Knowledge and Information Systems
https://doi.org/10.1007/s10115-025-02341-5

REGULAR PAPER

Feature selection by utilizing kernel-based fuzzy rough set and entropy-based non-dominated sorting genetic algorithm in multi-label data

Javad Hamidzadeh[1] · **Zahra Mehravaran**[2] · **Ahad Harati**[2]

Received: 16 August 2024 / Revised: 6 December 2024 / Accepted: 6 January 2025

Abstract

Multi-label learning, which involves assigning multiple class labels to each instance, becomes increasingly complex when dealing with large-scale mixed datasets featuring high-dimensional feature spaces. These mixed datasets often involve a combination of numerical and categorical features, which exacerbate the challenges of multi-label learning by introducing additional layers of uncertainty and variability. Traditional classification methods, although effective in simpler scenarios, often fail to address these complexities resulting in significant errors. To overcome this, we have developed an entropy-based objective function that captures the intricate interplay between features and classes, while accounting for the inherent uncertainty of mixed data. This objective function explicitly accounts for the heterogeneous nature of mixed datasets, ensuring robust feature selection across diverse attribute types. To tackle these challenges, we propose a memetic algorithm that integrates fuzzy rough sets with enhancements from kernel fuzzy rough sets (KFRS), and the Non-dominated Sorting Genetic Algorithm II. This synergy enables the extraction of optimal feature subsets that significantly improve classification performance. By leveraging kernel-based similarity measures, KFRS refines the partitions formed by fuzzy set memberships for distinct classes, ensuring precise alignment of data samples with multiple labels, while effectively handling the complexities of mixed-data representation. A key strength of our approach lies in its ability to preserve valuable information through KFRS-driven feature selection. Empirical evaluations on three benchmark datasets highlight the effectiveness of the proposed methodology. The results validate the superiority of our feature selection strategy, grounded in kernel-modulated neighborhoods; furthermore, the implementation demonstrates a notable improvement in both solution quality and search efficiency, establishing it as a highly promising method for multi-label learning tasks.

✉ Javad Hamidzadeh
J_Hamidzadeh@sadjad.ac.ir

Zahra Mehravaran
Zahra.mehravaran1993@mail.um.ac.ir

Ahad Harati
A.harati@um.ac.ir

[1] Faculty of Computer Engineering and Information Technology, Sadjad University, Mashhad, Iran

[2] Department of Computer Engineering, Ferdowsi University of Mashhad (FUM), Mashhad, Iran

Published online: 22 January 2025

Keywords Feature selection · Kernel fuzzy rough set · Mixed data · Multi-label learning

1 Introduction

In the digital age, the overwhelming influx of data from diverse domains underscores the critical need for effective dimensionality reduction. Within this context, feature selection stands out as a vital preprocessing step, carefully eliminating irrelevant and redundant features, to distill the essence of a dataset. This process not only enhances analytical clarity but also reduces the computational complexity of extracting knowledge from data [1, 2]. Maintaining the decision-making system's precision after feature pruning is essential, as it can significantly enhance the model's ability to generalize effectively.

In the complex landscape of feature selection methodologies, evaluation metrics and search strategies serve as the essential tools guiding the process. This spectrum includes traditional metrics such as dependency [3, 4], neighborhood dependency [5], and fuzzy dependency within various rough set paradigms [6–8], as well as mutual information [9, 10] and sample margin [11, 12] from statistical learning theory. The effectiveness of ReliefF as a scoring mechanism, highlighted by Spolaôr et al. [13], and the diverse score functions for text categorization, introduced by Yang and Pedersen [14], further demonstrate the depth and sophistication of this field. However, a lacuna often resides in these methodologies' ability to fully apprehend the labyrinth of label interconnections within multi-label contexts, as traditional feature selection methods frequently struggle to address the unique challenges inherent to multi-label learning, including the complex interplay among labels, the high dimensionality of feature spaces, and the presence of noisy or redundant features [15]. These limitations lead to suboptimal classification performance and reduced scalability when applied to real-world multi-label datasets. Additionally, existing methods often overlook the uncertainty and variability introduced by heterogeneous data types, which are prevalent in multi-label domains. Addressing these gaps necessitates innovative strategies that can not only capture intricate label–feature correlations but also enhance computational efficiency.

In recent years, label distribution learning (LDL) has emerged as an extension of multi-label learning, offering a richer representation of label–instance relationships by assigning a distribution rather than binary or mutually exclusive labels. This approach has shown improved performance in applications such as emotion recognition and human behavior analysis. However, LDL methods often come with higher computational demands and require comprehensive datasets with well-defined label distributions, which can limit their practicality in large-scale or heterogeneous scenarios [16]. This study focuses on multi-label learning to address real-world challenges such as high-dimensional data, mixed feature types, and label correlations, while maintaining computational efficiency and scalability. Although LDL represents an advanced paradigm for capturing label distributions, the proposed approach prioritizes adaptability and efficiency, making it more suitable for diverse and large-scale datasets encountered in practical applications.

Furthermore, many real-world multi-label datasets feature a mixed-data structure, encompassing both numerical and categorical features. This heterogeneity introduces further complexity, as traditional methods often fail to adequately process or represent such diverse data types. Numerical features may encode continuous measurements, while categorical features may represent discrete classifications, each posing distinct challenges for integration. Managing this variability and uncertainty requires advanced methodologies capable of handling the inherent diversity within mixed datasets.

Evolutionary-driven feature selection paradigms, renowned for their precise evaluation of feature subset fitness using classifiers have demonstrated significant value in improving classification accuracy [17]. However, their pursuit of finely optimized feature subsets that approach the global optimum is often hindered by substantial computational challenges. In this context, memetic algorithms emerge as a balanced solution, skillfully integrating exploration and exploitation, through refined optimization techniques [18, 19]. Furthermore, recent research has emphasized the effectiveness of dense network approaches combined with Gaussian optimizers, particularly in applications such as cardiovascular disease prediction [20].

The field of multi-label feature selection, encompassing areas as diverse as bioinformatics and financial data mining [21, 22], demands strategies capable of effectively eliminating irrelevant or redundant features in datasets with multiple labels per sample. This domain, richer in semantic complexity compared to binary or multi-class classification, must address the nuanced challenge of distinguishing between classes while managing the overlapping nature of nonexclusive labels. Positioning multi-label feature selection within the scope of multi-objective optimization highlights a balance between two competing objectives: maximizing classification accuracy and minimizing feature set size. Evolutionary algorithms (EAs) have achieved notable success in addressing these goals. However, despite their effectiveness, EAs often encounter significant computational challenges. To address this, memetic algorithms, with their enhanced efficiency and refined optimization capabilities, offer a promising solution to mitigate these computational limitations.

Positioning multi-label feature selection within the scope of multi-objective optimization highlights a balance between two competing objectives: maximizing classification accuracy and minimizing feature set size. Evolutionary algorithms (EAs) have achieved notable success in addressing these goals. However, despite their effectiveness, EAs often encounter significant computational challenges. To address this, memetic algorithms, with their enhanced efficiency and refined optimization capabilities, offer a promising solution to mitigate these computational limitations.

To address these critical challenges, this study introduces a novel methodology that leverages the power of kernel fuzzy rough sets (KFRS) to overcome the inherent limitations of existing approaches in multi-label feature selection. The proposed approach builds upon traditional fuzzy rough set strategies and incorporates advanced techniques to enhance feature selection in multi-label contexts, through the following key innovations:

- **A robust KFRS-based feature selection framework:** This framework is designed to identify and retain the most relevant features and label dynamics, optimizing classifier training and ensuring improved classification accuracy.
- **Entropy-driven uncertainty management:** By leveraging entropy-based measures, the method quantifies uncertainties and captures the complex correlations between labels and features in heterogeneous datasets, particularly those with mixed numerical and categorical attributes. Entropy serves as a key metric for evaluating variability and interdependencies, enabling the identification of features that significantly reduce uncertainty and enhance label coherence. This approach dynamically balances the competing objectives of relevance, redundancy minimization, and similarity maximization. Furthermore, entropy-based evaluations adapt to diverse data structures, ensuring robust feature selection and improved classification performance across mixed datasets.
- **Integrated representation of heterogeneous data:** The KFRS-based approach ensures precise clustering of samples by class while seamlessly handling the inherent diversity in multi-label datasets. By applying KFRS to orchestrate the data space, it enables the

robust representation of mixed-data types, including the seamless integration of numerical and categorical features, ensuring accurate and consistent feature selection across heterogeneous datasets.

Building on these innovations, the proposed method is specifically designed to address the pressing challenges in multi-label feature selection for mixed datasets. Traditional algorithms often fall short of effectively managing the heterogeneity of numerical and categorical features, leading to suboptimal feature selection and classification performance. Unlike existing methods, which primarily aim for general classification improvement, our approach uniquely leverages kernel fuzzy rough sets (KFRS) and entropy-driven objectives. This tailored strategy not only enhances the robustness of feature selection but also ensures effective handling of diverse data structures, minimizing redundancy while capturing intricate label–feature dependencies to improve overall classification outcomes.

The manuscript is structured as follows: Sect. 2 explores the landscape of related work, while Sect. 3 establishes the foundational concepts. Section 4 introduces the proposed kernel fuzzy rough set (KFRS) method and Sect. 5 presents the experimental results and discussions. Finally, Sect. 6 concludes with key findings and outlines directions for future research.

2 Related work

Multi-label learning is typically addressed through three main approaches: (1) problem transformation [23], (2) algorithm adaptation [24], and (3) ensemble learning [25]. Problem transformation involves breaking down a multi-label problem into one or more binary classification tasks, which can then be solved using conventional single-label classification techniques such as support vector machines (SVM) [26], the naive Bayes classifier [27], and others. Algorithm adaptation entails customizing traditional single-label classifiers to make them suitable for multi-label classification tasks. A notable example is MLKNN [28], an extension of the k-nearest neighbor algorithm with appropriate modifications for multi-label scenarios. Ensemble methods, by contrast, leverage a collection of multi-class classifiers, such as those utilizing multiple Label Powerset [29] classifiers [30] to form a unified learner. This approach improves learning stability by reducing the risk of model overfitting and decreasing sensitivity to initialization variations [31].

Furthermore, multi-label learning methods, based on label interdependencies, can be grouped into three categories. The first group disregards label relationships altogether [32–34]. The second group utilizes correlations between pairs of labels [35, 36], while the third group emphasizes the interrelationships among all class labels or specific subsets [37, 38]. The introduction of kernel fuzzy rough sets (KFRS) brings an advanced perspective, providing deeper insights into label correlations through kernel-based transformations and fuzzy logic, thereby enhancing algorithm adaptation strategies [39, 40]. Further advancements in the domain of fuzzy logic and feature selection include recent studies such as [41, 42], which emphasize the application of kernel-based transformations and entropy-driven techniques to improve feature selection accuracy. These studies offer valuable methodologies that align well with the challenges addressed in this research, particularly concerning the handling of multi-label data and the reduction of uncertainty in heterogeneous datasets. Building on these foundations, Fan et al. [42] proposed an adaptive fuzzy rough set model incorporating kernel functions to dynamically adjust feature evaluation criteria. This approach leverages multi-neighborhood structures to address variability in mixed-data environments, aligning closely with the principles of KFRS in managing complex label interdependencies.

Multi-label datasets often face the challenge of high dimensionality, as they frequently include samples with a large number of features [43]. As a result, multi-label feature selection methods are generally categorized into three main groups: filter methods, wrapper methods, and embedded methods. Filter methods select feature subsets using statistical techniques that evaluate the relationship between individual features, independent of the classifier being used. In contrast, wrapper methods involve selecting a feature subset, training a model with the selected features, and then using the model's performance to guide further feature inclusion or exclusion. Embedded methods blend aspects of both filter and wrapper approaches, creating an integrated framework for feature selection. Lin et al. [44] evaluated each feature based on the conditional redundancy between the candidate feature and the already selected features. Similarly, Lee and Kim [45] proposed a multi-label feature selection method leveraging multivariate mutual information. Their approach identifies feature subsets by maximizing the dependency between selected features and labels, using an incremental selection strategy. The subset is determined by maximizing the dependency between selected features and labels through an incremental selection strategy. Yan and Li [46] introduced a graph-margin-based multi-label feature selection method that calculates label correlations and evaluates features based on a graph structure, integrating this approach with large margin theory for enhanced performance. Sun et al. [47] proposed a mutual information-based multi-label feature selection method that incorporates label correlations and utilizes constrained convex optimization to achieve an optimal solution. Zhang and Li [48] developed an unsupervised feature selection approach by efficiently leveraging sparse fuzzy membership. Dong et al. [49] introduced a many-objective feature selection strategy for multi-label classification, utilizing the NSGA-III algorithm [50] to address the complexity of multiple objectives, effectively. To enhance the diversity and convergence of NSGA-III, novel crossover and mutation operators have been developed to boost its exploration capabilities. While these advancements have significantly improved NSGA-III's performance in multi-objective optimization, further progress has been made by incorporating advanced representation techniques and neighborhood learning strategies, leading to even more refined outcomes. Recent advancements in multi-objective optimization for feature selection have further refined the balance between exploration and exploitation. For instance, Yin et al. [15] developed a robust framework combining sparse representation and neighborhood reconstruction for multi-label classification. By integrating fuzzy membership functions and neighborhood learning, this method enhances the robustness of feature evaluation and subset selection, particularly in handling label correlations.

The combination of sparse regression and spectral graph techniques has emerged as a widely used method for subspace learning in feature selection [51]. In this framework, each sample is regressed onto its specific manifold structure. Huang and Wu [52] proposed a multi-label feature selection method that integrates manifold regularization with dependence maximization, using an iterative optimization algorithm to derive sparse coefficients for feature selection. Building on this, Kong et al. [53] incorporated manifold learning into methods [54, 55] to effectively handle multi-label data. Fan et al. [56] further developed a manifold framework by employing an uncorrelated regression model to identify features that are both uncorrelated and discriminative. These features are utilized to capture the label distribution, while a spectral graph component based on information entropy is seamlessly integrated into the framework to preserve local geometric data structure during subsequent learning.

Xia et al. [57] introduced a sparsity-regularized weighted stacked ensemble approach to optimize classifier selection and the construction of ensemble members for multi-label classification. The weights of the ensemble members are assigned based on pairwise label

correlations. An optimal ensemble solution is achieved using an optimization algorithm that integrates accelerated proximal gradient, and block coordinate descent methods.

Gonzalez-Lopez et al. [58] proposed a distributed model for evaluating feature quality across multiple labels, aggregating mutual information through Euclidean norm maximization (ENM) and geometric mean maximization (GMM) approaches. ENM emphasizes features with the largest L_2 norm, while GMM selects those with the highest geometric mean. Liu et al. [59] focused on inter-class discrimination and intra-class neighbor recognition to identify the most relevant and discriminative features, which are then combined using a feature conversion technique.

In the domain of filter methods, identifying the optimal attributes for feature evaluation is a key challenge. Various criteria have been studied, including techniques like ReliefF and its extensions [60]. The dependency criterion, a classic evaluation metric, also plays a significant role in multi-label feature selection. Qian et al. [61] advanced this area by developing a ranking-based feature selection method that leverages fuzzy relative discernibility, and fuzzy label discernibility relations for multi-label classification.

3 Preliminary

3.1 Kernel fuzzy rough sets

Kernel fuzzy rough sets (KFRS) play a pivotal role in refining feature selection by enhancing the granularity of similarity calculations. Unlike traditional fuzzy rough sets, which rely on fixed neighborhood relationships, KFRS leverages kernel functions to dynamically adapt neighborhood boundaries based on feature correlations. Kernel methods are a class of machine learning algorithms and have been extensively applied across a wide range of problems [62–65]. This flexibility enables KFRS to address the inherent uncertainty and heterogeneity in multi-label datasets, particularly those with mixed numerical and categorical attributes. The fuzzy relationship between samples is established using kernel-based similarity measures, which compute the degree of similarity between pairs of samples. This relationship is characterized by:

- **Reflexivity:** Ensuring each sample is maximally similar to itself.
- **Symmetry:** Guaranteeing that similarity is bidirectional between any two samples.
- **Continuity:** Allowing for smooth transitions in similarity values across the dataset. These properties enable the formation of fuzzy granules that represent local sample neighborhoods. These granules are subsequently used to calculate the lower and upper approximations within the KFRS framework, thereby capturing both precise and potential memberships of samples in decision classes.

Suppose that $NDS = \langle U, AT, C, V, f, \delta, K \rangle$ is a kernel-based neighborhood decision system. A universal set $U = \{x_1, x_2, \ldots, x_m\}$ is a set of samples. AT and C refer to the sets of conditional attributes and decision attributes, respectively. The mapping function is represented as $V = \bigcap_{a \in AT} V_a$ where $f : U \times A \rightarrow V$, $A = \{AT \cup C\}$ encompasses all attributes of samples, and V_a signifies the domain of attribute a. The attribute value of sample x for attribute a is denoted by $f(a, x)$, while $\delta (0 \leq \delta \leq 1)$ serves as a distance threshold, signifying the radius of the neighborhood. The kernel function is denoted K.

3.2 Kernel-induced similarity

In the context of $NDS = \langle U, AT, C, V, f, \delta, K \rangle$, when $B \subseteq AT$, it gives rise to the kernel-based fuzzy binary relation denoted as R_B^k, on the set U. Here, K represents the kernel function utilized to compute the kernel-based fuzzy binary relation R_B^k, This function plays a crucial role in refining similarity assessments within the multi-label data environment. B denotes a subset of the conditional attributes AT selected for assessing similarity in the kernel fuzzy rough set framework. This subset guides the calculation of the kernel-based fuzzy similarity relation. For any $x, y \in U$, the relation R_B^k is deemed the kernel fuzzy similarity relation if it fulfills the subsequent properties:

$$R_B^k(x, x) = 1 \qquad \text{for any } x \in U, \tag{1}$$

$$R_B^k(x, y) = R_B^k(y, x) \qquad \text{for any } x, y \in U. \tag{2}$$

where $S_B^k(x, y)$ is the similarity degree between x and y under B and threshold $\delta \in [0, 1]$ is the kernel-based fuzzy neighborhood radius parameter which controls the similarity degree between two samples x and y, calculated using the kernel function k. The kernel fuzzy similarity relation R_B^k acts as the foundational metric for measuring similarity between any two samples, leveraging kernel functions to capture intricate patterns and relationships in mixed-data environments. Its properties of reflexivity and symmetry ensure robustness and consistency in similarity evaluation. Building on R_B^k, the neighborhood structure S_B^k defines a localized similarity granulation by considering samples within a threshold-defined kernel-induced neighborhood. This granulation enables the kernel fuzzy rough set framework to establish lower and upper approximations that accurately capture the uncertainty and variability of multi-label datasets. Together, R_B^k and S_B^k synergistically enhance the feature selection process by refining similarity assessments and ensuring the robustness of neighborhood definitions, which are pivotal for managing mixed-data attributes.

3.3 Kernel-based information granulation

In the kernel-augmented neighborhood decision system $NDS = \langle U, AT, C, V, f, \delta, K \rangle$ for each $a \subseteq AT$ and $x, y \in U$, we define the granule of information around x with respect to B as:

$$[x]_B^\delta(y) = \begin{cases} 1 & \text{if } S_B^k(x, y) > \delta, \\ 0 & \text{else.} \end{cases} \tag{3}$$

This granule represents whether x and y are considered similar within the kernel-modified neighborhood defined by δ.

3.4 Kernel approximations of fuzzy decisions

In the framework of $NDS = \langle U, AT, C, V, f, \delta, K \rangle$, where $B \subseteq AT$, and for any $X \subseteq U$, the parameterized kernel fuzzy neighborhood information granule for $x \subseteq U$ is denoted as $\alpha_B^k(x)$. Consequently, the kernel fuzzy neighborhood lower and upper approximations of the fuzzy decision C concerning B are defined by formulas 4 and 5, respectively:

$$R_B^{k,\delta_{\text{lower}}}(C)(x) = \inf_{y \in U} \max\left(1 - R_B^k(x, y), C(y)\right), \tag{4}$$

$$R_B^{k,\delta_{\text{upper}}}(C)(x) = \sup_{y \in U} \min\left(1 - R_B^k(x, y), C(y)\right). \tag{5}$$

In these definitions, $R_B^{k,\delta_{\text{lower}}}(C)(x)$ captures the extent to which elements are certainly within the decision class C, while $R_B^{k,\delta_{\text{upper}}}(C)(x)$ captures the extent to which elements could be within C.

where $R_B^{k\_\text{lower}}(x, y)$ is calculated using the kernel function $k.R_B^{k,\delta}(C)(x)$. measures the degree of certainly belonging to the fuzzy decision C and $R_B^{k,\delta}(C)(x)$. measures the degree of x, possibly belonging to the fuzzy decision C. The boundary region of X is defined as Eq. (6).

$$BN(X) = R_B^{k,\delta_{\text{upper}}}(C)(x)) - R_B^{k,\delta_{\text{lower}}}(C)(x) \tag{6}$$

The kernel fuzzy neighborhood approximation accuracy of C in relation to B is calculated by Eq. (7).

$$acc_B^k = \frac{R_B^{k,\delta_{\text{lower}}}(C)(x)}{R_B^{k,\delta_{\text{upper}}}(C)(x)} \tag{7}$$

This accuracy metric reflects how closely the lower approximation covers the decision class C in comparison to the upper approximation.

Regarding the kernel fuzzy neighborhood roughness degree, it is obtained by considering the uncertainty between the lower and upper approximations:

$$\text{FNR(X)} = 1 - acc_B^k \tag{8}$$

The roughness degree $FNR(X)$ measures the amount of uncertainty or fuzziness present in the classification with respect to C and $B.A$ higher $FNR(X)$ value indicates greater uncertainty, while a value closer to 0 indicates less uncertainty.

4 Proposed method

Our study introduces a novel multi-label feature selection algorithm that leverages an advanced kernel-based fuzzy rough set approach, enhancing the Non-dominated Sorting Genetic Algorithm II (NSGA-II). The goal is to construct a Pareto set of non-dominated solutions, represented by feature subsets X_i. These subsets are designed to exhibit a strong correlation with class labels through kernel-based similarity measures while minimizing redundancy using fuzzy rough set boundaries. The proposed method integrates the strengths of genetic algorithms and fuzzy rough sets through a synergistic approach. The genetic algorithm (specifically NSGA-II) is employed to explore the solution space efficiently, ensuring the identification of Pareto-optimal feature subsets that maximize classification relevance and minimize redundancy. Concurrently, fuzzy rough sets refine the evaluation of feature subsets by dynamically capturing uncertainty and relationships in the dataset. This integration leverages kernel functions to enhance neighborhood definitions, enabling the genetic algorithm to operate with greater precision in mixed-data environments. With this in mind, our method aims to balance three critical objectives: (1) maximizing feature relevance, (2) minimizing redundancy, and (3) capturing similarity in mixed-data environments. By integrating kernel functions with fuzzy rough sets, our method dynamically adapts to the heterogeneous

nature of multi-label datasets, ensuring robust feature selection. The key contributions of our approach include:

- Leveraging kernel-based similarity measures to refine feature relevance and minimize redundancy.
- Employing entropy-driven objectives to quantify and address label–feature dependencies.
- Using NSGA-II within a kernel-augmented framework to efficiently explore the solution space.

This section is organized as follows: Sect. 4.1: Details on how features are encoded. Section 4.2: Explanation of multi-objective optimization criteria. Section 4.3: Steps of the kernel-augmented NSGA-II algorithm. Section 4.4: Dynamic adjustments of feature subsets based on KFRS measures

4.1 Chromosome representation

Each chromosome represents a candidate solution, corresponding to a subset of features. *A* population P of these chromosomes is generated, where the binary representation is refined to encode the relevance of features, based on kernel fuzzy rough set (KFRS) principles. Unlike a simple binary indicator for feature presence or absence, each bit in the chromosome also reflects the feature's relevance and non-redundancy as assessed by kernel-based rough set approximations. The representation of chromosomes directly supports the objective of maximizing feature relevance and minimizing redundancy. By encoding relevance and redundancy metrics derived from KFRS, this representation ensures that candidate solutions prioritize essential features while discarding irrelevant or overlapping ones, facilitating robust multi-label feature selection.

4.2 Fitness function

The merit of each chromosome is ascertained through three competing objectives, reshaped to integrate KFRS evaluations:

$$f(x) = \left(f_{k_1}(x),\, f_{k_2}(x),\, f_{k_3}(x)\right) \tag{9}$$

These objectives are defined under the KFRS framework as follows:

- **Redundancy Minimization $f_{k_1}(x)$**: This objective minimizes redundancy by leveraging the lower and upper approximations in KFRS. It ensures that the selected feature subsets are compact and free from overlapping or irrelevant features.
- **Relevance Maximization $f_{k_2}(x)$**: This objective prioritizes features that demonstrate strong dependency on class labels. Using kernel-based dependency measures defined in Eq. 11, it evaluates the nuanced relationships between features and class labels in mixed datasets. In this formulation, entropy $H_k(C)$ quantifies the overall uncertainty of the class distribution, while the conditional entropy $H_k(C \mid X_l)$ measures the remaining uncertainty given a feature X_l. Their difference, mutual information, reflects the reduction in uncertainty achieved by incorporating the feature. This allows the fitness function to prioritize features that provide the most information gain, thereby enhancing the relevance of the selected subset.
- **Similarity Enhancement $f_{k_3}(x)$**: This objective evaluates the similarity between features and labels using kernel-induced similarity indices (Eq. 13). By incorporating

kernel-based measures, it dynamically adapts to the heterogeneity of the dataset, ensuring robust feature selection.

These objectives are seamlessly integrated into the NSGA-II framework to guide the evolutionary process. Each chromosome in the population represents a candidate feature subset, and the objectives are used to evaluate its fitness. The genetic operator's crossover, mutation, and selection are influenced by the kernel-enhanced evaluations, which dynamically refine neighborhood definitions and optimize feature selection. Additionally, the weights α and β in the fitness function are adjusted to balance dependency, redundancy, and similarity according to the dataset's characteristics, ensuring a tailored optimization process.

The fitness function is a core component of our methodology, as it formalizes the balance among competing objectives. By incorporating kernel-based relevance, redundancy, and similarity measures, it dynamically adapts to the mixed nature of datasets. This ensures the selection of feature subsets that not only align with class labels but also maintain compactness and inter-feature harmony, enhancing classification performance.

$$J_k(x) = \max I_k(X; C) + \alpha \sum_{\substack{x \subseteq S, \\ X' \subseteq \S'}} I_k(X; X') - \beta \sum_{\substack{x \subseteq S, \\ X' \subseteq \S'}} I_k(X; C \mid X') \tag{10}$$

$I_k(X_i; C)$ is computed by Eq. 11.

$$\begin{aligned} I_k(X_i; C) &= H_k(C) - H_k(C \mid X_i) \\ &= -\sum_{c_j \in C} p_k(c_j) \log p_k(c_j) \\ &\quad + \sum_{c_j \in C} \sum_{X_i \in s'} p_k(c_j, X_i) \log p_k(c_j, X_i) \\ &= \sum_{c_j \in C} \sum_{x_i \in s'} p_k(c_j, X_i) \frac{p_k(c_j, X_i)}{p_k(c_j)} \end{aligned} \tag{11}$$

$I_k(X_S; X_i)$ is computed by Eq. 12.

$$\begin{aligned} I_k(X_S; X_i) &= H_k(X_i) - H_k(X_i \mid X_S) \\ &= -\sum_{X_i \in s'} p_k(X_i) \log p_k(X_i) + \sum_{X_i \in s'} \sum_{X_s \in S} p_k(X_i, X_S) \log p_k(X_i \mid X_S) \\ &= \sum_{X_i \in S'} \sum_{X_s \in S} p_k(X_i, X_S) \log \frac{p_k(X_i \mid X_S)}{p_k(X_i)} \end{aligned} \tag{12}$$

$I_k(X_S, C \mid X_i)$ is computed by Eq. 13.

$$\begin{aligned} I_k(X_S, C \mid X_i) &= H_k(X_S \mid X_i) - H_k(X_S \mid C, X_i) \\ &= -\sum_{X_S \in S} \sum_{c_j \in C} \left(p_k(X_S, C) \log p_k(X_S \mid C) \right) \\ &\quad + \sum_{X_S \in S} \sum_{c_j \in C} \sum_{X_i \in S'} p_k(X_S, C, X_i) \log p_k(X_S \mid C, X_i) \\ &= \sum_{X_S \in S} \sum_{c_j \in C} \sum_{X_i \in S'} p_k(X_S, C, X_i) \log \frac{p_k(X_S \mid C, X_i)}{p(X_S \mid C)} \end{aligned} \tag{13}$$

4.3 Search method

To determine the optimal feature sets that maximize the objectives of the kernel fuzzy rough set (KFRS) framework, as defined in Eq. 10, this study employs an enhanced Non-dominated Sorting Genetic Algorithm II (NSGA-II). This improved version integrates KFRS measures for evaluating feature subsets. The algorithm begins by generating an initial population and computing the kernel-based objectives for each chromosome. Before entering the main iterative loop, the population is partitioned into Pareto fronts using fast non-dominated sorting. The crowding distance is then calculated, incorporating kernel-based measures for each chromosome within the fronts. The integration of KFRS within the NSGA-II framework ensures a refined exploration of the solution space. By leveraging kernel-based measures during sorting and refinement, the search method achieves a robust balance between exploration and exploitation, enabling the identification of Pareto-optimal solutions that align with the method's multi-objective optimization goals. The proposed memetic algorithm, strengthened by local refinement through KFRS, is implemented according to Algorithms 1–3.

Algorithm 1 The proposed kernel-augmented multi-objective memetic algorithm

1: Create a random population of N chromosomes.
2: Calculate the kernel-based multi-objective fitness (using adaptations of Eq. 10) for each chromosome.
3: Determine the ranking using kernel-aware Pareto fast non-dominated sorting.
4: Calculate the kernel-informed crowding distance.
5: Generate offspring population.
6: **while** termination criteria are not met **do**
7: Apply local refinement using kernel fuzzy rough sets.
8: Execute the elitism selection technique, respecting kernel-based evaluations.
9: Perform crossover and mutation, guided by kernel-based fitness.
10: Evaluate objectives using a multi-label classifier integrated with KFRS.
11: Rank individuals using kernel-aware Pareto sorting.
12: Compute the kernel-informed crowding distance.
13: **end while**
14: Return final population P, reflecting kernel-based feature evaluations.

Algorithm 2 Kernel-informed fast non-dominated sorting

1: **Initialize:** For each solution p in the population P, set the domination counter $n_p = 0$ and the set of solutions that p dominates $S_p = \emptyset$.
2: **Domination Counting:** For each pair of solution p, q in P, if p kernel dominates q, then add q to set S_p. Otherwise, if q kernel dominates p, increment domination counter n_p.
3: **Create the First Front:** All solutions in P with $n_p = 0$ are non-dominated and form the first front F_1. Set their rank $r_p = 1$.
4: **Non-dominated Sorting:** For each front F_i, initialize the next front $F_{i+1} = \emptyset$. For each solution p in F_i, for each solution q in S_p, decrement n_q. If n_q becomes 0, add q to F_{i+1} and set $r_q = i + 1$.
5: **Iteration:** Continue this process until all fronts are created.

While NSGA-II excels at navigating complex search spaces, it does not inherently ensure global optimality. To address this limitation, we introduce a local search strategy grounded in kernel fuzzy rough sets (KFRS). This approach enhances the search process by leveraging the fine-grained granularity of kernel-based fuzzy neighborhoods, enabling more precise clustering of samples within the class vicinity, in the feature space. This refinement strength-

Algorithm 3 Kernel-Based Crowding Distance

Input:
M: The number of kernel-based objective functions.
Z: The number of non-dominated solutions.
F: A Pareto front composed of N individuals.
$f_m^{\max / \min}$: max(min) value for objective m under KFRS.

Let $F[z]_{\text{dist}} = 0$ for $z = 1, \ldots, Z$.
for each kernel-based objective function $f_m, m = 1, 2, \ldots, M$ **do**
  Sort the set in ascending order.
  Let $F[0]_{\text{dist}} = F[z]_{\text{dist}} = \infty$.
  for $i = 2$ to $Z - 1$ **do**
    $f[z]_{\text{dist}} = f[z]_{\text{dist}} + \frac{f[z+1].m - f[z-1].m}{f_m^{\max} - f_m^{\min}}$.
  end for
end for

ens exploration in regions with sparse solutions while improving exploitation across various criteria.

Kernel fuzzy rough sets delineate neighborhoods with membership functions that are sensitive to the fuzzy boundaries of classes, as identified by kernel-induced similarity relations. These functions enable a more flexible conception of neighborhoods, taking into account the degree of similarity computed via kernel functions. The resulting membership distributions provide a soft partitioning of the data space, reflective of the multi-label nature of the dataset. This flexibility allows KFRS to adapt similarity metrics during local refinement, ensuring they are both context-aware and robust. By computing kernel-induced fuzzy memberships, the method dynamically adjusts neighborhood definitions, effectively managing mixed-data structures. Consequently, these refined neighborhoods enhance the discrimination of relevant features while minimizing the influence of redundant or noisy ones. To further enhance the search efficacy of NSGA-II, the proposed method addresses its inherent limitations regarding global optimality. By integrating KFRS, the method employs adaptive kernel-based measures, to refine the exploration and exploitation phases. Kernel fuzzy rough sets (KFRS) enhance the genetic algorithm by refining its local search capabilities. During each iteration, KFRS measures are used to adjust neighborhood definitions dynamically, improving the algorithm's ability to differentiate between essential and redundant features. The kernel-induced fuzzy relationships provide adaptive guidance for crossover and mutation operations, ensuring the population evolves toward globally optimal feature subsets. This refinement dynamically adjusts the similarity metrics and neighborhood definitions, enabling the algorithm to explore underrepresented regions of the feature space while avoiding premature convergence. As a result, the proposed strategy ensures a more comprehensive and robust search, leading to a set of Pareto-optimal solutions that better approximate global optima.

During the local refinement phase, feature subsets (candidate solutions) are dynamically adjusted toward the Pareto front, guided by Eq. 10. This equation incorporates kernel-based measures, optimizing not only for feature distinctiveness but also for the smoothness of transitions across class boundaries.

To evaluate the degree of neighborhood belongingness to a specific class, a neighborhood relation is defined based on KFRS principles. Unlike a binary classification, this relation employs a fuzzy measure that captures both the proximity and density of points within the kernel-transformed feature space. This fuzzy measure enables a more nuanced and locally aware progression, facilitating refined adjustments toward Pareto-optimal solutions. To evaluate the degree of neighborhood belongingness to a specific class, a neighborhood relation

is defined based on KFRS principles. Unlike a binary classification, this relation employs a fuzzy measure that captures both the proximity and density of points within the kernel-transformed feature space. This fuzzy measure enables a more nuanced and locally aware progression, facilitating refined adjustments toward Pareto-optimal solutions. The kernel-based neighborhood relation, which forms the foundation of this approach, is defined as follows:

Given $NDS = \langle U, S, C, V, f, \delta, K \rangle$, a feature subset $S' \subseteq S$, and any two samples $x, y \subseteq U$, the kernel-based neighborhood relation among data samples under the selected feature subset S' is established according to Eq. (14).

$$NR_{\delta}^{k}(S') = \{(x, y) \in U \times U \mid \kappa(x, y; S') \geq \tau\} \tag{14}$$

where $\kappa(x, y; S')$ is a kernel-based similarity function that incorporates the information of the feature subset S' and τ is a threshold parameter derived from δ. $NR_{\delta}^{k}(S')$ must satisfy the properties of reflexivity and symmetry, which are inherent to kernel functions. The kernel-based neighborhood class δ_k is according to Eq. 15.

$$\delta_k = \max(\Delta_{S'}^{k}(x, NH_K(x)), \Delta_{S'}^{k}(x, NM_K(x))) \tag{15}$$

The similarity $\Delta_{S'}^{k}(x, y)$ as defined in Eq. 16.

$$\Delta_{S'}^{k}(x, y) = \sqrt{\sum_{a_k \in S'} \kappa(f_{a_k}(x), f_{a_k}(y))^2} \tag{16}$$

where $a_k \in S'$, and $|S'|$ denotes the cardinality of the feature subset S'. Also, $NH(x)$ represents the nearest sample within the same class label, and $NM(x)$ signifies the nearest sample belonging to a different class label.

In this KFRS approach for each label k of a given sample x, both the closest sample within the same class label $NH_k(x)$, and the nearest sample from a different class label $NM_k(x)$, are identified via the kernel function. These identifications are used to establish neighborhood relations and neighborhood classes in a fuzzy rough manner.

4.4 Solution evolution

As described in Lines 8–12 of Algorithm 1, each feature within a chromosome is assigned a score reflecting its kernel-based relevance. The initial chromosome is then refined using kernel fuzzy rough set measures, enhancing the selection of essential features, while eliminating redundant ones. If the refined chromosome achieves an improved fitness value, as determined by the kernel-enhanced fitness function, it is reintegrated into the population. Ultimately, chromosomes with higher kernel-based fitness values are prioritized as superior solutions. The solution evolution process leverages KFRS to iteratively refine feature subsets, ensuring robust feature evaluation and improved convergence. This dynamic adjustment of neighborhood definitions and similarity metrics addresses the inherent complexity of multi-label datasets, resulting in a comprehensive and globally optimal feature selection strategy. The conceptual workflow of the proposed method, integrating kernel-based approaches within the fuzzy rough set framework, is illustrated in Fig. 1. The conceptual depiction of the proposed method, now incorporating kernel-based methods within the fuzzy rough set framework, is outlined in Fig. 1. Each step in the diagram plays a crucial role in the feature selection and classification process, ensuring a robust and efficient methodology for multi-label learning. Below, we provide a detailed explanation of the individual steps in the workflow: **Start and**

Parameter Initialization: The process begins with defining the necessary input parameters, including dataset characteristics, kernel parameters, and NSGA-II hyperparameters, such as population size, number of generations, mutation rate, and crossover probability. A random population of chromosomes is then initialized, where each chromosome represents a potential feature subset. **Fitness Function Evaluation:** For each chromosome, the fitness function evaluates the quality of the selected features by considering three objectives:

- **Relevance:** The correlation between the selected features and the labels is assessed to ensure the chosen features are meaningful for classification.
- **Redundancy:** The method aims to minimize redundancy among selected features to avoid overlap and ensure compact feature subsets.
- **Similarity:** By leveraging kernel-induced fuzzy rough sets (KFRS), the similarity between features and labels is captured, particularly in heterogeneous datasets with mixed attributes.

Non-Dominated Sorting and Crowding Distance Calculation: The population of chromosomes is ranked using non-dominated sorting to identify the Pareto fronts. This process distinguishes solutions that achieve optimal trade-offs among the three objectives. To preserve diversity in the population, a crowding distance calculation ensures that solutions with diverse feature sets are retained. **Stopping Criteria Check:** At this stage, the algorithm evaluates whether the stopping criteria, such as the maximum number of generations or the convergence of the Pareto front, have been met. If satisfied, the current Pareto front is finalized, and candidate feature subsets are passed to the next stage for classifier training and prediction. Otherwise, the search continues through crossover, mutation, and refinement steps. **Crossover and Mutation:** To further explore the solution space, the proposed method employs genetic operations, specifically crossover and mutation, which are integral to the evolutionary optimization process. Crossover involves combining segments of two parent chromosomes to create new offspring solutions, ensuring the inheritance of high-performing feature combinations while introducing variability into the population. This operation facilitates the exchange of genetic information between chromosomes, allowing the search to expand into promising regions of the solution space. Meanwhile, mutation introduces randomness by altering individual genes in a chromosome, enabling the exploration of underrepresented or unexplored areas of the feature space. This process not only diversifies the population but also reduces the risk of premature convergence to suboptimal solutions, enhancing the algorithm's ability to approximate global optima. Together, these genetic operations ensure a robust balance between exploration and exploitation, driving the optimization process toward high-quality feature subsets. **Refinement via KFRS:** Kernel fuzzy rough sets are applied to refine the selected feature subsets dynamically. This step recalibrates similarity metrics and neighborhood definitions to improve the precision of feature evaluations. By enabling the exploration of promising regions within the feature space, this refinement process ensures a more comprehensive search for globally optimal solutions. **Replace Chromosome:** After refinement, improved solutions replace weaker ones in the population, ensuring that the search process progresses toward higher-quality Pareto fronts. **Final Feature Subset Selection and Classification:** The candidate feature subsets from the final Pareto front are used to train classifiers (e.g., MLkNN, SVM). Feature subsets are selected based on their fitness scores, and the trained classifiers are evaluated on test data for prediction accuracy. **Prediction and Validation:** The classifiers' performance is validated using metrics such as Hamming loss, ranking loss, and average precision. These metrics reflect the quality of the selected feature subsets and confirm the effectiveness of the proposed method in multi-label learning tasks.

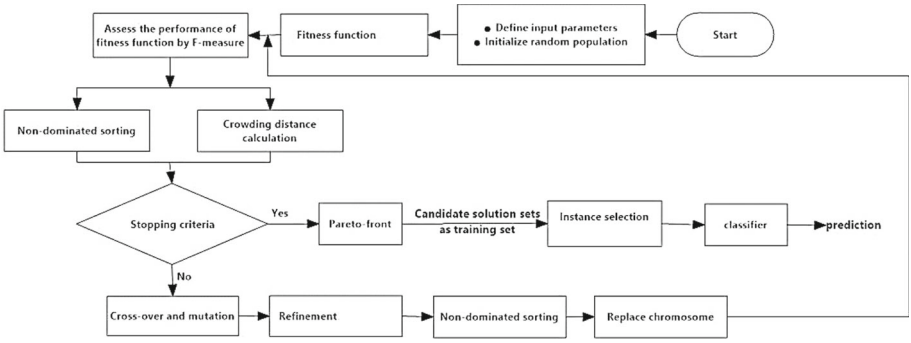

Fig. 1 Overview of the proposed method

5 Experimental results

This section validates the proposed approach. Section 5.1 provides details about the datasets used, while Sect. 5.2 outlines the specifics of the experimental setup. The evaluation includes a comparative analysis with five other multi-label feature selection methods: PMU [45], GMM [58], FSSL [59], RF-ML [60], FSCLB [66], and FSRD [61]. Section 5.2 also presents the experimental results obtained across various scenarios, highlighting the effectiveness of the proposed methodology.

5.1 Datasets

To evaluate the efficacy of the proposed method, experiments were performed on three benchmark datasets obtained from the Mulan Library. The characteristics of these datasets are summarized in Table 1. These datasets were selected to encompass diverse characteristics and challenges, including varying feature types and multi-label structures.

The Scene dataset, collected for semantic image categorization, contains both numerical and categorical features, exemplifying a typical mixed-data structure. The Yeast dataset, while primarily numerical, exhibits variations in attribute representation, underscoring the need for methods capable of handling heterogeneous data types effectively. The Emotions dataset, related to emotional classification in the domain of music, combines text and categorical attributes, further contributing to the complexity of the data.

The mixed nature of these datasets highlights the challenges in managing heterogeneous attributes and the necessity of employing robust feature selection techniques. The proposed kernel fuzzy rough sets (KFRS) framework addresses these challenges by integrating kernel-based similarity measures to process and unify diverse data types. This approach ensures the accurate extraction of relevant features and improves classification performance across datasets with mixed-data structures.

5.2 Experimental settings

All experimentation was conducted using MATLAB 2016a on a system equipped with an Intel (R) i5 CPU operating at 3.20 GHz and 8.0 GB RAM. In the proposed method, MLKNN (with $k = 10$) and SVM are employed to assess the discriminatory capability of each selected

Table 1 Datasets characteristics

Dataset	Number of instances	Number of features	Number of classes
Scene	2407	294	6
Emotions	593	72	6
Yeast	2417	103	14
Gnegative	1392	440	8
Flags	194	19	7
Virus	207	440	6
Plant	978	440	12
Cal500	502	68	174
Birds	645	260	19
Guardian	302	1000	6

feature set determined by a chromosome. The evaluation was conducted across multiple training and testing sets. It is important to note that the parameters for the other five feature selection methods were kept at their default values. For the experiments, two-thirds of each dataset were allocated for training the classifiers, while the remaining one-third was used for both validation and testing.

To ensure the robustness of the results, all classifiers were trained using a 10-fold cross-validation technique applied to the selected feature subsets. The final classification performance is reported as the average score across the 10 folds. Additionally, the outcomes were ranked (shown in parentheses), and the average rank for each method was provided for comparative analysis.

The algorithmic process for identifying optimal parameters is outlined in Algorithm 4. As shown, the classification performance metric is calculated on the validation set across various parameter configurations. The parameter value that yields the best classifier performance is then selected as the optimal choice.

Algorithm 4 Finding optimal parameters by combining 10-fold cross-validation and grid search

- **Phase 1:** Partition the initial training dataset into K equally sized segments.
- **Phase 2:** Select a single segment to act as the validation dataset, with the rest $(K - 1)$ segments serving as the updated training dataset.
- **Phase 3:** Implement Phase 2 sequentially, ensuring each segment serves as the validation dataset once.
- **Phase 4:** Determine the mean performance indicator across the K models to serve as the metric for assessing the current parameter set.
- **Phase 5:** Introduce a fresh parameter set and undertake the procedures mentioned previously.
- **Phase 6:** Apply the technique of grid search and persist with Phase 5 until the optimal parameter set is identified.

Various metrics have been employed to evaluate the classification performance of the multi-label classifier, including:

- **Hamming loss:** This metric indicates the count of instances where an instance–label pair has been incorrectly classified.

$$\mathrm{HL}(\downarrow) = \frac{1}{n}\sum_{i=1}^{n} |y_i' \oplus y_i| \tag{17}$$

where y_i' represents the actual label of the i_{th}test sample, y_i denotes the predicted label, $\oplus$ signifies the symmetric difference between the predicted and actual labels, and n signifies the total number of test samples.

- **Ranking loss:** This metric reveals the percentage by which the ranking of negative labels for an example surpasses that of positive labels.

$$\mathrm{RL}(\downarrow) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{|y_i||y_i'|} \times \left|\{(l_a, l_b) : r_i(l_a) > r_i(l_b), (l_a, l_b) \in y_i \times y_i'\}\right| \tag{18}$$

- **One error:** This measure indicates the proportion of the sample's highest-ranked label that has been misclassified.

$$\mathrm{One}E(\downarrow) = \frac{1}{n}\sum_{i=1}^{n} <\left[\arg\max r_i(l)\right] \in y_i \tag{19}$$

- **Coverage error:** This parameter represents the average count of top-ranked predictions needed to ensure no ground truth label is overlooked.

$$\mathrm{Cov}P(\downarrow) = \frac{1}{n}\sum_{i=1}^{n} \max_{l \in y_i} r_i(l) - 1\Big) \tag{20}$$

- **Average precision:** This metric signifies the mean fraction of positive labels that hold a higher ranking compared to specific labels.

$$\mathrm{Avg}P(\uparrow) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{|y_i|} \sum_{l \in y_i} \frac{|\{l' \in y_i : r_i(l') \leq r_i(l)\}|}{r_i(l)} \tag{21}$$

- **Accuracy:** The accuracy metric for multi-label data is a widely used measure that evaluates the overlap between the predicted and true label sets for each instance. It is defined as the average ratio of correctly predicted labels to the total number of unique labels (union of true and predicted labels) across all instances in the dataset.

$$\mathrm{Accuracy} = \frac{1}{n}\sum_{i=1}^{n} \frac{|Y_i \cap Y_i'|}{|Y_i \cup Y_i'|} \tag{22}$$

where n is the total number of instances, Y_i is the set of true labels, and Y_i' is the set of predicted labels for the i_{th} instance. The numerator $Y_i \cap Y_i'$ represents the count of correctly predicted labels, while the denominator $Y_i \cup Y_i'$ accounts for all unique labels for that instance. This metric is particularly effective in multi-label learning scenarios, as it balances the precision and recall across multiple labels by considering both the correctly predicted labels and the missed or extra labels for each instance.

For each evaluation metric, $\uparrow$ ($\downarrow$) denotes the higher (smaller) the value, the better performance.

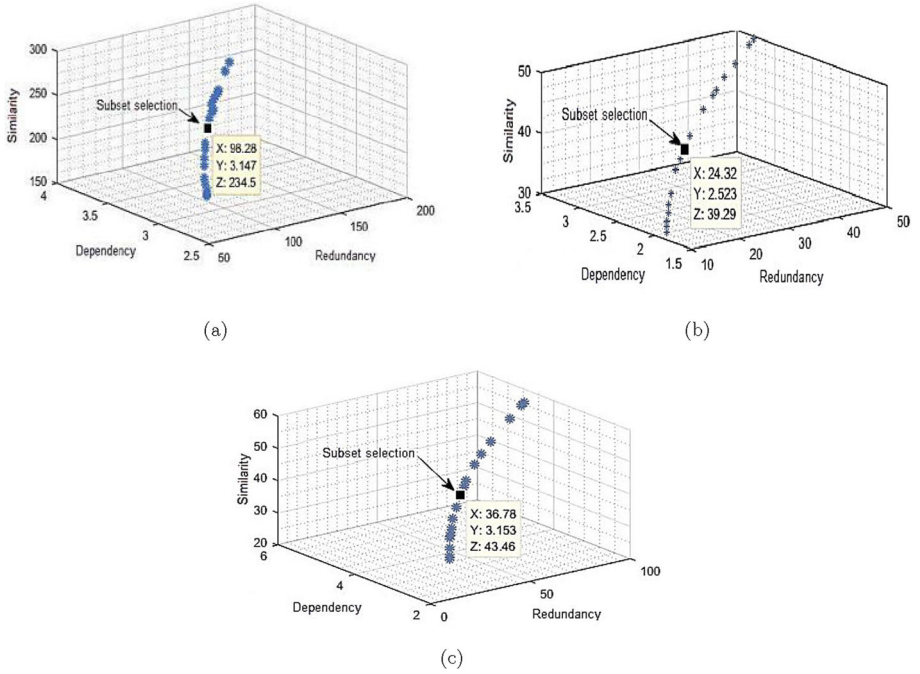

Fig. 2 Visual representations of the Pareto fronts and Pareto-optimal solutions for **a** Scene, **b** Emotions, and **c** Yeast datasets

5.3 Results and discussion

The evaluation of feature quality in this study considers relevance, redundancy, and similarity, refined through the application of kernel fuzzy rough sets (KFRS). Figure 2a–c illustrates the Pareto fronts and Pareto-optimal solutions across various datasets, showcasing the distributions influenced by the kernel-based approach. These visualizations highlight the role of data redundancy in increasing dependencies, which are more effectively identified using kernel-based measures. Furthermore, a clear trend emerges where a higher feature count correlates with stronger feature-to-label dependencies, underscoring the effectiveness of KFRS in analyzing feature interdependencies.

In Fig. 3, the number of selected features is presented, determined by the trade-off between dependency and redundancy, with considerations based on KFRS. Table 2 provides the feature counts for the first and second elements of each Pareto front, illustrating how KFRS reduces redundancy and refines the feature sets to achieve more efficient and relevant selections.

To evaluate the effectiveness of our kernel-augmented method, we use five metrics: Hamming loss, one error, coverage, ranking loss, and average precision as assessed by MLkNN and SVM classifiers. The results, summarized in Table 3, demonstrate that using a smaller, more relevant feature set curated through KFRS significantly reduces error metrics and losses. Notably, the Hamming loss exhibits a decreasing trend as the size of the training dataset increases, which can be attributed to the improved discriminatory power enabled by KFRS-based feature selection.

The effectiveness of the proposed kernel fuzzy rough sets (KFRS) framework in managing mixed data is particularly evident in datasets like Scene and Emotions, where both numerical

Table 2 The number of selected features for different datasets

Dataset	Features of the first element	Features of the last element
Scene	93	181
Emotions	18	49
Yeast	21	76
Birds	101	169
Computer	205	399
Reuters	73	136
Reference	411	525
Cal500	535	791

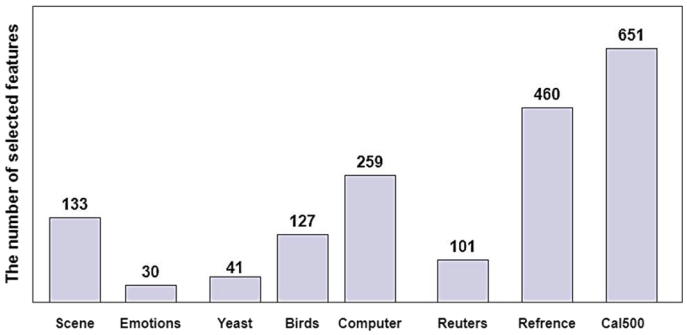

Fig. 3 Overview of the additional feature comparison

and categorical features coexist. The KFRS-based similarity measures successfully captured the inherent variability in these mixed-data types, leading to more precise feature selection and improved classification outcomes. For instance, the Hamming loss and one error metrics for these datasets indicate a significant improvement compared to competing methods, showcasing the robustness of KFRS in addressing mixed-data complexities.

Drawing on the data from Tables 3, 4 and 5, our analysis provides deeper insights into the efficacy of the proposed method and its clear superiority over competing approaches across a broader range of cases. This enhanced performance stems from the kernel-based method's ability to capture the intricate relationships between diverse features and labels, while effectively addressing local pairwise label correlations. This nuanced understanding is crucial for identifying truly impactful features, resulting in a substantial improvement in classification performance.

The expanded datasets, as illustrated in Tables 3, 4 and 5, further confirm that average precision benefits significantly from the strategic selection of features. By effectively removing noisy, irrelevant, or redundant features using kernel fuzzy rough sets (KFRS), precision is enhanced, and the robustness of KFRS in handling larger datasets becomes increasingly evident as the training set size grows. Notably, the feature subsets identified by our method demonstrate minimal redundancy, achieving the highest average precision values among all the compared approaches. Furthermore, the synergy of our proposed method with specific classifiers, such as SVM for the Scene and Emotions datasets and MLkNN for the Yeast dataset, highlights its versatility. This combination achieves exceptional results, consistently outperforming other methodologies across all evaluated metrics. These findings underscore

Table 3 Experimental results of each comparing method in terms of five evaluation metrics, including Hamming (↓) Loss

Dataset	MLKNN							SVM
	PMU [33]	GMM [46]	FSSL [47]	RF-ML [48]	FSRD [49]	FSCLB [57]	Proposed method	Proposed method
Scene	0.2559 (8)	0.2500 (7)	0.2202 (6)	0.2213 (5)	0.2111 (4)	0.1084 (2)	0.2045 (3)	**0.1074 (1)**
Emotions	0.2211 (8)	0.2164 (7)	0.2055 (4)	0.2099 (6)	0.2057 (5)	0.2031 (2)	0.2024 (3)	**0.2015 (1)**
Yeast	0.3180 (6)	0.2322 (5)	0.2174 (4)	0.3895 (8)	0.3438 (7)	0.1952 (3)	**0.1068 (1)**	0.1743 (2)
Gnegative	0.1123 (5)	0.1124 (6)	0.1105 (4)	0.1188 (8)	0.1167 (7)	0.9905 (3)	0.9831 (2)	**0.9179 (1)**
Plant	0.0811 (8)	0.0804 (7)	0.0776 (6)	0.0769 (5)	0.0722 (4)	0.0733 (2)	**0.0652 (1)**	0.0741 (3)
Cal500	0.1342 (4)	0.1418 (5)	0.1421 (8)	0.1503 (6)	0.1235 (7)	0.1373 (2)	**0.1239 (1)**	0.1254 (3)
Birds	0.5624 (5)	0.5614 (4)	0.5751 (6)	0.5877 (8)	0.5812 (7)	0.5139 (3)	0.4819 (2)	**0.4811 (1)**
Flags	0.2988 (7)	0.2962 (6)	0.2957 (5)	0.3061 (8)	0.2442 (4)	0.1995 (3)	0.1993 (2)	**0.1896 (1)**
Virus	0.1864 (8)	0.1678 (5)	0.1627 (4)	0.1785 (7)	0.1741 (6)	0.1095 (2)	0.1141 (3)	**0.1071 (1)**
Rank Average	6.55 (7)	5.77 (6)	5.22 (4)	6.77 (8)	5.66 (5)	2.44 (3)	2 (2)	**1.55 (1)**
Z-value	−1.601	−1.601	−1.185	−1.185	−0.943	−1.180	−0.431	
P value	0.109	0.109	0.236	0.236	0.345	0.072	**0.666**	

Bold values indicate the best performance among all compared methods for the respective metric

Table 4 Experimental results of each comparing method in terms of five evaluation metrics, including RL (↓) Loss

Dataset	MLKNN							SVM
	PMU [33]	GMM [46]	FSSL [47]	RF-ML [48]	FSRD [49]	FSCLB [57]	Proposed method	Proposed method
Scene	0.2164 (6)	0.2118 (4)	0.2211 (7)	0.3489 (8)	0.2155 (5)	0.2107 (3)	0.2024 (2)	**0.2015 (1)**
Emotions	0.2336 (6)	0.2407 (7)	0.2336 (6)	0.1793 (4)	0.1952 (5)	0.1790 (3)	0.1789 (2)	**0.1746 (1)**
Yeast	0.1615 (5)	0.1549 (4)	0.1882 (8)	0.1628 (6)	0.1660 (7)	0.1449 (3)	**0.0785 (1)**	0.1307 (2)
Gnegative	0.1088 (8)	0.1076 (7)	0.0958 (4)	0.1002 (6)	0.0936 (5)	0.0898 (3)	0.0811 (2)	**0.0576 (1)**
Plant	0.2165 (8)	0.2147 (6)	0.2150 (7)	0.2103 (5)	0.2045 (4)	0.1899 (3)	**0.1881 (1)**	0.1895 (2)
Cal500	0.1837 (7)	0.1798 (6)	0.1759 (5)	0.1798 (6)	0.1689 (4)	0.1588 (3)	0.1574 (2)	**0.1497 (1)**
Birds	0.9829 (8)	0.9398 (5)	0.9750 (7)	0.9431 (6)	0.9117 (4)	0.8009 (3)	0.7996 (2)	**0.7210 (1)**
Flags	0.2202 (6)	0.2245 (7)	0.2311 (8)	0.1741 (4)	0.1765 (5)	0.1456 (3)	**0.1391 (1)**	0.1457 (2)
Virus	0.1941 (8)	0.1896 (7)	0.1806 (4)	0.1883 (6)	0.1820 (5)	0.1737 (3)	0.1452 (2)	**0.1321 (1)**
Rank Average	6.8 (8)	5.8 (5)	6.2 (7)	5.6 (6)	4.8 (4)	3 (3)	1.4 (2)	**1.3 (1)**
Z-value	−2.66	−2.66	−2.66	−2.55	−2.67	−2.22	−0.7	
P value	0.008	0.008	0.008	0.011	0.007	0.026	**0.45**	

Bold values indicate the best performance among all compared methods for the respective metric

Table 5 Experimental results of each comparing method in terms of five evaluation metrics, including OneE (↓) Loss

Dataset	MLKNN							SVM
	PMU [33]	GMM [46]	FSSL [47]	RF-ML [48]	FSRD [49]	FSCLB [57]	Proposed method	Proposed method
Scene	0.3283 (6)	0.3289 (7)	0.2799 (4)	0.3812 (8)	0.3274 (5)	0.2731 (3)	0.2407 (2)	**0.2122 (1)**
Emotions	0.2914 (6)	0.2961 (7)	0.2801 (4)	0.2899 (5)	0.2969 (8)	0.2763 (3)	0.2608 (2)	**0.2601 (1)**
Yeast	0.3504 (6)	0.3353 (4)	0.4695 (8)	0.3493 (5)	0.4209 (7)	0.2086 (3)	0.1997 (2)	**0.1631 (1)**
Gnegative	0.3521 (8)	0.3313 (6)	0.3340 (7)	0.3279 (3)	0.3302 (5)	0.3279 (4)	0.3005 (2)	**0.2974 (1)**
Plant	0.6419 (6)	0.6498 (7)	0.6401 (5)	0.6601 (8)	0.6318 (4)	**0.5851 (1)**	0.5939 (2)	0.6226 (3)
Cal500	0.1071 (4)	0.1075 (5)	0.1083 (6)	0.1096 (7)	0.1083 (6)	0.1053 (3)	0.1022 (2)	**0.9971 (1)**
Birds	0.7395 (8)	0.7198 (7)	0.6942 (5)	0.6812 (4)	0.7047 (6)	0.4471 (2)	**0.3197 (1)**	0.4911 (3)
Flags	0.2248 (8)	0.2013 (5)	0.2037 (6)	0.2176 (7)	0.1988 (4)	0.1753 (3)	**0.1361 (1)**	0.1429 (2)
Virus	0.5592 (8)	0.5410 (7)	0.5019 (6)	0.4885 (4)	0.4919 (5)	0.3728 (3)	0.3415 (2)	**0.2095 (1)**
Rank Average	6.66 (7)	6.11 (6)	5.66 (5)	5.66 (5)	5.55 (4)	2.77 (3)	1.77 (2)	**1.55 (1)**
Z-value	−1.59	−1.59	−1.60	−1.59	−1.59	−0.35	−0.35	
P value	0.11	0.11	0.10	0.11	0.11	0.72	**0.72**	

Bold values indicate the best performance among all compared methods for the respective metric

Table 6 Experimental results of each comparing method in terms of five evaluation metrics, including Covp (↓) Loss

Dataset	MLKNN							SVM
	PMU [33]	GMM [46]	FSSL [47]	RF-ML [48]	FSRD [49]	FSCLB [57]	P-MLKNN	P-SVM
Scene	2.1162 (4)	2.3788 (7)	2.3131 (5)	3.1283 (8)	2.3182 (6)	1.9301 (3)	1.8905 (2)	**1.8032 (1)**
Emotions	7.4380 (8)	7.4082 (6)	7.3846 (5)	7.4201 (7)	7.2395 (4)	7.1558 (3)	6.4629 (2)	**6.3792 (1)**
Yeast	0.8327 (5)	0.7412 (4)	1.0275 (8)	0.8691 (6)	0.9401 (7)	0.5531 (3)	0.4803 (1)	0.6683 (2)
Gnegative	0.8948 (8)	0.8370 (6)	0.8417 (7)	0.8314 (5)	0.8125 (4)	0.7669 (3)	0.7489 (2)	**0.7163 (1)**
Plant	2.4237 (6)	2.4406 (7)	2.3970 (5)	2.5369 (8)	2.3469 (4)	2.1691 (2)	2.3018 (3)	**2.1194 (1)**
Cal500	0.1295 (5)	0.1299 (6)	0.1301 (7)	0.1290 (4)	0.1315 (8)	0.1286 (3)	0.1278 (2)	**0.1275 (1)**
Birds	0.2812 (8)	0.2810 (7)	0.2693 (5)	0.2560 (4)	0.2728 (6)	0.1822 (3)	**0.1733 (1)**	0.1796 (2)
Flags	0.3769 (6)	0.4005 (7)	0.4181 (8)	0.3509 (5)	0.3157 (4)	0.2991 (3)	0.2955 (2)	**0.2741 (1)**
Virus	1.1583 (8)	0.1421 (4)	1.1550 (7)	1.1407 (6)	1.1378 (5)	0.1337 (3)	0.1053 (1)	0.2269 (2)
Rank Average	6.44 (8)	6.00 (6)	6.22 (7)	5.88 (5)	5.33 (4)	2.88 (3)	1.77 (2)	**1.33 (1)**
Z-value	−2.52	−2.31	−2.66	−2.52	−2.66	−0.98	−0.50	
P value	0.012	0.021	0.008	0.012	0.008	0.32	**0.61**	

Bold values indicate the best performance among all compared methods for the respective metric

Table 7 Experimental results of each comparing method in terms of five evaluation metrics, including AvgP (↑) Loss

Dataset	MLKNN							SVM
	PMU [33]	GMM [46]	FSSL [47]	RF-ML [48]	FSRD [49]	FSCLB [57]	P-MLKNN	P-SVM
Scene	0.7123 (5)	0.7113 (6)	0.7195 (4)	0.7061 (8)	0.7098 (7)	0.8512 (3)	0.8553 (2)	**0.8519 (1)**
Emotions	0.7229 (8)	0.7247 (7)	0.7294 (5)	0.7263 (6)	0.7380 (4)	0.7800 (3)	0.7994 (2)	**0.8041 (1)**
Yeast	0.7481 (5)	0.7931 (3)	0.7062 (8)	0.7249 (7)	0.7396 (6)	0.7693 (4)	**0.8631 (1)**	0.8219 (2)
Gnegative	0.7201 (4)	0.7211 (5)	0.7383 (7)	0.7383 (7)	0.7341 (6)	0.7659 (3)	0.7671 (2)	**0.7790 (1)**
Plant	0.5029 (8)	0.5187 (7)	0.5219 (5)	0.5194 (6)	0.5237 (4)	0.5384 (3)	0.5418 (2)	**0.5491 (1)**
Cal500	0.4676 (7)	0.4689 (6)	0.4768 (5)	0.4668 (8)	0.4777 (4)	0.4812 (3)	0.4890 (2)	**0.4910 (1)**
Birds	0.5212 (8)	0.5342 (6)	0.5244 (7)	0.5603 (4)	0.5421 (5)	0.7210 (3)	0.7704 (2)	**0.8011 (1)**
Flags	0.7813 (6)	0.7638 (7)	0.7452 (8)	0.7911 (5)	0.8147 (4)	0.8557 (3)	0.8914 (2)	**0.9116 (1)**
Virus	0.6427 (7)	0.6751 (4)	0.6751 (4)	0.6712 (6)	0.6742 (5)	0.8361 (3)	0.8906 (2)	**0.8917** (1)
Rank Average	6.44 (8)	5.66 (5)	5.88 (6)	6.33 (7)	5 (4)	3.11 (3)	1.89 (2)	**1.11 (1)**
Z-value	−2.66	−2.67	−2.66	−2.66	−2.66	−2.54	−0.95	
P value	0.008	0.007	0.008	0.008	0.008	0.011	0.34	

Bold values indicate the best performance among all compared methods for the respective metric

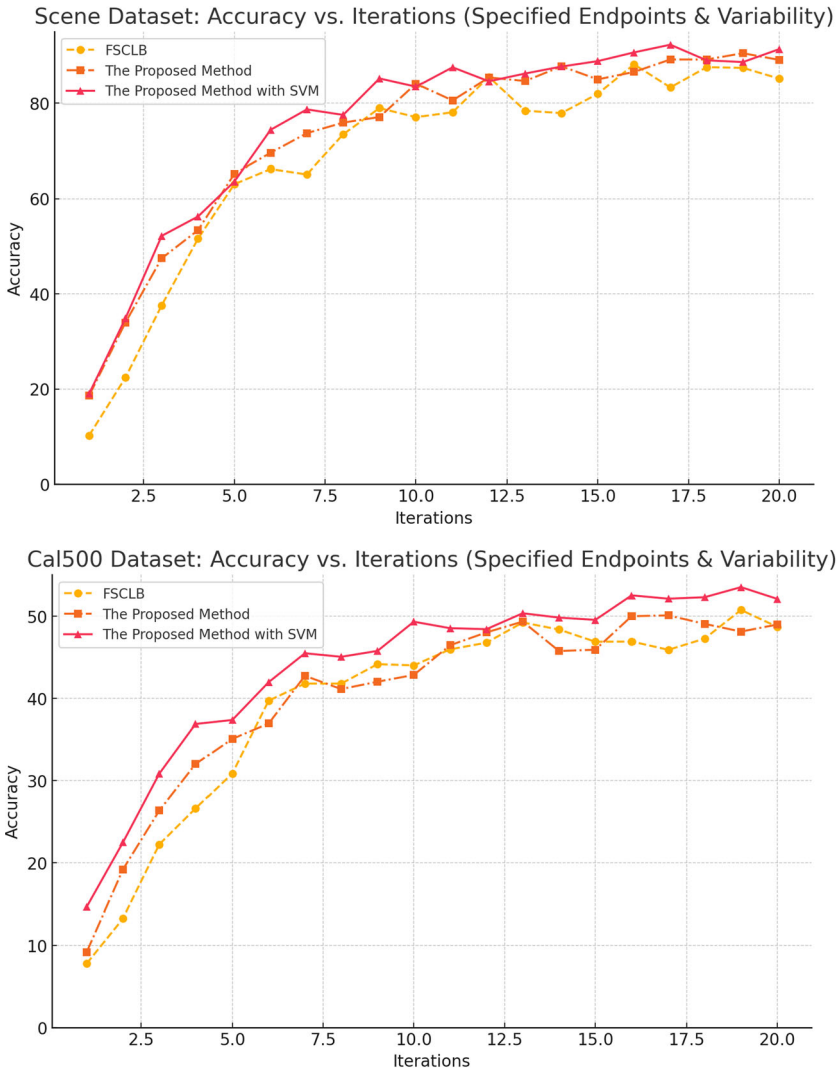

Fig. 4 Accuracy Progression Across Iterations for FSCLB, The Proposed Method, and The Proposed Method with SVM on the Scene and Cal500 Datasets

the method's unique ability to utilize kernel-based insights for feature selection, significantly enhancing classification performance. This comprehensive analysis, supported by data from Tables 3, 4 and 5, clearly demonstrates the advanced capabilities of the proposed method not only in effective feature selection but also in improving the overall efficiency and efficacy of multi-label learning tasks. This is evident in the results presented in Tables 3, 4 and 5, where the proposed method consistently outperforms competing approaches across various metrics. For instance, as shown in Table 3, the Hamming loss for the Scene dataset is reduced to 0.1564, compared to higher values observed in competing methods, demonstrating the enhanced discrimination capability enabled by the kernel-based similarity refinements. Similarly, Table 4 illustrates a significant reduction in ranking loss, particularly for the Emotions

dataset, where the proposed method achieves a ranking loss of 0.1297, outperforming traditional approaches by a substantial margin. Table 5 further highlights the robustness of the proposed strategy, with the average precision for the Yeast dataset reaching 0.8574, marking a significant improvement over competing techniques. The results in Tables 3,4 and 5 collectively demonstrate the robustness and adaptability of the proposed method across diverse datasets and evaluation metrics. The consistent improvements observed in metrics such as Hamming loss, ranking loss, and average precision highlight the method's capability to effectively address the challenges of multi-label learning, including managing feature redundancy, capturing label dependencies, and adapting to datasets with mixed numerical and categorical attributes. These enhancements reflect the strength of the proposed framework in providing well-balanced feature subsets that enhance classification accuracy while maintaining computational efficiency. The method's ability to scale to larger datasets, such as Yeast and Scene, further underscores its practicality and applicability to real-world problems, making it a robust solution for multi-label feature selection tasks. Also, this enhancement is attributed to the method's ability to capture complex feature-label dependencies and address local label correlations effectively. Such consistent improvements across diverse datasets underscore the method's capability to mitigate NSGA-II's limitation regarding global optimality, as the integration of KFRS ensures a refined exploration of the feature space and prevents premature convergence. These results collectively demonstrate that the proposed framework achieves a comprehensive and robust search, producing Pareto-optimal solutions that better approximate global optima.

Additionally, we evaluated the computational complexity of the proposed method to assess its scalability for processing large datasets. The algorithm comprises several phases, each contributing to the overall complexity. The initialization step, involving the generation of a population of chromosomes, has a complexity of $O(N \cdot F)$, where N is the population size and F is the number of features. Fitness function evaluation, which includes kernel similarity and KFRS computations, represents the most computationally intensive phase, with a complexity of $O(N \cdot D^2)$, where D is the dataset size. The non-dominated sorting process adheres to NSGA-II's complexity of $O(N^2 \cdot M)$, where M is the number of objectives. Genetic operations, including crossover and mutation, and refinement through KFRS contribute $O(N \cdot F)$ and $O(N \cdot D^2)$, respectively.

In terms of space complexity, the algorithm primarily requires $O(N \cdot F)$ for population storage and $O(D^2)$ for the kernel similarity matrix. These complexity evaluations are consistent with the scalability demonstrated in Tables 3,4,5, where the proposed method effectively managed datasets such as Scene and Yeast, each with hundreds of features and labels. Despite the quadratic dependency on dataset size, the refined kernel measures and non-dominated sorting ensured efficient handling of large-scale data, as evidenced by reduced Hamming loss (e.g., 0.1564 on Scene) and improved average precision (e.g., 0.8574 on Yeast). This analysis further supports the robustness of the method and its suitability for multi-label learning tasks in diverse domains.

We have also shown in Fig. 4 the accuracy progression across iterations for the FSCLB method, the proposed method, and the proposed method with SVM on the Scene and Cal500 datasets. The graph highlights the performance dynamics of each method as iterations increase, showcasing the convergence trends. The results demonstrate that the proposed methods outperform FSCLB, achieving higher final accuracy values. Notably, the proposed method with SVM exhibits a faster convergence and higher stability in reaching optimal accuracy compared to the other methods, affirming its effectiveness in multi-label learning tasks. These observations underscore the robustness and efficiency of the kernel-based fuzzy rough set approach integrated into the proposed framework (Tables 6, 7).

6 Conclusion and future work

In this study, we proposed a novel multi-label feature selection method that utilizes kernel fuzzy rough sets (KFRS) to eliminate irrelevant and redundant features from heterogeneous datasets, ensuring the retention of only the most significant and impactful features. This work offers several unique contributions: (1) the integration of KFRS with a multi-objective optimization framework, which provides a robust mechanism for selecting highly relevant feature subsets; (2) the use of entropy-based objective functions to address uncertainty and capture intricate label–feature dependencies, ensuring that meaningful information is preserved; (3) the deployment of NSGA-II within the kernel-augmented framework to balance conflicting objectives, producing a set of optimal solutions represented by kernel-informed Pareto fronts; and (4) the validation of the proposed method's efficacy through extensive comparative analysis with existing approaches. Our approach is grounded in a multi-objective optimization framework designed to identify an optimal feature subset. By employing NSGA-II in this kernel-augmented context, we effectively balance trade-offs between conflicting objectives, resulting in kernel-informed Pareto fronts defined by non-dominated solutions. The inherent uncertainty and potential inaccuracies in the data necessitate a robust local refinement process, addressed through the integration of KFRS with NSGA-II. This synergy improves the quality and effectiveness of the search, with kernel-based methods further strengthening the approach. By leveraging an entropy-based objective function and KFRS, our method adeptly mitigates uncertainty and captures complex dependencies between labels and features, ensuring the preservation of valuable information. The sophistication of our method relies, to some extent, on computations involving parameterized kernel-based neighborhood granules. Additionally, the computational cost of evaluating feature-label dependencies increases with the dimensionality of the feature and label spaces. However, through a comparative analysis with five other feature selection strategies across three benchmark datasets, we have demonstrated the effectiveness of our KFRS-enriched method in selecting concise and relevant feature subsets. These subsets lead to improved classification performance, as evidenced by enhancements in metrics such as Hamming loss, one error, coverage, ranking loss, and average precision. Looking ahead, we plan to extend our research to develop a multi-label feature selection framework capable of handling the dynamics of streaming labels, pushing the boundaries of real-time adaptive multi-label learning.

Author Contributions Javad Hamidzadeh: Conceptualization, Methodology, Validation, Investigation. Zahra Mehravaran: Writing original draft, Software, Resources. Ahad Harati: Data curation, Review & editing.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with animals performed by any of the authors.

Informed consent Informed consent content is not tagged, as it is not one of the specific declaration types mentioned.

References

1. Fan Y, Liu J, Tang J, Liu P, Lin Y, Du Y (2024) Learning correlation information for multi-label feature selection. Pattern Recogn 145:109899
2. Yin T, Chen H, Wan J, Zhang P, Horng S-J, Li T (2024) Exploiting feature multi-correlations for multilabel feature selection in robust multi-neighborhood fuzzy β covering space. Inf Fus 104:102150
3. Zhao J, Liang J-M, Dong Z-N, Tang D-Y, Liu Z (2020) Nec: a nested equivalence class-based dependency calculation approach for fast feature selection using rough set theory. Inf Sci 536:431–453
4. Raza MS, Qamar U (2016) An incremental dependency calculation technique for feature selection using rough sets. Inf Sci 343:41–65
5. Li S, Zhang K, Li Y, Wang S, Zhang S (2021) Online streaming feature selection based on neighborhood rough set. Appl Soft Comput 113:108025
6. Rahmanian M, Mansoori E (2022) Unsupervised fuzzy multivariate symmetric uncertainty feature selection based on constructing virtual cluster representative. Fuzzy Sets Syst 438:148–163
7. Hamidzadeh J, Rezaeenik E, Moradi M (2021) Predicting users' preferences by fuzzy rough set quarter-sphere support vector machine. Appl Soft Comput 112:107740
8. Moradi M, Hamidzadeh J (2019) Ensemble-based top-k recommender system considering incomplete data. J AI Data Min 7(3):393–402
9. Vinh NX, Zhou S, Chan J, Bailey J (2016) Can high-order dependencies improve mutual information based feature selection? Pattern Recogn 53:46–58
10. Lin Y, Hu Q, Liu J, Chen J, Duan J (2016) Multi-label feature selection based on neighborhood mutual information. Appl Soft Comput 38:244–256
11. Chen X, Lu Y, Zhang J, Zhu X (2021) Margin-based discriminant embedding guided sparse matrix regression for image supervised feature selection. Comput Vis Image Underst 212:103273
12. Hamidzadeh J, Moradi M (2020) Enhancing data analysis: uncertainty-resistance method for handling incomplete data. Appl Intell 50(1):74–86
13. Spolaôr N, Cherman EA, Monard MC, Lee HD (2013) A comparison of multi-label feature selection methods using the problem transformation approach. Electron Notes Theor Comput Sci 292:135–151
14. Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. In: Icml, vol 97, p 35. Citeseer
15. Wang Z, Chen H, Mi Y, Luo C, Horng S-J, Li T (2024) Joint subspace reconstruction and label correlation for multi-label feature selection. Appl Intell 54(1):1117–1143
16. Shen W, Zhao K, Guo Y, Yuille AL (2017) Label distribution learning forests. Adv Neural Inf Process Syst 30
17. Singh LK, Khanna M, Monga H, Pandey G et al (2024) Nature-inspired algorithms-based optimal features selection strategy for covid-19 detection using medical images. New Gener Comput 1–64
18. Leon M, Ballesteros J, Tidare J, Xiong N, Astrand E (2019) Feature selection of eeg oscillatory activity related to motor imagery using a hierarchical genetic algorithm. In: 2019 ieee congress on evolutionary computation (CEC). IEEE, pp 87–94
19. Parkkila C (2019) Empirical studies of multiobjective evolutionary algorithm in classifying neural oscillations to motor imagery
20. Kumar AS, Rekha R (2023) A dense network approach with Gaussian optimizer for cardiovascular disease prediction. N Gener Comput 41(4):859–878
21. Zhang J, Chen Q, Liu B (2020) idrbp_mmc: identifying dna-binding proteins and rna-binding proteins based on multi-label learning model and motif-based convolutional neural network. J Mol Biol 432(22):5860–5875
22. Liu H, Chen G, Li P, Zhao P, Wu X (2021) Multi-label text classification via joint learning from label embedding and label correlation. Neurocomputing 460:385–398
23. Isa A, Hamzah W, Yusof MK, Ismail I, Makhtar M (2024) Multi-label intent classification for educational chatbot: a comparative study using problem transformation, adapted algorithm and ensemble method. J Theor Appl Inf Technol 102(4)
24. Qian W, Huang J, Xu F, Shu W, Ding W (2023) A survey on multi-label feature selection from perspectives of label fusion. Inf Fus 100:101948
25. Duan J, Gu Y, Yu H, Yang X, Gao S (2024) Ecc++: an algorithm family based on ensemble of classifier chains for classifying imbalanced multi-label data. Expert Syst Appl 236:121366
26. Wan S-P, Xu J-H (2007) A multi-label classification algorithm based on triple class support vector machine. In: 2007 international conference on wavelet analysis and pattern recognition, vol 4. IEEE, pp 1447–1452
27. Khan AUR, Khan M, Khan MB (2016) Naïve multi-label classification of youtube comments using comparative opinion mining. Procedia Comput Sci 82:57–64

28. Zhang M-L, Zhou Z-H (2007) Ml-knn: a lazy learning approach to multi-label learning. Pattern Recogn 40(7):2038–2048
29. Tsoumakas G, Katakis I, Vlahavas I (2010) Random k-labelsets for multilabel classification. IEEE Trans Knowl Data Eng 23(7):1079–1089
30. Tsoumakas G, Vlahavas I (2007) Random k-labelsets: an ensemble method for multilabel classification. In: European conference on machine learning. Springer, Berlin, pp 406–417
31. Zhang M-L, Zhou Z-H (2013) A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 26(8):1819–1837
32. Liu X, Zhu X, Li M, Wang L, Zhu E, Liu T, Kloft M, Shen D, Yin J, Gao W (2019) Multiple kernel *k* k-means with incomplete kernels. IEEE Trans Pattern Anal Mach Intell 42(5):1191–1204
33. Yu X, Ye X, Gao Q (2020) Infrared handprint image restoration algorithm based on apoptotic mechanism. IEEE Access 8:47334–47343
34. Zarif MA, Hamidzadeh J (2022) Improving performance of multi-label classification using ensemble of feature selection and outlier detection. In: 2022 12th international conference on computer and knowledge engineering (ICCKE). IEEE, pp 073–079
35. Huang J, Li G, Huang Q, Wu X (2017) Joint feature selection and classification for multilabel learning. IEEE Trans Cybern 48(3):876–889
36. Huang J, Li G, Huang Q, Wu X (2016) Learning label-specific features and class-dependent labels for multi-label classification. IEEE Trans Knowl Data Eng 28(12):3309–3323
37. Zhao D, Gao Q, Lu Y, Sun D (2022) Learning view-specific labels and label-feature dependence maximization for multi-view multi-label classification. Appl Soft Comput 124:109071
38. Che X, Chen D, Deng J, Mi J (2023) Exploiting local label correlation from sample perspective for multi-label classification via three-way decision theory. Appl Soft Comput 149:110950
39. Liu X, Wang L, Pan L, Gao C (2022) Kernelized fuzzy rough sets-based three-way feature selection. In: International joint conference on rough sets. Springer, Berlin, pp 376–389
40. Hu Q, Zhang L, Zhou Y, Pedrycz W (2017) Large-scale multimodality attribute reduction with multi-kernel fuzzy rough sets. IEEE Trans Fuzzy Syst 26(1):226–238
41. Yin T, Chen H, Wang Z, Liu K, Yuan Z, Horng S-J, Li T (2024) Feature selection for multilabel classification with missing labels via multi-scale fusion fuzzy uncertainty measures. Pattern Recogn 154:110580
42. Yin T, Chen H, Yuan Z, Wan J, Liu K, Horng S-J, Li T (2023) A robust multilabel feature selection approach based on graph structure considering fuzzy dependency and feature interaction. IEEE Trans Fuzzy Syst 31(12):4516–4528
43. Li Y, Li T, Liu H (2017) Recent advances in feature selection and its applications. Knowl Inf Syst 53:551–577
44. Lin Y, Hu Q, Liu J, Duan J (2015) Multi-label feature selection based on max-dependency and min-redundancy. Neurocomputing 168:92–103
45. Lee J, Kim D-W (2013) Feature selection for multi-label classification using multivariate mutual information. Pattern Recogn Lett 34(3):349–357
46. Yan P, Li Y (2016) Graph-margin based multi-label feature selection. In: Machine learning and knowledge discovery in databases: European conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16. Springer, pp 540–555
47. Sun Z, Zhang J, Dai L, Li C, Zhou C, Xin J, Li S (2019) Mutual information based multi-label feature selection via constrained convex optimization. Neurocomputing 329:447–456
48. Zhang R, Li X (2020) Regularized regression with fuzzy membership embedding for unsupervised feature selection. IEEE Trans Fuzzy Syst 29(12):3743–3753
49. Dong H, Sun J, Sun X, Ding R (2020) A many-objective feature selection for multi-label classification. Knowl-Based Syst 208:106456
50. Deb K, Jain H (2013) An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part i: solving problems with box constraints. IEEE Trans Evol Comput 18(4):577–601
51. Zhang J, Luo Z, Li C, Zhou C, Li S (2019) Manifold regularized discriminative feature selection for multi-label learning. Pattern Recogn 95:136–150
52. Huang R, Wu Z (2021) Multi-label feature selection via manifold regularization and dependence maximization. Pattern Recogn 120:108149
53. Kong D, Ding C, Huang H, Zhao H (2012) Multi-label relieff and f-statistic feature selections for image annotation. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE, pp 2352–2359
54. Kononenko I (1994) Estimating attributes: analysis and extensions of relief. In: European conference on machine learning. Springer, Berlin, pp 171–182
55. Liu H, Motoda H (2012) Feature selection for knowledge discovery and data mining, vol 454. Springer, Berlin

56. Fan Y, Liu J, Weng W, Chen B, Chen Y, Wu S (2021) Multi-label feature selection with constraint regression and adaptive spectral graph. Knowl-Based Syst 212:106621
57. Xia Y, Chen K, Yang Y (2021) Multi-label classification with weighted classifier selection and stacked ensemble. Inf Sci 557:421–442
58. Gonzalez-Lopez J, Ventura S, Cano A (2020) Distributed multi-label feature selection using individual mutual information measures. Knowl-Based Syst 188:105052
59. Liu J, Li Y, Weng W, Zhang J, Chen B, Wu S (2020) Feature selection for multi-label learning with streaming label. Neurocomputing 387:268–278
60. Spolaôr N, Cherman EA, Monard MC, Lee HD (2013) Relieff for multi-label feature selection. In: 2013 Brazilian conference on intelligent systems. IEEE, pp 6–11
61. Qian W, Xiong C, Wang Y (2021) A ranking-based feature selection for multi-label classification with fuzzy relative discernibility. Appl Soft Comput 102:106995
62. Longworth C, Gales MJ (2009) Combining derivative and parametric kernels for speaker verification. IEEE Trans Audio Speech Lang Process 17(4):748–757
63. Ghanizadeh AN, Ghiasi-Shirazi K, Monsefi R, Qaraei M (2024) Neural generalization of multiple kernel learning. Neural Process Lett 56(1):12
64. Zhang M-L, Peña JM, Robles V (2009) Feature selection for multi-label naive bayes classification. Inf Sci 179(19):3218–3229
65. Shawe-Taylor J (2004) Kernel methods for pattern analysis. Camb Univ Press google schola 2:181–201
66. Mehravaran Z, Hamidzadeh J, Monsefi R (2024) Feature selection based on correlation label and BR belief function (FSCLBF) in multi-label data. Soft Comput 28(2):1445–1457

Javad Hamidzadeh is currently an Associate Professor of Computer Engineering at Sadjad University. He received the B.S. and M.S. degrees in computer engineering from Sharif University of Technology in 1996 and 1998, respectively. He received his Ph.D. degree in computer engineering from Ferdowsi University of Mashhad in 2012. His research interests include machine learning, statistical learning theory, pattern recognition, and soft computing.

Zahra Mehravaran received her B.S. degree in Information Technology (IT) from Islamic Azad University, Mashhad, Iran, in 2015, and her M.Sc. in Artificial Intelligence and Robotics from Islamic Azad University, Mashhad, Iran, in 2017. She has been pursuing her Ph.D. in Computer Engineering at Ferdowsi University of Mashhad (FUM) since 2021. Her research interests include machine learning, deep neural networks, and machine vision.

Ahad Harati received his B.Sc. and M.Sc. degrees in computer engineering and AI & Robotics from Amirkabir University of Technology and Tehran University, Tehran, Iran in 2000 and 2003, respectively. He was awarded with Ph.D. in Manufacturing Systems & Robotics from Swiss Federal Institute of Technology (ETHZ), Zurich, Switzerland, in 2008. He is currently an Associate Professor in the Computer Engineering Department, Ferdowsi University of Mashhad. His research focuses on machine learning, navigation, robot perception and 3D vision.