Contents lists available at ScienceDirect

Decision Analytics Journal

journal homepage: www.elsevier.com/locate/dajour

A new univariate feature selection algorithm based on the best–worst multi-attribute decision-making method

Dharyll Prince M. Abellana *, Demelo M. Lao

Department of Computer Science, College of Science, University of the Philippines Cebu, 6000 Cebu, Philippines

ARTICLE INFO

Keywords:
Filter-based feature selection
Best–worst method
Feature selection framework
Multi-attribute decision-making
Three-way analysis of variance

ABSTRACT

With the extensive applicability of machine learning classification algorithms to a wide spectrum of domains, feature selection (FS) becomes a relevant data preprocessing technique due to the high dimensionality of data used in these domains. While efforts have been made to study various filters for ranking features, scholars have paid little attention to developing a unified framework that can be used as an interface for any filter. The development of such a framework would formalize the understanding of filter-based FS. This helps put scholars in the same perspective when analyzing new FS algorithms. This study proposes a new filter-based FS framework based on the best–worst multi-attribute decision-making method. The proposed algorithm is compared to two control groups: (a) no FS and (b) randomized algorithm. Furthermore, two blocking variables are considered: (i) classifier and (ii) training dataset. The performance of the classifiers was measured using the area under the curve (AUC) of the receiver operating characteristics (ROC) curve. A three-way analysis of variance (ANOVA) is used to compare the proposed approach to the control groups considering the blocking variables. This paper offers several contributions to the literature. For one thing, it is one of the few works that put forward a framework for performing filter-based FS. To the best of the authors' knowledge, the study is the first to provide empirical evidence about the interaction between the factors considered in the literature for evaluating FS algorithms.

1. Introduction

Machine learning (ML) has been widely used in a variety of disciplines because of its utility in prediction, knowledge discovery, and modeling, among other things. Classification is one of the most notable uses of ML among the numerous tasks that could possibly be performed. It has been used in fields such as remote sensing, biomedical informatics, sports result prediction, machine fault diagnosis, power monitoring, crop categorization, email spam filtering, and seismic event classification [1]. Despite their utility, classification algorithms have various limitations in practice, particularly when complex data (e.g., unstructured data and streaming data) are involved [1]. For starters, most real-world data have a large number of features. Similarly, noise, irrelevant features, redundant features, and anomalies in the data can exist. This complexity in the dataset results in high dimensionality and sparseness, which are computationally expensive for classification algorithms [2,3]. In light of this, scholars in the literature have proposed dimensionality reduction and data preprocessing strategies.

Feature selection (FS) is a strategy for discovering subsets of the original feature set that efficiently represent the input data while decreasing the impacts of noise and irrelevant features but still providing relatively excellent results for the task [3,4]. FS has been demonstrated to increase classification performance in a variety of applications. For instance, [5] used FS for high-dimensional cancer microarray datasets.The paper by [6] used FS for Arabic text classification. The paper by [7] used FS for educational data mining. Moreover, [8] employed FS to predict heart disease using a bio-inspired optimization technique. So far, the usage of FS has been essential in applications, particularly when decision-makers lack a thorough understanding of the features' relevance in certain domains [9]. Based on how the algorithms accomplish the selection process, the literature on FS can be classified into three major areas: (i) filter, (ii) wrapper, and (iii) embedded methods [10]. Because of their computational efficiency, filter techniques have received the most attention of the three approaches [11]. For one thing, filter approaches are independent of classifiers. They are computationally efficient to implement since they are independent of the type of classifier utilized. As a result, they are well suited to challenges involving a high number of features, such as bioinformatics, natural language processing, and big data.

Filter approaches evaluate the usefulness of features by assessing data inherent properties such as distance, consistency, and correlation [11]. They have been used for data preprocessing in large datasets

* Corresponding author.
E-mail addresses: dmabellana@up.edu.ph (D.P.M. Abellana), dmlao1@up.edu.ph (D.M. Lao).

https://doi.org/10.1016/j.dajour.2023.100240
Received 21 September 2022; Received in revised form 22 April 2023; Accepted 24 April 2023
Available online 4 May 2023

due to their effectiveness in reducing dimensionality while retaining significant information. While scholars have put forward various works for filter-based FS, most of these efforts are focused on exploring measures for ranking features. However, very few scholars have paid attention to the development of general frameworks for performing filter-based FS. A general framework for filter-based FS would offer several advantages for the data mining literature. For one thing, it would enable scholars to understand filter-based FS in a holistic manner. In turn, the development of ranking measures and strategies can be tackled in a systematic and rigorous manner. Furthermore, a general framework would enable a formal treatment of feature-based FS, which is relatively under-explored in the current literature. Multi-criteria decision making (MCDM) is an active research area in soft computing focused on studying algorithms and mathematical models for making the best decisions. These models and algorithms have been widely applied in many real world domains such as sustainable supply chain management [12], sustainable mining [13], hospitality management [14], and technology management [15], among others.

Most applications of MCDM are focused on selection and prioritization problems. In light of this, there are several reasons to infer about the applicability of these algorithms to FS. For instance, these algorithms share a common goal with FS, which is ranking attributes according to some preference function. A significant advantage for adopting MCDM in FS is the availability of well-established frameworks. Therefore, viewing filter-based FS through the lenses of MCDM directly offers a formal representation that can be used as a framework for carrying out filter-based FS. Scholars in the literature, such as [16,17] substantiated this idea. This study aims to investigate the effectiveness of MCDM as a framework for filter-based FS. In particular, the study develops an FS algorithm based on the best–worst method (BWM) of [18]. The paper by [18] found that pairwise comparison-based MCDM approaches are prone to inconsistencies when dealing with large number of criteria. As such, [18] proposed the BWM to overcome such issues. Thus, the BWM was adopted in this paper over other MCDM approaches due to its promising scalability when dealing with large datasets, which would often be the case in FS problems.

The study offers several contributions to the literature. First, it is one of the very few works that puts forward a framework for performing filter-based FS. In fact, to the best of the authors' knowledge, this study is the first to investigate the use of MCDM in univariate FS. As such, it extends the applicability of MCDM to a novel application. The papers by [16,17] explored on a similar problem, but are focused on ensemble feature selection. Therefore, this study would complement the findings revealed by such works. In turn, the study would significantly contribute to the advancement of both the MCDM and data mining literature. This could provide an avenue for the cross-fertilization of these fields. Second, to the best of the authors' knowledge, this is the first paper to provide evidence of a significant interaction between the factors used for testing FS algorithms. Third, the paper is one of the very few papers in the current literature to adopt significance testing for assessing the proposed FS algorithm while majority of the papers adopt a descriptive approach. Finally, to the best of the authors' knowledge, the paper is the first to simultaneously include the *training dataset* and *classifiers* as blocking factors in comparing feature selection algorithms.

2. Literature review

2.1. Trends in machine learning and current challenges

ML is considered a subset of artificial intelligence (AI) concerned with building mathematical models based on data to generate predictions or decisions without being explicitly programmed to do so. With the extensive use of ML in the current literature, some drawbacks have also been pointed out by scholars in different fields. For instance, [19] pointed out that the processing of large-scale data produced by cyber–physical systems, internet of things (IoT), and other digital technologies, has been one of the greatest challenges in the integration of ML-based intelligent systems. Several scholars have also pointed out the issues of cost, algorithmic complexity, and high dimensionality of data for many ML applications. For instance, [4] claimed that high dimensionality (e.g., sheer volume, and unbounded length and patterns) and computationally expensive training of ML algorithms pose huge challenges for insider threat detection. Similarly, [20] pointed out that along with the increasing use of AI (e.g., ML, expert systems, natural language processing (NLP)) in global health studies, large datasets and the opaque nature of several algorithms are some of the major challenges faced by scholars in the current literature. Amidst such challenges, several solutions have been proposed in literature such as feature selection, and dimensionality reduction, among others [21,22]. One of the ways in which ML algorithms are categorized in literature is through how much human intervention is needed for training [23]. Due to having a relatively straightforward training and interpretation, classification problems are one of the most widely studied topics both in theory and application [24]. For brevity, this review will focus only on the application of FS in classification problems. While classification has played a significant role in modeling and prediction, current literature is posed with issues that undermine its usefulness in large-scale applications. In particular, most real-world applications involve a huge number of features, which contribute to the high dimensionality of data [2,3]. For instance, due to the high dimensionality, presence of noise, and huge size of text documents available on the web, text classification becomes computationally expensive [3]. Likewise, in DNA analysis, classification has posed a serious challenge due to the involvement of high dimensional micro-array data (usually involving hundreds of features) [25].

2.2. Trends in univariate filter-based feature selection

The scalability and computational efficiency of filter-based FS are one of the many facets that sets it apart from other FS methods. While filter-based FS has been shown to be significant when dealing with large datasets in real-world problems, there is not a one-size-fits-all approach for doing it [26]. In the literature, filter-based FS approaches can be generally divided into: (i) univariate filter-based approaches and (ii) multivariate filter-based approaches [27]. These approaches are differentiated based on the following factors: (a) number of filters employed and (b) selection strategy [28,29]. According to [28], univariate approaches rank features individually using some performance measures where the final feature subset can be determined by establishing a threshold value or specifying the number of features to retain. For example, [30] used a single filter (i.e., one-way analysis of variance (ANOVA)) to exclude nuisance features in analyzing network attacks. Furthermore, the selection strategy used a threshold value (i.e., significance of the features) to select the feature subset. On the contrary, multivariate approaches evaluate all feature subsets using a particular search strategy and some performance measures. The selected feature subset is the best performing among all other feature subsets. For instance, [29] emphasized the use of metaheuristics and nature-inspired algorithms for solving multivariate filter-based FS to find the best feature subsets while avoiding the intractability of combinatorial optimization problems. While it would appear that multivariate approaches are better than univariate approaches, there are some cases when univariate approaches are preferable. For one thing, some problems require looking at only a single performance measure. For example, judging the relevance of features based on their statistical significance (e.g., [30]). For another thing, univariate filters are more computationally efficient compared to multivariate filters. For these reasons, univariate filters have been applied in several circumstances. For example, [31] investigated the applicability of univariate filter FS in heart disease prediction. Similarly, [32] developed a filter algorithm based on distance correlation and applied it to various synthetic and benchmark data. Some of the common filters used in the literature include information gain, chi-square, and mutual information,

Table 1
Common univariate filter measures.

Name	Mathematical form	Description
Information gain	$IG(X\|Y) = H(X) - H(X\|Y)$	Given the variables X and Y, $H(X)$ is the entropy of X and $H(X\|Y)$ is the entropy of X given Y. See [36] for details on entropy.
Symmetrical Uncertainty	$SU(X,Y) = 2 \times \left[\frac{IG(X\|Y)}{H(X)+H(Y)}\right]$	For a pair of variables X and Y, $IG(X\|Y)$ is the information gain of X given Y, $H(X)$ and $H(Y)$ is the entropy of X and Y, respectively.
Chi-square	$\chi^2(t_k, c_i) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)}$	For the kth feature t_k and ith class c_i, A is the number of records in c_i that contain the feature t_k, B is the number of records that contain the feature t_k in other classes, C is the number of records in class c_i that do not contain the feature t_k, D is the number of records that do not contain the feature t_k in other classes, and N is the total number of records in the dataset.
Mutual information	$MI(t_k, c_i) = \frac{\log(P(t_k,c_i))}{P(t_k)\times P(c_i)}$	Feature t_k and class c_i, $P(t_k, c_i)$ is the joint probability of t_k and c_i, $P(t_k)$ is the marginal probability of t_k, and $P(c_i)$ is the marginal probability of c_i.

to name a few. These filters are presented in Table 1. The univariate filter approaches continue to be relevant since their development and new algorithms have been proposed based on these methods recently (e.g., [33,34], and [35]). Although filter approaches have the advantage of being computationally efficient, their ranking is too simple. While this is desirable from a computational perspective, this has several pitfalls from an analytical perspective [16]. For instance, the generated rankings could not provide one with insights to decide the number of features to retain. As such, deciding the threshold value or specifying the number of features to retain becomes too arbitrary. Thus, there is a need to look into more robust ranking or selection strategies that would serve as a framework for filter-based FS [16].

2.3. Multi-criteria decision-making as a promising feature selection framework

One of the limitations of conventional univariate filter-based FS is the lack of robust ranking or selection strategies [16]. Scholars in relevant fields have been addressing this drawback through the lenses of multi-criteria decision-making (MCDM). In fact, [16] pointed out that since FS is essentially about selecting the best features that characterize a dataset, it can be natural to view it through the lenses of MCDM. To this end, new developments have emerged on the use of MCDM for FS. For example, [16] applied the Evaluation Based on Distance from Average Solution (EDAS) method to aggregate various filters, such as chi-square, symmetric uncertainty (SU), and Relief-F. Furthermore, they found that EDAS was effective in reducing the number of features without significantly affecting the performance of the considered classifiers. Similarly, [17] used the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) to construct ensemble feature rankings in text categorization. While few works have begun investigating the applicability of MCDM to FS, these works focused on ensemble FS. There have been no works yet that apply MCDM to univariate FS despite being equally applicable (and invaluable) to it. As such, demonstrating the applicability of MCDM to univariate filter-based FS would help fill this gap in the existing literature. With this, the contribution of this paper as compared to existing works in the literature is presented in Table 2. Numerous applications of MCDM to various problems have been performed in the literature. For instance, [41] used the fuzzy Dombi EDAS model to evaluate the metaverse integration of freight fluidity measurement alternatives. The paper by [42] used a novel MCDM method based on wins in league for the prioritization of environmental deterioration strategies. Similarly, [43] used MCDM for developing a collaborative COVID-19 surge management strategy. Finally, [44] used the analytic hierarchy process as an assessment method for offshore wind resource. The availability of various MCDM methods (and their applications) could potentially provide a point of entry for applying MCDM to FS. In spite of that, [18] pointed out that MCDM methods that use a pairwise comparison approach would be prone to inconsistencies as the number of alternatives or criteria increase. Since FS is used for reducing large number of features, this could pose a problem. As such, an MCDM method that gets away with this problem would be desirable. The paper by [18] proposed the BWM, which gets away with the pairwise comparison approach. It has been applied to numerous problem domains due to its promising results. For example, [12] used the BWM for the evaluation of sustainable supply chain in the food industry. Similarly, [45] used the BWM for the prioritization of the critical success factors of sustainable lean six sigma in healthcare organizations. An investigation of the application of BWM to univariate filter-based FS could pave the way to the application of MCDM to the broader FS domain.

3. Methodology

3.1. Overview

This study proposes a univariate filter-based FS algorithm based on the BWM. In order to evaluate the performance of the proposed algorithm, it is tested using benchmark datasets for FS. A literature search was performed to determine the most common benchmark datasets used by scholars in the literature. As a result, this paper uses seven real world datasets obtained from public repositories, such as the University of California Irvine Machine Learning Repository, and Kaggle. Each dataset is selected according to : (i) number of features, (ii) number of instances, and (iii) number of class labels. These criteria are determined using a literature analysis. The datasets which have between five and 100 features are considered in order to provide breadth in the number of features that will be processed by the proposed FS algorithm. The number of instances are varied between 100 to 5000 instances for the same reason as with the number of features. Furthermore, only datasets with two classes are considered in order to satisfy the number of class labels in binary classification problems. The descriptions of these datasets are presented in Table 3. These datasets were trained on the classification algorithms presented in Section 3.2. To provide an overview of how the proposed FS algorithm works, a step-by-step procedure of the BWM is presented in Section 3.3. Afterward, the proposed FS algorithm is discussed in detail in Section 3.4. Finally, the experimental design, which presents how the performance of the proposed FS will be analyzed is presented in Section 3.5. The overall framework for testing the proposed FS algorithm is presented in Fig. 1.

3.2. Classification algorithms

3.2.1. K-nearest neighbors (k-NN)

The k-NN algorithm is a nonparametric method proposed by [52]. It finds k-nearest neighbors of the query from the training set and assigns a class label to the query through the majority voting rule [53]. Let $T = \{(y_i, c_i)\}_{i=1}^{N}$ denote a training set with N instances from M classes, where $y_i \in R^d$ is a training instance in a D-dimensional space, $c_i \in C$ is the corresponding class label of y_i for $C = \{c_1, c_2, \ldots, c_M\}$, and $(x, \tilde{c}) \notin T$ is an instance that is not in the training set where $\tilde{c}$ is an unknown class label. The pseudocode of the k-NN algorithm is given in Algorithm 1 (see [54]).

Table 2
Comparison of the proposed approach (*item no. 7*) to recent univariate filter-based feature selection studies.

Item No.	Selection strategy	Filter	Performance analysis	Author	Year
1	Relevance of features are judged based on the statistical significance of a feature	F-statistic	Descriptive (accuracy, precision, and detection rate)	[30]	2023
2	Features are ranked using a filter-relationship gain score	Distance correlation	Descriptive (average RMSE)	[32]	2022
3	Multiple univariate filters (i.e., but not as an ensemble) are used to select the best ranking features	Pearson r, cosine similarity, euclidean distance, and mutual information	One-way ANOVA	[37]	2022
4	Multiple univariate filters (i.e., but not as an ensemble) are used to select the best ranking features	Symmetric uncertainty, correlation, and unsupervised clustering	Descriptive (accuracy, TP rate, FP rate, precision, recall, F-measure, and ROC-AUC)	[38]	2021
5	Multiple univariate filters (i.e., but not as an ensemble) are used to select the best ranking features	Correlation-based filters (Pearson r, Spearman ρ, Kendall τ), Chi-square statistic, F-statistic, univariate ROC-AUC	Descriptive (precision, recall, F1-score, and accuracy)	[39]	2022
6	Progressive Ablation Feature Selection (PAFS) method with XGBoost	Correlation coefficient and Gain value	Descriptive (ROC-AUC)	[40]	2022
7	An MCDM method (i.e., BWM) is used to compare the strength of associations between each feature. The features are ranked based on the optimal weights generated using the BWM	Kendall τ, point biserial correlation, and Cramer's V	Three-way ANOVA	This paper	This paper

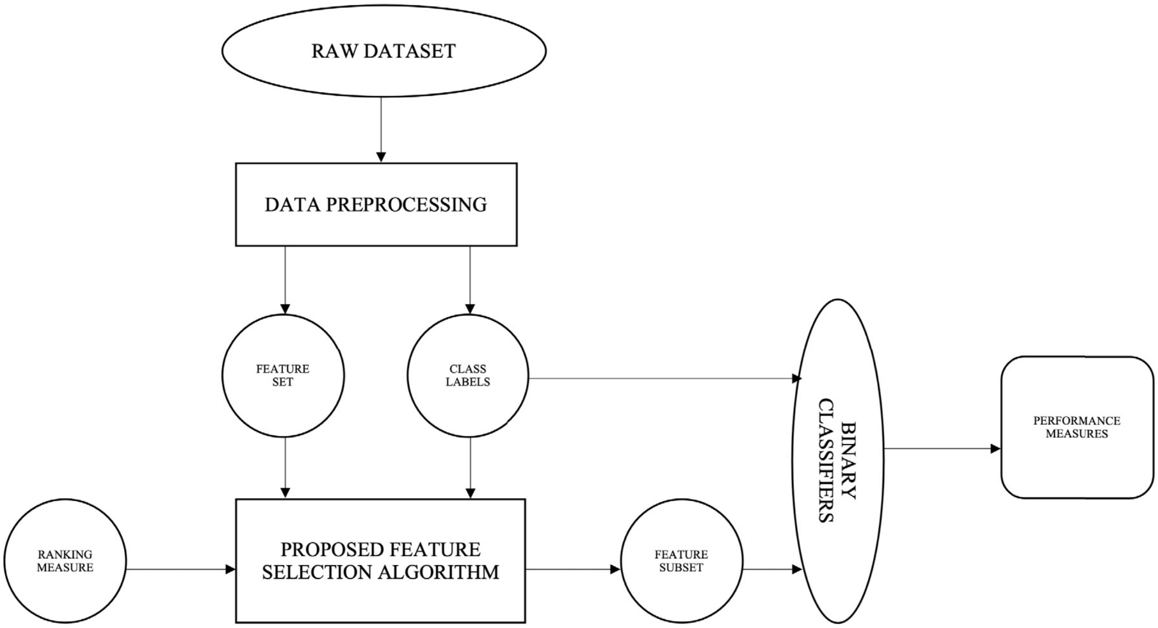


Fig. 1. Research framework. The circular nodes represent inputs and outputs during the algorithm's execution. The rectangular nodes represent a process for transforming inputs to outputs. The square nodes with curved edges represent data that will be used as response variable for the empirical evaluation in Section 3.5. The oval nodes represent factors that will be used as blocking variables in the empirical validation in Section 3.5.

3.2.2. *Decision trees*

The decision tree algorithm uses a tree-like structure, where each internal node represents a test on an attribute, each branch denotes the outcome of the attribute test and each leaf node denotes the class label [55]. The tree performs classification by splitting the branches of the tree, where each split represents a test on the data attribute [55]. The pseudocode in Algorithm 2 uses the C4.5 variant of the decision tree algorithm developed by [56].

3.2.3. *Logistic regression*

Logistic regression is a classification algorithm suitable for binary classification [57]. Using this model, the probability that an event occurs is expressed as a linear function of the input variables [57]. It

Table 3
Dataset description. Each dataset is obtained from public databases.

No.	Name	No. of features	No. of instances	Source
1	QSAR Biodegradation Data Set	41	1055	[46,47]
2	Pima Indians Diabetes Database	8	768	[47–49]
3	Heart Attack Prediction	13	303	[50]
4	Hepatitis Data Set	19	155	[47]
5	Ionosphere Data Set	35	351	[47]
6	Connectionist Bench (Sonar, Mines vs. Rocks) Data Set	60	208	[47]
7	Wine Quality Data Set	11	4898	[47,51]

Algorithm 1: Pseudocode of the classical k-Nearest Neighbor (k-NN) Classification Algorithm adopted from [54]

```
Data: $X$: the training data, $Y$: class labels of $X$, $k$: the number
      of nearest neighbors to consider, $x$ the test sample.
Result: $\tilde{c}$: the class label of $x$.
for $i = 1$ to $N$ do
    Calculate $d\left(y_i, x\right)$;
end
Compute Set $I$ which contains the minimum sets of $k$ distances
  $d\left(y_i, x\right)$;
$\tilde{c} \leftarrow$ majority label for $\{c_i | i \in I\}$
```

Algorithm 2: Very Fast C4.5 (VF-C4.5) algorithm for decision trees developed by [56].

```
Data: $A_{mn}$: matrix of continuous attributes
Result: $A_o$: the optimal attribute, $cp_o$: the optimal cut point
for $j = 1$ to $m$ do
    Calculate $mean_j$ and $median_j$ of $\left[A_{ij}\right]_{i=1}^{n}$;
    Find all potential cut points $ccp_{1j} = mean_j$, $ccp_{2j} = median_j$;
    for each cut point $cp_{ij}$ do
        Calculate information gain $Gain_{ccp_{ij}}$
    end
    Select the optimal cut point $cp_o$;
    Calculate splitting performance $Split\left(S, cp_{oj}\right)$;
    Calculate gain ratio $GainRatio\left(S, cp_{oj}\right)$
end
Select the optimal attribute $A_o$ and its cut point $cp_o$
```

is a particular case of the generalized linear models, where the link function is the logit or logistic function in Eq. (1) [58].

$$p_m(\vec{x}) = \frac{1}{1 + e^{-(\langle \vartheta, \vec{x} \rangle + b)}} \tag{1}$$

where p_m is the probability of being in class, c_m, $\vec{x}$ is the vector of features, ϑ is the vector of weights, and b is the bias. Let N be the number of training instances, M be the number of classes, d be the number of features, $C = \{c_m\}_{m=1}^{M}$ be the set of classes, $X = \{\vec{x}_i | \vec{x}_i \in R^d\}_{i=1}^{N}$ be the set of feature instances, $Y = \{c_i | c_i \in C\}_{i=1}^{N}$ be the set of class instances corresponding to X, and $T = \{\left(\vec{x}_i, c_i\right) | T \subset X \times Y\}_{i=1}^{N}$ be the set of training instances. The pseudocode of the logistic classification algorithm is presented in Algorithm 3 (see [59]).

3.2.4. Support vector machine (SVM)

The SVM [60] is an algorithm that classifies data by mapping the features into higher dimensions and divide the classes using a hyperplane while maximizing its distance with the classes. The task can be solved by minimizing an expression of the form

Algorithm 3: Pseudocode of a regularized logistic regression adopted from [59]

```
Data: Regularization parameter $\lambda > 0$
$m_i = 0$, $q_i = \lambda$. (Each weight $w_i$ has an independent prior
  $\mathcal{N}\left(m_i, q_i^{-1}\right)$);
for $t = 1$ to $T$ do
    Get a new batch of training data $\left(\vec{x}_j, y_j\right)$, $j = 1, \ldots, n.$;
    Find $\vec{w}$ as the minimizer of:
      $\frac{1}{2}\sum_{i=1}^{d} q_i \left(w_i - m_i\right)^2 + \sum_{j=1}^{n} \log\left(1 + \exp\left(-y_j \vec{w}^{\top} \vec{x}_j\right)\right)$;
    $m_i = w_i$;
    $q_i = q_i + \sum_{j=1}^{n} \sum_{j=1}^{n} x_{ij}^2 p_j \left(1 - p_j\right)$, $p_j = \left(1 + \exp\left(-\vec{w}^{\top} \vec{x}_j\right)\right)^{-1}$
end
```

$$\left[\frac{1}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i\left(\langle \vec{w}, \vec{x}_i \rangle - b\right)\right)\right] + \lambda \|\vec{w}\|^2 \tag{2}$$

The classical approach of implementing the SVM in Eq. (2) is by reducing it to a quadratic program in primal form as in Eq. (3).

$$\begin{aligned} &\min \frac{1}{n}\sum_{i=1}^{n} \zeta_i + \lambda \|\vec{w}\|^2 \\ &\text{subject to:} \\ &y_i\left(\langle \vec{w}, \vec{x}_i \rangle - b\right) \geq 1 - \zeta_i, \ \text{for all } i \\ &\zeta_i \geq 0, \ \text{for all } i \end{aligned} \tag{3}$$

Equivalently, Eq. (3) can be expressed in dual form as in Eq. (4)

$$\begin{aligned} &\max f\left(c_1, c_2, \ldots, c_n\right) = \sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \langle \vec{x}_i, \vec{x}_j \rangle y_i y_j \\ &\text{subject to:} \\ &\sum_{i=1}^{n} c_i y_i = 0, \ \text{for all } i \\ &0 \leq c_i \leq \frac{1}{2n\lambda}, \ \text{for all } i \end{aligned} \tag{4}$$

In other formulations, Eq. (4) is converted to a minimization problem by negating the objective function as in Eq. (5).

$$\begin{aligned} &\min f\left(c_1, c_2, \ldots, c_n\right) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \langle \vec{x}_i, \vec{x}_j \rangle y_i y_j - \sum_{i=1}^{n} c_i \\ &\text{subject to:} \\ &\sum_{i=1}^{n} c_i y_i = 0, \ \text{for all } i \\ &0 \leq c_i \leq \frac{1}{2n\lambda}, \ \text{for all } i \end{aligned} \tag{5}$$

Following from [61], a simple SVM classifier is implemented using the stochastic gradient descent (SGD) algorithm. Using SGD, the updating rule is

$$w^{(k+1)} = -\frac{1}{\lambda k}\sum_{j=1}^{k} v_j \ , \tag{6}$$

where v_j is a subgradient of the loss function at $w^{(j)}$ on the random example chosen at iteration j. Denoting $\theta^{(k)} = -\sum_{j<k} v_j$, $I = \{1, 2, ..M\}$, M is the number of training instances, and $T = \{\left(\vec{x}_i, y_i\right) | i \in I\}$ be the set of training instances (i.e., the training set), the pseudocode in Algorithm 4 is obtained.

3.2.5. Artificial neural network

An artificial neural network (ANN) is a computational system that imitates the organization of biological neurons and can learn from experimentally generated data or validated models [62]. ANNs are

Algorithm 4: Support Vector Machine (SVM) Algorithm implemented using Stochastic Gradient Descent (SGD). The pseudocode is adopted from [61].

```
Data: $T$: the training data, $K$: an arbitrary upper bound
Result: $\hat{w}$
$\theta^{(1)} \leftarrow \vec{0}$;
for $k = 1$ to $K$ do
    $\vec{w}^{(k)} := \frac{1}{\lambda k}\theta^{(k)}$;
    $i \leftarrow$ Choose uniformly at random from $I$;
    if $y_i\langle\vec{w}^{(k)}, \vec{x}_i\rangle < 1$ then
        $\theta^{(k+1)} := \theta^{(k)} + y_i\vec{x}_i$;
    else
        $\theta^{(k+1)} := \theta^{(k)}$;
    end
end
$\hat{w} := \frac{1}{K}\sum_{k=1}^{K}\vec{w}^{(k)}$;
```

modeled after biological neurons and operate on the principle that a network of simple processing units that are highly interconnected can acquire the ability to understand intricate connections between independent and dependent variables [62]. ANNs are a desirable method for modeling highly intricate, nonlinear systems that possess a vast number of inputs and outputs, utilizing parallel structures with extensive connections [62]. These networks typically have three layers, consisting of an input layer, a hidden layer, and an output layer [62]. The multilayer perceptron (MLP) is an example of a feedforward ANN that contains at least three layers, and it has found broad applicability in various classification problems [63]. The training algorithm for MLP is presented in Algorithm 5 (see [64]).

Algorithm 5: Pseudocode for training an MLP adopted from [64].

```
Data: $T$: the training data, $\xi$: stopping criterion
Result: $\theta$: trained parameters
Initialize the synaptic weights $\theta \leftarrow \theta^{[0]}$;
while $\xi$ is not satisfied do
    Forward Pass: Compute the input signals of the MLP by
      propagating forward from the input layer to the output
      layer of the MLP;
    Backward Pass: Compute the gradients of the weights $\theta$ in
      each layer using the chain rule;
    Adjust the weights $\theta$ using the gradients
end
```

3.3. Best–worst method

The BWM is an MCDM method proposed by [18]. This method was created to address the limitations, such as inconsistency, of MCDM techniques that rely on pairwise comparison matrices. Its effectiveness is evident in various domains, including the assessment of social sustainability in supply chains [12], expediting the implementation of technological innovations [15], and prioritization of the critical success factors of sustainable lean six sigma implementation [45]. The initial version of the BWM is based on a nonlinear optimization model. Therefore, when the number of alternatives or criteria increases, achieving convergence to a globally optimal solution becomes computationally demanding, as the problem becomes NP-hard [65]. The paper by [66] developed a linear formulation for the BWM to simplify the method and make it computationally tractable. In contrast to the original formulation of the BWM, the linear formulation does not allow for the existence of multiple optimal solutions, which could be beneficial when decision-makers seek alternative solutions. Despite this limitation, the linear formulation is computationally feasible and can be advantageous in many scenarios. An algorithm for performing the BWM using a linear formulation is presented in [66]. A comprehensive step-by-step discussion for performing the BWM can be found in [18,66]. Given a criteria set $C = \{c_1, c_2, \dots, c_n\}$, the best criteria $c_B \in C$, the worst criteria $c_W \in C$, the best-to-others vector $\vec{a}_B = (a_{B1}, a_{B2}, \dots, a_{Bn})$, and the others-to-worst vector $\vec{a}_W = (a_{1W}, a_{2W}, \dots, a_{nW})$, the linear formulation of the BWM can be expressed as follows:

$$\begin{aligned}
&\min \xi^L \\
&\text{s.t.} \\
&|w_B - a_{Bj}w_j| \leq \xi^L, \text{ for all } j \\
&|w_j - a_{jW}w_W| \leq \xi^L, \text{ for all } j \\
&\sum_j w_j = 1 \\
&w_j \geq 0, \text{ for all } j
\end{aligned} \tag{7}$$

The optimal weight vector, $\hat{\mathbf{w}}^*$ and consistency indicator, ξ^{L*} can be obtained by solving Eq. (7). According to [66], the consistency indicator, ξ^{L*}, can serve as a direct measure of consistency, which is different from the original model presented by [18], where the consistency index and consistency ratio are still utilized. Therefore, the closer the value of ξ^{L*} to zero, the greater the consistency of the solution obtained. The pseudocode for the linear BWM is presented in Algorithm 6.

Algorithm 6: Pseudocode of the best-worst method derived from [66]

```
Data: $C$: the criteria set, $c_B$: the best criterion, $c_W$: the worst
      criterion
Result: $\hat{\mathbf{w}}^*$: the optimal weight vector, $\xi^{L*}$: the consistency
        indicator
$\vec{a}_B \leftarrow$ initialize the best-to-others vector;
$\vec{a}_W \leftarrow$ initialize the others-to-worst vector;
for $j = 1$ to $n$ do
    $\vec{a}_{Bj} \leftarrow$ the rating of $c_B$ over $C_j$;
    $\vec{a}_{jW} \leftarrow$ the rating of $C_j$ over $c_W$;
end
Compute the optimal values $\hat{\mathbf{w}}^*$ and $\xi^{L*}$ by solving the linear
  BWM optimization model in Equation (7) given the ratings in
  $\vec{a}_B$ and $\vec{a}_W$.
```

3.4. Proposed feature selection approach

The proposed FS algorithm uses the BWM in ranking the features based on any choice of filter. Here the choice of filter is the criterion and the features are the alternatives in the context of MCDM. The paper by [16] provides a framework for looking at FS through the lenses of MCDM. The advantage of using BWM as a framework for doing univariate filter-based FS is that it allows for any filter to be used as the criterion in comparing the alternatives (or features). This allows the BWM to be useful regardless of the choice of filter desired by the user (e.g., information gain, correlation, mutual information). This process is explained in Step 3 of the procedure presented below. A pseudocode of the algorithm is provided in Algorithm 7. Each step in Algorithm 7 is explained as follows:

- **Step 1. Initialize the algorithm.** In this step, the variables, parameters, and configurations of the algorithm should be initialized. Determine the total set of features, the number of samples, and the class (or label) of each sample.
- **Step 2. Select the ranking criterion.** As with other univariate filter-based FS, the ranking measure facilitates how the features will be sorted according to some criterion (or measure). Although the other measures such as distance, correlation, and information theoretic, are equally valid to be selected, in this paper, we select

Table 4
Mapping for transforming distance to BWM rating.

Distance	Linguistic equivalent	Equivalent BWM rating
[0, 0.2)	Very weak	1
[0.2, 0.4)	Weak	3
[0.4, 0.6)	Fair	5
[0.6, 0.8)	Strong	7
[0.8, 1]	Very strong	9

correlation as the ranking measure as it is much easier understood. Since the proposed framework is concerned with binary classification problems, this paper offers a guide on selecting an appropriate type of correlation measure depending on the variable type of each feature as follows. If the feature is nominal, then the chi-square test for independence (see [67]) can be used to find the existence of a significant association between the feature and the class, while strength of association can be quantified using an effect size (e.g., Cramer's V) (see [68]). If the feature is ordinal, then Kendall τ can be used to find the strength of association between the feature and class. If the feature is continuous (i.e., interval or ratio), then the point-biserial correlation (see [69]) can be used to find the strength of association between the feature and class.

- **Step 3. Create a mapping between the ranking measure and the BWM's linguistic scale.** The obtained strength of association or effect size must be converted to an equivalent scale in the BWM method in order to carry out the proposed algorithm. There is no straightforward approach to perform the conversion. In this paper, we use the linguistic equivalent used in the literature as presented in Table 4. Such a mapping can also be modified by users (e.g., fuzzy/possibilistic distributions) depending on their preference.
- **Step 4. Obtain the priority weight of each feature.** By solving the BWM linear optimization model, the normalized priority weights and the consistency indicator can be obtained. The priority weight of a feature depicts its degree of importance relative to other features in the classification problem. It must be emphasized that the weights must be interpreted relative to the other features only. In other words, the weights do not imply an absolute meaning rather than simply a measure for ranking the features. Moreover, when the problem is solved repeatedly, the consistency indicator represents the degree at which the same solution is obtained. The closer the consistency indicator to zero, the greater the consistency of the solution [66].
- **Step 5. Select the features to retain.** With the priority weights obtained in the previous step, the features can be ranked from largest to smallest value. The feature/s with greatest weight receive/s the highest priority. To select the features to retain, the user has several options:

  1. Use a threshold that would serve as a cut for retaining features. Features that have priority weights lower than the threshold are dropped out. The threshold can be decided by experts or computed (e.g., mean of the weights)
  2. Retain only a predetermined number of features to retain. For example, select only the top 10 features.
  3. Use the available heuristics in literature such as the elbow method in cluster analysis, and scree plot in PCA. For convenience, this option is used in this paper.

  Other heuristics can also be adopted by users depending on their preference and the heuristic's appropriateness to the problem.
- **Step 6. Repeat steps 4 to 5 until desired performance of the classifier is achieved.** Due to the low computational cost of the proposed algorithm (i.e., linear programming is known in literature to have an algorithm bounded in polynomial time), the steps can be repeated until satisfactory performance is obtained. Doing so, however, may make the approach as expensive as wrapper methods because performing the next test may depend on the training time of the classifier. Repeating the algorithm many times on large-scale problems may not be preferable. Hence, this step is optional.

Algorithm 7: Pseudocode of the Proposed Feature Selection Algorithm.

```
Data: $D = (X, y)$: the original data matrix, $\phi$: mapping of the
      distance to the equivalent BWM rating, $\alpha$: threshold
      value
Result: $D'$: the filtered data matrix
$R \leftarrow [R_j]_{1\times M}$ // stores the correlation between
    each feature and the class label
for $X_j \in X,\ j = 1, \dots, M$ do
    $R_j \leftarrow$ correlation($X_j, y$);
end
$c_B \leftarrow \arg\max_{j=1,\dots,M}\{R_j\}$ // the best feature has the
    largest correlation
$c_W \leftarrow \arg\min_{j=1,\dots,M}\{R_j\}$ // the worst feature has the
    smallest correlation
$\vec{a}_B \leftarrow [a_{Bj}]_{1\times M}$ // initialize the best-to-others
    vector
$\vec{a}_W \leftarrow [a_{jW}]_{1\times M}$ // initialize the others-to-worst
    vector
for $j = 1$ to $M$ do
    $p_{Bj} \leftarrow$ distance$\left(R_{c_B}, R_j\right)$;
    $\vec{a}_{Bj} \leftarrow \phi\left(p_{Bj}\right)$ // transforms distance to the
        equivalent BWM rating as in Table 4
    $p_{jW} \leftarrow$ distance$\left(R_j, R_{c_W}\right)$;
    $\vec{a}_{jW} \leftarrow \phi\left(p_{jW}\right)$ // transforms distance to the
        equivalent BWM rating as in Table 4
end
Calculate the optimal priority weight vector $\hat{\mathbf{w}}^*$ using
  Algorithm 6;
Sort $\hat{\mathbf{w}}^*$ from highest to lowest to generate the feature ranking;
$D' \leftarrow$ filtered $D$ with the features having priority weights
  satisfying the threshold value $\alpha$
```

3.5. Statistical treatment of the data

In order to evaluate the effectiveness of the algorithm in reducing the dimensionality of large datasets, statistical analysis will be used to detect statistically significant differences. Consistent with the objectives, the paper aims to test if the proposed FS framework can reduce the dimensionality of the dataset without significantly reducing the performance of the classifiers. Furthermore, the paper aims to test if the proposed algorithm performs significantly better than a randomized algorithm (i.e., where features are selected randomly). To conduct the test, the experimental group (i.e., the proposed FS framework) will be compared to two control groups - (a) positive control group (i.e., the randomized algorithm), and (b) negative control group (i.e., no feature selection or all features retained). Two blocking factors are used - (a) Classifier, and (b) Training Dataset. The use of these blocking factors is consistent with the experimental design adopted in the current literature such as [70–72], and [73], among others. In summary, three factors are involved in this experiment - (i) FS algorithm (Factor A), (ii) Classifier (Factor B), and (iii) Training Dataset (Factor C). A three-way analysis of variance will be used to test significant differences between the experimental group and the control groups. The paper's research hypotheses are: (i) the experimental group is not statistically different from the negative control group, and (ii) the experimental

Table 5
Summary of the number of retained features for each dataset.

Training dataset	Proposed feature selection	Naive (all features)	Randomly selected features
Biodegradation	18	41	9
Diabetes	2	8	2
Heart Attack	8	13	3
Hepatitis	8	19	4
Ionosphere	4	35	7
Sonar	15	60	12
Wine Quality	2	11	3

group performs better than the positive control group. Preplanned contrasts are constructed to answer these hypothesis as follows. For Factor A, let $\bar{y}_j$ be the sample mean of the jth level, a_j be the contrast coefficient for the jth level, and $\hat{l}_k = \sum_j a_j \bar{y}_j$ be the kth contrast. For this experiment, the Helmert contrast is adopted to test the difference between the Experimental and Negative Control Group and the average of the Experimental and Negative Control Group with the Positive Control Group. There are three levels. Thus, we have $k = 2$ linear contrasts as shown in Eq. (8):

$$\begin{aligned} \hat{l}_1 &= [1, -1, 0] \cdot [\bar{y}_1, \bar{y}_2, \bar{y}_3] \\ \hat{l}_2 &= [1, 1, -2] \cdot [\bar{y}_1, \bar{y}_2, \bar{y}_3] \end{aligned} \tag{8}$$

4. Results

4.1. Feature selection

In this section, the results of the feature selection is presented. In Table 5, the number of retained features under the experimental group and the two control groups are presented. One important insight that can be inferred from Table 5 is that the number of features retained by the proposed FS algorithm is somewhat close to the number of features retained by the randomized algorithm. It is important to note that the proportion of features retained by the randomized algorithm is fixed at 20% following the Pareto principle. This principle was adopted due to the lack of available works dealing randomized FS in the current literature. As for the proposed FS, the number of features retained is decided based on their optimal weights. These optimal weights are plotted on an elbow plot to visually discriminate between the features, as in Figs. 2(a)–2(f). Afterwards, the number of features retained are those that belong to the highest group discriminated by the algorithm. For example, in Fig. 2(b), only *Glucose* and *BMI* are retained. In practice, decision-makers can choose a threshold arbitrarily depending on the problem. Looking closely into the results of Table 5 would imply that the proposed FS retains about 32% of the features on average. Moreover, the resulting elbow plots indicate that the algorithm is able to distinguish between the features.

4.2. Exploratory analysis

This study employs five classifiers: (i) ANN, (ii) decision trees, (iii) k-NN, (iv) logistic regression, and (v) SVM. The performance of a classifier is measured using the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. The ROC curve plots the true positive rate versus the false positive rate of a classifier. For binary classification, several scholars (e.g., [74,75]) agree that ROC is a good way of determining the performance of a classifier because it allows the selection of a suitable decision threshold. Using this criterion, an AUC equal to 1 is considered ideal. First, 1050 experimental units are randomly assigned to the experimental group and control groups considering the two blocking factors. The AUC is obtained from the classifiers under these conditions. A table of descriptive statistics is omitted for brevity. However, the mean for each combination is displayed in the interaction plot in Figs. 3(a)–3(b). By inspecting the interaction plots in Figs. 3(a) and 3(b), two insights can be obtained. First, the negative control group (i.e., *Naive*) and the experimental group (i.e., *Proposed FS*) appear to be very close. However, their average appears to be greater than the negative control group (i.e., *Random*). Moreover, it can be inferred from Figs. 3(a)–3(b) that interaction effects may be present. In order to compare the main effects of the experimental groups and the control groups in a straightforward manner, interaction terms involving factor A must not be statistically significant. In the next section, an iterative procedure for fixing the levels where interactions exist is presented.

Table 6
Summary of three-way ANOVA results. Factor A refers to the feature selection setup. Factor B refers to the type of classifiers. Factor C refers to the training dataset used.

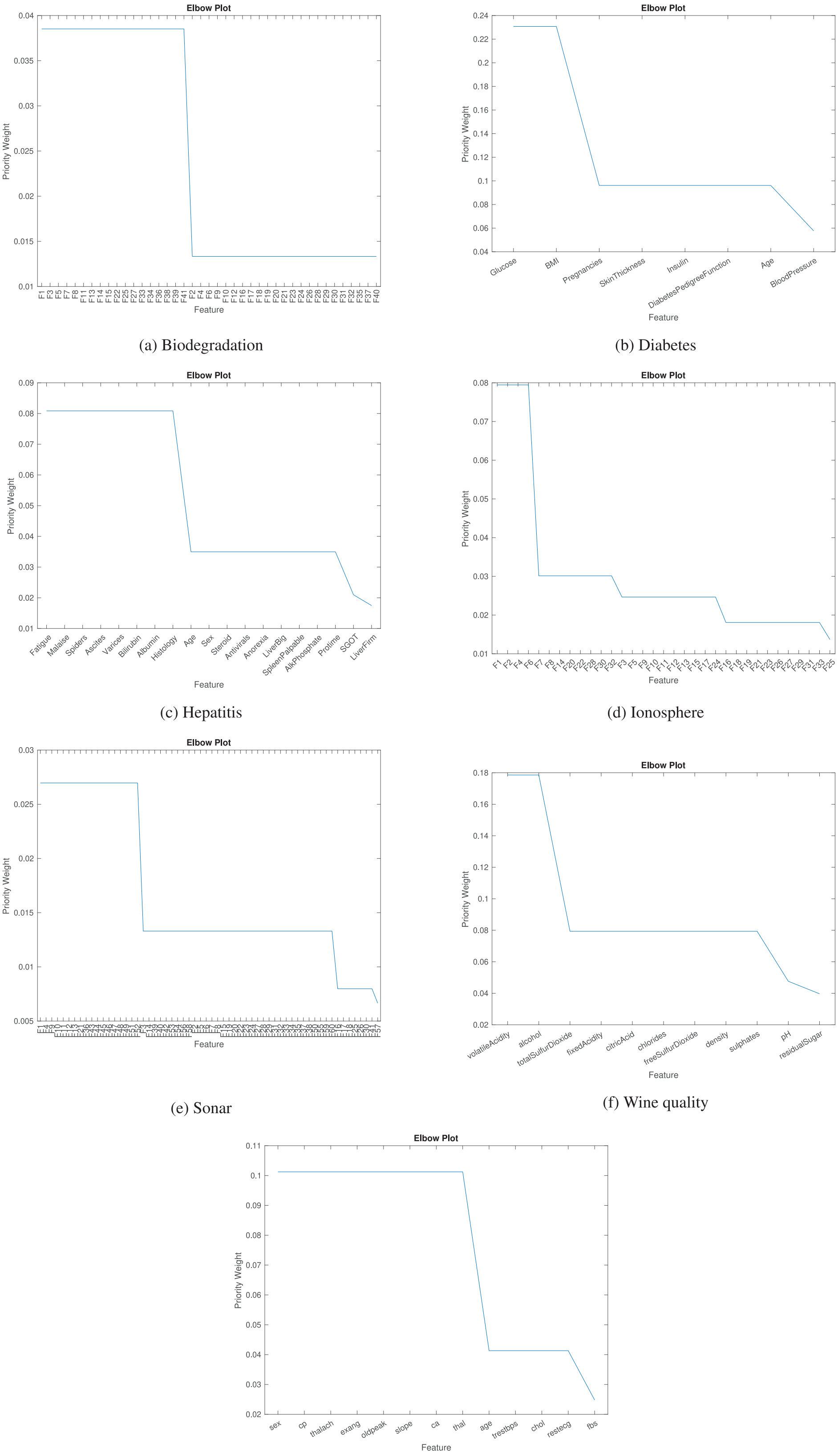
Factor	Df	SS	MS	F value	p-value	Significant
A	2	1.70844	0.85422	107.5185	<0.000001	Yes
B	2	0.01146	0.00573	0.7210	0.487279	No
C	2	1.98714	0.99357	125.0582	<0.000001	Yes
$A \times B$	4	0.03851	0.00963	1.2117	0.306394	No
$A \times C$	4	0.02566	0.00641	0.8073	0.521503	No
$B \times C$	4	0.13330	0.03332	4.1944	0.002654	Yes
$A \times B \times C$	8	0.11301	0.01413	1.7780	0.081958	No
Residuals	243	1.93060	0.00794			

Table 7
Summary of contrast results.

Contrast	F value	F critical value	Significant
$\hat{l}_1$	1.741008	5.200226	No
$\hat{l}_2$	213.296064	5.200226	Yes

4.3. Three-way analysis of variance

A three-way ANOVA is used to analyze the AUC for the experimental group and the control groups. However, results suggest that significant interaction effects involving Factor A exist due to some of the factor levels being intertwined. Since Factor B and Factor C would only serve as blocking factors, the levels that make the interaction terms significant are fixed with the help of the interaction plots in Figs. 3(a) and 3(b). The process is performed iteratively until the three-way interaction and the two interaction involving Factor A are not statistically significant. As a result, only three levels for Factor B and Factor C were used to analyze the main effect of Factor A. For Factor B (i.e., *Classifier*), the levels are: (i) ANN, (ii) decision tree, and (iii) k-NN. For Factor C (i.e., *Training Dataset*), the levels are: (i) Hepatitis, (ii) Ionosphere, and (3) Sonar. The interaction plots using this setup are presented in Figs. 4(a)–4(c). While this approach (i.e., fixing intertwined levels) is valid, the generalization is limited only to the levels included in the model. Because this paper is the first (to the best of the authors' knowledge) to report the presence of significant interactions between the considered factors in the context of filter-based FS, scholars must pay attention to how these drawbacks can be addressed in future works. Using this setup, the three-way ANOVA is conducted and the results are presented in Table 6. Since the interaction terms involving Factor A are not statistically significant, the main effect can be used to compare the experimental group and the control groups in a straightforward manner. Considering that the main effect of Factor A is statistically significant, the Helmert contrast in Section 3.5 can be used to compare the experimental and control groups as in Table 7. As such, results show that the experimental group (i.e., *Proposed FS*) and the negative control group (i.e., *Naive*) are not statistically different. However, the mean of the positive control group (i.e., *Random*) is statistically less than the experimental group (and the negative control group). These results provide evidence that the proposed FS algorithm reduces the dimensionality of the datasets without reducing classification performance.



(a) Biodegradation

(b) Diabetes

(c) Hepatitis

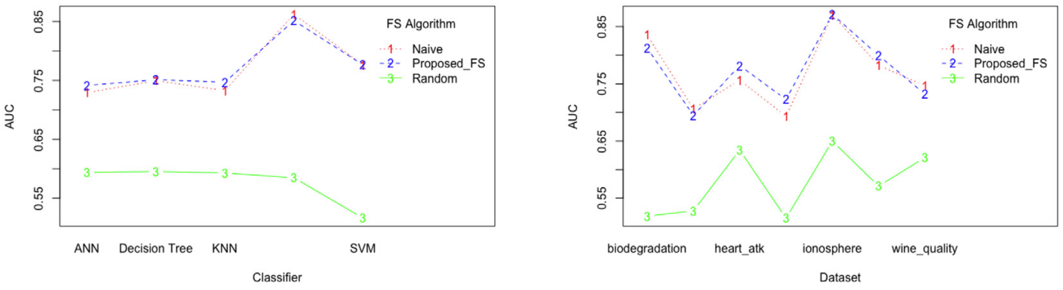
(d) Ionosphere

(e) Sonar

(f) Wine quality

(g) Heart attack

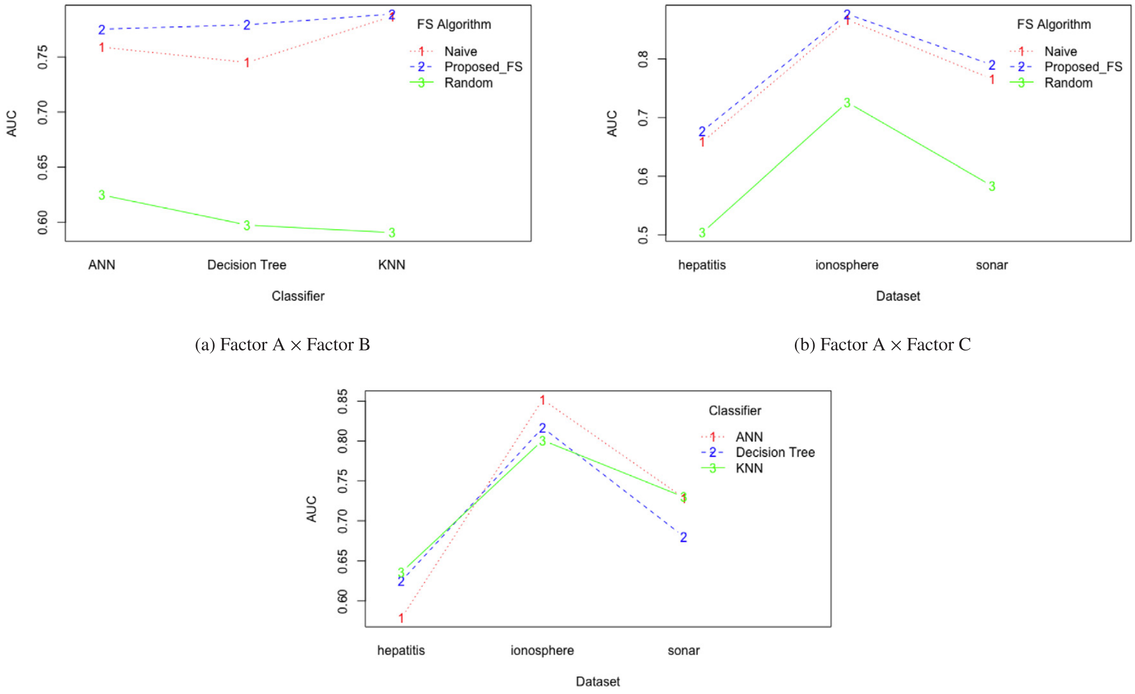
Fig. 2. Elbow plot of the optimal weights for each dataset using the proposed approach.



(a) Factor A × Factor B

(b) Factor A × Factor C

Fig. 3. Interaction plot between Factor A (FS algorithms) with Factor B (Classifiers) and Factor C (Training dataset). This plot presents how each FS algorithm vary across the considered blocking factors.



(a) Factor A × Factor B

(b) Factor A × Factor C

(c) Factor B × Factor C

Fig. 4. Interaction plot of between Factor A (FS algorithms), Factor B (Classifiers), and Factor C (Training Dataset) without the fixed levels. This plot is a result of disentangling the interacting levels in Fig. 3. This step is crucial for interpreting the ANOVA results in a more straightforward manner (i.e., without the interactions).

Moreover, the proposed FS algorithm performs significantly better than the randomized algorithm (i.e., randomly selecting features).

5. Discussion and implications

In this study, a filter-based FS framework is developed. Such a framework would serve as a general structure for doing FS using filters. In the current literature, filter-based FS simply ranks the features using the results given by the preferred criterion, such as chi-square, information gain, Spearman's correlation, and mutual information, among others. While the simplicity of the approach is appealing, several drawbacks can also result from it. This paper points out four major drawbacks:

1. There is no direct way of comparing one feature relative to all other features. In practice, decisionmakers must be able to judge features based on how they perform relative to all other features.
2. There is no direct way of deciding up to how many features should be retained. For example, should features with correlation of at least 60%. Bottom line, there is a significant lack of a methodological reason to retain features.
3. With the lack of a unifying framework, decisionmakers would have to learn low-level details such as the characteristics of correlation, chi-square, and mutual information, among others. Consequently, this would make the feature selection more complex and time consuming.
4. When attempting to compare different measures, interpretability becomes more challenging. For example, decisionmakers will need to compare information gain with chi-square. Due to difference in how these criteria calculate relevance, direct comparisons will need more expert knowledge. As a result, exploratory analysis by non-experts will be more difficult.

The proposed framework overcomes these drawbacks. First, by deriving an optimal weight for each feature based on how it performs compared to the best feature and the worst feature, each feature is given priority based on these two extreme values. Thus, a relative weight can be inferred from the result. Second, the paper demonstrates the use of an elbow plot to decide the features to be retained. This process is similar to the case in principal component analysis (PCA) and cluster analysis. Thus, the proposed approach provides a methodological basis that is well-understood in the relevant literature. Third, by returning a set of optimal weights, the proposed approach abstracts the *association process*. Hence, allowing decisionmakers to focus on the optimal weights instead of the individual association functions (e.g., chi-square, point biserial correlation). This makes exploratory analysis more straightforward for non-experts (as well experts). Finally, by focusing only on the optimal weights, the interpretation of the results become unified regardless of the specific measure (e.g., chi-square) used.

The paper also has significant implications for the empirical validation of feature selection algorithms. This paper considers the *Training Dataset* and *Classifier* as blocking factors. Moreover, positive and negative control groups are constructed to compare with the experimental group. In the presence of significant interactions, the paper fixed the intertwined levels with the aid of the interaction plot. While the approach used in this paper is empirically sound, very few works in the current literature resort to similar approaches. On the one hand, majority of the works in the current literature use only a descriptive approach in evaluating FS algorithms. While such an approach is valid, it is not powerful enough to determine if a newly proposed algorithm truly exhibits statistical significance. Thus, scholars should pay attention to developing a unified testing framework for evaluating new filter-based feature selection algorithms. This framework would help facilitate the analysis of FS algorithms.

6. Conclusion and future works

Classification algorithms are used in a wide spectrum of domains. With the presence of large data, these algorithms can be trained to perform prediction tasks. A major drawback for classification algorithms is the curse of dimensionality, due to sparse and high dimensional data. To address this drawback, feature selection is performed to reduce dimensionality without significantly reducing the performance of classifiers when trained using the reduced dataset. The focus of this paper is on filter-based feature selection. In this particular area, several gaps arise in the current literature. Specifically, the lack of a general framework for doing feature selection using filters. Such a framework offers several benefits for the literature. For one thing, it facilitates the development of a formal analysis on filter-based FS algorithms. In turn, it enables scholars to analyze new FS algorithms in a more rigorous setting. Moreover, instead of performing separate analyses for each filter, the framework will allow scholars to focus on a single analysis. In this paper, a general framework for univariate feature selection is introduced. The proposed framework is based on the BWM.

The advantage of MADM for filter-based FS is the availability of well-understood frameworks for performing ranking, selection, and prioritization problems. Results show that the proposed framework can reduce the dimensionality of the datasets without significantly affecting classification performance. Furthermore, the proposed framework performs significantly better than a randomized algorithm. This study offers several contributions to the literature. First, this paper is one of the very few to use MADM for feature selection. Second, it induces the cross-fertilization of MADM and feature selection, which would open new research avenues for scholars in both fields. Third, to the best of the authors' knowledge, this is the first paper to provide evidence of a significant interaction between the factors used for testing feature selection algorithms. Fourth, the paper is one of the very few papers in the current literature to adopt significance testing for assessing the proposed FS algorithm while majority of the papers adopt a descriptive approach. Finally, to the best of the authors' knowledge, the paper is the first to simultaneously include the *training dataset* and *classifiers* as blocking factors in comparing feature selection algorithms.

While the approach adopted in this paper is extensive, its generalization is limited to the experimental setup adopted. For one thing, the experimental design treats the blocking variables as fixed factors instead of random factors. This limitation is due to the interaction terms that need to be fixed in order to have a straightforward interpretation of the main effects. Furthermore, the lack of a well-agreed set of test datasets for feature selection restricts the randomization required for treating the blocking variables as random factors. For future works, scholars are encouraged to pay attention on developing a unified testing framework for testing new filter-based FS algorithms. A unified testing framework should include a set of classifiers and test datasets that are known not to exhibit interactions. Moreover, scholars should also find strategies for treating the blocking factors as random factors without exhibiting interactions. This paper focuses only on univariate filter-based FS. For future works, scholars may explore on a multivariate version that ranks features using several association criteria (e.g., information gain, chi-square, and correlation) simultaneously. Moreover, the concept of practical significance is loosely discussed in the current literature. Considering this, scholars must put forward efforts that would relate statistical significance and practical significance.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets used in this study are available in public repositories (UCI Machine Learning Repository, and Kaggle).

References

[1] Y.-J. Mao, H.-J. Lim, M. Ni, W.-H. Yan, D.W.-C. Wong, J.C.-W. Cheung, Breast tumour classification using ultrasound elastography with machine learning: A systematic scoping review, Cancers 14 (2) (2022) 367.

[2] P. Dhal, C. Azad, A comprehensive survey on feature selection in the various fields of machine learning, Appl. Intell. (2022) 1–39.

[3] H.H. Htun, M. Biehl, N. Petkov, Survey of feature selection and extraction techniques for stock market prediction, Financ. Innov. 9 (1) (2023) 26.

[4] M.N. Al-Mhiqani, R. Ahmad, Z.Z. Abidin, W. Yassin, A. Hassan, K.H. Abdulkareem, N.S. Ali, Z. Yunos, A review of insider threat detection: Classification, machine earning techniques, datasets, open challenges, and recommendations, Appl. Sci. 10 (15) (2020) 5208.

[5] W. Ali, F. Saeed, Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data, Processes 11 (2) (2023) 562.

[6] S.L. Marie-Sainte, N. Alalyani, Firefly algorithm based feature selection for Arabic text classification, J. King Saud Univ.-Comput. Inf. Sci. 32 (3) (2020) 320–328.

[7] M. Yang, L. Tan, X. Chen, Y. Luo, Z. Xu, X. Lan, Laws and regulations tell how to classify your data: A case study on higher education, Inf. Process. Manage. 60 (3) (2023) 103240.

[8] M. Pradhan, Cardiac image-based heart disease diagnosis using bio-inspired optimized technique for feature selection to enhance classification accuracy, in: Machine Learning and AI Techniques in Interactive Medical Image Analysis, IGI Global, 2023, pp. 151–166.

[9] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, H. Liu, Feature selection: A data perspective, ACM Comput. Surv. 50 (6) (2017) 1–45.

[10] K. Kalaiselvi, S.B. V.J. Sara, A hybrid filter wrapper embedded-based feature selection for selecting important attributes and prediction of chronic kidney disease, in: International Conference on Computing, Communication, Electrical and Biomedical Systems, Springer, 2022, pp. 137–153.

[11] A.J. Tallón-Ballesteros, J.C. Riquelme, R. Ruiz, Filter-based feature selection in the context of evolutionary neural networks in supervised machine learning, Pattern Anal. Appl. 23 (1) (2020) 467–491.

[12] Y. Tavakoli Haji Abadi, S. Avakh Darestani, Evaluation of sustainable supply chain risk: evidence from the Iranian food industry, J. Sci. Technol. Policy Manag. 14 (1) (2023) 127–156.

[13] M. Deveci, P.R. Brito-Parada, D. Pamucar, E.A. Varouchakis, Rough sets based Ordinal Priority Approach to evaluate sustainable development goals (SDGs) for sustainable mining, Resour. Policy 79 (2022) 103049.

[14] S. Vatankhah, M. Darvishmotevali, R. Rahimi, S.M. Jamali, N. Ale Ebrahim, MCDM in travel and tourism research since 1997: A bibliometric approach, Int. J. Contemp. Hosp. Manag. (2023).

[15] G. Soni, S. Kumar, R.V. Mahto, S.K. Mangla, M. Mittal, W.M. Lim, A decision-making framework for Industry 4.0 technology implementation: The case of FinTech and sustainable supply chain finance for SMEs, Technol. Forecast. Soc. Change 180 (2022) 121686.

[16] D.P.M. Abellana, R.R. Roxas, D.M. Lao, P.E. Mayol, S. Lee, Ensemble feature selection in binary machine learning classification: A novel application of the evaluation based on distance from average solution (EDAS) method, Math. Probl. Eng. 2022 (2022).

[17] G. Fu, B. Li, Y. Yang, C. Li, Re-ranking and TOPSIS-based ensemble feature selection with multi-stage aggregation for text categorization, Pattern Recognit. Lett. 168 (2023) 47–56.

[18] J. Rezaei, Best-worst multi-criteria decision-making method, Omega 53 (2015) 49–57.

[19] S. Hamdan, S. Almajali, M. Ayyash, H.B. Salameh, Y. Jararweh, An intelligent edge-enabled distributed multi-task learning architecture for large-scale IoT-based cyber–physical systems, Simul. Model. Pract. Theory 122 (2023) 102685.

[20] N. Schwalbe, B. Wahl, Artificial intelligence and the future of global health, Lancet 395 (10236) (2020) 1579–1586.

[21] I.H.V. Gue, A.T. Ubando, M.-L. Tseng, R.R. Tan, Artificial neural networks for sustainable development: a critical review, Clean Technol. Environ. Policy (2020) 1–17.

[22] R.J. Manoj, M.A. Praveena, K. Vijayakumar, An ACO–ANN based feature selection algorithm for big data, Cluster Comput. 22 (2) (2019) 3953–3960.

[23] M. Kang, N.J. Jameson, Machine learning: Fundamentals, in: Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things, Wiley Online Library, 2018, pp. 85–109.

[24] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, A.J. Aljaaf, A systematic review on supervised and unsupervised machine learning algorithms for data science, in: Supervised and Unsupervised Learning for Data Science, Springer, 2020, pp. 3–21.

[25] N. Naik, Y. Sharath Kumar, Efficient feature selection algorithm for gene classification, in: Cognition and Recognition: 8th International Conference, ICCR 2021, Mandya, India, December 30–31, 2021, Revised Selected Papers, Springer, 2023, pp. 165–189.

[26] Y. Lyu, Y. Feng, K. Sakurai, A survey on feature selection techniques based on filtering methods for cyber attack detection, Information 14 (3) (2023) 191.

[27] M. Rostami, K. Berahmand, E. Nasiri, S. Forouzandeh, Review of swarm intelligence-based feature selection methods, Eng. Appl. Artif. Intell. 100 (2021) 104210.

[28] H. Benhar, M. Hosni, A. Idri, Univariate and multivariate filter feature selection for heart disease classification, J. Inf. Sci. Eng. 38 (4) (2022).

[29] R. Abu Khurma, I. Aljarah, A. Sharieh, M. Abd Elaziz, R. Damaševičius, T. Krilavičius, A review of the modification strategies of the nature inspired algorithms for feature selection problem, Mathematics 10 (3) (2022) 464.

[30] S. Walling, S. Lodh, Performance evaluation of supervised machine learning based intrusion detection with univariate feature selection on NSL KDD dataset, 2023.

[31] H. Benhar, A. Idri, M. Hosni, Impact of threshold values for filter-based univariate feature selection in heart disease classification, in: HEALTHINF, 2020, pp. 391–398.

[32] H. Tan, G. Wang, W. Wang, Z. Zhang, Feature selection based on distance correlation: a filter algorithm, J. Appl. Stat. 49 (2) (2022) 411–426.

[33] G. Manikandan, S. Abirami, An efficient feature selection framework based on information theory for high dimensional data, Appl. Soft Comput. 111 (2021) 107729.

[34] S. Pande, A. Khamparia, D. Gupta, Feature selection and comparison of classification algorithms for wireless sensor networks, J. Ambient Intell. Humaniz. Comput. (2021) 1–13.

[35] K. Liu, D. Wang, W. Du, D.O. Wu, Y. Fu, Interactive reinforced feature selection with traverse strategy, Knowl. Inf. Syst. (2023) 1–28.

[36] L. Yu, H. Liu, Feature selection for high-dimensional data: A fast correlation-based filter solution, in: Proceedings of the 20th International Conference on Machine Learning (ICML-03), 2003, pp. 856–863.

[37] N.A. Samee, G. Atteia, S. Meshoul, M.A. Al-antari, Y.M. Kadah, Deep learning cascaded feature selection framework for breast cancer classification: Hybrid CNN with univariate-based approach, Mathematics 10 (19) (2022) 3631.

[38] S.P. Potharaju, et al., Design and implementation of feature selection approaches using filter based ranking methods, 2021.

[39] I.J. Jebadurai, G.J.L. Paulraj, J. Jebadurai, S. Silas, Experimental analysis of filtering-based feature selection techniques for fetal health classification, Serbian J. Electr. Eng. 19 (2) (2022) 207–224.

[40] A. Homayouni, T. Liu, T. Thieu, Diabetic retinopathy prediction using progressive ablation feature selection: A comprehensive classifier evaluation, Smart Health 26 (2022) 100343.

[41] M. Deveci, I. Gokasar, O. Castillo, T. Daim, Evaluation of Metaverse integration of freight fluidity measurement alternatives using fuzzy Dombi EDAS model, Comput. Ind. Eng. 174 (2022) 108773.

[42] S.S.H. Dehshiri, B. Firoozabadi, A new multi-criteria decision making approach based on wins in league to avoid rank reversal: A case study on prioritizing environmental deterioration strategies in arid urban areas, J. Clean. Prod. 383 (2023) 135438.

[43] I.E. Essoussi, M. Masmoudi, M.Z. Babai, Multi-criteria decision-making for collaborative COVID-19 surge management and inter-hospital patients' transfer optimisation, Int. J. Prod. Res. (2023) 1–30.

[44] X. Ma, Y. Liu, J. Yan, S. Han, L. Li, H. Meng, M. Deveci, K. Kölle, U. Cali, Assessment method of offshore wind resource based on a multi-dimensional indexes system, CSEE J. Power Energy Syst. (2022).

[45] V. Swarnakar, A. Bagherian, A. Singh, Prioritization of critical success factors for sustainable Lean Six Sigma implementation in Indian healthcare organizations using best-worst-method, TQM J. 35 (3) (2023) 630–653.

[46] K. Mansouri, T. Ringsted, D. Ballabio, R. Todeschini, V. Consonni, Quantitative structure–activity relationship models for ready biodegradability of chemicals, J. Chem. Inf. Model. 53 (4) (2013) 867–878.

[47] D. Dua, C. Graff, UCI machine learning repository, 2019, URL http://archive.ics.uci.edu/ml.

[48] National Institute of Diabetes, Pima Indians diabetes database, 2016, URL https://www.kaggle.com/uciml/pima-indians-diabetes-database.

[49] J.W. Smith, J. Everhart, W. Dickson, W. Knowler, R. Johannes, Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, in: Proceedings of the Annual Symposium on Computer Application in Medical Care, American Medical Informatics Association, 1988, p. 261.

[50] N. Anand, A. Janosi, W. Steinbrunn, P. Pfisterer, R. Detrano, Heart attack prediction, 2018, URL https://www.kaggle.com/imnikhilanand/heart-attack-prediction.

[51] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis, Modeling wine preferences by data mining from physicochemical properties, Decis. Support Syst. 47 (4) (2009) 547–553.

[52] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inform. Theory 13 (1) (1967) 21–27.

[53] Z. Pan, Y. Wang, Y. Pan, A new locally adaptive k-nearest neighbor algorithm based on discrimination class, Knowl.-Based Syst. (2020) 106185.

[54] M. Alloghani, A. Aljaaf, A. Hussain, T. Baker, J. Mustafina, D. Al-Jumeily, M. Khalaf, Implementation of machine learning algorithms to create diabetic patient re-admission profiles, BMC Med. Inform. Decis. Mak. 19 (2019) 1–16.

[55] Priyanka, D. Kumar, Decision tree classifier: a detailed survey, Int. J. Inf. Decis. Sci. 12 (3) (2020) 246–269.

[56] A. Cherfi, K. Nouira, A. Ferchichi, Very fast C4. 5 decision tree algorithm, Appl. Artif. Intell. 32 (2) (2018) 119–137.

[57] V. Kumar, Evaluation of computationally intelligent techniques for breast cancer diagnosis, Neural Comput. Appl. 33 (8) (2021) 3195–3208.

[58] A. Robles-Velasco, P. Cortés, J. Muñuzuri, L. Onieva, Prediction of pipe failures in water supply networks using logistic regression and support vector classification, Reliab. Eng. Syst. Saf. 196 (2020) 106754.

[59] O. Chapelle, L. Li, An empirical evaluation of Thompson Sampling, Adv. Neural Inf. Process. Syst. 24 (2011).

[60] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 1992, pp. 144–152.

[61] S. Shalev-Shwartz, S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.

[62] M. Kubat, Artificial neural networks, in: An Introduction to Machine Learning, Springer, 2015, pp. 91–111.

[63] H. Ramchoun, M.A.J. Idrissi, Y. Ghanou, M. Ettaouil, Multilayer perceptron: Architecture optimization and training, IJIMAI 4 (1) (2016) 26–30.

[64] S. Mishra, H.K. Tripathy, B.K. Mishra, Implementation of biologically motivated optimisation approach for tumour categorisation, Int. J. Comput. Aided Eng. Technol. 10 (3) (2018) 244–256.

[65] S. Burer, A.N. Letchford, Non-convex mixed-integer nonlinear programming: A survey, Surv. Oper. Res. Manag. Sci. 17 (2) (2012) 97–106.

[66] J. Rezaei, Best-worst multi-criteria decision-making method: Some properties and a linear model, Omega 64 (2016) 126–130.

[67] M.L. McHugh, The chi-square test of independence, Biochem. Med.: Biochem. Med. 23 (2) (2013) 143–149.

[68] H.-Y. Kim, Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test, Restor. Dent. Endod. 42 (2) (2017) 152–155.

[69] R.F. Tate, Correlation between a discrete and a continuous variable. Point-biserial correlation, Ann. Math. Stat. 25 (3) (1954) 603–607.

[70] C. Yao, Y.-F. Liu, B. Jiang, J. Han, J. Han, LLE score: A new filter-based unsupervised feature selection method based on nonlinear manifold embedding and its application to image recognition, IEEE Trans. Image Process. 26 (11) (2017) 5257–5269.

[71] A.K. Uysal, S. Gunal, A novel probabilistic feature selection method for text classification, Knowl.-Based Syst. 36 (2012) 226–235.

[72] R.H. Pinheiro, G.D. Cavalcanti, R.F. Correa, T.I. Ren, A global-ranking local feature selection method for text categorization, Expert Syst. Appl. 39 (17) (2012) 12851–12857.

[73] R. Cekik, A.K. Uysal, A novel filter feature selection method using rough set for short text data, Expert Syst. Appl. 160 (2020) 113691.

[74] S.S. Han, M.S. Kim, W. Lim, G.H. Park, I. Park, S.E. Chang, Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm, J. Invest. Dermatol. 138 (7) (2018) 1529–1538.

[75] K. Li, Y. Fang, W. Li, C. Pan, P. Qin, Y. Zhong, X. Liu, M. Huang, Y. Liao, S. Li, CT image visual quantitative evaluation and clinical classification of coronavirus disease (COVID-19), Eur. Radiol. 30 (8) (2020) 4407–4416.