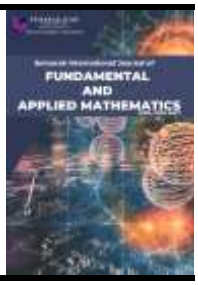
Semarak International Journal of
Fundamental and Applied Mathematics

Journal homepage:
https://semarakilmu.my/index.php/sijfam
ISSN: 3030-5527



Performance Evaluation Criteria for High Dimensional Classification Problems

Nor Aishah Ahad[1,*], Friday Zinzendoff Okwonu[2], Yik Siong Pang[3], Shuhairy Norhisham[4], Muhammad Fadhlullah Abu Bakar[4]

[1] Institute of Strategic Industrial Decision Modeling, School of Quantitative Sciences, College of Arts and Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia
[2] Department of Mathematics, Faculty of Science, Delta State University, P.M.B.1, Abraka, Nigeria
[3] School of Quantitative Sciences, College of Arts and Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia
[4] Department of Civil Engineering, College of Engineering, College of Engineering, Universiti Tenaga Nasional, 43000 Kajang, Selangor, Malaysia

ARTICLE INFO

Article history:
Received 3 November 2024
Received in revised form 17 November 2024
Accepted 1 December 2024
Available online 15 December 2024

Keywords:
Fisher linear classification; independent classification rule; misclassification rate; principal component analysis; variable selection

ABSTRACT

In high dimensional small sample (HDSS) classification problems, the issue of relevant and irrelevant data, the curse of singularity, and dimensionality persist. The presence of irrelevant variables has generated different problems in the classification domain such as computational time, misclassification rate, and performance evaluation criteria. The covariance-dependent classification methods such as the Fisher linear classification method (FLCM) are redundant as such, the independent classification rule (ICR) was coined to solve these problems. Yet, the training and validation of the ICR learned model depends on the relevant and irrelevant data in the variables. To overcome these problems, we applied the principal component analysis (PCA) for dimension reduction on the FLCM (PCA- FLCM), the ICR method (PCA-ICR), F-weighted PCA called W-PCA, and the proposed benchmark extraction method (BEM) to tackle the above mentioned HDSS classification problems. For this study, we investigated the number and percentage of relevant variables selected, computational time, and the probability of correct classification (PCC). To evaluate the performance of these methods, we applied the performance evaluation criteria (PEC) to analyse the probability of correct classification for HDSS classification problems based on the axioms of the probability concept. The results revealed that the W-PCA procedure is very sensitive to select the most vital few variables (Minimum number of vital variables) followed by the BEM procedure. The W-PCA variants have the best computational time while the BEM has the overall best PCC for the data set investigated. The findings demonstrated that the BEM approach outperformed other methods in terms of probability of correct classification while the W-PCA has the best optimal variable search and selection capabilities than the other methods.

* *Corresponding author.*
E-mail address: aishah@uum.edu.my

https://doi.org/10.37934/sijfam.4.1.6174b

1. Introduction

The advent of high dimensional small sample (HDSS) classification problems has rendered the traditional classification methods that depend on covariance matrix impossible due to the curse of singularity and dimensionality of the data matrix. Such massive HDSS data sets often emerged from different fields of study such as biomedical [1], computer vision [2], image and text classification [3], microarray gene expression data [4], and signal processing [5]. The high dimensional and sparsity of these data are severe challenges to data processing [6]. In practice, it is not always easy to extract useful information from the HDSS data set due to the presence of irrelevant data points in the features or variables [7,8]. The irrelevant data set in HDSS may hamper the classification performance of any good classifier and increase the computational time. As a result, different variable or feature selection techniques have been proposed to transform HDSS ($p > n$) to large sample low dimension (LSLD; $n > p$) to improve the classification accuracy and computational speed of any classifier [9,10]. Some variable selection methods often discard the irrelevant data sets in the variables [11].

Variable selection methods (VSM) have been applied as preprocessing methods to obtain better and efficient dimension reduction in HDSS problems [8,11-13]. The VSM is often applied to HDSS data to transform it to LSLD data, then a classifier is applied to LSLD to perform the classification tasks. The principal component analysis (PCA) is a classical VSM often applied for dimension reduction [14-16]. Besides the conventional PCA, improved versions of PCA exist for dimension reduction [17-21]. Apart from the above-mentioned dimension reduction techniques, variable selection is an indirect powerful dimension reduction technique because of the numerical strength of the data point of the variable selected. Variable selection is a step phase for the classifier to achieve robust classification results. Thus, dimension reduction or variable selection may lead to some vital information loss which may result in bias classification performance and analysis [22].

Different variable selection algorithms have been proposed for HDSS classification problems to infer relevant information from the data set [22]. The hybrid feature selection method based on data ranking has been proposed to extract useful information from the data [1]. The two-dimensional linear discriminant method (2DLDM) was discussed to solve HDSS classification problems [2,5,23]. This method was also applied to the LSLD, and it was observed to retain the data structure during variable extraction. It was also shown to exhibit lower computational time compared to their supervised learners such as the linear discriminant analysis [2].

For the HDSS classification problems, the coefficient of the classical classification methods that depends on the covariance matrix such as the Fisher linear classification method (FLCM) cannot be formulated [23] as such the independence classification rule (ICR), diagonal linear discriminant analysis (DLDA) and two-dimensional linear discriminant method (2DLDM) which depends on the diagonal of the covariance matrix and many more were coined. Apart from the diagonal transformation in the aforementioned methods, in the validation stage, all the variables are applied to validate the learned models. For the HDSS classification problems, many relevant and irrelevant data exist which may contribute to a high misclassification rate and high computational time. To solve these problems, we applied the PCA and F-Weighted-PCA (W-PCA) methods to FLCM and ICR methods (PCA-FLCM, PCA-ICR, W-PCA-FLCM, W-PCA-ICR). We also proposed the benchmark extraction method (BEM) to extract vital variables based on the data point numerical strength from the plethora of variables in the data set. Its main functions are to maximally separate samples of different groups, gather samples of the same group conveniently, search and extract or select the very vital few variables to perform the classification tasks. These methods are applied to determine the number, or the percentage of relevant variables selected to learn the models and perform further classification tasks. The main objectives of this study are to determine the number of vital few

variables extracted to build the models, computational time, the probability of correct classification (PCC), and the robustness of these methods based on the comparison between PCC and performance evaluation criteria (PEC).

Variants of variable or feature selection techniques often expunge the irrelevant variables thereby leading to significant loss of information. These variable selection procedures often are not robust because the variable selection techniques were not designed to detect and remedy the influence of outliers. The loss of information during feature selection may hamper the performance of the classifiers. To address the above pitfall of the existing variable selection, first, the conventional principal component analysis (PCA) was modified by introducing weight to reduce the influence of outliers before the PCA was applied to reduce the dimension of the variables from $p > n$ to $n > p$ before the covariance-dependent classifiers are applied but still significant loss of information is observed in the process. To minimize the loss of information, a new data point and variables extraction procedure was proposed, that is the benchmark extraction method (BEM). This method was designed to reduce the influence of outliers and identify the very vital data points that belong to the variables of interest. In other words, it reduces the dimension of the data set based on variable contributions. However, if the number of irrelevant data points contained in the variable is extremely minimal, the benchmark extraction method (BEM) retains the variable but if the number of irrelevant data points exceeds the benchmark, the variable is annihilated. The uniqueness of the new method lies in the fact that it preprocesses the data points based on variables individually and performs classification tasks simultaneously.

This paper is structured as follows. The classical principal component analysis (PCA) based on FLCM and ICR, W-PCA-FLCM, W-PCA-ICR, BEM, and data collection are described in Section 2. Results and discussion are contained in Section 3 followed by conclusions in Section 4.

2. Materials and Methods

2.1 Principal Component Analysis based on Fisher Linear Classification Method (PCA-FLCM)

The classical Fisher linear classification method (FLCM) cannot be applied to HDSS problems directly. To formulate the coefficient of the FLCM, the HDSS problems need to be transformed into LSLD (n > p) problems by using data dimension reduction techniques. In most cases, the principal component analysis (PCA) or variable selection methods are applied to perform this task before the FLCM coefficient can be formulated and further train the classifier [24-26]. The PCA has been applied to different areas for dimensionality reduction before suitable classifiers are applied to perform classification tasks [27-31].

In this consideration, to formulate the PCA-FLCM classifier, the HDSS (p > n) data need to be transformed to n > p problem by selecting the largest eigenvalue contribution which is considered the vital variable in the data [32,33]. The corresponding reduced data set are used to train the FLCM for classification purposes [34].

2.2 Weight Principal Component Analysis based on Fisher Linear Classification Method (W-PCA-FLCM)

In this subsection, we apply the F-weight as shown in Eq. (1)

$$W_i = \frac{C_i}{\partial}, C_i = X_i {X_i}^T, \delta = \sum C_i, i = 1,2 \quad (1)$$

to transform the data matrix before applying the PCA for data dimension reduction. The FLCM is trained and validated based on the selected variables from W-PCA. This method is referred to as W-PCA-FLCM.

2.3 Independent Classification Rule (ICR)

For the HDSS classification problem, the covariance matrix method would suffer the curse of singularity. To dodge the singularity problem, the independent classification rule (ICR) was proposed. The coefficient of this method is formed by taking the diagonal of the covariance matrix [35-37]. The ICR method has shown better performance for LSLD problems [36]. Pang and Tong [36] noted that the performance of ICR for HDSS problems is "not reliable" which may probably be due to the large p for training the classifier. On this note, we will apply PCA and W-PCA to the HDSS data to transform it into LSLD data before applying the ICR classifier to perform the classification task. The PCA-ICR and W-PCA-ICR are based on the data dimension reduction concept discussed above.

2.4 Benchmark Extraction Method (BEM)

The BEM applies the weighted mean approach to search and selects the very vital few variables with the highest contributions. In other words, it searches for vital variables in the data thereby extracting and reducing the column of the data matrix to fewer columns (very relevant variables) such that the covariance dependent classification techniques can be applied to train the classifier. Let $X_i = 1, 2$ be $p > n$ data matrix, where p is the data matrix dimension and n is the sample size. The BEM can be described as follows. The first step is to compute the group mean vectors, that is,

$$\begin{aligned} &\bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}, \bar{\bar{X}} = \frac{\sum_{i=1}^{k} \bar{X}_i}{k}, k = 2 \\ &\mathrm{C}_i = \begin{cases} 1, if\ \bar{X}_i > \bar{\bar{X}} \\ 0, otherwise \end{cases} \end{aligned} \tag{2}$$

Therefore, the weighting steps proceed as follows

$$\partial_i = \mathrm{C}_i \bar{X}_i \tag{3}$$

Based on Eq. (3), search and select the variable with the highest weight for the two groups which can be visualized as

$$\cup_i = \partial_i \partial_i \tag{4}$$

From Eq. (3) or Eq. (4) we can determine the number of relevant variables selected as follows

$$\cap_i = \sum_{j=1}^{n_i} \mathrm{C}_{ij} \tag{5}$$

Then transform the number of variables selected as follows

$$\emptyset_i = \frac{\partial_i}{\cap_i} \tag{6}$$

$$\alpha_i = \frac{\cup_i}{\cap_i} \tag{7}$$

$$\alpha = \frac{\sum_{i=1}^{k} \alpha_i}{k} \tag{8}$$

Step two variable search and selection continue as follows

$$C_{2i} = \begin{cases} 1, if\ \alpha_i > \alpha \\ 0, otherwise \end{cases} \tag{9}$$

$$\cap_{2i} = \sum_{j=1}^{n_i} C_{2ij} \tag{10}$$

Eq. (10) represents the number of variables selected for step 2. Note that each step of variable screening generates its benchmark heuristically based on the available data set in the variable. If the number of variables selected from Eq. (2) and Eq. (9) is different, continue until you find variable selection stability. Variable search and selection stability simply means an attempt to obtain the variables with the highest contributory weight which are the same for all attempts. Consider the following expression

$$\partial_{2i} = C_{2i}\alpha_i \tag{11}$$

From Eq. (11) we have

$$\emptyset_{2i} = \frac{\partial_{2i}}{\cap_{2i}} \tag{12}$$

where ∩2i = ni C2ij. We repeat steps 7 and 8 as follows

$$\alpha_{2i} = \frac{\cup 2_i}{\cap_{2i}} = \frac{\partial_{2i}\partial_{2i}}{\sum_{j=1}^{n_i} C_{2ij}} \tag{13}$$

$$\alpha_2 = \frac{\sum_{i=1}^{k} \alpha_{2i}}{k} \tag{14}$$

Repeat step (9) for new variable selection

$$C_{3i} = \begin{cases} 1, if\ \alpha_{2i} > \alpha_2 \\ 0, otherwise \end{cases} \tag{15}$$

If the number of variables searched and selected by Eq. (15) is equal to the number of variables selected by Eq. (2) and Eq. (9), then we conclude that the variables selected are stable [38], then stop and train the model coefficient as follows

$$\Delta_i = C_{3i}X_i \tag{16}$$

Eq. (16) implies stable numbers of variables have been extracted and indicates that the number of variables selected can be used to train the classifier to improve computational time and classification accuracy. The mean vectors of the extracted vital variables are

$$\bar{V}_1 = \frac{\sum_{j=1}^{n_1} \Delta_1}{n_1}, \bar{V}_2 = \frac{\sum_{j=1}^{n_2} \Delta_2}{n_2} \quad (17)$$

are the weighted screened variable mean vectors for the two groups, n_1, n_2 are the sample sizes for each group, at this point, we have transformed $p > n$ to $n > p$. Henceforth, we can apply any classical classification methods to perform classification tasks on the data set. Let determine the screen mean deviation, that is

$$D = (\bar{V}_1 - \bar{V}_2) \quad (18)$$

Then the group variance and pooled sample variance are defined as

$$S_i = \frac{\sum_{j=1}^{n_i} (\Delta_i - \bar{V}_i)^2}{n_i - 1} \text{ and} \quad (19)$$

$$S_p = \frac{\sum_{i=1}^{k} (n_i - 1) S_i}{\sum_{i=1}^{k} n_i - k}. \quad (20)$$

From these definitions, we formulate the model coefficient

$$\nabla = \frac{(\bar{V}_1 - \bar{V}_2)}{\frac{\sum_{i=1}^{k} (n_i - 1) S_i}{\sum_{i=1}^{k} n_i - k}} = D S_P^{-1}. \quad (21)$$

The classification scores are defined as

$$W_i = \nabla \Delta_i{}^T \quad (22)$$

The classification benchmark is defined as follows

$$\bar{D} = \frac{(\bar{V}_1 + \bar{V}_2)}{k}, \ \bar{W} = \bar{D} S_P^{-1}. \quad (23)$$

From Eq. (22) and Eq. (23) we can assign a new object, that is classified $\boldsymbol{\Delta_1}$ to group one if $\boldsymbol{W_1} \geq \overline{\boldsymbol{W}}$, otherwise, assign it to group two. The BEM selects the variables based on internally computed parameters by continuous search and screening. Once all the screening steps have repeatedly screened similar variables, the process automatically translates to the LSLD system with the original sample size preserved. This process may indicate higher variable selection in one group than in the other. This method can be categorized as the filter method because after the relevant features have been determined, a classifier is trained from the new data set.

2.5 Performance Evaluation Criteria (PEC)

Conventionally, the axiomatic concept of probability that is $0 \leq P \leq 1, P\epsilon[0,1]$ is often used to determine the performance of $p > n$ problems. Due to the irrelevant data points and variables associated with this type of data set, the irrelevant data points affect the performance of any classifier that belongs to this group because the evaluation measure is unsusceptible against irrelevant data points. To solve this problem, it was pertinent to propose evaluation criteria that require the application of the data point used in training and validating the classifier. However, the axiomatic approach focused only on the probability obtained from the classifiers with $P = \sum_{i=1}^{k} p_i = 1$ which in many instances gives a high rate of misclassification. To solve this problem, a new procedure was proposed to remedy the aforementioned. It is observed that the error of misclassification obtained based on this method is minimal compared to the error of misclassification obtained from the axiomatic concept. The proposed method is presented as follows.

Based on the concept of axioms of probability, the optimal probability of correct classification is

$$\Omega = \pi + \tau = 1, \tag{24}$$

where π is the probability of misclassification defined as

$$\pi = \left[\frac{(1-pcc)}{2\times pcc}\right] \times pcc. \tag{25}$$

where pcc denotes the probability of correct classification from the training and validation data [39]. Therefore, the probability of correct classification (τ) is defined as

$$\tau = \nabla = \Omega - \pi. \tag{26}$$

The error of misclassification due to Eq. (26) is minimal compared to the error of misclassification associated with the axiomatic concept.

2.6 Data Collection

In this study, we apply four real data sets from https://www.openml.org to investigate the performance of the methods discussed. Since the study focused on dimension reduction and the utilization of the most vital few relevant variables, we are elected to focus on the number of variables and the percentages of variables selected, or the corresponding percentage of eigenvalues based on the number of relevant variables selected, the computational time of the methods and the probability of correct classification. The first data set is based on mines and sonar rock signal [40], the second data set consists of scene image recognition [41], the third was obtained from https://www.openml.org/d/922, and the fourth Tecator meat data from https://www.openml.org/d/851.

i) Mines and sonar rock signal
This data originally consists of 104 sample size $(n = 104)$ with 60 variables $(p = 60)$. For this study, that is $p > nk$, we select $nk = 50$ for each group and $p = 60$, that is $(60 > 50)$. The results are reported in Table 1 and Figure 1 based on the study variables.

ii) Scene image recognition data set

For this study, the analysis focused on urban and non-urban real images based on the scene. The original data consists of 2407 samples and 294 variables [41], which is $n = 2407$, $p = 294$, where $n_1 = 1976$, $n_2 = 431$, $n = n_1 + n_2 = 2407$. For this study, that is $p > nk$, $nk = 200, n_{1k} = 100, n_{2k} = 100, 294 > nk$. Table 2 and Figure 2 contain the classification performance of the different methods based on the study variables.

iii) https://www.openml.org/d/922
The author and usage of this data are unknown. This data was obtained from https://www.openml.org/d/922. It consists of two classes defined as positive and negative with
$n_1 = 42$ and $n_2 = 58$, $n = \sum_{i=1}^{2} n_i = 100$. To satisfy the condition of this study $(p > n)$, we assumed an equal sample size for the two groups, hence $n_1 = n_2 = 42$, therefore $p > n_1, n_2$. Table 3 and Figure 3 contain the performance analysis of these methods.

iv) Tecator meat data (https://www.openml.org/d/851)
This data set was recorded using Tecator Infratech food and feed analyzer for 850nm-1,050nm wavelength by near-infrared transmission procedure. This data was used to predict the fat content of meat-based on its near-infrared absorbance spectrum (https://www.openml.org/d/851). The data set contains $n_1 = 102$ and $n_2 = 138$, $n = \sum_{i=1}^{2} n_i = 240, p = 124$. We assumed an equal sample size based on n_1 as such, $p > n_k = 102$. The classification result is reported in Table 4 and Figure 4 respectively.

3. Results and Discussion

The results in Table 1, showed that the classical PCA-FLCM, PCA-ICR utilized 46.67% of the relevant variables with corresponding 98% of the eigenvalues while the W-PCA-FLCM and W-PCA-ICR utilized 8.33% of the variables with corresponding 98.75% eigenvalues. The proposed method utilized only 38.33% of the variables. The implication of this is that the values in brackets are the percentage of relevant variables recognized by the respective methods. We observed that the average computational time for all methods is approximately 0.10 CPU time (seconds). The findings demonstrated that the proposed method (BEM) outperformed the PCA variants while the W-PCA-FLCM, and W-PCA-ICR outperformed the classical PCA-FLCM and PCA-ICR. From Table 1, the W-PCA methods identified only ten relevant variables which accounted for 98.75% of the eigenvalue contribution compared to the classical PCA with 98% from 56 variables for both groups. This implies that the W-PCA method is more susceptible to identify the relevant variables than the classical PCA.

Table 1
Analysis of the mines and sonar rock signal based on PCA variants and the BEM

Parameters	PCA-FLCM	PCA-ICR	W-PCA-FLCM	W-PCA-ICR	BEM
NV	56(46.67%)	56(46.67%)	10(8.33%)	10(8.33%)	46(38.33%)
EV(%)	98.00	98.00	98.75	98.75	NA
Time (CPU)	0.11	0.09	0.07	0.11	0.11
$\tau = \nabla$	0.77	0.77	0.79	0.79	**0.87**

NV: number of variables; EV: eigenvalues; τ=∇: the probability of correct classification

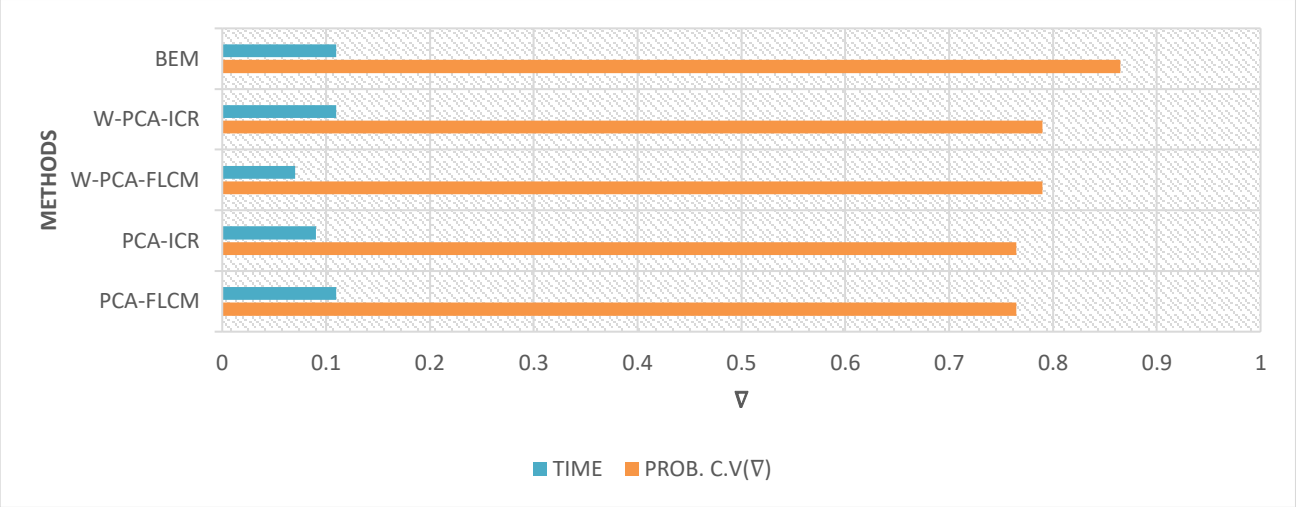


Fig. 1. Comparative analysis of the methods for mines and sonar rock signal based on PCC and CPU time

The results in Table 2 demonstrate the classification performance of these methods based on the number and percentages of variables selected. The classical PCA methods selected 198 variables representing 33.67% of the variables. The 198 variables accounted for 100% eigenvalue contributions with a 77% classification rate. The W-PCA methods selected 12 variables from the two groups which represent 2.04% of the variables, this 2.04% accounted for 97.5% eigenvalue contributions with a 78% classification rate. The proposed method (BEM) selected 210 variables which represent 35.71% of the original variables with a 98% correct classification rate. The average computational time for all the methods is 0.99 CPU time. Based on this data set, the proposed method selected the highest number of variables and it outperformed other methods.

Table 2
Analysis of scene image recognition based on PCA variants and the BEM

Parameters	PCA-FLCM	PCA-ICR	W-PCA-FLCM	W-PCA-ICR	BEM
NV	198(33.67%)	198(33.67%)	12(2.04%)	12(2.04%)	210(35.71%)
EV(%)	100	100	97.5	97.5	NA
Time (CPU)	1.06	1.07	0.95	0.92	0.95
$\tau = \nabla$	0.77	0.77	0.78	0.78	0.98

NV: number of variables; EV: eigenvalues; τ=∇: the probability of correct classification

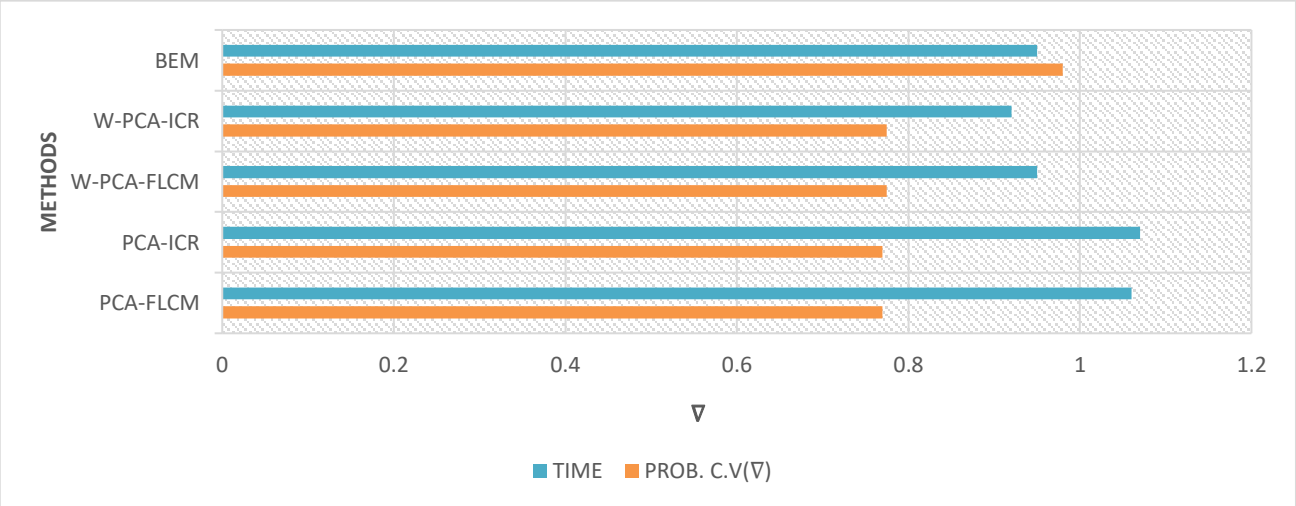


Fig. 2. Comparative analysis of the methods for scene image recognition based on PCC and CPU time

The performance analysis for the binarized data https://www.openml.org/d/922 indicates that the PCA-FLCM, PCA-ICR, and BEM performed comparably as shown in Table 3. The BEM method extracted 20% of the relevant variables while the PCA variants extracted 56% of the relevant variables which accounted for 96% of eigenvalues, also the W-PCA variants extracted 44% of the relevant variables with 97.87% eigenvalues. In terms of computational time and classification accuracy, Figure 3 shows that the conventional PCA and the BEM outperformed the other method meanwhile the W-PCA-ICR has the lowest computational time.

Table 3
Analysis of binarized data based on PCA variants and the BEM

Parameters	PCA-FLCM	PCA-ICR	W-PCA-FLCM	W-PCA-ICR	BEM
NV	56(56%)	56(56%)	44(44%)	44(44%)	20(20%)
EV(%)	96.01	96.01	97.87	97.87	NA
Time (CPU)	0.25	0.21	0.25	0.20	0.23
$\tau = \nabla$	0.80	0.80	0.78	0.78	0.80

NV: number of variables; EV: eigenvalues; τ=∇: the probability of correct classification

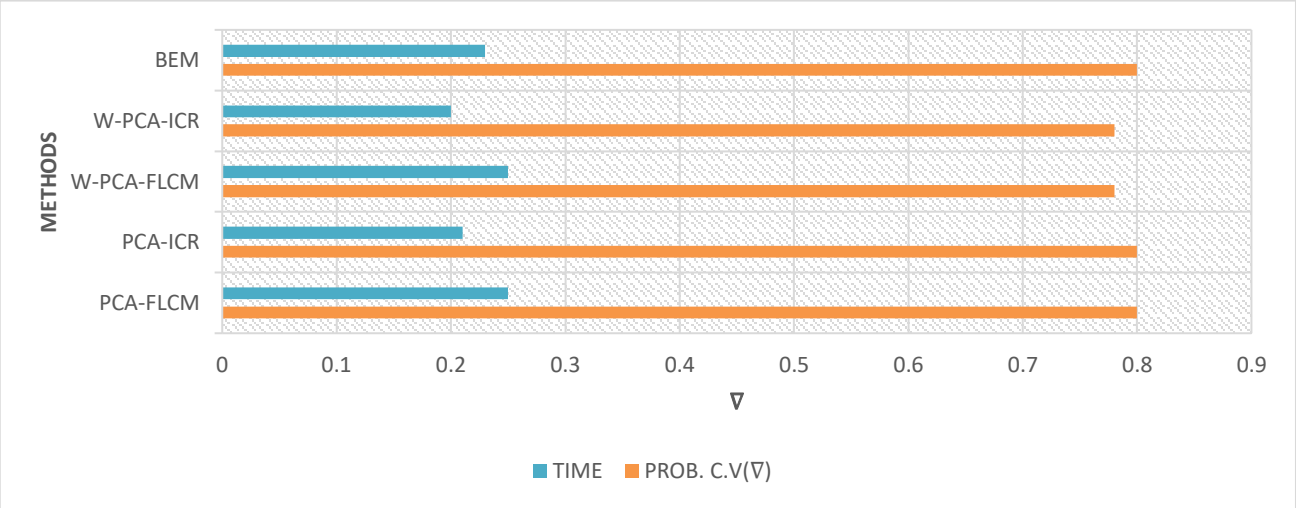


Fig. 3. Comparative analysis of the methods for binarized data based on PCC and CPU time

Based on the results from analyzing tecator meat data as presented in Table 4, it shows that PCA-FLCM and PCA-ICR revealed that 4.03% of the variables were selected with 95.01% eigenvalues meanwhile, the W-PCA-FLCM and W-PCA-ICR selected 1.62% relevant variables with corresponding 99.82% eigenvalues. The BEM selected 1.62% of relevant variables. Figure 4 revealed that the PCA-ICR and BEM have minimum computational time while the PCA-FLCM has the highest computational time. The BEM has the best classification performance followed by the W-PCA variants.

Table 4
Analysis of tecator meat data based on PCA variants and the BEM

Parameters	PCA-FLCM	PCA-ICR	W-PCA-FLCM	W-PCA-ICR	BEM
NV	10(4.03%)	10(4.03%)	4(1.62%)	4(1.62%)	4(1.62%)
EV(%)	95.01	95.01	99.82	99.82	NA
Time (CPU)	0.81	0.73	0.76	0.75	0.73
$\tau = \nabla$	0.77	0.77	0.78	0.78	0.82

NV: number of variables; EV: eigenvalues; τ=∇: the probability of correct classification

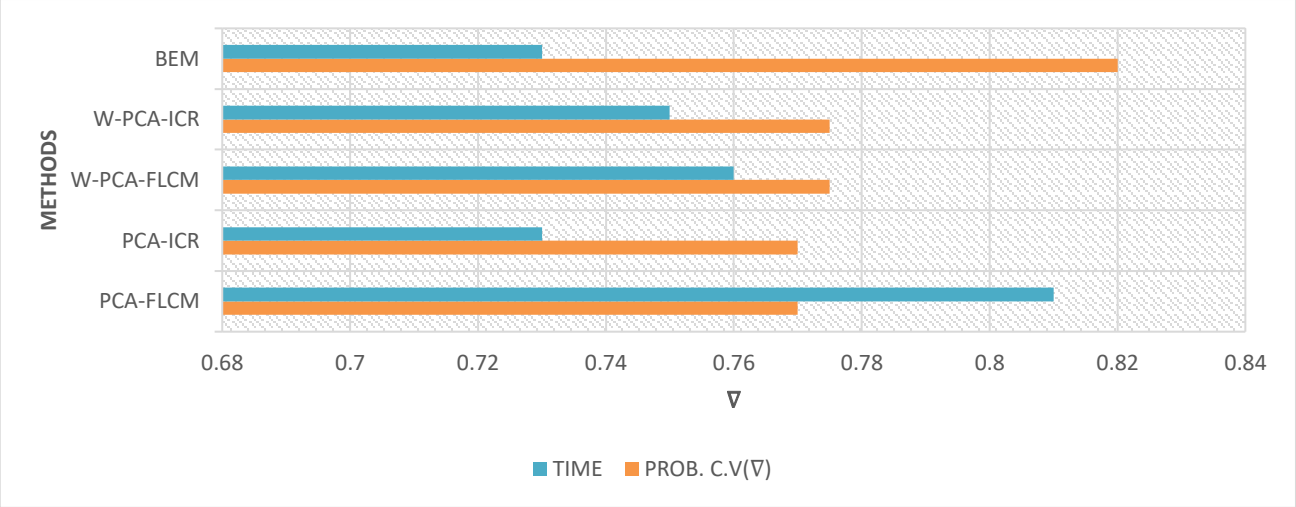


Fig. 4. Comparative analysis of the methods for tecator meat data based on PCC and CPU time

The results have demonstrated that a large chunk of the variables in HDSS problems do not contribute to classification accuracy rather they increase the misclassification rate [42-44]. However, the performance of dimension reduction and variable selection methods may depend strictly on data dependency theory. This theory simply states that the performance of any classification method strictly depends on the data structure and sign direction. The quality of the variables selected, numerical composition, and directions significantly affect the classification performance of any good classification method. The number of variables selected based on the percentage of eigenvalues does not guarantee better classification performance rather the composition and the structure of the data selected play a vital role in doing the classification tasks. For the data set considered in this study, the proposed method has demonstrated better classification performance than the other methods while the F-weighted PCA method has revealed its strength in variable selection and computational time. This study has demonstrated that eigenvalue contribution or the number of variables selected depends on the data composition, structure, and sign directions. The comparative performance analysis showed that PEC is unique in evaluating the performance of HDSS classification problems and very capable of diminishing the likelihood of overfitting occurring in any classification study results.

4. Conclusions

This study has shown the importance of selecting relevant variables in HDSS problems to perform classification tasks. The analysis demonstrated the importance of applying variable selection techniques to extract vital variables to perform classification tasks and to enhance the classification accuracy of the models for HDSS problems. The study further demonstrated that the proposed F-weighted PCA(W-PCA) is very sensitive in extracting the most vital few variables with excellent computational time compared to other methods discussed. The real data application revealed that the BEM procedure showed better classification accuracy based on the performance evaluation criteria (PEC). The study indicates that the performance of any classification model follows the data dependency theory which illustrated that the nature and structure of the data enhance the PEC values of any classification model. Therefore, this study concludes that the BEM is robust over the other methods, the computational time, the search, and selection of very vital few variables by the W-PCA procedure are more superior to the other methods investigated.

It is observed that even though the benchmark extraction method minimizes the loss of information, it can be improved upon by direct transformation of irrelevant data points to relevant data points. This will enhance the classifiers to carry out classification tasks without expunging data points or variables. This process will reduce computational time and increase the number of variables to be included in the data processing stage.

Acknowledgement
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Chaudhuri, A., and T. P. Sahu. "PROMETHEE-based hybrid feature selection technique for high-dimensional biomedical data: application to Parkinson's disease classification." *Electronics Letters* 56, no. 25 (2020): 1403-1406. https://doi.org/10.1049/el.2020.2517

[2] Sun, Weijun, Shengli Xie, and Na Han. "Robust discriminant analysis with adaptive locality preserving." *International Journal of Machine Learning and Cybernetics* 10 (2019): 2791-2804. https://doi.org/10.1007/s13042-018-00903-4

[3] Bolón-Canedo, Verónica, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. "Recent advances and emerging challenges of feature selection in the context of big data." *Knowledge-based Systems* 86 (2015): 33-45. https://doi.org/10.1016/j.knosys.2015.05.014

[4] Kim, HyunJi, Byong Su Choi, and Moon Yul Huh. "Booster in high dimensional data classification." *IEEE Transactions on Knowledge and Data Engineering* 28, no. 1 (2015): 29-40. https://doi.org/10.1109/TKDE.2015.2458867

[5] Lu, Yuwu, Chun Yuan, Zhihui Lai, Xuelong Li, David Zhang, and Wai Keung Wong. "Horizontal and vertical nuclear norm-based 2DLDA for image representation." *IEEE Transactions on Circuits and Systems for Video Technology* 29, no. 4 (2018): 941-955. https://doi.org/10.1109/TCSVT.2018.2822761

[6] Varghese, Jijo, and P. Tamil Selvan. "A Novel Clustering and Matrix Based Computation for Big Data Dimensionality Reduction and Classification." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 32, no. 1 (2023): 238-251. https://doi.org/10.37934/araset.32.1.238251

[7] Yu, Lei, and Huan Liu. "Efficient feature selection via analysis of relevance and redundancy." *The Journal of Machine Learning Research* 5 (2004): 1205-1224. https://dblp.org/rec/journals/jmlr/YuL04

[8] Jović, Alan, Karla Brkić, and Nikola Bogunović. "A review of feature selection methods with applications." In *2015 38th International Convention on Information And Communication Technology, Electronics And Microelectronics (MIPRO)*, pp. 1200-1205. Ieee, 2015. https://doi.org/10.1109/MIPRO.2015.7160458

[9] Zhang, Mengmeng, Wei Li, Qian Du, Lianru Gao, and Bing Zhang. "Feature extraction for classification of hyperspectral and LiDAR data using patch-to-patch CNN." *IEEE Transactions on Cybernetics* 50, no. 1 (2018): 100-111. https://doi.org/10.1109/TCYB.2018.2864670

[10] Wu, Yue, Steven CH Hoi, Tao Mei, and Nenghai Yu. "Large-scale online feature selection for ultra-high dimensional sparse data." ACM Transactions on Knowledge Discovery from Data (TKDD) 11, no. 4 (2017): 1-22. https://doi.org/10.1145/3070646

[11] Remeseiro, Beatriz, and Veronica Bolon-Canedo. "A review of feature selection methods in medical applications." *Computers in Biology and Medicine* 112 (2019): 103375. https://doi.org/10.1016/j.compbiomed.2019.103375

[12] Carvalho, Walisson Ferreira, and Luis Zarate. "Causal Feature Selection." In *Integration Challenges for Analytics, Business Intelligence, and Data Mining*, pp. 145-160. IGI Global, 2021. https://doi.org/10.4018/978-1-7998-5781-5.ch007

[13] Ghanbari, Najme. "A review of feature selection methods with the applications in pattern recognition in the last decade." In *Fundamental Research in Electrical Engineering: The Selected Papers of The First International Conference on Fundamental Research in Electrical Engineering*, pp. 163-171. Springer Singapore, 2019. https://doi.org/10.1007/978-981-10-8672-4_12

[14] Bair, Eric, Trevor Hastie, Debashis Paul, and Robert Tibshirani. "Prediction by supervised principal components." *Journal of the American Statistical Association* 101, no. 473 (2006): 119-137. https://doi.org/10.1198/016214505000000628

[15] Ma, Ji, and Yuyu Yuan. "Dimension reduction of image deep feature using PCA." *Journal of Visual Communication and Image Representation* 63 (2019): 102578. https://doi.org/10.1016/j.jvcir.2019.102578

[16] Li, Lingjun, Shigang Liu, Yali Peng, and Zengguo Sun. "Overview of principal component analysis algorithm." *Optik* 127, no. 9 (2016): 3935-3944. https://doi.org/10.1016/j.ijleo.2016.01.033
[17] Liu, Lydia T., Edgar Dobriban, and Amit Singer. "ePCA: High dimensional exponential family PCA." (2018): 2121-2150. https://doi.org/10.1214/18-AOAS1146
[18] Thomas, Minta, Kris De Brabanter, and Bart De Moor. "New bandwidth selection criterion for Kernel PCA: Approach to dimensionality reduction and classification problems." *BMC Bioinformatics* 15 (2014): 1-12. https://doi.org/10.1186/1471-2105-15-137
[19] Drees, Holger, and Anne Sabourin. "Principal component analysis for multivariate extremes." *Electronic Journal of Statistics* 15, no. 1 (2021): 908-943. https://doi.org/10.1214/21-EJS1803
[20] Sando, Keishi, and Hideitsu Hino. "Modal principal component analysis." *Neural Computation* 32, no. 10 (2020): 1901-1935. https://doi.org/10.1162/neco_a_01308
[21] Farrugia, Jessica, Sholeem Griffin, Vasilis P. Valdramidis, Kenneth Camilleri, and Owen Falzon. "Principal component analysis of hyperspectral data for early detection of mould in cheeselets." *Current Research in Food Science* 4 (2021): 18-27. https://doi.org/10.1016/j.crfs.2020.12.003
[22] Li, Yanxia, Yi Chai, Hongpeng Yin, and Bo Chen. "A novel feature learning framework for high-dimensional data classification." *International Journal of Machine Learning and Cybernetics* 12 (2021): 555-569. https://doi.org/10.1007/s13042-020-01188-2
[23] Yang, Jian, David Zhang, Xu Yong, and Jing-yu Yang. "Two-dimensional discriminant transform for face recognition." *Pattern recognition* 38, no. 7 (2005): 1125-1129. https://doi.org/10.1016/j.patcog.2004.11.019
[24] Liu, Chengjun, and Harry Wechsler. "Robust coding schemes for indexing and retrieval from large face databases." *IEEE Transactions on Image Processing* 9, no. 1 (2000): 132-137. https://doi.org/10.1109/83.817604
[25] Belhumeur, Peter N., Joao P. Hespanha, and David J. Kriegman. "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19, no. 7 (1997): 711-720. https://doi.org/10.1109/34.598228
[26] Karanwal, Shekhar. "A comparative study of 14 state of art descriptors for face recognition." *Multimedia Tools and Applications* 80, no. 8 (2021): 12195-12234. https://doi.org/10.1007/s11042-020-09833-2
[27] Jain, Rahul, Ram Kumar Karsh, and Abul Abbas Barbhuiya. "Face Recognition Using Computational Efficient Algorithms." In *2020 4th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, pp. 1-5. IEEE, 2020. https://doi.org/10.1109/IEMENTech51367.2020.9270115
[28] Zhao, Shiqi, Xuzhou Wu, X. Zhang, B. Li, J. Mao, and J. Xu. "Automatic gesture recognition with surface electromyography signal." *J. Xi'an Jiaotong Univ* 54 (2020): 149-156. http://doi.org/10.7652/xjtuxb202009017
[29] Zhou, An-li, Jin-hua Jiang, Chun-xiao Sun, Xin-zhong Xu, and Xin-ming Lu. "Identification of Different Origins of Hetian Jade Based on Statistical Methods of Multi-Element Content." *Guangpuxue yu Guangpu Fenxi/Spectroscopy and Spectral Analysis* 40 (2020): 3174-3178. https://www.gpxygpfx.com/EN/10.3964/j.issn.1000-0593(2020)10-3174-05
[30] Rapa, Mattia, Salvatore Ciano, Roberto Ruggieri, and Giuliana Vinci. "Bioactive compounds in cherry tomatoes (Solanum Lycopersicum var. Cerasiforme): Cultivation techniques classification by multivariate analysis." *Food Chemistry* 355 (2021): 129630. https://doi.org/10.1016/j.foodchem.2021.129630
[31] Ye, Wenjing, Weiwei Lu, Yanping Tang, Guoxi Chen, Xiaopan Li, Chen Ji, Min Hou et al. "Identification of COVID-19 clinical phenotypes by principal component analysis-based cluster analysis." *Frontiers in Medicine* 7 (2020): 570614. https://doi.org/10.3389/fmed.2020.570614
[32] Mahmoudi, Mohammad Reza, Mohammad Hossein Heydari, Sultan Noman Qasem, Amirhosein Mosavi, and Shahab S. Band. "Principal component analysis to study the relations between the spread rates of COVID-19 in high risks countries." *Alexandria Engineering Journal* 60, no. 1 (2021): 457-464. https://doi.org/10.1016/j.aej.2020.09.013
[33] Beattie, J. Renwick, and Francis WL Esmonde-White. "Exploration of principal component analysis: deriving principal component analysis visually using spectra." *Applied Spectroscopy* 75, no. 4 (2021): 361-375. https://doi.org/10.1177/0003702820987847
[34] Ricciardi, Carlo, Antonio Saverio Valente, Kyle Edmund, Valeria Cantoni, Roberta Green, Antonella Fiorillo, Ilaria Picone, Stefania Santini, and Mario Cesarelli. "Linear discriminant analysis and principal component analysis to predict coronary artery disease." *Health Informatics Journal* 26, no. 3 (2020): 2181-2192. https://doi.org/10.1177/1460458219899210
[35] Bickel, Peter J., and Elizaveta Levina. "Some theory for Fisher's linear discriminant function,naive Bayes', and some alternatives when there are many more variables than observations." *Bernoulli* 10, no. 6 (2004): 989-1010. https://doi.org/10.3150/bj/1106314847
[36] Pang, Herbert, and Tiejun Tong. "Recent advances in discriminant analysis for high-dimensional data classification". *Journal of Biometrics & Biostatistics* 3, no. 1 (2012): 1-2. https://doi.org/10.4172/2155-6180.1000e106

[37] Dudoit, Sandrine, Jane Fridlyand, and Terence P. Speed. "Comparison of discrimination methods for the classification of tumors using gene expression data." *Journal of the American Statistical Association* 97, no. 457 (2002): 77-87. https://doi.org/10.1198/016214502753479248

[38] Kalousis, Alexandros, Julien Prados, and Melanie Hilario. "Stability of feature selection algorithms: a study on high-dimensional spaces." Knowledge and Information Systems 12 (2007): 95-116. https://doi.org/10.1007/s10115-006-0040-8

[39] Okwonu, Friday Zinzendoff, Nor Aishah Ahad, Innocent Ejiro Okoloko, Joshua Sarduana Apanapudor, Saadi Ahmad Kamaruddin, and Festus Irimisose Arunaye. "Robust hybrid classification methods and applications." *Pertanika Journal of Science and Technology* 30, no. 4 (2022). https://doi.org/10.47836/pjst.30.4.29

[40] Gorman, R. Paul, and Terrence J. Sejnowski. "Analysis of hidden units in a layered network trained to classify sonar targets." *Neural Networks* 1, no. 1 (1988): 75-89. https://doi.org/10.1016/0893-6080(88)90023-8

[41] Boutell, Matthew R., Jiebo Luo, Xipeng Shen, and Christopher M. Brown. "Learning multi-label scene classification." *Pattern Recognition* 37, no. 9 (2004): 1757-1771. https://doi.org/10.1016/j.patcog.2004.03.009

[42] Fan, Jianqing, and Jinchi Lv. "Sure independence screening for ultrahigh dimensional feature space." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 70, no. 5 (2008): 849-911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

[43] Nandy, Debmalya, Francesca Chiaromonte, and Runze Li. "Covariate information number for feature screening in ultrahigh-dimensional supervised problems." *Journal of the American Statistical Association* 117, no. 539 (2022): 1516-1529. https://doi.org//10.1080/01621459.2020.1864380

[44] Zambom, Adriano Zanin, and Gregory J. Matthews. "Sure independence screening in the presence of missing data." *Statistical Papers* 62 (2021): 817-845. https://doi.org/10.1007/s00362-019-01115-w