COFS: Counterfactual Optimization for Reinforcement Learning Feature Selection

Zixing Chen[1,2], Xiaohan Huang[1,2], Yuanchun Zhou[1,2,*]
[1]Computer Network Information Center, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
zxchen@cnic.cn, xhhuang@cnic.cn, zyc@cnic.cn

***Abstract*—Reinforcement learning (RL) has recently been used for automated feature selection by exploring the exponentially large subset space. However, most RL-based methods update policies with a single subset-level reward, which yields coarse feedback and is insufficient for attributing rewards to individual features under complex interactions. We propose Counterfactual Optimization for Reinforcement Learning Feature Selection (COFS), which improves RL-based feature selection by providing feature-wise counterfactual advantages for policy optimization. Specifically, COFS trains a critic to estimate rewards and computes a counterfactual advantage for each selected feature. This is achieved by contrasting the predicted reward of the current subset with that of the leave-one-out variants formed by removing each feature, effectively distinguishing important features from redundant ones without repeatedly invoking the downstream model. Moreover, we introduce a pretraining stage that learns subset state representations based on reward signals rather than solely on data distribution. This strategy enhances the critic's prediction accuracy and facilitates more effective policy optimization. Finally, extensive experimental results demonstrate the effectiveness of our proposed method, showcasing notable enhancements in subset quality and compactness.**

***Index Terms*—Automated Feature Selection, Reinforcement Learning, Representation Learning**

I. Introduction

Feature selection aims to identify a compact subset of informative features that improves predictive performance while reducing model complexity and redundancy [1]. This process plays a pivotal role in many real-world applications, such as medical diagnosis [2]–[4] and gene panel selection [5], where identifying relevant biomarkers enhances interpretability and reliability, and recommender systems [6], where removing irrelevant dimensions reduces computational costs. However, identifying the optimal subset remains challenging due to the exponentially large search space. Traditional feature selection methods can be broadly categorized into three paradigms [7]. Filter methods select the top-K ranked features by statistical criteria. Embedded methods integrate feature selection into model training through a regularization term. Wrapper methods iteratively generate candidate subsets and evaluate them with a downstream model, selecting the best-performing subset [8]–[10].

Recently, reinforcement learning (RL) has offered a promising perspective on feature selection by enabling automated exploration in the combinatorial subset space [11]–[14]. However, existing RL-based methods often face a tension between modeling granularity and scalability. On the one hand, some methods assign independent policy networks to each feature to model their heterogeneity [12], [13], which increases parameters and training overhead as the dimension grows. On the other hand, approaches such as [14] utilize a unified policy network to generate actions sequentially, which compromises decision efficiency due to order-dependent exploration. More fundamentally, these approaches typically use only subset-level rewards to guide policy updates [15], [16]. As illustrated in Fig. 1, features within a selected subset often contribute unequally to the global performance, while complex interactions such as redundancy and complementarity are pervasive [17], [18]. Consequently, updating policies with such a coarse signal renders individual feature contributions indistinguishable [19]. This ambiguity prevents the policy from effectively identifying critical versus redundant features, thereby slowing convergence and leading to suboptimal solutions [20].

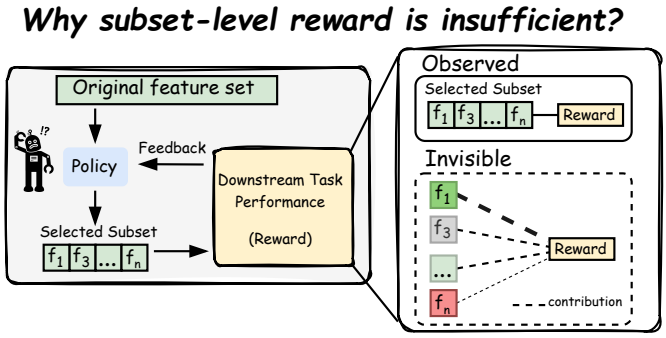


Fig. 1. The limitation of subset-level rewards. While a downstream model evaluates a subset, the performance typically emerges from complex feature interactions, where some features are informative while others are redundant and may impair performance. A single subset-level reward only provides coarse feedback, obscuring the contribution of each feature.

To overcome these limitations, we aim to derive feature-wise learning signals that reflect the contribution of each feature within the selected subset. Drawing inspiration from [21], the contribution of a specific feature can be estimated by comparing the reward of the current subset against a counterfactual baseline where that feature is removed.

In this paper, we propose **C**ounterfactual **O**ptimization for Reinforcement Learning **F**eature **S**election **(COFS)**, which

*Corresponding author.
This work is partially supported by National Natural Science Foundation of China (No.92470204).

performs policy optimization via feature-wise counterfactual advantages. To remain scalable while allowing heterogeneous decisions, COFS adopts a shared-parameter actor that outputs per-feature selections conditioned on feature embeddings. Adapting the strategy from [22], we further train a reward-predictive critic as a surrogate to estimate subset rewards efficiently without invoking the downstream model. The resulting counterfactual advantages provide informative signals to distinguish informative features from redundant ones under complex interactions. Moreover, we introduce a reward-aware pretraining stage that learns more informative subset state representations guided by downstream rewards rather than being dominated by data distribution effects, thereby mitigating cold-start issues and stabilizing policy optimization.

Overall, our contributions are summarized as follows:

- We identify a limitation of subset-level rewards in RL-based feature selection methods, where the resulting learning signal is often insufficient to reveal the effect of each selected feature under complex interactions.
- We propose COFS, a unified framework that integrates a permutation-invariant encoder with a shared actor to enable heterogeneous decisions. In addition, we incorporate a reward-aware pretraining strategy to learn better subset representations, facilitating stable policy optimization.
- We introduce feature-wise counterfactual advantages, providing informative signals for policy optimization. Extensive experiments demonstrate that our approach identifies compact feature subsets while maintaining strong predictive performance across multiple datasets and tasks.

To facilitate reproducibility, our code is publicly accessible at https://github.com/w1ndz321/COFS.

II. Problem Statement

Given a dataset $\mathcal{S} = \{(\mathbf{x}_k, y_k)\}_{k=1}^{K}$ consisting of K samples, where $\mathbf{x}_k \in \mathbb{R}^N$ denotes the feature vector of the k-th sample and y_k is the corresponding target label, and with N features $\mathcal{F} = \{f_1, \ldots, f_N\}$, a feature subset is represented by a binary mask $\mathbf{m} \in \{0,1\}^N$, where $m^i = 1$ indicates selecting feature f_i and $m^i = 0$ otherwise. For a given mask $\mathbf{m}$, we evaluate the corresponding subset with a downstream ML model and obtain a scalar reward $R(\mathbf{m})$. Our goal is to learn a parameterized selection policy π_θ that induces a distribution over masks. Starting from a randomly initialized mask, the policy iteratively samples a new binary mask conditioned on the current subset state to refine the selected subset. Each sampled mask is evaluated to obtain a respective reward, and the collected samples are used to update the policy. Accordingly, we optimize π_θ to maximize the expected reward:

$$\theta^* = \arg\max_{\theta} \ \mathbb{E}_{\mathbf{m}\sim\pi_\theta}\big[R(\mathbf{m})\big]. \tag{1}$$

This objective encourages the policy to favor compact subsets with strong predictive performance.

III. Methodology

A. Framework Overview

The overall framework of COFS is illustrated in Fig. 2. Our method consists of two stages: (1) reward-aware pretraining of a subset encoder and a reward-predictive critic, (2) and RL-based policy optimization for feature selection. In the pretraining stage, we compute per-feature statistical descriptors and learn feature embeddings to provide informative and heterogeneous inputs. We then randomly sample subset masks and obtain their rewards via downstream evaluation, forming subset-reward pairs. These pairs are used to jointly train a set-based subset encoder that maps the subset to a permutation-invariant state representation, and a reward-predictive critic that outputs the corresponding reward. Through pretraining, the encoder provides subset states for decision making and the critic serves as an efficient surrogate for reward prediction. In the policy optimization stage, the actor outputs a binary selection action for each feature conditioned on the current subset state and the feature embedding, yielding a new subset mask. The resulting subset is evaluated to obtain its reward, which is further used to refine the critic. To provide feature-wise learning signals, we construct leave-one-out counterfactual subsets by removing one selected feature at a time, and use the critic to predict their rewards efficiently in batches. Comparing the predicted rewards of the factual and counterfactual subsets yields feature-wise counterfactual advantages that approximate marginal effects under the current subset context, enabling more informative and stable policy updates than a single subset-level reward.

B. Reward-aware Pretraining

Why reward-aware pretraining matters. Representation learning is widely recognized as a key factor in stabilizing and improving efficiency in reinforcement learning, since the policy and the value function rely on the state to summarize the environment [23]. In the context of feature selection, the state representation should capture how a selected subset is expected to perform on the downstream task, rather than merely describing the data distribution induced by the subset. However, existing RL-based feature selection methods [12], [24] often use the fixed statistics of the selected features to construct the state. In this way, it may fail to distinguish subsets that are statistically similar but differ significantly in predictive performance, which makes it difficult for the policy to learn reliable selection behaviors and for the critic to generalize value estimates across subsets. To address this issue, we introduce a reward-aware pretraining stage that jointly trains a subset encoder and a critic to predict subset rewards. This encourages the learned state representations to align more closely with downstream performance, providing more informative representations for later policy learning and more robust value estimation.

Preparation of training data. We first compute a raw statistical descriptor vector $\mathbf{c}_i$ for each feature f_i, summarizing its distributional characteristics and statistical dependencies.

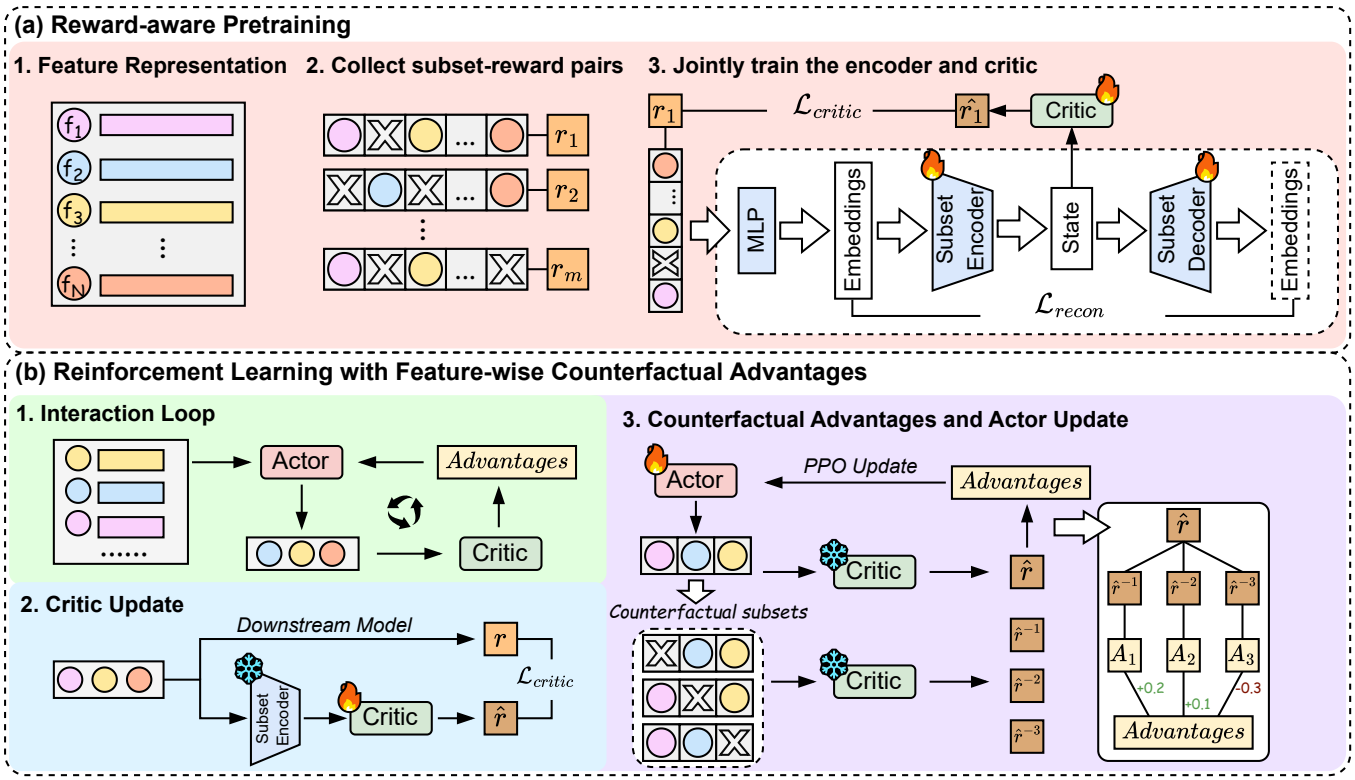


Fig. 2. An overview of the proposed framework. (a) Reward-aware Pretraining: we compute feature representations and collect subset–reward pairs to jointly train the subset encoder and a reward-predictive critic. (b) Reinforcement Learning with Feature-wise Counterfactual Advantages: the actor conditions on the current state and feature embeddings to produce binary actions, forming a new subset; the reward is used to refine the critic, whose predicted rewards on the factual subset and counterfactual subsets are further used to compute feature-wise advantages for updating the actor.

To distinguish features, we additionally assign each feature a learnable identifier embedding $\mathbf{e}_i$ and map the concatenation $[\mathbf{c}_i; \mathbf{e}_i]$ through an MLP [25] to obtain the feature embedding matrix $\boldsymbol{\beta} \in \mathbb{R}^{N\times d}$. We then randomly sample a batch of subsets represented by $\mathbf{m}_j \in \{0,1\}^N$, where $m_j^i = 1$ indicates selecting f_i. Finally, we compute the reward for $\mathbf{m}_j$: $r_j = \lambda p_j + (1-\lambda)\left(1 - \frac{\|\mathbf{m}_j\|_0}{N}\right)$, yielding pairs $\{(\mathbf{m}_j, r_j)\}$ for pretraining, where p_j is the predictive performance of subset $\mathbf{m}_j$ evaluated by a downstream ML model, $\|\mathbf{m}_j\|_0$ is the number of selected features, and $\lambda \in [0,1]$ balances predictive performance and sparsity.

Set-based Encoder. Since a feature subset is an unordered set whose semantics are invariant to the ordering of its elements, its state representation should be permutation-invariant [26]. Accordingly, we design a set-based encoder ω to map the masked feature embeddings to a global subset state, $\mathbf{s}_j = \omega(\mathbf{m}_j \odot \boldsymbol{\beta})$, where $\odot$ denotes element-wise multiplication. To capture pairwise interactions in subsets, ω adopts an attention-based set architecture. However, applying standard self-attention to N feature embeddings incurs $O(N^2)$ complexity. Inspired by the efficient design in [26], we adopt the Induced Set Attention Block (ISAB) [27], which introduces a small set of trainable inducing points $\mathbf{I} \in \mathbb{R}^{M\times d}$ ($M \ll N$) as intermediaries, reducing the complexity to $O(NM)$ per block. The resulting $\mathbf{s}_j$ serve as a compact subset representation, which is used for policy learning and reward prediction.

Critic for reward prediction. Evaluating the reward of a subset requires repeatedly invoking the downstream model, which becomes prohibitively expensive when scoring a large number of counterfactual subsets. Inspired by [22], we introduce a reward-predictive critic that approximates the subset reward from its state representation. Given the encoded state $\mathbf{s}$, the critic outputs a reward prediction $\hat{r} = V_{\phi_c}(\mathbf{s})$, where $V_{\phi_c}(\cdot)$ is a neural network parameterized by ϕ_c. The critic is first trained on sampled subset-reward pairs during the pretraining stage, and is further refined during policy optimization as new subsets are evaluated. With this design, we can efficiently predict rewards for many counterfactual subsets in batches, enabling computation of feature-wise counterfactual advantages.

Optimization objectives. To learn informative and reward-aligned subset representations, we jointly pretrain the subset encoder ω_{ϕ_e} and the reward-predictive critic V_{ϕ_c} on sampled subset-reward pairs. Let $\mathcal{D} = \{(\mathbf{m}_j, r_j)\}$ denote the pretraining set of masks and their corresponding rewards, and define the encoded subset state as $\mathbf{s}_j = \omega_{\phi_e}(\mathbf{m}_j \odot \boldsymbol{\beta})$, where $\boldsymbol{\beta} \in \mathbb{R}^{N\times d}$ is the matrix of embeddings of features and $\mathbf{m}_j$ denotes the corresponding subset. We define two losses:

$$\ell_{\mathrm{r}}(j) \triangleq \left(V_{\phi_c}(\mathbf{s}_j) - r_j\right)^2, \tag{2}$$

$$\ell_{\mathrm{rec}}(j) \triangleq \|\psi(\mathbf{s}_j) - (\mathbf{m}_j \odot \boldsymbol{\beta})\|_F^2, \tag{3}$$

where ψ is an auxiliary decoder and $\|\cdot\|_F$ denotes the Frobenius norm. The critic is trained by minimizing the

expected reward-prediction loss:

$$\mathcal{L}_{\text{critic}}(\phi_c) = \mathbb{E}_{(\mathbf{m}_j, r_j)\sim\mathcal{D}}\left[\ell_{\text{r}}(j)\right]. \tag{4}$$

The encoder is optimized with a composite objective that combines reward prediction and subset reconstruction:

$$\mathcal{L}_{\text{enc}}(\phi_e) = \mathbb{E}_{(\mathbf{m}_j, r_j)\sim\mathcal{D}}\left[\ell_{\text{rec}}(j) + \ell_{\text{r}}(j)\right]. \tag{5}$$

Overall, ℓ_{r} aligns the learned state with subset quality, while ℓ_{rec} regularizes the representation to preserve structural information beyond a scalar score.

C. Reinforcement Learning with Feature-wise Counterfactual Advantages.

Why feature-wise counterfactual advantages matter. Existing RL-based feature selection methods often optimize policies using a single subset-level reward [13], [14], [24]. While effective at ranking subsets, such global signals suffer from a credit assignment problem as they provide little guidance on the specific contribution of each selected feature to the final reward, especially under complex feature interactions [28]. As a result, redundant features can be reinforced whenever they co-occur with informative ones, slowing down convergence and degrading subset compactness. To obtain a more informative learning signal, we construct feature-wise advantages that quantify the marginal contribution of each feature, inspired by counterfactual baseline mechanism [21]. Instead of treating the subset as an indivisible entity, we define a counterfactual intervention on the action space: we construct leave-one-out subsets by removing one selected feature at a time and compare their predicted rewards against the original subset. This design serves as a computationally tractable approximation of Shapley values [29], which provide a principled attribution by averaging marginal contributions over all coalitions but are prohibitively expensive to compute exactly. Furthermore, as exploration increasingly concentrates on higher-quality regions of the subsets, these counterfactual signals yield more task-relevant and lower-variance guidance for policy optimization.

Shared-parameter Actor. We use a single shared actor to produce individualized selection actions for each feature while remaining scalable to high-dimensional feature spaces. Specifically, each feature f_i is associated with a learnable embedding $\boldsymbol{\beta}_i$, initialized from its statistical descriptor to encode feature identity and distributional characteristics. Conditioned on the current subset state $\mathbf{s}$ and the feature embedding $\boldsymbol{\beta}_i$, the actor outputs a selection probability for f_i through the shared network. A binary action $m^i \in \{0, 1\}$ is then sampled for each feature according to its probability, and the actions are concatenated to form the new subset mask $\mathbf{m}$. Because the actor shares parameters across features while receiving different $\boldsymbol{\beta}_i$, it can produce distinct selection behaviors for different features under the same subset context.

Feature-wise Counterfactual Advantages. During the reinforcement learning stage, the actor samples a binary mask $\mathbf{m} = [m^1, \ldots, m^N]$ conditioned on the representation of the previous subset, where $m^i \in \{0, 1\}$ indicates whether feature f_i is selected. To obtain feature-wise learning signals, we construct leave-one-out counterfactual masks for selected features. Specifically, for each selected feature with $m^i = 1$, we define a counterfactual mask $\mathbf{m}^{-i}$ by setting $m^i = 0$ while keeping all other entries unchanged. We then encode the factual subset and counterfactual subsets into latent states $\mathbf{s}$ and $\mathbf{s}^{-i}$, respectively. Leveraging the reward-predictive critic, the feature-wise counterfactual advantage is computed as:

$$A_i = V_{\phi_c}(\mathbf{s}) - V_{\phi_c}(\mathbf{s}^{-i}). \tag{6}$$

This quantity measures how the predicted subset reward would change if feature f_i were removed from the current subset, under the same context defined by the remaining selected features. Consequently, these advantages provide more informative guidance for policy updates than a single subset-level reward, and can reflect context-dependent utility when feature interactions are present, helping the search favor compact, high-quality subsets.

Policy Optimization. To optimize the shared actor with the derived feature-wise counterfactual advantages, we adopt Proximal Policy Optimization (PPO) [30], which stabilizes policy-gradient updates via a clipped surrogate objective [31]. For each feature, we compute the ratio $\rho_i(\theta) = \frac{\pi_\theta(m^i|\mathbf{s},\boldsymbol{\beta}_i)}{\pi_{\theta_{\text{old}}}(m^i|\mathbf{s},\boldsymbol{\beta}_i)}$, and the corresponding clipped surrogate term

$$\ell_i(\theta) = \min(\rho_i(\theta)\, A_i,\ \text{clip}(\rho_i(\theta), 1-\epsilon, 1+\epsilon)\, A_i), \tag{7}$$

where ϵ is the clipping hyperparameter. The actor objective averages the surrogate terms over features and the sampled mini-batch:

$$J_{\text{actor}}(\theta) = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\ell_i(\theta)\right]. \tag{8}$$

Maximizing J_{actor} increases the probability of feature-selection actions with positive counterfactual advantages and decreases that of actions with negative advantages, yielding stable updates under stochastic subset sampling.

IV. Experiments

A. Experimental Setup

Datasets. We evaluate our method on 19 publicly available datasets from OpenML [32], UCI [33], LibSVM [34], the Feature Selection Benchmark [35], and the NCBI Gene Expression Omnibus [36], covering binary classification (C), multi-class classification (MC), and regression (R) tasks.

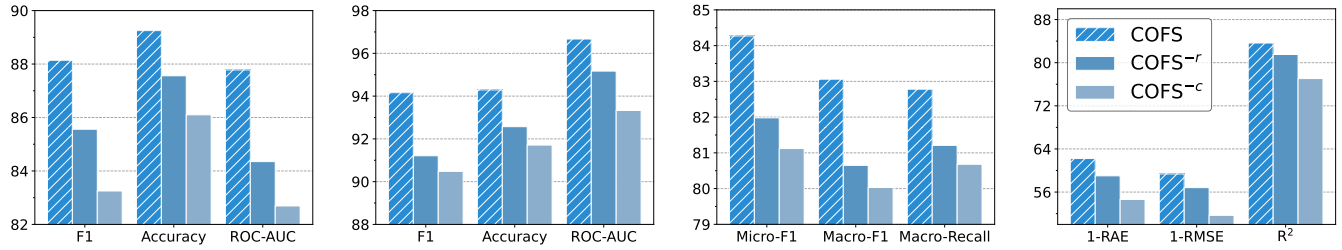
Evaluation Metrics. Following prior work [37], we evaluate the quality of selected feature subsets using Random Forest as a downstream ML model. For each dataset, we adopt a 20% hold-out validation split and perform five-fold cross-validation to reduce evaluation variance. We report F1-score, Accuracy, and ROC-AUC for binary classification; Micro-F1, Macro-F1, and Macro-Recall for multi-class classification; and R^2, 1-Root Mean Square Error (1-RMSE), and 1-Relative Absolute Error (1-RAE) [38] for regression tasks. For all metrics, higher values indicate better downstream performance.

Baselines. We compare our proposed method with nine representative feature selection algorithms, which can be broadly

TABLE I

OVERALL PERFORMANCE COMPARISON. WE REPORT F1-SCORE FOR CLASSIFICATION TASKS (C), MICRO-F1 FOR MULTI-CLASS CLASSIFICATION TASKS (MC), AND 1-RAE FOR REGRESSION TASKS (R). THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Dataset	Task	#Samples	#Features	Original	KBest	LASSO	LASSONet	SAFS	CompFS	RFE	MARLFS	GAINS	HRLFS	COFS
SpectF	C	267	44	75.96	81.41	75.96	75.96	82.36	83.69	80.80	79.16	80.55	87.56	$\mathbf{88.14}^{\pm 0.51}$
SVMGuide3	C	1243	21	77.81	76.38	77.64	76.44	77.65	74.41	78.53	75.92	77.53	78.90	$\mathbf{83.16}^{\pm 1.38}$
Credit Default	C	30000	25	80.19	79.87	77.86	79.87	80.29	73.90	80.17	80.11	80.05	80.56	$\mathbf{80.87}^{\pm 0.12}$
SpamBase	C	4601	57	90.68	91.58	90.73	89.63	88.05	89.70	91.59	90.04	91.71	92.60	$\mathbf{92.70}^{\pm 0.11}$
Ionosphere	C	351	34	91.33	92.74	89.92	94.14	87.21	91.08	95.69	90.18	92.24	94.27	$\mathbf{97.10}^{\pm 0.82}$
Darwin	C	174	451	79.76	85.35	82.86	79.76	82.49	85.59	88.35	85.48	83.76	87.59	$\mathbf{94.16}^{\pm 2.59}$
Scene	C	2407	299	79.34	82.03	79.60	82.15	78.18	80.97	81.55	79.70	82.19	81.59	$\mathbf{84.54}^{\pm 1.42}$
Mice-Protein	MC	1080	77	74.99	80.54	81.51	78.23	78.23	78.96	80.10	80.56	81.04	83.79	$\mathbf{84.27}^{\pm 0.84}$
Coil-20	MC	1440	400	96.53	96.88	96.54	94.80	96.53	95.16	97.57	96.19	97.22	97.56	$\mathbf{97.92}^{\pm 0.09}$
MNIST fashion	MC	10000	784	80.15	79.25	79.80	79.45	79.20	78.30	80.50	79.90	81.00	79.95	$\mathbf{82.20}^{\pm 0.66}$
Jannis	MC	83733	54	62.10	65.77	65.09	65.78	64.42	66.61	67.54	66.70	65.15	67.91	$\mathbf{70.16}^{\pm 1.21}$
Han	MC	2746	20670	76.36	76.18	73.55	72.57	75.79	73.20	79.27	74.69	74.03	81.45	$\mathbf{84.36}^{\pm 2.01}$
Handwritten	MC	5620	64	95.82	96.00	92.44	96.35	96.71	66.55	96.26	96.09	96.00	96.17	$\mathbf{96.89}^{\pm 0.23}$
Isolet	MC	7797	617	92.05	90.45	92.18	91.99	89.17	83.72	91.67	90.64	92.38	92.31	$\mathbf{92.82}^{\pm 1.13}$
Openml_586	R	1000	25	54.95	57.68	60.67	56.03	53.15	55.49	58.10	55.02	59.36	61.75	$\mathbf{62.01}^{\pm 0.16}$
Openml_607	R	1000	50	51.73	53.03	58.10	53.63	51.82	53.18	54.39	53.94	60.44	61.70	$\mathbf{62.22}^{\pm 0.42}$
Openml_618	R	1000	50	46.89	51.33	47.41	50.69	45.89	48.55	53.64	49.31	55.68	56.40	$\mathbf{58.37}^{\pm 0.93}$
Wave-Perth	R	36043	148	94.23	95.21	87.99	85.99	87.09	88.56	96.21	94.25	96.20	96.12	$\mathbf{96.23}^{\pm 0.17}$
CT-Image	R	53500	386	92.59	91.79	90.51	92.41	92.24	86.74	92.49	92.80	92.96	92.42	$\mathbf{93.00}^{\pm 0.07}$



(a) SpectF (b) Darwin (c) Mice-Protein (d) Openml_607

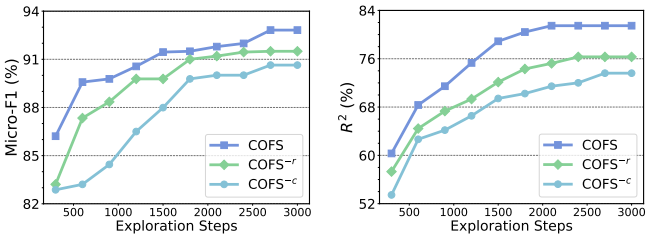
Fig. 3. The impact of removing reward-aware pretraining (COFS^{-r}) and feature-wise counterfactual advantages (COFS^{-c}) on performance.

categorized into three groups: (1) Filter methods, such as KBest [39], which select features based on statistical scores independently of downstream models. (2) Embedded methods, including LASSO [9], LASSONet [40], and SAFS [41], which integrate feature selection into model training through regularization. (3) Wrapper methods, which perform subset search by repeatedly training and evaluating a downstream model. This category includes CompFS [42], RFE [43], MARLFS [12], GAINS [37], and HRLFS [13], covering both search-based approaches and learning-based strategies such as reinforcement learning or gradient-guided optimization.

Implementation Details. To pretrain the encoder and critic, we collect 500 subset-reward pairs as training data. The encoder consists of a single ISAB layer with four attention heads. The dimensions of individual feature embeddings and the global subset state are set to 64 and 128, respectively. During the reinforcement learning stage, we perform 800 exploration steps in total and update the policy every 32 steps. Following [44], we employ Proximal Policy Optimization (PPO) with a learning rate of 3×10^{-4} for the actor and 1×10^{-3} for the critic. The clipping ratio is set to 0.2, and the reward trade-off hyperparameter λ is set to 0.9 to balance the performance and sparsity. All experiments are conducted on an Ubuntu 24.04.3 LTS system equipped with an Intel Xeon Platinum 8380 CPU and an NVIDIA A800 GPU, with the framework of Python 3.11.13 and PyTorch 2.1.0.

B. Experimental Results

Overall Performance. We conduct this experiment to evaluate the overall effectiveness of the proposed method. Table I presents the overall performance comparison between COFS and nine feature selection baselines, where all methods are evaluated under identical settings to ensure a fair comparison. The results show that COFS achieves consistently competitive performance across different domains and tasks. This observation may be attributed to three key design choices. The permutation-invariant set encoder captures feature interactions and yields higher-quality subset states for both the policy and the reward predictor. Reward-aware pretraining further aligns these states with downstream performance, providing a more reliable surrogate signal during optimization. Finally, computing feature-wise counterfactual advantages provides finer-grained learning signals than a single subset-level reward, enabling the policy to better distinguish informative features from redundant ones and to more effectively account for interaction effects under the current subset context.



(a) Isolet (b) Openml_618

Fig. 4. Best-so-far performance over exploration steps for COFS, COFS^{-r}(without reward-aware pretraining), and COFS^{-c}(without feature-wise counterfactual advantages).

TABLE II
ROBUSTNESS CHECK WITH DIFFERENT DOWNSTREAM ML MODELS ON JANNIS DATASET IN TERMS OF F1-SCORE.

	LightGBM	RF	SVM	DT	KNN
Raw	0.636	0.621	0.612	0.508	0.523
KBest	0.656	0.658	0.656	0.553	0.607
LASSO	0.652	0.651	0.654	0.548	0.616
SAFS	0.654	0.644	0.620	0.532	0.534
RFE	0.673	0.675	0.673	0.567	0.611
HRLFS	0.684	0.679	0.689	0.563	0.631
COFS	**0.696**	**0.702**	**0.702**	**0.581**	**0.664**

Ablation Study. This experiment is conducted to investigate the effectiveness of key components in the proposed method. We implement two variant models: (1) COFS^{-r} removes the reward-aware supervision during pretraining and trains the encoder using the reconstruction objective only; and (2) COFS^{-c} discards the feature-wise counterfactual advantages and instead uses the same subset-level reward to update the policy. We randomly select four datasets spanning different task types and report the final subset performance in Fig. 3. In addition, we examine the optimization behavior on Isolet and Openml_618 by plotting the best subset performance achieved so far against exploration steps in Fig. 4. The results indicate that the full model achieves better performance than the two variants. A plausible explanation is that reward-aware pretraining encourages the subset state to encode performance-relevant differences among subsets. In addition, using a single subset-level reward as the learning signal may provide limited guidance on the effect of individual feature decisions and making it harder to identify redundant features. Overall, these observations provide empirical evidence that both reward-aware pretraining and feature-wise counterfactual advantages contribute to effective subset search in COFS.

Study of Subset Size and Performance. We conduct this experiment to examine whether COFS can select compact feature subsets while maintaining strong predictive performance. We randomly select six datasets and, for each dataset, compare our method with the best baseline method identified in Table I. Figure 5 reports both the downstream performance and the relative subset size, measured as the ratio of selected features to the original feature set. We can observe that COFS achieves better performance while selecting smaller subsets across these datasets. A plausible explanation is that feature-wise counterfactual advantages provide more informative signals

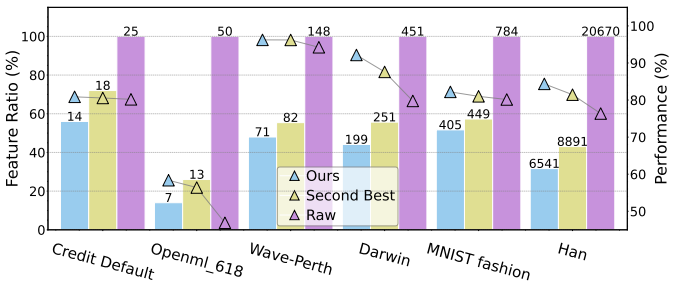
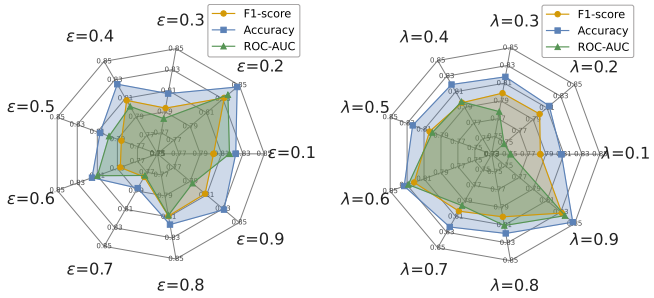


Fig. 5. Comparison of the selected subset size and performance between COFS (Ours), the best baseline on each dataset (Second Best), and using all features (Raw). Subset size is reported as the feature ratio relative to the original feature set.



(a) Clipping Ratio (b) Reward Trade-off

Fig. 6. Hyperparameter sensitivity test on SVMGuide3.

during policy updates, making it easier to identify redundant features that contribute little under the current subset context and to retain features that are useful through interactions. Overall, this experiment demonstrates that the subsets selected by COFS can improve predictive performance while reducing computational expense by using fewer features.

Study of Robustness. To evaluate whether the feature subsets selected by our proposed method remain effective under different downstream ML models, we conduct experiments on the Jannis dataset and compare against several baseline methods. For each method, the selected subset is evaluated by five downstream ML models (LightGBM, Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), and KNN), and Micro-F1 is reported as the performance metric. Table II summarizes the results and indicates that the subsets produced by COFS perform consistently well across these models. Overall, these results provide empirical evidence that COFS can identify feature subsets with relatively stable discriminative utility, leading to robust performance when the downstream model changes.

Study of Hyperparameters. This experiment is conducted to analyze the sensitivity of COFS to the clipping ratio ϵ and the reward trade-off λ. We vary both parameters from 0.1 to 0.9 and report the performance on SVMGuide3 using F1-score, Accuracy, and ROC-AUC. As shown in Fig. 6, the proposed method achieves relatively better performance when ϵ is set to 0.2 and λ is set to 0.9. According to [30], a clipping ratio of 0.2 generally leads to more stable policy updates, which is consistent with our empirical observations. The parameter λ balances predictive performance and subset

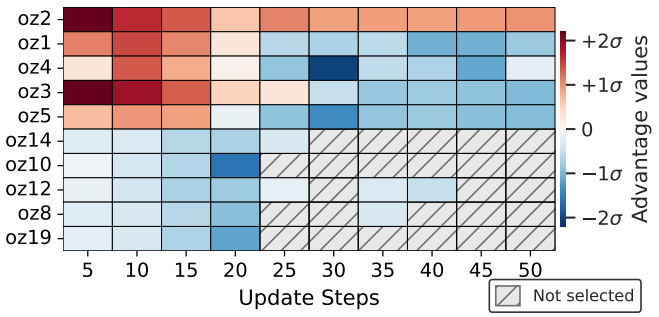


Fig. 7. Visualization of feature-wise counterfactual advantages during training on Openml_586. Colors indicate standardized advantage values at different policy update steps for the top-10 features ranked by advantage, and hatched cells indicate the feature is not selected within the corresponding interval.

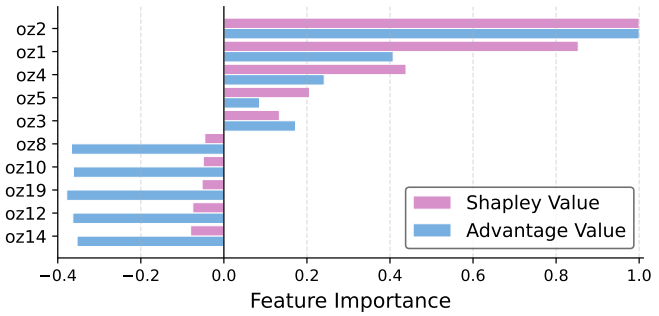


Fig. 8. Comparison between Shapley values and aggregated feature-wise counterfactual advantages on Openml_586 (computed on the top-10 features for tractability).

sparsity in the reward function. When λ is set to 0.1, excessive emphasis on subset compactness may result in overly small feature subsets, potentially affecting downstream predictive performance. Overall, this experiment provides empirical guidance on how these hyperparameters influence the performance and how to choose optimal values from different perspectives.

Case Study: Interpreting Feature-wise Counterfactual Advantages During Training. We conduct this case study to assess whether the proposed feature-wise counterfactual advantages provide informative signals for policy optimization. Experiments are performed on Openml_586, an artificially constructed dataset with 25 features, where the first five features are designed to be the optimal subset and the remaining ones are noise features. We perform 50 policy updates (one update every 32 interaction steps) and aggregate the standardized advantage values over each block of 5 consecutive updates for visualization. Figure 7 visualizes how feature-wise counterfactual advantages evolve during training. For clarity, we display the top-10 features ranked by their aggregated advantage scores from top to bottom. The heatmap indicates that the first five informative features tend to receive consistently higher advantage values, while noise features are gradually assigned lower values and become rarely selected in later stages of training, suggesting that the policy increasingly focuses on informative features. Figure 8 further compares the aggregated advantage scores with Shapley values [29] computed on the same top-10 features for tractability. The resulting rankings are broadly consistent: the informative features are ranked at the top by our aggregated advantages, and the overall ordering is similar to that induced by Shapley values. Overall, these observations provide empirical evidence that the proposed feature-wise counterfactual advantages offer discriminative feature-level supervision for policy optimization, steering learning toward compact, informative subsets.

V. Related Work

Feature selection aims to identify informative subsets to enhance predictive performance and efficiency. Traditional paradigms include filter methods [8], [39], which use statistical criteria; embedded methods [9], [40], [41], which integrate selection into training; and wrapper methods [42], [43], which iteratively search the best-performing subset. While filter and embedded approaches are efficient, they often overlook complex feature interactions. Conversely, wrapper methods account for interactions but suffer from high computational complexity in large combinatorial spaces. Recently, RL-based methods [12]–[14] have emerged to adaptively navigate the subset space. However, they typically rely on coarse, subset-level rewards, which obscure individual feature contributions and fail to provide fine-grained guidance when features exhibit strong redundancy or complementarity [17].

VI. Conclusion

In this paper, we propose a reinforcement learning framework for feature selection that provides feature-wise learning signals while retaining parameter sharing for per-feature decisions. Concretely, it encodes the current subset into a permutation-invariant representation and uses a shared actor conditioned on feature embeddings to output heterogeneous selection actions. To avoid repeated downstream evaluations, the framework trains a reward-predictive critic to estimate subset rewards and computes feature-wise counterfactual advantages by comparing the predicted reward of the current subset with that of the subset obtained by removing one selected feature. Moreover, we introduce a reward-aware pretraining stage that learns reward-aligned subset state representations rather than relying solely on the data distribution, thereby stabilizing subsequent policy optimization. Extensive experiments demonstrate the effectiveness of our proposed method in identifying compact, high-quality feature subsets.

References

[1] D. Wang, Y. Huang, W. Ying, H. Bai, N. Gong, X. Wang, S. Dong, T. Zhe, K. Liu, M. Xiao *et al.*, "Towards data-centric ai: A comprehensive survey of traditional, reinforcement, and generative approaches for tabular data transformation," *ACM Transactions on Knowledge Discovery from Data*, 2025.

[2] X. Ye, M. Xiao, Z. Ning, W. Dai, W. Cui, Y. Du, and Y. Zhou, "Needed: Introducing hierarchical transformer to eye diseases diagnosis," in *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. SIAM, 2023, pp. 667–675.

[3] X. Huang, M. Xiao, C. Qin, Q. Long, J. Chen, Y. Zhou, and H. Zhu, "Scihorizon-gene: Benchmarking llm for life sciences inference from gene knowledge to functional understanding," *arXiv preprint arXiv:2601.12805*, 2026.

[4] C. Qin, X. Chen, C. Wang, P. Wu, X. Chen, Y. Cheng, J. Zhao, M. Xiao, X. Dong, Q. Long *et al.*, "Scihorizon: Benchmarking ai-for-science readiness from scientific data to large language models," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 2025, pp. 5754–5765.

[5] M. Xiao, W. Zhang, X. Huang, H. Zhu, M. Wu, X. Li, and Y. Zhou, "Knowledge-guided gene panel selection for label-free single-cell rna-seq data: A reinforcement learning perspective," *IEEE Transactions on Computational Biology and Bioinformatics*, 2025.

[6] W. Lin, X. Zhao, Y. Wang, T. Xu, and X. Wu, "Adafs: Adaptive feature selection in deep recommender system," in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2022, pp. 3309–3317.

[7] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM computing surveys (CSUR)*, vol. 50, no. 6, pp. 1–45, 2017.

[8] N. Sánchez-Maroño, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter methods for feature selection–a comparative study," in *International conference on intelligent data engineering and automated learning*. Springer, 2007, pp. 178–187.

[9] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[10] N. El Aboudi and L. Benhlima, "Review on wrapper feature selection approaches," in *2016 international conference on engineering & MIS (ICEMIS)*. IEEE, 2016, pp. 1–5.

[11] K. Liu, Y. Fu, L. Wu, X. Li, C. Aggarwal, and H. Xiong, "Automated feature selection: A reinforcement learning perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 3, pp. 2272–2284, 2021.

[12] K. Liu, Y. Fu, P. Wang, L. Wu, R. Bo, and X. Li, "Automating feature subspace exploration via multi-agent reinforcement learning," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 207–215.

[13] W. Zhang, X. Huang, Y. Du, Z. Qiao, Q. Long, Z. Meng, Y. Zhou, and M. Xiao, "Comprehend, divide, and conquer: Feature subspace exploration via multi-agent hierarchical reinforcement learning," *arXiv preprint arXiv:2504.17356*, 2025.

[14] K. Liu, P. Wang, D. Wang, W. Du, D. O. Wu, and Y. Fu, "Efficient reinforced feature selection via early stopping traverse strategy," in *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2021, pp. 399–408.

[15] M. Xiao, D. Wang, M. Wu, Z. Qiao, P. Wang, K. Liu, Y. Zhou, and Y. Fu, "Traceable automatic feature transformation via cascading actor-critic agents," in *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. SIAM, 2023, pp. 775–783.

[16] D. Wang, M. Xiao, M. Wu, Y. Zhou, Y. Fu *et al.*, "Reinforcement-enhanced autoregressive feature transformation: Gradient-steered search in continuous space for postfix expressions," *Advances in Neural Information Processing Systems*, vol. 36, pp. 43 563–43 578, 2023.

[17] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[18] M. Xiao, D. Wang, M. Wu, K. Liu, H. Xiong, Y. Zhou, and Y. Fu, "Traceable group-wise self-optimizing feature transformation learning: A dual optimization perspective," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 4, pp. 1–22, 2024.

[19] X. Huang, D. Wang, Z. Ning, Z. Qiao, Q. Long, H. Zhu, Y. Du, M. Wu, Y. Zhou, and M. Xiao, "Collaborative multi-agent reinforcement learning for automated feature transformation with graph-driven path optimization," *arXiv preprint arXiv:2504.17355*, 2025.

[20] D. T. Nguyen, A. Kumar, and H. C. Lau, "Credit assignment for collective multiagent rl with global rewards," *Advances in neural information processing systems*, vol. 31, 2018.

[21] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[22] T. He, X. Huang, Y. Du, Q. Long, Z. Qiao, M. Wu, Y. Fu, Y. Zhou, and M. Xiao, "Fastft: Accelerating reinforced feature transformation via advanced exploration strategies," in *2025 IEEE 41st International Conference on Data Engineering (ICDE)*. IEEE Computer Society, 2025, pp. 4120–4133.

[23] C. Le Lan, S. Tu, A. Oberman, R. Agarwal, and M. G. Bellemare, "On the generalization of representations in reinforcement learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 4132–4157.

[24] W. Fan, K. Liu, H. Liu, Y. Ge, H. Xiong, and Y. Fu, "Interactive reinforcement learning for feature selection with decision tree in the loop," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 2, pp. 1624–1636, 2021.

[25] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[26] R. Liu, R. Xie, Z. Yao, Y. Fu, and D. Wang, "Continuous optimization for feature selection with permutation-invariant embedding and policy-guided search," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 2025, pp. 1857–1866.

[27] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 3744–3753.

[28] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of machine learning research*, vol. 5, no. Oct, pp. 1205–1224, 2004.

[29] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[30] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[31] Z. Tong, C. Qin, C. Fang, K. Yao, X. Chen, J. Zhang, C. Zhu, and H. Zhu, "From missteps to mastery: Enhancing low-resource dense retrieval through adaptive query generation," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 2025, pp. 1373–1384.

[32] Public, "Openml dataset download," [EB/OL], 2022, https://www.openml.org.

[33] Public, "Uci dataset download," [EB/OL], 2022, https://archive.ics.uci.edu/.

[34] L. Chih-Jen, "Libsvm dataset download," [EB/OL], 2022, https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.

[35] V. Cherepanova, R. Levin, G. Somepalli, J. Geiping, C. B. Bruss, A. G. Wilson, T. Goldstein, and M. Goldblum, "A performance-driven benchmark for feature selection in tabular deep learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 41 956–41 979, 2023.

[36] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: Ncbi gene expression and hybridization array data repository," *Nucleic acids research*, vol. 30, no. 1, pp. 207–210, 2002.

[37] M. Xiao, D. Wang, M. Wu, P. Wang, Y. Zhou, and Y. Fu, "Beyond discrete selection: Continuous embedding space optimization for generative feature selection," in *2023 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2023, pp. 688–697.

[38] D. Wang, Y. Fu, K. Liu, X. Li, and Y. Solihin, "Group-wise reinforcement feature generation for optimal and explainable representation space reconstruction," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 1826–1834.

[39] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the fourteenth international conference on machine learning*, 1997, pp. 412–420.

[40] I. Lemhadri, F. Ruan, and R. Tibshirani, "Lassonet: Neural networks with feature sparsity," in *International conference on artificial intelligence and statistics*. PMLR, 2021, pp. 10–18.

[41] T. Yasuda, M. Bateni, L. Chen, M. Fahrbach, G. Fu, and V. Mirrokni, "Sequential attention for feature selection," in *The Eleventh International Conference on Learning Representations*, 2025.

[42] F. Imrie, A. Norcliffe, P. Liò, and M. van der Schaar, "Composite feature selection using deep ensembles," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 142–36 160, 2022.

[43] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, "Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products," *Chemometrics and intelligent laboratory systems*, vol. 83, no. 2, pp. 83–90, 2006.

[44] Y. F. Wu, W. Zhang, P. Xu, and Q. Gu, "A finite-time analysis of two time-scale actor-critic methods," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 617–17 628, 2020.