White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Leeds.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/210541/

Version: Accepted Version

Article:
Fabiyi, S., Murray, P., Zabalza, J. et al. (2024) A New Hybridized Dimensionality Reduction Approach Using Genetic Algorithm and Folded Linear Discriminant Analysis Applied to Hyperspectral Imaging for Effective Rice Seed Classification. IEEE Transactions on AgriFood Electronics, 2 (1). 151 -164. ISSN: 2771-9529

https://doi.org/10.1109/TAFE.2024.3374753

This is an author produced version of an article published in IEEE Transactions on AgriFood Electronics, made available under the terms of the Creative Commons Attribution License (CC-BY), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

A new Hybridized Dimensionality Reduction Approach Using Genetic Algorithm and Folded Linear Discriminant Analysis Applied to Hyperspectral Imaging for Effective Rice Seed Classification

Samson Damilola Fabiyi, Paul Murray, Jaime Zabalza, Christos Tachtatzis, Hai Vu, and Trung Kien Dao

***Abstract*—Hyperspectral imaging (HSI) has been reported to produce promising results in the classification of rice seeds. However, HSI data often requires the use of dimensionality reduction techniques for the removal of redundant data. Folded Linear Discriminant Analysis (F-LDA) is an extension of Linear Discriminant Analysis (LDA, a commonly used technique for dimensionality reduction) and was recently proposed to address the limitations of LDA, particularly its poor performance when dealing with a small number of training samples which is a usual scenario in HSI applications. This paper presents an improved version of F-LDA, exploring the feasibility of hybridizing a Genetic Algorithm (GA) and F-LDA for effective dimensionality reduction in HSI-based rice seeds classification. The proposed approach, inspired by the previous combination of GA with Principal Component Analysis (PCA), is evaluated on rice seed datasets containing 256 spectral bands. Experimental results show that, in addition to attaining promising classification accuracies of up to 96.21%, this novel combination of GA and F-LDA (GA+F-LDA) can further reduce the computational complexity and memory requirement in the standalone F-LDA. It is worth noting that these benefits are not without a slight reduction in classification accuracy when evaluated against those reported for the standard F-LDA (up to 96.99%).**

***Index Terms*— Folded Linear Discriminant Analysis (F-LDA), Genetic Algorithm (GA), Hyperspectral Imaging (HSI), rice seed variety**

I. INTRODUCTION

Rice seeds serve as nutritional sources and are used, as seeds, for planting in many countries of the world. Hence, both the farmers and consumers in those countries place a high importance on the quality and purity of rice seeds [1]-[3], being rice seed classification a very important step in the quality evaluation, purity inspection, origin verification and specie identifications of this product. Human operators visually inspect rice seeds using the variation in their physical features such as shape, length, width, size, colour etc.[2],[4],[5]. Besides being subjective, visual inspection requires the deployment of experienced personnel and the training of new staff [1],[5],[6]. Visual inspection of the seeds, and other conventional approaches such as the use of High-Performance Liquid-Chromatography (HPLC) [7] and Gas Chromatography/Mass Spectrometry (GC-MS) [8] can also be exhausting, tedious, expensive and time-consuming [2],[9],[5]. RGB imaging and spectroscopy are alternative approaches which have been used separately for rice seeds classification in order to automate the inspection process and overcome the challenges limiting the conventional approaches [9],[5]. However, RGB imaging can only provide information on the morphology, colour and texture of acquired seeds and not their molecular composition [4]. Similarly, spectroscopy can provide information on the chemical constituents [4] but no information in the spatial context.

Spatial, textural and spectral information can be acquired simultaneously using Hyperspectral Imaging (HSI) systems [2],[4],[5]. This facilitates the utilization of both the chemical and physical profiles of samples in acquired hyperspectral images for enhanced classification [9],[6]. Due to the rich spectral information which are present in hyperspectral images [10]-[12], HSI solutions have already led to very promising results in the classification, quality inspection, and evaluation of Agri-Tech products such as beef [13], tea [14], lamb [15], cake [16], strawberry [17], and more specifically rice seeds [1],[5],[6],[18]. In [1], 3 varieties of rice seeds were classified using hyperspectral images in the Visible-Near Infrared (VIS/NIR) range of 400-1000 nm. The authors in [1] utilized an artificial neural network to achieve an accuracy of 94.45%. Applying a Random Forest (RF) on a combination of spectral and spatial features, the authors in [5] achieved a precision of 84% on rice seeds data containing 6 varieties. In [6], 4 rice seeds varieties were classified using a Support Vector Machines

For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission *(Corresponding author: Samson Damilola Fabiyi).*

Samson Damilola Fabiyi is with the School of Computing, University of Leeds, Leeds, LS2 9BW, U.K. (e-mail: s.d.fabiyi@leeds.ac.uk).

Paul Murray, Jaime Zabalza, and Christos Tachtatzis are with the Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, G1 1XW, U.K. (paul.murray@strath.ac.uk; j.zabalza@strath.ac.uk; christos.tachtatzis@strath.ac.uk).

Hai Vu and Trung Kien Dao are with the School of Electronics and Electrical Engineering , Hanoi University of Science and Technology, Hanoi 10000, Vietnam. (hai.vu@hust.edu.vn; trung-kien.dao@hust.edu.vn).

(SVM). The authors in [6] reported a classification accuracy of 91.67% when the classification model was applied on combined spectral, texture and morphological features. A RF was trained in [18] to classify 6 - 90 varieties of rice seeds. The authors in [18] reported F1 scores of 78.27% - 98.17% using a combination of spectral features acquired from hyperspectral images at a VIS/NIR range of ~ (385 – 1000) nm and spatial features acquired from RGB images.

However, HSI still faces some challenges that limit its potential for classification and quality inspection of rice seeds. Firstly, the labelled samples in the hyperspectral data are usually not enough for training [19],[20]. Secondly, hyperspectral data contains many spectral bands which are usually in order of hundreds, and most of which are highly correlated [21]. This results in noise and data redundancy [22],[23]. Hughes phenomenon therefore limits the performance of conventional machine learning models in HSI data classification resulting in reduced classification accuracy due to the lack of enough samples for training and the high dimensionality of the HSI data [24]-[26]. The Hughes phenomenon refers to a situation in classification where the classification accuracy initially increases as more spectral bands are added, reaches a peak, and then starts to decline [27]-[29]. Hence, the development of innovative data dimensionality reduction techniques for redundant data removal while retaining important information becomes crucial for enhancing classification tasks [30]-[32]. For instance, an end-to-end unsupervised band selection method known as a Dual Global–Local Attention Network (DGLAnet), was proposed in [33] where both global and local features were extracted from HSI data for effective dimensionality reduction and improved recognition accuracy. Also, in [34], the dimensionality reduction of HSI data was achieved by applying Enhancing Transformation Reduction (ETR) where an error matrix was initially subtracted from the covariance matrix to increase the proximity and uniformity between variables before the extraction of top eigenvectors and transformation of the data into a new subspace. Besides being a dimensionality reduction tool, ETR incorporates a facility to rectify pixels' misplacement and reduce noise, according to the authors in [34].

Dimensionality reduction techniques can be classified into two categories: feature extraction and feature selection. Feature extraction involves transforming the data from its original high-dimensional feature space to a lower-dimensional space. Various feature extraction techniques have been used in HSI data classification, achieving promising results. Examples of these are Generalized Discriminant Analysis (GDA) [35], Nonparametric Weighted Feature Extraction (NWFE) [36], isometric mapping [37], random projection [38], maximum noise fraction [39], sparse and low-rank near-isometric linear embedding [40], and wavelet dimensionality reduction [41]. More specifically, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are two prominent feature extraction tools that have been greatly applied in the classification and quality inspection of rice seed using HSI [1],[5],[6],[20],[18]. PCA is a conventional technique commonly used in many HSI applications [42] and is noted to be the most regularly applied for dimensionality reduction in HSI-based rice seeds classifications (as shown in Table I). However, the work in [20] showed that LDA can give a better performance as a dimensionality reduction tool than PCA for rice seeds classification and quality inspection. Nevertheless, the performance of LDA can be limited by the lack of enough samples available for training. Therefore, in a recent work [43], Folded Linear Discriminant Analysis (F-LDA), an extension of LDA, was proposed for an improved dimensionality reduction under small training sample size scenarios. The proposed F-LDA performed better than the traditional LDA in terms of classification accuracy, computational complexity and memory requirement when applied on small training samples.

On the other hand, feature selection, also known as band selection in the context of HSI, involves selecting a subset of features from the original feature set. Feature selection tools have also found applications in the dimensionality reduction of HSI data. Examples of these are DGLAnet [33], multiscale spectral features graph fusion [44], and fast and latent low-rank subspace clustering [45]. Genetic algorithms (GAs), are another feature selection tool. GAs operate using a wrapper-based approach where an optimal subset of features is selected by analysing the relationship between the entire feature sets and classification models for increased classification performance. GAs have been applied for feature selection and reduction, providing very promising results in a number of applications including vowel [46], learning [47], and cancer data classification [48], fish species prediction [49], biometrics authentication [50], activity recognition [51], spam email detection [52], and emotion and genre detection [53].

Based on the current state of the art (see Table I), no techniques appear to have used GAs for feature reduction in automated rice seed inspection tasks. At the same time, literature has reported very promising results achieved by the combination of GA and PCA [50], [51], and also GA and LDA [54], [55] in other applications different to rice seeds classification. However, to the best of the authors' knowledge, the idea of combining GA with FLDA for rice seed analysis has not been explored. Thus a conceptual gap exists in this area. This, along with the LDA superiority over PCA as a dimensionality reduction technique for rice seeds classification [20], has inspired us to propose a novel hybridization of GA and F-LDA, where a RF is employed for classification of the resulting feature vectors, allowing a comparison with other existing methods in Table I.

This work (part of the first author's PhD thesis [56]) is aimed at introducing a novel feature extraction approach based on hybridizing GAs and F-LDA, evaluating its effectiveness for rice seeds classification. To the best of the authors' knowledge, we are the first to explore this approach.

The novel contributions of our paper lie in two key aspects, the first being the introduction of an improved version of F-LDA, a new hybridized dimensionality reduction approach using a GA and F-LDA, for HSI data classification. Using the hybridized approach, F-LDA can be applied on a reduced dataset – the number of features is significantly reduced, compared to the original dimensionality. The reduction in

computational complexity and memory requirements achieved by the hybridized approach, compared to the standalone F-LDA, makes it a theoretical improvement over the conventional approach. Secondly, this paper evaluates the effectiveness of our hybridized dimensionality reduction approach for rice seeds classification. The experimental results demonstrate the effectiveness of the hybridized approach in dealing with the challenges of high dimensionality in HSI data classification specific to this application domain. The continued requirement for automation and improvement of rice seed screening tasks [1],[57]-[59], makes rice seeds classification (a crucial stage) a suitable field of application for F-LDA and the proposed GA+F-LDA, where they would be deployed to enhance the seeds classification.

II. Methods And Materials

A. Data Acquisition and Description

Two groups of rice seed datasets are used in this work for performance evaluation of the proposed approach. The first group contains 10 different rice seeds (subsets), each containing 10 varieties. There are another 10 different rice seeds (subsets) in the second group, each containing 20 varieties. The rice seed subsets in the two groups are randomly selected from the main rice seed datasets of 90 varieties in [60]. We chose to use the two groups of rice seed datasets to demonstrate the consistency of the proposed approach by averaging the classification results obtained using the different randomly selected rice seeds subsets. Information on the varieties in each subset for the two groups of rice seeds is illustrated in Table II.

The rice seed datasets contain 256 spectra bands that were extracted from hyperspectral images plus 6 spatial features (area, major axis length, minor axis length, aspect ratio, perimeter over area ratio, and eccentricity), which were extracted from RGB images. The hyperspectral images were collected using a VIS/NIR range HSI system operating at ~ (385 – 1000) nm while the RGB images were collected using a Fujifilm X-M1 with a 35mm/F2.0 lens. Description of the procedure and processing steps used when extracting the spatial and spectral features are provided in [18]. While the focus of this work is on addressing the high dimensionality problem in the spectral domain, spatial features are utilized in the rice seed analysis to demonstrate their potential to further improve classification results when used in conjunction with spectral features.

B. Spectral Response of Rice Seeds

Plots of the average spectra of each of the species in two rice seed datasets containing 10 and 20 varieties are obtained and illustrated in Fig. 1 and Fig. 2. There are differences in the average spectral profiles that can be attributed to the differences in chemical constituents (property) of the rice seeds species in the NIR region (700-1000 nm) and physical property (colour variation) in the visible region (385-700 nm) [2],[3],[61]. Consequently, the spectral differences in these regions are exploited by models, reducing the dimensionality of the data by retaining only key features able to discriminate the rice seed species.

C. Proposed Approach

1) **Concept**

The proposed framework for our novel GA + F-LDA technique consists of the following three main steps (illustrated in Fig. 3): 1) data acquisition; 2) dimensionality reduction; and 3) classification. The dimensionality reduction is performed in two steps: 1) optimal feature subset selection using a GA, and 2) feature extraction using F-LDA. The GA is used to select optimal spectral features in order to extract a data subset from each of the original spectral datasets. F-LDA is then used to further reduce the features of the extracted data (the implementation details is provided in the next sub section). This facilitates the application of F-LDA on a reduced dataset (number of features is significantly reduced with relation to the original dimensionality), while also improving the classification performance of the conventional LDA. It is also expected that the use of fewer features will bring about further reduction in the computational complexity and memory requirement of F-LDA. The next subsection explains the implementation details of the dimensionality reduction and classification steps.

2) **Implementation Details**

First, implementation of the GA is carried out following the 3 key steps that are motivated by the principles of natural selection and genetics namely selection, crossover and mutation [46]-[49]. The algorithm begins by randomly initiating a population of candidate solutions. It then estimates the fitness of the current set of candidate solutions using an objective function and discards unfit candidate solutions. Finally, it produces the next generation offspring solutions (a new population of candidate solutions) by mating the fitter ones using crossover and mutation. The above steps are repeated over many generations until the specified maximum generation is reached or an optimal solution is attained. Offspring solutions produced by crossover contain genes which are located in the parents' chromosomes - each solution refers to a chromosome with a set of genes. When offspring solutions are produced by mutations, they contain genes which cannot be found in both parents. During mutation, a global optimal solution is achieved through exploration of the search space. In this work, the crossover was set to a probability of 0.6 while a probability of 0.2 was specified for the mutation. The initial population of candidate solutions and maximum number of generations were both set to 100 [48],[49]. Implementation of the algorithm was carried out in Python using a scikit-learn module namely sklearn-genetic [62] which selects the feature subset that maximizes an objective function. RF and accuracy are the classifier and fitness (objective) function used in this work respectively.

Second, F-LDA is applied to the extracted data subsets to further reduce its optimal spectral features. F-LDA achieves this by initially converting each spectral vector in the hyperspectral data matrix in to a 2D matrix. The new (resulting) data is now a set of 2D matrices (folded vectors), each having a size of $G \times B = f$, where f denotes the number of features in

the hyperspectral data. F-LDA then uses the conventional LDA steps namely within class matrix computation, between class matrix computation and data projection on the new data. In the final step, F-LDA unfolds the projected samples, after which the unfolded samples are presented to models for classification. Unlike the conventional LDA which processes a set of spectral vectors, F-LDA deals with a set of folded spectral vectors. Different configurations (dimensions, $G \times B$) of the 2D matrices are explored in this work, selecting the one that produces the best classification results. Mathematically, F-LDA is implemented as follows. The first step is the conversion of a hyperspectral data cube into a data matrix denoted as X with dimensions $s * f$ where s and f denotes the number of samples and features in the hyperspectral data respectively. Denoting each spectral vector (sample) in the data matrix as $\boldsymbol{x_n} = [x_{n1}\, x_{n2}\, x_{n3} \ldots\ldots\ldots x_{nf}]$ where $n \epsilon [1, s]$, each $\boldsymbol{x_n}$ is folded (converted) into a 2D matrix denoted as $\boldsymbol{P_n}$ using (1).

$$\boldsymbol{P_n} = \begin{bmatrix} p_{n(1,1)} & \cdots & p_{n(1,B))} \\ \vdots & \ddots & \vdots \\ p_{n(G,1)} & \cdots & p_{n(G,B)} \end{bmatrix} \quad (1)$$

Denoting each element in the matrix P_n as $p_{n(h+1,i)}$, it can be computed using (2).

$$p_{n(h+1,i)} = x_{(h*B)+i} \quad (2)$$

where $h \epsilon [0, G-1]$ and $i \epsilon [1, B]$. Let c and c_j denote the number of classes and jth class in $\boldsymbol{X}$ respectively, the number of samples in each class can be denoted as N_j. The mean of all $\boldsymbol{P_n}$ (folded samples) in each class c_j can be denoted as $\boldsymbol{M_j}$ where $\epsilon [1, c]$, and computed using (3)

$$\boldsymbol{M_j} = \frac{1}{N_j} \sum_{i=1}^{N_j} P_{ij} \;\; \boldsymbol{M_j} \epsilon \Re^{G \times B} \quad (3)$$

where $\boldsymbol{P_{ij}}$ is the ith converted matrix (folded sample) in class c_j and $i \epsilon [1, N_j]$. Similarly, the overall mean of all $\boldsymbol{P_n}$ (folded samples) can be denoted as $\boldsymbol{M}$ and computed using (4).

$$\boldsymbol{M} = \sum_{j=1}^{c} \frac{N_j}{s} M_j, \;\; \boldsymbol{M} \epsilon \Re^{G \times B} \quad (4)$$

Using (5) and (6), the within-class variance $\boldsymbol{V_{PW}}$ and the between-class variance $\boldsymbol{V_{PB}}$ of the data matrix can be computed respectively.

$$\boldsymbol{V_{PW}} = \sum_{j=1}^{c} \sum_{i=1}^{N_j} (\boldsymbol{P_{ij}} - \boldsymbol{M_j})(\boldsymbol{P_{ij}} - \boldsymbol{M_j})^T \quad (5)$$

$$\boldsymbol{V_{PB}} = \sum_{j=1}^{c} N_j (\mathbf{M_j} - \mathbf{M})(\mathbf{M_j} - \mathbf{M})^T \quad (6)$$

where $\boldsymbol{V_{PW}} \epsilon \Re^{G \times G}$ and $\boldsymbol{V_{PB}} \epsilon \Re^{G \times G}$. As shown in (7), the transformation matrix, $\boldsymbol{T_P}$ can be computed using the between-class variance, $\boldsymbol{V_{PB}}$, and the within-class variance, $\boldsymbol{V_{PW}}$. Eigenvalues and eigenvectors are computed from Eigen Value Decomposition (EVD) of $\boldsymbol{T_P}$. Ranking the eigenvectors, of size $G \times G$, in descending order according to their corresponding eigenvalues, the first d columns of the ranked eigenvectors can be selected and the rest discarded, reducing the eigenvectors to a submatrix, denoted as $\boldsymbol{V_{Pd}}$, of size $G \times d$. The data can then be projected into a lower dimensional space using (8) where $\boldsymbol{Y_n}$ is the projected matrix of each sample.

$$\boldsymbol{T_P} = \boldsymbol{V_{PW}}^{-1} \boldsymbol{V_{PB}}, \boldsymbol{T_P} \epsilon \Re^{G \times G} \quad (7)$$

$$\boldsymbol{Y_n} = \boldsymbol{P_n}^T \boldsymbol{V_{Pd}}, \boldsymbol{V_{Pd}} \epsilon \Re^{G \times d'}, \; \boldsymbol{Y_n} \epsilon \Re^{G \times d'} \quad (8)$$

The flowchart illustrating the implementation of F-LDA is presented in Fig. 4.

Finally, a RF model is used for classification of the rice seeds data. The selection of RF model for this work is motivated by the promising results achieved when it was applied on HSI data of rice seeds in related papers [5],[9],[63]. To classify the rice seeds using the RF, the data is split into training and testing sets. The number of decision trees used in the RF, D, is varied starting from 100 up to 1000 (in steps of 100). A k-fold (k=5) cross validation is used on the training set to determine the optimal values of D. That is, the training and validation of the RF classifier was carried out k times. In each case, k-1 of the folds was used to train the classifier while the remaining fold is used for its validation. Average of the accuracies recorded in all cases is then outputted by the cross validation. The next section reports and discusses the classification results obtained during the final evaluation on the test set.

IV. Results And Discussion

Accuracy is a widely used evaluation metric in related works [64]-[66] and is therefore used to evaluate and compare the proposed approach with different feature schemes using the RF classifier. Classification models are trained using the original spectral feature sets or selected/reduced spectral feature sets obtained from the dimensionality reduction process. Therefore, in this work, we utilize different feature schemes and compare their performances with that of our proposed new feature extraction scheme, namely GA+F-LDA. The other feature schemes included for comparison are:

- Raw spectral features - original spectral features.
- GA only features - an optimal feature subset selected from the raw spectral feature sets using the GA only.
- LDA features - features extracted from the raw spectral feature sets using LDA.
- PCA features - features extracted from the raw spectral feature sets using PCA
- GA + PCA features - features extracted following the application of PCA on the GA outputs.
- GA+ LDA features - features extracted following the application of LDA on the GA outputs.
- F-LDA features - features extracted by the standard F-LDA, i.e., without GA.

A. Spectra Features Selection

The number of features selected from the application of GA on the raw spectral features varied for different datasets as illustrated in Table III. For the group of rice seed datasets with 10 species, the number of spectral features selected varies from 96 to 177. The number of spectral features selected for the group of rice seed datasets with 20 species varies from 94 to 225. For visualisation of the regions of the spectral range that cover the selected spectral features and are useful for the classification tasks in this work, the average spectra of sample species taken from data subsets containing 10 and 20 rice seed species are plotted and illustrated in Fig. 5. As can be seen in Fig. 5, the selected features spread across the spectral range of the rice seeds and cannot be separated into distinct clusters. This shows the relevance of both regions (Visible and NIR) of the utilized spectral range for the rice seed classification task. Specifically, it shows the usefulness of the differences in chemical property (composition) in the lower bands of the NIR region and physical property (colour variation among the species) in the visible region of the spectral profiles in discriminating the rice seed species [2],[3],[61]. Notably, the number of features produced by the GA are still high (above a hundred in majority of the cases). This explains why Fig. 5 looks as though pretty much every band across the spectrum are being utilized. Notwithstanding, the goal here is to apply GA as a forerunner to F-LDA with the expectation that the application of F-LDA on the GA features (which are fewer in numbers when compared with the raw features) will cause further reduction in computational complexity and memory requirement of F-LDA.

B. Analysing the Performance on the Rice Seeds Datasets

First, before performing any dimensionality reduction of the rice seed spectral data, we used the 10 random subsets of 10 rice seed varieties to train the RF classifier. We also extracted sub datasets with reduced feature subset from each of the random subsets using GA and use this to train the RF classifier. We apply F-LDA on the raw spectral datasets and the datasets with the reduced features (selected GA features). We used the outputs of the F-LDA obtained in both cases separately to train the RF classifier. For performance comparison of the proposed approach, (GA+F-LDA) with other techniques (LDA, GA+LDA, PCA, and GA+PCA), firstly, we applied the outputs of PCA on the raw spectral data and then on the data with the selected GA features (starting from the first principal components up to the 10^{th} in the dataset). Secondly, we applied the outputs of LDA on the raw spectral data and the data with the selected GA features, starting from the first LDA component up to the $(c-1)^{th}$ where c is the number of species in the dataset. Finally, we separately combined the outputs of F-LDA and GA+F-LDA with their corresponding 6 spatial features. In all the cases considered, we vary the size of the training samples and compute the average of classification results using the 10 random subsets of rice seed varieties as presented in Fig. 6.

From the classification results presented in Fig. 6, we observed that the accuracy with the selected GA features is an improvement on the one obtained when the RF was trained with the spectral features only. Expectedly, significantly higher results are achieved when the RF is trained with the output of PCA and LDA applied to the raw spectral data. Ultimately, as can be seen in Fig 6, the classification accuracy obtained when we utilized the LDA features extracted from the spectral data (with high training to testing samples ratio) is greater than the accuracy obtained when we used the PCA features. This validates our motivation for proposing the hybridization of GA and F-LDA for dimensionality reduction of HSI data in rice seed classification.

We also observed that the classification accuracy increases with the training sample size when we trained the RF classifier with LDA outputs. Besides, Fig. 6 shows that GA+LDA, F-LDA and GA+FLDA feature schemes all compensate for the inability of LDA to give comparable classification results when applied in small training samples scenarios. We observed from Fig. 6 that the proposed GA+F-LDA gave the second highest accuracy, which is lower than those given by F-LDA. When the standard deviation reported in Fig. 6 is given consideration, the accuracy achieved by both GA+F-LDA and F-LDA can be seen to be comparable. The reduction in classification performance of GA+F-LDA (when compared with F-LDA only) can therefore be considered insignificant. As can be seen in Fig. 6, higher classification results are reported for F-LDA and GA+F-LDA when their outputs are combined with the spatial features.

Though the standard F-LDA slightly outperforms the proposed GA+F-LDA in terms of accuracy, the proposed GA+F-LDA can reduce the computational complexity and memory requirement at the different stages of F-LDA (in the majority of the cases considered) as illustrated in Table IV and Table V. Preference may be given to these reduction in computational complexity and memory requirement in practice, though at the expense of slight reduction in classification performance (accuracy). The question of whether to use F-LDA or GA+FLDA would then depend on the application. The key point is that F-LDA performs well whether it is applied on original or datasets with reduced features i.e. datasets containing GA selected featured. We also illustrate the average feature extraction time (s) of F-LDA (when applied on the random datasets of 10 varieties) in both cases in Table VIII. We observed from Table VIII that the time used by F-LDA for features extraction when applied on the reduced datasets is less. Again, it is noteworthy that the reduction in feature extraction time is insignificant when the standard deviation reported in the tables are given consideration. Besides, the time used by F-LDA in both cases is less than 0.1s as can be seen in Table VIII. The reduction in computational complexity and memory requirement achieved by the application of F-LDA on datasets with the selected GA features therefore demonstrates the potential of the proposed approach (GA+F-LDA).

We repeated the above process on the rice seed datasets of 20 varieties and present the classification results in Fig. 7. Again, we observed from Fig. 7 that GA+LDA, F-LDA and GA+F-LDA feature schemes continue to compensate for the inability of LDA to perform well in small training samples scenarios. We observed from Fig. 7 that the accuracy achieved by GA+F-LDA is very high and is only surpassed by the one reported for F-LDA. Again, when the standard deviation reported in Fig. 7 is given consideration, the accuracy achieved by both GA+F-LDA and F-LDA can be seen to be comparable.

The reduction in classification performance of GA+F-LDA (when compared with F-LDA) can therefore be considered insignificant. We further observed from Table VI and Table VII that reduction in computational complexity and memory requirement are achieved for the majority of the cases in F-LDA when applied on the reduced datasets (datasets containing the selected GA features). We also illustrated the average feature extraction time (s) of F-LDA when applied on the full and GA features in Table VIII. As can be seen in Table VIII, the time used by F-LDA in extracting the features when applied on the reduced datasets is less. It is worth noting that this reduction in feature extraction time is insignificant when the standard deviation reported for each case is given consideration. Besides, the time used by F-LDA in both cases is less than 0.25s as can be seen in Table VIII.

V. Conclusion

In this paper, a novel dimensionality reduction approach which is based on hybridizing GA and F-LDA (GA+F-LDA) has been introduced and its effectiveness evaluated for HSI data. Rice seed hyperspectral datasets were used to evaluate performance of the proposed approach. The experimental results achieved by GA+F-LDA are promising and establish the potential of applying F-LDA on HSI datasets reduced by GA (further reduction in computational complexity and memory requirement can be obtained). It is worth noting that these benefits are accompanied by a slight reduction in classification accuracy when evaluated against those reported for the standard F-LDA. While this paper has explored dimensionality reduction approaches for effective rice seeds classification using HSI, visually classifying hyperspectral and multispectral images of rice seeds clusters is a work that could be investigated in the future for valuable insights.

Appendix

Data Availability: The data that support the findings of this study are available in Zenodo with the identifier(s) 10.5281/zenodo.3241923 [60].

References

[1] 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 Anal Methods*, vol. 8, no. 2, pp. 515–523, 2015, doi: 10.1007/s12161-014-9916-5.

[2] Z. Qiu, J. Chen, Y. Zhao, S. Zhu, Y. He, and C. Zhang, "Variety Identification of Single Rice Seed Using Hyperspectral Imaging Combined with Convolutional Neural Network," *Applied Sciences*, vol. 8, no. 2, p. 212, Jan. 2018, doi: 10.3390/app8020212.

[3] W. Liu, Ch. Liu, F. Ma, X. Lu, J. Yang, and L. Zheng, "Online Variety Discrimination of Rice Seeds Using Multispectral Imaging and Chemometric Methods," *J Appl Spectrosc*, vol. 82, no. 6, pp. 993–999, Jan. 2016, doi: 10.1007/s10812-016-0217-1.

[4] W. Kong, C. Zhang, F. Liu, P. Nie, and Y. He, "Rice seed cultivar identification using near-infrared hyperspectral imaging and multivariate data analysis," *Sensors (Basel)*, vol. 13, no. 7, pp. 8916–8927, 2013, doi: 10.3390/s130708916.

[5] H. Vu *et al.*, "Spatial and spectral features utilization on a Hyperspectral imaging system for rice seed varietal purity inspection," in *2016 IEEE RIVF International Conference on Computing and Communication Technologies: Research, Innovation, and Vision for the Future, RIVF 2016 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., Dec. 2016, pp. 169–174. doi: 10.1109/RIVF.2016.7800289.

[6] J. Sun, X. Lu, H. Mao, X. Jin, and X. Wu, "A Method for Rapid Identification of Rice Origin by Hyperspectral Imaging Technology," *J Food Process Eng*, vol. 40, no. 1, p. e12297, Feb. 2017, doi: 10.1111/jfpe.12297.

[7] F. Huebner, A. Hussain, G. Lookhart, J. Bietz, W. Bushuk, and B. Juliano, "Discrimination of sister-line IR rice varieties by polyacrylamide gel electrophoresis and reversed-phase high-performance liquid chromatography," *Cereal Chem*, vol. 68, no. 6, pp. 583–588, 1991.

[8] W. J *et al.*, "Metabolite Profiles of Rice Cultivars Containing Bacterial Blight-Resistant Genes Are Distinctive From Susceptible Rice," *Acta Biochim Biophys Sin (Shanghai)*, vol. 44, no. 8, 2012, doi: 10.1093/ABBS/GMS043.

[9] W. Kong, C. Zhang, F. Liu, P. Nie, and Y. He, "Rice seed cultivar identification using near-infrared hyperspectral imaging and multivariate data analysis," *Sensors (Basel)*, vol. 13, no. 7, pp. 8916–8927, 2013, doi: 10.3390/s130708916.

[10] H. Fu, G. Sun, J. Zabalza, A. Zhang, J. Ren, and X. Jia, "A Novel Spectral-Spatial Singular Spectrum Analysis Technique for near Real-Time in Situ Feature Extraction in Hyperspectral Imaging," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 13, pp. 2214–2225, 2020, doi: 10.1109/JSTARS.2020.2992230.

[11] P. Ghamisi *et al.*, "Advances in Hyperspectral Image and Signal Processing: A Comprehensive Overview of the State of the Art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4. Institute of Electrical and Electronics Engineers Inc., pp. 37–78, Dec. 01, 2017. doi: 10.1109/MGRS.2017.2762087.

[12] Q. Yang, Y. Xu, Z. Wu, and Z. Wei, "Hyperspectral and Multispectral Image Fusion Based on Deep Attention Network," in *Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing*, IEEE Computer Society, Sep. 2019. doi: 10.1109/WHISPERS.2019.8920825.

[13] T. Qiao, J. Ren, C. Craigie, J. Zabalza, Ch. Maltin, and S. Marshall, "Quantitative Prediction of Beef Quality Using Visible and NIR Spectroscopy with Large Data Samples Under Industry Conditions," *J Appl Spectrosc*, vol. 82, no. 1, pp. 137–144, Mar. 2015, doi: 10.1007/s10812-015-0076-1.

[14] T. Kelman, J. Ren, and S. Marshall, "Effective classification of Chinese tea samples in hyperspectral imaging," *Artif Intell Res*, vol. 2, no. 4, p. 87, Oct. 2013, doi: 10.5430/air.v2n4p87.

[15] M. Kamruzzaman, G. Elmasry, D. W. Sun, and P. Allen, "Application of NIR hyperspectral imaging for discrimination of lamb muscles," *J Food Eng*, vol. 104, no. 3, pp. 332–340, Jun. 2011, doi: 10.1016/j.jfoodeng.2010.12.024.

[16] A. Polak, F. K. Coutts, P. Murray, and S. Marshall, "Use of hyperspectral imaging for cake moisture and hardness prediction," *IET Image Process*, vol. 13, no. 7, pp. 1152–1160, May 2019, doi: 10.1049/iet-ipr.2018.5106.

[17] G. ElMasry, N. Wang, A. ElSayed, and M. Ngadi, "Hyperspectral imaging for nondestructive determination of some quality attributes for strawberry," *J Food Eng*, vol. 81, no. 1, pp. 98–107, Jul. 2007, doi: 10.1016/j.jfoodeng.2006.10.016.

[18] S. D. Fabiyi *et al.*, "Varietal Classification of Rice Seeds Using RGB and Hyperspectral Images," *IEEE Access*, vol. 8, pp. 22493–22505, 2020, doi: 10.1109/ACCESS.2020.2969847.

[19] Y. Zhang, G. Cao, A. Shafique, and P. Fu, "Label Propagation Ensemble for Hyperspectral Image Classification," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 12, no. 9, pp. 3623–3636, Sep. 2019, doi: 10.1109/JSTARS.2019.2926123.

[20] S. D. Fabiyi *et al.*, "Comparative Study of PCA and LDA for Rice Seeds Quality Inspection," in *2019 IEEE AFRICON*, IEEE, Sep. 2019, pp. 1–4. doi: 10.1109/AFRICON46755.2019.9134059.

[21] X. Wei, W. Zhu, B. Liao, and L. Cai, "Matrix-based margin-maximization band selection with data-driven diversity for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 12, pp. 7294–7309, Dec. 2018, doi: 10.1109/TGRS.2018.2849981.

[22] A. Zhang *et al.*, "Hyperspectral band selection using crossover-based gravitational search algorithm," *IET Image Process*, vol. 13, no. 2, pp. 280–286, Feb. 2019, doi: 10.1049/iet-ipr.2018.5362.

[23] J. Zabalza *et al.*, "Novel two-dimensional singular spectrum analysis for effective feature extraction and data classification in hyperspectral imaging," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 8, pp. 4418–4433, Aug. 2015, doi: 10.1109/TGRS.2015.2398468.

[24] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5408–5423, Sep. 2018, doi: 10.1109/TGRS.2018.2815613.

[25] X. Dai and W. Xue, "Hyperspectral Remote Sensing Image Classification Based on Convolutional Neural Network," in *Chinese Control Conference, CCC*, IEEE Computer Society, Oct. 2018, pp. 10373–10377. doi: 10.23919/ChiCC.2018.8484034.

[26] J. Leng, T. Li, G. Bai, Q. Dong, and H. Dong, "Cube-CNN-SVM: A Novel Hyperspectral Image Classification Method," in *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, San Jose, CA: Institute of Electrical and Electronics Engineers Inc., 2016, pp. 1027–1034. doi: 10.1109/ICTAI.2016.0158.

[27] M. Pal and G. M. Foody, "Feature selection for classification of hyperspectral data by SVM," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 5, pp. 2297–2307, May 2010, doi: 10.1109/TGRS.2009.2039484.

[28] M. Kamandar and H. Ghassemian, "Linear feature extraction for hyperspectral images based on information theoretic learning," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 4, pp. 702–706, 2013, doi: 10.1109/LGRS.2012.2219575.

[29] H. Yu, L. Gao, J. Li, S. Li, B. Zhang, and J. Benediktsson, "Spectral-Spatial Hyperspectral Image Classification Using Subspace-Based Support Vector Machines and Adaptive Markov Random Fields," *Remote Sens (Basel)*, vol. 8, no. 4, p. 355, Apr. 2016, doi: 10.3390/rs8040355.

[30] G. Yang, K. Huang, W. Sun, X. Meng, D. Mao, and Y. Ge, "Enhanced mangrove vegetation index based on hyperspectral images for mapping mangrove," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 189, pp. 236–254, Jul. 2022, doi: 10.1016/J.ISPRSJPRS.2022.05.003.

[31] W. Sun, K. Ren, X. Meng, G. Yang, J. Peng, and J. Li, "Unsupervised 3-D Tensor Subspace Decomposition Network for Spatial-Temporal-Spectral Fusion of Hyperspectral and Multispectral Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, 2023, doi: 10.1109/TGRS.2023.3324028.

[32] T. Hou, W. Sun, C. Chen, G. Yang, X. Meng, and J. Peng, "Marine floating raft aquaculture extraction of hyperspectral remote sensing images based decision tree algorithm," *International Journal of Applied Earth Observation and Geoinformation*, vol. 111, p. 102846, Jul. 2022, doi: 10.1016/J.JAG.2022.102846.

[33] K. He *et al.*, "A Dual Global-Local Attention Network for Hyperspectral Band Selection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022, doi: 10.1109/TGRS.2022.3169018.

[34] D. Al-Alimi, Z. Cai, and M. A. A. Al-Qaness, "FHIC: Fast Hyperspectral Image Classification Model Using ETR Dimensionality Reduction and ELU Activation Function," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, 2023, doi: 10.1109/TGRS.2023.3314619.

[35] G. Yang, X. Yu, and X. Zhou, "HYPERSPECTRAL IMAGE FEATURE EXTRACTION BASED ON GENERALIZED DISCRIMINANT ANALYSIS," in *XXIst ISPRS Congress*, Beijing, 2008, pp. 285–290.

[36] B. C. Kuo and D. A. Landgrebe, "Nonparametric weighted feature extraction for classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 5, pp. 1096–1105, May 2004, doi: 10.1109/TGRS.2004.825578.

[37] W. Li, L. Zhang, L. Zhang, and B. Du, "GPU Parallel Implementation of Isometric Mapping for Hyperspectral Classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 9, pp. 1532–1536, Sep. 2017, doi: 10.1109/LGRS.2017.2720778.

[38] V. Menon, Q. Du, and J. E. Fowler, "Fast SVD with Random Hadamard Projection for Hyperspectral Dimensionality Reduction," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 9, pp. 1275–1279, Sep. 2016, doi: 10.1109/LGRS.2016.2581172.

[39] J. Xia, P. Du, X. He, and J. Chanussot, "Hyperspectral remote sensing image classification based on rotation forest," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 1, pp. 239–243, 2014, doi: 10.1109/LGRS.2013.2254108.

[40] W. Sun, G. Yang, B. Du, L. Zhang, and L. Zhang, "A sparse and low-rank near-isometric linear embedding method for feature extraction in hyperspectral imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 4032–4046, Jul. 2017, doi: 10.1109/TGRS.2017.2686842.

[41] L. M. Bruce, C. H. Koger, and J. Li, "Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 10, pp. 2331–2338, Oct. 2002, doi: 10.1109/TGRS.2002.804721.

[42] Shwetank, K. Jain, and K. Bhatia, "Hyperspectral Data Compression Model Using SPCA (Segmented Principal Component Analysis) and Classification of Rice Crop Varieties," Springer, Berlin, Heidelberg, 2010, pp. 360–372. doi: 10.1007/978-3-642-14834-7_34.

[43] S. D. Fabiyi, P. Murray, J. Zabalza, and J. Ren, "Folded LDA: Extending the Linear Discriminant Analysis Algorithm for Feature Extraction and Data Reduction in Hyperspectral Remote Sensing," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 14, pp. 12312–12331, 2021, doi: 10.1109/JSTARS.2021.3129818.

[44] W. Sun *et al.*, "A Multiscale Spectral Features Graph Fusion Method for Hyperspectral Band Selection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022, doi: 10.1109/TGRS.2021.3102246.

[45] W. Sun, J. Peng, G. Yang, and Q. Du, "Fast and Latent Low-Rank Subspace Clustering for Hyperspectral Band Selection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 6, pp. 3906–3915, Jun. 2020, doi: 10.1109/TGRS.2019.2959342.

[46] Y. Korkmaz, A. Boyacı, and T. Tuncer, "Turkish vowel classification based on acoustical and decompositional features optimized by Genetic Algorithm," *Applied Acoustics*, vol. 154, pp. 28–35, Nov. 2019, doi: 10.1016/J.APACOUST.2019.04.027.

[47] J.-Y. Kim and S.-B. Cho, "Exploiting deep convolutional neural networks for a neural-based learning classifier system," *Neurocomputing*, vol. 354, pp. 61–70, Aug. 2019, doi: 10.1016/J.NEUCOM.2018.05.137.

[48] M. Jansi Rani and D. Devaraj, "Two-Stage Hybrid Gene Selection Using Mutual Information and Genetic Algorithm for Cancer Data Classification," *J Med Syst*, vol. 43, no. 8, p. 235, Aug. 2019, doi: 10.1007/s10916-019-1372-8.

[49] R. Zarkami, Z. Darizin, R. Sadeghi Pasvisheh, A. Bani, and A. Ghane, "Use of data-driven model to analyse the occurrence patterns of an indicator fish species in river: A case study for Alburnoides eichwaldii (De Filippi, 1863) in Shafaroud River, north of Iran," *Ecol Eng*, vol. 133, pp. 10–19, Aug. 2019, doi: 10.1016/J.ECOLENG.2019.04.018.

[50] L. Ghoualmi, A. Draa, and S. Chikhi, "An efficient feature selection scheme based on genetic algorithm for ear biometrics authentication," in *2015 12th International Symposium on Programming and Systems (ISPS)*, IEEE, Apr. 2015, pp. 1–5. doi: 10.1109/ISPS.2015.7244991.

[51] A. M. Abo El-Maaty and A. G. Wassal, "Hybrid GA-PCA Feature Selection Approach for Inertial Human Activity Recognition," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, Nov. 2018, pp. 1027–1032. doi: 10.1109/SSCI.2018.8628702.

[52] H. Faris *et al.*, "An intelligent system for spam detection and identification of the most relevant features based on evolutionary Random Weight Networks," *Information Fusion*, vol. 48, pp. 67–83, Aug. 2019, doi: 10.1016/J.INFFUS.2018.08.002.

[53] R. Geetha Ramani and K. Priya, "Improvised emotion and genre detection for songs through signal processing and genetic algorithm," *Concurr Comput*, vol. 31, no. 14, p. e5065, Jul. 2019, doi: 10.1002/cpe.5065.

[54] M. Cui, S. Prasad, M. Mahrooghy, L. M. Bruce, and J. Aanstoos, "Genetic algorithms and linear discriminant analysis based dimensionality reduction for remotely sensed image analysis," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2011, pp. 2373–2376. doi: 10.1109/IGARSS.2011.6049687.

[55] K. Belattar, S. Mostefai, and A. Draa, "A Hybrid GA-LDA Scheme for Feature Selection in Content-Based Image Retrieval," *International Journal of Applied Metaheuristic Computing*, vol. 9, no. 2, pp. 48–71, Feb. 2018, doi: 10.4018/ijamc.2018040103.

[56] S. D. Fabiyi, "New data analysis and dimensionality reduction methods for hyperspectral imagery," 2022, doi: 10.48730/KQED-S704.

[57] X. Hou, Z. Jie, J. Wang, X. Liu, and N. Ye, "Application of terahertz spectroscopy combined with feature improvement algorithm for the identification of adulterated rice seeds," *Infrared Phys Technol*, vol. 131, p. 104694, 2023, doi: 10.1016/j.infrared.2023.104694.

[58] S. A. Wadood *et al.*, "Rice authentication: An overview of different analytical techniques combined with multivariate analysis," *Journal of Food Composition and Analysis*, vol. 112, p. 104677, Sep. 2022, doi: 10.1016/J.JFCA.2022.104677.

[59] K. Kiratiratanapruk *et al.*, "Development of Paddy Rice Seed Classification Process using Machine Learning Techniques for Automatic Grading Machine," *J Sens*, vol. 2020, pp. 1–14, Jul. 2020, doi: 10.1155/2020/7041310.

[60] H. Vu *et al.*, "RGB and VIS/NIR Hyperspectral Imaging Data for 90 Rice Seed Varieties," Dec. 2019, doi: 10.5281/ZENODO.3241923.

[61] C. Liu, W. Liu, X. Lu, W. Chen, J. Yang, and L. Zheng, "Nondestructive determination of transgenic Bacillus thuringiensis rice seeds (Oryza sativa L.) using multispectral imaging and chemometric methods," *Food Chem*, vol. 153, pp. 87–93, Jun. 2014, doi: 10.1016/j.foodchem.2013.11.166.

[62] M. Calzolari, "manuel-calzolari/sklearn-genetic: sklearn-genetic 0.2 | Zenodo." 2019. doi: doi:10.5281/zenodo.3348077.

[63] P. T. T. Hong, T. T. T. Hai, L. T. Lan, V. T. Hoang, V. Hai, and T. T. Nguyen, "Comparative Study on Vision Based Rice Seed Varieties Identification," in *Proceedings - 2015 IEEE International Conference on Knowledge and Systems Engineering, KSE 2015*, Institute of Electrical and Electronics Engineers Inc., Jan. 2016, pp. 377–382. doi: 10.1109/KSE.2015.46.

[64] J. Onmankhong, T. Ma, T. Inagaki, P. Sirisomboon, and S. Tsuchikawa, "Cognitive spectroscopy for the classification of rice varieties: A comparison of machine learning and deep learning approaches in analysing long-wave near-infrared hyperspectral images of brown and milled samples," *Infrared Phys Technol*, vol. 123, no. February, 2022, doi: 10.1016/j.infrared.2022.104100.

[65] J. Sun, L. Zhang, X. Zhou, K. Yao, Y. Tian, and A. Nirere, "A method of information fusion for identification of rice seed varieties based on hyperspectral imaging technology," *J Food Process Eng*, vol. 44, no. 9, 2021, doi: 10.1111/jfpe.13797.

[66] B. Jin *et al.*, "Determination of viability and vigor of naturally-aged rice seeds using hyperspectral imaging with machine learning," *Infrared Phys Technol*, vol. 122, no. February, p. 104097, 2022, doi: 10.1016/j.infrared.2022.104097.

[67] Z. Liu *et al.*, "Comparison of Spectral Indices and Principal Component Analysis for Differentiating Lodged Rice Crop from Normal Ones. 5th Computer and Computing Technologies in Agriculture (CCTA)," pp. 84–92, 2011, doi: 10.1007/978-3-642-27278-3_10ï.

[68] J. Onmankhong, T. Ma, T. Inagaki, P. Sirisomboon, and S. Tsuchikawa, "Cognitive spectroscopy for the classification of rice varieties: A comparison of machine learning and deep learning approaches in analysing long-wave near-infrared hyperspectral images of brown and milled samples," *Infrared Phys Technol*, vol. 123, p. 104100, Jun. 2022, doi: 10.1016/J.INFRARED.2022.104100.

[69] Y. Yang, J. Chen, Y. He, F. Liu, X. Feng, and J. Zhang, "Assessment of the vigor of rice seeds by near-infrared hyperspectral imaging combined with transfer learning," *RSC Adv*, vol. 10, no. 72, pp. 44149–44158, Nov. 2020, doi: 10.1039/d0ra06938h.

[70] B. Jin *et al.*, "Determination of viability and vigor of naturally-aged rice seeds using hyperspectral imaging with machine learning," *InPhT*, vol. 122, p. 104097, May 2022, doi: 10.1016/J.INFRARED.2022.104097.

TABLE I
SUMMARY OF DIMENSIONALITY REDUCTION TECHNIQUES (DRT) FOR HSI-BASED RICE SEEDS CLASSIFICATION

Reference	Varieties	DRT	Classifier	Performance	Year
[61]	2	PCA	PCA, PLSDA, LS-SVM, PCA-BPNN	100%	2014
[3]	5	PCA	LS-SVM, PLSDA, PCA-BPNN	94%	2016
[42]	5	PCA, Segmented PCA	Spectral Angle Mapper (SAM)	-	2010
[20]	20	PCA, LDA	RF	85.94%	2019
[18]	90	LDA	RF	78.27%	2020
[6]	4	PCA	SVM	91.67%	2017
[9]	4	PCA	PLS-da, SIMCA, KNN, SVM AND RF	100 %	2013
[5]	6	PCA	SVM, RF	84%	2016
[1]	3	PCA	PCA, BPNN	94.45%	2015
[67]	2	PCA	PNN	97.8%	2011
[68]	3	PCA	SVM, CNN	95.5%	2022
[65]	5	Bootstrapping soft shrinkage	SVM	99.44%	2021
[69]	2	PCA	CNN, RNN, PLS-DA, SVM	99.50%	2020
[70]	3	PCA	SVM, CNN, LR	89.61%	2022

TABLE II
NAMES OF VARIETIES IN EACH OF THE 10 RANDOM SUBSET FOR THE FIRST AND SECOND GROUPS OF RICE SEEDS

Dataset	Random subsets	Names of varieties in each subset
Group 1 (Each subset contains 10 species)	1st	DT52, DV108, GS55R, KhangDan18, NPT1, NT16, NTP, PC10, PD211, VS1
	2nd	DaiThom8, KhangDan18, LocTroi183, MyHuong88, N54, NPT3, NepCoTien, NepThomHungYen, TC10, VS6
	3rd	9d, DA1, DT52, DT66, DaiThom8, HoangLong, KhangDan18, MyHuong88, NC7, TruongXuanHQ
	4th	CL61, HoangLong, KB6, KimCuong111, LocTroi183, MyHuong88, NC7, ND9, NepCoTien, NepKB19
	5th	DV108, HongQuang15, NBT1, NC7, NT16, NepCoTien, SHPT1, TQ14, TruongXuan1, VS1
	6th	91RH, DT66, HaNa39, KB6, LDA8, MyHuong88, NBK, NH92, NM14, NT16
	7th	DTH155, DTL2, HT18, HaPhat28, HoangLong, N54, NT16, PC10, R998KBL, VietThom8
	8th	CS6, CTX30, MT15, N54, NBK, NBP, NPT3, NTP, PC10, TQ14
	9th	BQ10, BacThomSo7, CNC12, CTX30, DTH155, HT18, NepCoTien, NepThomHungYen, TC10, ThuanViet2
	10th	CNC12, DT66, N98, NBK, NBP, NTP, NepThomHungYen, R998KBL, TC10, VietHuong8
Group 2 (Each subset contains 20 species)	1st	91RH, BT6, CL61, CNC12, DaiThom8, GS55R, H229, HungDan1, KimCuong111, MT15, MyHuong88, N54, NC7, NDC1, NM14, NV1, NepHongNgoc, NepKB19, TruongXuan1, TruongXuanHQ
	2nd	9d, AH1000, CTX30, DT66, HaPhat28, KhangDan18, KimCuong111, MyHuong88, N97, N98, NC2, NH92, NN4B, NPT3, NT16, PD211, SHPT1, TB14, ThuanViet2, VietThom8
	3rd	AH1000, BacThomSo7, CL61, DMV58, DT66, HS1, HT18, HongQuang15, KB27, KB6, KL25, KN5, KhangDan18, KimCuong111, NBK, NBP, NPT1, PD211, SVN1, VS5
	4th	A128, BC15, CS6, CTX30, DTL2, HoangLong, KhangDan18, N54, N97, NBK, NBT1, NH92, NM14, NN4B, NTP, NepPhatQuy, TB13, TB14, TC11-2, VietThom8
	5th	A128, AH1000, DA1, DMV58, DT66, DV108, DaiThom8, GS55R, HaNa39, KhangDan18, N54, N97, NC7, ND9, NM14, NTP, NepHongNgoc, NepThomHungYen, R998KBL, VS5
	6th	BT6, BacThomSo7, CH12, CS6, DMV58, DT52, DTH155, DaiThom8, GiaLoc301, HS1, HoangLong, KB6, KL25, KhangDan18, MyHuong88, ND9, PD211, SHPT1, SVN1, TQ14
	7th	9d, CTX30, H229, HT18, HoangLong, KL25, MyHuong88, NDC1, NH92, NTP, NepKB19, NepPhatQuy, NepThomBacHai, PC10, PD211, SHPT1, TB14, TC10, TQ36, VinhPhuc1
	8th	AH1000, BacThomSo7, CL61, CS6, CTX30, DMV58, DaiThom8, GS55R, HaNa39, KL25, KimCuong111, NC2, NTP, NepPhatQuy, R998KBL, TQ14, TQ36, TruongXuan1, VS5, VietThom8
	9th	A128, CH12, CT286, DT66, LocTroi183, MT15, N98, NBP, NC2, NC7, ND9, NPT3, NTP, NepCoTien, NepPhatQuy, NepThomBacHai, TQ14, ThuanViet2, TruongXuan1, VinhPhuc1
	10th	CH12, CTX30, DMV58, DT66, DaiThom8, GiaLoc301, HT18, HaPhat28, HoangLong, HungDan1, LDA8, N98, NBK, NT16, NV1, PC10, TB14, TC10, TC11-2, VietHuong8

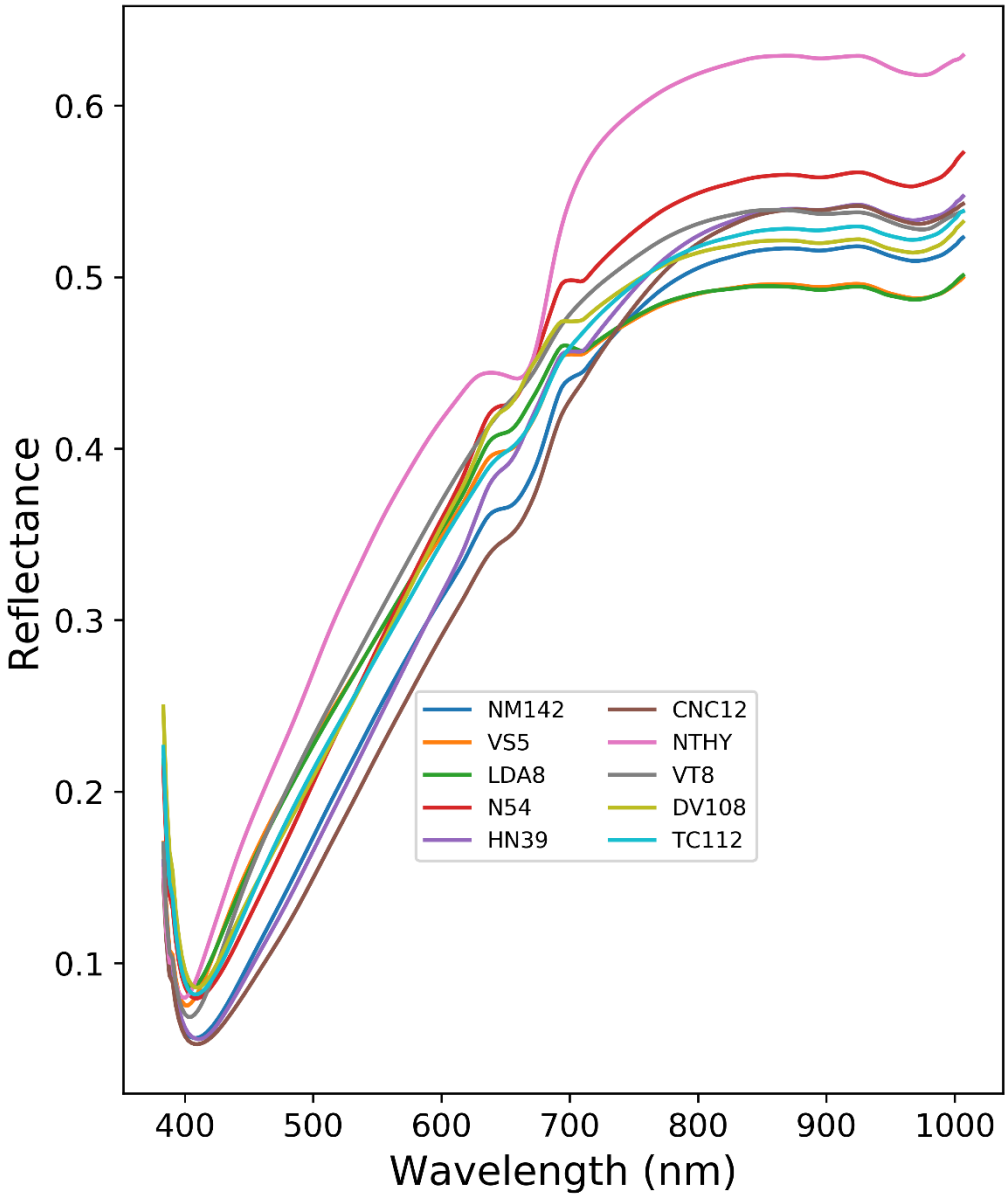

Fig. 1. The average spectral profiles of 10 rice seeds species (more information available in [60]).

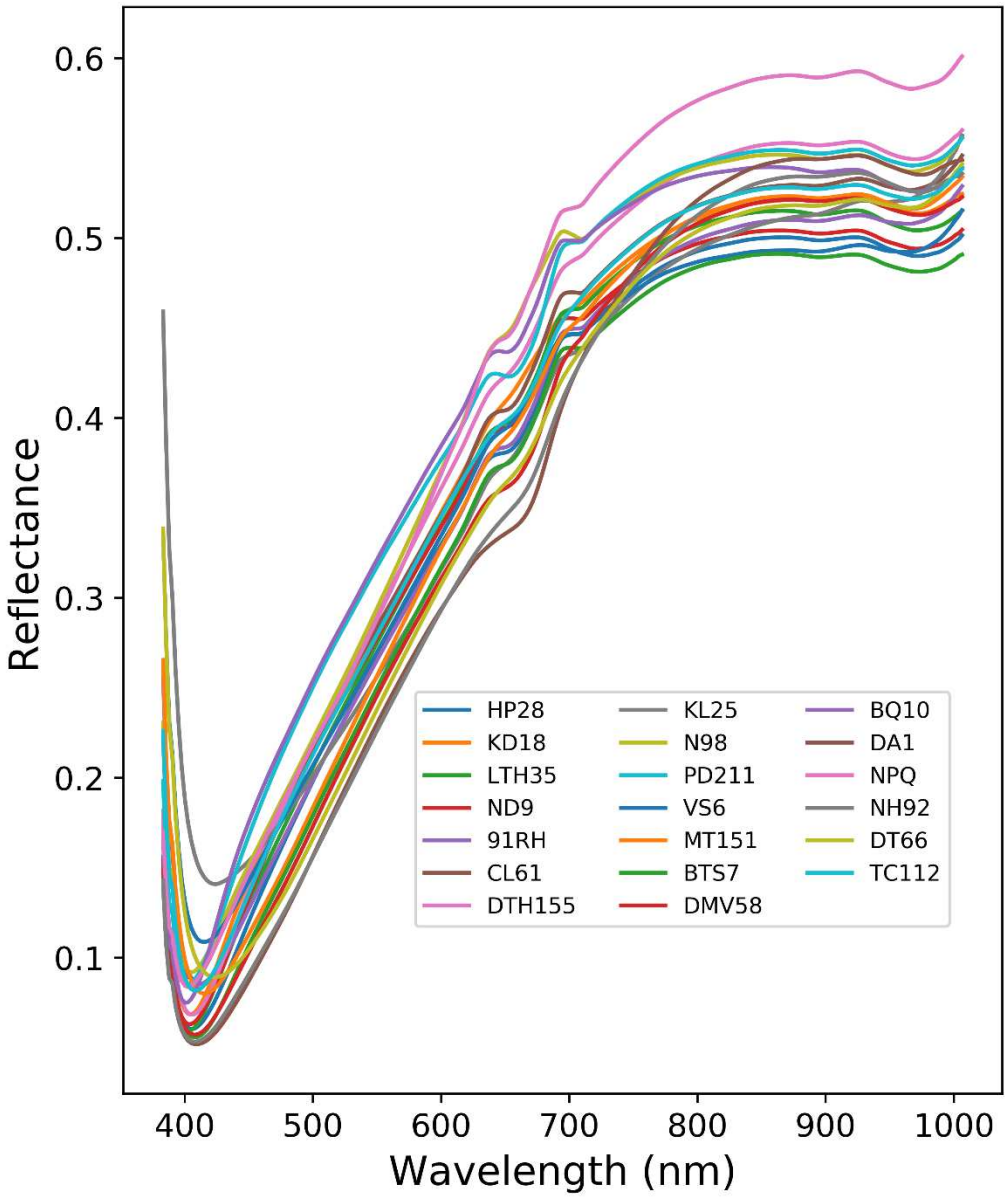

Fig. 2. The average spectral profiles of 20 rice seeds species (more information available in [60]).

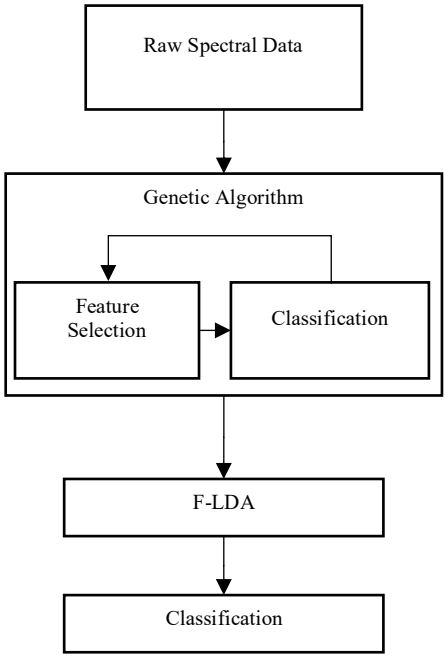

Fig. 3. Proposed technique

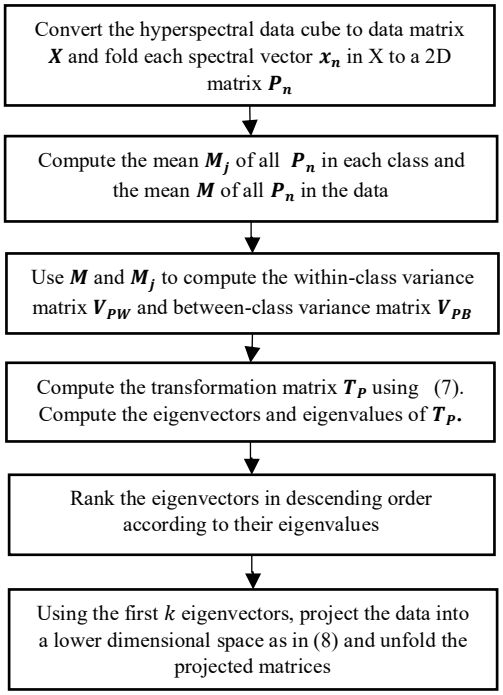

Fig. 4. A flowchart illustrating the implementation of F-LDA [43]

TABLE III
NUMBER OF FEATURES IN EACH DATASET BEFORE AND AFTER THE SELECTION OF OPTIMAL FEATURES USING GA

Dataset	Random subsets	Number of spectral features	Number of selected features
Group 1 Rice seeds (10 species)	1st	256	139
	2nd		96
	3rd		114
	4th		177
	5th		145
	6th		147
	7th		132
	8th		110
	9th		111
	10th		103
Group 2 Rice seeds (20 species)	1st	256	155
	2nd		123
	3rd		183
	4th		225
	5th		212
	6th		157
	7th		137
	8th		94
	9th		140
	10th		128

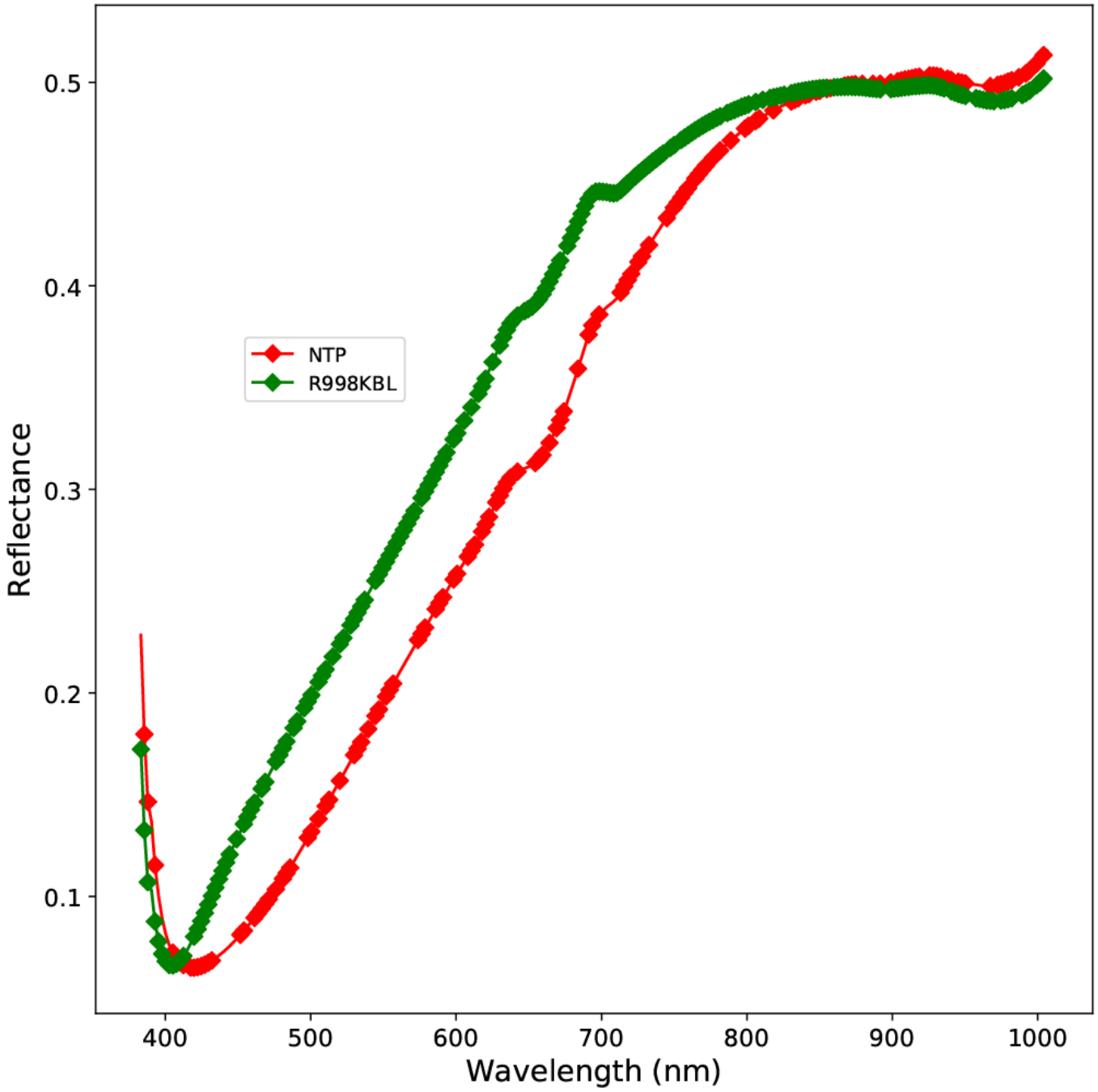

Fig. 5. Average spectra of sample species (NTP and R998KBL) taken from the data subsets containing 10 and 20 rice seed species respectively

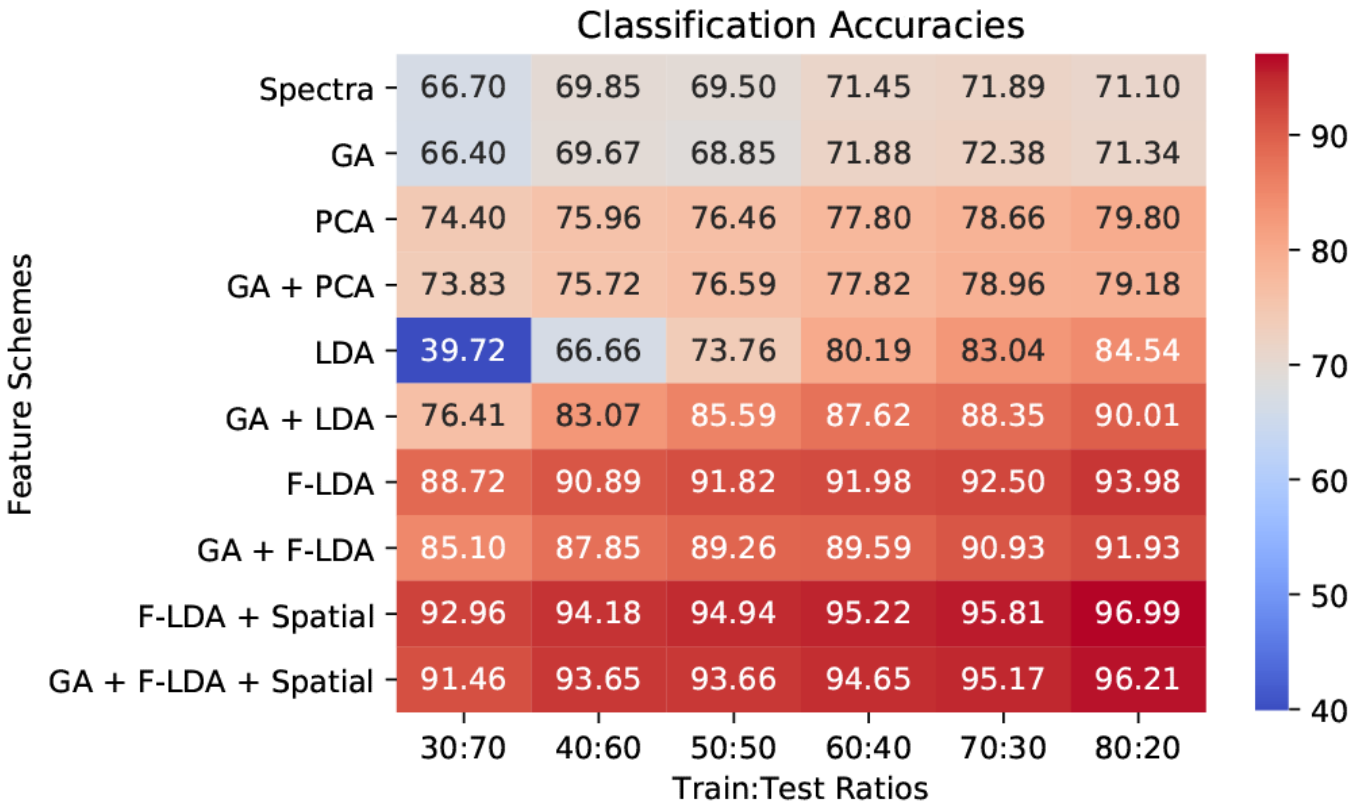

Fig. 6. Accuracy heatmap of the rice seed datasets with 10 varieties

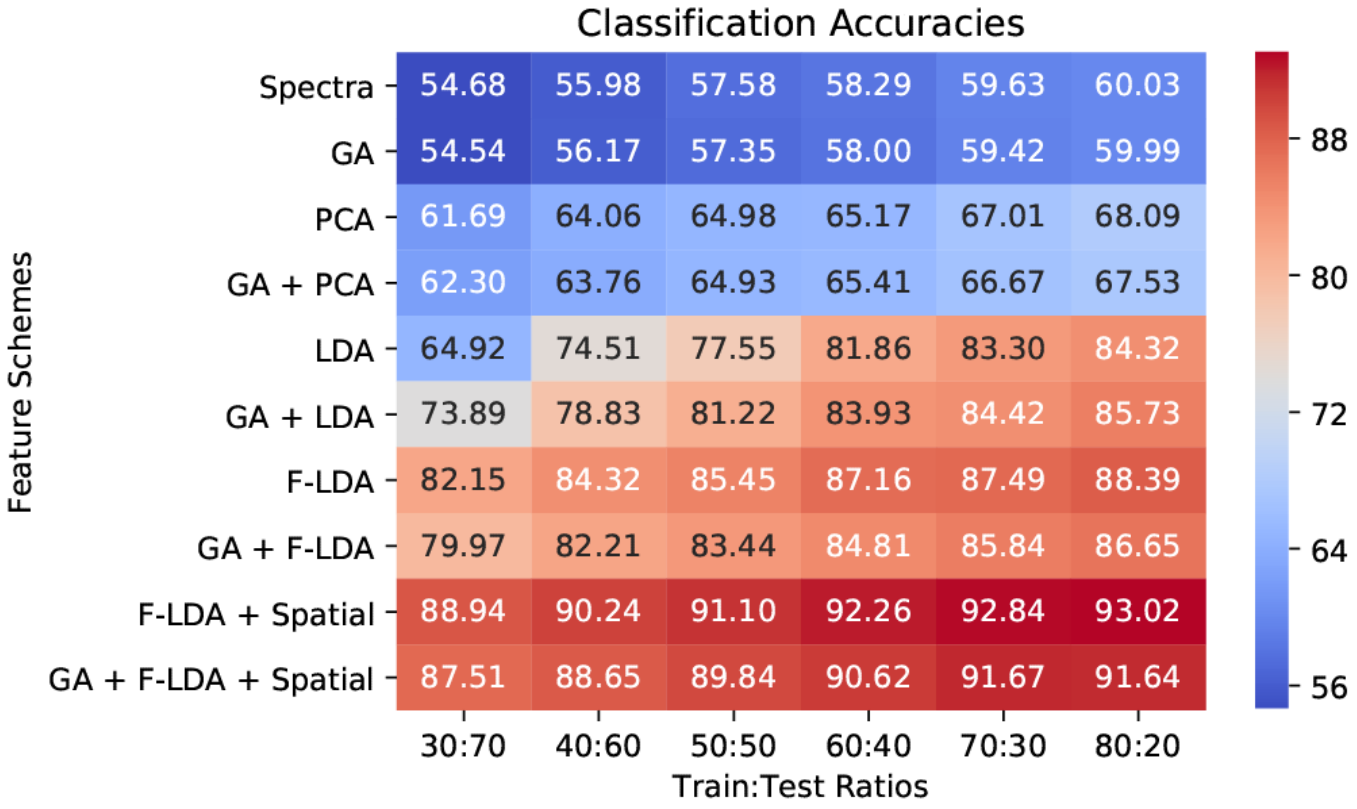

Fig. 7. Accuracy heatmap of the rice seed datasets with 20 varieties

TABLE IV
COMPUTATIONAL COMPLEXITY (CONTENT CONSUMPTION) FOR THE DIFFERENT STAGES OF THE F-LDA WHEN APPLIED SEPARATELY ON THE FULL AND SELECTED GA FEATURES USING THE RICE SEED DATASETS OF 10 SPECIES (N_j AND d REPRESENT THE NUMBER OF SAMPLES PER SPECIES AND THE NUMBER OF FEATURES EXTRACTED BY F-LDA RESPECTIVELY).

Dataset (ratio of training to testing sample size used is 80:20)		Best configuration based on the classifier's accuracy ($G \times B$)	Within-class variance matrix	Between-class variance matrix	Transformation matrix	Eigen problem	Data projection
Random subsets	Computational complexity (formulated for F-LDA in [43])	-	$o(cN_jG^2B)$	$o(cG^2B)$	$o(G^3)$	$o(G^3)$	$o(sGd)$
1st	F-LDA (full features)	128 * 2	o(327680N_j)	o(327680)	o(2097152)	o(2097152)	o(128sd)
	F-LDA (GA features)	139 * 1	o(193210N_j)	o(193210)	o(2685619)	o(2685619)	o(139 sd)
	Saving factors	-	**1.70**	**1.70**	0.78	0.78	0.92
2nd	F-LDA (full features)	64 * 4	o(163840N_j)	o(163840)	o(262144)	o(262144)	o(64 sd)
	F-LDA (GA features)	24 * 4	o(23040N_j)	o(23040)	o(13824)	o(13824)	o(24 sd)
	Saving factors	-	**7.11**	**7.11**	**18.96**	**18.96**	**2.67**
3rd	F-LDA (full features)	32 * 8	o(81920N_j)	o(81920)	o(32768)	o(32768)	o(32 sd)
	F-LDA (GA features)	57 * 2	o(64980N_j)	o(64980)	o(185193)	o(185193)	o(57 sd)
	Saving factors	-	**1.26**	**1.26**	0.18	0.18	0.56
4th	F-LDA (full features)	64 * 4	o(163840N_j)	o(163840)	o(262144)	o(262144)	o(64 sd)
	F-LDA (GA features)	59 * 3	o(104430N_j)	o(104430)	o(205379)	o(205379)	o(59 sd)
	Saving factors	-	**1.57**	**1.57**	**1.28**	**1.28**	**1.08**
5th	F-LDA (full features)	64 * 4	o(163840N_j)	o(163840)	o(262144)	o(262144)	o(64 sd)
	F-LDA (GA features)	29 * 5	o(42050N_j)	o(42050)	o(42050)	o(42050)	o(29 sd)
	Saving factors	-	**3.90**	**3.90**	**6.23**	**6.23**	**2.21**
6th	F-LDA (full features)	64 * 4	o(163840N_j)	o(163840)	o(262144)	o(262144)	o(64 sd)
	F-LDA (GA features)	49 * 3	o(72030N_j)	o(72030)	o(117649)	o(117649)	o(49 sd)
	Saving factors	-	**2.27**	**2.27**	**2.23**	**2.23**	**1.31**
7th	F-LDA (full features)	64 * 4	o(163840N_j)	o(163840)	o(262144)	o(262144)	o(64 sd)
	F-LDA (GA features)	66 * 2	o(87120N_j)	o(87120)	o(287496)	o(287496)	o(66 sd)
	Saving factors	-	**1.88**	**1.88**	0.91	0.91	0.97
8th	F-LDA (full features)	128 * 2	o(327680N_j)	o(327680)	o(2097152)	o(2097152)	o(128 sd)
	F-LDA (GA features)	110 * 1	o(121000N_j)	o(121000)	o(1331000)	o(1331000)	o(110 sd)
	Saving factors	-	**2.71**	**2.71**	**1.58**	**1.58**	**1.16**
9th	F-LDA (full features)	64 * 4	o(163840N_j)	o(163840)	o(262144)	o(262144)	o(64 sd)
	F-LDA (GA features)	37 * 3	o(41070N_j)	o(41070)	o(50653)	o(50653)	o(37 sd)
	Saving factors	-	**3.99**	**3.99**	**5.18**	**5.18**	**1.73**
10th	F-LDA (full features)	128 * 2	o(327680N_j)	o(327680)	o(2097152)	o(2097152)	o(128 sd)
	F-LDA (GA features)	103 * 1	o(106090N_j)	o(106090)	o(1092727)	o(1092727)	o(103 sd)
	Saving factors	-	**3.09**	**3.09**	**1.92**	**1.92**	**1.24**

TABLE V
MEMORY REQUIREMENT (CONTENT CONSUMPTION) AT THE DIFFERENT STAGES OF THE F-LDA WHEN APPLIED SEPARATELY ON THE FULL AND SELECTED GA FEATURES USING THE RICE SEED DATASETS OF 10 SPECIES (N_j AND d REPRESENT THE NUMBER OF SAMPLES PER SPECIES AND THE NUMBER OF FEATURES EXTRACTED BY F-LDA RESPECTIVELY).

Dataset (ratio of training to testing sample size used is 80:20)		Best configuration based on the classifier's accuracy ($G \times B$)	Data matrix size	Within-class variance matrix size	Between-class variance matrix size	Transformation matrix size	Projection matrix size
Random subsets	Memory requirement (formulated for F-LDA in [43])	-	$G \times B$	$G \times G$	$G \times G$	$G \times G$	$G \times {}^{d}/_{B}$
1st	F-LDA (full features)	128 * 2	256	16384	16384	16384	64d
	F-LDA (GA features)	139 * 1	139	19321	19321	19321	139d
	Saving factors	-	**1.84**	0.85	0.85	0.85	0.46
2nd	F-LDA (full features)	64 * 4	256	4096	4096	4096	16d
	F-LDA (GA features)	24 * 4	96	576	576	576	6d
	Saving factors	-	**2.67**	**7.11**	**7.11**	**7.11**	**2.67**
3rd	F-LDA (full features)	32 * 8	256	1024	1024	1024	4d
	F-LDA (GA features)	57 * 2	114	114	114	114	28.50d
	Saving factors	-	**2.25**	**8.98**	**8.98**	**8.98**	0.14
4th	F-LDA (full features)	64 * 4	256	4096	4096	4096	16d
	F-LDA (GA features)	59 * 3	177	3481	3481	3481	19.67d
	Saving factors	-	**1.45**	**1.18**	**1.18**	**1.18**	0.81
5th	F-LDA (full features)	64 * 4	256	4096	4096	4096	16d
	F-LDA (GA features)	29 * 5	145	841	841	841	5.8
	Saving factors	-	**1.77**	**4.87**	**4.87**	**4.87**	**2.76**
6th	F-LDA (full features)	64 * 4	256	4096	4096	4096	16d
	F-LDA (GA features)	49 * 3	147	2401	2401	2401	16.33
	Saving factors	-	**1.74**	**1.71**	**1.71**	**1.71**	0.98
7th	F-LDA (full features)	64 * 4	256	4096	4096	4096	16
	F-LDA (GA features)	66 * 2	132	4356	4356	4356	4356
	Saving factors	-	**1.94**	0.94	0.94	0.94	0.00
8th	F-LDA (full features)	128 * 2	256	16384	16384	16384	64
	F-LDA (GA features)	110 * 1	110	12100	12100	12100	110
	Saving factors	-	**2.33**	**1.35**	**1.35**	**1.35**	0.58
9th	F-LDA (full features)	64 * 4	256	4096	4096	4096	16d
	F-LDA (GA features)	37 * 3	111	1369	1369	1369	12.33d
	Saving factors	-	**2.31**	**2.99**	**2.99**	**2.99**	**1.30**
10th	F-LDA (full features)	128 * 2	256	16384	16384	16384	64d
	F-LDA (GA features)	103 * 1	103	10609	10609	10609	103d
	Saving factors	-	**2.49**	**1.54**	**1.54**	**1.54**	0.62

TABLE VI

COMPUTATIONAL COMPLEXITY (CONTENT CONSUMPTION) FOR THE DIFFERENT STAGES OF THE F-LDA WHEN APPLIED SEPARATELY ON THE FULL AND SELECTED GA FEATURES USING THE RICE SEED DATASETS OF 20 SPECIES (N_j AND d REPRESENT THE NUMBER OF SAMPLES PER SPECIES AND THE NUMBER OF FEATURES EXTRACTED BY F-LDA RESPECTIVELY)

Dataset (ratio of training to testing sample size used is 80:20)		Best configuration based on the classifier's accuracy ($G \times B$)	Within-class variance matrix	Between-class variance matrix	Transformation matrix	Eigen problem	Data projection
Random Subsets	Computational complexity (formulated for F-LDA in [43])	-	$o(cN_jG^2B)$	$o(cG^2B)$	$o(G^3)$	$o(G^3)$	$o(sGd)$
1st	F-LDA (full features)	128 * 2	o(655360N_j)	o(655360)	o(2097152)	o(2097152)	o(128sd)
	F-LDA (GA features)	31 * 5	o(96100N_j)	o(96100)	o(29791)	o(29791)	o(31sd)
	Saving factors	-	**6.82**	**6.82**	**70.40**	**70.40**	**4.13**
2nd	F-LDA (full features)	64 * 4	o(327680N_j)	o(327680)	o(262144)	o(262144)	o(64sd)
	F-LDA (GA features)	41 * 3	o(100860N_j)	o(100860)	o(68921)	o(68921)	o(41sd)
	Saving factors	-	**3.25**	**3.25**	**3.80**	**3.80**	**1.56**
3rd	F-LDA (full features)	128 * 2	o(655360N_j)	o(655360)	o(2097152)	o(2097152)	o(128sd)
	F-LDA (GA features)	183 * 1	o(669780N_j)	o(669780)	o(6128487)	o(6128487)	o(183sd)
	Saving factors	-	0.98	0.98	0.34	0.34	0.70
4th	F-LDA (full features)	128 * 2	o(655360N_j)	o(655360)	o(2097152)	o(2097152)	o(128sd)
	F-LDA (GA features)	75 * 3	o(337500N_j)	o(337500)	o(421875)	o(421875)	o(75sd)
	Saving factors	-	**1.94**	**1.94**	**4.97**	**4.97**	**1.71**
5th	F-LDA (full features)	128 * 2	o(655360N_j)	o(655360)	o(2097152)	o(2097152)	o(128sd)
	F-LDA (GA features)	106 * 2	o(449440N_j)	o(449440)	o(1191016)	o(1191016)	o(106sd)
	Saving factors	-	**1.46**	**1.46**	**1.76**	**1.76**	**1.21**
6th	F-LDA (full features)	128 * 2	o(655360N_j)	o(655360)	o(2097152)	o(2097152)	o(128sd)
	F-LDA (GA features)	157 * 1	o(492980N_j)	o(492980)	o(3869893)	o(3869893)	o(157sd)
	Saving factors	-	**1.33**	**1.33**	0.54	0.54	0.82
7th	F-LDA (full features)	128 * 2	o(655360N_j)	o(655360)	o(2097152)	o(2097152)	o(128sd)
	F-LDA (GA features)	137 * 1	o(375380N_j)	o(375380)	o(2571353)	o(2571353)	o(137sd)
	Saving factors	-	**1.75**	**1.75**	0.82	0.82	0.93
8th	F-LDA (full features)	128 * 2	o(655360N_j)	o(655360)	o(2097152)	o(2097152)	o(128sd)
	F-LDA (GA features)	94 * 1	o(176720N_j)	o(176720)	o(830584)	o(830584)	o(94sd)
	Saving factors	-	**3.71**	**3.71**	**2.52**	**2.52**	**1.36**
9th	F-LDA (full features)	128 * 2	o(655360N_j)	o(655360)	o(2097152)	o(2097152)	o(128sd)
	F-LDA (GA features)	70 * 2	o(196000N_j)	o(196000)	o(343000)	o(343000)	o(70sd)
	Saving factors	-	**3.34**	**3.34**	**6.11**	**6.11**	**1.83**
10th	F-LDA (full features)	64 * 4	o(327680N_j)	o(327680)	o(262144)	o(262144)	o(64sd)
	F-LDA (GA features)	128 * 1	o(327680N_j)	o(327680)	o(2097152)	o(2097152)	o(128sd)
	Saving factors	-	**1.00**	**1.00**	0.13	0.13	0.50

TABLE VII
MEMORY REQUIREMENT (CONTENT CONSUMPTION) AT THE DIFFERENT STAGES OF THE F-LDA WHEN APPLIED SEPARATELY ON THE FULL AND SELECTED GA FEATURES USING THE RICE SEED DATASETS OF 20 SPECIES (N_j AND d REPRESENT THE NUMBER OF SAMPLES PER SPECIES AND THE NUMBER OF FEATURES EXTRACTED BY F-LDA RESPECTIVELY).

Dataset (ratio of training to testing sample size used is 80:20)		Best configuration based on the classifier's accuracy ($G \times B$)	Data matrix size	Within-class variance matrix size	Between-class variance matrix size	Transformation matrix size	Projection matrix size
Random Subsets	Memory requirement (formulated for F-LDA in [43])	-	$G \times B$	$G \times G$	$G \times G$	$G \times G$	$G \times {}^{d}/_{B}$
1st	F-LDA (full features)	128 * 2	256	16384	16384	16384	$64d$
	F-LDA (GA features)	31 * 5	155	155	155	155	$6.20d$
	Saving factors	-	**1.65**	**105.70**	**105.70**	**105.70**	**10.32**
2nd	F-LDA (full features)	64 * 4	256	4096	4096	4096	$16d$
	F-LDA (GA features)	41 * 3	123	1681	1681	1681	$13.67d$
	Saving factors	-	**2.08**	**2.44**	**2.44**	**2.44**	**1.17**
3rd	F-LDA (full features)	128 * 2	256	16384	16384	16384	$64d$
	F-LDA (GA features)	183 * 1	183	33489	33489	33489	$183d$
	Saving factors	-	**1.40**	0.49	0.49	0.49	0.35
4th	F-LDA (full features)	128 * 2	256	16384	16384	16384	$64d$
	F-LDA (GA features)	75 * 3	225	5625	5625	5625	$25d$
	Saving factors	-	**1.14**	**2.91**	**2.91**	**2.91**	**2.56**
5th	F-LDA (full features)	128 * 2	256	16384	16384	16384	$64d$
	F-LDA (GA features)	106 * 2	212	11236	11236	11236	$53d$
	Saving factors	-	**1.21**	**1.46**	**1.46**	**1.46**	**1.21**
6th	F-LDA (full features)	128 * 2	256	16384	16384	16384	$64d$
	F-LDA (GA features)	157 * 1	157	24649	24649	24649	$157d$
	Saving factors	-	**1.63**	0.66	0.66	0.66	0.41
7th	F-LDA (full features)	128 * 2	256	16384	16384	16384	$64d$
	F-LDA (GA features)	137 * 1	137	18769	18769	18769	$137d$
	Saving factors	-	**1.87**	0.87	0.87	0.87	0.47
8th	F-LDA (full features)	128 * 2	256	16384	16384	16384	$64d$
	F-LDA (GA features)	94 * 1	94	8836	8836	8836	$94d$
	Saving factors	-	**2.72**	**1.85**	**1.85**	**1.85**	0.68
9th	F-LDA (full features)	128 * 2	256	16384	16384	16384	$64d$
	F-LDA (GA features)	70 * 2	140	4900	4900	4900	$35d$
	Saving factors	-	**1.83**	**3.34**	**3.34**	**3.34**	**1.83**
10th	F-LDA (full features)	64 * 4	256	4096	4096	4096	$16d$
	F-LDA (GA features)	128 * 1	128	16384	16384	16384	$128d$
	Saving factors	-	**2.00**	0.25	0.25	0.25	0.13

TABLE VIII
FEATURE EXTRACTION TIME (S) OF F-LDA WHEN APPLIED SEPARATELY ON THE FULL AND SELECTED GA FEATURES (THE RATIO OF TRAINING TO TESTING SAMPLE SIZE USED IS 80:20)

F-LDA Approach	Rice seed datasets (10 species)	Rice seed datasets (20 species)
F-LDA (using the full features)	0.0739 ± 0.0314	0.2055 ± 0.0393
F-LDA (using the GA features)	**0.0581 ± 0.0319**	**0.1798 ± 0.0844**