Iranian Journal of Fuzzy Systems

Volume *, Number *, (****), pp. **

Original paper

Reducing dimensionality using copula-based extropy

F. Arad [1], A. Sheikhi [2] and F. Keynia [3]

[1,2] *Department of Statistics, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran.*
[3] *Department of Energy Management and Optimization, Institute of Science and High Technology and Environmental Sciences, Graduate University of Advanced Technology, Kerman, Iran.*

fereshteh_arad@math.uk.ac.ir, sheikhy.a@uk.ac.ir, fkeynia@gmail.com

Abstract

When dealing with high-dimensional datasets, dimensionality reduction techniques become particularly important. Feature selection is one such method, aiming to select a subset of features such that the selected features have the highest relevance to the target variable and the lowest redundancy with each other. In this paper, we intend to utilize extropy, as a complement to entropy, for feature selection. Since common feature selection methods often ignore the dependencies between variables, we will consider these dependencies in the selection process by using copula functions. To this end, a feature selection algorithm based on copula functions and extropy is introduced, and the results of its implementation on real and simulated datasets are presented. More specifically, we assess our proposed approach using three scenarios in simulation study and we apply it to feature selection in three real-world datasets. These results demonstrate that while our proposed feature selection method significantly outperforms other alternatives, it benefits from capturing the nonlinear dependencies.

Keywords: Dimensionality reduction, copula functions, extropy.

2020 MSC: 62H25, 62H05.

1 Introduction

In the real world, data sets contain a large number of features or variables. A high number of features creates problems such as increased processing time, overfitting and ultimately problems in interpreting the data. One of the solutions to deal with such problems is dimensionality reduction, which overcomes these problems by selecting the most important features and reducing their number while preserving important information [14]. On the other hand, by reducing the dimensions of the data, one may can better identify the relationships between the data and ultimately improve the performance of data analysis methods [5].

Dimensionality reduction can be carried out with two main approaches. 1. Feature extraction 2. Feature selection. In feature extraction, we seek to construct new features from the original features so that the new features have as much information as the original features. This selection is beneficial in two ways. 1. The selected variables have the most information. 2. Unnecessary features are eliminated. Nanga et al. (2021) [19], after reviewing 341 articles, divide dimension reduction techniques into two categories: linear and nonlinear, and on the other hand, show that these techniques are divided into three categories of learning types: unsupervised, supervised, and semi-supervised learning we refer to [10, 24, 26, 35].

As a matter of fact, feature selection seeks to find a subset of predictor variables that meet two important requirements: 1. The selected variables have the highest correlation with the target variable 2. The selected variables have the lowest redundancy among themselves. This is done by the Minimum Redundancy Maximum Relevance (MRMR) method [35]. Billah and Waheed (2020) [3] used MRMR to identify polyps in the digestive tract; Wang et al. (2022) [36]

Corresponding Author: A. Sheikhi
Received: June 2026; Revised: September 2026; Accepted: November 2026.
https://doi.org/10.22111/ijfs.2026.55828.9912, https://ijfs.usb.ac.ir, Print ISSN: 1735-0654

showed that MRMR can be used to diagnose heart disease with more accuracy, and also showed that features selected in this method are important that were previously ignored in previous methods.

One challenge with MRMR is that it does not consider dependencies between variables. Accounting for inter-variable dependencies is essential, as variables inherently interact. These dependencies intensify within complex structures and nonlinear models, and neglecting them will compromise the model's optimal performance [34]. Shad Muhammad et al. (2014) [17] deal with modeling nonlinear behaviors between independent and dependent variables and show that the introduced model is more flexible than other models. By introducing a new method based on independent components, hierarchical model, and mixture of Laplacian distributions, Park and Lee (2004) [22] showed that nonlinear and complex dependencies between a different parts of an image can be identified with this method. In fact, with this method, the image can be better segmented. In this work, we consider this dependency using copula functions. Lall et al. (2021) [16] improved the MRMR method using a copula function, introducing the Copula-Based Feature Selection Algorithm. Femmam et al. (2023) [8] generalized the MRMR method using a copula function, introducing the GBCFS. See also other articles, including Peng et al. (2023) [23] and Femmam and Femmam (2022) [9] which have addressed feature selection and, consequently, dimensionality reduction using copulas. In addition to methods such as mRMR, other models are also used for feature selection.

Random Forest and LASSO are among these methods. For example, some studies such as the Díaz-Uriarte and Alvarez de Andrés (2006) [7] have shown that using the Random Forest algorithm, important features also be selected; also Reif et al. (2006) [27] used the copula-based Random Forest algorithm as a feature selection approach to identify important genes and proteins in high-dimensional genomic and proteomic datasets. We can also mention the study by Fonti and Belitser (2017) [11], in which the LASSO method was used to select features in regression models in high-dimensional data.

On the other hand, in recent years, a new concept called extropy has been introduced, which is known as the complement of entropy. Entropy and extropy are two sides of the same coin, or better said, entropy is a measure of uncertainty and extropy is a measure of certainty in a system. Entropy indicates disorder or randomness, while extropy indicates order [15]. In fact, in today's world, better decisions can be made by recognizing patterns, which is why we need extropy, which helps identify patterns in data. For example, by recognizing consumer behavioral patterns from the market, better decisions can be made to predict consumer behavior in the future. Therefore, by calculating extropy, predictions can be made more accurately. Many refer to [30] and [18] for more information.

Overall, understanding these two concepts and the complex relationship between them helps us develop statistical models, thereby strengthening our ability to predict and make better decisions. Nowadays, Extropy is also used to identify patterns and increase accuracy [4]. Toomaj et al. (2023) [33] discuss several advantages of extropy in their paper. As stated, copula functions are used to account for dependencies between variables. Furthermore, copula functions are employed to improve extropy estimations, particularly when the dependence between variables is non-linear and complex. For example, Saha and Kayal (2026) [28] considered the copula density to calculate the extropy of two variables.

Since LASSO, mRMR, mRMR-Copula, and Random Forest employ distinct methodologies, they each have their own disadvantages. In the LASSO method, the algorithm tends to perform well when linear relationships exist; however, it begins to fail when it is confronted with non-linear relationships or complex interactions among variables [12]. In Random Forest, the structure of the trees influences which variables are selected, which may lead to some low-importance variables being incorrectly identified as important [32]. Also, in mRMR-Copula, numerical instability may occur due to the presence of logarithms in the calculations when dealing with noisy or small-sized data [13]. In addition, this method focuses more on information content than on the distance from independence.

In contrast, the proposed algorithm in this paper, titled Feature Selection using Extropy Copula (FSEC), largely overcomes these limitations. In fact, the main novelty of this paper is that the feature selection process is performed based on a new criterion, which we denote as IEC. Specifically, this criterion differs from traditional copula-based criteria, in that previous methods mainly focus only on the dependence structure between two random variables, such as the criterion proposed by Saha and Kayal (2026) [28]. In contrast, the proposed criterion considers information about marginal distributions in addition to dependence; therefore, it can be considered a generalization of copula-based criteria in the existing literature. Also, from a relational perspective, the proposed criterion establishes a formal relation between the dependence structure and marginal distributions, which has not been fully considered in previous criteria. On the other hand, the FSEC algorithm, which operates based on the proposed IEC criterion, can efficiently select appropriate features from high-dimensional data, in a way that both useful information is preserved and the computational burden is reduced. As a result, this method allows for more accurate identification of nonlinear and complex relationships between features.

The proposed method shows good performance; however, there are a few practical considerations to be noted. First, computing double integrals for copula functions usually requires the use of numerical approximations, which can

increase computational time and cost. However, given the processing power of today's computers, this is manageable in most applications. Second, the performance of the method may depend to some extent on the distribution structure and dependencies of the data, although in most cases the results are stable and no significant changes are observed.

This article consists of several main sections: First, we describe the necessary background and basic discussions to introduce the method in question. Then, in Section 3, we introduce and explain the FSEC algorithm for feature selection. Finally, Section 4 is dedicated to the implementation of this algorithm in numerical studies.

2 Background

In this section, we will briefly review the methods used in this article, including copula theory, entropy and extropy and neural networks.

2.1 Copula function

In the context of the possible situation, the copula $C : [0,1]^d \longrightarrow [0,1]$ is a joint cumulative distribution function of a d-dimensional random vector $(X_1, X_2, ..., X_d)$ with uniform marginals on the unit cube $[0,1]^d$ [20]. According to Sklar's theorem [31], every multivariate distribution function $H(x_1, x_2, ..., x_d)$ can be expressed in terms of its marginals $F_1(x_1), F_2(x_2), ..., F_d(x_d)$ as

$$H(x_1, x_2, ..., x_d) = C(F_1(x_1), F_2(x_2), ..., F_d(x_d)). \tag{1}$$

Indeed, if H has a density h, then it relates to the density of the copula c, such that

$$h(x_1, ..., x_d) = c(F_1(x_1), ..., F_d(x_d))f_1(x_1)...f_d(x_d), \tag{2}$$

where $f_1, f_2, ..., f_d$ are the marginal densities of $X_1, X_2, ..., X_d$ respectively.

2.2 Extropy

Consider an observable quantity X with possible N values in the range $R(X) = (x_1, x_2, ..., x_N)$ and the probability vector $\mathbf{p}_N = (p_1, p_2, ..., p_N)$ represents probabilities associated with N possible occurrences in X. It is well known that the entropy function has been defined as $H(X) = H(p_N) = -\sum_{i=1}^{N} p_i \log(p_i)$. In this case, from Lad et al. [15], extropy is defined as follows:

Definition 2.1. The extropy in X or in p_N equals

$$J(X) = J(p_N) = -\sum_{i=1}^{N}(1-p_i)\log(1-p_i). \tag{3}$$

Assume another probability vector, $\mathbf{t}_N = (t_1, t_2, ..., t_N)$, distinct from $\mathbf{p}_N$, is defined on X, in this case, the relative extropy, which is defined as the function complement of the Kullback-Leibler divergence, is as follows [15]:

$$D_c(\mathbf{p}_N \| \mathbf{t}_N) = \sum_{i=1}^{N}(1-p_i)\log\left(\frac{1-p_i}{1-t_i}\right). \tag{4}$$

Definition 2.2. Consider a continuous random variable X with a probability density function $f(\cdot)$; the extropy of $f(\cdot)$ is defined as [15]:

$$j(X) = -\frac{1}{2}\int f^2(x)dx < 0. \tag{5}$$

Furthermore, the relative extropy between two density functions $f(\cdot)$ and $g(\cdot)$ defined on the interval $[a,b]$ can be defined as

$$d^c(f||g) = \frac{1}{2}\int_a^b [f(x) - g(x)]^2 dx. \tag{6}$$

In addition to the above definition, the concept of bivariate and multivariate extropy has been introduced in the literature by Balakrishnan et al. (2022) [2].

Definition 2.3. If X and Y are absolutely continuous random variables with joint density $f(x,y)$, the bivariate version of extropy is denoted by

$$j(X,Y) = \frac{1}{4}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f^2(x,y), dx, dy, \tag{7}$$

where $f(x,y)$ is the joint density of (X,Y) [28]. For a general d-dimensional vector,

$$j(X_1,\ldots,X_d) = \left(-\frac{1}{2}\right)^d \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f^2(x_1,\ldots,x_d), dx_1\cdots dx_d. \tag{8}$$

2.3 Neural network

The design of a multi-layer perceptron (MLP) is similar to the human brain's nervous system. It has a high ability to model nonlinear behaviors, which is very suitable for nonlinear prediction problems [1]. On the other hand, the method is such that it first learns the steps necessary to solve a problem and then discovers the relationships between these steps. To achieve this, of course, a large amount of data is given to the network in the training phase and many patterns are discovered by the network that are used to produce the correct output. The structure of the first layer of an MLP neural network is as

$$\mathbf{l}_1 = g_1(\mathbf{w}_{lk}\mathbf{k}_d + \mathbf{c}_1), \tag{9}$$

where $\mathbf{k}_d$ denotes the input vector containing d input variables, where d represents the dimensionality of the input space, and $\mathbf{w}_{lk}$ signifies the weights connecting the input layer to the first hidden layer's nodes. The activation function g_1 introduces non-linearity, and $\mathbf{c}_1$ represents the bias term. Subsequent hidden layers are determined using the following recursive relation:

$$\mathbf{l}_h = g_h(\mathbf{w}_{h,h-1}\mathbf{l}_{h-1} + \mathbf{c}_h) \quad h = 2,3,..., \tag{10}$$

where g_h denotes the activation function for layer h, $\mathbf{w}_{h,h-1}, h = 2,3,...$, represents the weights connecting the h-th and $(h-1)$-th layers, and $\mathbf{c}_h$ is the bias term for layer h. The network's output $\hat{\mathbf{y}}$ is generated from the final hidden layer through:

$$\hat{\mathbf{y}} = f_w(\mathbf{w}_{ly}\mathbf{l}_m + \mathbf{c}_y). \tag{11}$$

In this output equation, f_w is a potentially nonlinear output function, $\mathbf{w}_{ly}$ are the weights connecting the final hidden layer to the output node and $\mathbf{l}_m$ denotes the output of the final hidden layer, where m represents the index of the final hidden layer, and $\mathbf{c}_y$, denotes the bias term [21].

3 Suggested method for feature selection

In this section, we introduce a feature selection method based on extropy and copulas. Suppose the set of original features is represented as follows:

$$S = \{f_1, f_2, ..., f_p\},$$

where p denotes the total number of predictor variables. The goal is to perform a feature selection by identifying a subset of this set, denoted by s, as

$$s = \{f_1, f_2, ..., f_k\},$$

where $k < p$ such that the performance of the predictor variables selected in this set is better or similar to that of the variables in set S in predicting the target variable Y. As a result, the proposed method performs dimensionality reduction by reducing the original feature space from p dimensions to k dimensions while preserving the most relevant information associated with the target variable. Before introducing the FSEC method, we state the following theorem.

Theorem 3.1. *If $X_1, X_2, ..., X_d$ have distributions $F_{X_1}(x_1), F_{X_2}(x_2), ..., F_{X_d}(x_d)$ respectively and are linked through copula $C_{X_1,...,X_d}$, then the marginal-weighted copula extropy is computed as follows:*

$$j_{C_{X_1X_2\ldots X_d}} = \left(-\frac{1}{2}\right)^d \int_0^1\cdots\int_0^1 c^2(u_1,u_2,...,u_d) f_{X_1}(F_{X_1}^{-1}(u_1))\cdots f_{X_d}(F_{X_d}^{-1}(u_d)) du_1\cdots du_d, \tag{12}$$

where $u_1 = F_{X_1}(x_1), ..., u_d = F_{X_d}(x_d)$.

Proof. By extending the definition of extropy given in equation (8) to the support $\mathbb{R}^d$, we have

$$\begin{aligned} j_{C_{X_1X_2\ldots X_d}} &= \left(-\frac{1}{2}\right)^d \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f^2(x_1, x_2, ..., x_d)\, dx_1 \cdots dx_d \\ &= \left(-\frac{1}{2}\right)^d \int_0^1 \cdots \int_0^1 c^2(u_1, u_2, ..., u_d) f_{X_1}(F_{X_1}^{-1}(u_1)) \cdots f_{X_d}(F_{X_d}^{-1}(u_d))\, du_1 \cdots du_d. \end{aligned}$$

□

Now, if we want to write information theory based on extropy and copula, we have the following theorem.

Theorem 3.2. *If X and Y have distributions $F_X(x)$ and $F_Y(y)$, respectively and are linked through copula $C_{XY}(u, v)$, the amount of mutual information based on extropy and copula between these two variables is as follows*

$$I_{EC}(X;Y) = \frac{1}{4} \int_0^1 \int_0^1 (c(u,v) - 1)^2 f_X(F_X^{-1}(u)) f_Y(F_Y^{-1}(v)) du\, dv. \tag{13}$$

where I_{EC} denotes a measure for assessing mutual information extropy copula and here we will use it as an extropy copula-based mutual information.

Proof. Using Equations (2) and (6), we readily have

$$\begin{aligned} I_{EC}(X;Y) &= d^c(f_{X,Y} \| f_X f_Y) \\ &= \frac{1}{4} \int_{\mathbb{R}} \int_{\mathbb{R}} (f_{X,Y}(x,y) - f_X(x) f_Y(y))^2\, dx\, dy \\ &= \frac{1}{4} \int_{\mathbb{R}} \int_{\mathbb{R}} f_X^2(x) f_Y^2(y) (c(F_X(x), F_Y(y)) - 1)^2\, dx\, dy. \end{aligned}$$

Now, we define the transformation $u = F_X(x)$, $v = F_Y(y)$, which maps $\mathbb{R}^2$ into $(0,1)^2$, so that $du = f_X(x)\, dx$, $dv = f_Y(y)\, dy$, and equivalently

$$dx = \frac{du}{f_X(F_X^{-1}(u))}, \qquad dy = \frac{dv}{f_Y(F_Y^{-1}(v))}.$$

Substituting these expressions into the integral yields

$$I_{EC}(X;Y) = \frac{1}{4} \int_0^1 \int_0^1 (c(u,v) - 1)^2 f_X(F_X^{-1}(u)) f_Y(F_Y^{-1}(v))\, du\, dv,$$

finishing the proof. □

In the existing literature, including Saha and Kayal (2026) [28], the $J_c(X,Y)$ (Copula extropy) measure has been introduced as a measure based on the density of the copula in the uniform space $(0,1)^2$. This measure mainly describes the intensity of the dependence between variables through the copula structure and does not consider the effect of marginal distributions. In contrast, the novelty of our paper is to present a new copula-based information framework that is developed in line with the concept of "distance from independence" and inspired by the concept of mutual information. In this approach, instead of focusing solely on the copula structure, the proposed measure measures the degree to which the joint distribution deviates from the state of independence. This formulation allows for the simultaneous combination of the copula structure and marginal information through a weighting.

From a theoretical perspective, it is shown that if the variables are transformed into a copula space with uniform marginal distributions on the interval $(0,1)$, the proposed measure $I_{EC}(X;Y)$ is an affine transformation of the classical extropy copula measure $J_c(X,Y)$ according to Saha and Kayal (2026) [28], as established in the following lemma. Therefore, it can be concluded that I_{EC} is a generalization of the extropy copula measures available in the literature (see the proof of Theorem 3.2).

Lemma 3.3. *If X and Y have uniform marginal distributions on $(0,1)$, then*

$$I_{EC}(X;Y) = J_c(X,Y) - \frac{1}{4}. \tag{14}$$

Consequently, $I_{EC}(X;Y)$ and $J_c(X,Y)$ provide the same ranking of features.

Proof. Since X and Y have uniform marginal distributions on $(0,1)$, we have $f_X(F_X^{-1}(u)) = f_Y(F_Y^{-1}(v)) = 1$. Thus,

$$\begin{aligned} I_{EC}(X;Y) &= \frac{1}{4}\int_0^1\int_0^1 (c(u,v)-1)^2\,du\,dv \\ &= \frac{1}{4}\int_0^1\int_0^1 c^2(u,v)\,du\,dv - \frac{1}{4}, \end{aligned}$$

where $\int_0^1\int_0^1 c(u,v)\,du\,dv = 1$ and $J_c(X,Y) = \frac{1}{4}\int_0^1\int_0^1 c^2(u,v)\,du\,dv$. We obtain

$$I_{EC}(X;Y) = J_c(X,Y) - \frac{1}{4}.$$

□

In a feature selection method based on copula and extropy function, the main goal is to identify a set of features that have the highest dependency on the target variable. In this regard, a feature f_i is considered more relevant to Y than f_j if f_i has more mutual information extropy copula with Y compared to f_j. Mathematically speaking

$$I_{EC}(f_i;Y) > I_{EC}(f_j;Y). \tag{15}$$

This principle is considered the relevancy test for features.

Maximum-Relevancy-Extropy-Copula (MREC)

The first step of the proposed method focuses on identifying the variables that show the strongest dependence on the response variable. Accordingly, the maximum relevance criterion can be expressed as:

$$s_{MREC} = \arg\max_{f_i \in S} I_{EC}(f_i, Y), \tag{16}$$

where $S = \{f_1, f_2, ..., f_p\}$ represents the complete set of available features. The feature that produces the highest dependency value is selected as the most relevant variable.

Minimum-Redundancy-Extropy-Copula (mREC)

Selecting features solely based on their relevance may result in including variables that carry similar information. Therefore, in addition to maximizing relevance, it is necessary to control redundancy among the selected features.

Suppose that at a given iteration, the set of selected variables is $S_{sel} = \{f_1, f_2, ..., f_m\}$. For each remaining candidate feature $f_i \in (S \setminus S_{sel})$, its association with the target variable is defined by $R(f_i) = I_{EC}(f_i, Y)$, while its average redundancy with the currently selected variables is defined as

$$D(f_i) = \frac{1}{|S_{sel}|}\sum_{f_j \in S_{sel}} I_{EC}(f_i, f_j). \tag{17}$$

To simultaneously consider these two criteria, the following evaluation function is introduced:

$$Q(f_i) = R(f_i) - \alpha D(f_i), \tag{18}$$

where α controls the degree of redundancy penalty. Then, the feature that provides the largest score is selected based on

$$f^\star = \arg\max_{f_i \in (S \setminus S_{sel})} Q(f_i). \tag{19}$$

At convergence, the final selected set is denoted by $s = S_{sel}$, where $|S_{sel}| = k$. Hence, the proposed framework seeks variables that are highly informative about the target variable while containing the least overlapping information with the pre-selected features. The term $R(f_i)$ encourages the selection of relevant variables, while $D(f_i)$ discourages redundant variables. This iterative procedure continues until

$$\frac{Q^\star}{R_{\max}} < \tau, \tag{20}$$

where $Q^\star$ is the highest score among the candidate features, and $R_{\max}$ is the maximum extropy-copula relevance value. This stopping rule prevents the inclusion of variables with poor information and leads to a compact subset of selected features.

The main novelty of this paper is the introduction of a feature selection algorithm based on the proposed I_{EC} criterion. This algorithm has significant advantages over existing methods and overcomes some of the limitations of previous methods. In particular, this algorithm is designed to automatically determine and output the optimal number of selected features for each dataset. In fact, this method can simultaneously model the dependence between variables through the copula structure and include information about the marginal functions of the variables in the feature selection process. As a result, this algorithm selects a set of features that, while having a smaller size, have the largest amount of information related to the target variable and can help predict the target variable with higher accuracy.

In this section, we introduce our feature selection algorithm based on extropy and copula functions in detail. We first prove some theoretical concepts.

Consider two random variables X and Y are related by a copula function, C_{XY}. Inspired by Nelsen (2006) [20], if we apply functions γ and η to X and Y respectively, that creating $\gamma(X)$ and $\eta(Y)$, then the copula between $\gamma(X)$ and $\eta(Y)$ can be expressed as follows

- for both strictly increasing functions γ and η,

$$C_{\gamma(X)\eta(Y)}(u,v) = C_{XY}(u,v); \tag{21}$$

- for strictly increasing function γ and strictly decreasing function η,

$$C_{\gamma(X)\eta(Y)}(u,v) = u - C_{XY}(u,1-v); \tag{22}$$

- for strictly decreasing function γ and strictly increasing function η,

$$C_{\gamma(X)\eta(Y)}(u,v) = v - C_{XY}(1-u,v); \tag{23}$$

- for both strictly decreasing functions γ and η,

$$C_{\gamma(X)\eta(Y)}(u,v) = v + u - 1 - C_{XY}(1-u,1-v). \tag{24}$$

The above statement is used to examine the behavior of I_{EC} under monotone transformations. Although the copula remains invariant under strictly increasing transformations, I_{EC} is generally not invariant due to its dependence on marginal densities.

Theorem 3.4. *Let γ and η be strictly increasing differentiable functions applied to continuous random variables X and Y, respectively. Then the copula remains invariant under these transformations. However, the mutual information based on extropy and copula I_{EC} is generally not invariant under such transformations. In particular, if $\gamma(X) = aX$ and $\eta(Y) = bY$, where $a, b > 0$, then*

$$I_{EC_{aX,bY}} = \frac{1}{ab} I_{EC_{XY}}. \tag{25}$$

Proof. Moreover, using Equation (21) and the marginal densities under these transformations, which are given by

$$f_{\gamma(X)}\left(F^{-1}_{\gamma(X)}(u)\right) = \frac{f_X\left(F_X^{-1}(u)\right)}{\gamma'\left(F_X^{-1}(u)\right)},$$

and

$$f_{\eta(Y)}\left(F^{-1}_{\eta(Y)}(v)\right) = \frac{f_Y\left(F_Y^{-1}(v)\right)}{\eta'\left(F_Y^{-1}(v)\right)}.$$

Therefore, using Equation (13), we obtain

$$I_{EC_{\gamma(X),\eta(Y)}} = \frac{1}{4}\int_0^1\int_0^1 (c_{XY}(u,v)-1)^2 \frac{f_X\left(F_X^{-1}(u)\right) f_Y\left(F_Y^{-1}(v)\right)}{\gamma'\left(F_X^{-1}(u)\right)\eta'\left(F_Y^{-1}(v)\right)}\,du\,dv.$$

In particular, for the linear transformations $\gamma(X) = aX$ and $\eta(Y) = bY$, where $a, b > 0$, we have

$$\gamma'(x) = a, \qquad \eta'(y) = b.$$

Consequently,

$$I_{EC_{aX,bY}} = \frac{1}{4ab}\int_0^1\int_0^1 (c_{XY}(u,v)-1)^2 f_X\left(F_X^{-1}(u)\right) f_Y\left(F_Y^{-1}(v)\right)\,du\,dv.$$

Therefore,

$$I_{EC_{aX,bY}} = \frac{1}{ab} I_{EC_{XY}}.$$

Hence, I_{EC} is not invariant under linear rescaling, although the copula remains invariant under strictly increasing transformations. □

Corollary 3.5. *If $\gamma(X) = X + a$ and $\eta(Y) = Y + b$, where $a, b \in \mathbb{R}$, then $\gamma'(x) = \eta'(y) = 1$. Therefore,*

$$I_{EC_{X+a,Y+b}} = I_{EC_{XY}}. \tag{26}$$

Thus, I_{EC} is invariant under translations, although it is not invariant under linear rescaling.

Proof. The proof is immediate. □

Theorem 3.6. *Let γ be a strictly increasing differentiable function and η be a strictly decreasing differentiable function applied to continuous random variables X and Y, respectively. Then the mutual information based on extropy and copula I_{EC} is generally not invariant under these transformations. In particular,*

$$I_{EC_{\gamma(X),\eta(Y)}} = \frac{1}{4}\int_0^1\int_0^1 (c_{XY}(u,v)-1)^2 \frac{f_X(F_X^{-1}(u))f_Y(F_Y^{-1}(v))}{\gamma'(F_X^{-1}(u))\left|\eta'(F_Y^{-1}(v))\right|}\,du\,dv. \tag{27}$$

Proof. Using Equation (22) and the marginal densities under these transformations, which are given by

$$f_{\gamma(X)}\left(F^{-1}_{\gamma(X)}(u)\right) = \frac{f_X\left(F_X^{-1}(u)\right)}{\gamma'\left(F_X^{-1}(u)\right)},$$

and

$$f_{\eta(Y)}\left(F^{-1}_{\eta(Y)}(v)\right) = \frac{f_Y\left(F_Y^{-1}(1-v)\right)}{\left|\eta'\left(F_Y^{-1}(1-v)\right)\right|}.$$

Therefore, using Equation (13), we obtain

$$I_{EC_{\gamma(X),\eta(Y)}} = \frac{1}{4}\int_0^1\int_0^1 (c_{XY}(u,1-v)-1)^2 \frac{f_X(F_X^{-1}(u))f_Y(F_Y^{-1}(1-v))}{\gamma'(F_X^{-1}(u))\left|\eta'(F_Y^{-1}(1-v))\right|}\,du\,dv.$$

Now set $m = 1 - v$, so that $dm = -dv$ while u remains unchanged. Since this transformation is bijective on $(0,1)$, it preserves the integration domain. Therefore,

$$I_{EC_{\gamma(X),\eta(Y)}} = \frac{1}{4}\int_0^1\int_0^1 (c_{XY}(u,m)-1)^2 \frac{f_X(F_X^{-1}(u))f_Y(F_Y^{-1}(m))}{\gamma'(F_X^{-1}(u))\left|\eta'(F_Y^{-1}(m))\right|}\,du\,dm.$$

Rewriting the dummy variable m as v, we obtain

$$I_{EC_{\gamma(X),\eta(Y)}} = \frac{1}{4}\int_0^1\int_0^1 (c_{XY}(u,v)-1)^2 \frac{f_X(F_X^{-1}(u))f_Y(F_Y^{-1}(v))}{\gamma'(F_X^{-1}(u))\left|\eta'(F_Y^{-1}(v))\right|}\,du\,dv.$$

Hence, in general,

$$I_{EC_{\gamma(X),\eta(Y)}} \neq I_{EC_{XY}}.$$

Thus, I_{EC} is generally not invariant under transformations in which one variable is transformed by a strictly increasing function and the other by a strictly decreasing function. □

Theorem 3.7. *Let γ be a strictly decreasing differentiable function and η be a strictly increasing differentiable function applied to continuous random variables X and Y, respectively. Then the mutual information based on extropy and copula I_{EC} is generally not invariant under these transformations. In particular,*

$$I_{EC_{\gamma(X),\eta(Y)}} = \frac{1}{4}\int_0^1\int_0^1 (c_{XY}(u,v)-1)^2 \frac{f_X(F_X^{-1}(u))f_Y(F_Y^{-1}(v))}{\left|\gamma'(F_X^{-1}(u))\right|\eta'(F_Y^{-1}(v))}\,du\,dv. \tag{28}$$

Proof. The proof is similar to that of Theorem 3.6. □

Theorem 3.8. *Let γ and η be strictly decreasing differentiable functions applied to continuous random variables X and Y, respectively. Then the copula transforms according to the corresponding reflection transformations. However, the mutual information based on extropy and copula I_{EC} is generally not invariant under such transformations. In particular, if $\gamma(X) = -aX$ and $\eta(Y) = -bY$, where $a, b > 0$, then*

$$I_{EC_{-aX,-bY}} = \frac{1}{ab} I_{EC_{XY}}. \tag{29}$$

Proof. The proof is similar to that of Theorem 3.4. □

3.1 Dimensionality reduction mechanism

Although the proposed method does not perform dimensionality reduction through linear visualization techniques, it achieves dimensionality reduction through feature subset selection. Specifically, from an original feature space containing p variables, the proposed MREC and mREC criteria identify a subset of informative variables with size $k < p$. As a result, the data representation is reduced from $\mathbb{R}^p$ to $\mathbb{R}^k$, while preserving the most relevant dependency information with respect to the target variable. The reduction process is achieved by maximizing relevance and minimizing redundancy, along with a stopping criterion $Q^\star < \tau R_{\max}$, which avoids the inclusion of features with weak information, where X and Y are related by a copula function, C_{XY}.

3.1.1 Dimensionality reduction rate

Dimensionality reduction rate (DR) is used to measure the achieved dimensionality reduction expressed as

$$DR = \frac{p-k}{p} \times 100, \tag{30}$$

where p denotes the number of original features and k denotes the number of selected features. A larger value of DR indicates a greater reduction in the dimensionality of the original feature space.

3.2 Theoretical properties of IEC

3.2.1 Property 1 (Symmetry)

For any two random variables X and Y,

$$I_{EC}(X;Y) = I_{EC}(Y;X). \tag{31}$$

Proof. This follows directly from the symmetric form of Equation (13). □

3.2.2 Property 2 (Independence)

If X and Y are independent, then

$$I_{EC}(X;Y) = 0. \tag{32}$$

Proof. According to Sklar's theorem, if the two variables X and Y are independent, then we have:

$$C(u,v) = uv,$$

then

$$c(u,v) = 1.$$

Substituting into (13) gives

$$I_{EC}(X;Y) = 0.$$

Therefore, I_{EC} correctly identifies complete independence. □

3.2.3 Property 3 (Non-negativity)

For any pair of random variables X and Y,

$$I_{EC}(X;Y) \geq 0. \tag{33}$$

Since

$$(c(u,v)-1)^2 \geq 0,$$

and

$$f_X(\cdot) \geq 0, \qquad f_Y(\cdot) \geq 0,$$

the integrand in (13) is nonnegative, implying that I_{EC} is always nonnegative.

3.2.4 Property 4 (Sensitivity to Dependence Strength)

If two variables are independent, then $I_{EC}(X;Y) = 0$. Therefore, as the dependence between two variables increases, and given that the terms in the I_{EC} formula are nonnegative, the value of $I_{EC}(X;Y)$ also increases. As a result, the $I_{EC}(X;Y)$ measure has an increasing copula-based relationship with the intensity of the dependence between the variables.

Algorithm 1 Feature Selection using Extropy-Copula (FSEC)

Input: $S = \{f_1, \ldots, f_p\}$, target variable Y, redundancy parameter α, stopping parameter τ

Output: S_{sel}

Standardize predictors: $f_i^* \leftarrow \dfrac{f_i - \bar{f}_i}{\sigma_{f_i}}, \quad i = 1, \ldots, p$

Initialization: $S_{sel} \leftarrow \emptyset$

Compute initial relevance:

for $i = 1$ **to** p **do**

  $R(f_i^*) \leftarrow I_{EC}(f_i^*, Y)$

$R_{\max} \leftarrow \max_{f_i \in S} R(f_i^*)$

Select first feature: $f^\dagger \leftarrow \arg\max_{f_i \in S} R(f_i^*)$

$S_{sel} \leftarrow S_{sel} \cup \{f^\dagger\}$

$Q^\dagger \leftarrow R_{\max}$

Iterative selection:

while $S \setminus S_{sel} \neq \emptyset$ ***and*** $\dfrac{Q^\dagger}{R_{max}} \geq \tau$ **do**

  for $f_i \in (S \setminus S_{sel})$ **do**

    $D(f_i^*) \leftarrow \dfrac{1}{|S_{sel}|} \sum_{f_j \in S_{sel}} I_{EC}(f_i^*, f_j^*)$

    $Q(f_i^*) \leftarrow R(f_i^*) - \alpha D(f_i^*)$

  $f^\dagger \leftarrow \arg\max_{f_i \in (S \setminus S_{sel})} Q(f_i^*)$

  $Q^\dagger \leftarrow Q(f^{\dagger *})$

  if $\dfrac{Q^\dagger}{R_{max}} < \tau$ **then**

    break

  $S_{sel} \leftarrow S_{sel} \cup \{f^\dagger\}$

return S_{sel}

After the FSEC algorithm determines the feature set, a regression-based neural network is employed to predict the response variable, as the target variable Y is continuous.

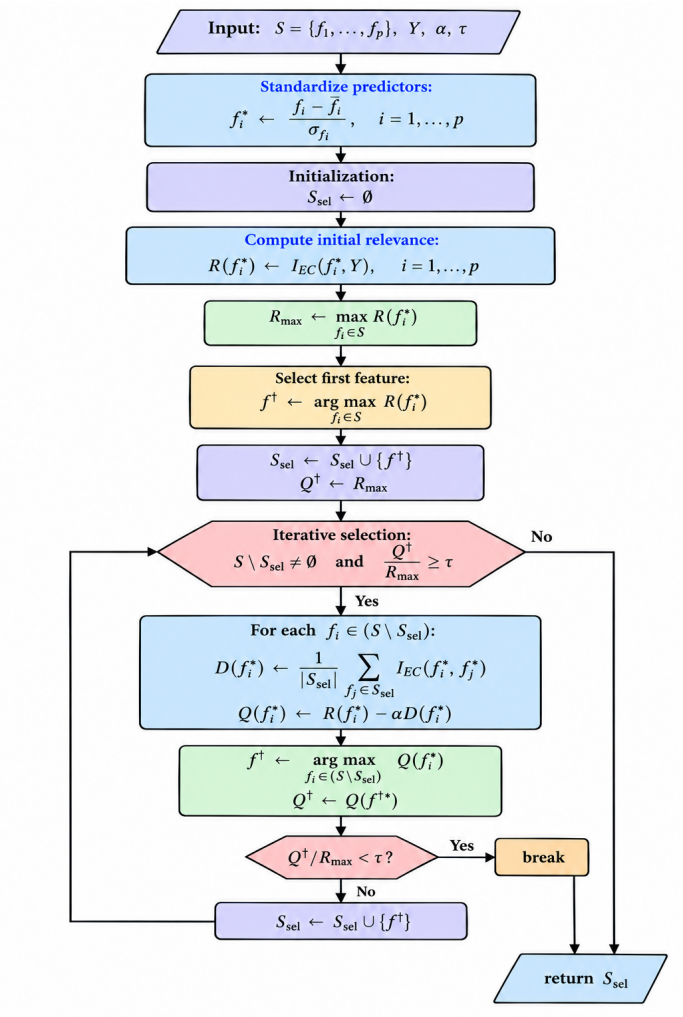


Figure 1: Overview of the proposed Feature Selection using Extropy Copula (FSEC) framework.

3.3 Numerical study

In this section, the performance of the proposed extropy and copula-based feature selection framework is investigated through three simulation scenarios specifically designed to evaluate the efficiency of the method in different data structures. In the first scenario, the ability of the proposed method to identify effective predictor variables in the presence of redundant and noisy variables is evaluated. In the second scenario, the performance of the method in high-dimensional data with a large number of features is studied. In the third scenario, the proposed method is evaluated under different dependency structures generated by different copula families. In addition to simulation studies, the proposed FSEC algorithm is also implemented on several real datasets including Microbe, Breast Cancer, and Compressive Strength data to evaluate its practical performance and compare it with other feature selection methods.

For copula estimation and model selection, the VineCopula package in R was employed. In particular, the BiCopSelect procedure was used to automatically select the most appropriate copula family according to the Akaike Information Criterion (AIC), while copula parameters were estimated using the Maximum Likelihood Estimation (MLE) method.

After feature selection, prediction performance was evaluated using artificial neural networks. Model evaluation was performed using five-fold cross-validation to assess the stability, robustness, and generalization ability of the competing feature-selection methods. In each fold, feature selection was performed using the training data only, and the selected features were subsequently used to construct the prediction model, which was then evaluated on the held-out test data. The proposed FSEC method was compared with four widely used benchmark approaches:

1. Copula-Based Feature Selection (CBFS)
2. LASSO

3. Random Forest (RF)
4. Minimum Redundancy Maximum Relevance (mRMR).

For a fair comparison, the benchmark methods were configured to select the same number of features as the proposed FSEC procedure in the main simulation experiments.

3.4 Performance measures

1. The Root Mean Squared Error (RMSE) is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

2. The Mean Absolute Error (MAE) is

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|.$$

3. The coefficient of determination is

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

3.5 Parameter sensitivity analysis and selection

The parameter α controls the trade-off between feature relevance and redundancy, whereas τ determines the stopping criterion and therefore influences the final size of the selected feature subset. To investigate the impact of these tuning parameters, a sensitivity analysis was conducted by considering different combinations of α and τ. For each parameter combination, the proposed FSEC procedure was evaluated in terms of predictive performance and the number of selected features.

3.6 Simulation study

In this section, we examine the method proposed in this article using three simulated scenarios.

3.6.1 Scenario 1: Noisy data with nonlinear signal structure

The first simulation scenario was designed to evaluate the ability of the proposed FSEC method to identify informative variables in the presence of nonlinear relationships, correlated predictors, and irrelevant noise variables. A dataset with $n = 500$ observations and $p = 50$ predictor variables was generated. Among these variables, X_1, X_2, X_3, and X_4 were specified as the truly informative variables, while X_5 and X_6 were introduced as highly correlated redundant predictors. Specifically, X_5 and X_6 were generated to have a correlation of 0.95 with X_1 and X_2, respectively. The remaining variables, $X_7, \ldots, X_{50}$, were generated as irrelevant noise variables. All predictor variables were initially generated from standard normal distributions.

The response variable was generated according to the following nonlinear model:

$$Y = 8X_1^2 + 6\sin(X_2) + 5X_3 + 3X_4^2 + \varepsilon, \qquad \varepsilon \sim N(0, 0.75^2).$$

This response structure incorporates both nonlinear and linear effects. In particular, X_1 contributes through a quadratic effect, X_2 through a sinusoidal transformation, X_3 through a linear effect, and X_4 through an additional quadratic effect. Thus, the scenario provides a challenging setting for feature selection methods that must identify nonlinear and linear signal components in the presence of correlated and irrelevant predictors.

The generated dataset was randomly divided into training, validation, and test sets containing 60%, 20%, and 20% of the observations, respectively. The training set was used for feature selection and model fitting, while the validation set was used for tuning the FSEC parameters. The independent test set was not used during feature selection, parameter tuning, or model development and was reserved exclusively for the final evaluation of predictive performance.

The FSEC algorithm was applied according to the procedure described in Algorithm 1, including the required preprocessing and standardization steps. In the sensitivity analysis, two tuning parameters, α and τ, were varied over the following grids:

$$\alpha \in \{0.05, 0.10, 0.20, 0.50\}, \qquad \tau \in \{0.05, 0.10\}.$$

Thus, eight combinations of (α, τ) were considered. For each parameter combination, FSEC was run independently on the training data. The number of selected variables, denoted by k, was not fixed in advance; instead, it was automatically determined by FSEC according to its stopping rule. The selected variables obtained under each parameter combination were then used to train a neural network on the training set, and the fitted model was evaluated on the validation set. The validation RMSE was used as the primary criterion for comparing the parameter configurations, while validation MAE and R^2 were also recorded.

Table 1 summarizes the sensitivity analysis and reports the evaluated configurations in ascending order of validation RMSE. The configuration $(\alpha, \tau) = (0.05, 0.05)$ achieved the minimum validation RMSE among the evaluated configurations, with $\text{RMSE}_{\text{VALID}} = 0.95$, $\text{MAE}_{\text{VALID}} = 0.78$, and $R^2_{\text{VALID}} = 0.99$. Under this parameter configuration, FSEC automatically selected $k = 4$ variables. Therefore, the final number of selected variables was determined by FSEC rather than being specified as an input parameter.

Table 1: Sensitivity analysis of FSEC parameters for Scenario 1, ordered by validation RMSE.

α	τ	k	$\text{RMSE}_{\text{VALID}}$	$\text{MAE}_{\text{VALID}}$	R^2_{VALID}
0.05	0.05	4	0.95	0.78	0.99
0.20	0.05	4	0.96	0.78	0.99
0.20	0.10	4	0.96	0.78	0.99
0.50	0.05	2	3.97	3.30	0.88
0.50	0.10	2	3.97	3.30	0.88
0.05	0.10	3	4.24	3.46	0.87
0.10	0.05	3	4.24	3.46	0.87
0.10	0.10	2	7.81	5.93	0.54

Using the selected parameter configuration, FSEC reduced the dimensionality from 50 variables to 4 selected variables, corresponding to a 92% reduction in the number of predictors. The selected variables were X_1, X_4, X_3, and X_2. Thus, FSEC successfully identified all four truly informative variables, namely X_1, X_2, X_3, and X_4.

For a fair comparison, the final number of variables selected by FSEC was subsequently used as the target selection size for the competing methods. Accordingly, CBFS, LASSO, random forest (RF), and mRMR were constrained to select exactly four variables. Importantly, this constraint was imposed only after FSEC had determined its final number of selected variables and was not used as an input to FSEC.

Table 2: Selected variables by the feature selection methods in Scenario 1.

Method	Selected Features	k
FSEC	X_1, X_4, X_3, X_2	4
CBFS	X_3, X_2, X_8, X_{40}	4
LASSO	X_2, X_3, X_{34}, X_{40}	4
RF	X_1, X_4, X_2, X_5	4
mRMR	X_1, X_2, X_3, X_{19}	4

To evaluate predictive performance, the selected features from each method were used as inputs to a neural network model with 10 hidden neurons. The final neural network models were fitted using the combined training and validation sets after the feature selection and parameter-tuning stages had been completed. Predictive performance was then evaluated on both the combined training–validation set and the independent test set using RMSE, MAE, and R^2.

As shown in Table 3, FSEC achieved the best predictive performance overall. It obtained RMSE values of 0.71 and 0.75 on the combined training–validation and test sets, respectively, with corresponding R^2 values of 0.99 and 0.99. The MAE values were 0.56 and 0.54, indicating stable predictive performance across the two evaluation sets. mRMR produced a comparable training–validation performance, but its test RMSE was slightly higher than that of FSEC.

In contrast, RF showed a substantial performance gap between the combined training–validation and test sets, with RMSE increasing from 5.34 to 7.82 and R^2 decreasing from 0.82 to 0.62. CBFS and LASSO exhibited considerably

Table 3: Predictive performance of feature selection methods for Scenario 1.

Method	$\text{RMSE}_{\text{TRAIN+VALID}}$	$\text{RMSE}_{\text{TEST}}$	$\text{MAE}_{\text{TRAIN+VALID}}$	MAE_{TEST}	$R^2_{\text{TRAIN+VALID}}$	R^2_{TEST}
FSEC	0.71	0.75	0.56	0.54	0.99	0.99
CBFS	8.99	13.59	6.41	10.36	0.50	-0.12
LASSO	8.04	19.06	5.80	11.83	0.60	-1.22
RF	5.34	7.82	3.94	5.47	0.82	0.62
mRMR	0.71	0.76	0.56	0.55	0.99	0.99

poorer generalization, with test-set R^2 values of -0.12 and -1.22, respectively. Overall, FSEC provided the lowest test RMSE and one of the highest test R^2 values, demonstrating its ability to identify a compact set of informative predictors while maintaining strong predictive generalization in the presence of nonlinear effects, correlated variables, and irrelevant noise.

3.6.2 Scenario 2: High-dimensional data with nonlinear signal structure

The second simulation scenario was designed to evaluate the performance of the proposed FSEC method in a high-dimensional setting characterized by nonlinear signal structure, redundant predictors, and a large number of irrelevant variables. A dataset with $n = 1200$ observations and $p = 200$ predictor variables was generated. The first ten variables, $X_1, \cdots, X_{10}$, were specified as the truly informative variables. The variables $X_{11}, \cdots, X_{20}$ were generated as noisy surrogate variables, with each surrogate variable correlating 0.65 with its corresponding informative variable. The remaining variables, $X_{21}, \cdots, X_{200}$, were generated as pure noise variables.

The predictor variables were initially generated from a standard normal distribution. The response variable was generated according to

$$\begin{aligned} Y = {} & 7\sin(X_1) + 6.5\cos(X_2) + 6X_3^2 + 5.5\exp(-0.5X_4^2) \\ & + 5\log(X_5^2 + 1) + 4.5\sin(1.5X_6) + 4X_7^2 \\ & + 3.5\cos(1.5X_8) + 3\sin(X_9) + 2.5X_{10}^2 + \varepsilon, \end{aligned} \tag{1}$$

where $\varepsilon \sim N(0, 0.3^2)$. This response structure incorporates several nonlinear functional forms, including sinusoidal, cosine, quadratic, exponential, and logarithmic effects, providing a challenging setting for evaluating feature selection methods in the presence of redundant and irrelevant predictors.

The generated dataset was randomly divided into training, validation, and test sets using a 60%–20%–20% split. The training set was used for feature selection and model fitting, whereas the validation set was used for tuning the FSEC parameters. The independent test set was kept separate from the feature selection and parameter-tuning procedures and was used only for the final evaluation of predictive performance.

The sensitivity analysis was conducted by considering combinations of the two FSEC tuning parameters, α and τ. Specifically, nine values of α and fifteen values of τ were considered:

$$\alpha \in \{0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80\}, \tag{2}$$

$$\tau \in \{0.001, 0.005, 0.010, 0.020, 0.030, 0.050, 0.075, 0.100, 0.150, 0.200, 0.250, 0.300, 0.400, 0.500, 0.600\}. \tag{3}$$

These values resulted in a total of $9 \times 15 = 135$ (α, τ) combinations. For each (α, τ) combination, the FSEC algorithm was applied independently to the training data. Importantly, the number of selected variables, k, was not specified as an input parameter. Instead, FSEC determined both the selected feature subset and the corresponding value of k according to its stopping rule for each parameter combination. The selected features were then used to fit a neural network using the training data, and the fitted model was evaluated on the validation set. The validation RMSE was used as the primary criterion for tuning α and τ, while validation MAE and R^2 were also recorded. The resulting value of k was retained for assessing the dimensionality reduction achieved by FSEC.

For concise presentation, Table 4 reports representative parameter combinations and their corresponding validation results, ordered from the lowest to the highest validation RMSE. The table is intended to summarize the behavior of the sensitivity analysis rather than to report all 135 parameter combinations.

Table 4: Summary of representative sensitivity results for FSEC in Scenario 2, ordered by validation RMSE.

α	τ	k	$\text{RMSE}_{\text{VALID}}$	$\text{MAE}_{\text{VALID}}$	R^2_{VALID}
0.05	0.05	10	4.48	3.45	0.88
0.40	0.10	10	4.49	3.47	0.87
0.50	0.20	10	4.49	3.47	0.87
0.20	0.30	10	4.50	3.47	0.87
0.60	0.05	10	4.51	3.49	0.86
0.50	0.30	7	4.81	3.84	0.85
0.30	0.30	8	4.84	3.86	0.84
0.60	0.30	6	4.99	4.00	0.83
0.05	0.40	4	6.28	4.80	0.74
0.20	0.40	4	6.28	4.80	0.74
0.50	0.40	4	6.28	4.80	0.74
0.05	0.50	2	8.54	6.62	0.51
0.05	0.60	2	8.54	6.62	0.51
0.20	0.50	2	8.54	6.62	0.51
0.20	0.60	2	8.54	6.62	0.51

As shown in Table 4, the number of selected variables varied from 2 to 10 across the representative combinations of α and τ. These results show that the FSEC stopping rule responded to changes in the tuning parameters and consequently produced different feature subset sizes. In general, configurations producing smaller feature subsets were associated with larger validation RMSE values, whereas configurations retaining more variables generally achieved better validation performance. For example, $(\alpha, \tau) = (0.05, 0.50)$ resulted in $k = 2$ and a validation RMSE of 8.54, whereas $(\alpha, \tau) = (0.05, 0.05)$ resulted in $k = 10$ and the lowest reported validation RMSE of 4.48.

The minimum validation RMSE observed in the sensitivity analysis was 4.48, obtained for $(\alpha, \tau) = (0.05, 0.05)$ with $k = 10$ selected variables. This configuration also yielded an MAE of 3.45 and an R^2 of 0.88 on the validation set. Accordingly, $(0.05, 0.05)$ was retained for the subsequent feature-selection comparison. Under this setting, FSEC selected ten variables according to its stopping rule, reducing the dimensionality from 200 to 10 predictors, corresponding to a 95% reduction in the original feature space.

Table 5: Selected features obtained by different feature selection methods in Scenario 2.

Method	Selected Features	k
FSEC	$X_3, X_1, X_7, X_6, X_{10}, X_{13}, X_{11}, X_8, X_2, X_4$	10
CBFS	$X_1, X_{11}, X_6, X_{147}, X_9, X_{19}, X_{162}, X_{93}, X_{72}, X_{16}$	10
LASSO	$X_1, X_6, X_9, X_{147}, X_{162}, X_{58}, X_{40}, X_{28}, X_{185}, X_{93}$	10
RF	$X_3, X_1, X_7, X_6, X_{10}, X_{11}, X_8, X_{19}, X_9, X_{13}$	10
mRMR	$X_1, X_6, X_9, X_{162}, X_{147}, X_{11}, X_{93}, X_{72}, X_{28}, X_{58}$	10

The selected features obtained by the different feature selection methods are presented in Table 5. FSEC selected eight of the ten truly informative variables, namely X_1, X_2, X_3, X_4, X_6, X_7, X_8, and X_{10}. It also selected two surrogate variables, X_{11} and X_{13}. Thus, FSEC recovered 80% of the truly informative variables while reducing the feature space from 200 to 10 predictors. RF selected seven of the ten truly informative variables, namely X_1, X_3, X_6, X_7, X_8, X_9, and X_{10}, corresponding to a recovery rate of 70%. In comparison, CBFS, LASSO, and mRMR each selected three of the ten truly informative variables, corresponding to a recovery rate of 30%. These methods also included several variables from the noise region, indicating a weaker ability to distinguish informative predictors from irrelevant variables in the high-dimensional setting.

For a fair comparison, after a representative FSEC parameter combination among those achieving the minimum validation RMSE had been identified and the corresponding number of selected variables had been obtained as $k = 10$, the competing methods CBFS, LASSO, Random Forest, and mRMR were constrained to select exactly ten variables. This constraint was imposed only on the competing methods and was not provided to FSEC as an input. Each competing method independently selected its own best ten predictors according to its corresponding selection criterion.

To evaluate predictive performance, the selected features from each method were used as inputs to a neural network model with 15 hidden neurons. The final neural network models were fitted using the combined training and validation sets after completion of the feature selection and parameter-tuning stages. Predictive performance was evaluated on both the combined training–validation set and the independent test set using RMSE, MAE, and R^2. The results are

reported in Table 6.

Table 6: Predictive performance of feature selection methods for Scenario 2.

Method	$\text{RMSE}_{\text{TRAIN+VALID}}$	$\text{RMSE}_{\text{TEST}}$	$\text{MAE}_{\text{TRAIN+VALID}}$	MAE_{TEST}	$R^2_{\text{TRAIN+VALID}}$	R^2_{TEST}
FSEC	4.75	6.58	3.59	5.17	0.86	0.79
CBFS	8.24	14.07	6.16	11.00	0.57	0.02
LASSO	8.29	14.74	6.23	11.24	0.56	-0.07
RF	5.08	8.16	3.89	6.14	0.84	0.67
mRMR	8.24	14.75	6.22	11.11	0.57	-0.07

As shown in Table 6, FSEC achieved the best predictive performance among the evaluated feature selection methods. Its RMSE increased from 4.75 on the combined training–validation set to 6.58 on the independent test set, while the corresponding MAE increased from 3.59 to 5.17. The R^2 value decreased from 0.86 to 0.79, indicating comparatively good generalization performance. RF achieved the second-best predictive performance, with RMSE values of 5.08 and 8.16 and R^2 values of 0.84 and 0.67 on the combined training–validation and test sets, respectively. In contrast, CBFS, LASSO, and mRMR exhibited substantially poorer generalization performance, with test-set R^2 values of 0.02, -0.07, and -0.07, respectively.

Overall, FSEC provided the lowest test-set RMSE and MAE and the highest test-set R^2 among all evaluated methods. These results demonstrate that FSEC can effectively reduce a high-dimensional feature space while retaining a substantial proportion of the underlying nonlinear signal and achieving comparatively strong predictive generalization.

3.6.3 Scenario 3: Mixed copula dependency structure

The third simulation scenario was designed to evaluate the performance of FSEC under a heterogeneous dependence structure generated from multiple copula families. A total of $N = 500$ observations and $P = 40$ predictor variables were generated, including eight truly informative variables and 32 independent Gaussian noise variables. The informative variables were generated in four dependent pairs: (X_1, X_2) from a Gaussian copula with $\rho = 0.60$, (X_3, X_4) from a Clayton copula with $\theta = 2.00$, (X_5, X_6) from a Gumbel copula with $\theta = 2.00$, and (X_7, X_8) from a Frank copula with $\theta = 5.00$. The remaining variables, $X_9, \cdots, X_{40}$, were generated independently from a standard normal distribution.

The response variable was generated from a nonlinear combination of marginal and interaction effects involving the eight informative variables. Specifically, the response was defined as

$$Y = 10X_1X_2 + 8X_2^3 + 6X_4 + 7\sin(\pi X_5X_6) + 6\exp(X_7) + 5\log(|X_8| + 1) + \varepsilon, \qquad \varepsilon \sim N(0, 0.50^2). \tag{4}$$

Thus, the response incorporates different types of nonlinear relationships, including an interaction term between X_1 and X_2, a cubic effect of X_2, a linear effect of X_4, a nonlinear interaction between X_5 and X_6, and nonlinear marginal effects of X_7 and X_8. This construction provides a heterogeneous nonlinear signal in which the informative variables are associated with different dependence structures.

The generated dataset was randomly divided into training, validation, and test sets containing 60%, 20%, and 20% of the observations, respectively. The training set was used for feature selection and model fitting, the validation set was used for tuning the FSEC parameters, and the independent test set was reserved exclusively for the final evaluation of predictive performance. The FSEC algorithm was applied according to the procedure described in Algorithm 1, including the required preprocessing and standardization steps.

To investigate the sensitivity of FSEC to its tuning parameters, nine values of α and eleven values of τ were considered. Specifically, the parameter grids were defined as

$$\alpha \in \{0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80\}, \tag{5}$$

$$\tau \in \{0.001, 0.005, 0.010, 0.020, 0.030, 0.050, 0.075, 0.100, 0.150, 0.200, 0.250\}. \tag{6}$$

These values resulted in a total of 99 (α, τ) combinations. For each parameter combination, FSEC was applied independently to the training data. Importantly, the number of selected variables, k, was not specified as an input parameter. Instead, FSEC determined both the selected feature subset and the corresponding value of k according to its stopping rule for each (α, τ) combination. The resulting selected features were then used to fit a neural network on the training data, and predictive performance was evaluated on the validation set. The validation RMSE was used as the primary criterion for tuning α and τ, while validation MAE and R^2 were also recorded. Thus, the value of k

reported for each configuration represents the number of variables automatically selected by FSEC under that particular parameter combination.

For concise presentation, Table 7 reports representative parameter configurations together with the corresponding number of variables selected by FSEC and their validation performance. The configurations are ordered from the lowest to the highest validation RMSE.

Table 7: Summary of representative sensitivity results for FSEC in Scenario 3, ordered by validation RMSE.

α	τ	k	$\text{RMSE}_{\text{VALID}}$	$\text{MAE}_{\text{VALID}}$	R^2_{VALID}
0.05	0.001	8	1.38	1.09	0.43
0.40	0.150	8	1.41	1.10	0.43
0.05	0.050	8	1.43	1.14	0.43
0.05	0.001	5	1.45	1.18	0.43
0.05	0.050	5	1.45	1.18	0.43
0.05	0.150	5	1.45	1.18	0.43
0.40	0.001	5	1.45	1.18	0.43
0.40	0.001	3	1.47	1.14	0.45
0.05	0.001	3	1.48	1.16	0.45
0.05	0.050	3	1.48	1.14	0.45
0.05	0.150	3	1.48	1.14	0.45

As shown in Table 7, the FSEC stopping rule yielded different subset sizes of $k = 3$, $k = 5$, and $k = 8$ under different (α, τ) configurations. Across the reported configurations, the validation RMSE decreased from 1.47–1.48 for $k = 3$ to 1.45 for $k = 5$ and to 1.38 for $k = 8$. The minimum validation RMSE of 1.38 was obtained for $(\alpha, \tau) = (0.05, 0.001)$, under which FSEC automatically selected eight variables. This configuration was therefore retained for the subsequent comparison, with $k = 8$ determined by the FSEC stopping rule rather than being pre-specified.

Table 8: Selected features by different methods in Scenario 3.

Method	Selected Features	k
FSEC	$X_7, X_5, X_6, X_8, X_4, X_2, X_3, X_1$	8
CBFS	$X_7, X_5, X_8, X_6, X_{24}, X_4, X_{10}, X_{13}$	8
LASSO	$X_7, X_5, X_6, X_4, X_{24}, X_3, X_{13}, X_1$	8
Random Forest	$X_5, X_7, X_6, X_8, X_3, X_4, X_{25}, X_{10}$	8
mRMR	$X_7, X_6, X_4, X_8, X_{10}, X_5, X_{13}, X_2$	8

For a fair comparison, after the representative FSEC parameter combination had been identified and FSEC had automatically determined the corresponding number of selected variables as $k = 8$, the competing methods CBFS, LASSO, Random Forest, and mRMR were constrained to select exactly eight variables. This constraint was imposed only after FSEC had determined its final subset size and was not provided to FSEC as an input. Each competing method independently selected its own best eight predictors according to its corresponding selection criterion.

The selected features obtained by the different feature selection methods are presented in Table 8. FSEC successfully identified all eight truly informative variables, namely $X_1, \ldots, X_8$. CBFS selected five of the eight informative variables and three noise variables, whereas LASSO, Random Forest, and mRMR each selected six informative variables and two noise variables. Thus, FSEC achieved a recovery rate of 100% for the truly informative variables, compared with 62.5% for CBFS and 75% for LASSO, Random Forest, and mRMR.

To evaluate predictive performance, the selected features from each method were used as inputs to a neural network model with one hidden layer containing 15 neurons. The final neural network models were fitted using the combined training and validation sets after completion of the feature selection and parameter-tuning stages. Predictive performance was evaluated on both the combined training–validation set and the independent test set using RMSE, MAE, and R^2. The results are reported in Table 9.

Table 9: Comparison of feature selection methods in Scenario 3.

Method	$\text{RMSE}_{\text{TRAIN+VALID}}$	$\text{RMSE}_{\text{TEST}}$	$\text{MAE}_{\text{TRAIN+VALID}}$	MAE_{TEST}	$R^2_{\text{TRAIN+VALID}}$	R^2_{TEST}
FSEC	0.68	1.86	0.51	1.56	0.89	0.02
CBFS	0.78	1.97	0.59	1.61	0.86	-0.11
LASSO	0.70	2.08	0.53	1.69	0.89	-0.23
Random Forest	0.71	2.08	0.54	1.55	0.88	-0.24
mRMR	0.69	1.98	0.52	1.61	0.89	-0.12

As reported in Table 9, FSEC achieved the best predictive performance on the independent test set in terms of RMSE and R^2. It obtained the lowest test-set RMSE of 1.86 and the highest test-set R^2 of 0.02. FSEC also achieved the lowest RMSE on the combined training–validation set, with a value of 0.68, and a training–validation R^2 of 0.89. Random Forest achieved a slightly lower test-set MAE of 1.55 compared with 1.56 for FSEC. However, its test-set RMSE was higher at 2.08 and its test-set R^2 was negative at -0.24. CBFS, LASSO, and mRMR also produced negative test-set R^2 values of -0.11, -0.23, and -0.12, respectively. These results indicate substantially weaker generalization performance compared with FSEC.

Overall, FSEC provided the strongest generalization performance in this heterogeneous dependence setting, achieving the lowest test-set RMSE and the highest test-set R^2 among the evaluated methods. Its ability to recover all eight truly informative variables further indicates that the proposed dependence-based feature selection procedure can effectively identify informative variables arising from different copula-based dependence structures.

4 Real data

In this section, we plan to test our method using real datasets.

4.1 Real data: Microbes dataset

The Microbes dataset was used to evaluate the proposed FSEC method on a real-world morphological dataset. The response variable was *Area*, which was retained throughout the analysis as the response variable. During the preprocessing stage, to avoid information leakage, variables that were directly related to the response variable were excluded from the predictor set. Specifically, *FilledArea*, *ConvexArea*, *EquivDiameter*, *MajorAxisLength*, *MinorAxisLength*, and *microorganisms* were removed from the set of candidate predictors, while *Area* was retained as the response variable. Consequently, 17 predictor variables were retained for the feature selection analysis: *Solidity*, *Eccentricity*, *Extrema*, *Extent*, *Orientation*, *EulerNumber*, *BoundingBox1*, *BoundingBox2*, *BoundingBox3*, *BoundingBox4*, *ConvexHull1*, *ConvexHull2*, *ConvexHull3*, *ConvexHull4*, *Centroid1*, *Centroid2*, and *raddi*. After removing incomplete observations, 1000 complete observations were used. The data were randomly divided into training, validation, and independent test sets containing 600, 200, and 200 observations, respectively. All standardization parameters were estimated from the training set only and subsequently applied unchanged to the validation and test sets.

To reduce boundary bias in density estimation near the boundaries of the unit interval, reflection kernel density estimation (KDE) was used for both the marginal and copula densities. The numerical integration was performed over $[\varepsilon, 1-\varepsilon]^2$ with $\varepsilon = 0.01$, using a 20×20 grid and the composite trapezoidal rule.

To examine the sensitivity of FSEC to its two tuning parameters, α and τ, an exhaustive grid search was performed using the training and validation sets. The candidate values were

$$\alpha \in \{0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90\},$$

$$\tau \in \{0.001, 0.005, 0.010, 0.020, 0.030, 0.050, 0.075, 0.100, 0.150, 0.200, 0.250, 0.300\}.$$

Thus, 132 (α, τ) combinations were evaluated. For each combination, FSEC was independently applied to the training data according to its predefined stopping rule. The number of selected variables, k, was not specified or fixed as an input parameter. Instead, FSEC automatically determined the selected feature subset and the resulting value of k for each (α, τ) combination. The selected features were then used to fit a neural network on the training data, and predictive performance was evaluated on the validation set. Validation RMSE was used as the primary criterion for selecting the optimal (α, τ) combination, while validation MAE and R^2 were also recorded. The independent test set was not used during this tuning procedure.

For concise presentation, Table 10 reports the top parameter configurations together with the number of variables automatically selected by FSEC and their corresponding validation performance. Thus, k is reported as an outcome of the FSEC stopping rule rather than as an independently tuned parameter.

Table 10: Top FSEC configurations based on validation performance for the Microbes dataset, sorted by the number of selected variables.

Rank	α	τ	k	$\text{RMSE}_{\text{VALID}}$	$\text{MAE}_{\text{VALID}}$	R^2_{VALID}
1	0.40	0.010	5	0.3042920	0.1584727	0.9619915
2	0.20	0.005	5	0.3120735	0.1480333	0.9600227
3	0.40	0.150	5	0.3268981	0.1605044	0.9561344
4	0.20	0.020	5	0.3303400	0.1748500	0.9552058
5	0.40	0.100	5	0.3909543	0.1760789	0.9372590
6	0.01	0.100	5	0.3935657	0.1989857	0.9364180
7	0.10	0.150	5	0.3951722	0.1835635	0.9358979
8	0.20	0.100	5	0.4021904	0.1783604	0.9336008
9	0.80	0.050	4	0.3300220	0.1895244	0.9552920
10	0.70	0.300	4	0.3829498	0.2021658	0.9398018

The best validation performance was obtained for $\alpha = 0.40$ and $\tau = 0.010$, with a validation RMSE of 0.3042920, a validation MAE of 0.1584727, and an R^2 value of 0.9619915. Under this configuration, the FSEC stopping rule automatically selected five variables. Therefore, $k = 5$ was an outcome of the FSEC procedure and was not imposed in advance.

After the optimal (α, τ) configuration had been determined, the resulting feature-selection procedure was reapplied to the combined training–validation set. The FSEC stopping rule automatically selected $k = 5$ predictors. This value was then used only to establish a common feature-selection size for the competing methods. Each competing method independently selected its own five predictors according to its respective selection criterion. The selected feature subsets are reported in Table 11.

Table 11: Features selected by the feature selection methods for the Microbes dataset.

Method	Selected Features	k
FSEC	EulerNumber, BoundingBox4, BoundingBox3, Solidity, ConvexHull1	5
LASSO	EulerNumber, BoundingBox3, BoundingBox4, raddi, Solidity	5
mRMR	EulerNumber, BoundingBox3, BoundingBox4, raddi, Eccentricity	5
CBFS	EulerNumber, BoundingBox3, BoundingBox4, BoundingBox2, BoundingBox1	5
Random Forest	EulerNumber, BoundingBox3, BoundingBox4, Eccentricity, raddi	5

The resulting dependence measures and selected copula families are summarized in Table 11. The selected predictors exhibited varying degrees of dependence with the response variable. EulerNumber, BoundingBox3, and BoundingBox4 were consistently selected by all five feature-selection methods. The results show that FSEC identified a compact subset consisting of EulerNumber, BoundingBox4, BoundingBox3, Solidity, and ConvexHull1. Under the selected FSEC configuration, the stopping rule automatically selected five variables; therefore, $k = 5$ was determined by the FSEC procedure rather than being imposed in advance.

To evaluate predictive performance, the selected features from each method were used as inputs to the same neural network architecture with 15 hidden neurons, a weight-decay parameter of 10^{-4}, and a maximum of 3000 iterations. After feature selection and parameter tuning were completed, the final neural network models were fitted using the combined training and validation sets and evaluated on the independent test set. The coefficient of determination was calculated as $R^2 = 1 - \text{SSE}/\text{SST}$. The resulting predictive performance is presented in Table 12.

Table 12: Predictive performance of the feature selection methods on the Microbes dataset.

Method	Features	$\text{RMSE}_{\text{Train+Valid}}$	$\text{RMSE}_{\text{Test}}$	$\text{MAE}_{\text{Train+Valid}}$	MAE_{Test}	$R^2_{\text{Train+Valid}}$	R^2_{Test}
FSEC	5	0.1631851	0.48694225	0.1028043	0.2126775	0.9840692	0.87917234
LASSO	5	0.1641636	0.6366654	0.1074314	0.2135636	0.9838775	0.80025523
mRMR	5	0.2090983	0.7323011	0.1365888	0.2557517	0.9738436	0.73573952
CBFS	5	0.2217705	1.3993529	0.1525541	0.3662843	0.9705771	0.03504432
Random Forest	5	0.2204149	0.5032548	0.1391081	0.2272974	0.9709357	0.87519594

FSEC achieved the lowest test RMSE among the evaluated feature selection methods, with a test RMSE of 0.48694225. It also achieved the lowest test MAE of 0.2126775 and the highest test R^2 value of 0.87917234. Random Forest followed with a test RMSE of 0.5032548 and a test R^2 of 0.87519594, while LASSO achieved a test RMSE of 0.6366654, a test MAE of 0.2135636, and a test R^2 of 0.80025523. On the combined training–validation set, FSEC achieved an RMSE of 0.1631851, an MAE of 0.1028043, and an R^2 of 0.9840692. The mRMR and CBFS methods showed lower test R^2 values of 0.73573952 and 0.03504432, respectively. Overall, FSEC provided the lowest test RMSE and MAE and the highest test R^2 among the evaluated feature selection methods while using only five selected predictors.

4.2 Breast cancer dataset

To evaluate the applicability of the proposed FSEC framework to a real-world medical dataset, the Breast Cancer dataset was considered. The original dataset contains diagnostic measurements extracted from digitized fine-needle aspiration specimens of breast tumors. During preprocessing, observation-level identifier variables and the diagnostic variable were excluded from the predictor set. Only complete observations with numeric variables were retained, and the first 150 observations were used for the analysis. The variable *texture mean* was considered as the response variable, while the remaining numeric variables associated with cell nucleus characteristics were used as candidate predictors. The data were randomly divided into training, validation, and independent test sets using a 60%–20%–20% split, respectively. All predictor variables were standardized using the training set only, and the same transformation was subsequently applied to the validation and test sets.

To examine the sensitivity of FSEC to its two tuning parameters, α and τ, an exhaustive grid search was performed using the training and validation sets. The candidate values were

$$\alpha \in \{0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80\},$$

$$\tau \in \{0.01, 0.02, 0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90\}.$$

A total of 117 (α, τ) combinations were evaluated. For each combination, FSEC was applied to the training data, and the stopping rule of the algorithm automatically determined the number of selected predictors, k. Thus, the value of k was not fixed in advance. The selected features were then used as inputs to a neural network regression model fitted on the training set, and its predictive performance was evaluated on the validation set. The optimal (α, τ) configuration was determined according to the lowest validation RMSE. The independent test set was not used during parameter tuning.

Table 13: Top FSEC configurations based on validation performance for the Breast Cancer dataset.

Rank	α	τ	k	$\text{RMSE}_{\text{VALID}}$	FSEC Total Score
1	0.70	0.10	9	1.562531	0.107232
2	0.50	0.10	4	1.597216	0.065082
3	0.05	0.05	4	1.756507	0.035533
4	0.80	0.05	26	1.824243	0.187408
5	0.20	0.01	29	1.832073	0.082372
6	0.20	0.02	29	1.832073	0.082372
7	0.05	0.01	28	1.870146	0.053849
8	0.80	0.15	2	1.878989	0.069423
9	0.80	0.20	2	1.878989	0.069423
10	0.80	0.30	2	1.878989	0.069423

As shown in Table 13, the lowest validation RMSE was obtained for $\alpha = 0.70$ and $\tau = 0.10$, with a validation RMSE of 1.562531. Under this optimal configuration, the FSEC stopping rule automatically selected nine predictors, with a corresponding FSEC total score of 0.107232. Therefore, $k = 9$ was determined by the FSEC procedure rather than being imposed in advance. This value was subsequently used to establish a common feature-selection size for the competing methods.

Each competing method independently selected its own nine predictors according to its respective feature-selection criterion. The selected feature subsets are reported in Table 14.

Table 14: Features selected by the feature selection methods for the Breast Cancer dataset.

Method	Selected Features	k
FSEC	texture_worst, concave points_worst, texture_se, concave points_se, symmetry_worst, symmetry_se, radius_mean, compactness_worst, area_se	9
LASSO	texture_worst, texture_se, concavity_se, fractal_dimension_worst, area_se, symmetry_worst, fractal_dimension_se, symmetry_se, perimeter_se	9
mRMR	texture_worst, concavity_se, texture_se, area_mean, smoothness_worst, symmetry_se, radius_mean, concavity_worst, fractal_dimension_worst	9
CBFS	texture_worst, texture_se, concavity_worst, area_mean, smoothness_worst, symmetry_se, compactness_se, symmetry_worst, fractal_dimension_worst	9
Random Forest	texture_worst, texture_se, concavity_worst, compactness_worst, concave points_se, radius_se, area_worst, area_se, smoothness_worst	9

As shown in Table 14, texture_worst and texture_se were selected by all five feature selection methods, indicating their consistently strong predictive relevance for the response variable. The remaining selected predictors varied across the methods, reflecting differences in their respective feature-selection criteria. FSEC selected texture_worst, concave points_worst, texture_se, concave points_se, symmetry_worst, symmetry_se, radius_mean, compactness_worst, and area_se.

To evaluate predictive performance, the selected features from each method were used as inputs to the same neural network architecture with 5 hidden neurons, a weight-decay parameter of 0.01, and a maximum of 300 iterations. After feature selection and parameter tuning were completed, the final neural network regression models were fitted using the combined training and validation sets and evaluated on the independent test set. The coefficient of determination was calculated as $R^2 = 1 - \text{SSE/SST}$. The resulting predictive performance is presented in Table 15.

Table 15: Predictive performance of the feature selection methods on the Breast Cancer dataset.

Method	Features	$\text{RMSE}_{\text{Train+Valid}}$	$\text{RMSE}_{\text{Test}}$	$\text{MAE}_{\text{Train+Valid}}$	MAE_{Test}	$R^2_{\text{Train+Valid}}$	R^2_{Test}
FSEC	9	0.506852	1.588661	0.398690	1.229826	0.979853	0.860125
LASSO	9	0.575989	2.045605	0.389411	1.361758	0.973982	0.768089
mRMR	9	0.795203	2.199357	0.636387	1.590388	0.950410	0.731917
CBFS	9	0.627989	2.402118	0.490968	1.788026	0.969073	0.680209
Random Forest	9	0.709372	2.322641	0.546876	1.717080	0.960537	0.701020

FSEC achieved the best independent test-set performance, with a test RMSE of 1.588661, a test MAE of 1.229826, and a test R^2 of 0.860125. It also achieved the lowest RMSE on the combined training–validation set, with a value of 0.506852, and the highest corresponding R^2 of 0.979853. LASSO obtained the second-best test-set performance, with a test RMSE of 2.045605, a test MAE of 1.361758, and a test R^2 of 0.768089. The remaining methods showed lower predictive performance on the independent test set. Specifically, mRMR, Random Forest, and CBFS obtained test R^2 values of 0.731917, 0.701020, and 0.680209, respectively, with corresponding test RMSE values of 2.199357, 2.322641, and 2.402118. Overall, FSEC provided the strongest generalization performance among the evaluated feature selection methods, achieving the lowest test RMSE and MAE and the highest test R^2 while using only nine selected predictors. These results demonstrate that the proposed copula-extropy-based feature selection framework can effectively identify a compact and informative predictor subset and improve regression performance on the real-world Breast Cancer dataset.

4.3 Concrete compressive strength dataset

To conduct a more practical evaluation of the proposed FSEC framework, a real-world concrete compressive strength dataset was analyzed. The dataset consists of 1,030 observations and 9 variables, including the response variable. In this study, the compressive strength of concrete is considered as the response variable, while Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, and Age are used as predictor variables. This dataset was included to further assess the applicability and effectiveness of the proposed method in a real engineering context.

Only complete observations were retained for the analysis, and the data were randomly divided into training, validation, and independent test sets using a 60%–20%–20% split, respectively. All predictor variables were standardized using the training set only, and the same transformation was subsequently applied to the validation and test sets.

To examine the sensitivity of FSEC to its two tuning parameters, α and τ, an exhaustive grid search was performed using the training and validation sets. The candidate values were

$$\alpha \in \{0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80\},$$

$$\tau \in \{0.01, 0.02, 0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90\}.$$

A total of 117 (α, τ) combinations were evaluated. For each combination, FSEC was applied to the training data, and the stopping rule of the algorithm automatically determined the number of selected predictors, k. Thus, the value of k was not fixed in advance. The selected features were then used as inputs to a neural network regression model fitted on the training set, and its predictive performance was evaluated on the validation set. The optimal (α, τ) configuration was determined according to the lowest validation RMSE. The independent test set was not used during parameter tuning.

Table 16: Top FSEC configurations based on validation performance for the Concrete dataset.

Rank	α	τ	k	$\text{RMSE}_{\text{VALID}}$	$\text{MAE}_{\text{VALID}}$	R^2_{VALID}	FSEC Total Score
1	0.20	0.02	5	5.377461	4.175375	0.9064492	1.582766
2	0.10	0.30	5	5.604172	4.329204	0.8983948	1.615466
3	0.05	0.15	5	5.643448	4.219966	0.8969656	1.649610
4	0.50	0.20	5	5.717125	4.493661	0.8942578	1.281367
5	0.80	0.70	5	5.840435	4.547571	0.8896472	1.227719
6	0.05	0.20	3	5.881105	4.393261	0.8881049	1.649610
7	0.20	0.50	3	5.891858	4.436874	0.8876954	1.582766
8	0.10	0.40	3	5.935139	4.497768	0.8860393	1.615466
9	0.05	0.40	3	5.956879	4.569222	0.8852029	1.649610
10	0.05	0.50	3	5.993448	4.619197	0.8837892	1.649610

As shown in Table 16, the lowest validation RMSE was obtained for $\alpha = 0.20$ and $\tau = 0.02$, with a validation RMSE of 5.377461, an MAE of 4.175375, and an R^2 of 0.9064492. Under this optimal configuration, the FSEC stopping rule automatically selected five predictors, with a corresponding FSEC total score of 1.582766. Therefore, $k = 5$ was determined by the FSEC procedure rather than being imposed in advance. This value was subsequently used to establish a common feature-selection size for the competing methods.

Each competing method independently selected its own five predictors according to its respective feature-selection criterion. The selected feature subsets are reported in Table 17.

Table 17: Selected predictors by different feature selection methods for the Concrete dataset.

Method	k	Selected Predictors
FSEC	5	Cement, Age, Blast Furnace Slag, Water, Superplasticizer
LASSO	5	Cement, Blast Furnace Slag, Age, Fly Ash, Water
mRMR	5	Age, Cement, Superplasticizer, Coarse Aggregate, Fine Aggregate
CBFS	5	Cement, Superplasticizer, Age, Coarse Aggregate, Fine Aggregate
Random Forest	5	Age, Cement, Water, Blast Furnace Slag, Fine Aggregate

As shown in Table 17, Cement and Age were selected by all five feature selection methods, indicating their consistently strong predictive relevance for concrete compressive strength. Blast furnace slag was selected by FSEC, LASSO, and Random Forest, whereas superplasticizer was selected by FSEC, mRMR, and CBFS. The remaining selected predictors varied across the methods, reflecting differences in their respective feature-selection criteria. FSEC selected Cement, Age, Blast Furnace Slag, Water, and Superplasticizer.

To evaluate predictive performance, the selected features from each method were used as inputs to the same neural network regression architecture with 10 hidden neurons, a weight-decay parameter of 0.01, and a maximum of 1000 iterations. After feature selection and parameter tuning were completed, the final neural network regression models were fitted using the combined training–validation set and evaluated on the independent test set. The coefficient of determination was calculated as $R^2 = 1 - \text{SSE}/\text{SST}$. The resulting predictive performance is presented in Table 18.

Table 18: Predictive performance of different feature selection methods for the Concrete dataset.

Method	k	$RMSE_{\text{Train+Valid}}$	$RMSE_{\text{Test}}$	$MAE_{\text{Train+Valid}}$	MAE_{Test}	$R^2_{\text{Train+Valid}}$	R^2_{Test}
FSEC	5	5.162288	6.141721	3.756639	4.286176	0.9062492	0.8492782
LASSO	5	5.542472	6.177048	4.295269	4.712624	0.8919319	0.8475393
mRMR	5	7.288548	7.378915	5.709324	5.744569	0.8131158	0.7824391
CBFS	5	7.419191	7.592816	5.686047	5.924980	0.8063561	0.7696428
Random Forest	5	5.762818	6.502521	4.376767	5.001743	0.8831684	0.8310495

As shown in Table 18, FSEC achieved the best predictive performance on the independent test set, with a test RMSE of 6.141721, a test MAE of 4.286176, and a test R^2 of 0.8492782. It also achieved the lowest RMSE on the combined training–validation set, with a value of 5.162288, the lowest MAE of 3.756639, and the highest corresponding R^2 of 0.9062492. LASSO provided the second-best test-set performance, with a test RMSE of 6.177048, a test MAE of 4.712624, and a test R^2 of 0.8475393. Random Forest also showed relatively competitive performance, obtaining a test RMSE of 6.502521, a test MAE of 5.001743, and a test R^2 of 0.8310495. In contrast, mRMR and CBFS showed lower predictive performance, with test R^2 values of 0.7824391 and 0.7696428, respectively. Overall, FSEC provided the strongest generalization performance among the evaluated feature selection methods, achieving the lowest test RMSE and MAE and the highest test R^2 while using only five selected predictors. These results indicate that the proposed copula-extropy-based feature selection framework can effectively identify a compact and informative predictor subset while maintaining strong predictive performance for concrete compressive strength prediction.

Table 19: Dimensionality reduction achieved by the proposed FSEC method.

Dataset	Original Features (p)	Selected Features (k)	DR (%)
Scenario 1	50	4	92
Scenario 2	200	10	95
Scenario 3	40	8	80
Microbes Dataset	17	5	70.59
Breast Cancer Dataset	29	9	68.97
Concrete Strength Dataset	8	5	37.50

The table above summarizes the dimensionality reduction achieved by the proposed FSEC method based on the number of original and selected predictors. FSEC substantially reduced the predictor space across all evaluated datasets. The largest reduction was obtained in Scenario 2, where the number of predictors was reduced from 200 to 10, corresponding to a 95% reduction. In Scenario 3, the dimensionality was reduced from 40 to 8 predictors, corresponding to an 80% reduction, while Scenario 1 was reduced from 50 to 4 predictors, corresponding to a 92% reduction. For the real-world datasets, FSEC reduced the predictor space from 17 to 5 variables for the Microbes dataset, from 29 to 9 variables for the Breast Cancer dataset, and from 8 to 5 variables for the Concrete Strength dataset.

5 Conclusion

In classical Mutual Information, the dependence between variables is defined by the ratio of the joint density to the product of the marginal densities. However, the presence of the logarithm function in this measure can, in practice, cause numerical sensitivity and instability in the estimation. Specifically, this problem can arise in the presence of noise and/or small sample sizes. On the other hand, the copula-based Mutual information approach, focuses primarily on the dependence structure between variables via the copula function, often without fully integrating the marginal distributions into the dependence measure. As a result, if only copula-based dependence is considered without incorporating marginal information, an incomplete description of the statistical structure of the data may be obtained.

The dependency structure as well as the marginal behaviors are explicitly incorporated into our formulation. This contribution enables us to obtain more informative results for dimensionality reduction. In fact, another innovation of this work is the use of the proposed measure for feature selection. Specifically, based on this measure, the FSEC algorithm is developed, which can automatically select an optimal subset of features such that both the dimensionality of the data is reduced and the predictive performance is improved compared with existing methods.

In the future, this idea can be developed in several directions. One important direction is to optimize the parameter α in the FSEC algorithm so that a better balance between redundancy and relevance to the target variable can be achieved and the performance of the algorithm can be optimized in a data-driven manner. Adaptive methods can also

be developed to automatically determine the number of selected features (k) based on the structure of each dataset. Furthermore, this framework can be extended to classification problems, and its combination with various machine learning models, such as neural networks, support vector machines, and ensemble methods, can be explored. Finally, given that in this paper the performance of the proposed method has been evaluated on simulated data and real-world datasets under a copula- and extropy-based framework, the use of more diverse real-world data from different domains can help to more accurately examine the stability, generalizability, and statistical behavior of the I_{EC} criterion under more complex conditions of dependence and noise, and provide a more comprehensive evaluation of the proposed method along the lines of studies such as [6, 25, 29].

Declarations

Declaration of Generative AI and AI-Assisted Technologies: During the preparation of this manuscript, the authors used ChatGPT (OpenAI) solely for language editing, improving clarity, readability, and polishing the English language. The authors reviewed and verified all AI-assisted output and took full responsibility for the accuracy, originality, validity, integrity, and scientific content of the manuscript. No generative AI or AI-assisted technology was used to generate, modify, fabricate, manipulate, or interpret the primary data, scientific results, statistical analyses, or conclusions reported in this manuscript.

References

[1] A. H. Abdel-aziem, T. H. M. Soliman, *A multi-layer perceptron (MLP) neural networks for stellar classification: A review of methods and results*, International Journal of Advances in Applied Computational Intelligence, **3**(2) (2023), 29-37. https://doi.org/10.54216/IJAACI.030203

[2] N. Balakrishnan, F. Buono, M. Longobardi, *On weighted extropy*, Communications in Statistics - Theory and Methods, **51**(18) (2022), 6250-6267. https://doi.org/10.1080/03610926.2020.1860222

[3] M. Billah, S. Waheed, *Minimum redundancy maximum relevance (mRMR) based feature selection from endoscopic images for automatic gastrointestinal polyp detection*, Multimedia Tools and Applications, **79**(33-34) (2020), 23633-23643. https://doi.org/10.1007/s11042-020-09151-7

[4] F. Buono, O. Kamari, M. Longobardi, *Interval extropy and weighted interval extropy*, Ricerche di Matematica, **72**(1) (2023), 283-298. https://doi.org/10.48550/arXiv.2112.01152

[5] M. Cord, P. Cunningham, *Machine learning techniques for multimedia: Case studies on organization and retrieval*, Springer Berlin, Heidelberg, 2008. https://doi.org/10.1007/978-3-540-75171-7

[6] A. Dhindsa, S. Bhatia, S. Agrawal, B. S. Sohi, *An improvised machine learning model based on mutual information feature selection approach for microbes classification*, Entropy, **23**(2) (2021), 257. https://doi.org/10.3390/e23020257

[7] R. Díaz-Uriarte, S. Alvarez de Andrés, *Gene selection and classification of microarray data using random forest*, BMC Bioinformatics, **7** (2006), 3. https://doi.org/10.1186/1471-2105-7-3

[8] K. Femmam, B. Brahimi, S. Femmam, *An optimized feature selection technique based on bivariate copulas "GBCFS"*, Journal of Combinatorial Optimization, **45**(2) (2023), 74. https://doi.org/10.1007/s10878-023-01006-9

[9] K. Femmam, S. Femmam, *Fast and efficient feature selection method using bivariate copulas*, Journal of Advances in Information Technology, **13**(3) (2022), 301-305. https://doi.org/10.12720/jait.13.3.301-305

[10] I. K. Fodor, *A survey of dimension reduction techniques*, Lawrence Livermore National Laboratory (LLNL), Livermore, CA, United States, 2002.

[11] V. Fonti, E. Belitser, *Feature selection using LASSO*, VU Amsterdam Research Paper in Business Analytics, **30** (2017), 1-25.

[12] T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*, Springer Series in Statistics, 2nd ed., Springer, New York, 2009. https://doi.org/10.1007/978-0-387-84858-7

[13] M. Jian, S. Zengqi, *Mutual information is copula entropy*, Tsinghua Science and Technology, **16**(1) (2011), 51-54. https://doi.org/10.1016/S1007-0214(11)70008-6

[14] I. T. Jolliffe, J. Cadima, *Principal component analysis: A review and recent developments*, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, **374**(2065) (2016), 20150202. https://doi.org/10.1098/rsta.2015.0202

[15] F. Lad, G. Sanfilippo, G. Agró, *Extropy: Complementary dual of entropy*, Statistical Science, **30**(1) (2015), 40-58. https://doi.org/10.1214/14-STS430

[16] S. Lall, D. Sinha, A. Ghosh, D. Sengupta, et al., *Stable feature selection using copula based mutual information*, Pattern Recognition, **112** (2021), 107697. https://doi.org/10.1016/j.patcog.2020.107697

[17] Y. S. Muhammad, M. Iqbal, I. Hussain, S. Kamal, et al., *Modeling non-linear behavior of independent variables*, International Journal of Business and Social Science, **5**(13) (2014), 192-200.

[18] R. D. Nair, E. I. A. Sathar, *Some useful results related to various measures of extropy and their interrelationship*, Statistics and Probability Letters, **193** (2023), 109729. https://doi.org/10.1016/j.spl.2022.109729

[19] S. Nanga, A. T. Bawah, B. A. Acquaye, R. Bill, et al., *Review of dimension reduction methods*, Journal of Data Analysis and Information Processing, **9**(3) (2021), 189-231. https://doi.org/10.4236/jdaip.2021.93013

[20] R. B. Nelsen, *An introduction to copulas*, Springer Series in Statistics, 2nd ed., Springer, New York, 2006. https://doi.org/10.1007/0-387-28678-0

[21] M. A. Nielsen, *Neural networks and deep learning*, Determination Press, San Francisco, CA, USA, 2015.

[22] H. J. Park, T. W. Lee, *Modeling nonlinear dependencies in natural images using mixture of Laplacian distribution*, Advances in Neural Information Processing Systems, **17** (2004), 1041-1048.

[23] H. Peng, G. Fang, P. Li, *Copula for instance-wise feature selection and ranking*, Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence, **216** (2023), 1651-1661. https://doi.org/10.5555/3625834.3625989

[24] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, J. M. O'Sullivan, *A review of feature selection methods for machine learning-based disease risk prediction*, Frontiers in Bioinformatics, **2** (2022), 927312. https://doi.org/10.3389/fbinf.2022.927312

[25] M. A. Rahman, M. Hamada, S. Sharmin, T. A. Rimi, et al., *Enhancing early breast cancer detection through advanced data analysis*, IEEE Access, **12** (2024), 161941-161953. https://doi.org/10.1109/ACCESS.2024.3483095

[26] P. Ray, S. S. Reddy, T. Banerjee, *Various dimension reduction techniques for high dimensional data analysis: A review*, Artificial Intelligence Review, **54**(5) (2021), 3473-3515. https://doi.org/10.1007/s10462-020-09928-0

[27] D. M. Reif, A. A. Motsinger, B. A. McKinney, J. E. Crowe, et al., *Feature selection using a random forests classifier for the integrated analysis of multiple data types*, Proceedings of the 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, (2007), 1-8. https://doi.org/10.1109/CIBCB.2006.330987

[28] S. Saha, S. Kayal, *Copula-based extropy measures, properties, and dependence in bivariate distributions*, Communications in Statistics - Theory and Methods, **55**(1) (2026), 189-216. https://doi.org/10.1080/03610926.2025.2491720

[29] A. Sheikhi, R. Mesiar, M. Holeňa, *A dimension reduction in neural network using copula matrix*, International Journal of General Systems, **52**(2) (2023), 131-146. https://doi.org/10.1080/03081079.2022.2108029

[30] G. Shi, Y. Sheng, H. Ahmadzade, S. Tahmasebi, *Extropy: Dual of entropy for uncertain random variables and its applications*, American Institute of Mathematical Sciences, **21**(5) (2025), 4025-4040. https://doi.org/10.3934/jimo.2025041

[31] A. Sklar, *Fonctions de répartition á n dimensions et leurs marges*, Publications de l'Institut Statistique de l'Université de Paris, **8** (1959), 229-231.

[32] C. Strobl, A. L. Boulesteix, A. Zeileis, T. Hothorn, *Bias in random forest variable importance measures: Illustrations, sources and a solution*, BMC Bioinformatics, **8**(1) (2007), 25. https://doi.org/10.1186/1471-2105-8-25

[33] A. Toomaj, M. Hashempour, N. Balakrishnan, *Extropy: Characterizations and dynamic versions*, Journal of Applied Probability, **60**(4) (2023), 1333-1351. https://doi.org/10.1017/jpr.2023.7

[34] A. Ullah, T. Wang, W. Yao, *Nonlinear modal regression for dependent data with application for predicting COVID-19*, Journal of the Royal Statistical Society Series A: Statistics in Society, **185**(3) (2022), 1424-1453. https://doi.org/10.1111/rssa.12849

[35] B. Venkatesh, J. Anuradha, *A review of feature selection and its methods*, Cybernetics and Information Technologies, **19**(1) (2019), 3-26. https://doi.org/10.2478/cait-2019-0001

[36] G. Wang, F. Lauri, A. H. El Hassani, *Feature selection by mRMR method for heart disease diagnosis*, IEEE Access, **10** (2022), 100786-100796. https://doi.org/10.1109/ACCESS.2022.3207492