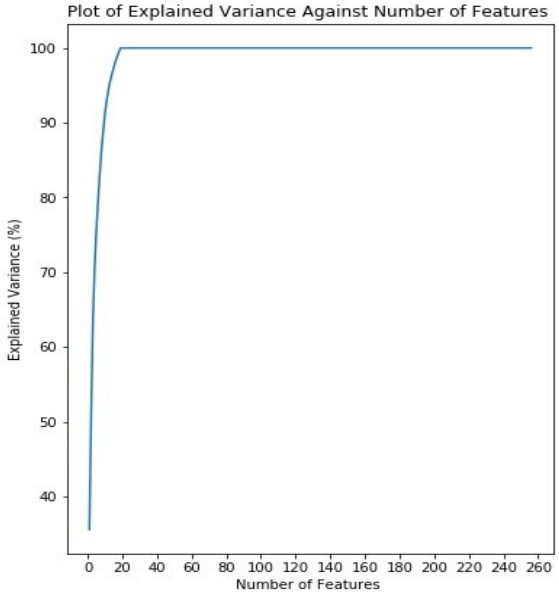

Figure 2 Result of LDA on spectra

average precision, average recall and average f1 score of 61.09%, 61.45% and 60.71% achieved respectively. We then apply PCA and LDA on the dataset to reduce dimensionality of the spectra features which was originally 256. In order to determine which of the two dimensionality reduction techniques is more effective for the rice seeds classification, we used the outputs of PCA and LDA on the spectra features to train the RF classifier. We present the classification results in Figure 3 and 4. We observed that application of the dimensionality reduction techniques significantly improved performance of the RF classifier.

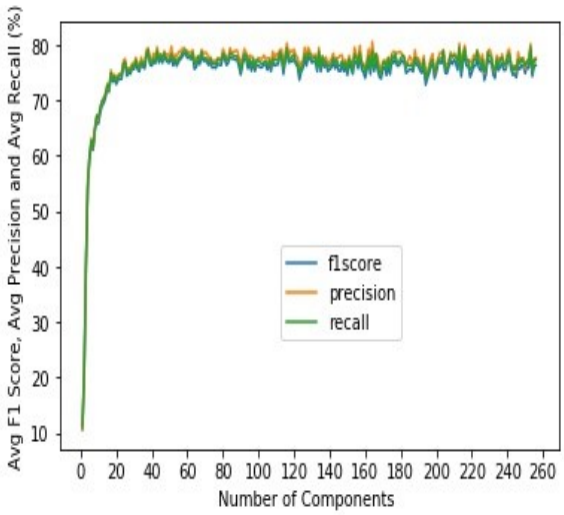

Figure 3 Results with output of PCA

From Figure 3, we observed that the best feature combination scheme using the output of PCA is 253 pcs when average precision, average recall and average f1 score of 80.14%, 79.55% and 78.96% were attained respectively. Similarly, from Figure 4, we observed that first 19 features of spectra gave the best classification results using the output of LDA with average precision, average recall and average f1 score of 85.94%, 86.28% and 85.86% attained respectively.

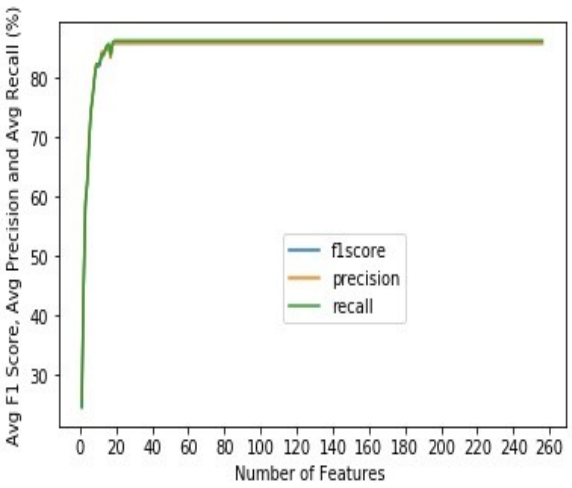

Figure 4 Results with output of LDA

Comparing results presented in Figure 3 and 4, we observed that RF classifier gave a better performance with the outputs of LDA. Hence, we selected 19 features which was obtained using LDA as the overall best feature scheme. These results show that LDA which maximizes the separability of categories in the rice seeds dataset is superior to PCA for purity inspection of rice seeds.

IV. Conclusion

Comparison of the performance of LDA and PCA for dimensionality reduction in quality inspection of rice seeds has been presented. Results obtained show that LDA is a superior dimensionality reduction technique to PCA for inspection of rice seeds quality.

Acknowledgment

The authors are thankful to the Newton Research Collaboration Programme (NRCP1516/1/65) for the financial support.

References

[1] H. Vu et al., 'Rice seed varietal purity inspection using hyperspectral imaging', in Hyperspectral Imaging and Applications Conference, Coventry, United Kingdom, 2016.

[2] H. Vu et al., "Spatial and spectral features utilization on a hyperspectral imaging system for rice seed varietal purity inspection," 2016 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), Hanoi, 2016, pp. 169-174. doi: 10.1109/RIVF.2016.7800289

[3] L. Wu, C. Yuan, A. Lin, and B. Zheng, "Identification of early moldy rice samples by PCA and PNN," Communications in Computer and Information Science Communications and Information Processing, pp. 506–514, 2012.

[4] Shwetank, K. Jain, and K. Bhatia, "Hyperspectral data compression model using SPCA (Segmented Principal Component Analysis) and classification of rice crop varieties," Communications in Computer and Information Science Contemporary Computing, pp. 360–372, 2010.

[5] L. Lu, S. Deng, Z. Zhu, and S. Tian, "Classification of rice by combining electronic tongue and nose," Food Analytical Methods, vol. 8, no. 8, pp. 1893–1902, 2015.

[6] L. Wang, D. Liu, H. Pu, D.-W. Sun, W. Gao, and Z. Xiong, "Use of hyperspectral imaging to discriminate the variety and quality of rice," Food Analytical Methods, vol. 8, no. 2, pp. 515–523, 2015.

[7] W. Liu, C. Liu, F. Ma, X. Lu, J. Yang, and L. Zheng, "Online variety discrimination of rice seeds using multispectral imaging and chemometric methods," Journal of Applied Spectroscopy, vol. 82, no. 6, pp. 993–999, 2016.

[8] Z. Liu, C. Li, Y. Wang, W. Huang, X. Ding, B. Zhou, H. Wu, D. Wang, and J. Shi, “Comparison of spectral indices and principal component analysis for differentiating lodged rice crop from normal ones,” Computer and Computing Technologies in Agriculture V IFIP Advances in Information and Communication Technology, pp. 84–92, 2012.

[9] J. Yang, J. Sun, L. Du, B. Chen, Z. Zhang, S. Shi, and W. Gong, “Monitoring of paddy rice varieties based on the combination of the Laser-Induced Fluorescence and Multivariate Analysis,” Food Analytical Methods, vol. 10, no. 7, pp. 2398–2403, 2017.

[10] G. A. Camelo-Méndez, B. H. Camacho-Díaz, A. A. D. Villar-Martínez, M. L. Arenas-Ocampo, L. A. Bello-Pérez, and A. R. Jiménez-Aparicio, “Digital image analysis of diverse Mexican rice cultivars,” Journal of the Science of Food and Agriculture, vol. 92, no. 13, pp. 2709–2714, 2012.

[11] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, “Linear discriminant analysis: A detailed tutorial,” AI Communications, vol. 30, no. 2, pp. 169–190, 2017.