Multi-Label Feature Selection via Non-Additive Fusion of Embedded and Spectral Evidence

Expert Systems With Applications 333 (2027) 134236

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Multi-label feature selection via non-additive fusion of embedded and spectral evidence

Filippo Casu [a], Antonio Costantino Marceddu [b], Giuseppe A. Trunfio [a,*]

[a] Department of Engineering, University of Sassari, Sassari, Italy
[b] Department of Control and Computer Engineering (DAUIN), Politecnico di Torino, Turin, Italy

ARTICLE INFO

Keywords:
Multi-label learning
Feature selection
Convex optimisation
Spectral graph diffusion
Score-level fusion
Choquet integral

ABSTRACT

Multi-label feature selection (MLFS) is a critical task for handling high-dimensional datasets where labels exhibit complex interdependencies. While embedded methods effectively identify discriminative features by optimizing predictive performance through structured sparsity, they often overlook the underlying topological structure of the label space. Conversely, spectral methods capture label correlations via similarity structures but lack direct optimization for prediction. To bridge this gap, this paper introduces CFIFS (Choquet Fuzzy Integral Feature Selection), a framework that models MLFS as a non-additive, score-level information fusion problem. The proposed pipeline integrates two complementary perspectives: a group-sparse embedded score for predictive relevance and a spectral score—combining Dirichlet smoothness and the Hilbert-Schmidt Independence Criterion (HSIC)—for topological consistency. Rather than relying on traditional additive aggregation, CFIFS employs a two-source Choquet fuzzy integral to fuse these rankings. The fuzzy capacity is learned through an automated data-driven procedure, allowing the framework to adaptively weigh the synergy or redundancy between input scores. Furthermore, the model provides an interpretable interaction index that quantifies the relationship between the predictive and topological evidence. Extensive experiments on 20 benchmark datasets demonstrate that CFIFS significantly outperforms seven state-of-the-art MLFS baselines. Rigorous statistical validation via Friedman and Holm tests confirms the robustness and superior generalization of the proposed fusion approach across diverse application domains.

1. Introduction

Multi-Label Learning (MLL) addresses classification tasks in which each instance can be associated with several class labels at once, rather than a single one (Tsoumakas & Katakis, 2007; Zhang & Zhou, 2014). Such tasks arise naturally in text categorisation (Read et al., 2008), image annotation, gene function prediction, and music emotion recognition (Tsoumakas et al., 2009), where the joint co-occurrence patterns among labels carry structural information that single-label formulations cannot represent. As the dimensionality of the feature space grows, retaining all measured features becomes increasingly problematic: irrelevant or redundant features inflate the computational cost of downstream classifiers, weaken generalisation through overfitting, and make it harder to identify the variables that are genuinely informative (Guyon & Elisseeff, 2003). Multi-Label Feature Selection (MLFS) addresses this problem by producing a compact feature ranking—or subset—that preserves the information needed for accurate multi-label prediction while simultaneously improving model interpretability, since the practitioner can inspect which features the selection process deems important and why.

The MLFS literature has developed along three broad paradigms (Kashef et al., 2018; Pereira et al., 2018). Filter methods score features through statistical criteria computed independently of any learning algorithm, such as mutual information (Lee & Kim, 2015a,b; Lin et al., 2015) and relevance–redundancy measures (Zhang et al., 2020); they are efficient but may overlook complex feature interactions. Wrapper methods evaluate candidate subsets by training and testing a full multi-label classifier, achieving strong predictive performance at a computational cost that limits scalability (Zhang et al., 2017). Embedded methods occupy a middle ground: they learn a predictive model whose internal structure—for instance an $\ell_{2,1}$-norm group-sparsity penalty (Nie et al., 2010)—directly encodes feature importance, offering both discriminative power and moderate cost. Among recent embedded methods, non-negative matrix factorisation (Gao et al., 2023; Hu et al., 2020) and manifold-regularised regression (Huang & Wu, 2021; Zhang et al., 2019) have further

* Corresponding author.
E-mail addresses: fcasu1@uniss.it (F. Casu), antonio.marceddu@polito.it (A.C. Marceddu), trunfio@uniss.it (G.A. Trunfio).

https://doi.org/10.1016/j.eswa.2026.134236
Received 31 March 2026; Received in revised form 27 June 2026; Accepted 29 August 2026
Available online 1 September 2026

enriched the range of structural priors available. However, these methods typically rely on a fixed parametric mapping from features to labels, which may fail to capture the non-linear topological structure of the data space.

A separate line of research exploits graph-based and spectral properties. Methods in this family build a graph on the instances or labels—using Laplacian regularisation (Huang et al., 2018), manifold constraints (Bai & Wu, 2024), or diffusion processes (Sun et al., 2008)—and favour features that are smooth or informative with respect to that graph structure. The resulting scores capture topological dependencies that an embedded regression model is not designed to represent directly, whereas embedded methods tie feature importance more closely to predictive objectives through an explicit optimisation model. These two perspectives are therefore largely complementary. Recent work has also explored aggregation-oriented and ensemble-based MLFS formulations, including Multi-Criteria Decision Making (MCDM) and ranking aggregation schemes (Hashemi et al., 2020, 2021; Huang et al., 2023; Miri et al., 2022; Mohanrasu et al., 2026). These approaches rely on different combination mechanisms, including MCDM-based aggregation, order-statistics-based ranking, and distance-based rank aggregation (Hashemi et al., 2020, 2021; Miri et al., 2022; Mohanrasu et al., 2026). To the best of our knowledge, however, they do not explicitly model interaction effects between complementary embedded and spectral score channels through a non-additive fusion operator.

This article introduces *Choquet Fuzzy Integral Feature Selection (CFIFS)*, a unified MLFS framework able to address this limitation by treating embedded and spectral rankings as distinct, yet interacting, information channels. Given a training fold, the pipeline computes two complementary score vectors.

The first is a *convex embedded score*, derived from a group-sparse model that couples a PLST-style label embedding (Tai & Lin, 2012) with an embedding-aware logistic term and standard inverse-frequency instance reweighting for label imbalance (Charte et al., 2015; He & Garcia, 2009). The second is a *spectral graph-diffusion score*, which blends Dirichlet smoothness on a Jaccard label-affinity graph (He et al., 2005) with a kernel-based Hilbert–Schmidt Independence Criterion (HSIC) relevance criterion (Gretton et al., 2005), capturing global label-graph dependencies that the embedded model cannot access. These normalised score vectors are then fused via a *two-source Choquet fuzzy integral* (Grabisch, 1996; Grabisch & Labreuche, 2007; Murofushi & Sugeno, 1989). Unlike traditional additive fusion, the Choquet integral employs a fuzzy capacity to model the contribution of each channel and their mutual interaction, allowing the framework to dynamically adapt between conjunctive (agreement-seeking) and disjunctive (complementarity-seeking) behaviour. By learning this capacity through an inner Cross-Validation (CV) loop, the fusion adapts to the specific characteristics of each dataset.

The individual components of CFIFS are well-established; the value of this work lies in how they are combined and calibrated. In summary, the main contributions of this article are the following:

- *A non-additive score-level fusion framework for MLFS.* We formulate multi-label feature selection as a two-source information-fusion problem and combine an embedded (group-sparse, label-embedding-aware) score with a spectral (label-graph Dirichlet + HSIC) score through a two-source Choquet fuzzy integral, going beyond the additive, order-statistics-based, and distance-based aggregation rules used by existing ensemble and MCDM-style selectors.
- *A data-driven, capacity-learning mechanism.* The fuzzy capacity governing the fusion is selected by an inner CV loop that is entirely contained within each training fold and recomputes both scoring stages on every inner-training partition, so the fusion behaviour adapts to each dataset without any per-dataset manual tuning and without inner-validation labels ever entering the fused scores.
- *Interpretability of the fusion behaviour.* The learned capacity yields an interaction index—whose behavioural semantics we establish formally through an exact Shapley decomposition of the two-source Choquet integral—that characterises, for each dataset, whether the two channels are combined conjunctively (agreement-seeking) or disjunctively (redundancy-aware).
- *A comprehensive empirical validation.* Experiments on 20 standard benchmarks under a rigorous 10-fold CV protocol—complemented by ablation, robustness-to-noise, downstream-classifier, and higher-dimensional analyses—show that CFIFS achieves the best average rank on all seven evaluation metrics and is Holm-significant against every baseline on all primary metrics, including both Micro-F1 and Macro-F1.

The remainder of the article is organised as follows. Section 2 reviews related work. Section 3 formalises the proposed method. Section 4 presents the experimental evaluation, and Section 5 concludes the article.

2. Related work

MLFS has evolved along several complementary research streams; here we trace the main lines of development and highlight the gap addressed by our proposal. For broader coverage, we refer the reader to the surveys of Qian et al. (2023), Pereira et al. (2018), and Hancer et al. (2025).

The earliest MLFS methods mainly follow a *filter* strategy, in which features are scored by statistical criteria that are independent of the downstream predictor. Simple Binary Relevance (BR) decompositions evaluate each feature–label pair separately and then aggregate the resulting scores (Spolaor et al., 2016), while more principled approaches rely on Mutual Information (MI) and its multivariate decompositions. Along this line, Lee and Kim introduced D2F (Lee & Kim, 2015b) and SCLS (Lee & Kim, 2017)—the latter specifically designed for large label sets—and related criteria such as PMU (Lee & Kim, 2013) and FIMF (Lee & Kim, 2015a) further refined information-theoretic ranking. Lin et al. (2015) proposed MDMR, which maximises dependency while minimising redundancy, and Zhang et al. (2020) later introduced GRRO, a global relevance-and-redundancy optimisation criterion accounting for both feature–label and feature–feature interactions. Although computationally attractive, filter methods usually evaluate features either in isolation or through limited-order interactions, which constrains their ability to capture the joint discriminative structure revealed by a predictive model.

This limitation naturally motivates *embedded* methods, where feature importance is inferred from the parameters of a structured predictor, typically under group-sparsity constraints. The $\ell_{2,1}$-norm minimisation framework of Nie et al. (2010), which induces row-sparsity in the coefficient matrix, laid much of the foundation for this family. Building on this template, SCMFS (Hu et al., 2020) incorporates shared common-mode label structure, SSFS (Gao et al., 2023) introduces constrained latent structure, and MDFS (Zhang et al., 2019) combines regression with manifold-based discriminative regularisation; MRDM (Huang & Wu, 2021) further couples manifold regularisation with dependence maximisation. More recent developments continue to enrich this sparsity-plus-regularisation paradigm: RFSFS (Li et al., 2023) uses robust flexible sparse regularisation, LRMFS (Fan et al., 2025) introduces label relaxation, and SRFS (Li et al., 2025) integrates sparse supplementation with information fusion. The most recent work in this stream couples robust structure learning with discriminative label correlations (Jia et al., 2024), relaxes the binary label matrix into continuous regression targets (Fan et al., 2025), reconstructs the feature representation itself before selection (Fan et al., 2026), and jointly performs feature reconstruction and label-correlation learning (Zhang et al., 2025a); specialised variants also target imbalanced label distributions (Xu & Li, 2025) and feature noise in partially labelled data (Wu et al., 2025). A related thread explicitly models label correlations within the embedded objective: Li et al. (2022) study label-correlation variation, Fan et al.

(2024b) learn correlation information during selection, and Zhang et al. (2023) adaptively estimate label correlation. Despite their discriminative power, these methods still derive a single ranking from a single optimisation model, and therefore do not explicitly exploit complementary structural evidence available from alternative scoring mechanisms.

A separate line of work seeks precisely such structural evidence through *graph-based* and *spectral* formulations. The Laplacian Score of He et al. (2005) is a foundational unsupervised filter that favours features whose signals are smooth on a nearest-neighbour graph; its multi-label adaptation MCLS (Huang et al., 2018) injects label information through a manifold-based constraint Laplacian. Subsequent work has enriched both graph construction and regularisation: DSMFS (Hu et al., 2022) employs dynamic subspace dual-graph regularisation, Zhang and Ma (2022) combine non-negativity with dynamic graph updates, and LRDG (Zhang et al., 2024) couples latent representation learning with dynamic graph constraints. Other recent contributions include graph-based selection with latent representation learning (Bai & Wu, 2024), SCNMF (He et al., 2024), which imposes similarity constraints through non-negative matrix factorisation on a feature graph, and sparse selection with pseudo-label learning under dynamic graph constraints (Zhang et al., 2025b). At the intersection of graph methods and statistical dependence, the HSIC (Gretton et al., 2005) has also been used as a relevance measure in feature selection since HSIC-Lasso (Yamada et al., 2014), a line of work that remains very active: Wang et al. (2023) provide a unified view of HSIC-based selection criteria, and a recent survey (Wang, 2026) documents the breadth of current HSIC-Lasso variants. However, in the multi-label setting it has mostly appeared as a single-criterion filter rather than as part of a fused discriminative architecture.

These complementary directions have stimulated growing interest in *aggregation-oriented*, *ensemble-based*, and *hybrid* MLFS methods. Beyond sequential hybrids, the literature already includes explicit information-fusion and multi-criteria formulations. For example, MFS-MCDM models MLFS as a multi-criteria decision-making problem and ranks features through a TOPSIS-based information-fusion procedure (Hashemi et al., 2020), while VMFS extends the same perspective through a VIKOR-based ranking strategy for multi-target feature selection (Hashemi et al., 2021). Other works aggregate feature rankings more directly: Miri et al. (2022) propose an ensemble method for multi-label text classification based on intelligent order statistics, and Huang et al. (2023) combine label enhancement with the analytic hierarchy process to derive feature importance under heterogeneous label contributions. These studies clearly show that aggregation in MLFS is not restricted to simple filter–embedded pipelines. At the same time, the dominant combination mechanisms remain additive, order-statistics-based, or distance-based, and therefore do not explicitly model synergy or redundancy between scoring sources.

Orthogonally to the scoring mechanism, several works address high label-space dimensionality through *label embedding*. The PLST of Tai and Lin (2012) projects the label matrix into a low-rank code space via singular value decomposition, and feature-aware extensions such as CPLST (Chen & Lin, 2012) further improve the embedding quality. Although originally conceived for classification, such embeddings can also serve as compact regression targets in embedded feature selection — a connection that our framework exploits explicitly by using a PLST-style embedding as the supervision signal for the convex embedded stage.

Taken together, the literature suggests that MLFS benefits from combining multiple perspectives, but current aggregation mechanisms still leave an important modelling gap. Embedded methods provide discriminative rankings tied to predictive objectives, whereas graph-based methods capture topological and spectral structure that may be highly informative yet only indirectly related to the predictive model. In the broader information-fusion literature, non-additive aggregators such as the Choquet fuzzy integral offer a principled way to reconcile such heterogeneous sources by explicitly modelling synergy and redundancy through fuzzy capacities (Grabisch, 1996; Grabisch & Labreuche, 2007). Moreover, Choquet-based fusion has already proved effective in classifier aggregation and ensemble design (Batista et al., 2022; Cao et al., 2012; Pacheco & Krohling, 2018), and this family of aggregation operators remains a very active research topic: recent work investigates 2-additive Choquet integrals as interpretable predictive models (Huang & Chen, 2024) and generalised Choquet-inspired aggregation functions (Bustince et al., 2026), while in multi-label learning fuzzy fusion mechanisms have recently been used to handle missing labels (Dai et al., 2025) and to aggregate multi-view embedded scores (Hao et al., 2025). However, to the best of our knowledge, this idea has not been explicitly investigated in MLFS as a score-level fusion mechanism between complementary *embedded* and *spectral* channels.

The proposed CFIFS framework is designed precisely in this space. Rather than treating embedded and spectral scores as competing alternatives, CFIFS interprets them as interacting information sources and combines them in a late-fusion architecture through a two-source Choquet integral. This yields a data-driven and explicitly non-additive combination rule, capable of adapting to agreement, redundancy, or complementarity between the channels. Accordingly, the contribution of this work lies not only in the individual scoring stages, but in the study of *how* and *when* their outputs should be fused, thereby opening a distinct point in the MLFS design space beyond conventional additive aggregation.

3. Choquet fuzzy integral feature selection

CFIFS produces a training-only feature ranking by combining two complementary scoring mechanisms—one *embedded* and one *spectral*—within a single unified pipeline.

In order to formalize the proposed approach, let $X \in \mathbb{R}^{n\times d}$ be the feature matrix and $Y \in \{0,1\}^{n\times L}$ the label matrix for one training fold, where n is the number of instances, d the number of features, and L the number of labels. We denote by $x_i^\top$ the ith row of X, by $y_i^\top$ the ith row of Y, and by $x_f = X_{:,f} \in \mathbb{R}^n$ the fth feature column.

3.1. Convex embedded scoring

The first scoring stage builds on the $\ell_{2,1}$ group-sparse regression framework (Nie et al., 2010) and the PLST label embedding (Tai & Lin, 2012). It learns a convex model and extracts per-feature importance from the group norms of the optimised coefficient matrix.

The procedure begins by computing a training-only label embedding $V \in \mathbb{R}^{L\times r}$ whose columns are the top-r right singular vectors of the label matrix Y (PLST Tai & Lin, 2012), and forming the embedded regression target

$$T = YV \in \mathbb{R}^{n\times r}. \tag{1}$$

Given this target, the embedded stage fits a shared group-sparse coefficient matrix $W \in \mathbb{R}^{d\times r}$ and a label bias $b \in \mathbb{R}^L$ by minimising

$$\min_{W,b} \ \frac{1-\alpha}{2}\left\|D_s\left(XW-T\right)\right\|_F^2 + \alpha\sum_{i=1}^{n} s_i \sum_{\ell=1}^{L} \mathrm{CE}\left(Y_{i\ell},\, \sigma\left(x_i^\top W v_\ell + b_\ell\right)\right) + \beta_{\mathrm{sp}}\|W\|_{2,1} + \rho\|W\|_F^2, \tag{2}$$

where $\mathrm{CE}(y,\hat{y}) = -[y\log\hat{y} + (1-y)\log(1-\hat{y})]$ is the binary cross-entropy, $\sigma(z) = 1/(1+e^{-z})$ is the logistic sigmoid, $v_\ell \in \mathbb{R}^r$ denotes the ℓth row of V, and $\alpha \in [0,1]$ blends the embedded reconstruction loss with the embedding-aware logistic term. The remaining ingredients of (2) deserve a separate explanation, since each of them affects the final ranking:

- *Inverse-frequency instance weighting.* The diagonal matrix $D_s = \mathrm{diag}(\sqrt{s})$ applies a standard inverse-frequency instance reweighting (Charte et al., 2015; He & Garcia, 2009) to mitigate the well-known label-imbalance bias in multi-label data: the reconstruction residual of each instance is re-weighted by $s_i = 1 + \kappa\sum_{\ell=1}^{L} Y_{i\ell}\, p_\ell$, where $p_\ell \propto (f_\ell+1)^{-\gamma}$ is a power-law inverse-frequency prior over labels and $f_\ell = \sum_i Y_{i\ell}$ is the number of positive instances of label ℓ.

Instances carrying rare labels therefore receive weights larger than 1, counteracting the tendency of the squared loss to be dominated by frequent labels. The same weights s_i multiply the per-instance logistic term.

- *Reweighting hyper-parameters.* The exponent $\gamma \geq 0$ controls how aggressively the prior concentrates on rare labels (γ=0 yields uniform weights; larger γ emphasises rarer labels more strongly), while $\kappa \geq 0$ bounds the overall strength of the reweighting: each weight satisfies $1 \leq s_i \leq 1 + \kappa$, so κ caps the maximum boost an instance can receive.
- *Sparsity and stability.* The penalty $\|W\|_{2,1} = \sum_{j=1}^{d} \|W_{j:}\|_2$ enforces group-sparsity across the rows of W, driving entire feature rows to zero and thereby inducing feature selection, while the small ridge term $\rho\|W\|_F^2$ guarantees strong convexity and numerical stability.

Since V is fixed within a fold, (2) is convex in (W, b) and solved by accelerated proximal gradient descent.

Once the optimisation has converged, the per-feature importance is read off as the row norm $g_j^{\text{emb}} = \|W_{j:}\|_2$, which directly reflects the contribution of feature j across all embedding dimensions.

3.2. Spectral graph-diffusion scoring

The second scoring stage extends the Laplacian Score of He et al. (2005) and the HSIC relevance criterion of Gretton et al. (2005) to the multi-label setting. Instead of fitting a predictive model, it evaluates each feature through the lens of a label-induced sample graph: features that vary smoothly on this graph and are statistically dependent on the label structure receive high scores. The computational cost is $O(n^2(L + d))$ with no iterative solver.

The starting point is an $n \times n$ symmetric affinity matrix A constructed from the training labels via Jaccard similarity:

$$A_{ij} = \frac{y_i^\top y_j}{\|y_i\|_1 + \|y_j\|_1 - y_i^\top y_j}, \qquad A_{ii} = 0. \tag{3}$$

Jaccard is scale-invariant and naturally handles the varying label cardinalities found in multi-label data.

From this affinity we derive the degree matrix $D = \text{diag}\left(\sum_j A_{ij}\right)$ and the normalised Laplacian

$$\mathcal{L} = I_n - D^{-1/2} A D^{-1/2}. \tag{4}$$

The Laplacian gives rise to a natural smoothness criterion: for each feature f, the Dirichlet energy on the label graph measures how much the feature signal varies between similarly-labelled instances:

$$E(f) = \frac{x_f^\top \mathcal{L} x_f}{x_f^\top x_f + \epsilon}, \tag{5}$$

where $\epsilon = 10^{-10}$ is a small regulariser. A *low* Dirichlet energy indicates that the feature varies smoothly over the label graph, suggesting relevance for label prediction. Fig. 1 illustrates the concept on the *Flags* dataset, one of the standard benchmark datasets detailed later in Section 4: the smooth feature (E=0.268) exhibits a coherent trend along the spectral ordering of the label-affinity graph, while the noisy feature (E=0.956) oscillates randomly, and the full distribution of Dirichlet energies separates the two regimes clearly.

While smoothness captures how well a feature aligns with the label graph, it does not directly measure statistical dependence. To address this, we complement it with the (biased) HSIC between each feature and the label matrix, a kernel dependence measure with a long and still very active record in feature selection (Gretton et al., 2005; Wang, 2026; Wang et al., 2023; Yamada et al., 2014). Using the label affinity A as the label kernel, we apply standard double-centering:

$$K_Y^c = H A H, \qquad H = I_n - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top, \tag{6}$$

and compute the per-feature HSIC score:

$$H(f) = \frac{x_f^\top K_Y^c x_f}{x_f^\top x_f + \epsilon}. \tag{7}$$

A *high* $H(f)$ indicates strong statistical dependence between the feature and the label structure.

Finally, both raw score vectors are min–max normalised to $[0, 1]$ and combined into a single spectral score:

$$g_f^{\text{spec}} = (1 - \alpha_S)\left(1 - \tilde{E}(f)\right) + \alpha_S \tilde{H}(f), \tag{8}$$

where $\tilde{E}$ and $\tilde{H}$ are the normalised Dirichlet and HSIC scores, and $\alpha_S \in [0, 1]$ controls the relative importance of the two criteria. In all experiments we set $\alpha_S = 0.70$, giving more weight to HSIC relevance. Both (5) and (7) are computed in vectorised form as the diagonals of $X^\top \mathcal{L} X$ and $X^\top K_Y^c X$, respectively; the dominant cost is $O(n^2\, d)$, fully parallelisable on Graphics Processing Unit (GPU).

3.3. Data-driven fusion via inner cross-validation

The embedded and spectral scores capture complementary aspects of feature quality: the former exploits group-sparse discriminative structure in a convex model, while the latter leverages global label-graph topology. CFIFS combines them via a score-level fusion whose balance is determined *automatically* from the training data.

Let $\hat{g}^{\text{emb}}, \hat{g}^{\text{spec}} \in [0, 1]^d$ be the *rank-normalised* embedded and spectral score vectors, obtained by applying an empirical Cumulative Distribution Function (CDF) transform (average ranks) on the training fold. This maps each channel to a comparable, outlier-robust scale before fusion. We fuse them with a two-source Choquet fuzzy integral (Grabisch, 1996; Grabisch & Labreuche, 2007; Murofushi & Sugeno, 1989). Let μ be a (possibly non-additive) capacity on $\{\text{emb}, \text{spec}\}$ with $\mu(\emptyset)$=0 and $\mu(\{\text{emb}, \text{spec}\})$=1, and denote $\mu_e = \mu(\{\text{emb}\})$ and $\mu_s = \mu(\{\text{spec}\})$. For two sources the Choquet integral has a closed form; the fused per-feature score is

$$g_j = C_\mu\left(\hat{g}_j^{\text{emb}}, \hat{g}_j^{\text{spec}}\right) = \begin{cases} \hat{g}_j^{\text{spec}} + \left(\hat{g}_j^{\text{emb}} - \hat{g}_j^{\text{spec}}\right)\mu_e, & \hat{g}_j^{\text{emb}} \geq \hat{g}_j^{\text{spec}}, \\ \hat{g}_j^{\text{emb}} + \left(\hat{g}_j^{\text{spec}} - \hat{g}_j^{\text{emb}}\right)\mu_s, & \text{otherwise}, \end{cases} \quad j = 1, \ldots, d, \tag{9}$$

which always lies between $\min(\hat{g}_j^{\text{emb}}, \hat{g}_j^{\text{spec}})$ and $\max(\hat{g}_j^{\text{emb}}, \hat{g}_j^{\text{spec}})$. When $\mu_e + \mu_s = 1$, (9) reduces to an additive (linear) weighted mean; when $\mu_e + \mu_s \neq 1$, it becomes non-additive and can express redundancy or complementarity between the two channels.

Remark 1 (Interaction index)**.** For two sources, the Murofushi–Soneda (Shapley) interaction index (Grabisch, 1996; Murofushi & Soneda, 1993) of the pair $\{\text{emb}, \text{spec}\}$ reduces to

$$I = \mu(\{\text{emb}, \text{spec}\}) - \mu_e - \mu_s = 1 - \mu_e - \mu_s. \tag{10}$$

Note that on a two-element universe *every* capacity is trivially 2-additive, so (10) coincides exactly with the standard interaction index of 2-additive measures, with no further assumption required.

The behavioural reading of I that we use throughout the article—I>0 as agreement-seeking (conjunctive) fusion and I<0 as disjunctive fusion—is not a heuristic interpretation but follows from an exact decomposition of the two-source Choquet integral.

Proposition 1 (Shapley decomposition of the two-source Choquet integral)**.** *Let μ be any capacity on $\{\text{emb}, \text{spec}\}$ with singleton values μ_e, μ_s and $\mu(\{\text{emb}, \text{spec}\})$=1, and let $I = 1 - \mu_e - \mu_s$ be the interaction index (10). Then, for all $e, s \in [0, 1]$,*

$$C_\mu(e, s) = \phi_e\, e + \phi_s\, s - \frac{I}{2}\,|e - s|, \phi_e = \frac{1 + \mu_e - \mu_s}{2}, \phi_s = \frac{1 + \mu_s - \mu_e}{2}, \tag{11}$$

where $\phi_e, \phi_s \geq 0$ are the Shapley values of the two sources and $\phi_e + \phi_s = 1$.

Proof. Assume $e \geq s$ (the case $e < s$ is symmetric). By (9), $C_\mu(e, s) = s + (e - s)\mu_e = \mu_e e + (1 - \mu_e)s$. Substituting ϕ_e, ϕ_s and I into the right-hand

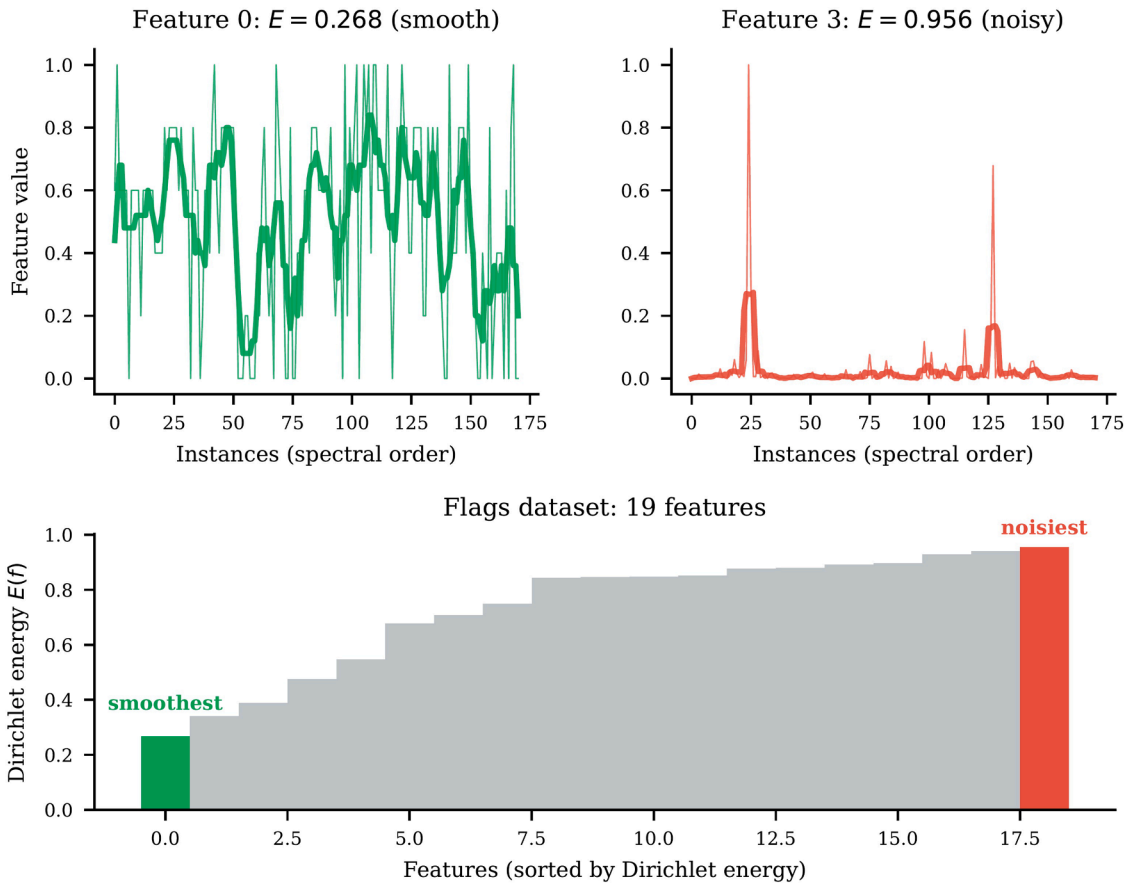

Fig. 1. Spectral smoothness illustration on the *Flags* dataset (d=19). **Top row:** feature signal plotted over instances sorted by the Fiedler vector (spectral ordering of the label-affinity graph). The smooth feature (left, E=0.268) shows a coherent trend—its values align with the graph structure—while the noisy feature (right, E=0.956) oscillates randomly. **Bottom:** Dirichlet energy distribution for all 19 features; the two highlighted features are marked.

side of (11) with $|e-s| = e-s$ gives $\phi_e e + \phi_s s - \frac{I}{2}(e-s) = \left(\phi_e - \frac{I}{2}\right)e + \left(\phi_s + \frac{I}{2}\right)s = \mu_e e + (1-\mu_e)s$, since $\phi_e - \frac{I}{2} = \frac{(1+\mu_e-\mu_s)-(1-\mu_e-\mu_s)}{2} = \mu_e$. □

Property 1 makes the semantics of I explicit: the fused score is a fixed additive (Shapley-weighted) mean of the two channels, *corrected by a disagreement term* proportional to $|e-s|$. When $I > 0$ the correction is a penalty: features on which the two channels disagree are demoted, so a high fused score requires *joint* support—a conjunctive, agreement-seeking behaviour. When $I < 0$ the correction is a bonus: disagreement is rewarded, so a single strongly supportive channel suffices—a disjunctive behaviour suited to redundant or unevenly reliable sources. When $I = 0$ the correction vanishes and (11) reduces to the additive weighted mean.

The same decomposition clarifies *why* non-additive fusion is necessary at all: the disagreement term cannot be reproduced by any linear rule, and there are ranking behaviours that no additive aggregation can express.

Proposition 2 (Limits of additive fusion). *The feature ranking induced on* $[0,1]^2$ *by the conjunctive two-source Choquet integral* $C_{\min}(e,s) = \min(e,s)$ *(the case* $\mu_e = \mu_s = 0$*,* $I = 1$*) cannot be reproduced by any additive aggregation* $A_w(e,s) = we + (1-w)s$ *with* $w \in [0,1]$*: for every weight* w *there exist score pairs whose order under* $C_{\min}$ *is reversed by* A_w*. The disjunctive extreme* $C_{\max}(e,s) = \max(e,s)$ *(*$I = -1$*) is non-reproducible as well.*

Proof. Fix an arbitrary $w \in [0,1]$ and consider the three score pairs $a = (0.9, 0.1)$, $b = (0.45, 0.45)$, $c = (0.1, 0.9)$, used here only as a witness. Under $C_{\min}$, $C_{\min}(b) = 0.45 > 0.1 = C_{\min}(a) = C_{\min}(c)$, so b is ranked strictly above both a and c. Under A_w, however, $A_w(a) + A_w(c) = (0.1 + 0.8w) + (0.9 - 0.8w) = 1$, hence $\max\{A_w(a), A_w(c)\} \geq \frac{1}{2} > 0.45 = A_w(b)$, so A_w places a or c above b and reverses the order. As w was arbitrary, no additive rule reproduces the $C_{\min}$ ranking. For the disjunctive case, the reflection $(e,s) \mapsto (1-e, 1-s)$ sends $C_{\min}$ to $1 - C_{\max}$ and A_w to $1 - A_w$, so reproducing the $C_{\max}$ ranking by an additive rule would reproduce the $C_{\min}$ one, a contradiction. □

The practical consequence is that whenever the informative features are those *jointly* supported by the embedded and spectral channels (synergy), or conversely those strongly supported by *either* channel (redundancy), a linear combination—however well tuned—cannot reproduce the desired ordering, while the one-parameter family swept by I in (11) interpolates continuously between these regimes. Whether the data call for $I \neq 0$ is an empirical question, which is precisely what the inner cross-validation procedure described next is designed to answer; the analysis of the learned capacities in Section 4.3 shows that $|I| \geq 0.15$ on 9 of the 20 benchmarks.

In order to apply Eq. (9), a practical question is how to set the capacity parameters (μ_e, μ_s). A fixed choice is simple but cannot adapt to dataset-specific characteristics. In fact, on some datasets the spectral channel is noisy and should be discounted, while on others it carries complementary information that the embedded model cannot capture. Rather than relying on a closed-form heuristic, CFIFS determines (μ_e, μ_s) via an *inner CV* loop that is entirely contained within the training fold. The protocol is designed so that the data used to *evaluate* a candidate capacity never contribute to the scores being fused: within the inner loop, both scoring stages are recomputed from scratch on each inner-training partition, so no statistic derived from inner-validation labels—in particular, neither the label-affinity graph of Section 3.2 nor the embedded model of Section 3.1—enters the candidate evaluation. Specifically, the

inner cross-validation protocol is executed through the following five steps:

1. Split the training fold into K_{in} inner folds.
2. For each inner split, recompute the embedded and spectral scores on the *inner-training partition only*, and rank-normalise them; denote these split-local vectors $\hat{g}^{\text{emb}}_{(i)}, \hat{g}^{\text{spec}}_{(i)}$ for inner split i.
3. For each candidate (μ_e, μ_s) on a discrete search grid Ω_μ and each inner split i:
   a Fuse $\hat{g}^{\text{emb}}_{(i)}, \hat{g}^{\text{spec}}_{(i)}$ via (9) and select the top fraction p_{frac} of features.
   b Train a lightweight Multi-label k-Nearest Neighbor (ML-kNN) classifier (with k neighbours and cosine distance) on the inner-training partition and evaluate it on the inner-validation instances, recording both Micro-F1 and Macro-F1.
4. Compute the *Geometric Mean* (GM) criterion $C(\mu_e, \mu_s) = \sqrt{\text{Micro-F1}(\mu_e, \mu_s) \cdot \text{Macro-F1}(\mu_e, \mu_s)}$, averaging each metric over the K_{in} inner splits, and select $(\mu_e^*, \mu_s^*) = \arg\max C$.
5. Fuse the full-fold score vectors $\hat{g}^{\text{emb}}, \hat{g}^{\text{spec}}$ (computed once on the whole training fold) with (μ_e^*, μ_s^*) and output the final ranking.

Note that using the full training fold in the last step is legitimate: once the capacity has been selected, the final ranking may exploit all available training data, exactly as a model is refit on the full training set after hyper-parameter selection. Three properties of this procedure are worth making explicit. First, the search space is not an arbitrary heuristic restriction: for two sources, a normalised capacity is fully determined by the pair $(\mu_e, \mu_s) \in [0,1]^2$ (monotonicity is automatically satisfied since $0 \le \mu_e, \mu_s \le \mu(\{\text{emb}, \text{spec}\}) = 1$), so Ω_μ is an exhaustive uniform discretisation of the *entire* space of admissible capacities, and the procedure performs empirical risk minimisation of the validated downstream criterion over that space. Second, because Ω_μ is finite and is enumerated exhaustively, the selected pair (μ_e^*, μ_s^*) is the *exact* global maximiser of the inner-CV criterion over Ω_μ; no iterative optimisation, and hence no convergence issue, is involved. Moreover, the criterion is a piecewise-constant function of (μ_e, μ_s): by (9) the fused scores are piecewise linear in the capacity, so the selected top-p_{frac} subset—and with it the criterion value—changes only when the capacity crosses one of finitely many boundaries where two features swap fused-score order. The grid resolution is therefore the only approximation, and refining it beyond $\Delta\mu$=0.20 produced no measurable change in our experiments (Section 4.1). Third, the criterion itself is chosen on principled grounds. The geometric mean balances instance-level accuracy (Micro-F1, dominated by frequent labels) with label-level accuracy (Macro-F1, which weights all labels equally), and among the standard means it has two properties that suit capacity selection: (i) it vanishes whenever either component vanishes, acting as a veto against degenerate capacities that maximise one metric by sacrificing the other (e.g., ignoring all rare labels); and (ii) by the AM–GM inequality, for a fixed sum of the two metrics the GM strictly increases as they become more balanced, so it systematically prefers balanced solutions over lopsided ones with the same aggregate level. We do not claim that GM is the uniquely optimal choice; it is a transparent, parameter-free compromise, and Section 4.4 reports an empirical comparison with single-metric criteria (Micro-F1 only, Macro-F1 only), showing that GM matches the best single-metric criterion on its favoured metric while avoiding its weakness on the other. This procedure is fully automatic: no fusion parameters are fixed across datasets. Importantly, the scoring stages are recomputed once per inner split—not once per capacity candidate—because the score vectors do not depend on (μ_e, μ_s): each candidate evaluation then only performs fusion, feature masking, and an ML-kNN fit–predict. The cost of the inner-CV is therefore K_{in} additional runs of the two scoring stages plus $K_{\text{in}} \times |\Omega_\mu|$ lightweight ML-kNN evaluations on the selected feature subset, and the latter dominate in practice (see Section 3.4 for a detailed breakdown).

Remark 2 (Independence of the capacity-selection signal)**.** Recomputing the scores on each inner-training partition keeps the capacity-selection signal independent of the data on which it is evaluated: the label-affinity graph (3), the Laplacian (4), the HSIC kernel (6), and the embedded model (2) are all built without the inner-validation instances, whose labels are used exclusively to compare candidate capacities. The outer test fold, in turn, is never accessed by any stage of the pipeline, so the reported estimates are unbiased by construction.

Beyond accuracy, the learned capacity provides interpretability: (μ_e^*, μ_s^*) indicates how much each channel is trusted when it dominates, while the interaction index (10) summarises whether the fusion behaves more conjunctively or disjunctively on a dataset.

The empirical motivation for this design is visible in Fig. 2: on three representative benchmark datasets (detailed later in Section 4), the scatter plots reveal a substantial cloud of features in the top-20% of one channel but not the other (orange and blue points), whereas only a minority of features are jointly top-ranked (green points). The phenomenon is especially pronounced on *Science* (d=743), where the two channels identify almost disjoint top-feature sets. Fig. 3 quantifies this: the overlap between the top-k embedded and top-k spectral rankings remains below 50% at the k/d=20% operating point on all three datasets, and drops below 30% on *Science*. The low overlap persists across the full range of selection ratios, confirming that the two scoring mechanisms identify structurally different informative features—a necessary condition for score-level fusion to outperform either channel alone.

Algorithm 1 summarises the complete CFIFS training-fold procedure. Conceptually, the pipeline is a three-stage cascade. The first two stages—convex embedded scoring and spectral graph-diffusion scoring—run independently on the full training fold and produce two d-dimensional importance vectors. Because these stages share no parameters, they can in principle be executed in parallel. The third stage performs the data-driven Choquet fusion described above: it rank-normalises the two score vectors, sweeps the capacity grid Ω_μ via inner-CV, and selects the capacity that maximises the geometric-mean criterion. The output is a single fused score vector from which the final feature ranking is obtained by descending sort.

Two aspects of this design deserve emphasis. First, the inner-CV loop is lightweight by construction: the scoring stages are recomputed only once per inner split (K_{in} extra runs in total), while each of the $K_{\text{in}} \times |\Omega_\mu|$ candidate evaluations performs only fusion, feature masking, and an ML-kNN fit–predict. Second, the entire pipeline is *training-only*: the test fold is never accessed during scoring or capacity selection, so the final ranking is produced without any access to the test data; and, within the training fold, the inner-validation labels used to compare capacity candidates never enter the score vectors being fused (Remark 2).

3.4. Computational complexity

Since CFIFS fuses two independent scoring stages and adds a data-driven inner-CV loop, its per-fold cost is inherently higher than that of any single-paradigm baseline. We analyse each component in turn and provide empirical GPU timings to contextualise the overhead.

The *embedded* stage is dominated by a truncated Singular Value Decomposition (SVD) of the $n \times L$ label matrix, costing $O(n\,L\,r)$ where r is the embedding rank, followed by T proximal-gradient iterations at $O(n\,d\,r)$ each; the total is $O(n\,L\,r + T\,n\,d\,r)$, with $T \le 120$ in all our experiments. The *spectral* stage builds the $n \times n$ Jaccard affinity matrix in $O(n^2\,L)$ and computes the Dirichlet energies and HSIC scores for all d features as diagonals of $X^\top \mathcal{L}\, X$ and $X^\top K_Y^c\, X$, each in $O(n^2\,d)$; the total is $O(n^2(L+d))$ with no iterative solver. The combined asymptotic cost of the two stages is therefore $O(Tndr + n^2(L+d))$.

The inner-CV adds two contributions. First, the scoring stages are recomputed once per inner split on the inner-training partition ($n_{\text{in}} \approx (1 - 1/K_{\text{in}})\,n$ instances), adding K_{in} runs of the $O(Tn_{\text{in}}dr + n_{\text{in}}^2(L+d))$ cost above; the score vectors do not depend on the capacity, so they are *not* recomputed per candidate. Second, for each (μ_e, μ_s, inner split) triple, only fusion, feature selection, and a lightweight ML-kNN evaluation are performed: with K_{in} inner folds and $|\Omega_\mu|$ capacity candidates,

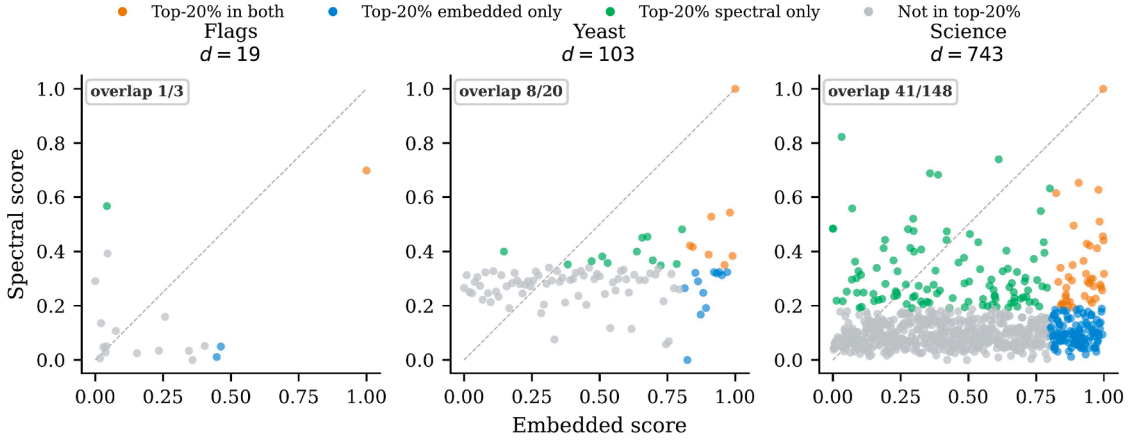

Fig. 2. Embedded vs. spectral score complementarity on three datasets of increasing dimensionality: *Flags* (d=19), *yeast* (d=103), and *Science* (d=743). Each point is one feature; colour encodes top-20% membership: green = selected by both channels, orange = embedded only, blue = spectral only, grey = neither. The substantial fraction of single-channel features (orange and blue) motivates score-level fusion.

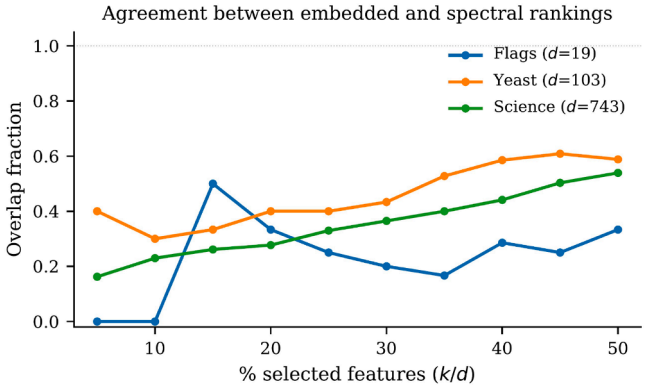

Fig. 3. Overlap fraction between the top-k embedded and top-k spectral feature rankings as a function of the selection ratio k/d, shown for *Flags* (d=19), *yeast* (d=103), and *Science* (d=743). On all three datasets the overlap stays well below 1.0, confirming that the two scoring mechanisms identify substantially different feature subsets, especially at tight budgets ($k/d \leq 20\%$).

this amounts to $K_{\text{in}} \times |\Omega_\mu|$ ML-kNN fit–predict calls on feature subsets of size $\lfloor p\,d \rfloor$, each costing $O(n_{\text{in}}^2\, k)$ for the distance computation, for a total of $O(K_{\text{in}}\, |\Omega_\mu|\, n_{\text{in}}^2\, k)$. With the settings reported in Section 4.1 (K_{in}=3, $|\Omega_\mu|$=36, k=10), this amounts to 108 evaluations per fold, which dominate the 3 scoring recomputations in practice.

On a single NVIDIA RTX 4090 GPU, the per-fold timings for representative datasets are reported in Table 1. As can be seen, both scoring stages are fast on GPU: even for *Corel16k1* (n_{tr}=12,405) scoring takes under 0.5 s. Second, the dominant cost is the inner-CV loop (88–94% of total time), within which the $K_{\text{in}} \times |\Omega_\mu|$ ML-kNN evaluations (108 with our settings) far outweigh the K_{in} scoring recomputations. Nevertheless, the total wall-clock time per fold remains under 8 s even for the largest dataset, making the inner-CV overhead modest in absolute terms. By comparison, a single baseline evaluation (ML-kNN at a fixed p on the same GPU) takes 0.1–1.0 s depending on dataset size, so the full CFIFS pipeline is roughly 5–10× slower than evaluating a single pre-computed ranking—a price that the quality gains documented in Section 4.6 amply justify.

In terms of memory, the spectral stage requires storing a small constant number of dense $n \times n$ matrices (the label affinity and its normalised/centred variants), and this $O(n^2)$ footprint is the binding constraint on the scalability of the exact method. On the largest benchmark used here, *Corel16k1* (n_{tr}=12,405 per training fold), one such matrix in float32 occupies approximately 0.6 GB, which fits comfortably on a single GPU. The quadratic growth, however, makes the exact formulation impractical beyond a few tens of thousands of instances: at n=50,000 a single matrix already requires ∼10 GB, and at n=100,000 it reaches ∼40 GB, exceeding the memory of standard accelerators. The exact CFIFS pipeline presented in this article therefore targets the small-to-moderate instance regime that characterises the vast majority of public multi-label benchmarks (hundreds to low tens of thousands of instances). Extending it to substantially larger corpora is *not* a mere implementation detail: it requires replacing the dense label graph with an approximation—k-NN sparsification of the affinity matrix, Nyström or anchor-graph low-rank factorisations, or block-streaming computation of the two quadratic forms—each of which reduces memory to near-linear in n but introduces an approximation error whose effect on the final ranking must be quantified. We regard this as a separate research question and discuss it as future work in Section 5.

Overall, it is important to note that feature selection is an *offline, one-time* operation: once the selected subset is identified, all subsequent inference runs—which may number in the millions in a deployed system—use only the reduced feature set at no extra cost. The selection phase is therefore amortised over the entire lifetime of the downstream model. For the ordinary-sized datasets typical of multi-label benchmarks (hundreds to low tens of thousands of instances, hundreds to low thousands of features), investing a few additional seconds per fold to fuse complementary scoring paradigms and calibrate their combination in a data-driven manner is a sensible trade-off: it yields a higher-quality feature subset and, equally important, richer insight into the phenomenon under study, since the practitioner obtains both an embedded-relevance and a spectral-dependency view of each feature. In absolute terms, the fusion design trades a moderate increase in wall-clock time—seconds, not minutes—for statistically significant accuracy gains. The overhead arises primarily from the inner-CV (not from running two scoring stages), and could be reduced by coarsening the Ω_μ, restricting the capacity to be additive, or lowering K_{in} with only marginal impact on capacity selection quality. It is worth noting that, like the spectral stage, the inner-CV evaluations are quadratic in the number of instances (each ML-kNN call costs $O(n_{\text{in}}^2\, k)$), so on much larger datasets the loop would inherit the same scaling limits discussed above. Three complementary mitigations are available without altering the method: (i) the $K_{\text{in}} \times |\Omega_\mu|$ evaluations are embarrassingly parallel, since each one is an independent fit–predict on fixed score vectors; (ii) the capacity can be selected on a fixed-size subsample of the training fold, because it is a single discrete pair rather than a high-dimensional parameter; and (iii) the exact k-NN search can be replaced by an approximate index with negligible effect on the relative comparison of capacity candidates.

Algorithm 1: CFIFS: training-fold feature ranking.

Input: Training fold (X, Y); embedded hyper-parameters; spectral parameter α_S; inner-CV folds K_{in}, search grid Ω_μ, selection fraction p_{frac}, neighbours k.
Output: Feature ranking π (descending importance).

`// Stage 1: Convex embedded scoring (Section 3.1)`
Compute label embedding V, target T, inverse-frequency instance weights s
Solve convex objective (2) $\rightarrow W^*$
Extract embedded scores $g_j^{\text{emb}} = \|W_{j:}^*\|_2$ for $j = 1, \ldots, d$

`// Stage 2: Spectral graph-diffusion scoring (Section 3.2)`
Build label-affinity graph A via Jaccard (3)
Form normalised Laplacian $\mathcal{L}$ (4)
Compute Dirichlet energy $E(f)$ (5) and HSIC $H(f)$ (7) for all features
Blend into spectral scores g_f^{spec} (8)

`// Stage 3: Inner-CV Choquet fusion (Section 3.3)`
Rank-normalise both score vectors on the *full* training fold $\rightarrow$ $\hat{g}^{\text{emb}}, \hat{g}^{\text{spec}}$
Split (X, Y) into K_{in} inner folds
for *each inner split* i **do**
  Recompute Stages 1–2 on the inner-training partition only; rank-normalise $\rightarrow$ $\hat{g}_{(i)}^{\text{emb}}, \hat{g}_{(i)}^{\text{spec}}$
end
for *each* (μ_e, μ_s) *in* Ω_μ **do**
  for *each inner split* i **do**
    Fuse $\hat{g}_{(i)}^{\text{emb}}, \hat{g}_{(i)}^{\text{spec}}$ via (9), select top-p_{frac} features
    Train ML-kNN(k) on inner-train, evaluate on inner-val $\rightarrow$ Micro-F1(μ_e, μ_s), Macro-F1(μ_e, μ_s)
  end
end
$C(\mu_e, \mu_s) \leftarrow \sqrt{\text{Micro-F1}(\mu_e, \mu_s) \cdot \text{Macro-F1}(\mu_e, \mu_s)}$
$(\mu_e^*, \mu_s^*) \leftarrow \arg\max C(\mu_e, \mu_s)$
Fuse with (μ_e^*, μ_s^*) via (9) $\rightarrow g$
return $\pi = \text{argsort}(-g)$

Table 1
Empirical wall-clock times (in seconds) per fold on a single NVIDIA RTX 4090 GPU for representative datasets. n_{tr}, d, and L refer to the *training fold* (90% of the instances under 10-fold cross-validation); e.g., *Corel16k1* has 13,766 instances in total and 12,405 per training fold. "Scoring" includes the embedded and spectral stages on the full training fold, while "ICV" measures the complete inner cross-validation loop: the K_{in} per-split scoring recomputations plus the $K_{\text{in}} \times |\Omega_\mu|$ ML-kNN evaluations (108 with our settings).

Dataset	n_{tr}	d	L	Scoring	ICV	Total
Birds	580	260	19	0.10	0.75	0.85
Science	4,508	743	40	0.17	2.44	2.61
Slashdot	3,410	1,079	22	0.16	2.33	2.48
Corel16k1	12,405	500	153	0.47	6.32	6.79

4. Experiments

4.1. Experimental setup

We evaluate on 20 standard multi-label benchmarks spanning text categorisation, image annotation, audio classification, and biology (Table 2). The collection covers a wide range of dimensionalities (d from 19 to 1,449), label-space sizes (L from 5 to 174), and label densities (from 0.018 to 0.485), providing a stringent testbed for feature-selection methods. Specifically, the suite includes an audio classification dataset (*Birds*, d=260, L=19) and two biology datasets (*Human*, d=440, L=14; *Genbase*, d=1185, L=27) that test performance on different feature and label regimes; nine medium-dimensional text-categorisation benchmarks from the Yahoo corpus (Arts through Social, $d \in [438, 1047]$) that share a common domain but vary in label count and density; two image-annotation datasets with large label spaces (*Corel16k1*, L=153; *Corel16k2*, L=164) that stress label-correlation modelling; and several high-dimensional or specialised datasets (*Slashdot*, d=1079; *Genbase*, d=1185; *Medical*, d=1449; *Yelp*, n=10,806) that test scalability. In addition, *Image* (d=294, L=5) provides a compact vision benchmark, and *Yeast* (d=103, L=14) contributes a high-density biological dataset. This diversity is deliberately chosen to reveal whether a method's advantage is broad or confined to a narrow data regime. All features are min–max normalised to $[0, 1]$ on the training fold, and the same transformation is applied to the test fold.

Table 2
Summary of the 20 benchmark datasets. n: total instances; d: features; L: labels; Card.: label cardinality (mean labels per instance); Dens.: label density (Card./L).

Dataset	Domain	n	d	L	Card.	Dens.
Arts	text	5,000	462	26	1.65	0.064
Birds	audio	645	260	19	1.01	0.053
Business	text	5,000	438	30	1.60	0.053
Computers	text	5,000	681	33	1.51	0.046
Corel16k1	image	13,766	500	153	2.86	0.019
Corel16k2	image	13,761	500	164	2.88	0.018
Education	text	5,000	550	33	1.45	0.044
Entertain	text	5,000	640	21	1.41	0.067
Flags	image	194	19	7	3.39	0.485
Health	text	5,000	612	32	1.64	0.051
Recreation	text	5,000	606	22	1.43	0.065
Science	text	5,000	743	40	1.45	0.036
Slashdot	text	3,782	1,079	22	1.18	0.054
Social	text	5,000	1,047	39	1.27	0.033
Yelp	text	10,806	671	5	1.64	0.328
Human	biology	3,106	440	14	1.19	0.085
Image	vision	2,000	294	5	1.24	0.247
Genbase	biology	662	1,185	27	1.25	0.046
Medical	text	978	1,449	45	1.25	0.028
Yeast	biology	2,417	103	14	4.24	0.303

Using the datasets above, we investigate CFIFS and compare it against seven recent MLFS methods spanning filter, embedded, and graph-based paradigms: GRRO (Zhang et al., 2020) (filter, global relevance–redundancy optimisation), SRFS (Li et al., 2025) (embedded, sparse supplementation with information fusion), RFSFS (Li et al., 2023) (embedded, robust flexible sparse regularisation), LRMFS (Fan et al., 2025) (embedded, label-relaxed matrix factorisation), LSMFS (Fan et al., 2024a) (embedded, label relaxation with shared information), LRDG (Zhang et al., 2024) (graph-based, latent representation with dynamic graph), and SCNMF (He et al., 2024) (graph-based, similarity-constrained non-negative matrix factorisation). All baselines were executed using the authors' original code under the same data splits and evaluation protocol, with the default hyper-parameter configurations recommended in the respective code releases, kept fixed across all datasets. This mirrors the treatment of CFIFS, whose scoring-stage hyper-parameters are likewise fixed across datasets; the only adaptive component of CFIFS is the capacity pair, selected inside each training fold. The implications of this configuration policy—in particular for the baselines that collapse on specific datasets—are analysed explicitly in Section 4.6.

Specifically, each method produces a training-only feature ranking per fold; we then select the top-k features and evaluate a downstream ML-kNN classifier (k=10, smoothing s=1.0, cosine distance) on the reduced test set. ML-kNN is the standard evaluation classifier in the MLFS literature (Zhang & Zhou, 2007): it is parameter-light, does not internally perform feature selection, and therefore serves as a neutral probe

of the ranking quality. We adopt pre-frozen iterative-stratified 10-fold cross-validation splits (Sechidis et al., 2011; Szymański & Kajdanowicz, 2017) (90%/10% train/test ratio), ensuring that all methods are evaluated on exactly the same data partitions. We report seven metrics: Micro-F1, Macro-F1, and Hamming Loss as primary metrics, together with Micro/Macro PR-AUC, example-based Average Precision, and One-Error as additional ranking-quality indicators.

For compact tabular comparisons and statistical tests we use p=20% of selected features as the primary operating point, which is standard in the MLFS literature. Statistical significance is assessed via Friedman tests followed by paired Wilcoxon post-hoc comparisons with Holm correction. Since the Friedman statistic is sensitive to ties, rankings are assigned as *mid-ranks*: when two or more methods achieve the same value on a dataset, each receives the average of the rank positions they jointly occupy, and the tie-corrected version of the Friedman statistic is used. In practice ties are rare (they occur on two of the 20 datasets per metric, and only among baselines) and do not affect any of the conclusions. In the post-hoc Wilcoxon tests, zero differences are discarded following Wilcoxon's original treatment, and tied non-zero differences receive average ranks. Besides the representative feature budget, we evaluate a p-grid with $p \in \{5, 10, 15, \ldots, 50\}$% selected features for all methods and datasets, and report mean curves across datasets.

Finally, the key hyperparameters of CFIFS are as follows. *Embedded stage:* regression–logistic blend α=0.35, embedding rank r=50 (capped to L when $L < 50$), group-sparsity $\beta_{sp}=10^{-2}$, Frobenius ridge $\rho=10^{-4}$, proximal gradient with T=120 iterations, inverse-frequency instance weights (γ=2.0, κ=1.5). *Spectral stage:* HSIC–smoothness blend α_S=0.70. *Fusion:* rank-normalise each score channel via empirical CDF; K_{in}=3 inner folds; Choquet capacity grid $\mu_e, \mu_s \in \{0.00, 0.20, \ldots, 1.00\}$ (36 pairs, free/non-additive capacity); inner-CV GM selection criterion at p=20%. All embedded and spectral parameters are fixed across all 20 datasets; no per-dataset tuning is performed. The only component that adapts to the training data is the inner-CV capacity selection, which is entirely contained within each training fold.

We briefly justify the main fixed choices. The embedding rank r=50 follows the PLST default (Tai & Lin, 2012) and is capped to L for small label spaces, so it introduces no free parameter. The HSIC–smoothness blend α_S=0.70 reflects the empirical observation that HSIC provides a more discriminative per-feature signal than Dirichlet energy alone, which tends to assign similar scores to many features; the 70:30 ratio was selected on a small held-out pilot study (three datasets) and kept frozen for all experiments. The capacity grid resolution $\Delta\mu$=0.20 (yielding 6×6=36 candidates) balances modelling flexibility of the non-additive search with the computational cost of the inner-CV; a finer grid ($\Delta\mu$=0.10) gave no measurable improvement in preliminary tests while tripling the inner-CV evaluations.

4.2. Ablation study

To isolate the contribution of the key components of CFIFS, we perform a twofold ablation study on the p=20% operating point. We test both the structural validity of the rank-based transformation and the mathematical validity of the non-additive Choquet fusion. All configurations share the same embedded and spectral scoring stages; they differ in how scores are normalised and fused.

(i) Structural Ablation (Fusion vs No-Fusion, and Ranking). We first verify the necessity of the fusion strategy itself by dropping the spectral features (*Embedded only*). As shown in Tables 3 and 4 (Structural Ablation), dropping the fusion causes a statistically significant degradation: compared to the proposed CFIFS (CFI with rank normalisation), the Embedded-only approach degrades on both Micro-F1 (Holm $p=2.93\times10^{-4}$) and Macro-F1 (Holm $p=7.75\times10^{-5}$), and the Spectral-only approach degrades even more sharply (Holm $p<4\times10^{-6}$ on both metrics). This confirms that the multi-source fusion contributes materially to performance. Furthermore, we compare our rank-normalised fusion (CFIFS) against standard Min–Max normalisation (*Choquet (Min/Max)*) in the Capacity Ablation tables (Tables 5 and 6). The rank-normalised variant is statistically superior to Min–Max normalisation on both Micro-F1 (18/0/2 wins, Holm $p=1.25\times10^{-2}$) and Macro-F1 (17/1/2 wins, Holm $p=2.53\times10^{-3}$), indicating that nonparametric ranking helps protect minority classes from being suppressed by extreme calibration bounds during fusion.

Table 3
Performance comparison evaluated using the Micro-F1 (Component Ablation) metric (higher is better). Mean results ± standard deviation are reported across 10 cross-validation folds. The best value for each dataset is highlighted in bold. * indicates statistical significance according to the Wilcoxon paired test with Holm correction ($p < 0.05$).

Dataset	Embedded Only	Spectral Only	+ Choquet Fusion
Arts	0.1380 ± 0.0288	0.0490 ± 0.0093	**0.1703** ± 0.0045
Birds	0.2502 ± 0.0800	0.2735 ± 0.0798	**0.3602** ± 0.0678
Business	0.6873 ± 0.0091	0.6753 ± 0.0100	**0.6922** ± 0.0091
Computers	0.3758 ± 0.1183	**0.3841** ± 0.0917	0.3728 ± 0.0970
Corel16k1	**0.0337** ± 0.0034	0.0075 ± 0.0034	0.0336 ± 0.0052
Corel16k2	0.0379 ± 0.0061	0.0130 ± 0.0042	**0.0380** ± 0.0047
Education	0.1492 ± 0.0156	0.0154 ± 0.0087	**0.1642** ± 0.0145
Entertain	0.2938 ± 0.0218	0.2087 ± 0.0286	**0.3048** ± 0.0312
Flags	0.6629 ± 0.0257	0.6475 ± 0.0643	**0.6856** ± 0.0138
Health	0.4386 ± 0.0438	0.3189 ± 0.0289	**0.4709** ± 0.0265
Human	0.1294 ± 0.0374	0.1159 ± 0.0451	**0.1604** ± 0.0332
Image	0.5635 ± 0.0303	0.2516 ± 0.0293	**0.5756** ± 0.0230
Recreation	**0.1965** ± 0.0181	0.0509 ± 0.0184	0.1910 ± 0.0136
Science	**0.1270** ± 0.0180	0.0446 ± 0.0169	0.1262 ± 0.0214
Slashdot	0.3817 ± 0.0270	0.2158 ± 0.0241	**0.4799** ± 0.0321
Social	0.3587 ± 0.0284	0.3146 ± 0.0167	**0.4128** ± 0.0218
Yelp	0.6638 ± 0.0103	0.5673 ± 0.0132	**0.6690** ± 0.0101
Genbase	0.9522 ± 0.0158	0.4396 ± 0.0255	**0.9528** ± 0.0157
Medical	0.7757 ± 0.0273	0.4127 ± 0.0576	**0.7797** ± 0.0323
Yeast	0.6147 ± 0.0187	0.6014 ± 0.0137	**0.6239** ± 0.0156
Average	0.3915	0.2804	0.4132
Avg Rank	1.90	2.85	1.25
+ Choquet Fusion W/T/L	16/0/4	19/0/1	–
Holm p-value	**2.93e-04**	**3.81e-06**	–

Table 4
Performance comparison evaluated using the Macro-F1 (Component Ablation) metric (higher is better). Mean results ± standard deviation are reported across 10 cross-validation folds. The best value for each dataset is highlighted in bold. * indicates statistical significance according to the Wilcoxon paired test with Holm correction ($p < 0.05$).

Dataset	Embedded Only	Spectral Only	+ Choquet Fusion
Arts	0.0553 ± 0.0058	0.0219 ± 0.0043	**0.0679** ± 0.0093
Birds	0.1139 ± 0.0451	0.1131 ± 0.0445	**0.1792** ± 0.0563
Business	0.0806 ± 0.0154	0.0374 ± 0.0049	**0.0827** ± 0.0211
Computers	0.0254 ± 0.0097	0.0325 ± 0.0105	**0.0600** ± 0.0215
Corel16k1	**0.0115** ± 0.0022	0.0014 ± 0.0007	0.0113 ± 0.0022
Corel16k2	0.0116 ± 0.0017	0.0051 ± 0.0030	**0.0129** ± 0.0030
Education	0.0421 ± 0.0048	0.0119 ± 0.0109	**0.0602** ± 0.0119
Entertain	0.1316 ± 0.0158	0.0894 ± 0.0122	**0.1440** ± 0.0200
Flags	0.4853 ± 0.0563	0.5031 ± 0.0990	**0.5153** ± 0.0671
Health	0.1227 ± 0.0105	0.0712 ± 0.0175	**0.1475** ± 0.0163
Human	0.0390 ± 0.0096	0.0345 ± 0.0111	**0.0541** ± 0.0073
Image	0.5610 ± 0.0326	0.2400 ± 0.0277	**0.5740** ± 0.0258
Recreation	0.1180 ± 0.0147	0.0366 ± 0.0126	**0.1273** ± 0.0116
Science	0.0496 ± 0.0110	0.0189 ± 0.0053	**0.0511** ± 0.0125
Slashdot	0.2253 ± 0.0166	0.0882 ± 0.0103	**0.2927** ± 0.0591
Social	0.0740 ± 0.0185	0.0386 ± 0.0133	**0.0752** ± 0.0158
Yelp	0.6061 ± 0.0165	0.4375 ± 0.0314	**0.6102** ± 0.0241
Genbase	**0.5388** ± 0.0278	0.2186 ± 0.0240	**0.5388** ± 0.0278
Medical	0.2823 ± 0.0218	0.0902 ± 0.0149	**0.2843** ± 0.0243
Yeast	0.3341 ± 0.0196	0.2998 ± 0.0094	**0.3494** ± 0.0220
Average	0.1954	0.1195	0.2119
Avg Rank	2.00	2.90	1.10
+ Choquet Fusion W/T/L	18/1/1	20/0/0	–
Holm p-value	**7.75e-05**	**1.91e-06**	–

Table 5
Performance comparison evaluated using the Micro-F1 (Capacity Ablation) metric (higher is better). Mean results ± standard deviation are reported across 10 cross-validation folds. The best value for each dataset is highlighted in bold. * indicates statistical significance according to the Wilcoxon paired test with Holm correction ($p < 0.05$).

Dataset	Fixed Cap. (0.50, 0.50)	Additive Cap.	Choquet (Min/Max)	**Choquet (Rank)**
Arts	0.1618 ± 0.0130	0.1633 ± 0.0184	0.1642 ± 0.0229	**0.1703** ± 0.0045
Birds	0.3545 ± 0.0902	0.3505 ± 0.0603	0.3400 ± 0.0977	**0.3602** ± 0.0678
Business	0.6918 ± 0.0117	0.6902 ± 0.0110	0.6909 ± 0.0129	**0.6922** ± 0.0091
Computers	0.3676 ± 0.1043	0.3651 ± 0.1099	0.3395 ± 0.1173	**0.3728** ± 0.0970
Corel16k1	0.0291 ± 0.0051	0.0334 ± 0.0044	0.0316 ± 0.0099	**0.0336** ± 0.0052
Corel16k2	0.0361 ± 0.0097	0.0348 ± 0.0105	0.0336 ± 0.0077	**0.0380** ± 0.0047
Education	0.1606 ± 0.0143	**0.1742** ± 0.0143	0.1575 ± 0.0209	0.1642 ± 0.0145
Entertain	**0.3152** ± 0.0330	0.3146 ± 0.0286	0.3018 ± 0.0268	0.3048 ± 0.0312
Flags	0.6848 ± 0.0489	0.6674 ± 0.0386	0.6775 ± 0.0527	**0.6856** ± 0.0138
Health	0.4614 ± 0.0225	0.4692 ± 0.0190	0.4650 ± 0.0315	**0.4709** ± 0.0265
Human	0.1384 ± 0.0268	0.1580 ± 0.0270	0.1578 ± 0.0296	**0.1604** ± 0.0332
Image	0.5515 ± 0.0153	0.5616 ± 0.0232	0.5680 ± 0.0321	**0.5756** ± 0.0230
Recreation	0.1777 ± 0.0266	0.1902 ± 0.0270	0.1908 ± 0.0168	**0.1910** ± 0.0136
Science	0.1305 ± 0.0110	0.1299 ± 0.0181	**0.1480** ± 0.0082	0.1262 ± 0.0214
Slashdot	**0.4966** ± 0.0384	0.4877 ± 0.0343	0.4787 ± 0.0283	0.4799 ± 0.0321
Social	0.4084 ± 0.0171	0.4067 ± 0.0227	0.4071 ± 0.0263	**0.4128** ± 0.0218
Yelp	0.6473 ± 0.0117	0.6641 ± 0.0098	0.6671 ± 0.0131	**0.6690** ± 0.0101
Genbase	0.9442 ± 0.0108	0.9528 ± 0.0157	**0.9620** ± 0.0084	0.9528 ± 0.0157
Medical	0.7426 ± 0.0325	0.7699 ± 0.0272	0.7748 ± 0.0307	**0.7797** ± 0.0323
Yeast	0.6198 ± 0.0144	0.6237 ± 0.0201	0.6203 ± 0.0173	**0.6239** ± 0.0156
Average	0.4060	0.4104	0.4088	0.4132
Avg Rank	2.95	2.70	2.85	1.50
Choquet (Rank) W/T/L	17/0/3	15/1/4	18/0/2	–
Holm p-value	**1.53e-02**	**4.55e-02**	**1.25e-02**	–

Table 6
Performance comparison evaluated using the Macro-F1 (Capacity Ablation) metric (higher is better). Mean results ± standard deviation are reported across 10 cross-validation folds. The best value for each dataset is highlighted in bold. * indicates statistical significance according to the Wilcoxon paired test with Holm correction ($p < 0.05$).

Dataset	Fixed Cap. (0.50, 0.50)	Additive Cap.	Choquet (Min/Max)	**Choquet (Rank)**
Arts	0.0673 ± 0.0054	0.0641 ± 0.0070	0.0643 ± 0.0067	**0.0679** ± 0.0093
Birds	0.1753 ± 0.0723	0.1707 ± 0.0437	0.1741 ± 0.0769	**0.1792** ± 0.0563
Business	0.0780 ± 0.0170	0.0788 ± 0.0113	0.0824 ± 0.0209	**0.0827** ± 0.0211
Computers	**0.0638** ± 0.0217	0.0613 ± 0.0155	0.0547 ± 0.0193	0.0600 ± 0.0215
Corel16k1	0.0103 ± 0.0034	**0.0117** ± 0.0027	0.0110 ± 0.0022	0.0113 ± 0.0022
Corel16k2	0.0117 ± 0.0033	0.0123 ± 0.0027	0.0113 ± 0.0028	**0.0129** ± 0.0030
Education	**0.0623** ± 0.0170	0.0609 ± 0.0124	0.0568 ± 0.0115	0.0602 ± 0.0119
Entertain	0.1508 ± 0.0202	**0.1538** ± 0.0243	0.1356 ± 0.0198	0.1440 ± 0.0200
Flags	0.5052 ± 0.0789	0.4950 ± 0.0652	0.5029 ± 0.0841	**0.5153** ± 0.0671
Health	**0.1532** ± 0.0236	0.1466 ± 0.0178	0.1464 ± 0.0166	0.1475 ± 0.0163
Human	0.0376 ± 0.0061	0.0537 ± 0.0083	**0.0541** ± 0.0111	**0.0541** ± 0.0073
Image	0.5478 ± 0.0171	0.5609 ± 0.0264	0.5660 ± 0.0342	**0.5740** ± 0.0258
Recreation	**0.1299** ± 0.0313	0.1230 ± 0.0117	0.1235 ± 0.0154	0.1273 ± 0.0116
Science	0.0524 ± 0.0101	0.0504 ± 0.0096	**0.0536** ± 0.0070	0.0511 ± 0.0125
Slashdot	0.2814 ± 0.0170	0.2705 ± 0.0160	0.2636 ± 0.0225	**0.2927** ± 0.0591
Social	0.0698 ± 0.0162	0.0684 ± 0.0112	0.0721 ± 0.0183	**0.0752** ± 0.0158
Yelp	0.5571 ± 0.0179	0.6051 ± 0.0148	0.5932 ± 0.0232	**0.6102** ± 0.0241
Genbase	0.4859 ± 0.0137	0.5388 ± 0.0278	**0.5490** ± 0.0265	0.5388 ± 0.0278
Medical	0.2498 ± 0.0227	0.2786 ± 0.0236	0.2749 ± 0.0271	**0.2843** ± 0.0243
Yeast	0.3461 ± 0.0146	**0.3506** ± 0.0222	0.3490 ± 0.0191	0.3494 ± 0.0220
Average	0.2018	0.2078	0.2069	0.2119
Avg Rank	2.70	2.70	2.90	1.70
Choquet (Rank) W/T/L	14/0/6	14/1/5	17/1/2	–
Holm p-value	**1.96e-02**	**1.96e-02**	**2.53e-03**	–

(ii) Capacity Ablation and Robustness to Source Imbalance. Next, we evaluate the contribution of the Choquet Integral's fuzzy measure (μ) compared to simpler aggregation rules: an *Additive Cap.* mechanism (where the capacity is constrained to $\mu_e + \mu_s = 1$ but still optimised via the same inner CV), and a *Fixed Cap. (0.50, 0.50)* baseline (a naive arithmetic mean without inner cross-validation).

As observed in Tables 5 and 6, the data-driven **Choquet (Rank)** fusion records the best average rank on both Micro-F1 (1.50) and Macro-F1 (1.70) and is Holm-significantly superior to every alternative aggregation rule: the fixed arithmetic mean (Micro-F1 p=1.53×10^{-2}, Macro-F1 p=1.96×10^{-2}), the learned Additive capacity (Micro-F1 p=4.55×10^{-2}, Macro-F1 p=1.96×10^{-2}), and the Min–Max-normalised Choquet (Micro-F1 p=1.25×10^{-2}, Macro-F1 p=2.53×10^{-3}). The margin over the learned Additive capacity is the narrowest of the three, which is consistent with the formulation rather than a shortcoming: the Choquet integral acts as a safe, non-additive generalisation of the weighted mean, and on datasets

where the embedded and spectral channels exhibit weak non-linear interaction the learned fuzzy capacity naturally converges towards additivity, closely matching the additive baseline. What the comparison isolates clearly is the value of *learning* the fusion: both learned variants dominate the fixed arithmetic mean, while the non-additive freedom of the full Choquet capacity pays off precisely under source asymmetry (below) and under label noise (Section 4.5), where it widens the gap.

A fixed arithmetic mean forces the noisy signal to contaminate the reliable one, whereas the inner-CV Choquet integral detects the mismatch and suppresses the failing channel. On *Yelp*, the Spectral-only Macro-F1 drops to 0.4375 compared to the Embedded-only 0.6061; consequently, the Fixed Capacity fusion is dragged down to 0.5571. The Choquet integral automatically suppresses the spectral influence (μ_s=0.24), recovering 0.6102. The same rescue pattern appears on *Genbase*, where the Spectral source (0.2186) drags the fixed average to 0.4859, but the learned capacity shuts it off (μ_s=0.00) and recovers exactly the Embedded-level performance of 0.5388; and on *Medical*, where a spectral drop to 0.0902 reduces the average to 0.2498, but the capacity recovers to 0.2843, slightly above the Embedded-only score.

Importantly, in more symmetric settings where both channels are informative, the non-additive capacity matches the simpler baselines without signs of overfitting, indicating that the adaptive fusion mechanism does not introduce a performance penalty in benign regimes.

4.3. Parameter analysis: interpreting the fuzzy measure

Because the Choquet Integral learns a dataset-specific capacity μ via inner cross-validation, we can extract the interaction index $I = 1 - \mu_e - \mu_s$ (Remark 1) to characterise how the two scoring channels combine on each benchmark, with the exact behavioural semantics established by Property 1. Fig. 4 reports I averaged over 10 folds for each of the 20 datasets. Out of the 20 benchmarks, 9 exhibit $|I| \geq 0.15$, confirming that the two sources are often not well described by an additive (independent) combination. The embedded channel receives more capacity on average ($\bar{\mu}_e$=0.59 vs. $\bar{\mu}_s$=0.47), but the spectral source becomes the primary contributor on datasets such as *Human* (μ_s=0.86, μ_e=0.34), *Flags* (μ_s=0.68, μ_e=0.18), *Slashdot* (μ_s=0.68, μ_e=0.50), and *Birds* (μ_s=0.60, μ_e=0.48).

The most synergistic (agreement-seeking) interactions appear on *Arts* and *Yeast* ($I = +0.18$) and *Genbase* ($I = +0.16$), where by (11) the fusion penalises features on which the two channels disagree. The strongest disjunctive behaviour appears on *Yelp* ($I = -0.24$) and on *Image* and *Human* ($I = -0.20$), where a single high score from either channel suffices. On the "safety net" datasets discussed in Section 4.2, the learned allocation of capacity is clearly visible: on *Yelp* the spectral capacity is strongly reduced (μ_s=0.24, μ_e=1.00), and on *Genbase* it is entirely suppressed (μ_s=0.00, μ_e=0.84). These patterns cannot be expressed by a fixed additive aggregation, which treats every dataset identically. The error bars in Fig. 4 also reveal a non-trivial fold-to-fold variability of the selected pair (μ_e^*, μ_s^*). This is an expected consequence of the geometry of the selection problem: the inner-CV criterion is piecewise constant in the capacity (Section 3.3), and near the optimum several neighbouring capacities induce identical or nearly identical top-p feature subsets, so the arg max may legitimately land on any of them in different folds. What matters is the stability of the *outcome* rather than of the selected coordinates, and the outer-fold standard deviations reported in the main tables confirm that the resulting rankings perform consistently.

4.4. Sensitivity analysis

Inner-CV selection criterion. We first assess the choice of the GM criterion empirically, re-running the full pipeline with single-metric inner criteria. Averaged over the 20 benchmarks at p=20%, optimising Micro-F1 alone yields Micro-F1 0.4103 but Macro-F1 0.2065; optimising Macro-F1 alone yields Macro-F1 0.2099 but Micro-F1 0.4067; the GM criterion attains 0.4132/0.2119, matching or improving on each single-metric criterion on its own favoured metric while avoiding the weakness each shows on the other. This confirms the intended behaviour discussed in Section 3.3: the GM acts as a balanced compromise rather than a third objective with its own bias.

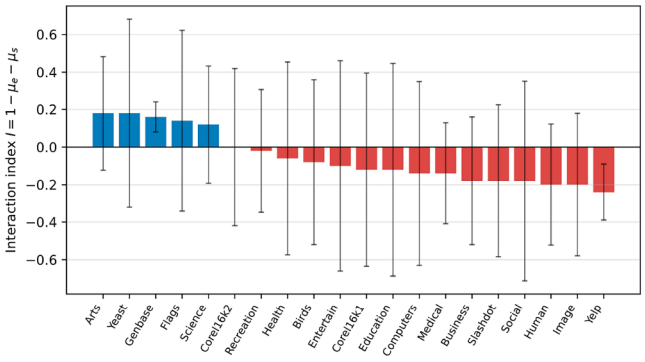

Fig. 4. Interaction index $I = 1 - \mu_e - \mu_s$ averaged over 10 folds for each benchmark. Blue bars ($I \geq 0$) indicate synergy (agreement-seeking fusion); red bars ($I < 0$) indicate redundancy (disjunctive fusion). Error bars show the standard deviation across folds.

Exposed inner-CV hyper-parameters. We evaluate the stability of CFIFS across its two primary user-exposed hyper-parameters: the neighbourhood size k of the ML-kNN classifier used inside the inner CV to score capacity candidates, and the validation scope p_{frac} (the percentage of top-ranked features used by the internal cross-validation to probe the Choquet capacities). Fig. 5 visualises the performance response across these two dimensions. The leftmost column reports the global average over all 20 datasets, while the remaining columns isolate three critical benchmarks (*Yelp*, *Genbase*, and *Medical*) where we previously observed strong asymmetries between the embedded and spectral channels.

Regarding the validation scope p_{frac} (top row), the global average response has a broad, shallow optimum: Micro-F1 stays within 0.002 of the default for $p_{\text{frac}} \in [0.15, 0.30]$, and only the extreme settings lose ground (global Micro-F1 0.397 at p_{frac}=0.05 and 0.401 at 0.50, vs 0.410 at the default 0.20). This indicates that the internal validation module does not require a precisely tuned feature budget to identify the best fusion strategy: evaluating the capacity over any reasonable probing subset (the top 15% to 30%) reliably exposes the underlying synergy or redundancy between the scoring sources, while probing only a sliver of the ranking (5%) or half of it (50%) makes the selection signal less discriminative.

The neighbourhood size k of the inner-CV ML-kNN (bottom row) shows an even flatter response: global Micro-F1 varies by less than 0.006 over $k \in [3, 30]$, with a mild decay only at the largest values ($k \geq 30$), where the validation classifier oversmooths rare labels and the capacity-selection signal loses resolution. The default k=10 sits at the top of the curve, and any choice in $[5, 25]$ is statistically equivalent.

Fixed hyper-parameters. We additionally swept, one at a time, every fixed hyper-parameter of the two scoring stages and of the fusion grid, re-running the full pipeline on all 20 datasets for each value (Fig. 6). The framework is remarkably insensitive to all of them. The HSIC–smoothness blend α_S moves global Micro-F1 by at most 0.006 over its entire range $[0, 1]$, with a shallow optimum around 0.6–0.8 that contains the default 0.70. The regression–logistic blend α is flat for $\alpha \in [0, 0.7]$ (within 0.005); only the pure-logistic extreme α=1 degrades appreciably (Micro-F1 0.390), confirming that the embedded reconstruction term carries useful signal. The group-sparsity strength β_{sp} is flat across two orders of magnitude (10^{-3} to 10^{-1}, within 0.003), as the downstream ranking depends on the ordering of the row norms rather than on their absolute scale. Finally, the capacity grid resolution confirms the choice

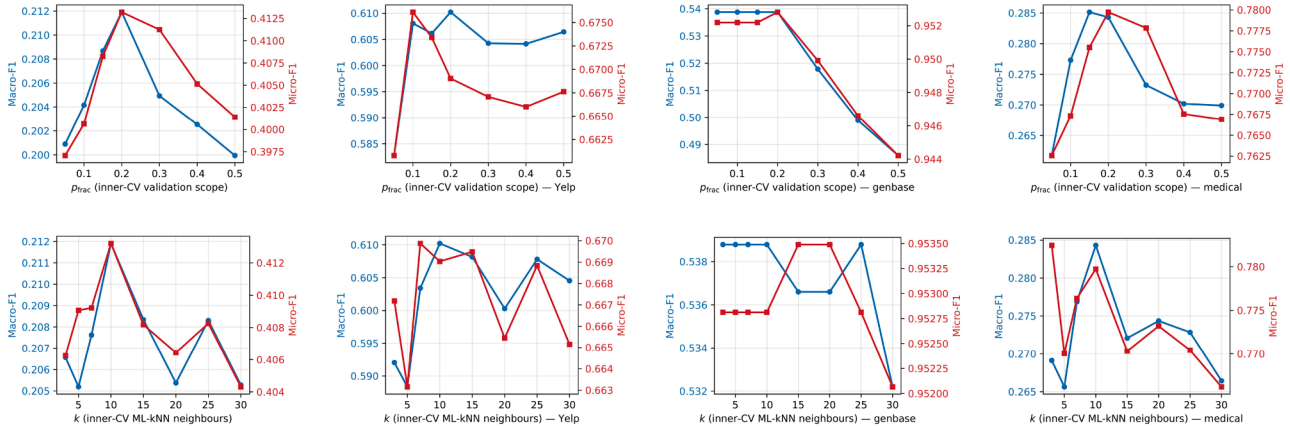

Fig. 5. Sensitivity analysis of CFIFS for its exposed inner-CV hyper-parameters: validation scope p_{frac} (top row) and ML-kNN neighbourhood size k (bottom row); Micro-F1 (red squares, right axis) and Macro-F1 (blue circles, left axis) at the p=20% operating point. Columns: global average across all 20 datasets (left), followed by Yelp, Genbase, and Medical. The average curves show broad plateaus around the defaults (p_{frac}=0.20, k=10), whereas datasets with source asymmetry exhibit wider variation, confirming the relevance of the capacity-learning mechanism in these regimes.

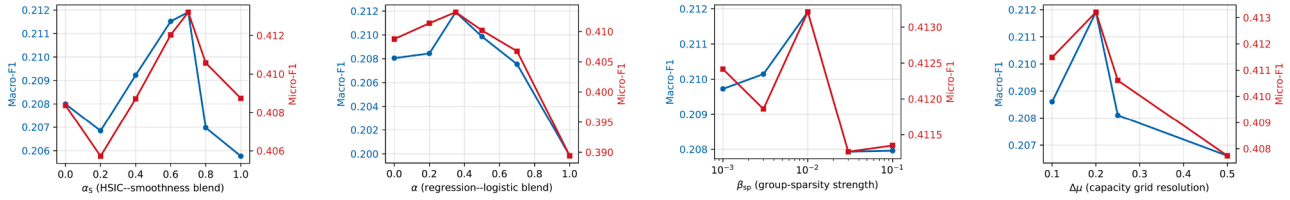

Fig. 6. One-parameter-at-a-time sensitivity of the fixed CFIFS hyper-parameters (global average over the 20 benchmarks at p=20%; Micro-F1 in red squares, right axis; Macro-F1 in blue circles, left axis). From left to right: HSIC–smoothness blend α_S, regression–logistic blend α, group-sparsity strength β_{sp} (log scale), and capacity grid resolution $\Delta\mu$. All curves exhibit broad plateaus containing the defaults (α_S=0.70, α=0.35, β_{sp}=10^{-2}, $\Delta\mu$=0.20).

discussed in Section 4.1: refining $\Delta\mu$ from 0.20 to 0.10 changes global Micro-F1 by +0.0015 (at triple the inner-CV cost), while coarsening to 0.50 loses 0.002.

Finally, while the inner-CV capacities could theoretically be aggregated across candidates via either hard-max selection or softmax-weighted averaging, preliminary tests showed that softening the selection ($\tau = 0.02$) provides only negligible stabilising regularisation over hard routing. Consequently, we fix the routing mechanism and do not expose it as a tunable dimension.

Taken together, these results show that the *values* of CFIFS's fixed hyper-parameters need not be carefully tuned: each sits on a broad plateau around its default, and the two exposed inner-CV parameters tolerate wide variation. This insensitivity should not be mistaken for irrelevance of the underlying components. On the contrary, the ablation of Section 4.2 shows that removing the spectral channel or replacing the learned capacity with a fixed average degrades performance significantly, and the plateaus reported here are *bounded*: performance does drop at the extremes, as the pure-logistic case α=1—which discards the embedded reconstruction term—makes explicit. In other words, each component contributes, but the precise value of its controlling parameter is not critical. What the sweep establishes is therefore robustness rather than redundancy: a single default configuration generalises across 20 diverse datasets, so practitioners can deploy CFIFS without costly per-dataset grid searches.

4.5. Robustness to contaminated features and label noise

The ablation study showed that the learned capacity acts as a safety net when one channel is structurally weak on a dataset. A natural follow-up question is whether this adaptive mechanism remains reliable when the training data themselves are corrupted—a scenario in which the inner-CV signal that drives capacity selection is also degraded. We therefore stress-test CFIFS on six representative benchmarks (*Birds*, *Science*, *Yeast*, *Yelp*, *Genbase*, *Medical*; 10 folds each) under two controlled perturbations: (i) *feature contamination*, in which a fraction (20% or 40%) of feature columns is independently permuted across instances in both the training and the test fold, preserving each feature's marginal distribution while destroying its association with the labels; and (ii) *training-label noise*, in which each training-label bit is flipped independently with probability 10% or 20%, while test labels remain clean. Table 7 compares CFIFS against its non-adaptive ablations—the fixed-capacity fusion and the two single channels—all derived from the same channel scores.

Three observations emerge. First, under feature contamination all variants degrade gracefully and remain within a narrow band: the rank-normalised scores are intrinsically robust to features whose scores collapse to noise levels, so the fusion rule matters relatively little in this regime. Second, under label noise — which attacks both scoring stages *and* the inner-CV selection signal simultaneously — the value of the learned capacity becomes unambiguous: at a 20% flip rate CFIFS retains substantially more performance (Micro-F1 0.485, Macro-F1 0.306) than the fixed arithmetic mean (0.441/0.240) and than either single channel ($\lesssim$ 0.395/0.270). In relative terms, CFIFS loses 16% of its clean Micro-F1 at the harshest noise level, against 23% for the fixed mean and roughly 30% for the single channels. Third, and most tellingly, the adaptive mechanism rescues *opposite* failure modes on different datasets: on *Genbase* label noise destroys the spectral channel (Macro-F1 0.33, since the Jaccard label graph is rebuilt from corrupted labels), and CFIFS suppresses it, retaining 0.676, on par with the unaffected embedded channel; on *Medical* it is the *embedded* channel that collapses (0.065), and CFIFS leans on the spectral source, retaining 0.264—four times the embedded-only score. No fixed fusion rule can produce both behaviours, since they require shifting capacity in opposite directions. Crucially, this adaptation is visible directly in the *learned capacity* and not only in the downstream scores: across the thirty clean-and-contaminated conditions of Table 7, the capacity assigns the larger mass to whichever channel has

Table 7
Robustness under contaminated features and training-label noise (Micro-F1/Macro-F1 at p=20%, averaged over *Birds*, *Science*, *Yeast*, *Yelp*, *Genbase*, and *Medical*, 10 folds each). Feature contamination permutes a fraction of feature columns across instances (train and test); label noise flips each training-label bit with the given probability, while test labels remain clean. All variants share the two scoring stages; capacity-learning variants use the leakage-free inner CV.

	Clean		Feat. 20%		Feat. 40%		Labels 10%		Labels 20%	
Variant	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
CFIFS (learned capacity)	**0.5812**	**0.4030**	0.5525	**0.3777**	0.5248	**0.3458**	**0.5525**	**0.3661**	**0.4854**	**0.3062**
Fixed (0.50, 0.50)	0.5737	0.3760	**0.5541**	0.3621	**0.5310**	0.3374	0.5250	0.3137	0.4406	0.2400
Embedded only	0.5632	0.3904	0.5439	0.3749	0.5108	0.3402	0.4509	0.3212	0.3946	0.2695
Spectral only	0.5588	0.3593	0.5391	0.3480	0.5127	0.3199	0.4803	0.2606	0.3920	0.2026

the stronger single-channel Macro-F1 in 29 of the 30 cases—the sole exception being a near-exact tie between the two channels—and the capacity gap $\mu_e - \mu_s$ correlates at ρ=0.52 with the channel-quality gap. The most pointed case is *Medical* at 20% label noise, the only condition in the whole study in which the embedded channel becomes the weaker of the two: there the capacity reverses accordingly (μ_e=0.34 < μ_s=0.44). The inner-CV selection thus downweights whichever channel is genuinely less informative, even when the very signal that drives the selection is itself degraded by the corruption.

4.6. Main comparison

Having described the design choices and the non-additive fusion mechanism, we now report the unified comparison of CFIFS against the seven state-of-the-art baselines.

Tables 8 to 10 report the main performance comparison extracting p=20% features (reporting the mean ± standard deviation over 10 cross-validation folds). To rigorously assess the statistical significance of the results, we report the paired Wilcoxon signed-rank test comparing each baseline against CFIFS, corrected for multiple comparisons using the conservative Holm step-down procedure.

As shown in Table 8, on Micro-F1 CFIFS achieves the best average score (0.4132) and the best average rank (1.55). Moreover, in terms of direct comparisons, CFIFS wins on 17–20 out of 20 datasets against every baseline with Holm-corrected p-values all below 2.0×10^{-4}. The closest competitors on this metric are RFSFS (avg 0.3784, rank 4.00), SRFS (avg 0.3773, rank 4.00) and GRRO (avg 0.3773, rank 3.73), but even against these CFIFS wins on 17/20, 17/20 and 19/20 datasets, respectively. The gains are most pronounced on medium-to-high-dimensional text classification tasks: on *Arts* (0.1703 vs 0.1280 for GRRO), *Education* (0.1642 vs 0.0991 for LRDG), *Recreation* (0.1910 vs 0.1470 for LRMFS), and *Slashdot* (0.4799 vs 0.4744 for SRFS), CFIFS outperforms all baselines on these datasets. Similarly, on the vision-domain benchmark *Image* (0.5756 vs 0.5552 for LRMFS), the Choquet fusion appears to retain complementary features that single-source methods discard. Conversely, on the *Corel16k* image-descriptor datasets, SRFS (0.0340) and RFSFS (0.0398) narrowly outperform CFIFS (0.0336 and 0.0380), suggesting that for spatially structured descriptors, pseudo-label regularisation may better preserve spatial clusters than rank-based fusion. Methods relying on graph-manifold learning without a tailored label-affinity structure (LSMFS, LRDG, SCNMF) show marked performance drops on high-dimensional and imbalanced datasets—SCNMF drops to near-zero performance on *Genbase* and *Medical*—whereas CFIFS remains comparatively stable.

The exact 0.0000±0.0000 of SCNMF on *Genbase* has a concrete and instructive explanation. *Genbase* is extremely sparse: on each training fold roughly 1,090 of its 1,185 feature columns are identically zero, leaving only ∼95 informative features. With its default hyper-parameters SCNMF ranks the constant all-zero columns *above* the informative ones, so the top-20% subset (237 features) consists entirely of all-zero columns on every fold; the downstream ML-kNN then receives a degenerate (all-zero) representation and predicts no positive label, yielding identically zero F1 with no variance. This is a ranking failure on a pathologically sparse dataset, not a property of the metric or of the evaluation protocol, and it motivates checking whether tuning removes it. We therefore verified explicitly whether these collapses are an artefact of untuned hyper-parameters rather than an inherent property of the algorithms.

For the two collapsing methods we performed an *oracle* per-dataset sweep on *Genbase* and *Medical*: for SCNMF, the 5×5 grid published by its own authors ($\lambda_1 \in \{10^{-4}, \dots, 5{\cdot}10^{-2}\}$, $\lambda_2 \in \{10^{-2}, \dots, 10^{2}\}$); for LSMFS, $\alpha, \beta \in \{10^{-3}, \dots, 10\}$ with the remaining parameters at their defaults; in both cases the configuration was selected *a posteriori* on test performance, an upper bound no validated tuning procedure can exceed. The outcome is twofold. On the one hand, the collapses are indeed a configuration artefact: oracle-tuned SCNMF reaches Micro-F1 0.9523 on *Genbase* (from 0.0000 with defaults) and 0.7679 on *Medical* (from 0.0205); oracle-tuned LSMFS behaves similarly. On the other hand, even against these oracle upper bounds, CFIFS—with a single configuration shared by all 20 datasets and a training-only adaptive component—matches or exceeds the tuned baselines on the same datasets (Micro-F1 0.9528 on *Genbase*, 0.7797 on *Medical*; Macro-F1 0.2843 vs 0.2690 for the best oracle on *Medical*). The practical reading is that some baselines require careful per-dataset tuning—for which their original papers prescribe no validated procedure—to avoid catastrophic configurations, whereas the data-driven component of CFIFS achieves the same adaptation automatically and without touching validation or test labels. This robustness is consistent with the late-fusion architecture and the Choquet capacity's ability to reduce the influence of a weak source when its contribution is not supported by inner-validation performance.

Macro-F1, Table 9, reveals failure modes that Micro-F1 often masks. CFIFS leads with an average of 0.2119 and rank 1.80; the runner-up is RFSFS (0.1933, rank 3.40), and all Holm-corrected p-values remain below 4.8×10^{-3}. On *Genbase*, SCNMF collapses entirely (0.0) with its default configuration (see the oracle-tuning analysis above) and LSMFS and LRDG score below 0.09; by contrast, CFIFS (0.5388) almost matches the best stable baseline (SRFS, 0.5413). On *Birds*, CFIFS (0.1792) outperforms the second-best GRRO (0.1395) by a +28% relative gain, highlighting the value of fusing the spectral channel for small datasets with many rare labels. The W/T/L profiles remain clearly favourable (e.g., 17/0/3 vs RFSFS), and no baseline approaches the overall generalisation capability of CFIFS.

The Hamming Loss results (Table 10) confirm that these discriminative gains do not come at the cost of higher absolute error rates: CFIFS achieves the lowest average (0.0756) and the best rank (2.00), with Holm-corrected p-values ranging from 9.4×10^{-5} to 1.92×10^{-2}, all significant at the α=0.05 level. This consistency across all three metrics—a frequency-weighted measure, a label-balanced measure, and an error-rate measure—rules out the possibility that CFIFS's advantage is merely metric-specific. Across the three metrics, CFIFS achieves average ranks of 1.55, 1.80, and 2.00, while no baseline falls below rank 3.40 on any metric; all 21 Holm-corrected p-values (7 baselines × 3 metrics) lie below the 0.02 threshold.

The analysis at p=20% raises a natural question: does CFIFS's advantage depend on that particular operating point, or does it generalise across the entire feature-budget spectrum? Fig. 9 plots the global average across all 20 datasets for each metric as a function of p, while Figs. 7 and 8 show per-dataset trajectories.

Table 8
Performance comparison evaluated using the Micro-F1 metric (higher is better). Mean results ± standard deviation are reported across 10 cross-validation folds. The best value for each dataset is highlighted in bold. * indicates statistical significance according to the Wilcoxon paired test with Holm correction ($p < 0.05$).

Dataset	GRRO	SRFS	RFSFS	LRMFS	LSMFS	LRDG	SCNMF	**CFIFS**
Arts	0.1280 ± 0.0231	0.1121 ± 0.0227	0.1129 ± 0.0243	0.1072 ± 0.0137	0.0317 ± 0.0113	0.0414 ± 0.0111	0.0650 ± 0.0090	**0.1703** ± 0.0045
Birds	0.2795 ± 0.0745	0.2340 ± 0.0760	0.2420 ± 0.0478	0.2811 ± 0.0559	0.2417 ± 0.0567	0.2309 ± 0.0644	0.1646 ± 0.0373	**0.3602** ± 0.0678
Business	0.6868 ± 0.0115	0.6824 ± 0.0103	0.6805 ± 0.0092	0.6816 ± 0.0072	0.6733 ± 0.0084	0.6803 ± 0.0106	0.6817 ± 0.0085	**0.6922** ± 0.0091
Computers	0.3406 ± 0.1004	0.3348 ± 0.0955	0.3346 ± 0.0964	0.3535 ± 0.0912	0.3352 ± 0.1651	0.3291 ± 0.1239	0.3060 ± 0.1713	**0.3728** ± 0.0970
Corel16k1	0.0215 ± 0.0064	**0.0340** ± 0.0054	0.0339 ± 0.0057	0.0154 ± 0.0047	0.0195 ± 0.0053	0.0135 ± 0.0040	0.0215 ± 0.0050	0.0336 ± 0.0052
Corel16k2	0.0248 ± 0.0073	0.0382 ± 0.0068	**0.0398** ± 0.0078	0.0259 ± 0.0081	0.0229 ± 0.0054	0.0164 ± 0.0051	0.0276 ± 0.0064	0.0380 ± 0.0047
Education	0.0743 ± 0.0070	0.0871 ± 0.0139	0.0898 ± 0.0086	0.0956 ± 0.0066	0.0354 ± 0.0062	0.0991 ± 0.0317	0.0816 ± 0.0143	**0.1642** ± 0.0145
Entertain	0.2909 ± 0.0183	0.2863 ± 0.0235	0.2872 ± 0.0197	0.2991 ± 0.0231	0.1420 ± 0.0363	0.2761 ± 0.0232	0.1321 ± 0.0217	**0.3048** ± 0.0312
Flags	0.6883 ± 0.0344	**0.7026** ± 0.0316	**0.7026** ± 0.0316	0.7006 ± 0.0589	0.6951 ± 0.0380	0.6364 ± 0.0593	0.6962 ± 0.0242	0.6856 ± 0.0138
Health	0.4290 ± 0.0203	0.4064 ± 0.0318	0.4070 ± 0.0344	0.4012 ± 0.0270	0.3733 ± 0.0364	0.3314 ± 0.0367	0.4570 ± 0.0211	**0.4709** ± 0.0265
Human	0.1461 ± 0.0235	0.1218 ± 0.0371	0.1264 ± 0.0337	0.1460 ± 0.0333	0.1201 ± 0.0348	0.1039 ± 0.0367	0.1185 ± 0.0203	**0.1604** ± 0.0332
Image	0.5099 ± 0.0332	0.4401 ± 0.0206	0.4613 ± 0.0241	0.5552 ± 0.0359	0.5221 ± 0.0309	0.5347 ± 0.0293	0.4968 ± 0.0372	**0.5756** ± 0.0230
Recreation	0.1455 ± 0.0222	0.1402 ± 0.0167	0.1372 ± 0.0207	0.1470 ± 0.0250	0.0394 ± 0.0147	0.1405 ± 0.0245	0.0757 ± 0.0172	**0.1910** ± 0.0136
Science	0.0873 ± 0.0198	0.0818 ± 0.0162	0.0830 ± 0.0156	0.0812 ± 0.0173	0.0272 ± 0.0103	0.0766 ± 0.0103	0.0637 ± 0.0159	**0.1262** ± 0.0214
Slashdot	0.3884 ± 0.0349	0.4744 ± 0.0150	0.4736 ± 0.0174	0.3276 ± 0.0790	0.1820 ± 0.0395	0.1070 ± 0.0191	0.1483 ± 0.0255	**0.4799** ± 0.0321
Social	0.3885 ± 0.0132	0.3916 ± 0.0212	0.3821 ± 0.0268	0.3712 ± 0.0206	0.2439 ± 0.0260	0.3075 ± 0.0287	0.2432 ± 0.0342	**0.4128** ± 0.0218
Yelp	0.6429 ± 0.0060	0.6319 ± 0.0094	0.6343 ± 0.0098	0.6661 ± 0.0136	0.6378 ± 0.0449	0.6558 ± 0.0102	0.6198 ± 0.0110	**0.6690** ± 0.0101
Genbase	0.9523 ± 0.0156	0.9523 ± 0.0156	0.9523 ± 0.0156	0.9523 ± 0.0156	0.2041 ± 0.2532	0.0334 ± 0.0511	0.0000 ± 0.0000	**0.9528** ± 0.0157
Medical	0.7059 ± 0.0518	0.7779 ± 0.0286	0.7724 ± 0.0281	0.7206 ± 0.0322	0.1172 ± 0.0192	0.0907 ± 0.0469	0.0205 ± 0.0198	**0.7797** ± 0.0323
Yeast	0.6159 ± 0.0132	0.6157 ± 0.0149	0.6146 ± 0.0186	0.6084 ± 0.0117	0.5903 ± 0.0131	0.6125 ± 0.0216	**0.6274** ± 0.0150	0.6239 ± 0.0156
Average	0.3773	0.3773	0.3784	0.3768	0.2627	0.2659	0.2524	0.4132
Avg Rank	3.73	4.00	4.00	3.92	6.45	6.30	6.05	1.55
CFIFS W/T/L	19/0/1	17/0/3	17/0/3	19/0/1	19/0/1	20/0/0	18/0/2	–
Holm p-value	**1.43e-05**	**1.05e-04**	**1.05e-04**	**5.44e-05**	**1.14e-05**	**6.68e-06**	**3.81e-05**	–

Table 9
Performance comparison evaluated using the Macro-F1 metric (higher is better). Mean results ± standard deviation are reported across 10 cross-validation folds. The best value for each dataset is highlighted in bold. * indicates statistical significance according to the Wilcoxon paired test with Holm correction ($p < 0.05$).

Dataset	GRRO	SRFS	RFSFS	LRMFS	LSMFS	LRDG	SCNMF	**CFIFS**
Arts	0.0510 ± 0.0087	0.0499 ± 0.0101	0.0498 ± 0.0116	0.0468 ± 0.0081	0.0146 ± 0.0066	0.0199 ± 0.0041	0.0316 ± 0.0049	**0.0679** ± 0.0093
Birds	0.1395 ± 0.0438	0.1323 ± 0.0625	0.1326 ± 0.0298	0.1380 ± 0.0419	0.1017 ± 0.0368	0.1103 ± 0.0413	0.0879 ± 0.0212	**0.1792** ± 0.0563
Business	0.0702 ± 0.0133	0.0626 ± 0.0115	0.0573 ± 0.0108	0.0551 ± 0.0071	0.0368 ± 0.0042	0.0501 ± 0.0078	0.0583 ± 0.0119	**0.0827** ± 0.0211
Computers	0.0577 ± 0.0213	0.0704 ± 0.0197	**0.0705** ± 0.0208	0.0631 ± 0.0163	0.0179 ± 0.0085	0.0571 ± 0.0236	0.0254 ± 0.0102	0.0600 ± 0.0215
Corel16k1	0.0071 ± 0.0028	0.0105 ± 0.0022	0.0106 ± 0.0021	0.0087 ± 0.0034	0.0086 ± 0.0024	0.0064 ± 0.0022	0.0077 ± 0.0016	**0.0113** ± 0.0022
Corel16k2	0.0094 ± 0.0028	0.0117 ± 0.0022	0.0121 ± 0.0027	0.0103 ± 0.0027	0.0103 ± 0.0035	0.0058 ± 0.0026	0.0120 ± 0.0029	**0.0129** ± 0.0030
Education	0.0515 ± 0.0065	0.0509 ± 0.0058	0.0510 ± 0.0070	0.0467 ± 0.0056	0.0193 ± 0.0081	0.0466 ± 0.0146	0.0274 ± 0.0078	**0.0602** ± 0.0119
Entertain	0.1392 ± 0.0149	0.1421 ± 0.0220	0.1425 ± 0.0183	**0.1521** ± 0.0192	0.0665 ± 0.0195	0.1347 ± 0.0154	0.0750 ± 0.0133	0.1440 ± 0.0200
Flags	**0.5415** ± 0.0644	0.5343 ± 0.0342	0.5343 ± 0.0342	0.5307 ± 0.0762	0.5208 ± 0.0372	0.4391 ± 0.0788	0.5339 ± 0.0363	0.5153 ± 0.0671
Health	**0.1549** ± 0.0229	0.1299 ± 0.0239	0.1354 ± 0.0246	0.1252 ± 0.0180	0.0694 ± 0.0110	0.1028 ± 0.0174	0.1373 ± 0.0195	0.1475 ± 0.0163
Human	0.0493 ± 0.0134	0.0331 ± 0.0098	0.0337 ± 0.0108	0.0403 ± 0.0126	0.0297 ± 0.0090	0.0268 ± 0.0098	0.0321 ± 0.0073	**0.0541** ± 0.0073
Image	0.5018 ± 0.0351	0.4075 ± 0.0227	0.4414 ± 0.0251	0.5521 ± 0.0375	0.5150 ± 0.0374	0.5290 ± 0.0311	0.4845 ± 0.0407	**0.5740** ± 0.0258
Recreation	0.1180 ± 0.0179	0.1195 ± 0.0174	0.1163 ± 0.0229	0.1118 ± 0.0205	0.0252 ± 0.0100	0.1006 ± 0.0176	0.0473 ± 0.0115	**0.1273** ± 0.0116
Science	0.0300 ± 0.0079	0.0312 ± 0.0086	0.0356 ± 0.0098	0.0256 ± 0.0071	0.0109 ± 0.0039	0.0315 ± 0.0049	0.0290 ± 0.0103	**0.0511** ± 0.0125
Slashdot	0.2015 ± 0.0203	0.2661 ± 0.0078	0.2632 ± 0.0109	0.1681 ± 0.0478	0.0816 ± 0.0222	0.0748 ± 0.0187	0.0765 ± 0.0192	**0.2927** ± 0.0591
Social	0.0662 ± 0.0123	0.0582 ± 0.0125	0.0576 ± 0.0150	0.0550 ± 0.0145	0.0289 ± 0.0103	0.0348 ± 0.0067	0.0394 ± 0.0092	**0.0752** ± 0.0158
Yelp	0.5622 ± 0.0141	0.5686 ± 0.0167	0.5736 ± 0.0131	0.6044 ± 0.0182	0.5614 ± 0.0472	0.5886 ± 0.0212	0.5508 ± 0.0095	**0.6102** ± 0.0241
Genbase	**0.5413** ± 0.0262	**0.5413** ± 0.0262	**0.5413** ± 0.0262	**0.5413** ± 0.0262	0.0893 ± 0.1525	0.0111 ± 0.0170	0.0000 ± 0.0000	0.5388 ± 0.0278
Medical	0.2313 ± 0.0336	0.2775 ± 0.0219	0.2717 ± 0.0254	0.2374 ± 0.0197	0.0336 ± 0.0180	0.0269 ± 0.0224	0.0056 ± 0.0058	**0.2843** ± 0.0243
Yeast	0.3397 ± 0.0181	0.3341 ± 0.0122	0.3364 ± 0.0194	0.3180 ± 0.0099	0.2885 ± 0.0134	0.3287 ± 0.0253	**0.3515** ± 0.0153	0.3494 ± 0.0220
Average	0.1932	0.1916	0.1933	0.1915	0.1265	0.1363	0.1307	0.2119
Avg Rank	3.58	3.70	3.40	4.33	6.95	6.35	5.90	1.80
CFIFS W/T/L	17/0/3	17/0/3	17/0/3	16/0/4	19/0/1	20/0/0	18/0/2	–
Holm p-value	**2.15e-03**	**2.15e-03**	**2.15e-03**	**1.42e-03**	**2.86e-05**	**6.68e-06**	**9.06e-05**	–

In the tight-budget regime ($p \leq 15\%$), which is the most challenging and practically relevant scenario for high-dimensional data, CFIFS opens a clear gap over all baselines already at p=5% and maintains it throughout. This indicates that the fused ranking reliably places the most informative features at the very top of the list — consistent with of the Choquet capacity's ability to exploit synergistic signals from both the embedded and spectral channels. The advantage is especially visible on high-dimensional benchmarks where random or single-source orderings waste the scarce budget: on *Arts*, *Education*, *Recreation*, and *Science*, CFIFS at p=5% already matches or exceeds the p=20% performance of several baselines.

As the budget grows into the moderate range (15%–35%), most methods have entered their plateau, yet the ranking among them shows no crossover: on *Medical*, *Slashdot*, and *Social*, CFIFS remains ahead throughout the range, confirming that its advantage is not an artefact of a single favourable budget. Some baselines exhibit instability in this regime—LSMFS and SCNMF show non-monotonic oscillations on *Health* and *Computers*—whereas CFIFS's trajectory remains comparatively smooth. This suggests that the capacity-optimised fusion produces a structurally coherent ranking rather than a noise-sensitive reordering. Beyond p=35%, the feature subset becomes progressively less selective and all methods tend to converge toward the performance of the full fea-

Table 10
Performance comparison evaluated using the Hamming Loss metric (higher is better). Mean results ± standard deviation are reported across 10 cross-validation folds. The best value for each dataset is highlighted in bold. * indicates statistical significance according to the Wilcoxon paired test with Holm correction ($p < 0.05$).

Dataset	GRRO	SRFS	RFSFS	LRMFS	LSMFS	LRDG	SCNMF	**CFIFS**
Arts	0.0613 ± 0.0014	0.0617 ± 0.0017	0.0617 ± 0.0018	0.0619 ± 0.0017	0.0634 ± 0.0016	0.0630 ± 0.0016	0.0626 ± 0.0016	**0.0609** ± 0.0017
Birds	0.0486 ± 0.0043	0.0512 ± 0.0039	0.0509 ± 0.0025	0.0491 ± 0.0035	0.0491 ± 0.0033	0.0492 ± 0.0042	0.0513 ± 0.0033	**0.0467** ± 0.0049
Business	**0.0278** ± 0.0009	0.0283 ± 0.0010	0.0283 ± 0.0008	0.0283 ± 0.0007	0.0287 ± 0.0009	0.0280 ± 0.0009	0.0280 ± 0.0007	0.0279 ± 0.0010
Computers	**0.0411** ± 0.0019	0.0414 ± 0.0017	0.0415 ± 0.0016	0.0413 ± 0.0018	0.0440 ± 0.0019	0.0420 ± 0.0023	0.0441 ± 0.0011	0.0424 ± 0.0021
Corel16k1	0.0187 ± 0.0002	**0.0187** ± 0.0001	0.0187 ± 0.0001	0.0187 ± 0.0002	0.0187 ± 0.0002	0.0187 ± 0.0002	0.0187 ± 0.0002	0.0187 ± 0.0002
Corel16k2	0.0176 ± 0.0001	0.0176 ± 0.0001	0.0176 ± 0.0001	0.0176 ± 0.0002	0.0176 ± 0.0001	0.0176 ± 0.0001	0.0176 ± 0.0001	**0.0176** ± 0.0001
Education	0.0431 ± 0.0008	0.0428 ± 0.0010	0.0429 ± 0.0010	0.0426 ± 0.0009	0.0435 ± 0.0009	0.0426 ± 0.0009	0.0429 ± 0.0008	**0.0419** ± 0.0012
Entertain	0.0586 ± 0.0017	0.0587 ± 0.0018	0.0588 ± 0.0017	**0.0581** ± 0.0017	0.0629 ± 0.0020	0.0589 ± 0.0021	0.0647 ± 0.0020	0.0582 ± 0.0022
Flags	0.2970 ± 0.0352	0.2961 ± 0.0271	0.2961 ± 0.0271	**0.2918** ± 0.0544	0.2977 ± 0.0361	0.3333 ± 0.0368	0.3159 ± 0.0249	0.3153 ± 0.0329
Health	0.0427 ± 0.0013	0.0449 ± 0.0017	0.0448 ± 0.0015	0.0443 ± 0.0015	0.0476 ± 0.0013	0.0471 ± 0.0011	0.0441 ± 0.0016	**0.0419** ± 0.0010
Human	0.0824 ± 0.0009	0.0831 ± 0.0010	0.0831 ± 0.0010	0.0833 ± 0.0013	**0.0822** ± 0.0013	0.0841 ± 0.0008	0.0832 ± 0.0017	0.0827 ± 0.0014
Image	0.1922 ± 0.0094	0.2042 ± 0.0067	0.1990 ± 0.0067	0.1784 ± 0.0079	0.1847 ± 0.0109	0.1848 ± 0.0125	0.1917 ± 0.0097	**0.1726** ± 0.0075
Recreation	0.0617 ± 0.0026	0.0621 ± 0.0024	0.0621 ± 0.0024	0.0620 ± 0.0025	0.0646 ± 0.0020	0.0621 ± 0.0023	0.0635 ± 0.0025	**0.0609** ± 0.0021
Science	0.0358 ± 0.0006	0.0358 ± 0.0007	0.0358 ± 0.0007	0.0359 ± 0.0005	0.0364 ± 0.0006	0.0359 ± 0.0007	0.0360 ± 0.0006	**0.0356** ± 0.0008
Slashdot	0.0453 ± 0.0017	0.0428 ± 0.0014	0.0425 ± 0.0014	0.0477 ± 0.0030	0.0509 ± 0.0013	0.0527 ± 0.0011	0.0525 ± 0.0012	**0.0420** ± 0.0016
Social	0.0274 ± 0.0008	0.0274 ± 0.0005	0.0275 ± 0.0007	0.0277 ± 0.0007	0.0300 ± 0.0007	0.0291 ± 0.0007	0.0295 ± 0.0010	**0.0265** ± 0.0009
Yelp	0.2090 ± 0.0045	0.2139 ± 0.0050	0.2126 ± 0.0046	**0.2006** ± 0.0052	0.2129 ± 0.0183	0.2043 ± 0.0044	0.2197 ± 0.0033	0.2016 ± 0.0051
Genbase	0.0043 ± 0.0014	0.0043 ± 0.0014	0.0043 ± 0.0014	0.0043 ± 0.0014	0.0400 ± 0.0121	0.0456 ± 0.0016	0.0464 ± 0.0013	**0.0042** ± 0.0014
Medical	0.0142 ± 0.0022	0.0112 ± 0.0017	0.0114 ± 0.0014	0.0135 ± 0.0016	0.0271 ± 0.0013	0.0273 ± 0.0012	0.0279 ± 0.0013	**0.0110** ± 0.0017
Yeast	0.2040 ± 0.0039	0.2032 ± 0.0042	0.2043 ± 0.0051	0.2072 ± 0.0061	0.2122 ± 0.0056	0.2052 ± 0.0069	**0.2027** ± 0.0038	0.2029 ± 0.0060
Average	0.0766	0.0775	0.0772	0.0757	0.0807	0.0816	0.0821	0.0756
Avg Rank	3.62	4.20	4.50	4.03	6.20	5.55	5.90	2.00
CFIFS W/T/L	16/0/4	17/0/3	17/0/3	16/0/4	18/0/2	18/0/2	18/0/2	–
Holm p-value	**1.92e-02**	**9.58e-03**	**8.44e-03**	**1.92e-02**	**2.54e-03**	**9.35e-05**	**9.35e-05**	–

ture set, governed by the downstream classifier rather than the ranking quality; this convergence reconfirms that the primary value of feature selection lies precisely in the tight-to-moderate-budget regime where CFIFS excels.

Several per-dataset trajectories in Figs. 7 and 8 deserve specific comment. On *Genbase* and *Medical*, LSMFS, LRDG, and SCNMF collapse to near-zero performance across the entire p-grid—not just at p=20%—a failure of their fixed default configurations on small-sample biomedical datasets (see the oracle-tuning analysis above), while CFIFS and the sparse-regularisation baselines (GRRO, SRFS, RFSFS) remain robust under the same fixed-configuration policy. On *Yeast*, one of the rare datasets where SCNMF is competitive, CFIFS matches it closely across all budgets. The slight disadvantage of CFIFS on *Corel16k1/2* at p=20% is marginal (within 0.002 of the best baseline) but consistent across the p-grid, suggesting that for these specific image descriptors, pseudo-label regularisation provides a small domain-specific structural advantage that the Choquet fusion does not fully recover. Finally, on *Yelp*, CFIFS and LRMFS track closely and are statistically indistinguishable; notably, these two are the only methods that avoid the premature performance plateau observed for LSMFS and LRDG below p=15%.

4.7. Generalisation across downstream classifiers

ML-kNN is the standard evaluation classifier in the MLFS literature, but a natural question is whether the comparison depends on this choice. We therefore re-evaluated the stored rankings of all methods with two structurally different downstream classifiers at p=20%: BR ℓ_2-regularised logistic regression (a linear model) and a 100-tree random forest (a non-linear tree ensemble). Table 11 summarises the results, which support three conclusions.

Table 11
Generalisation across downstream classifiers (mean over the 20 benchmarks at p=20%, 10 folds; lower rank is better). "ML-kNN-aligned" is the default CFIFS, whose capacity is selected by inner-CV ML-kNN performance; "classifier-aligned" selects the capacity with the *target* classifier (BR-logistic) inside the same leakage-free inner CV. * marks values where the reference CFIFS variant of each panel is superior to the method with Holm-corrected Wilcoxon $p < 0.05$ (the star is placed on the dominated method's row).

	Micro-F1		Macro-F1		Hamming Loss	
Method	Mean	Rank	Mean	Rank	Mean	Rank
Downstream: Binary-Relevance logistic regression						
GRRO	0.3897*	3.73	0.2235*	3.33	0.0742*	3.88
SRFS	0.3786*	4.55	0.2123*	4.35	0.0756*	5.05
RFSFS	0.3804*	4.00	0.2145*	3.85	0.0755*	4.35
LRMFS	0.3834*	4.38	0.2142*	4.92	0.0737*	3.52
LSMFS	0.2618*	7.85	0.1230*	8.15	0.0802*	7.75
LRDG	0.2727*	7.05	0.1378*	7.05	0.0790*	6.25
SCNMF	0.2342*	7.35	0.1166*	7.45	0.0801*	7.05
CFIFS (ML-kNN-aligned)	0.3919*	4.33	0.2245*	3.98	0.0743*	4.47
CFIFS (classifier-aligned)	0.4033	1.77	0.2305	1.93	0.0729	2.67
Downstream: random forest						
GRRO	0.4234	3.35	0.2409	2.75	0.0673	3.05
SRFS	0.4128	3.52	0.2313	3.83	0.0699	3.62
RFSFS	0.4151	3.12	0.2339	3.33	0.0696	3.27
LRMFS	0.4124*	4.10	0.2303	3.80	0.0697	3.95
LSMFS	0.2827*	6.15	0.1483*	6.75	0.0747*	6.15
LRDG	0.2987*	6.45	0.1597*	6.15	0.0743*	6.15
SCNMF	0.2746*	6.50	0.1495*	6.15	0.0763*	6.55
CFIFS (ML-kNN-aligned)	0.4252	2.80	0.2360	3.25	0.0694	3.25

First, the qualitative structure of the comparison is classifier-independent: the three methods that collapse under ML-kNN (LSMFS, LRDG, SCNMF) remain far behind under both alternative classifiers, confirming that their failure reflects ranking quality rather than an ML-kNN idiosyncrasy. Second, the default CFIFS ranking—whose capacity is selected by inner-CV *ML-kNN* performance—transfers well: under the random forest it achieves the best average Micro-F1 (0.4252) and the best Micro-F1 rank (2.80), and under logistic regression it remains in the leading group, never statistically worse than any baseline. Third, and most importantly, the capacity-selection mechanism is *classifier-aware by design*: the inner CV can evaluate candidate capacities with whatever classifier the practitioner intends to deploy. When the inner criterion is aligned with the downstream task ("classifier-aligned" in Table 11, using BR-logistic inside the same inner CV; a one-option change with no other tuning), CFIFS becomes Holm-significantly superior to *all seven* baselines on *all three* metrics

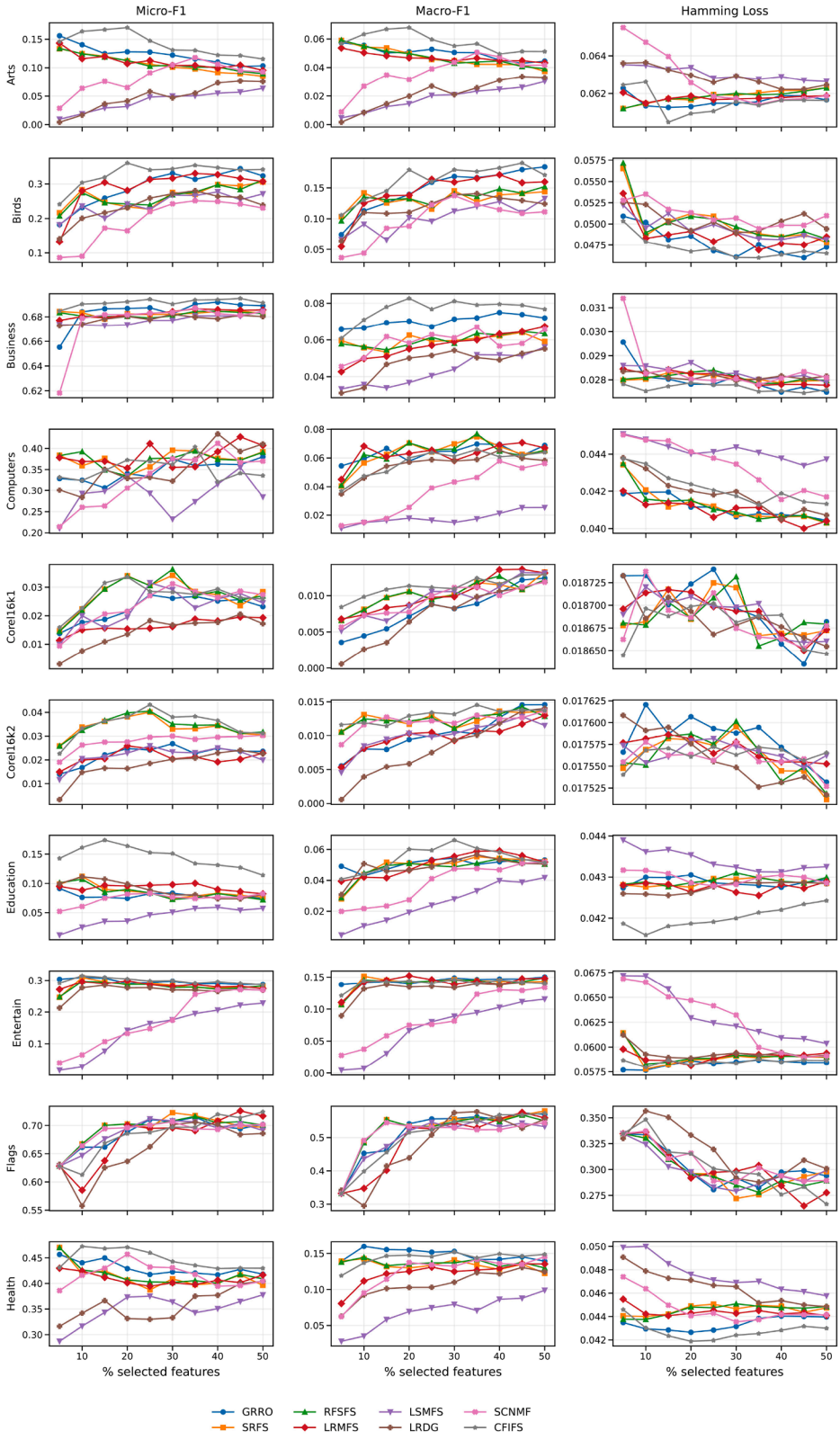

Fig. 7. Performance trajectories (Micro-F1, Macro-F1, Hamming Loss) across varying percentages of features selected ($p \in [5\%, 50\%]$) for the first subset of datasets.

under the logistic classifier as well (average ranks 1.77/1.93/2.67). This experiment highlights a structural advantage of validated capacity learning over fixed scoring rules: the fusion adapts not only to the dataset but also to the consumer of the ranking.

4.8. Extension to higher-dimensional benchmarks

To probe the behaviour of CFIFS beyond the moderate dimensionalities of the main suite, we extended the evaluation to eight

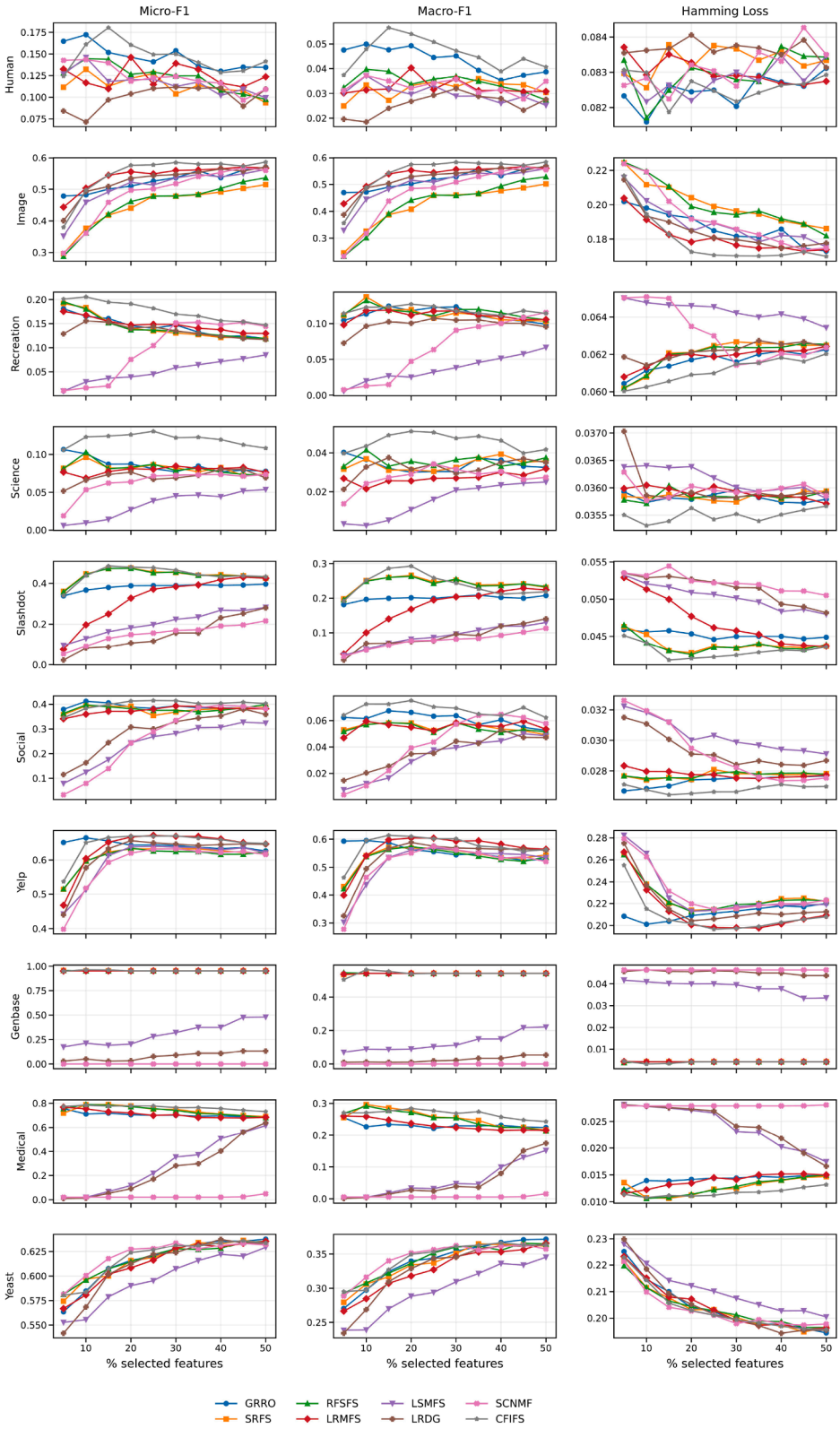

Fig. 8. Performance trajectories (Micro-F1, Macro-F1, Hamming Loss) across varying percentages of features selected ($p \in [5\%, 50\%]$) for the second subset of datasets.

additional public benchmarks: *Bibtex* (d=1,836, L=159), *Enron* (d=1,001, L=53), *Corel5k* (L=374, the largest label space considered), and the five *Rcv1* subsets (each d=10,000, L=101), which provide genuine high-dimensional text bag-of-words inputs and, being five statistically comparable corpora, let us check whether any trend is systematic rather than dataset-specific.

These benchmarks differ from the main suite in one respect that directly affects the embedded stage: their label spaces are much larger

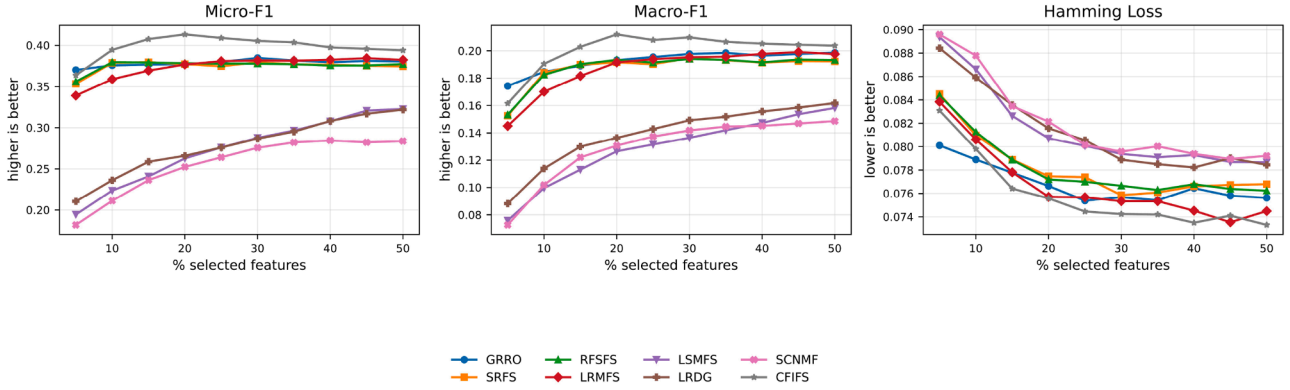

Fig. 9. Average performance across all datasets varying the percentage of features selected ($p \in [5\%, 50\%]$).

Table 12
Extension to higher-dimensional and larger-label-space benchmarks (mean over 10 folds at p=20%; best per row in bold; average mid-ranks over the eight datasets at the bottom, lower is better). The suite comprises *Bibtex* (d=1,836, L=159), *Enron* (d=1,001, L=53), *Corel5k* (d=499, L=374), and five *Rcv1* subsets (each d=10,000, L=101), the latter providing genuine high-dimensional text bag-of-words inputs. On this label-rich suite CFIFS uses the PLST embedding rank $r = \min(200, L)$ (see Section 4.8); all other settings are unchanged.

Dataset	Metric	GRRO	SRFS	RFSFS	LRMFS	LSMFS	LRDG	SCNMF	**CFIFS**
Bibtex	Micro-F1	0.3906	0.3970	**0.4008**	0.3626	0.2518	0.2935	0.3723	0.3980
	Macro-F1	0.1999	0.2063	0.2090	0.1745	0.0778	0.1315	0.1877	**0.2092**
Enron	Micro-F1	0.5077	0.5204	**0.5236**	0.4850	0.4073	0.5182	0.5220	0.5130
	Macro-F1	0.0982	0.1084	0.1099	0.0967	0.0760	0.1055	**0.1107**	0.1020
Corel5k	Micro-F1	0.0236	0.0540	**0.0553**	0.0230	0.0378	0.0139	0.0342	0.0543
	Macro-F1	0.0061	0.0088	**0.0090**	0.0048	0.0067	0.0028	0.0059	0.0086
Rcv1-s1	Micro-F1	0.3499	**0.3753**	0.3679	0.3678	0.1934	0.2205	0.0324	0.3705
	Macro-F1	0.1541	0.1638	0.1607	**0.1689**	0.0736	0.0841	0.0135	0.1641
Rcv1-s2	Micro-F1	0.2728	0.2762	0.2775	0.2785	0.0896	0.1158	0.0215	**0.2786**
	Macro-F1	0.1006	0.1153	**0.1163**	0.1111	0.0492	0.0478	0.0175	0.1125
Rcv1-s3	Micro-F1	0.3905	0.4021	0.4053	0.3993	0.0649	0.1830	0.0167	**0.4087**
	Macro-F1	0.1382	0.1513	0.1489	0.1506	0.0307	0.0559	0.0093	**0.1539**
Rcv1-s4	Micro-F1	0.4678	0.4882	0.4858	0.4814	0.1097	0.1962	0.0254	**0.4905**
	Macro-F1	0.1648	0.1738	0.1748	0.1758	0.0337	0.0565	0.0143	**0.1823**
Rcv1-s5	Micro-F1	0.4172	0.4293	**0.4337**	0.4238	0.1063	0.2171	0.0318	0.4262
	Macro-F1	0.1428	0.1613	**0.1616**	0.1524	0.0345	0.0594	0.0114	0.1604
Avg. rank (Micro-F1)		5.12	2.62	**1.88**	4.75	6.88	6.12	6.50	2.12
Avg. rank (Macro-F1)		5.00	2.62	**2.25**	4.25	6.75	6.25	6.50	2.38

(L up to 374 versus a maximum of 174 in Table 2). The PLST rank r=50 used in the main experiments is an efficiency *cap* appropriate for moderate label counts; when $L \gg r$, the truncated SVD discards label-space structure beyond the first r components and the embedded score loses information. We therefore set the embedding rank on this suite by the label-count rule $r = \min(200, L)$—the standard PLST regime, in which the rank tracks the available label dimensions—applied uniformly to all eight datasets and fixed *a priori*, without inspecting performance. This is the only setting that differs from Section 4.1; the fusion stage and all other hyper-parameters are unchanged, and we quantify below the effect of this single choice. Table 12 reports Micro- and Macro-F1 at p=20% together with the average mid-ranks over the eight datasets.

In this regime CFIFS is highly competitive: it attains the second-best average rank on both Micro-F1 (2.12) and Macro-F1 (2.38), behind only RFSFS (1.88/2.25) and ahead of SRFS (2.62/2.62), and it achieves the best per-dataset score on six of the sixteen dataset–metric cells (both metrics on *Rcv1-s3* and *Rcv1-s4*, Macro-F1 on *Bibtex*, and Micro-F1 on *Rcv1-s2*). A paired Wilcoxon test with Holm correction over the eight datasets finds CFIFS *statistically inseparable* from the two sparse-regression leaders RFSFS and SRFS on all three primary metrics, while it is significantly better than the filter baseline GRRO and than the three methods whose default configurations collapse at scale (LSMFS, LRDG, SCNMF; e.g. SCNMF drops to Micro-F1 0.0324 on *Rcv1-s1*). The five Rcv1 subsets confirm that this ordering is systematic rather than an artefact of a single corpus. Thus, even on genuinely high-dimensional, label-rich text—the regime most favourable to dedicated sparse-regression selectors—the late-fusion design remains on par with the strongest baselines and never exhibits the collapses of the graph/NMF methods. This regime-dependence is the expected behaviour of a fusion method: its advantage is largest where the embedded and spectral channels are complementary and no single specialised scorer already captures most of the signal—the conditions that hold across the moderate-dimensional main suite (Section 3.3), where CFIFS ranks first overall—and narrows in the ultra-sparse, very-high-dimensional regime, where purpose-built sparse-regression selectors already account for most of the sparse-linear signal.

Two analyses clarify the role of the embedding rank and of the fusion. First, the rank rule is decisive but not a form of accuracy tuning: had we kept the main-suite cap r=50 on this suite, CFIFS would drop to mid-ranks 3.62/3.50 (Micro/Macro), i.e. from second to mid-cluster, with the gap concentrated entirely on the large-label datasets (L=159–374) and

absent on *Rcv1* (L=101, already close to $\min(200, L)$). This is exactly the signature of SVD truncation, and it is corrected by the label-count rule with no change to the fusion. Second, a channel decomposition confirms that the fusion is not the limiting factor: CFIFS matches or exceeds the better of its own two channels almost everywhere, and on the Rcv1 subsets it outperforms *both*—for instance on *Rcv1-s1* it reaches Micro-F1 0.3705 against 0.3199 (embedded-only) and 0.3455 (spectral-only), exploiting an unusually informative spectral channel (μ_s=0.72). The residual edge of RFSFS/SRFS on a few datasets therefore reflects the strength of their embedded scorers on sparse text, not a weakness of the Choquet combination.

Finally, two practical points. Consistently with the tight-budget behaviour observed on the main suite, CFIFS retains an advantage in the most selective regimes, where feature selection matters most: it attains the best Micro-F1 at $p \leq 10\%$ on *Bibtex* and at p=5% on *Corel5k*, the baselines' parity materialising only at the larger p=20% budget. And the method remains computationally practical at scale: on *Rcv1* (d=10,000) the complete pipeline takes about one minute per fold on a single GPU, since the $O(n^2)$ spectral cost depends on the (moderate) instance count rather than on d, and the d-dependent costs are linear.

5. Conclusion and future work

This article introduced CFIFS, a multi-label feature selection framework that combines embedded and spectral scoring methods as complementary information channels through a two-source Choquet integral. By means of rank-based normalisation, the framework maps two heterogeneous score distributions onto a common scale. Rather than relying on a fixed linear aggregation rule, CFIFS learns the non-additive capacity of the Choquet integral through an inner cross-validation loop. This allows the model to adapt the fusion behaviour to the relationship between the two channels, capturing synergy or redundancy when needed, while also providing an interpretable interaction index at no additional inference cost.

Experiments on 20 benchmarks under a rigorous iteratively stratified 10-fold cross-validation protocol showed that CFIFS achieved the best average rank on all seven evaluation metrics at the standard p=20% operating point, with statistically significant improvements over all seven state-of-the-art baselines under the Holm correction on every primary metric. The feature-budget analysis further showed that these gains remain consistent across the tight-to-moderate budget range, rather than being tied to a single operating point. The ablation study highlighted the importance of the fusion stage: removing the spectral channel led to a statistically significant drop in performance, while replacing the learned capacity with a fixed average removed the adaptive behaviour that limits its performance degradation when one source is less informative—an effect especially visible on datasets with strong source asymmetry such as *Yelp*, *Genbase*, and *Medical*. Three further analyses corroborated the framework: under training-label noise, the validated capacity learning rescued opposite channel-failure modes that no fixed rule can handle; the rankings transferred to structurally different downstream classifiers, and aligning the inner criterion with the target classifier restored statistically significant superiority over all baselines under a linear classifier as well; and on eight higher-dimensional benchmarks (up to d=10,000, five of them genuine text bag-of-words corpora) CFIFS attained the second-best average rank—statistically on par with the strongest sparse-regression baselines and clearly above the methods that collapse at scale—once the PLST embedding rank was scaled to the larger label spaces, a single label-count-driven adjustment that a channel-level analysis confirms acts on the embedded scorer rather than on the fusion stage.

Several directions remain open for future work. First, the $O(n^2)$ memory cost of the spectral stage limits scalability on datasets with very large numbers of instances; Nyström approximations, anchor-graph constructions, or distributed formulations of the selection problem (Ju et al., 2026) could help alleviate this bottleneck, although the effect of the resulting approximation error on the final ranking would need careful quantification. Second, although the current framework fuses only two channels, k-additive capacities (Grabisch et al., 2008) could extend CFIFS to combine three or more heterogeneous sources, such as mutual-information filters or deep neural embeddings, while keeping the number of capacity parameters manageable. Third, replacing the convex embedded stage with a non-linear encoder could enable the framework to capture higher-order feature interactions before fusion, while deriving a data-driven stopping criterion from the fused scores could remove the need to specify the feature budget p in advance. Finally, since the fuzzy capacity is learned from validation performance rather than from the score distributions themselves, studying whether such capacities can transfer across related domains or support few-shot multi-label settings would be a natural next step.

Overall, CFIFS suggests that multi-label feature selection can also be viewed as a score-level fusion problem, rather than only as the design of a single scoring function. From this perspective, different MLFS methods can be treated as complementary modules whose combination may provide a more robust ranking than either source alone or a simple additive aggregation.

CRediT authorship contribution statement

Filippo Casu: Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing; **Antonio Costantino Marceddu:** Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing; **Giuseppe A. Trunfio:** Supervision, Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

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

Data availability

A complete implementation of CFIFS is publicly available at: https://github.com/ailab-uniss/CFIFS. All 20 datasets stem from standard public benchmarks (including the MULAN (Tsoumakas et al., 2011) repository and the Yahoo! corpus).

References

Bai, J., & Wu, Y. (2024). Graph-based multi-label feature selection with dynamic graph constraints and latent representation learning. *Applied Intelligence*, *55*(3). https://doi.org/10.1007/s10489-024-06116-3

Batista, T. V. V., Bedregal, B. R. C., & Moraes, R. M. (2022). Constructing multi-layer classifier ensembles using the choquet integral based on overlap and quasi-overlap functions. *Neurocomputing*, *500*, 413–421. https://doi.org/10.1016/j.neucom.2022.05.080

Bustince, H., Mesiar, R., Dimuro, G. P., Fernandez, J., Lafuente, J., & De Baets, B. (2026). Choquet-inspired aggregation functions. *Information Fusion*, *126*, 103499. https://doi.org/10.1016/j.inffus.2025.103499

Cao, Y., Wang, C., Li, M., Huang, J., & Wang, H. (2012). Aggregating multiple classification results using choquet integral for financial distress early warning. *Expert Systems with Applications*, *39*(4), 4082–4089. https://doi.org/10.1016/j.eswa.2011.08.067

Charte, F., Rivera, A. J., Jesus, M. J. d., & Herrera, F. (2015). Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, *163*, 3–16. https://doi.org/10.1016/j.neucom.2014.08.091

Chen, Y.-N., & Lin, H.-T. (2012). Feature-aware label space dimension reduction for multi-label classification. In *Advances in neural information processing systems 25 (neurIPS 2012)* (pp. 1529–1537).

Dai, J., Li, M., & Zhang, C. (2025). Multi-label feature selection with missing labels by weak-label fusion fuzzy discernibility pair. *Information Fusion*, *117*, 102921. https://doi.org/10.1016/j.inffus.2024.102921

Fan, Y., Chen, X., Luo, S., Liu, P., Liu, J., Chen, B., & Tang, J. (2024a). Label relaxation and shared information for multi-label feature selection. *Information Sciences*, *671*, 120662. https://doi.org/10.1016/j.ins.2024.120662

Fan, Y., Liu, J., Liu, P., Du, Y., Lan, W., & Wu, S. (2025). Multi-label feature selection via label relaxation. *Applied Soft Computing*, *175*, 113047. https://doi.org/10.1016/j.asoc.2025.113047

Fan, Y., Liu, J., Tang, J., & Liu, P. (2024b). Learning correlation information for multi-label feature selection. *Pattern Recognition*, *145*, 109899. https://doi.org/10.1016/j.patcog.2023.109899

Fan, Y., Liu, P., & Liu, J. (2026). Reconstructing data representation for multi-label feature selection. *Pattern Recognition*, *169*, 111941. https://doi.org/10.1016/j.patcog.2025.111941

Gao, W., Li, Y., & Hu, L. (2023). Multilabel feature selection with constrained latent structure shared term. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(3), 1253–1262. https://doi.org/10.1109/TNNLS.2021.3104230

Grabisch, M. (1996). The application of fuzzy integrals in multicriteria decision making. *European Journal of Operational Research*, *89*(3), 445–456. https://doi.org/10.1016/0377-2217(95)00176-x

Grabisch, M., Kojadinovic, I., & Meyer, P. (2008). A review of methods for capacity identification in choquet integral based multi-attribute utility theory: Applications of the kappalab r package. *European Journal of Operational Research*, *186*(2), 766–785. https://doi.org/10.1016/j.ejor.2007.02.025

Grabisch, M., & Labreuche, C. (2007). A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid. *4OR*, *6*(1), 1–44. https://doi.org/10.1007/s10288-007-0064-2

Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with Hilbert–Schmidt norms. In *Proceedings of the 16th international conference on algorithmic learning theory (ALT)* (pp. 63–77). Springer. https://doi.org/10.1007/11564089_7

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hancer, E., Xue, B., & Zhang, M. (2025). A survey on evolutionary feature selection in multi-label classification. *IEEE Transactions on Evolutionary Computation*, . https://doi.org/10.1109/TEVC.2025.3546780

Hao, P., Gao, W., & Hu, L. (2025). Embedded feature fusion for multi-view multi-label feature selection. *Pattern Recognition*, *157*, 110888. https://doi.org/10.1016/j.patcog.2024.110888

Hashemi, A., Dowlatshahi, M. B., & Nezamabadi-pour, H. (2020). MFS-MCDM: Multi-label feature selection using multi-criteria decision making. *Knowledge-Based Systems*, *206*, 106365. https://doi.org/10.1016/j.knosys.2020.106365

Hashemi, A., Dowlatshahi, M. B., & Nezamabadi-pour, H. (2021). VMFS: A vikor-based multi-target feature selection. *Expert Systems with Applications*, *182*, 115224. https://doi.org/10.1016/j.eswa.2021.115224

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

He, X., Cai, D., & Niyogi, P. (2005). Laplacian score for feature selection. In *Advances in neural information processing systems 18 (neurIPS 2005)* (pp. 507–514).

He, Z., Lin, Y., Lin, Z., & Wang, C. (2024). Multi-label feature selection via similarity constraints with non-negative matrix factorization. *Knowledge-Based Systems*, (p. 111948). https://doi.org/10.1016/j.knosys.2024.111948

Hu, J., Li, Y., Xu, G., & Gao, W. (2022). Dynamic subspace dual-graph regularized multi-label feature selection. *Neurocomputing*, *467*, 184–196. https://doi.org/10.1016/j.neucom.2021.10.022

Hu, L., Li, Y., Gao, W., Zhang, P., & Hu, J. (2020). Multi-label feature selection with shared common mode. *Pattern Recognition*, *104*, 107344.

Huang, J., Qian, W., Vong, C.-M., Ding, W., Shu, W., & Huang, Q. (2023). Multi-label feature selection via label enhancement and analytic hierarchy process. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *7*(5), 1377–1393. https://doi.org/10.1109/TETCI.2022.3231655

Huang, J.-J., & Chen, C.-Y. (2024). Leveraging the hierarchical symmetric 2-additive Choquet integral: Enhancing explainability and parallelizability in predictive models. *Information Sciences*, *678*, 121031. https://doi.org/10.1016/j.ins.2024.121031

Huang, R., Jiang, W., & Sun, G. (2018). Manifold-based constraint Laplacian score for multi-label feature selection. *Pattern Recognition Letters*, *112*, 346–352.

Huang, R., & Wu, Z. (2021). Multi-label feature selection via manifold regularization and dependence maximization. *Pattern Recognition*, *120*, 108149.

Jia, Q., Deng, T., Wang, Y., & Wang, C. (2024). Discriminative label correlation based robust structure learning for multi-label feature selection. *Pattern Recognition*, *154*, 110583. https://doi.org/10.1016/j.patcog.2024.110583

Ju, H., Tao, X., Ding, W., Lu, Z., Xu, S., Qiao, L., & Yang, X. (2026). Distributed multi-label feature selection via feature-label information granulation. *Information Sciences*, *742*, 123334. https://doi.org/10.1016/j.ins.2026.123334

Kashef, S., Nezamabadi-pour, H., & Nikpour, B. (2018). Multilabel feature selection: A comprehensive review and guiding experiments. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(2), e1240. https://doi.org/10.1002/widm.1240

Lee, J., & Kim, D.-W. (2013). Feature selection for multi-label classification using multivariate mutual information. *Pattern Recognition Letters*, *34*(3), 349–357. https://doi.org/10.1016/j.patrec.2012.10.005

Lee, J., & Kim, D.-W. (2015a). Fast multi-label feature selection based on information-theoretic feature ranking. *Pattern Recognition*, *48*(9), 2761–2771. https://doi.org/10.1016/j.patcog.2015.04.009

Lee, J., & Kim, D.-W. (2015b). Mutual information-based multi-label feature selection using interaction information. *Expert Systems with Applications*, *42*(4), 2013–2025. https://doi.org/10.1016/j.eswa.2014.09.063

Lee, J., & Kim, D.-W. (2017). SCLS: Multi-label feature selection based on scalable criterion for large label set. *Pattern Recognition*, *66*, 342–352. https://doi.org/10.1016/j.patcog.2017.01.014

Li, Y., Hu, L., & Gao, W. (2022). Label correlations variation for robust multi-label feature selection. *Information Sciences*, *609*, 1075–1097. https://doi.org/10.1016/j.ins.2022.07.154

Li, Y., Hu, L., & Gao, W. (2023). Multi-label feature selection via robust flexible sparse regularization. *Pattern Recognition*, *134*, 109074. https://doi.org/10.1016/j.patcog.2022.109074

Li, Y., Wang, X., Yang, X., Gao, W., Ding, W., & Li, T. (2025). Fusion-enhanced multi-label feature selection with sparse supplementation. *Information Fusion*, *117*, 102813.

Lin, Y., Hu, Q., Liu, J., & Duan, J. (2015). Multi-label feature selection based on max-dependency and min-redundancy. *Neurocomputing*, *168*, 92–103. https://doi.org/10.1016/j.neucom.2015.06.010

Miri, M., Dowlatshahi, M. B., Hashemi, A., Rafsanjani, M. K., Gupta, B. B., & Alhalabi, W. (2022). Ensemble feature selection for multi-label text classification: An intelligent order statistics approach. *International Journal of Intelligent Systems*, *37*(12), 11319–11341. https://doi.org/10.1002/int.23044

Mohanrasu, S. S., Rakkiyappan, R., & Gochoo, M. (2026). A framework for classifier-neutral ensemble multi-label feature selection via swarm optimization. *Results in Engineering*, *29*, 108756. https://doi.org/10.1016/j.rineng.2025.108756

Murofushi, T., & Soneda, S. (1993). Techniques for reading fuzzy measures (III): interaction index. *Proceedings of the 9th Fuzzy System Symposium*, (pp. 693–696).

Murofushi, T., & Sugeno, M. (1989). An interpretation of fuzzy measures and the Choquet integral as an integral with respect to a fuzzy measure. *Fuzzy Sets and Systems*, *29*(2), 201–227. https://doi.org/10.1016/0165-0114(89)90194-2

Nie, F., Huang, H., Cai, X., & Ding, C. (2010). Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. *Advances in Neural Information Processing Systems*, *23*, 1813–1821.

Pacheco, A. G. C., & Krohling, R. A. (2018). Aggregation of neural classifiers using Choquet integral with respect to a fuzzy measure. *Neurocomputing*, *292*, 151–164. https://doi.org/10.1016/j.neucom.2018.03.002

Pereira, R. B., Plastino, A., Zadrozny, B., & Mello, L. H. A. (2018). Categorizing feature selection methods for multi-label classification. *Artificial Intelligence Review*, *49*(1), 57–78. https://doi.org/10.1007/s10462-016-9516-4

Qian, W., Huang, J., Xu, F., Shu, W., & Ding, W. (2023). A survey on multi-label feature selection from perspectives of label fusion. *Information Fusion*, *100*, 101948. https://doi.org/10.1016/j.inffus.2023.101948

Read, J., Pfahringer, B., & Holmes, G. (2008). Multi-label classification using ensembles of pruned sets. In *2008 Eighth IEEE international conference on data mining* (pp. 995–1000). IEEE. https://doi.org/10.1109/icdm.2008.74

Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the stratification of multi-label data. In *Machine learning and knowledge discovery in databases (ECML PKDD)* (pp. 145–158). Springer. https://doi.org/10.1007/978-3-642-23808-6_10

Spolaor, N., Cherman, E., Monard, M. C., & Lee, H. D. (2016). Efficient multi-label feature selection using entropy-based label selection. *Entropy*, *18*(11), 405. https://doi.org/10.3390/e18110405

Sun, L., Ji, S., & Ye, J. (2008). Hypergraph spectral learning for multi-label classification. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* KDD08 (pp. 668–676). ACM. https://doi.org/10.1145/1401890.1401971

Szymański, P., & Kajdanowicz, T. (2017). A network perspective on stratification of multi-label data. In *Proceedings of the first international workshop on learning with imbalanced domains: Theory and applications (LIDTA)* (pp. 22–35). PMLR (*vol. 74*). Proceedings of Machine Learning Research.

Tai, F., & Lin, H.-T. (2012). Multilabel classification with principal label space transformation. *Neural Computation*, *24*(9), 2508–2542. https://doi.org/10.1162/NECO_a_00328

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, *3*(3), 1–13. https://doi.org/10.4018/jdwm.2007070101

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2009). Mining multi-label data. In *Data mining and knowledge discovery handbook* (pp. 667–685). Springer US. https://doi.org/10.1007/978-0-387-09823-4_34

Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., & Vlahavas, I. (2011). MULAN: A java library for multi-label learning. *Journal of Machine Learning Research*, *12*, 2411–2414.

Wang, T. (2026). A survey on Hilbert–Schmidt independence criterion Lasso. *Knowledge-Based Systems*, *342*, 115934. https://doi.org/10.1016/j.knosys.2026.115934

Wang, T., Hu, Z., & Liu, H. (2023). A unified view of feature selection based on Hilbert-Schmidt independence criterion. *Chemometrics and Intelligent Laboratory Systems*, *236*, 104807. https://doi.org/10.1016/j.chemolab.2023.104807

Wu, Y., Li, P., & Zou, Y. (2025). Partial multi-label feature selection with feature noise. *Pattern Recognition*, *162*, 111310. https://doi.org/10.1016/j.patcog.2024.111310

Xu, W., & Li, Y. (2025). Multi-label feature selection for imbalanced data via KNN-based multi-label rough set theory. *Information Sciences*, *715*, 122220. https://doi.org/10.1016/j.ins.2025.122220

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., & Sugiyama, M. (2014). High-dimensional feature selection by feature-wise kernelized Lasso. *Neural Computation*, *26*(1), 185–207. https://doi.org/10.1162/NECO_a_00537

Zhang, J., Lin, Y., Jiang, M., Li, S., Tang, Y., & Tan, K. C. (2020). Multi-label feature selection via global relevance and redundancy optimization. In *Proceedings of the twenty-ninth international joint conference on artificial intelligence* (pp. 2512–2518). IJCAI.

Zhang, J., Luo, Z., Li, C., Zhou, C., & Li, S. (2019). Manifold regularized discriminative feature selection for multi-label learning. *Pattern Recognition*, *95*, 136–150.

Zhang, M.-L., & Zhou, Z.-H. (2007). ML-kNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, *40*(7), 2038–2048. https://doi.org/10.1016/j.patcog.2006.12.019

Zhang, M.-L., & Zhou, Z.-H. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *26*(8), 1819–1837. https://doi.org/10.1109/TKDE.2013.39

Zhang, N., Wang, A., Lu, P., Feng, T., Xu, Y., & Du, G. (2025a). Multi-label feature selection with feature reconstruction and label correlations. *Expert Systems with Applications*, *285*, 127993. https://doi.org/10.1016/j.eswa.2025.127993

Zhang, Y., Gong, D.-w., Sun, X.-y., & Guo, Y.-n. (2017). A PSO-based multi-objective multi-label feature selection method in classification. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-00416-0

Zhang, Y., Huo, W., & Tang, J. (2024). Multi-label feature selection via latent representation learning and dynamic graph constraints. *Pattern Recognition*, *151*, 110411. https://doi.org/10.1016/j.patcog.2024.110411

Zhang, Y., & Ma, Y. (2022). Non-negative multi-label feature selection with dynamic graph constraints. *Knowledge-Based Systems*, *238*, 107924.

Zhang, Y., Tang, J., Cao, Z., & Chen, H. (2025b). Sparse multi-label feature selection via pseudo-label learning and dynamic graph constraints. *Information Fusion*, *118*, 102975. https://doi.org/10.1016/j.inffus.2025.102975

Zhang, Z., Ma, X.-A., Liu, J., Zhang, P., & Gao, W. (2023). Multi-label feature selection via adaptive label correlation estimation. *ACM Transactions on Knowledge Discovery from Data*, *17*(9), 1–28. https://doi.org/10.1145/3604560