STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 2025, pp 0–16.
Published online in International Academic Press (www.IAPress.org)

Feature Selection via Fuzzy Rough Set Theory for Robust Classification: a Review and Comparative study

Zineb Khaldoun*, Hasna Chamlal, Tayeb Ouaderhman

Computer Science and Systems Laboratory (LIS), Faculty of Sciences Ain Chock, University Hassan II of Casablanca, Morocco

Abstract
Despite a variety of powerful classifiers available in machine learning today, most of them struggle with processing large-scale real-world datasets. Usually, these datasets contain irrelevant and redundant information that can negatively affect the model's performance. To overcome this, feature selection has become a commonly used strategy to improve model performance by reducing dataset size while retaining essential information. Some feature selection techniques tend to require more information than what is provided in the given dataset, making them impractical in some cases. Alternatively, completely data-driven methods may lose critical information, as they can mistake vagueness or imprecision in the dataset for irrelevant or redundant features. Fuzzy-rough set theory offers a robust paradigm for tackling uncertainties, having been utilised across various domains, with feature selection being one of its most prominent applications. This paper presents an extensive review of feature selection methodologies grounded in fuzzy-rough set theory, accompanied by an empirical evaluation of multiple techniques to evaluate their effectiveness.

Keywords Feature selection, Dimensionality reduction, Fuzzy rough set, Rough set.

AMS 2010 subject classifications 62-07, 97K80, 68T20, 65C60

DOI: 10.19139/soic-2310-5070-2540

1. Introduction

High-dimensional data presents a significant challenge to machine learning models, as it increases the likelihood of models overfitting on noise and irrelevant information, thereby misclassifying data and leading to poor performance [8, 32]. Feature selection [1] attempts to solve these issues by identifying and extracting pertinent attributes (or features) for model construction. Besides this, the process of feature selection reduces training time, enhances interpretability, and improves classification accuracy [31]. Moreover, feature selection proves its importance across diverse fields [9, 33], including credit scoring [2], image classification [3], and medical diagnosis [33], by enhancing data quality and making the decision-making process more precise in these critical domains. A conceptual overview of a feature selection process is illustrated in Figure 1.

A substantial body of research has documented numerous feature selection methodologies across existing literature, yet a significant majority fails to adequately accommodate ambiguity and data inconsistency. These approaches may categorize such data as irrelevant and exclude it, yielding a resultant dataset with diminished informational potency. Rough Set Theory [5] offers a robust groundwork for handling datasets of disparate complexities without reliance on supplementary information beyond what the dataset

*Correspondence to: Zineb Khaldoun (Email: zineb.khaldoun-etu@etu.univh2c.ma). Department of Mathematics and Informatics, Faculty of Sciences Ain Chock, University Hassan II of Casablanca, Morocco.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

provides. However, one key restriction of rough set theory is that all data should be discrete, which diminishes its usability over real-valued data. A widely used approach for overcoming this challenge is the discretization of numerical data, which often leads to information loss.

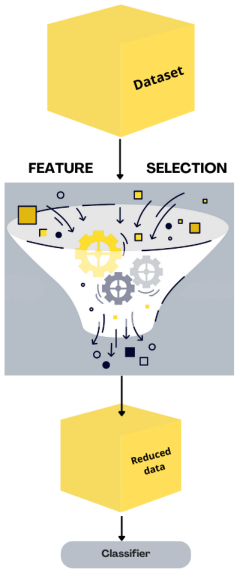


Figure 1. Overview of Feature selection

To address the above limitations, Dubois and Prade [6] presented a new framework that integrates fuzzy set theory and rough set theory into a singular, unified concept known as fuzzy-rough set theory. With such a hybrid theory, it is possible to work with continuous and discrete simultaneously without the need to discretize the data, which protects data against the risk of information loss in the process of reducing the dimensionality. With the incorporation of fuzzy membership functions into rough set definitions, fuzzy-rough set theory becomes particularly effective in feature selection tasks where data may contain imprecise or vague information. This theory has garnered substantial attention in the research world attributable to its capacity to process diverse data types and handle complex, uncertain data effectively. Nonetheless, existing literature exhibits a dearth of in-depth examinations regarding feature selection methods founded in fuzzy-rough set theory.

This paper presents a comprehensive examination of feature selection methods grounded in fuzzy-rough set theory, beginning with fundamental theoretical principles and essential definitions. The study concludes with an experimental evaluation across multiple datasets, assessing both the efficacy and advancements of these methods. The investigation particularly highlights their practical utility and demonstrated capability to handle complex, real-world datasets.

This review is structured in the following manners: Section 2 provides a comprehensive exposition on the theoretical base of both fuzzy sets and rough sets and thus lays the ground for the subsequent examination of fuzzy-rough set theory. Section 3 provides a critical assessment of feature selection approaches grounded in fuzzy-rough sets. Section 4 presents a comparative experimental analysis aimed at investigating the effectiveness of these methods. Finally, Section 5 distils the key outcomes of this study into a concise conclusion.

2. Preliminaries

In order to understand the subsequent section, the primary principles and operators of fuzzy set theory, rough set theory, and the unified framework of fuzzy-rough set theory will be introduced.

Within the theoretical frameworks examined in this paper, a dataset is represented as tuple $\langle \mathbb{U}, \mathbb{A} \rangle$ called information system [5], where $\mathbb{U}$ is the set of objects called the universe and $\mathbb{A}$ is the set of features in order that for every feature $c \in \mathbb{A}$, $c : \mathbb{U} \to V_c$. Here, V_c represents the set of possible values that feature c can take.

2.1. Fuzzy Set Theory

Fuzzy set theory is a theoretical framework proposed by Lotfi Zadeh in 1965. Inspired by human reasoning, it generalizes the classical set theory, by allowing an element to have a membership degree in multiple sets rather than being restricted to inclusion or exclusion within a single set. It excels at managing imprecision and uncertainty in data due to its reasoning flexibility. The utility of this theory shows in scenarios where classical set theory proves insufficient when dealing with complex, nuanced real-world data.

Definition 1 ([13]). A fuzzy set can be conceptually represented through a collection of ordered pairs:

$$A = \{(x, \mu_A(x)) \mid x \in \mathbb{U}\}$$

The function $\mu_A(x)$ is called the membership function for A, mapping each element of the universe $\mathbb{U}$ to a membership degree in the range [0.1].

In order to facilitate the manipulation of fuzzy sets, various operations have been created. The commonly applied fuzzy set operations are listed in Table 1, as detailed in this work [10].

Operator	Definition
Intersection	$\mu_{A\cap B}(x) = \min(\mu_A(x), \mu_B(x))$
Union	$\mu_{A\cup B}(x) = \max(\mu_A(x), \mu_B(x))$
Complement	$\mu_{\overline{A}}(x) = 1 - \mu_A(x)$

Table 1. Fuzzy Set Operators

2.2. Fuzzy Relations

A fuzzy relation R may be represented as a fuzzy set over the Cartesian product of the universe U, mapping each ordered pair (x, y) from $\mathbb{U} \times \mathbb{U}$ to the interval $[0, 1]$. The membership value $R(x, y)$ assigned to each pair indicates the degree of association between the elements x and y under the specified relation R.

Let R denote a fuzzy relation on the Cartesian product $\mathbb{U} \times \mathbb{U}$. In this work [7], R is referred to as a fuzzy T-similarity relation on $\mathbb{U}$ for all $x, y \in \mathbb{U}$ if R satisfies:

- R is reflexive $\Leftrightarrow R(x, x) = 1$
- R is symmetric $\Leftrightarrow R(x, y) = R(y, x)$
- R is T-transitive $\Leftrightarrow R(x, z) \geq T(R(x, y), R(y, z))$

In the particular case where triangular norm $T = min$ then R is fuzzy equivalence relation on U. Furthermore, they defined a fuzzy relation over any subset $J \subseteq \mathbb{A}$ by:

$$R_J = \bigcap_{a_k \in J} R_{a_k}. \tag{1}$$

Example of Fuzzy Similarity Relation

There are numerous fuzzy similarity relations proposed in the literature. In this article, we adopt a commonly used definition for numerical attributes. Given a numerical attribute a, the fuzzy similarity relation $R_a(x_i, x_j)$ between two instances x_i and x_j is defined as:

$$R_a(x_i, x_j) = 1 - \frac{|a(x_i) - a(x_j)|}{\max(a) - \min(a)} \tag{2}$$

This relation quantifies the similarity between instances based on their attribute values, where the value of $R_a(x_i, x_j)$ lies in the interval $[0, 1]$; a value of 1 indicates complete similarity, while 0 indicates no similarity.

Example:

Consider an information system that contains two features a_1 and a_2:

$\mathbb{U}$	a_1	a_2
x_1	2.0	10.0
x_2	3.5	12.0
x_3	4.0	14.0

Using Equation 2, the fuzzy similarity relations for the two features are calculated as:

$$\begin{aligned} R_{a_1}(x_1,x_2) &= 1 - \tfrac{1.5}{2.0} = 0.25, & R_{a_2}(x_1,x_2) &= 1 - \tfrac{2.0}{4.0} = 0.5 \\ R_{a_1}(x_1,x_3) &= 1 - \tfrac{2.0}{2.0} = 0.0, & R_{a_2}(x_1,x_3) &= 1 - \tfrac{4.0}{4.0} = 0.0 \\ R_{a_1}(x_2,x_3) &= 1 - \tfrac{0.5}{2.0} = 0.75, & R_{a_2}(x_2,x_3) &= 1 - \tfrac{2.0}{4.0} = 0.5 \end{aligned}$$

These can be represented in matrix form as:

$$R_{a_1} = \begin{bmatrix} 1 & 0.25 & 0 \\ 0.25 & 1 & 0.75 \\ 0 & 0.75 & 1 \end{bmatrix} \quad R_{a_2} = \begin{bmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0.5 \\ 0 & 0.5 & 1 \end{bmatrix}$$

By applying Equation 1, we can calculate the fuzzy similarity relation over the subset $\{a_1, a_2\}$:

$$R_{\{a_1,a_2\}} = \begin{bmatrix} 1 & \min(0.25, 0.5) = 0.25 & \min(0.0, 0.0) = 0.0 \\ 0.25 & 1 & \min(0.75, 0.5) = 0.5 \\ 0.0 & 0.5 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.25 & 0.0 \\ 0.25 & 1 & 0.5 \\ 0.0 & 0.5 & 1 \end{bmatrix}$$

This matrix represents the fuzzy similarity between instances x_1, x_2, x_3 based on the combination of features a_1 and a_2.

2.3. Rough Set Theory

Rough Set Theory, introduced by Zdzisław Pawlak in the 1980s, has been extensively utilized as a methodological tool by researchers to elicit data dependencies and to effect attribute reduction in datasets independently, devoid of supplementary information. Particularly created for datasets characterised by discrete features, aiming to reduce dimensionality by identifying a subset of features termed ***reduct*** that provides a minimal yet informative representation of the original data. We recall some key definitions from [14].

Let $\langle \mathbb{U}, \mathbb{A} \rangle$ be an information system. For any subset $J \subseteq \mathbb{A}$, there is an associated equivalence relation $IND(J)$:

$$\text{IND}(J) = \left\{ (x, y) \in \mathbb{U}^2 \mid \forall c \in J,\ c(x) = c(y) \right\} \tag{3}$$

$\mathrm{IND}(J)$ is known as *J-indiscernibility relation*, where objects that belong to $\mathrm{IND}(J)$ are indistinguishable with respect to J. This indiscernibility relation generates partitions of the universe U, denoted as $\mathbb{U}/\mathrm{IND}(J)$.

Given a subset K of the universe $\mathbb{U}$, where $\mathbb{U} = \{x_1, \ldots, x_n\}$, the set K can be approximated utilizing the information contained in set J by generating the J-lower and J-upper approximations of K as follows:

$$\underline{J}K = \{x_k \mid [x_k]_J \subseteq K\} \tag{4}$$

and

$$\overline{J}K = \{x_k \mid [x_k]_J \cap K \neq \emptyset\} \tag{5}$$

where $[x_k]_J$ represent the equivalence classes of the J-indiscernibility relation. Objects in $\underline{J}K$ can be classified with absolute certainty as elements of set K, whereas objects in $\overline{J}K$ can be classified as potential elements of set K within the constraints of knowledge base J [5].

The discovery of dependencies among attributes is one of the central tasks of data analysis. RST offers a prominent tool employed for this purpose, the dependency function, which quantifies the degree to which a set of attributes E depends on another set J. The dependency between J and E is expressed as [15]:

$$\gamma_J(E) = \frac{|POS_J(E)|}{|\mathbb{U}|} \tag{6}$$

where:

$$\mathrm{POS}_J(E) = \bigcup_{K \in \mathbb{U}/E} \underline{J}K \tag{7}$$

Here, $|\cdot|$ denotes the cardinality of a set. The value of $\gamma_J(E)$ is always situated in the interval $[0, 1]$. A value nearer to 1 suggests that E is highly dependent on J, while a value closer to 0 suggests a weaker dependence of E on J.

2.4. Fuzzy-Rough Set Theory

Fuzzy-rough set theory, proposed by Dubois and Prade, extends the traditional framework of rough set theory by incorporating the concept of a membership function from fuzzy set theory. This integration makes it possible to utlze the definitions of rough set theory in various types of data such as continued, discrete, and mixed datasets and not be limited only to discrete data.

Let $\langle \mathbb{U}, \mathbb{A} \rangle$ denote a fuzzy decision system, where the attribute set $\mathbb{A}$ is composed of a union of conditional attribute set $\mathbb{C}$ and decision attribute set $\mathbb{D}$. A fuzzy decision system is a type of fuzzy information system equipped with decision attributes.

In this context, the foundational work of Dubois and Prade [6] introduces key definitions related to fuzzy-rough sets, particularly in describing the lower and upper approximations of a fuzzy set X based on a T-similarity relation R, as follows:

$$\underline{R}_B X(x) = \inf_{y \in U} \max\{1 - R_B(x, y), X(y)\} \tag{8}$$

$$\overline{R}_B X(x) = \sup_{y \in U} \min\{R_B(x, y), X(y)\} \tag{9}$$

The function $\underline{R}_B X(x)$ measures the degree to which x certainly belongs to the fuzzy set X, based on the knowledge contained in B, while $\overline{R}_B X(x)$ measures the degree to which x possibly belongs to X, also based on the knowledge contained in B.

Furthermore, given $B \subseteq \mathbb{C}$. The fuzzy positive region of $\mathbb{D}$ with respect to B is defined as:

$$\mathrm{POS}_B(\mathbb{D}) = \bigcup_{X \in \mathbb{U}/\mathbb{D}} \underline{R_B} X \tag{10}$$

Building on the definition of the fuzzy positive region, this groundbreaking work [11] defines the fuzzy dependency function as follows:

$$\gamma'_{\mathbb{C}}(\mathbb{D}) = \frac{\sum_{x \in \mathbb{U}} \mathrm{POS}_B(\mathbb{D})(x)}{|\mathbb{U}|} \tag{11}$$

Similar to the dependency function in rough set theory framework, when the value of $\gamma'_{\mathbb{C}}(\mathbb{D})$ approaches 1, it indicates a strong dependency between the decision variable $\mathbb{D}$ and the feature subset $\mathbb{C}$. Conversely, a value closer to 0 suggests a weaker dependency.

2.5. Fuzzy Entropy

Fuzzy entropy, a metric developed by De Luca and Termini drawing upon Shannon's entropy [16], serves as a quantitative indicator of the degree of fuzziness within data. Specifically, it reflects the average quantifiable information resident within data that underpins object classification. According to this work [17], the fuzzy entropy H is given by:

$$H = -\frac{1}{n} \sum_{i=1}^{n} \left(\mu_i \log(\mu_i) + (1 - \mu_i) \log(1 - \mu_i) \right), \tag{12}$$

where μ_i represents the membership function of the fuzzy set, and log denotes the base 10 logarithm.

3. Fuzzy Rough Set for Selecting Features

Fuzzy rough set theory has become widespread in such fields as data mining, pattern recognition, and decision-making. Especially, this theory has witnessed quite significant development in feature selection, leading to a number of efficient and handy methodologies, capable of operating noisy and large-scale datasets. Quite a significant number of feature selection studies, based on the fuzzy rough set theory, has proved that efficient dimensionality reduction is possible without the loss of information.

3.1. Feature Selection techniques foundationed on Fuzzy Rough Set

Chouchoulas and Shen's seminal work, as reported [18], pioneered the application of rough set theory for feature selection. Through the introduction of the QUICKREDUCT algorithm, they devised a computational framework that harnesses a dependency function to isolate a reduct. Despite its efficacy, QUICKREDUCT is constrained by its discrete data constrictions and fails to ensure the identification of an optimal minimal subset. Hence, Jensen and Shen's [20] Fuzzy Rough Feature Selection (FRFS) algorithm, as illustrated in Algorithm 1, was promulgated to mitigate these limitations. FRFS uses a fuzzy-rough dependency function and selects relevant features iteratively based on it. Meaning features are incrementally included only when they increase the dependency measure, thereby conclusively determining the optimal subset.

This methodology constituted a substantial improvement in the application of fuzzy rough sets to feature selection, allowing for the management of datasets comprising both discrete and real-valued attributes. Nonetheless, the approach's efficacy deteriorates as dataset size increase, resulting in escalating processing durations, thereby compromising its operational efficiency. The primary impediment to this methodology's effectiveness is the substantial algorithmic complexity of the Cartesian product of fuzzy equivalence classes within the algorithm.

Algorithm 1: FRFS [20]

Input: Decision system DS
Output: R: the selected feature subset
(1) $R \leftarrow \emptyset$; $\gamma'_{\text{best}} \leftarrow 0$; γ'_{prev};
(2) **do** ;
(3) $T \leftarrow R$;
(4) $\gamma'_{\text{prev}} \leftarrow \gamma'_{\text{best}}$;
(5) **for each** $a \in (\mathbb{C} - R)$ **do** ;
(6) **if** $\gamma'_{R\cup\{a\}}(\mathbb{D}) > \gamma'_T(\mathbb{D})$ **then** ;
(7) $T \leftarrow R \cup \{a\}$;
(8) $\gamma'_{\text{best}} \leftarrow \gamma'_T(\mathbb{D})$;
(9) $R \leftarrow T$;
(10) **until** $\gamma'_{\text{best}} = \gamma'_{\text{prev}}$;
(11) **return** R;

In [20], Jensen and Shen introduced the Fuzzy Discernibility Matrix (FDM) algorithm, utilizing the fuzzy discernibility matrix for feature selection, which employs this criterion to identify reducts through evaluation of each feature's role in class differentiation. However, FDM is becoming computationally expensive as datasets grow bigger. Chen et al. [21] proposed the Sample Pair Selection (SPS) algorithm, which focuses solely on the essential elements deduced from the fuzzy discernibility matrix, by utilizing discernibility relations to facilate the identification of reducts by reducing drastically the computational complexity and considering only those elements that are essential for effective classification.

Other recent work on fuzzy dependency functions includes the Max-Relevance Max-Significance algorithm (MRMS) [22], which draws on principles of the Max-Relevance Min-Redundancy algorithm [23]. Notably, the MRMS algorithm is designed to simultaneously enhance both the relevance and the significance of the extracted attributes. Zhang et al. [25] introduce the Filter-Wrapper Approach Reduction Algorithm (FWARA), an instance-based feature selection method. Unlike conventional approaches that operate on the entire dataset, FWARA employs a two-stage selection process: it first identifies an initial feature subset by computing the fuzzy dependency function using only representative instances - the most discriminative data points determined through fuzzy relations; then it performs wrapper-based backward elimination, where a classifier iteratively refines the feature subset to retain the combination yielding the highest classification accuracy.The MRMS and FWARA algorithms are formally presented in Algorithms 2 and 3, respectively.

Numerous fuzzy rough feature selection methods employ upper and lower approximations of a fuzzy set based on similarity relation. For instance, Wang et al. [24] have developed a parameterized fuzzy relation. Thus, a new dependency function is formed on which a heuristic algorithm called NFRS is developed. The NFRS algorithm guarantees the best dependence of a sample's class with low uncertainty but maintains the inter-feature interactions which may lead to the under-specified features having important discriminative information.

Building on these foundations, De Luca and Termini [17] extended Shannon entropy towards fuzzy rough sets, enabling the effective extraction of information from fuzzy sets. Hu et al. [35] presented a fuzzy entropy-based approach named FEAR to measure the ability of fuzzy relations to distinguish among objects, as illustrated in Algorithm 4. Much later, Pasi Luukka [30] proposed a feature selection technique that employs fuzzy entropy coupled with a similarity classifier, called the FSFEmSC algorithm. In this approach, feature removal is guided by their entropy values, assuming that features having the highest entropy contribute the least to class discrimination, while the most informative features have the lowest entropy values. Additionally, Wang et al. [26] proposed a Dynamic Interaction Feature Selection

Algorithm 2: MRMS [22]

Input: Decision system DS
Output: R: reduct of the attributes A
(1) $R \leftarrow \emptyset$;
(2) Construct the FEPM M_{A_i} for each $A_i \in \mathbb{C}$;
(3) Calculate relevance $\gamma_{\{A_i\}}(D)$ for each $A_i \in \mathbb{C}$;
(4) Select A_i with highest $\gamma_{\{A_i\}}(D)$, add to R, and update $\mathbb{C} \leftarrow \mathbb{C} \setminus \{A_i\}$;
(5) **repeat** ;
(6) **for each** $A_j \in \mathbb{C}$ and $A_i \in R$ construct FEPM $M_{\{A_i,A_j\}}$;
(7) Compute $\sigma_{\{A_i,A_j\}}(D, A_j) = \gamma_{\{A_i,A_j\}}(D) - \gamma_{A_i}(D)$;
(8) **if** $\sigma_{\{A_i,A_j\}}(D, A_j) = 0$ for any $A_i \in R$, then remove A_j from $\mathbb{C}$;
(9) From remaining $A_j \in \mathbb{C}$, select feature maximizing:

$$\omega\gamma_{\{A_j\}}(D) + \frac{(1-\omega)}{|S|} \sum_{A_i \in S} \sigma_{\{A_i,A_j\}}(D, A_j)$$

(10) Add selected A_j to S and update $\mathbb{C} \leftarrow \mathbb{C} \setminus \{A_j\}$;
(11) **until** $\mathbb{C} = \emptyset$ **or** $|R| = d$ desired number of features;
(12) **return** R;

Algorithm 3: FWARA [25]

Input: Decision system DS, one minimal fuzzy granular rule set $R^*(A, D)$, and the representative instance set U^*.
Output: Reduct B of $(U, A \cup D)$.
(1) Initialize $B \leftarrow \emptyset$, threshold $\leftarrow -1$;
(2) Compute $\gamma^*_A(D)$;
(3) **for each** $a \in A \setminus B$, compute $\gamma^*_{B\cup\{a\}}(D)$;
(4) Let a_{i_0} be the attribute satisfying:

$$\gamma^*_{B\cup\{a_{i_0}\}}(D) = \max_{a\in A\setminus B} \gamma^*_{B\cup\{a\}}(D)$$

if $\gamma^*_{B\cup\{a_{i_0}\}}(D) >$ threshold, then update:

$$B \leftarrow B \cup \{a_{i_0}\}, \quad \text{threshold} \leftarrow \gamma^*_{B\cup\{a_{i_0}\}}(D)$$

(5) **if** threshold $< \gamma^*_A(D)$, go to Step (3); otherwise, output B and proceed;
(6) Proceed by applying a wrapper-based phase to refine the selected subset B. In this phase, different combinations of features in B are evaluated using a classifier, and the subset yielding the highest classification accuracy is retained as the final optimal feature subset.

method based on fuzzy entropy (DIFS-FRS), which selects features based on their relevancy, redundancy, and taking into consideration the interaction between features. Later, an incremental feature selection algorithm was developed by Dong et al. [28] for dynamic datasets, termed as ASIRA. It can handle the increases of both the number of samples and features simultaneously. However, this method is still suffering from its computation burden when datasets expand.

Most recently, Zhao et al. [29] devised a consistency approximation framework through their CAIFS algorithm to augment incremental feature selection. This framework is leveraged to accelerate

Algorithm 4: FEAR [35]

Input: Decision system DS
Output: One reduct red of DS
Step 1: For all $a \in A$, compute the equivalence relation;
Step 2: red $\leftarrow \emptyset$;
Step 3: foreach $a_i \in C \setminus red$ **do**
  $H_i \leftarrow SIG(a_i, \text{red}, d)$;
Step 4: Let a such that $SIG(a, \text{red}, d) = \max_i(H_i)$;
Step 5: if $SIG(a, red, d) > 0$ **then**
  red $\leftarrow$ red $\cup \{a\}$;
  Go to Step 3;
else
  return red;

computational efficiency through a tri-accelerator mechanism, which involves two major steps: identifying the most informative samples, followed by feature evaluation using the significance measure. Hence, this approach not only improves the model performance but also reduces the time complexity challenges encountered in large-scale and high-dimensional datasets.

3.2. Summary

The development of fuzzy rough set-based feature selection has provided powerful tools for handling dimensionality reduction, primarily underpinned by the discernibility matrix, dependency degree, and fuzzy entropy theoretical frameworks.

- **Discernibility matrix-based** approaches aim to reduce the dimensionality of datasets by removing those attributes that fail to discern between decision classes.
- **Dependency degree-based** strategies determines the importance of each feature, based on quantifying the dependency between attributes and decision attributes.
- **Fuzzy entropy-based** methodologies, which are based on the principle of information entropy for the purpose of refine feature selection by quantifying the information gained from fuzzy sets.

A summary of the categorization, advantages, and limitations of the fuzzy rough set-based feature selection methods reviewed in this paper is presented in Table 2.

4. Experimental Evaluation

In this section, we compare the performance of several fuzzy rough set-based algorithms in terms of subset optimality and classification accuracy. The experiments were conducted on a system equipped with an Intel(R) Core(TM) i7-8650U processor (4.20 GHz), 16 GB of RAM, and running the Windows 11 operating system. All algorithms were implemented and executed using Python version 3.12.7.

4.1. Experimental Analysis

The performance of four methods—FRFS, FEAR, FWARA, and MRMS—is evaluated. The experiments mainly focus on selecting the best feature subset using these methods and comparing them in terms of computational time, the cardinality of the selected feature subsets, and the classification performance of the selected subsets. To accomplish this, six datasets were downloaded from the UCI Machine Learning

Table 2. Summary of Fuzzy Rough Set-Based Feature Selection Methods

Method	Category	Advantages	Limitations
FRFS [20]	Dependency-based	Handles mixed data; selects only features improving dependency	High computational cost due to Cartesian product of fuzzy equivalence classes; less efficient on large datasets
FDM [20]	Discernibility-based	Identifies reducts using class-distinguishing features	Computationally expensive for large-scale data
SPS [21]	Discernibility-based	Reduces complexity by selecting only essential discernibility elements	May ignore weak but useful features
MRMS [22]	Dependency-based	Balances relevance and significance; controls redundancy	Can be sensitive to the choice of the membership function
FWARA [25]	Dependency-based	Efficient two-stage process; high accuracy with reduced subsets	The performance is classifier-dependent and unsuitable for large-scale data
NFRS [24]	Parameterized Dependency	Reduces uncertainty while preserving interactions	Risk of selecting underspecified features
FEAR [35]	Fuzzy entropy-based	Measures discrimination ability via fuzzy entropy	Performance decreases with dataset size due to entropy dilution
FSFEmSC [30]	Fuzzy entropy-based	Simple and effective entropy-based filter with similarity classifier	Assumes high-entropy features are non-informative, which may not always hold
DIFS-FRS [26]	Fuzzy entropy-based	Considers feature relevance, redundancy, and interaction	Computationally demanding as feature space grows
ASIRA [28]	Discernability-based	Handles dynamic datasets with increasing size	High computational cost in large-scale dynamic environments
CAIFS [29]	Dependency-based	Fast and consistent selection via tri-accelerator mechanism	Relies heavily on sample quality and significance thresholds

Repository [34] and OpenML [46]. The datasets are briefly described in Table 3. Prior to analysis, a simple preprocessing step was applied where any missing values were replaced by the mean of the respective feature, as all datasets are numeric.

Performance assessment was conducted using 5-fold cross-validation with an $80\% - 20\%$ train-test split per fold, where for each fold, the feature selection methods were applied to the training partition, and the selected features were evaluated using both a K-Nearest Neighbors (KNN, $K = 3$) classifier and a Support Vector Machine (SVM) classifier on the test partition.

Table 3. Characteristics of the datasets used in the study

Dataset	Objects	Features	Number of Classes	Source
Sonar	207	60	2	UCI
Glass	214	9	6	UCI
Ionosphere	230	34	2	UCI
WDBC	569	30	2	UCI
Colon	62	2000	2	OpenML
Prostate	149	12600	2	OpenML

4.2. Experimental Results

To compare the effectiveness of the chosen fuzzy-rough-set feature methods, we analyzed their performance across six diverse datasets, where an assessment of the algorithms was conducted in terms of classification accuracy, size of the selected feature subset, dimensionality reduction rate, and execution time. Tables 4 and 5 present the average classification accuracies and the cardinalities of the selected feature subsets obtained by each method using SVM and 3NN, respectively.

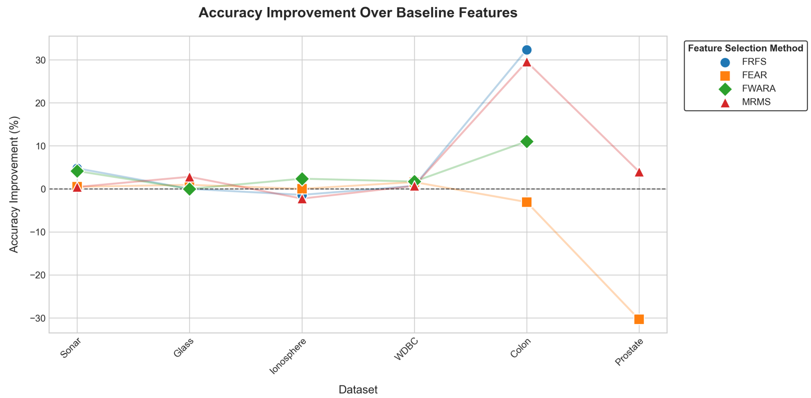
Table 4. Comparison of feature selection methods across datasets using SVM classifier

Dataset	Original Accuracy	FRFS		FEAR		FWARA		MRMS	
		$\|R\|$	Accuracy	$\|R\|$	Accuracy	$\|R\|$	Accuracy	$\|R\|$	Accuracy
Sonar	75.05	45	79.83	38	75.53	18	79.20	49	75.49
Glass	60.78	9	60.78	6	61.72	9	60.78	8	63.59
Ionosphere	88.29	30	86.90	30	88.31	26	90.66	9	86.03
WDBC	94.90	22	95.60	25	96.48	22	96.60	14	95.60
Colon	66.02	22	98.37	4	62.95	5	77.06	14	95.60
Prostate	86.23	-	-	18	55.90	-	-	20	90.23

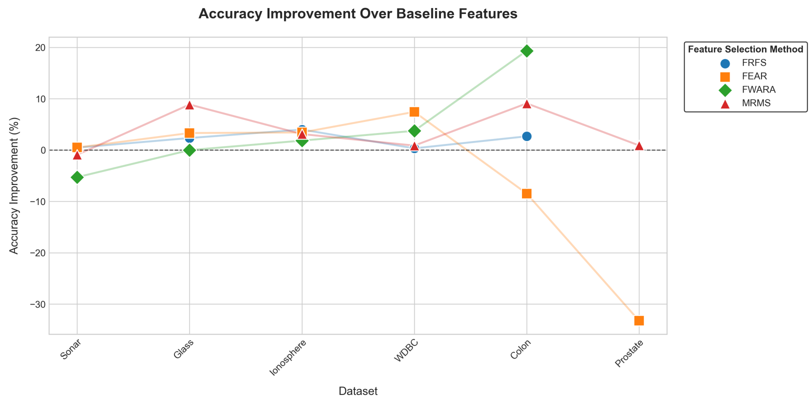
Table 5. Comparison of feature selection methods across datasets using 3NN classifier

Dataset	Original Accuracy	FRFS		FEAR		FWARA		MRMS	
		$\|R\|$	Accuracy	$\|R\|$	Accuracy	$\|R\|$	Accuracy	$\|R\|$	Accuracy
Sonar	81.72	24	82.20	38	82.23	18	76.45	50	80.82
Glass	65.43	7	67.79	7	68.76	9	65.43	8	74.31
Ionosphere	84.89	8	88.89	12	88.33	26	86.74	7	88.02
WDBC	91.91	24	92.26	28	99.36	22	95.67	26	92.79
Colon	72.82	17	75.51	12	64.36	5	92.17	10	81.92
Prostate	76.33	-	-	42	53.00	-	-	20	87.14

Figure 3 provides a visual overview of the feature reduction capability of each method. It can be observed that the majority of methods significantly reduce the number of features while preserving or even improving accuracy. Additionally, Table 6 reports the average running time for each method per fold, revealing that FRFS and FWARA are computationally more efficient than FEAR and MRMS on most datasets.



(a) SVM



(b) 3NN

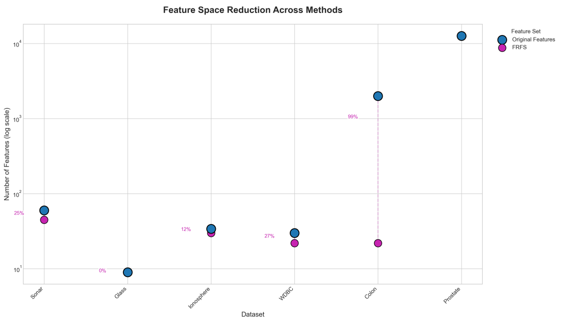
Figure 2. Accuracy improvement after feature selection: performance comparison using SVM and 3NN classifiers.

Table 6. Average running time (seconds) of finding one reduct (per fold)

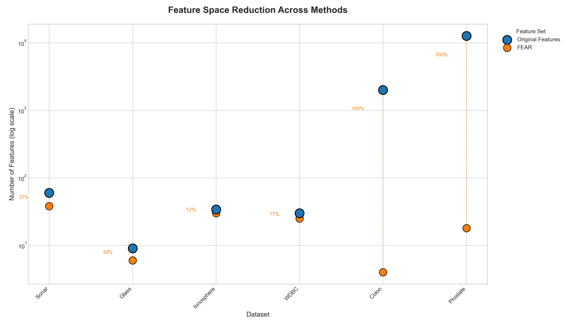
Dataset	FRFS	FEAR	FWARA	MRMS
Sonar	14.37	28.04	15.84	9.55
Glass	0.06	4.44	12.16	3.19
Ionosphere	7.10	29.62	18.13	37.62
WDBC	12.27	76.31	19.32	62.90
Colon	14.87	76.28	15.41	41.31
Prostate	-	1361.28	-	692.44

The comparative evaluation of the methods highlights several performance trends where it can be observed from the results that FRFS consistently improves SVM accuracy (up to +4.8% over baseline), especially in low-dimensional settings. However, it fails to generate valid subsets for the high-dimensional Prostate dataset. In the other hand the FEAR shows superior performance with 3NN, yielding up to +3.9% accuracy gain, demonstrating the advantage of its entropy-driven criteria in neighborhood-based classification. For the filter-wrapper FWARA method, it can be seen that it achieves the most substantial dimensionality reduction (approximately 82% fewer features on average).However, FWARA's performance varies across classifiers, which can be explained by the method's dependency on the classifier used. In

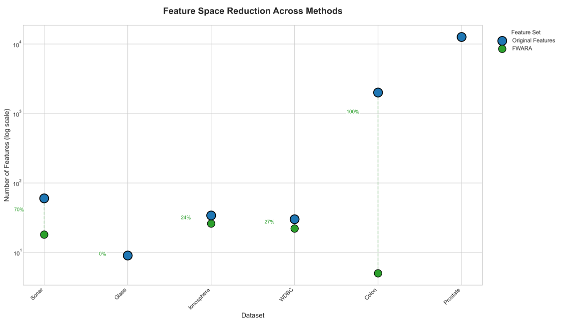
SVM Classifier Results



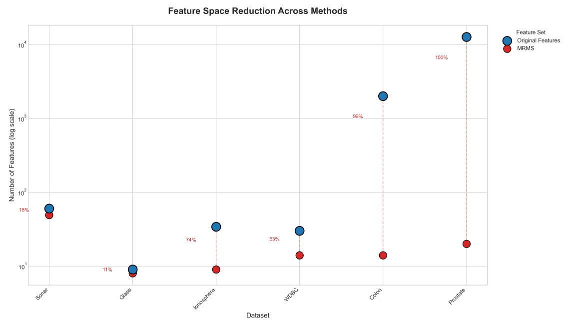
(a) FRFS Method



(b) FEAR Method

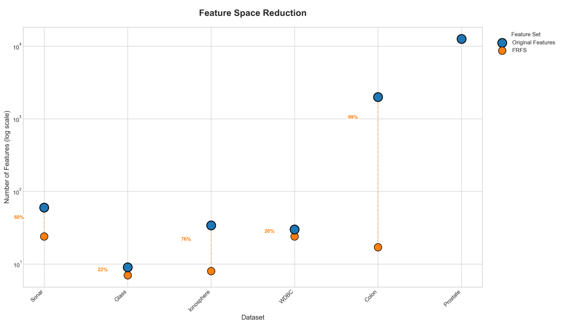


(c) FWARA Method

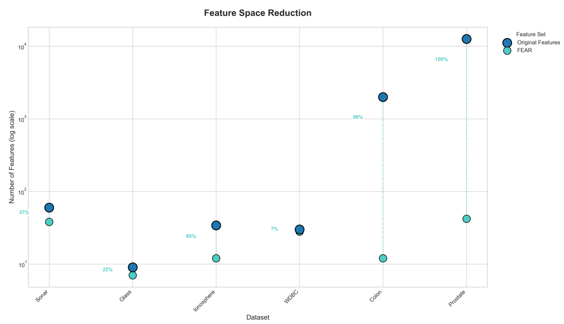


(d) MRMS Method

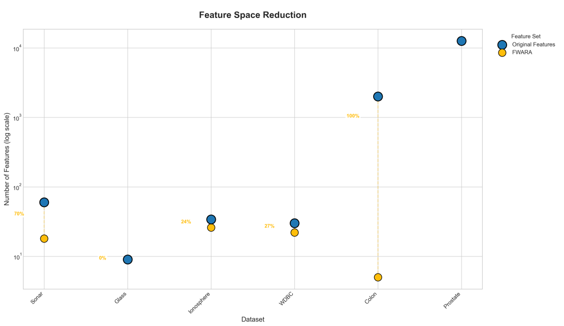
3NN Classifier Results



(e) FRFS Method



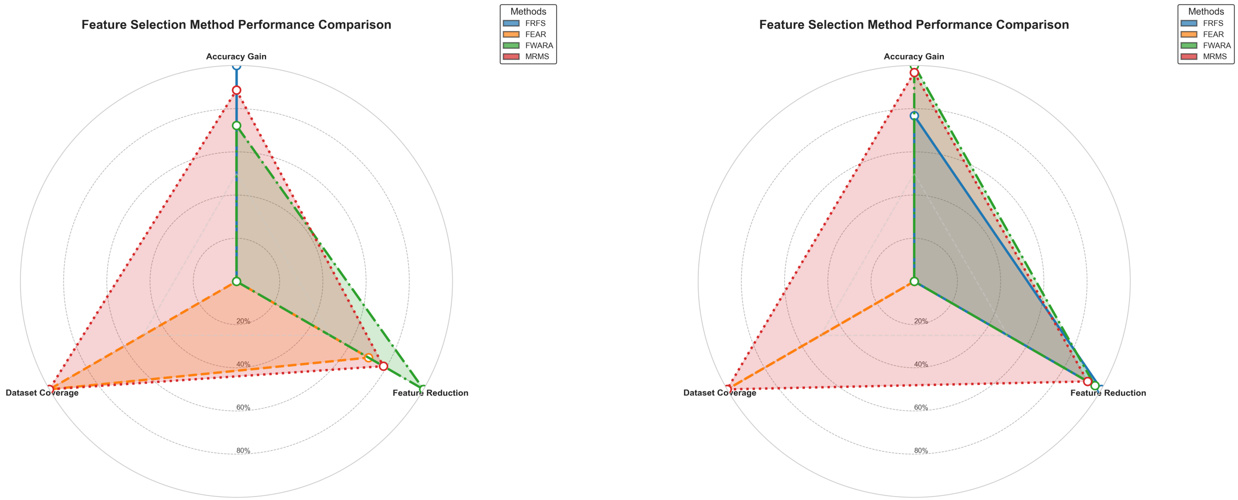
(f) FEAR Method



(g) FWARA Method

(h) MRMS Method

Figure 3. Feature reduction results of all methods for SVM and 3NN classifiers.



(a) SVM

(b) 3NN

Figure 4. Comparative performance of feature selection methods using SVM (a) and 3NN (b) classifiers.

contrast, MRMS is compatible with all the datasets that were studied and ensures consistent accuracy improvements for both classifiers (SVM: +3.1%, 3NN: +2.9%).

These results demonstrate that the optimal method depends on dataset characteristics. Methods like FRFS and FWARA tend to be more compatible with smaller datasets, while FEAR and MRMS generally serve as reliable default choices for diverse datasets. However, although FEAR is designed to handle larger datasets, its performance tends to decline as the significance calculated by fuzzy entropy becomes less distinct when the dataset size increases, leading to a potential loss of discriminatory power. with FRFS/FEAR preferred for accuracy-focused applications on small datasets, FWARA for maximal reduction when validated, and MRMS as a reliable default choice for diverse datasets.

4.3. Statistical significance test

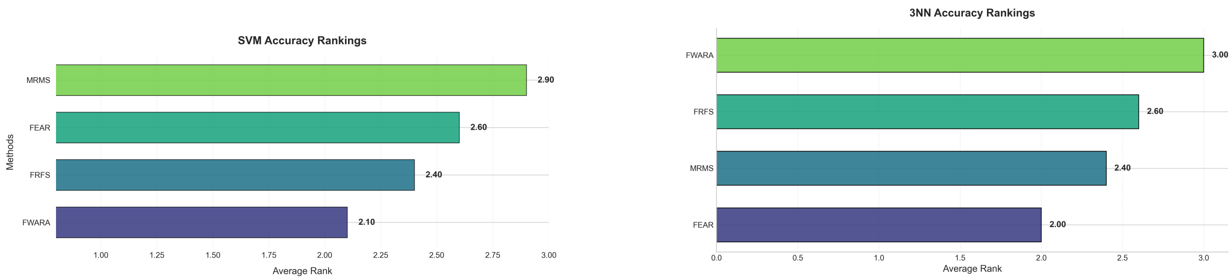
To evaluate whether the performance differences among the four feature selection methods (FRFS, FEAR, FWARA, MRMS) are statistically significant, we conducted the non-parametric Friedman test, followed by the computation of Kendall's coefficient of concordance (W) to assess the level of agreement in the rankings. The test was applied using both SVM and 3NN classifiers. The results are presented in Table 7.

Table 7. Comparative Friedman test results for SVM and 3NN classifiers

Classifier	Kendall's W	χ^2 (Q)	df	p-value
SVM	0.071	1.063	3	0.786
3NN	0.104	1.560	3	0.669

As shown in Table 7, for the SVM classifier, the Friedman test yields a test statistic of $\chi^2 = 1.063$ with 3 degrees of freedom and a p-value of 0.786. Similarly, for the 3NN classifier, the test yields $\chi^2 = 1.560$ with a p-value of 0.669. Both p-values are well above the commonly used significance threshold of 0.05, indicating that there is no statistically significant difference in the classification accuracies across the four feature selection methods.

Furthermore, the Kendall's W values are 0.071 for SVM and 0.104 for 3NN, which suggest a very weak level of agreement in the ranking of methods across datasets. This further supports the notion that the methods perform inconsistently across datasets and classifiers, and no single method consistently outperforms the others.



(a) SVM Rankings

(b) 3NN Rankings

Figure 5. Diagrams comparing average ranks of feature selection methods for SVM (a) and 3NN (b) classifiers.

These findings suggest that while the methods may differ in certain datasets, their overall performance is statistically comparable, and the choice of method may need to be guided by dataset-specific characteristics rather than global performance superiority.

5. Conclusion

While numerous techniques for fuzzy rough feature selection have been developed, ongoing research into fuzzy rough sets remains crucial to address the evolving challenges posed by real-world applications. This paper reviews some of these methods, highlighting their significance, potential, and the advancements that this field has witnessed. The inherent complexity of real-world problems necessitates continuous refinement and advancement in fuzzy rough set theory (FRST) and its associated feature selection methodologies. Despite being a relatively new field, FRST has already demonstrated impressive capabilities, particularly in its ability to process diverse data types—discrete, continuous, or mixed—without requiring additional information. By relying solely on the intrinsic structure of the data, FRST proves its strength in extracting meaningful information with minimal representation of knowledge in data, making it a highly effective tool for dimensionality reduction. Its capacity to preserve information in the data and improve the quality of data underscores its potential as a robust and versatile approach to feature selection, paving the way for further exploration and innovation in this domain.

Acknowledgement

The authors are in agreement with the finalised manuscript version for publication. We extend our appreciation to the reviewers for their utilisation of time and provision of insightful commentary in the evaluation of our research.

REFERENCES

1. Isabelle Guyon and André Elisseeff, *An introduction to variable and feature selection*, JMLR, vol. 3, pp. 1157–1182, 2003.
2. B. Mehdi, H. Chamlal, T. Ouaderhman, et al., *Intelligent credit scoring system using knowledge management*, IAES International Journal of Artificial Intelligence, vol. 8, no. 4, pp. 391, 2019.
3. K. Huang, S. Aviyente, *Wavelet feature selection for image classification*, IEEE Transactions on Image Processing, vol. 17, no. 9, pp. 1709-1720, 2008.

4. H. Chamlal, F. Aaboub, T. Ouaderhman, *A preordonance-based decision tree method and its parallel implementation in the framework of Map-Reduce*, Applied Soft Computing, vol. 167, pp. 112261, 2024. https://doi.org/10.1016/j.asoc.2024.112261
5. J. Komorowski, Z. Pawlak, L. Polkowski, and A. Skowron, *Rough sets: A tutorial*, Rough Fuzzy Hybridization: A New Trend in Decision-Making, pp. 3-98, 1999.
6. D. Dubois, H. Prade, *Rough fuzzy sets and fuzzy rough sets*, International Journal of General System, vol. 17, no. 2-3, pp. 191-209, 1990.
7. De Cock, M., Kerre, E.: On (un)suitable fuzzy relations to model approximate equality. Fuzzy Sets and Systems 133(2), 137–153 (2003).
8. F. Aaboub, H. Chamlal, and T. Ouaderhman, *Statistical analysis of various splitting criteria for decision trees*, Journal of Algorithms & Computational Technology, vol. 17, p. 17483026231198181, 2023.
9. F. Z. Janane, T. Ouaderhman, and H. Chamlal, *A filter feature selection for high-dimensional data*, Journal of Algorithms & Computational Technology, vol. 17, p. 17483026231184171, 2023.
10. L. A. Zadeh, *The role of fuzzy logic in the management of uncertainty in expert systems*, Fuzzy Sets and Systems, vol. 11, no. 1-3, pp. 199-227, 1983.
11. R. Jensen and Q. Shen, *Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches*, IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 12, pp. 1457-1471, 2004.
12. J. J. Buckley and E. Eslami, *An Introduction to Fuzzy Logic and Fuzzy Sets*, Springer Science and Business Media, 2002.
13. L. A. Zadeh, *Fuzzy sets*, Information and Control, vol. 8, no. 3, pp. 338–353, 1965.
14. Z. Pawlak, *Rough sets*. International journal of computer & information sciences, vol. 11, pp. 341-356, 1982.
15. Z. Pawlak, *Rough sets: Theoretical aspects of reasoning about data* (Vol. 9), Springer Science & Business Media, 2012.
16. C. Shannon, *A mathematical theory of communication*, Bell Syst. Tech. J., vol. 27, pp. 379-423, 1948.
17. A. D. Luca and S. Termini, *Entropy of L-fuzzy set*, Information and Control, vol. 24, no. 1, pp. 55-73, 1974.
18. A. Chouchoulas and Q. Shen, *Rough set-aided keyword reduction for text categorization*, Applied Artificial Intelligence, vol. 15, no. 9, pp. 843-873, 2001.
19. Y. Yang, D. Chen, H. Wang, and X. Wang, *Incremental perspective for feature selection based on fuzzy rough sets*, IEEE Transactions on Fuzzy Systems, vol. 26, no. 3, pp. 1257-1273, 2017.
20. R. Jensen and Q. Shen, *New approaches to fuzzy-rough feature selection*, IEEE Transactions on Fuzzy Systems, vol. 17, no. 4, pp. 824-838, 2008.
21. D. Chen, L. Zhang, S. Zhao, Q. Hu, and P. Zhu, *A novel algorithm for finding reducts with fuzzy rough sets*, IEEE Transactions on Fuzzy Systems, vol. 20, no. 2, pp. 385-389, 2011.
22. P. Maji and P. Garai, *On fuzzy-rough attribute selection: criteria of max-dependency, max-relevance, and min-redundancy, and max-significance*, Applied Soft Computing, vol. 13, no. 9, pp. 3968-3980, 2013.
23. H. Peng, F. Long, and C. Ding, *Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226-1238, 2005.
24. C. Wang, et al., *A fitting model for feature selection with fuzzy rough sets*, IEEE Transactions on Fuzzy Systems, vol. 25, no. 4, pp. 741-753, 2016.
25. X. Zhang, et al *A fuzzy rough set-based feature selection method using representative instances*, Knowledge-Based Systems 151 (2018): 216-229.
26. J. Wan, H. Chen, T. Li, X. Yang, and B. Sang, *Dynamic interaction feature selection based on fuzzy rough set*, Information Sciences, vol. 581, pp. 891-911, 2021.
27. J. Wan, H. Chen, T. Li, X. Yang, and B. Sang, *Dynamic interaction feature selection based on fuzzy rough set*, Information Sciences, vol. 581, pp. 891-911, 2021.
28. L. Dong, R. Wang, and D. Chen, *Incremental feature selection with fuzzy rough sets for dynamic data sets*, Fuzzy Sets and Systems, vol. 467, p. 108503, 2023.
29. J. Zhao, D. Wu, J. Wu, W. Ye, F. Huang, J. Wang, and E. W. K. See-To, *Consistency approximation: Incremental feature selection based on fuzzy rough set theory*, Pattern Recognition, p. 110652, 2024.
30. P. Luukka, *Feature selection using fuzzy entropy measures with similarity classifier*, Expert Systems with Applications, vol. 38, no. 4, pp. 4600-4607, 2011.
31. H. Chamlal, A. Benzmane, and T. Ouaderhman, *Elastic net-based high dimensional data selection for regression*, Expert Systems with Applications, vol. 244, p. 122958, 2024.
32. F. Aaboub, H. Chamlal, and T. Ouaderhman, *Analysis of the prediction performance of decision tree-based algorithms*, 2023 International Conference on Decision Aid Sciences and Applications (DASA), Annaba, Algeria, 2023.
33. T. Ouaderhman, H. Chamlal, and F. Z. Janane, *A new filter-based gene selection approach in the DNA microarray domain*, Expert Systems with Applications, vol. 240, p. 122504, 2024.
34. C. L. Blake and C. J. Merz, *UCI Repository of Machine Learning Databases*, 1998. [Online]. Available: http://archive.ics.uci.edu/
35. Q. Hu, D. Yu, and Z. Xie, *Information-preserving hybrid data reduction based on fuzzy-rough techniques*, Pattern Recognition Letters, vol. 27, no. 5, pp. 414-423, 2006.
36. A. Beck, and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences, vol. 2, no. 1, pp. 183–202, 2009.
37. J. Bioucas-Dias, and M. Figueiredo, *A new TwIST: Two-step iterative thresholding algorithm for image restoration*, IEEE Transactions on Image Processing, vol. 16, no. 12, pp. 2992–3004, 2007.
38. E. Candés, and Y. Plan, *Near-ideal model selection by L1 minimization*, Annals of Statistics, vol. 37, pp. 2145–2177, 2008.

39. E. Candés, and J. Romberg, *Practical signal recovery from random projections*, Wavelet Applications in Signal and Image Processing XI, Proc. SPIE Conf. 5914, 2004.
40. E. Candés, J. Romberg, and T. Tao, *Robust uncertainty rinciples: Exact signal reconstruction from highly incomplete frequency information*, IEEE Transactions on Information Theory, vol. 52, no. 2, pp. 489–509, 2006.
41. A. Chambolle, and P. L. Lions, *Image recovery via total variation minimization and related problems*, Numerische Mathematik, vol. 76, pp. 167–188, 1997.
42. T. F. Chan, and S. Esedoglu, *Aspects of total variation regularized ℓ_1 function approximation*, SIAM Journal on Applied Mathematics, vol. 65, pp. 1817–1837, 2005.
43. T. F. Chan, S. Esedoglu, F. Park, and A. Yip, *Total variation image restoration: Overview and recent developments*, in Handbook of Mathematical Models in Computer Vision, edited by N. Paragios, Y. Chen, and O. Faugeras, Springer-Verlag, New York, pp. 17–31, 2006.
44. D. Donoho, *Compressed sensing*, IEEE Transactions on Information Theory, vol. 52, no. 4, pp. 1289–1306, 2006.
45. T.G. Dietterich *Approximate statistical tests for comparing supervised classification learning algorithms*, Neural computation 10.7, 1895-1923 (1998).
46. J. Vanschoren, J.N.Van Rijn, B.Bischl, L. Torgo *OpenML: networked science in machine learning* ACM SIGKDD Explorations Newsletter 15(2), 49–60 (2014). https://www.openml.org/