IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 18, 2019, accepted May 1, 2019, date of publication May 8, 2019, date of current version May 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2915519

Dimensionality Reduction in Gene Expression Data Sets

JOVANI TAVEIRA DE SOUZA[1], ANTONIO CARLOS DE FRANCISCO[1], AND DAYANA CARLA DE MACEDO[2]
[1]Department of Production Engineering, Federal University of Technology, Paraná 84016-210, Brazil
[2]Department of Administration, Midwest University of Paraná, Paraná 84500-000, Brazil

Corresponding author: Jovani Taveira de Souza (jovanisouza5@gmail.com)

This work was developed under the financial support of the Fundação Araucária Paraná/Brazil, the Coordination of Improvement of Higher Education Personnel (CAPES), and the National Council for Scientific and Technological Development (CNPq).

ABSTRACT Dimensionality reduction is used in microarray data analysis to enhance prediction quality, reduce computing time, and construct more robust models. In addition, the algorithm learning performance involves an expressive number of attributes (genes) relative to the classes (samples). Therefore, in this paper, we conducted a detailed comparison of two reduction methods, attribute selection and principal component analysis, to analyze gene expression data sets. Both reduction methods were employed in the pre-processing stage and then evaluated experimentally. Furthermore, we introduced a combination of consistency-based subset evaluation (CSE) and minimum redundancy maximum relevance (mRMR), which we referred to as CSE-mRMR, to improve classification efficiency. The results indicated a significant increase in classifier hit rates with both methods, compared to using all attributes. By employing cross-validation, attribute selection outperformed PCA consistently across classifiers and datasets, and CSE-mRMR demonstrated good classification performance in the data sets. Taken together, the literature and current results suggest that the attribute selection may be relevant in the analysis and future prediction of gene expression data sets.

INDEX TERMS Attribute selection, consistency-based subset evaluation, gene expression, minimum redundancy maximum relevance, principal component analysis.

I. INTRODUCTION

For over a decade, microarray technology has been widely used in cancer research, simultaneously profiling expression levels of thousands of genes towards improving tumor classification and detection [1], [2]. However, one common drawback to this technology is high dimensionality, resulting from an expressive number of irrelevant genes and a small sample size [3]–[5].

According to Archetti *et al.* [6], studies involving gene expression analyses using microarray techniques found that the number of attributes (genes) usually results in data sets larger than instances (samples), thereby requiring effective reduction techniques to identify reliable and relevant associations between these interactions. In this sense, machine learning algorithms and advanced statistical techniques can help, since they are critical for high-dimensional data processing [7]. A good way to accomplish this is through dimensionality reduction (DR), which aims to reduce the volume of information contained in data sets and more specifically, the attributes, thereby enhancing the functioning capability of the learning methods by eliminating inconsistent data.

DR is a very important issue in the processing of high-dimensional data [7]. It is essential to select the most relevant attributes [8], since they facilitate data analysis and visualization, taking actions based on the knowledge obtained [9]. However, due to the nature of specific problems in each database, methods must be chosen with caution [10].

There are a few techniques employed for DR in databases, for instance, attribute selection (AS), which is essential for identifying the relevant attributes for a specific task. AS selects a subset of relevant attributes that seek to produce comparable or better results as opposed to cases where all attributes are used [11]. AS can be split into filter and wrapper approaches. The difference between the filter and wrapper approaches is that in the former, the attribute subsets are evaluated according to an independent measurement, while in the latter, the subset utilizes the classification algorithm for the evaluation.

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Asif Naeem.

Another important technique is the Principal Component Analysis (PCA), which is one of the most traditional techniques used for reduction, and can predict a sequence of the best linear combinations based on the original attributes of a certain set [12]. PCA reveals a new and reduced set of variables, determined to be the principal components, while also ensuring that little data from the original set is excluded from the analysis [12]. The principal components are ordered according to highest variance based on the original attributes [13].

Previous research, such as the study by Borges and Nievola [9] and Macedo *et al.* [14] has compared AS with random projection and methods based on framework (DRM-F), respectively, to verify the applicability of these methods in the field of gene expression. However, these studies have failed to present other important filtering measures, such as the fast correlation-based filter (FCBF) and the minimum redundancy maximum relevance (mRMR), which is highly relevant to biomedical data [15]. As criterion, this filter has the minimum redundancy and maximum relevance of features [16]. Additionally, it completes calculations more quickly when compared to other approaches [17].

New combinations for size reduction in microarray data have been performed in recent years, such as the study by Akadi and Amine [18], who developed a combination of mRMR and genetic algorithm; or Alshamham *et al.* [16], who proposed a new model combining mRMR and artificial bee colony. The method proposed by Ebrahimpour and Eftekhari [19] involved the combination of mRMR and Hesitant Fuzzy Sets. Both techniques were performed to form subsets of genes and analyze classification accuracy.

Besides providing a thorough analysis of the use of two DR methods and drawing a comparison between them in gene expression data sets [20], this paper also introduces and evaluates the CSE-mRMR, which is a combination of consistency-based subset evaluation (CSE) and mRMR. The CSE is a metric that eliminates redundant data from the assessment of the degree of consistency of the values of the class attribute [21]. To the best of our knowledge, no published articles exist that propose an evaluation of this combination. Cross-validation will be employed as an evaluation criterion, using classification accuracy and standard deviation.

This paper has the following structure: Section II and III describes the DR methods employed, while Section IV introduces the research methods. Section V displays results and discussion, and finally, Section VI provides the main conclusions of this paper.

II. ATTRIBUTE SELECTION (AS)

AS is one of the most important techniques used to reduce dimensionality, as its objective is to remove redundant attributes that are considered irrelevant to the data mining process [22] by identifying a smaller set of attributes that produce comparable or better ranking results than the case in which all attributes are used [11]. This technique has been the focus of attention because of its great potential in several applications, including bio-computing, medicine, data processing, and object recognition [23]. AS can be divided into filter and wrapper approaches, both of which are described below.

A. FILTER APPROACH

The filter approach uses the data features to perform the evaluation and select the subsets of features without involving a mining algorithm [24], [25]. First, the method selects the data; subsequently, it performs the classification process without considering the interactions among the attributes. The filter usually ranks features based on statistical calculations [26]. In our work, we employed the techniques used most commonly to evaluate subsets, such as correlation-based feature selection (CFS), consistency-based subset evaluation (CSE), minimum redundancy maximum relevance (mRMR), and fast correlation-based filter (FCBF). In addition, we introduced a combination of CSE and mRMR, to which we referred as CSE-mRMR.

1) CFS AND CSE

CFS classifies the subsets generated according to a correlation function based on a heuristic reward of evaluation. The subset of attributes is evaluated through the individual prediction of each and the degree of correlation between them [27]. The evaluation usually refers to subsets that correlate highly with the class [28], and is calculated with the following equation:

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{rff}}}. \tag{1}$$

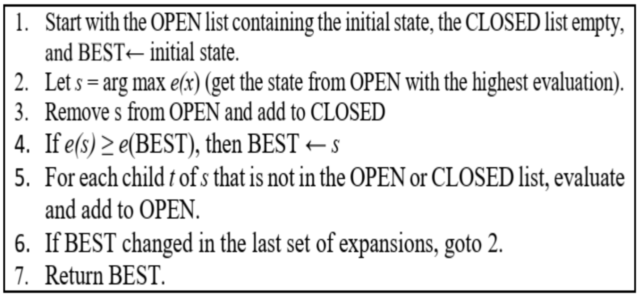
in which $Merit_s$ is the heurist merit of attribute subsets s containing k features, $\overline{r_{cf}}$ is the mean correlation between each resource and the class labels in s, and $\overline{rff}$ is the mean correlation between two features [29]. Therefore, in our case, the primary goal was to identify the subset in the initial set of microarray data that has highly correlated genes.

In relation to CSE, the attribute set is reduced preserving the original consistency, and the value of an attribute subset is assessed by the degree of consistency among the values verified for each class when the training instances are projected onto the attribute subset [27]. The evaluation method is calculated by following equation:

$$Consistency_s = 1 - \frac{\sum_{i=0}^{j} |D_i| - |M_i|}{N}. \tag{2}$$

where s represents an attribute subset, j denotes the number of distinct combinations of attribute values for s, $|D_i|$ expresses the number of occurrences of the ith attribute value combination, $|M_i|$ presents the cardinality of the majority class for an ith attribute value combination, and N is the total number of instances in the data set [30].

Based on preliminary experiments, we adopted the best-first algorithm as a search strategy to search through the space



```
1. Start with the OPEN list containing the initial state, the CLOSED list empty, and BEST← initial state.
2. Let $s = \arg\max e(x)$ (get the state from OPEN with the highest evaluation).
3. Remove s from OPEN and add to CLOSED
4. If $e(s) \geq e(\text{BEST})$, then BEST $\leftarrow s$
5. For each child $t$ of $s$ that is not in the OPEN or CLOSED list, evaluate and add to OPEN.
6. If BEST changed in the last set of expansions, goto 2.
7. Return BEST.
```

FIGURE 1. **The best-first search algorithm.**

of attribute subsets for both the CFS and the CSE. According to the Mursalin *et al.* [22], the best-first search starts with an empty set of features and then checks all possible single feature expansions. The subset with the maximum evaluation is assigned and expanded in the same manner by adding single features. If expanding the subset results in no improvement, the best-first search algorithm reverts to the next-best unexpanded feature subset and continues from there. The best-first search algorithm searches the candidate feature subset space in this manner and returns the best subset found when the search terminates. Fig.1 summarizes the main aspects of the best-first search algorithm.

2) MRMR

The mRMR measure is designed to analyze the quality and provide the best predictive performance of a subset of variables in view of the output variable (class attribute). According to Rana *et al.* [31], mRMR maximizes the relevance of the selected features with class labels while concurrently minimizing the redundancy among the selected features after considering mutual information (MI). This measure is frequently used for biomedical data [15]. For its calculation, it is first necessary to present the concept of MI. Considering two variables, x and y, their MIs are grounded in relation to their probabilistic density functions.

$$I\left(x;y\right)=\iint p\left(x,y\right)\log\frac{p\left(x,y\right)}{p\left(x\right)p\left(y\right)}dxdy, \tag{3}$$

thus, relevance and redundancy (high and low MIs, respectively) are calculated as:

$$Relevance\left(S,c\right)=\frac{1}{|S|}\sum_{f_i\in S}I\left(f_i;c\right), \tag{4}$$

$$Redundancy\left(S\right)=\frac{1}{|S|^2}\sum_{f_i,f_j\in S}I\left(f_i;c\right), \tag{5}$$

in which S refers to the feature set, c is the class, and f represents individual feature in S. As reported, the mRMR rates resources by concurrently maximizing relevancy and minimizing redundancy. In this sense, this operation is designated by the operator, Φ. From this, we obtain:

$$max\,\Phi\left(Relevance, Redundancy\right)=Relevance-Redundancy. \tag{6}$$

One important factor is combining mRMR with other approaches [32]; this combination can improve mRMR's accuracy and speed. In this sense, we will present a combination of CSE and mRMR in this section.

3) FCBF

FCBF developed by Yu and Liu [33] is used to find an adequate measure of correlations between features and a procedure to choose features based on the correlation between two random variables [34]. FCBF chooses the feature that most correlates to the class attribute. As a measure of correlation, the concept of symmetric uncertainty (SU) is used. SU can be calculated based on entropy (measure of the uncertainty of a random variable). To define this measure, some concepts must be presented first. The entropy of a feature X, for example, is defined as follows:

$$H\left(X\right)=-\sum\nolimits_i P\left(x_i\right)\log_2\left(P\left(x_i\right)\right), \tag{7}$$

and the entropy of X after observing values of another variable Y is defined as follows:

$$H\left(X\mid Y\right)=-\sum\nolimits_i P\left(y_i\right)\sum\nolimits_i P\left(x_i|\,y_i\right)\log_2\left(P\left(x_i|\,y_i\right)\right), \tag{8}$$

where $P\left(x_i\right)$ is the prior probability for all values of X, $P\left(x_i|\,y_i\right)$ is the posterior probabilities of X given the values of Y. In this sense, the amount by which the entropy of X decreases reflects additional information about X provided by Y is called information gain (IG) [35]. IG is defined as follows:

$$IG\left(X\mid Y\right)=H\left(X\right)-H\left(X\mid Y\right), \tag{9}$$

thus, according to the definitions reported above, SU is defined as follows:

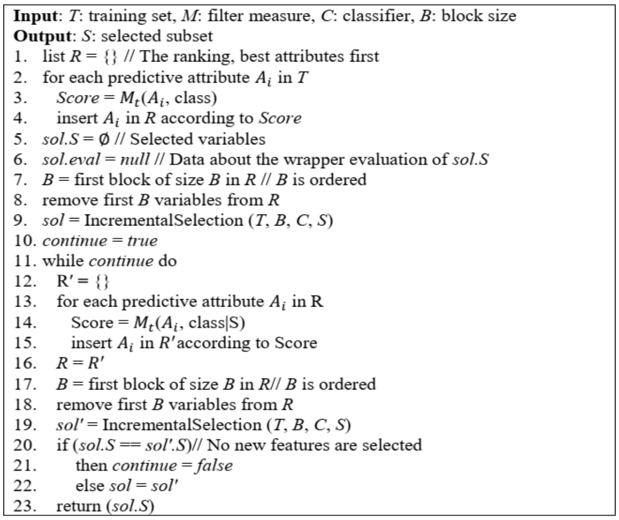
$$SU\left(X,Y\right)=2\left[\frac{IG\left(X\mid Y\right)}{H\left(X\right)+H(Y)}\right]. \tag{10}$$

when the value of SU is equal to 0 it is denoted that the variables X and Y are independent, whereas a value equal to 1 implies that the knowledge about one variable is enough to perfectly determine the other [36].

4) CSE-MRMR

In addition to the traditional methods reported above, we sought to combine two of the employed methods (CSE and mRMR) in order to verify their applicability to the gene expression data sets, as well as improving the accuracy of classification. As far as we know, no published studies have addressed this combination. The combination involved one method being nested inside the other one. First, the evaluation of the subset examined the consistency of the values for each class, posteriorly using the relevance and redundancy of the subsets (concepts have already been reported).

We used the re-ranking search [37] (Fig. 2) as a search strategy for mRMR, which uses IG to evaluate its information. In general, the search focuses primarily on creating a classification of all attributes in descending order, using



```
Input: T: training set, M: filter measure, C: classifier, B: block size
Output: S: selected subset
list R = {} // The ranking, best attributes first
for each predictive attribute $A_i$ in T
    Score = $M_t(A_i$, class)
    insert $A_i$ in R according to Score
sol.S = ∅ // Selected variables
sol.eval = null // Data about the wrapper evaluation of sol.S
B = first block of size B in R // B is ordered
remove first B variables from R
sol = IncrementalSelection (T, B, C, S)
continue = true
while continue do
    R' = {}
    for each predictive attribute $A_i$ in R
        Score = $M_t(A_i$, class|S)
        insert $A_i$ in R'according to Score
    R = R'
    B = first block of size B in R// B is ordered
    remove first B variables from R
    sol' = IncrementalSelection (T, B, C, S)
    if (sol.S == sol'.S)// No new features are selected
        then continue = false
        else sol = sol'
    return (sol.S)
```

FIGURE 2. Re-ranking search.

the attribute evaluation metric (i.e., information gain or symmetric uncertainty). Subsequently, the classification is fragmented into blocks based on its size, and the attribute selection search algorithm is applied to the first block. From there on, the rest of the ranking is modified according to an approximation metric of each attribute. The attribute selection search algorithm is executed again for the first block, and the process is completed when a new block does not modify the selected subset. In this step, one can choose any search algorithm used for attribute selection. Also, three different approach methods are executed to approximate the re-ranking of the remaining attributes in the ranking. In this case, these methods are the conditional mutual information maximization, the mutual information-based feature selection, and the max-relevance and min-redundancy [37].

B. WRAPPER APPROACH (WA)

The WA utilizes a predefined mining algorithm, with its performance as the evaluation criterion. It requires the termination of a mining algorithm, being its criterion utilized for subset evaluation [24], [25]. The WA seeks the best of features, aiming to enhance algorithm performance, but having a higher computational value than the filter model [38].

III. PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal Component Analysis (PCA) is a statistical technique used to study several applications with high dimensionality. It primarily aims to reduce the dimensionality of each data set, thereby creating a new set of variables while preserving the original information as much as possible. Put differently, the method transforms, orthogonally, a set of correlated variables into a set of linearly non-correlated variables, which are known as principal components. For better visualization, the steps of the PCA can be represented by a flow chart, as illustrated in Fig. 3.

In short, the principal components are obtained by calculating the eigenvalues and eigenvectors of the covariance matrix, defined by C, as presented in (11):

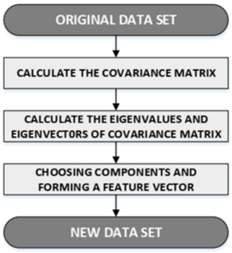

ORIGINAL DATA SET
CALCULATE THE COVARIANCE MATRIX
CALCULATE THE EIGENVALUES AND EIGENVECTORS OF COVARIANCE MATRIX
CHOOSING COMPONENTS AND FORMING A FEATURE VECTOR
NEW DATA SET

FIGURE 3. Flow chart of the PCA.

$$Cvi = \lambda_i v_i, \tag{11}$$

The covariance matrix of vectors of the original data X is represented by C. λ_i refers to the eigenvalues of matrix C, and v_i corresponds to the matching eigenvectors. Consecutively, the eigenvectors k, which correspond to the largest eigenvalues k, need to be computed [39] in order to reduce the data dimensionality. Considering $E_k = [v_1, v_2, v_3, \ldots\ldots., v_k]$ and $\Lambda = [\lambda_1, \lambda_2, \lambda_3 \ldots.\lambda_k]$, we have $CE_k = E_K \Lambda$. Finally, it is possible to obtain the following equation:

$$X^{PCA} = E_K^T X. \tag{12}$$

Regarding (12), the features of the original data matrix X are reduced by multiplying with the matrix dxk, which has eigenvectors k corresponding to the largest eigenvalues k. The result matrix is X^{PCA}, which corresponds to the new data set formed.

IV. RESEARCH METHODS

To test and evaluate the quality of the methods in this research, four key steps were performed: i) database selection, ii) pre-processing, iii) data mining and iv) post-processing. Some of these experiments have been performed by Souza *et al*. [20]. Fig. 4 shows the experimental process.

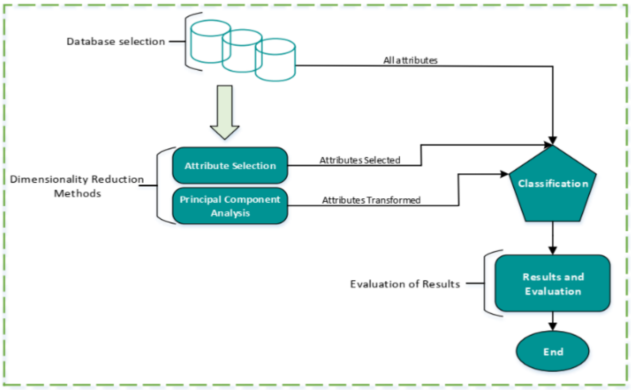


FIGURE 4. Experimental outline.

Firstly, the databases for study were selected. Subsequently, a classification process was utilized considering all database attributes. The next step referred to the AS and PCA methods, using the selected database attributes for classification. The software used was Waikato Environment for

Knowledge Analysis (WEKA) version 3.8. The steps are detailed below.

A. DESCRIPTION OF DATABASES

Databases were selected from a biomedical data repository [40]. The study employed three databases.

The first, the LungCancer-Michigan database [41] was composed of 96 samples. Of these, 86 were primary lung adenocarcinoma samples and the remaining 10 were non-neoplastic pulmonary samples. The set contained 7,129 attributes (genes).

The second, the LungCancer-Ontario database [42] was composed of 39 non-small cell lung cancer (NSCLC) samples, corresponding to non-small cell carcinomas. Of the 39 samples, 24 were related to patients who already had tumor relapses or metastases (labeled in the database as "Relapse"). The remaining samples comprised 15 non-neoplastic pulmonary samples (labeled "Non-Relapse"). The set contained 2,880 attributes (genes).

The third and final database, was the LungCancer-Harvard [43] database, and was composed of 203 samples, with 139 lung adenocarcinoma samples (labeled "ADEN"), 21 squamous cell lung cancer samples (labeled "SQUA"), 20 carcinoid lung tumor samples (labeled "COID"), 6 small cell lung carcinoma samples (labeled "SCLC") and 17 normal lung samples (labeled "NORMAL"). The set contained 12,600 attributes (genes).

B. PRE-PROCESSING: ATTRIBUTE SELECTION (AS)

For the AS process, two approaches were employed: filter and wrapper. The filter approach utilized CFS, CSE, mRMR, and FCBF measures. Furthermore, we combined CSE and mRMR filtering measures, which we referred to as CSE-mRMR. For this, we used the *Attribute Selected Classifier*, which combined selection methods with any algorithm for classification analysis. The WA engaged the classifier algorithms Naive Bayes (NB), J48, SVM, 1-NN, 3-NN, 5-NN and 7-NN. After utilizing the method, each approach generated new data subsets, containing only the relevant attributes [20].

C. PRE-PROCESSING: PRINCIPAL COMPONENT ANALYSIS (PCA)

This method submitted the databases to the PCA process. It is important to report that the method ignores the attribute class, also named attribute meta, when the principal components are computed. Therefore, they are new attributes derived from the original group (principal components). The utilization criteria for the PCA method were defined according to the variance percentage in the original data. The selected percentages were 90%, 95% and 99%.

D. DATA MINING

This stage is regarded as the core of knowledge discovery, and focused on extracting patterns from the data. In this stage, the methods and algorithms performed searches for useful knowledge in the data sets. This study applied the AS and PCA methods to three gene expression data analyses with seven classifiers (NB, J48, SVM, 1-NN, 3-NN, 5-NN, and 7-NN).

E. POST-PROCESSING

In the final stage, the information from the previous stages was analyzed. Here, the results were evaluated using (I) the classification accuracy (CA) or hit rate, which refers to the number of correctly sorted instances divided by the total number of instances; and (II) the standard deviation of 10 runs with 10-fold cross validation. For this calculation (referring to CA metric), the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) were considered. The equation for these metrics are as follows:

$$CA = \frac{TP + TN}{TP + TN + FP + FN} \times 100, \tag{13}$$

In additional, the sample standard deviation was calculated. Using the selected class ($i = 1, 2, 3, \ldots.., k$) and the selected attribute ($j = 1, 2, 3, \ldots.., k$), we obtain:

$$s_{ij} = \sqrt{\frac{\sum_{l=1}^{N}(x_{ij} - \overline{x_{ij}})^2}{N-1}}. \tag{14}$$

where s_{ij} is the sample standard deviation of all the samples in class i, x_{ij} represents the sample value of the jth attribute from the lth in class i, and $\overline{x_{ij}}$ is the mean value.

V. RESULTS AND DISCUSSION

This section outlines the results of each method as well as draws a comparison between these methods. The methods were compared to select the one that achieved the best average (performance) in the analyzed data sets. To verify their applicability, it was necessary to use the original database and the subsets generated by the methods. The three selected databases were submitted to the seven classifiers.

Regarding the AS method, we employed CFS, CSE, FCBF, and mRMR measures (filter approach). Additionally, we showed the values that correspond to the CSE-mRMR measure. The CSE-mRMR was not considered in calculation of the average filter approach, only the AS method was mentioned. For the PCA method, we used the variance percentage of the original data as the criterion, which was defined as 90%, 95% and 99%.

The NB algorithm was defined as the base algorithm. The italicized and underlined values in the table indicate that the result is significantly better or shows a slight improvement than the base algorithm.

A. RESULTS FOR SELECTED DATABASES

To measure the performance of the generated predictive models, the classifier efficiency was verified for both subsets through the mean and standard deviation of the hit rate. To evaluate them, the models were first executed on the bases with all attributes for the appropriate comparisons.

Table 1 presents the results related to the LungCancer-Michigan database.

TABLE 1. Results of the attribute selection and principal component analysis methods for the Lungcancer-Michigan database.

	Algorithms						
Subsets	**Naive Bayes**	**J48**	**SVM**	**1-NN**	**3-NN**	**5-NN**	**7-NN**
All attributes	100.0 ± 0	99.00 ± 3.16	*100.0 ± 0*	*100.0 ± 0*	98.89 ± 3.51	98.89 ± 3.51	*100.0 ± 0*
CFS	100.0 ± 0	99.00 ± 3.16	99.0 ± 3.16	*100.0 ± 0*	*100.0 ± 0*	*100.0 ± 0*	*100.0 ± 0*
CSE	96.78 ± 5.20	*99.00 ± 3.16*	95.89 ± 5.32	98.00 ± 4.22	*99.00 ± 3.16*	*99.00 ± 3.16*	*99.00 ± 3.16*
mRMR	100.0 ± 0	99.00 ± 3.16	99.00 ± 3.16	*100,0 ± 0*	*100.0 ± 0*	*100.0 ± 0*	*100.0 ± 0*
FCBF	97.78 ± 4.68	98.89 ± 3.51	94.89 ± 5.40	98.89 ± 3.51	98.89 ± 3.51	98.89 ± 3.51	*100.0 ± 0*
WA	100.0 ± 0	98.89 ± 3.51	*100.0 ± 0*	*100.0 ± 0*	99.00 ± 3.16	*100.0 ± 0*	*100.0 ± 0*
CSE-mRMR	97.78 ± 4.68	98.89 ± 3.51	94.89 ± 5.40	98.89 ± 3.51	98.89 ± 3.51	98.89 ± 3.51	*100.0 ± 0*
90% of attributes	99.00 ± 3.16	98.89 ± 3.51	*100,0 ± 0*	95.78 ± 7.22	96.78 ± 5.20	92.67 ± 8.69	93.78 ± 7.15
95% of attributes	99.00 ± 3.16	98.89 ± 3.51	98.00 ± 4.22	96.89 ± 5.02	96.89 ± 5.02	92.89 ± 4.92	92.89 ± 4.92
99% of attributes	99.00 ± 3.16	98.89 ± 3.51	91.67 ± 4.42	91.67 ± 4.42	89.56 ± 0.57	89.56 ± 0.57	89.56 ± 0.57

In reference to the AS method, the data from Table 1 shows that the approach with the best average performance was the WA with a hit rate of 99.70%, against a hit rate of 98.86% of the filter approach. The CSE-mRMR measure presented 98.32% of the average.

Table 1 also shows the algorithms with statistically significant averages, which include the k-NN (k = 1, k = 3, k = 5, and k = 7) algorithms for the CFS and mRMR measures, the J48 and k-NN algorithms for the CSE and FCBF measures. The averages are based on a comparison with the base algorithm (Naive Bayes).

With regards to the WA, only the J48 and 3-NN algorithms showed lower hit rates than the NB algorithm. Furthermore, in relation to the CSE-mRMR, the J48 and k-NN algorithms denoted higher values than the base algorithm.

Finally, only the SVM algorithm with 90% of the attributes from the original database presented a hit rate that is higher than the base algorithm for the PCA method. The subset that produced the best average was the subset with 90% of the attributes, with a hit rate of 96.70%.

Comparing the methods employed, an average hit rate of 99.54% was obtained for the utilization with all the database's attributes. The AS method reached 98.98%, while the PCA method obtained an average hit rate of 95.35%.

Therefore, the utilization with all the database's attributes showed the highest hit rate. However, the AS method obtained a value close to that of the original database, and it is worth mentioning that the PCA method also presented excellent results with a hit rate over 90%, indicating that both methods are well equipped for application in this database.

Table 2 presents the information related to the LungCancer-Ontario database.

The data presented in Table 2 shows that for the AS method, the WA presented the highest hit rate (88.33%), compared to 66.40% with the filter approach. The CSE-mRMR measure had an average of 67.74%, and thus was better than the filter approach.

TABLE 2. Results of the attribute selection and principal component analysis methods for the Lungcancer-Ontario database.

	Algorithms						
Subsets	**Naive Bayes**	**J48**	**SVM**	**1-NN**	**3-NN**	**5-NN**	**7-NN**
All attributes	67.50± 35.45	*84.17 ± 13.86*	*78.33 ± 22.97*	56.67 ± 19.95	55.83 ± 34.26	57.50 ± 39.18	58.33 ± 32.63
CFS	74.17± 20.58	71.67± 21.94	69.17± 15.74	44.17± 27.51	59.17± 31.29	61.67 ± 33.61	64.17± 33.58
CSE	74.17± 23.12	71.67± 14.27	61.67± 12.55	*79.17± 16.32*	74.17± 20.58	74.17 ± 20.58	74.17 ± 20.58
mRMR	76.67 ± 18.76	74.17 ± 23.72	71.67 ± 14.27	64.17 ± 23.91	56.67 ± 28.54	64.17 ± 26.66	59.17 ± 31.29
FCBF	69.17 ± 15.74	*76.67 ± 22.15*	63.33 ± 19.72	61.67 ± 23.96	61.67 ± 23.96	56.67 ± 28.54	54.17 ± 25.23
WA	84.17± 18.19	85.00± 17.48	82.50± 16.87	*95.00± 10.54*	*95.00± 10.54*	*92.50 ± 12.08*	84.17 ± 18.19
CSE-mRMR	76.67 ± 18.76	74.17 ± 23.72	66.67 ± 11.79	61.67 ± 29.19	61.67 ± 29.19	64.17 ± 20.81	69.17 ± 22.92
90% of attributes	69.17± 10.43	67.50± 28.99	63.33± 19.72	68.33± 25.09	*73.33± 23.17*	63.33 ± 19.72	58.33 ± 25.46
95% of attributes	75.83± 24.67	57.50± 28.99	55.83± 21.89	54.17± 27.85	60.83± 22.92	60.83 ± 25.78	53.33 ± 20.86
99% of attributes	58.33± 28.05	62.50± 29.46	62.50± 27.00	55.83± 21.89	*74.17± 16.87*	*64.17± 12.45*	58.33± 15.21

In relation to the CFS, mRMR, and CSE-mRMR measures, both classifier algorithms presented significantly worse results than the base algorithm. The best-performing algorithm was the 1-NN and J48 algorithms, for the CSE and FCBF measures, respectively. For the WA, only the SVM algorithm presented a hit rate lower than the base algorithm.

Regarding the PCA method, the algorithm with the best hit rate was 3-NN, for subsets with 90% and 99% of attributes from the original database, in comparison with the NB algorithm. In addition, the J48, SVM, and 5-NN algorithms showed statistically significant results considering subsets with 99% of the attributes. The best-performing subset was that conducted with 90% of the database's attributes, which obtained an average hit rate of 66.19%.

When comparing the performance of the two methods, the AS method was found to be superior with a hit rate of 70.38%, compared with 62.74% for the PCA method. The average hit rate for all of the database's attributes was 65.48%. It is important to note that even though the PCA method presented a lower average hit rate than that of all of the attributes it presented a higher average performance in some cases. Thus, the data above indicated that the AS method performed statistically better than the PCA method.

Table 3 displays information regarding the LungCancer-Harvard database.

Table 3 shows that the AS method presented values higher than the PCA method. Regarding the two approaches of AS, the WA offered the best average, with a hit rate of 97.62%, compared with 93.97% for the filter approach. The CSE-mRMR measure had an average of 92.60%.

TABLE 3. Results of the attribute selection and principal component analysis methods for the Lungcancer-Harvard database.

	Algorithms						
Subsets	**Naive Bayes**	**J48**	**SVM**	**1-NN**	**3-NN**	**5-NN**	**7-NN**
All attributes	80.38 ± 6.16	*93.12* *± 7.50*	*95.07* *± 3.34*	*89.71* *± 5.27*	*91.21* *± 4.89*	*89.74* *± 5.70*	*89.24* *± 5.40*
CFS	95.60 ± 5.45	93.14 ± 7.36	*97.05* *± 3.47*	95.07 ± 4.72	*97.07* *± 2.52*	*97.55* *± 2.59*	*97.05* *± 3.47*
CSE	90.19 ± 3.16	89.67 ± 7.45	82.33 ± 7.13	*93.07* *± 4.19*	90.67 ± 5.82	89.19 ± 5.54	87.74 ± 5.61
mRMR	96.10 ± 4.52	91.14 ± 7.25	96.10 ± 4.52	94.71 ± 5.26	96.57 ± 3.34	96.55 ± 3.35	96.55 ± 4.73
FCBF	95.57 ± 5.45	90.14 ± 6.20	*97.07* *± 5.20*	94.60 ± 4.93	*97.57* *± 2.56*	*97.07* *± 2.52*	96.07 ± 4.58
WA	98.05 ± 4.12	94.60 ± 3.56	97.02 ± 2.58	*99.50* *± 1.58*	*99.02* *± 2.06*	97.07 ± 3.45	98.05 ± 2.52
CSE-mRMR	93.57 ± 5.32	93.60 ± 3.35	90.14 ± 6.20	92.14 ± 4.60	94.12 ± 4.45	94.12 ± 4.45	92.64 ± 4.62
90% of attributes	79.45 ±10.50	*84.21* *± 8.10*	*87.69* *± 2.55*	75.86 ± 3.68	77.90 ± 5.40	75.40 ± 3.66	75.40 ± 3.73
95% of attributes	74.98 ±10.53	*84.71* *± 7.97*	*79.31* *± 3.10*	73.38 ± 4.27	73.93 ± 3.68	72.95 ± 4.39	72.98 ± 4.23
99% of attributes	69.55 ± 7.57	*84.71* *± 8.95*	*72.95* *± 4.45*	69.95 ± 3.58	68.50 ± 3.88	68.50 ± 2.00	68.50 ± 2.00

The best-performing algorithms were SVM, 3-NN, 5-NN, and 7-NN for the CFS measure and 1-NN and 3-NN for the CSE measure. The 3-NN, 5-NN, and 7-NN algorithms denoted higher values than the base algorithm for the mRMR, while the SVM, 3-NN, 5- NN, and 7-NN for the FCBF.

For the WA, the 1-NN and 3-NN algorithms produced higher values than the NB algorithm, and regarding the CSE-mRMR, the J48, 3-NN, and 5-NN algorithms showed higher values in relation to the base algorithm.

Regarding the PCA method, the J48 and SVM algorithms showed higher performances in the three subsets (90%, 95% and 99%) compared with the NB algorithm. The highest average performance was obtained with a subset comprising 90% of the attributes, with a hit rate of 79.42%.

The data above indicated that the database containing all of the attributes obtained an average hit rate of 89.78%. The AS method and the PCA method obtained hit rates of 94.40% and 75.80%, respectively. Therefore, the AS method appeared to be the optimal method among the databases analyzed. Lower values were obtained with the PCA method, but in most cases, a good hit rate was obtained.

Finally, our experiments corroborate with previous results, such as in the studies of Borges and Nievola [9] and Macedo *et al.* [14], who compared the random projection and DRM-F methods with AS. AS method presented better results, an average hit rate of 90%, which was similar to that found in the present study demonstrating the applicability and effectiveness of this analysis method for these types of data. PCA showed good results in some experiments, however according to Bartenhagen *et al.* [44], due PCA just considers the variance and don't use further information of the data like class labels, in some cases there are nonconformities, since the first principal components are not always sufficient in interpreting the original base, which makes the results difficult to explain.

With regards to the proposed combination, we found that the values designed by the CSE-mRMR measure were highly satisfactory and that, for some algorithms, the values were higher. In general, the measure showed a higher average hit rate than CSE and than all percentage of variance considered in the PCA method for the LungCancer-Michigan database. Regarding the LungCancer-Ontario database, the CSE-mRMR presented a higher average hit rate to the results using all attributes, mRMR, FCBF, and the subsets that correspond to the PCA method. For the LungCancer-Harvard database, the CSE-mRMR presented a higher average hit rate than the results using all attributes, CSE and subsets of the PCA method.

We also compared CSE-mRMR with combinations proposed by other studies, such as mRMR-GA [18], mRMR-ABC [16], and mRMR-HFS [19]. Regarding the classification accuracy, the CSE-mRMR presented 86.32% on average, considering the three databases and seven algorithms (previously described), mRMR-GA presented 94.09% on average (five databases and two algorithms: NB and SVM), mRMR-ABC presented 99.46% (six databases and one algorithm: SVM), and mRMR-HFS presented 84.00% on average (nine databases and three algorithms: C4.5, NB and SVM). This data further supports the idea that combining other measures with mRMR is valid for biomedical data. Therefore, we concluded that the proposed combination, CSE-mRMR, can be used for gene expression data sets, since it improved the classification accuracy and obtained an average hit rate of over 80%, which makes a measure viable for the theme in question.

VI. CONCLUSION

In this paper, two reduction methods (AS and PCA) were applied with the purpose of comparing their performances in the selected databases. The methods were applied to gene expression data sets, which presented difficulties for data processing, due to the expressive number of genes and little information related to the samples. Furthermore, we have introduced a combination of CSE and mRMR measures, CSE-mRMR.

Significant improvement opportunities were found. By employing the two approaches of AS (filter and wrapper), it was possible to claim that the classification accuracy was significantly improved when compared to the ones obtained in all the attributes of three databases investigated. In addition, there was a huge decline in the number of attributes when employing this method. We found that, although it demanded high computing time, the WA produced the best success rate. The CSE-mRMR presented excellent classification performance and presented better results than some traditional filters.

The best results were found in PCA method experiences with 90% of the attributes from the original database, that is, with a variance percentage lower than the other analyzed experiences. The PCA method presented hit rates of over 80% in some of the cases, thereby becoming a viable reduction method.

The algorithms that performed best were NB, J48, 3-NN, and 5-NN.

Thus, it is recommended that the AS method is applied in gene expression data analysis due to its presentation of consistent results for this type of domain. PCA can be used as an alternative method, as it can provide better results than the original set of attributes.

Therefore, in addition to comparing the methods described, the main contribution of the paper was to show the paths and alternatives that researchers can take when using some of these methods through the algorithms used, thus providing better knowledge about the data, and also to provide a new filter to form subsets and improve the classification rate of microarray data.

Future works are encouraged to compare other DR methods for the same classifier algorithms, or selecting other algorithms or even utilizing gene expression data sets related to other diseases. This large set of experiments aims at facilitating future comparative studies when a researcher proposes a new method. It is in our interest to apply the CSE-mRMR to other fields to verify its applicability.

REFERENCES

[1] T. Nguyen, A. Khosravi, D. Creighton, and S. Nahavandi, "Hierarchical gene selection and genetic fuzzy system for cancer microarray data classification," *PLoS ONE*, vol. 10, no. 3, 2015, Art. no. e0120364.

[2] S. Tabakhi, A. Najafi, R. Ranjbar, and P. Moradi, "Gene selection for microarray data classification using a novel ant colony optimization," *Neurocomputing*, vol. 168, pp. 1024–1036, Nov. 2015.

[3] C. Chira, J. Sedano, J. R. Villar, M. Camara, and C. Prieto, "Gene clustering for time-series microarray with production outputs," *Soft Comput.*, vol. 20, no. 11, pp. 4301–4312, Nov. 2016.

[4] C.-M. Lai, W.-C. Yeh, and C.-Y. Chang, "Gene selection using information gain and improved simplified swarm optimization," *Neurocomputing*, vol. 218, pp. 331–338, Dec. 2016.

[5] Y. Chen, Z. Zhang, J. Zheng, Y. Ma, and Y. Xue, "Gene selection for tumor classification using neighborhood rough sets and entropy measures," *J. Biomed. Inform.*, vol. 67, pp. 59–68, Mar. 2017.

[6] F. Archetti, I. Giordani, and L. Vanneschi, "Genetic programming for anticancer therapeutic response prediction using the NCI-60 dataset," *Comput. Oper. Res.*, vol. 37, no. 8, pp. 1395–1405, 2010.

[7] Q. Lu and X. Qiao, "Sparse Fisher's linear discriminant analysis for partially labeled data," *Stat. Anal. Data Mining*, vol. 11, no. 1, pp. 17–31, 2018.

[8] L. B. Romdhane, N. Fadhel, and B. Ayeb, "An efficient approach for building customer profiles from business data," *Expert Syst. Appl.*, vol. 37, no. 2, pp. 1573–1585, 2010.

[9] H. B. Borges and J. C. Nievola, "Comparing the dimensionality reduction methods in gene expression databases," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 10780–10795, 2012.

[10] S. Liang, A. J. Ma, S. Yang, Y. Wang, and Q. Ma, "A review of matched-pairs feature selection methods for gene expression data analysis," *Comput. Struct. Biotechnol. J.*, vol. 16, pp. 88–97, Feb. 2018.

[11] K. C. Tan, E. J. Teoh, Q. Yu, and K. C. Goh, "A hybrid evolutionary algorithm for attribute selection in data mining," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8616–8630, 2009.

[12] F. Artoni, A. Delorme, and S. Makeig, "Applying dimension reduction to EEG data by principal component analysis reduces the quality of its subsequent independent component decomposition," *NeuroImage*, vol. 175, pp. 176–187, Jul. 2018.

[13] I. T. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Springer-Verlag, 2002.

[14] D. C. Macedo, E. C. M. Ishikawa, C. B. Santos, S. N. Matos, H. B. Borges, and A. C. Francisco, "Proposed method for dimensionality reduction based on framework in gene expression domain," *Genet. Mol. Res.*, vol. 13, no. 4, pp. 10582–10591, 2014.

[15] A. Balodi, M. L. Dewal, R. S. Anand, and A. Rawat, "Texture based classification of the severity of mitral regurgitation," *Comput. Biol. Med.*, vol. 73, pp. 157–164, Jun. 2016.

[16] H. Alshamlan, G. Badr, and Y. Alohali, "mRMR-ABC: A hybrid gene selection algorithm for cancer classification using microarray gene expression profiling," *BioMed. Res. Int.*, vol. 2015, pp. 1–15, Mar. 2015.

[17] X. Yan and M. Jia, "Intelligent fault diagnosis of rotating machinery using improved multiscale dispersion entropy and mRMR feature selection," *Knowl.-Based Syst.*, vol. 163, pp. 450–471, Jan. 2019.

[18] A. E. Akadi, A. Amine, A. E. Ouardighi, and D. Aboutajdine, "A two-stage gene selection scheme utilizing MRMR filter and GA wrapper," *Knowl. Inf. Syst.*, vol. 26, no. 3, pp. 487–500, 2011.

[19] M. K. Ebrahimpour and M. Eftekhari, "Ensemble of feature selection methods: A hesitant fuzzy sets approach," *Appl. Soft Comput.*, vol. 50, pp. 300–312, Jan. 2017.

[20] J. T. Souza, "Methods of attribute selection and principal component analysis: A comparative study," Ph.D dissertation, Federal Technol. Univ.-Parana, Ponta Grossa, Brazil, 2017, p. 73. [Online]. Available: http://repositorio.utfpr.edu.br/jspui/handle/1/2387

[21] A. Onan, "A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer," *Expert Syst. Appl.*, vol. 42, no. 20, pp. 6844–6852, 2015.

[22] M. Mursalin, Y. Zhang, Y. Chen, and N. V. Chawla, "Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier," *Neurocomputing*, vol. 241, pp. 204–214, Jun. 2017.

[23] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Adv. Bioinf.*, vol. 2015, May 2015, Art. no. 198363. [Online]. Available: https://www.hindawi.com/journals/abi/2015/198363/

[24] Y. Ma, Y. Ding, and T. Zheng, "Feature subspace learning based on local point processes patterns," *Stat. Anal. Data Mining*, vol. 11, no. 1, pp. 32–50, 2018.

[25] J. Wang, L. Wu, J. Kong, Y. Li, and B. Zhang, "Maximum weight and minimum redundancy: A novel framework for feature subset selection," *Pattern Recognit.*, vol. 46, no. 6, pp. 1616–1627, 2013.

[26] A. Alkuhlani, M. Nassef, and I. Farag, "Multistage feature selection approach for high-dimensional cancer data," *Soft Comput.*, vol. 21, no. 22, pp. 6895–6906, 2017.

[27] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. San Mateo, CA, USA: Morgan Kaufmann, 2016.

[28] H. M. Alshamlan, "Co-ABC: Correlation artificial bee colony algorithm for biomarker gene discovery using gene expression profile," *Saudi J. Biol. Sci.*, vol. 25, no. 5, pp. 895–903, 2018.

[29] A. Wosiak and D. Zakrzewska, "Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis," *Complexity*, vol. 2018, Oct. 2018, Art. no. 2520706. [Online]. Available: https://www.hindawi.com/journals/complexity/2018/2520706/

[30] M. A. Hall and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 6, pp. 1437–1447, Nov./Dec. 2003.

[31] B. Rana *et al.*, "Relevant 3D local binary pattern based features from fused feature descriptor for differential diagnosis of Parkinson's disease using structural MRI," *Biomed. Signal Process. Control*, vol. 34, pp. 134–143, Apr. 2017.

[32] Y. Jiang and C. Li, "mRMR-based feature selection for classification of cotton foreign matter using hyperspectral imaging," *Comput. Electron. Agricult.*, vol. 119, pp. 191–200, Nov. 2015.

[33] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 856–863.

[34] O. Hadjerci, A. Hafiane, D. Conte, P. Makris, P. Vieyres, and A. Delbos, "Computer-aided detection system for nerve identification using ultrasound images: A comparative study," *Inform. Med. Unlocked*, vol. 3, pp. 29–43, Jan. 2016.

[35] S. S. Kannan and N. Ramaraj, "A novel hybrid feature selection via symmetrical uncertainty ranking based local memetic search algorithm," *Knowl.-Based Syst.*, vol. 23, no. 6, pp. 580–585, Aug. 2010.

[36] R. Taormina, S. Galelli, G. Karakaya, and S. D. Ahipasaoglu, "An information theoretic approach to select alternate subsets of predictors for data-driven hydrological models," *J. Hydrol.*, vol. 542, pp. 18–34, Nov. 2016.

[37] P. Bermejo, L. de la Ossa, J. A. Gámez, and J. M. Puerta, "Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking," *Knowl.-Based Syst.*, vol. 25, no. 1, pp. 35–44, 2012.

[38] X.-Y. Liu, Y. Liang, S. Wang, Z.-Y. Yang, and H.-S. Ye, "A hybrid genetic algorithm with wrapper-embedded approaches for feature selection," *IEEE Access*, vol. 6, pp. 22863–22874, 2018.

[39] X. Xu and X. Wang, "An adaptive network intrusion detection method-based on PCA and support vector machines," in *Proc. Int. Conf. Adv. Data Mining Appl.* Berlin, Germany: Springer, 2005, pp. 696–703.

[40] K. Ridge. (2018). *Kent Ridge Bio-medical Dataset*. [Online]. Available: http://leo.ugr.es/elvira/DBCRepository/

[41] D. G. Beer *et al.*, "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature Med.*, vol. 8, no. 8, pp. 816–824, Aug. 2002.

[42] D. A. Wigle *et al.*, "Molecular profiling of non-small cell lung cancer and correlation with disease-free survival," *Cancer Res.*, vol. 62, no. 11, pp. 3005–3008, Jun. 2002.

[43] A. Bhattacharjee *et al.*, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 24, pp. 13790–13795, 2001.

[44] C. Bartenhagen, H.-U. Klein, C. Ruckert, X. Jiang, and M. Dugas, "Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data," *Bioinformatics*, vol. 11, p. 567, Dec. 2010.

JOVANI TAVEIRA DE SOUZA received the master's degree in industrial engineering from the Federal University of Technology (UTFPR-PG), Brazil, where he is currently pursuing the Ph.D. degree in industrial engineering. He is a specialist in industrial engineering (UTFPR-PG), with a degree in mathematics from the State University of Ponta Grossa (UEPG). His areas of research and interests include data mining, knowledge management, and sustainability. He is a member of the Sustainable Production Systems Laboratory (LESP/UTFPR-PG).

ANTONIO CARLOS DE FRANCISCO is received the Ph.D. degree in industrial engineering from the Federal University of Santa Catarina (UFSC), Brazil. He is a Professor with the Federal University of Technology (UTFPR-PG), where he is also a Coordinator of the Industrial Engineering Postgraduation Program (Ponta Grossa Campus), and a Coordinator of the Sustainable Production Systems Laboratory (LESP/UTFPR-PG). His areas of research and interests include sustainable systems, life cycle thinking, circular economy, quality of life and quality of life at work, and data mining.

DAYANA CARLA DE MACEDO received the master's degree in industrial engineering and the Ph.D. degree in industrial engineering from the Federal University of Technology (UTFPR-PG), Brazil. She is a Specialist in industrial engineering (UTFPR-PG), with a degree in food technology from UTFPR-PG and the degree in administration from the State University of Ponta Grossa (UEPG). She is also a Professor with the Midwest University of Paraná, Brazil.