Learning Local Feature Masks with Variational Information Bottleneck

Lu Sun[1,3*], **Jun Sakuma**[1,2]

[1]Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo 103-0027, Japan
[2]School of Computing, Institute of Science Tokyo, Tokyo 152-8550, Japan
[3]Unprecedented-scale Data Analytics Center, Tohoku University, Sendai 980-0845, Japan
lu.sun.e4@tohoku.ac.jp, sakuma@c.titech.ac.jp

Abstract

Instance-wise feature selection (IWFS) identifies informative features for each instance, improving generalization by discarding irrelevant information and enhancing interpretability through personalized explanations. Most IWFS methods adopt a selector–predictor architecture, where a selector generates instance-specific masks to guide prediction. This often leads to *co-adaptation*, in which the selector encodes label information into the mask, resulting in spurious correlations and unfaithful explanations. Existing methods also struggle to capture diverse local patterns, which is critical for IWFS under heterogeneous sparsity. We propose **VIBMask**, a unified IWFS framework by the variational information bottleneck. **VIBMask** mitigates co-adaptation by penalizing mutual information between unselected features and the label, and improves expressivity via an ensemble of diverse selectors that capture heterogeneous sparse patterns. We further derive a novel variational lower bound for discrete masks, enabling efficient end-to-end training through reparameterization. Experiments on synthetic and real datasets show that **VIBMask** consistently outperforms state-of-the-art IWFS methods in both predictive accuracy and informative feature discovery.

1 Introduction

Achieving both strong predictive performance and interpretability is a central goal in modern machine learning, particularly in high-stakes domains. Identifying informative features improves generalization while enabling practitioners to understand how inputs influence predictions [Molnar, 2020; Sokar *et al.*, 2022; Gorishniy *et al.*, 2021]. Instance-wise feature selection (IWFS) [Yoon *et al.*, 2018; Chen *et al.*, 2018; Jiang *et al.*, 2024; Oosterhuis *et al.*, 2025] addresses this challenge by selecting features adaptively for each instance, yielding fine-grained, context-aware explanations. IWFS is especially effective on heterogeneous datasets [Ucar *et al.*, 2021; Si *et al.*, 2024], where feature relevance varies across samples. For example, in electronic health records, the importance of lab tests or medications depends on patient age and medical history [Choi *et al.*, 2016; Ma *et al.*, 2018], while in e-commerce, fraud, and anomaly detection, feature informativeness depends on user behavior or transaction patterns [Yang *et al.*, 2022; Liu *et al.*, 2023].

Most IWFS methods adopt a data-dependent selector–predictor architecture [Chen *et al.*, 2018; Yoon *et al.*, 2018], where instance-specific masks guide prediction using only locally relevant features. This design enables individualized, interpretable explanations, and recent studies [Yang *et al.*, 2022; Oosterhuis *et al.*, 2025; Jiang *et al.*, 2024] show that such architectures can improve both interpretability and accuracy, challenging the presumed trade-off between the two. By adapting feature relevance per instance, IWFS offers a principled framework for transparent and reliable modeling.

Despite their empirical success, existing IWFS methods exhibit fundamental limitations. First, co-adaptation between the selector and predictor can undermine interpretability: the selector may encode label information directly into the instance-wise mask, allowing the predictor to decode the mask rather than infer the label from the selected features. This leakage under joint end-to-end training induces shortcut learning [Geirhos *et al.*, 2020], enabling high accuracy with semantically uninformative features (e.g., 96% MNIST accuracy from a single pixel [Chen *et al.*, 2018; Jethani *et al.*, 2021]), and resulting in unfaithful explanations that fail to generalize to independently trained predictors. Second, reliance on a single selector limits the ability to capture heterogeneous sparsity patterns, reducing diversity in local feature selection. Third, most methods are driven primarily by architectural design within the selector–predictor framework, lacking an information-theoretic perspective and a unified formulation that could strengthen both performance and interpretability. Recent advances partially address these issues: REAL-X [Jethani *et al.*, 2021] decouples training but sacrifices end-to-end explainability; ProtoGate [Jiang *et al.*, 2024] enforces global and local relevance but is constrained by a non-parametric predictor; and SUWR [Oosterhuis *et al.*, 2025] reduces shortcuts via sequential selection but incurs high cost and error propagation. These limitations motivate the need for a more principled and unified approach.

To address IWFS challenges, we propose **VIBMask**, which formulates instance-specific mask learning within the Variational Information Bottleneck (VIB) framework (Fig. 1). A

*Corresponding author.

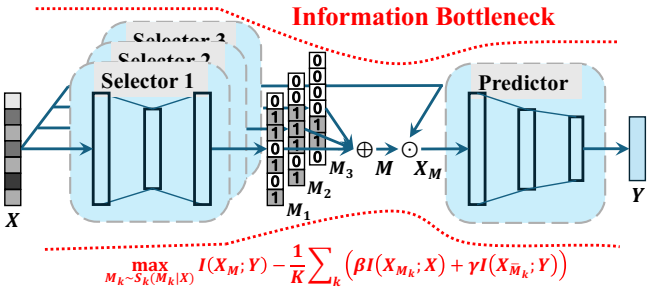

Figure 1: **VIBMask** performs IWFS by sampling K local masks $\{M_k\}_k$ from selectors $S_k(M_k \mid X)$ and aggregating them into a final mask $M = \frac{1}{K}\sum_k M_k$. Guided by the information bottleneck principle, it seeks to maximize $I(X_M; Y)$ while regularizing (i) *redundancy* via $\beta I(X_{M_k}; X)$ and (ii) *co-adaptation* via $\gamma I(X_{\overline{M}_k}; Y)$. The aggregated mask M thus emphasizes diverse and instance-relevant features for predicting Y.

core difficulty in joint training is co-adaptation: the selection mask can become a shortcut communication channel where the predictor decodes labels from the mask structure rather than the feature values, undermining interpretability. **VIBMask** mitigates this by penalizing mutual information between unselected features and the label. This enforces a necessary condition for faithfulness that masked features retain no predictive information, thereby collapsing mask-based shortcut solutions. Furthermore, **VIBMask** employs an ensemble of selectors to capture heterogeneous patterns and a unified VIB formulation for discrete masks, ensuring stable training via continuous reparameterization. Extensive experiments show consistent improvements in predictive performance and explanation quality over state-of-the-art IWFS methods.

Contributions. (1) We propose **VIBMask**, a unified IWFS framework grounded in the variational information bottleneck, which mitigates co-adaptation through information-theoretic regularization and captures diverse local features via a multi-selector mechanism. (2) We derive a novel variational lower bound for discrete masks, enabling efficient and stable optimization through reparameterization. (3) We achieve state-of-the-art performance with more informative feature selection on both synthetic and real datasets.

2 Related Work

Instance-Wise Feature Selection. IWFS aims to identify instance-specific relevant features, improving interpretability and generalization through context-aware explanations. Early models such as TabNet [Arik and Pfister, 2021] integrated selection and prediction end-to-end, while recent methods adopt a *selector–predictor* framework [Chen *et al.*, 2018; Yang *et al.*, 2022; Oosterhuis *et al.*, 2025], where a selector produces instance-dependent masks and a predictor operates on selected local features. Learning differentiable sparse masks remains challenging and is addressed via Gumbel-softmax [Chen *et al.*, 2018; Masoomi *et al.*, 2020], reinforcement learning [Yoon *et al.*, 2018], stochastic gates [Yamada *et al.*, 2020; Yang *et al.*, 2022], influence-aware attention [Liu *et al.*, 2023], and Gaussian copulas [Peng *et al.*, 2023]. However, many approaches treat feature selection independently, making them vulnerable to spurious correlations from degenerate selector–predictor interactions and limited expressivity under heterogeneous sparsity. In contrast, **VIBMask** reformulates IWFS within the VIB framework, mitigating spurious correlations by constraining co-adaptation and addressing heterogeneous sparsity through an ensemble of diverse selectors, resulting in robust and faithful instance-wise explanations.

Mitigating Co-Adaptation. Co-adaptation in IWFS arises when the selector and predictor jointly encode and decode label information through the selection mask, yielding shortcut solutions with high accuracy but unfaithful explanations. In this regime, the selected features are predictive only for the coupled predictor and fail to transfer to external classifiers, violating a core requirement of feature selection. Existing methods mitigate co-adaptation primarily through architectural or training decoupling: REAL-X [Jethani *et al.*, 2021] employs an amortized explanation model, ProtoGate [Jiang *et al.*, 2024] relies on a separately trained prototype-based predictor, and SUWR [Oosterhuis *et al.*, 2025] limits shortcuts via sequential greedy selection. In contrast, VIBMask addresses co-adaptation at the objective level by enforcing a necessary condition for faithful selection: unselected features must not retain predictive information about the target. This is implemented by penalizing the mutual information between those features and the label. This principled constraint suppresses label leakage without modifying the predictor, promoting transferable and semantically meaningful explanations.

VIB-Based Feature Selection. The VIB principle [Barber and Agakov, 2004; Mohamed and Jimenez Rezende, 2015] promotes compact yet predictive representations by minimizing input–latent mutual information while preserving target relevance, making it well suited for feature selection. Existing VIB-inspired methods, including Concrete Autoencoder [Balın *et al.*, 2019], Gated IB [Alesiani *et al.*, 2023], and Drop-Bottleneck [Kim *et al.*, 2021], mainly focus on *global* feature selection with a shared subset across samples. IWFS introduces additional challenges, including severe selector–predictor co-adaptation and limited expressivity under per-instance sparsity constraints. InterpreTabNet [Si *et al.*, 2024] extends TabNet [Arik and Pfister, 2021] with variational sparsity but neither explicitly mitigates co-adaptation nor avoids high-variance relaxations. To address IWFS, **VIBMask** introduces a variational objective that enforces feature faithfulness by ensuring unselected features carry no predictive information. By incorporating a decorrelation regularizer, it simultaneously prevents selector–predictor co-adaptation and enables stable, reparameterized optimization.

3 Methodology

3.1 An Information-Theoretic View of IWFS

The Information Bottleneck (IB) principle [Tishby *et al.*, 2000] seeks a latent representation Z of input $X \in \mathbb{R}^d$ that is informative about the target $Y \in \{1, \ldots, c\}$ while discarding irrelevant details. It is formalized as $\max_Z I(Y; Z) - \beta I(X; Z)$, where the trade-off parameter β balances relevance and compression, and mutual information $I(\cdot; \cdot)$ quantifies uncertainty reduction. The variational IB (VIB) [Alemi *et al.*, 2017]

approximates this objective using variational bounds, reparameterization [Kingma *et al.*, 2014], and Monte Carlo sampling [Kalos and Whitlock, 2009]. IB has been applied to improve robustness and generalization, including adversarial learning [Kuang *et al.*, 2023] and out-of-distribution generalization [Ahuja *et al.*, 2021].

We reinterpret IWFS through the IB lens to ground our method, **VIBMask**. Unlike standard VIB with continuous latents, IWFS selects discrete feature subsets per instance. We model this using a binary mask $M \in \{0,1\}^d$, where $M_i = 1$ indicates that feature X_i is selected. The masked input X_M is drawn from a distribution $S(M \mid X)$, and the objective balances task relevance and redundancy with the full input:

$$\max_S I(X_M;Y) - \beta I(X_M;X) \quad \text{s.t. } X_M = X \odot M,\ M \sim S(M \mid X), \tag{1}$$

where $\odot$ denotes the element-wise product. The mutual information is taken in expectation over masks sampled from $S(M \mid X)$, making Eq. (1) a *discrete, conditional* variant of the IB, with X_M representing a hard-selected feature subset. However, the discreteness of M hinders gradient-based optimization. To overcome this, we derive a variational lower bound for discrete latents and introduce a differentiable relaxation of the selection process, enabling efficient end-to-end training while preserving sparsity and interpretability.

3.2 Mitigating Co-Adaptation in IWFS

Recent IWFS methods [Jiang *et al.*, 2024; Jethani *et al.*, 2021] show that directly optimizing Eq. (1) can induce co-adaptation, a degenerate regime where the selector encodes label information into the mask M. The predictor "decodes" the label from the mask rather than the feature values, creating shortcut solutions that undermine interpretability. Crucially, this occurs because task-relevant information is shifted to the mask, leaving the discarded features to retain a residual predictive signal. Consequently, co-adaptation violates a core requirement of faithful selection: once features are masked, the unselected subset must have zero mutual information with the label.

To mitigate this effect, we explicitly penalize the mutual information between the *unselected* features and the label. Thus, **VIBMask** eliminates shortcut solutions in which the mask acts as a label-encoding channel and forces the selector to surface features that are intrinsically informative for prediction. Concretely, we augment Eq. (1) with a decorrelation penalty:

$$\max_S \quad I(X_M;Y) - \beta I(X_M;X) - \gamma I(X_{\overline{M}};Y) \quad \text{s.t.} \quad X_M = X \odot M, \quad M \sim S(M \mid X), \tag{2}$$

where $\overline{M} = 1 - M$ denotes the complement mask and $X_{\overline{M}}$ the unselected features. The hyperparameter $\gamma \geq 0$ controls the strength of the decorrelation loss. When informative features are omitted or label information is encoded in the mask, $I(X_{\overline{M}};Y)$ increases, penalizing such degenerate solutions and encouraging faithful, feature-based prediction.

Many IWFS methods, such as L2X [Chen *et al.*, 2018], STG [Yamada *et al.*, 2020], LSPIN [Yang *et al.*, 2022], and c-STG [Sristi *et al.*, 2024], can be interpreted as special cases of our discrete IB framework. Further details are provided in Sec. 4. However, unlike **VIBMask**, these methods neither target maximal compression of informative features nor explicitly mitigate co-adaptation, limiting their ability to provide compact and faithful explanations.

3.3 Integrating Multiple Selectors

While **VIBMask** mitigates co-adaptation via variational masking, it remains limited in capturing diverse sparsity patterns. To address this, we introduce *multi-selector integration*, which enhances both expressiveness and the ability to model heterogeneous feature relevance.

Specifically, we extend Eq. (2) by incorporating K selectors $\{S_1, \ldots, S_K\}$, each producing a local mask $M_k \sim S_k(M_k \mid X)$ for $k \in [K]$. The final mask is formed by averaging, $M = \frac{1}{K}\sum_k M_k$, and applied to the input to obtain $X \odot M$ for prediction. This design offers three advantages: (i) diverse selectors capture complementary data aspects, (ii) averaging improves robustness when individual selectors underperform, and (iii) architectural flexibility allows selectors to vary in width, depth, or network type. Accordingly, combining multiple selectors with (2) yields the final **VIBMask** objective:

$$\max_S I(X_M;Y) - \frac{1}{K}\sum_k \Big(\beta I(X_{M_k};X) + \gamma I(X_{\overline{M}_k};Y)\Big) \quad \text{s.t. } M = \frac{1}{K}\sum_k M_k,\ M_k \sim S_k(M_k \mid X). \tag{3}$$

To encourage selector diversity and reduce redundancy, several strategies are possible: (i) heterogeneous selector architectures with different capacities; (ii) explicit regularization that penalizes mutual information or enforces orthogonality among $\{M_k\}$; and (iii) bootstrap aggregating (bagging) [Breiman, 1996] to train selectors on different resampled datasets. For simplicity and to avoid extra hyperparameters, we adopt bagging in practice. Each selector is trained on a distinct bootstrap sample containing approximately 63.2% unique data points, naturally inducing diversity. By leveraging this ensemble of input-conditioned selectors, **VIBMask** enables more effective, scalable, and flexible instance-wise feature masking.

3.4 Variational Lower Bound of VIBMask

The objective in (3) is intractable due to: (i) the inaccessibility of the posterior $p(Y|X_M)$, and (ii) the exponential search space (over 2^d subsets) of binary masks $M \in \{0,1\}^d$. To address this, we derive a variational lower bound (this subsection) and apply a continuous relaxation via reparameterization in Sec. 4. Variational bounds for each term are given in Lemmas 1–3, with the full bound in Theorem 1. The proofs are provided in Sec. A2 of the technical appendix.

Lemma 1. *For a variational approximation $q(Y|X_M)$ of the true posterior $p(Y|X_M)$, the mutual information $I(Y;X_M)$ in* (2) *satisfies:*

$$I(Y;X_M) \geq \mathbb{E}_{p(Y,X_M)}[\log q(Y|X_M)] - \mathbb{E}_{p(Y)}[\log p(Y)].$$

The last term is constant w.r.t. the mask M and can be ignored during optimization.

Lemma 2. *Let $\pi_{k,i} = \mathbb{P}(M_{k,i} = 1 \mid X)$ denote the conditional probability that the k-th selector ($k \in [K]$) chooses*

feature X_i given input X. Then, for every input X, we have

$$\frac{\beta}{K}\sum_{k=1}^{K} I(X_{M_k};X) \;\;\leq\;\; \frac{\beta}{K}\sum_{k=1}^{K}\sum_{i=1}^{d} H(X_i)\,\pi_{k,i},$$

where $H(X_i)$ is the entropy of the input feature X_i.

Lemma 3. *For every input X, we have*

$$\frac{\gamma}{K}\sum_{k=1}^{K} I\big(Y;X_{\overline{M}_k}\big) \;\;\leq\;\; \frac{\gamma}{K}\sum_{k=1}^{K}\sum_{i=1}^{d} I(X_i;Y)\big(1-\pi_{k,i}\big),$$

where $\pi_{k,i}=\mathbb{P}(M_{k,i}=1\mid X)$ is the conditional selection probability of feature X_i by selector S_k, and $I(X_i;Y)$ denotes the mutual information between feature X_i and the label Y.

The lower bound in Lemma 1 follows from the non-negativity of the KL divergence. The upper bounds in Lemmas 2 and 3 are derived using the chain rule of mutual information under mild conditional independence assumptions shown in Sec. A2 of the technical appendix.

Theorem 1. *By combining the bounds from Lemmas 1–3 and discarding terms that are independent of the selection probabilities, we obtain the following variational lower bound for VIBMask in* (3)*:*

$$\mathcal{L}(q,\pi)=\mathbb{E}_{X,Y}\,\mathbb{E}_{M|X}\Big[\log q(Y\mid X_M)-\tfrac{1}{K}\sum_{k,i}\omega_i\pi_{k,i}\Big],$$

where $M=\frac{1}{K}\sum_k M_k$ and $\omega_i=\beta H(X_i)-\gamma I(X_i;Y)$. For the k-th selector ($k\in[K]$), each feature i is independently retained with probability $\pi_{k,i}\in[0,1]$, which depends on the input X.

Building on Theorem 1, our method is inspired by the IB principle [Balın *et al.*, 2019; Kim *et al.*, 2021; Alesiani *et al.*, 2023] but differs in three key aspects: (i) it enables *instance-wise feature selection* for personalized, sample-level interpretability; (ii) it explicitly penalizes mutual information between unselected features and the target, *mitigating co-adaptation*; and (iii) it integrates multiple selectors to capture *diverse local feature relevance*.

4 Optimization

Theorem 1 replaces the intractable posterior $p(Y\mid X_M)$ with a variational approximation $q(Y\mid X_M)$ and relaxes the binary mask $M\in\{0,1\}^d$ to a continuous vector π, enabling gradient-based optimization while preserving a probabilistic notion of feature relevance. We parameterize π with a selector $g_\psi(X)$ and q with a predictor $f_\theta(X_M)$, and optimize the bound $\mathcal{L}(q,\pi)$ over ψ and θ.

Continuous Relaxation. Directly optimizing the objective in Theorem 1 is challenging due to the discreteness of the binary mask M, which leads to high-variance gradients [Mnih and Rezende, 2016]. Although estimators such as REINFORCE [Williams, 1992], soft attention [Bahdanau, 2014], and hard concrete [Louizos *et al.*, 2018] alleviate this issue, they remain unstable for feature selection [Yamada *et al.*, 2020]. We adopt a continuous relaxation inspired by stochastic gates [Yamada *et al.*, 2020], approximating M_k with a relaxed Bernoulli vector $\widetilde{M}_k\in[0,1]^d$:

$$\widetilde{M}_k=\texttt{clip}(\mu_k+\epsilon,0,1),\;\mu_k=g_{\psi_k}(X),\;\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I}),$$

where μ_k is generated by the k-th selector network g_{ψ_k} and $\texttt{clip}(x,0,1)=\max(0,\min(1,x))$. This enables the reparameterization trick [Figurnov *et al.*, 2018], reducing gradient variance. The regularizer of $\mathcal{L}(q,\pi)$ in Theorem 1 becomes:

$$\sum_i \omega_i\pi_{k,i}=\sum_i \omega_i\mathbb{P}(\widetilde{M}_{k,i}>0)=\sum_i \omega_i\Phi\left(\frac{\mu_{k,i}}{\sigma}\right),$$

where Φ is the standard Gaussian CDF. In practice, we estimate of $H(X_i)$ and $I(X_i;Y)$ through discretizations, as detailed in the supplementary materials. A detailed derivation and choice of σ are discussed in Sec. A4 and Fig. A1 of the technical appendix.

Optimization Algorithm. Based on the continuous relaxation and the lower bound in Theorem 1, the differentiable objective of **VIBMask** becomes:

$$\min_{\theta,\psi}\mathbb{E}_{X,Y}\mathbb{E}_{\widetilde{M}|X}\Big[L(f_\theta(X_{\widetilde{M}}),Y)+\tfrac{1}{K}\sum_{k,i}\omega_i\Phi\left(\frac{\mu_{k,i}}{\sigma}\right)\Big]$$

$$\text{s.t. } X_{\widetilde{M}}=X\odot\widetilde{M},\;\widetilde{M}=\tfrac{1}{K}\sum_k\widetilde{M}_k,\;\mu_k=g_{\psi_k}(X),$$

$$\widetilde{M}_k=\texttt{clip}(\mu_k+\epsilon,0,1),\;\epsilon_i\sim\mathcal{N}(0,\sigma^2\mathbf{I}),\qquad(4)$$

where L is the loss function and f_θ denotes the predictor network parameterized by θ. Gradients w.r.t. θ are computed directly, while those w.r.t. ψ_k are estimated via Monte Carlo [Kalos and Whitlock, 2009]:

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{d}\left[\frac{\partial L}{\partial\widetilde{M}_{k,i}^{(n)}}\frac{\partial\widetilde{M}_{k,i}^{(n)}}{\partial\mu_{k,i}}+\frac{\omega_i}{K}\frac{\partial\Phi(\mu_{k,i}/\sigma)}{\partial\mu_{k,i}}\right]\frac{\partial\mu_{k,i}}{\partial\psi_k},$$

where N denotes the number of instances. Therefore, the entire **VIBMask** model is trained end-to-end using gradient descent with the continuous reparameterization trick. A detailed algorithmic procedure and complexity analysis are provided in Sec. A5 of the technical appendix.

Inference. At inference, we remove stochasticity by setting $\epsilon=0$, which yields a deterministic mask $\widetilde{M}_k=\texttt{clip}(\mu_k,0,1)$. Most elements of $\widetilde{M}_k$ converge to binary values. We avoid rounding intermediate values—often arising under weak training signals (e.g., overlapping classes)—since rounding may distort feature magnitudes and degrade performance. For interpretability, we threshold each M_k at 0.5 and apply majority voting across the K local masks $\{M_k\}$ to identify the selected instance-wise features.

Remarks. Notably, our **VIBMask** formulation in (4) generalizes several prior IWFS methods as special cases when restricted to a single selector ($K=1$). Specifically, L2X [Chen *et al.*, 2018] corresponds to $\gamma,\beta=0$; STG [Yamada *et al.*, 2020] to constant $\gamma,\lambda=0$ with π independent of X; and both LSPIN [Yang *et al.*, 2022] and c-STG [Sristi *et al.*, 2024] to constant $\gamma,\lambda=0$. This highlights that **VIBMask** offers strong generalization capability and flexibility, making it applicable across a wide range of IWFS settings. A detailed comparison with existing IWFS approaches is presented in Table A1 in the technical appendix.

5 Experiments

We evaluate **VIBMask** to address three key questions: **Q1 Interpretability**—Can it identify informative features, mitigate co-adaptation, and enhance prediction? **Q2 Performance**—Does it achieve efficient, competitive results in prediction and feature selection vs. SOTA IWFS methods? **Q3 Ablation**—What is the contribution of each component (co-adaptation avoidance, multi-selector integration) to overall performance? To this end, we first assess **VIBMask**'s interpretability on six synthetic datasets (Sec. 5.1), then compare it with sixteen methods on eight real-world and two MNIST datasets (Sec. 5.2). Finally, we analyze the impact of each component of **VIBMask** (Sec. 5.3).

Datasets. We use six benchmark *synthetic* datasets [Chen *et al.*, 2018; Yoon *et al.*, 2018; Yang *et al.*, 2022; Jiang *et al.*, 2024] with 11-dimensional Gaussian features ($X \sim \mathcal{N}(0, I)$) and Bernoulli labels generated by varying logit(X). Each has 5,000 training and 5,000 testing samples. For *real-world* evaluation, we use ten popular datasets [Yang *et al.*, 2022; Jiang *et al.*, 2024]: six from scikit-feature (**Relathe**, **PCMAC**, **Colon**, **Tox-171**, **Lung**, **Prostate**), two from TabZilla (**Cnae-9**, **Mfeat-Fourier**), and **MNIST** [Deng, 2012] plus **Fashion-MNIST** [Xiao *et al.*, 2017]. MNIST and Fashion-MNIST each contain 70,000 grayscale 28×28 images in 10 classes; we use 6,000 training and 1,000 testing samples.

Comparison methods. We compare **VIBMask**[1] with sixteen methods: four *baselines* (**Ridge regression** [Cox, 1958], **KNN** [Fix, 1985], **SVM** [Cortes, 1995], **MLP** [Gorishniy *et al.*, 2021]), four *global Feature Selection (FS)* methods (**Lasso** [Tibshirani, 1996], **Random Forest** [Breiman, 2001], **XGBoost** [Chen and Guestrin, 2016], **STG** [Yamada *et al.*, 2020]), and eight *local (instance-wise) FS* methods (**INVASE** [Yoon *et al.*, 2018], **L2X** [Chen *et al.*, 2018], **LSPIN**, **LLSPIN** [Yang *et al.*, 2022], **ProtoGate** [Jiang *et al.*, 2024], **SUWR** [Oosterhuis *et al.*, 2025], **InterpreTabNet** [Si *et al.*, 2024], **REAL-X** [Jethani *et al.*, 2021]).

Configuration. We perform 5-fold cross-validation and report the average results over 5 independent runs. Within each training fold, 10% of the samples are held out for validation. Hyperparameters are selected based on aggregated validation performance across all folds. Additional details on datasets, experimental setup, and complementary results are provided in Sec. A6 of the technical appendix.

5.1 Q1: Experiments on Synthetic Datasets

We evaluate interpretability on six synthetic datasets (Syn1–Syn6) with known ground-truth informative features, which enable a precise assessment of *faithfulness*—whether the model consistently selects the features that are causally responsible for the label. Each dataset has 11 standard-Gaussian features and a Bernoulli label whose logit is a temperature-scaled nonlinear function of a designated feature subset; Syn4–Syn6 are conditional, with X_{11} as a control-flow switch. Full setup and per-dataset best configurations are in Secs. A6.1.1 and A6.3.6 of the technical appendix. Table 1 reports TPR↑, FDR↓, and Accuracy↑; for conditional datasets we additionally report the Control-Flow Selection Rate (CFSR↑) [Oosterhuis *et al.*, 2025], the fraction of test samples whose mask includes X_{11}. A low CFSR with high TPR is the diagnostic signature of co-adaptation.

Across all six datasets, **VIBMask** attains the highest accuracy and is at least tied for the lowest FDR. On the unconditional datasets (Syn1–Syn3) it achieves perfect feature recovery (TPR $= 100.0$, FDR $= 0.0$), with the largest accuracy gains on the harder distributions: $+5.0$ pp on Syn3 (0.950 vs. SUWR 0.903) and Syn6 (0.951 vs. SUWR 0.896), and $+1.1$ and $+2.7$ pp on Syn4 and Syn5. On the conditional datasets Syn4 and Syn6, **VIBMask** selects the control-flow feature on every (Syn4) or nearly every (Syn6, 99.9%) instance, confirming that the decorrelation objective surfaces the gating variable rather than routing label information through the mask. The remaining trade-offs are local: on Syn5 the CFSR and TPR trail REAL-X/SUWR by $\sim$8 pp while accuracy still leads, and on Syn6 a higher FDR (15.0) accompanies the largest accuracy jump in the table. Importantly, all top-3 CFSR slots on every conditional dataset are occupied by methods with explicit co-adaptation mitigation (REAL-X, ProtoGate, SUWR, VIBMask), validating that closing the mask shortcut is the operative mechanism behind faithful conditional-dependency capture across all six benchmarks.

5.2 Q2: Experiments on Real-World Datasets

Classification Accuracy. Table 2 reports balanced accuracy on the eight real-world datasets. **VIBMask** ranks first on Lung (95.34%, $+1.90$ pp over the strongest IWFS baseline), Relathe (89.54%, $+4.87$ pp over SUWR), and PCMAC (89.43%, $+1.54$ pp over ProtoGate), second on Cnae-9 (96.83%), and third on Colon and Prostate; on TOX-171 it lands within 0.5 pp of the best IWFS baseline. The only column with a meaningful gap is Mfeat-Fourier (87.73%), where the strongest method, SUWR, retains a 4.8 pp lead. Aggregated across the eight datasets, **VIBMask** attains the highest mean balanced accuracy (90.41% vs. 84.32% for the strongest published IWFS), confirming the benefit of unifying co-adaptation avoidance and multi-selector integration within a single information-bottleneck objective.

Two patterns are worth highlighting. First, **VIBMask** dominates on the two text-classification datasets (Relathe $+8.93$ pp, PCMAC $+4.32$ pp over the previous IWFS best), where the relevant features differ markedly across documents and an instance-adaptive mask is essential. Second, on the HDLSS biomedical datasets (Colon, Lung, Prostate, TOX-171) the spread among the top IWFS methods is small (≤ 2 pp), reflecting both the high variance of small test folds and the fact that several methods already operate near the Bayes-optimal regime. In this regime **VIBMask** remains competitive while still delivering compact and instance-adaptive masks. Competing IWFS approaches share related goals but lack a unified information-theoretic design: ProtoGate extends global to local masking without explicit information-bottleneck regularization; SUWR and REAL-X encourage diversity or conditional importance but do not directly mitigate selector–predictor co-adaptation; L2X and LSPIN can be viewed as restricted variants of **VIBMask**'s objective.

[1] Code and technical appendix: https://github.com/futuresun912/VIBMask.

Methods	Syn1			Syn2			Syn3			Syn4				Syn5				Syn6			
	TPR↑	FDR↓	ACC↑	TPR↑	FDR↓	ACC↑	TPR↑	FDR↓	ACC↑	CFSR↑	TPR↑	FDR↓	ACC↑	CFSR↑	TPR↑	FDR↓	ACC↑	CFSR↑	TPR↑	FDR↓	ACC↑
w/o FS	100.0	80.4	0.677	100.0	62.4	0.779	100.0	64.3	0.855	100.0	100.0	63.2	0.540	100.0	100.0	64.1	0.646	100.0	100.0	54.9	0.671
Oracle	100.0	0.0	0.789	100.0	0.0	0.885	100.0	0.0	0.900	100.0	100.0	0.0	0.814	100.0	100.0	0.0	0.822	100.0	100.0	0.0	0.890
L2X	65.7	29.6	0.770	88.1	23.2	0.861	83.9	12.9	0.878	58.2	83.2	31.0	0.770	50.5	87.6	40.0	0.780	36.0	81.4	22.1	0.857
INVASE	**100.0**	*4.6*	0.792	**100.0**	*0.8*	0.880	95.4	0.1	0.872	54.6	92.5	17.0	0.815	37.5	79.9	13.7	0.784	58.3	86.6	8.2	0.867
REAL-X	**100.0**	24.8	0.763	**100.0**	18.2	0.792	**100.0**	7.7	0.883	98.5	**99.5**	36.6	0.761	**100.0**	*98.2*	49.4	0.777	*99.5*	97.0	8.5	0.847
LSPIN	90.4	24.5	0.770	95.9	4.4	0.852	96.2	8.6	0.894	55.9	95.5	33.2	0.777	62.4	96.4	31.6	0.802	54.0	95.5	4.8	0.863
LLSPIN	*92.0*	22.0	0.780	96.5	3.0	0.860	97.0	7.0	0.898	67.5	96.0	30.0	0.780	56.8	97.0	30.0	0.810	58.2	96.5	*4.0*	0.870
IntTabNet	91.1	17.6	0.767	98.5	8.0	0.885	*97.1*	10.1	0.903	*98.0*	94.5	26.2	0.789	**100.0**	94.9	25.7	0.791	*99.5*	*97.6*	5.4	0.870
ProtoGate	99.0	5.0	0.805	97.8	1.5	*0.890*	98.5	*5.0*	*0.902*	97.8	*98.0*	17.0	*0.812*	99.0	98.5	*6.0*	0.820	98.7	98.2	1.0	*0.895*
SUWR	**100.0**	3.4	*0.803*	*98.1*	0.4	0.893	**100.0**	**0.0**	0.903	**100.0**	*98.0*	*20.0*	0.810	**100.0**	**99.0**	2.3	*0.816*	**100.0**	97.4	**0.4**	0.896
VIBMask	**100.0**	**0.0**	**0.809**	**100.0**	**0.0**	**0.903**	**100.0**	**0.0**	**0.950**	**100.0**	99.1	**4.5**	**0.831**	*91.3*	91.6	**2.1**	**0.852**	99.9	**99.4**	15.0	**0.951**

Table 1: Performance on six synthetic datasets. For metrics with ↑/↓, higher/lower values are better. **First**, Second, and *Third* rankings (excluding Oracle and w/o FS) are marked. **VIBMask** achieves the highest accuracy on all six datasets and the lowest FDR on five out of six. Detailed setup, logit functions, and per-dataset best configurations are in Secs. A6.1.1 and A6.3.6 of the technical appendix.

	Methods	Colon	Lung	Mfeat	Cnae-9	Prostate	Relathe	Tox-171	PCMAC
Baseline	Ridge	76.57 ± 12.67	92.94 ± 6.29	89.33 ± 6.52	90.62 ± 6.59	86.99 ± 9.34	75.43 ± 3.40	*91.67 ± 5.99*	68.33 ± 5.30
	SVM	72.11 ± 10.22	72.51 ± 9.50	90.43 ± 2.47	78.96 ± 5.45	86.15 ± 6.34	77.13 ± 3.22	66.55 ± 6.48	69.29 ± 4.52
	KNN	71.53 ± 11.51	90.64 ± 6.20	84.37 ± 5.43	82.79 ± 7.95	78.70 ± 9.25	75.23 ± 4.33	83.82 ± 7.53	72.39 ± 3.83
	MLP	80.14 ± 9.70	95.25 ± 3.99	87.81 ± 4.50	96.17 ± 2.59	86.83 ± 7.66	80.13 ± 2.34	85.77 ± 4.55	61.10 ± 6.38
Global FS	Lasso	78.83 ± 11.20	*94.38 ± 4.21*	88.08 ± 4.67	93.60 ± 2.84	**91.10 ± 5.72**	77.36 ± 3.01	90.56 ± 6.80	69.10 ± 5.50
	RF	80.02 ± 10.35	91.70 ± 6.73	83.98 ± 2.12	90.36 ± 4.75	90.67 ± 6.90	75.70 ± 4.48	80.03 ± 7.03	75.00 ± 6.50
	XGBoost	74.45 ± 11.85	86.75 ± 7.41	89.44 ± 3.05	92.65 ± 5.40	81.69 ± 9.02	78.93 ± 2.37	70.09 ± 7.46	84.90 ± 2.42
	STG	80.29 ± 9.25	93.50 ± 6.19	88.15 ± 5.46	76.13 ± 8.19	89.99 ± 5.15	79.66 ± 2.06	87.77 ± 5.08	85.95 ± 3.11
Local FS (IWFS)	L2X	59.10 ± 15.21	52.70 ± 12.42	74.27 ± 9.40	78.35 ± 11.47	63.40 ± 10.90	52.30 ± 6.29	44.56 ± 12.83	77.99 ± 5.18
	INVASE	65.47 ± 10.15	91.45 ± 6.00	81.48 ± 7.35	91.70 ± 6.84	80.62 ± 7.65	80.94 ± 2.83	80.25 ± 6.83	77.90 ± 4.44
	REAL-X	78.33 ± 11.11	93.12 ± 4.61	91.51 ± 3.52	95.59 ± 3.04	86.85 ± 7.84	*84.43 ± 2.37*	90.99 ± 4.25	82.00 ± 4.37
	LSPIN	81.31 ± 8.28	77.34 ± 9.58	83.45 ± 7.50	**97.18 ± 3.16**	87.34 ± 6.59	81.00 ± 2.22	85.50 ± 7.52	79.04 ± 3.42
	LLSPIN	80.33 ± 8.87	73.48 ± 10.82	86.70 ± 6.25	95.50 ± 3.60	88.85 ± 5.40	84.05 ± 1.20	83.57 ± 8.07	83.19 ± 3.30
	IntTabNet	81.50 ± 9.70	90.60 ± 8.81	85.71 ± 6.72	93.57 ± 5.62	89.50 ± 5.70	82.92 ± 2.23	90.17 ± 5.70	85.55 ± 6.57
	ProtoGate	**83.50 ± 11.23**	93.43 ± 6.10	*90.85 ± 5.49*	96.10 ± 3.60	90.20 ± 5.60	84.11 ± 1.19	91.80 ± 5.60	87.89 ± 3.10
	SUWR	83.00 ± 9.60	93.10 ± 6.05	**92.52 ± 3.11**	*96.44 ± 3.75*	90.30 ± 5.55	84.67 ± 2.07	**92.10 ± 5.55**	*87.50 ± 3.34*
	VIBMask	*82.40 ± 10.12*	**95.34 ± 7.11**	87.73 ± 2.49	96.83 ± 3.11	*90.38 ± 6.30*	**89.54 ± 2.05**	91.63 ± 4.52	**89.43 ± 1.25**

Table 2: Balanced classification accuracy (%) on eight real-world datasets, mean ± std. over 5×5 stratified runs. **First**, second, and *third*-best accuracies are highlighted. Statistical significance tests are reported in Sec. A6.7.1 of the technical appendix.

Feature Selection Sparsity and Computational Efficiency. We evaluate selector–predictor IWFS methods in terms of sparsity and inference latency, reporting results averaged over eight datasets (Fig. 2). INVASE and L2X select relatively large feature subsets (75.43% and 51.24%, respectively) yet achieve lower predictive performance (81.73% and 62.58%). In contrast, **VIBMask** attains the highest mean accuracy (90.41%) while selecting only 38.67% of features on average, demonstrating its ability to identify compact and informative subsets that improve both interpretability and reliability. The sparsity–accuracy frontier in Fig. 2 (left) shows that **VIBMask** sits at the top of the accuracy axis while keeping a mask fraction comparable to LSPIN and clearly below those of LLSPIN, L2X, and INVASE. From a computational perspective, L2X and SUWR exhibit the highest inference latency (2.24 ms and 2.54 ms), with L2X also yielding the lowest accuracy. ProtoGate and LLSPIN achieve a favorable trade-off, maintaining accuracy above 84% with inference times below 0.7 ms. Although **VIBMask** employs multiple selectors and thus has more trainable parameters than ProtoGate, LLSPIN, and L2X, its parameter count remains comparable to INVASE. Importantly, **VIBMask** achieves state-of-the-art accuracy with the lowest inference time among the compared IWFS methods (0.51 ms), about 20% faster than ProtoGate, underscoring its practical scalability for downstream deployment.

Performance on MNIST. We evaluate **VIBMask** on image classification using MNIST [Deng, 2012] and Fashion-MNIST [Xiao *et al.*, 2017]. Following [Oosterhuis *et al.*, 2025], methods select 3 × 3 patches on Fashion-MNIST to enhance interpretability. Compared to recent IWFS methods (excluding L2X and INVASE due to known unfaithfulness [Jethani *et al.*, 2021]), including **w/o FS**, **LSPIN**, **LLSPIN**, **ProtoGate**, and **SUWR**, **VIBMask** consistently ranks second in accuracy, excelling especially when selecting very few pixels or patches (Fig. 3). While **w/o FS** (using all features) achieves the highest accuracy, **VIBMask** better bal-

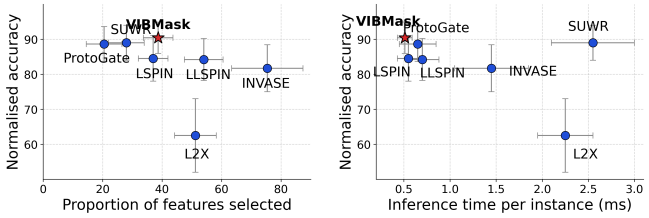

Figure 2: Normalized performance of selector–predictor IWFS methods. **Left:** trade-off between feature proportion and accuracy, showing sparsity–accuracy balance. **Right:** inference time vs. accuracy, indicating efficiency. **VIBMask** consistently achieves high accuracy with fewer features and low latency, balancing interpretability and efficiency.

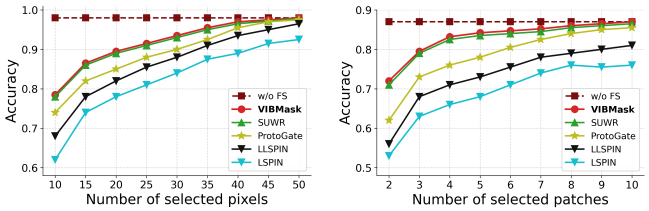

Figure 3: Classification accuracy on MNIST and Fashion-MNIST as a function of the mean number of selected pixels or patches. **w/o FS** (using all features) achieves the highest accuracy, while **VIBMask** consistently delivers strong performance with more efficient feature selection, particularly when very few pixels or patches are selected.

ances interpretability and performance. SUWR performs comparably, ProtoGate is competitive but slightly behind, and LSPIN/LLSPIN lag with less effective selection. These results demonstrate **VIBMask**'s superior ability to select robust, informative features while maintaining strong predictive accuracy.

5.3 Q3: Ablation Study

We assess two key design choices in **VIBMask**: the number of selectors K and the co-adaptation regularization strength γ. As shown on the left of Fig. 4, the cross-dataset mean accuracy peaks at $K = 3$ (87.20%), exceeding both K=1 (86.74%) and larger ensembles (K=4: 84.97%, K=5: 85.52%); the standard deviation also grows sharply for $K \geq 4$, indicating that overly large ensembles enlarge the optimization burden. Per-dataset optima vary with the underlying sparsity, but the cross-dataset average consistently favors $K = 3$, which we adopt as a balance between expressivity and efficiency.

The right panel of Fig. 4 shows the γ-ablation on the Colon dataset, a small-sample setting where co-adaptation is hard to diagnose by accuracy alone. We report a min-max-normalized $\tilde{I}(X_{\overline{M}}; Y) \in [0, 1]$ so that the value at $\gamma = 0$ sits at the top of the scale and full leakage suppression at 0. Activating the decorrelation term progressively drives $\tilde{I}(X_{\overline{M}}; Y)$ down, from ≈ 0.95 at $\gamma = 0$ to 0.00 at $\gamma = 10^2$; the rebound at $\gamma = 10^3$ reflects over-regularization, where the mask becomes nearly all-selecting and the complement set collapses. Meanwhile, the accuracy curve actually *rises*

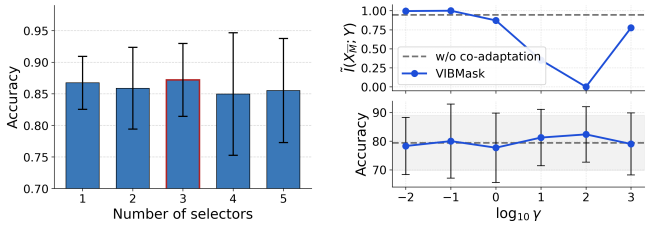

Figure 4: Ablation of **VIBMask**'s components. **Left:** cross-dataset accuracy vs. number of selectors $K = 1$–5 on eight datasets; $K = 3$ (highlighted) attains the best mean. **Right:** min-max-normalized $\tilde{I}(X_{\overline{M}}; Y)$ (top) and accuracy (bottom) on Colon as $\log_{10} \gamma$ varies from -2 to 3; the dashed line is the γ=0 baseline.

above the unregularized baseline ($79.40 \pm 9.60\%$) for moderate regularization, peaking at 82.35% for $\gamma = 10^2$. Suppressing the mask-encoded shortcut thus not only restores faithfulness but also marginally improves predictive accuracy, by forcing the predictor to rely on the genuine feature signal rather than on label leakage [Jiang *et al.*, 2024; Jethani *et al.*, 2021]. Together with the per-dataset predictor-replacement diagnostic in the appendix, these results demonstrate that both multi-selector integration and explicit co-adaptation mitigation are critical for robust and interpretable feature selection, validating the two core design choices of our information-bottleneck framework.

6 Conclusion

We introduce **VIBMask**, a novel IWFS method grounded in a unified framework that generalizes several existing IWFS approaches. Central to **VIBMask** is a newly derived variational bound that is broadly applicable to the discrete VIB problem, enabling principled and efficient feature selection. **VIBMask** addresses co-adaptation by minimizing mutual information between unselected features and the target, thus retaining only the most relevant features. Its design also leverages multiple selectors to capture diverse sparse patterns while fostering robustness and diversity. Using continuous reparameterization, **VIBMask** supports efficient end-to-end optimization and consistently outperforms existing IWFS methods across diverse synthetic and real-world datasets in terms of both accuracy and feature precision.

Despite these strengths, **VIBMask** has several limitations. It focuses on instance-wise mask selection and may overlook global feature patterns in datasets with an underlying sparse global structure. The method also introduces multiple hyperparameters requiring careful tuning, which may limit scalability to very high-dimensional inputs. In addition, **VIBMask** currently lacks explicit mechanisms to model complex instance–feature interactions such as context-dependent relevance. Future work will address these limitations through global feature awareness, adaptive hyperparameter tuning, and interaction-aware components (e.g., attention mechanisms) to further enhance the expressivity and applicability of **VIBMask** across diverse application domains.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and constructive feedback, which helped improve the quality of this paper. This work was supported in part by JST under Grant Nos. JPMJNX25C2, JPMJKP24C3, and JPMJCR21D3, and by JSPS KAKENHI under Grant Nos. 23H00483, 120251002, and 26K21320.

References

[Ahuja *et al.*, 2021] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[Alemi *et al.*, 2017] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

[Alesiani *et al.*, 2023] Francesco Alesiani, Shujian Yu, and Xi Yu. Gated information bottleneck for generalization in sequential environments. *Knowledge and Information Systems*, 65(2):683–705, 2023.

[Arik and Pfister, 2021] Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6679–6687, 2021.

[Bahdanau, 2014] Dzmitry Bahdanau. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[Balın *et al.*, 2019] Muhammed Fatih Balın, Abubakar Abid, and James Zou. Concrete autoencoders: Differentiable feature selection and reconstruction. In *International conference on machine learning*, pages 444–453. PMLR, 2019.

[Barber and Agakov, 2004] David Barber and Felix Agakov. The im algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16(320):201, 2004.

[Breiman, 1996] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[Breiman, 2001] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[Chen and Guestrin, 2016] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[Chen *et al.*, 2018] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International conference on machine learning*, pages 883–892. PMLR, 2018.

[Choi *et al.*, 2016] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter Stewart, and Jimeng Sun. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pages 3504–3512, 2016.

[Cortes, 1995] Corinna Cortes. Support-vector networks. *Machine Learning*, 1995.

[Cox, 1958] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232, 1958.

[Deng, 2012] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[Figurnov *et al.*, 2018] Michael Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit reparameterization gradients. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 439–450, Red Hook, NY, USA, 2018. Curran Associates Inc.

[Fix, 1985] Evelyn Fix. *Discriminatory analysis: nonparametric discrimination, consistency properties*, volume 1. USAF school of Aviation Medicine, 1985.

[Geirhos *et al.*, 2020] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[Gorishniy *et al.*, 2021] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943, 2021.

[Jethani *et al.*, 2021] Neil Jethani, Mukund Sudarshan, Yindalon Aphinyanaphongs, and Rajesh Ranganath. Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations. In *International Conference on Artificial Intelligence and Statistics*, pages 1459–1467. PMLR, 2021.

[Jiang *et al.*, 2024] Xiangjian Jiang, Andrei Margeloiu, Nikola Simidjievski, and Mateja Jamnik. Protogate: Prototype-based neural networks with global-to-local feature selection for tabular biomedical data. In *Forty-first International Conference on Machine Learning*, 2024.

[Kalos and Whitlock, 2009] Malvin H Kalos and Paula A Whitlock. *Monte carlo methods*. John Wiley & Sons, 2009.

[Kim *et al.*, 2021] Jaekyeom Kim, Minjung Kim, Dongyeon Woo, and Gunhee Kim. Drop-bottleneck: Learning discrete compressed representation for noise-robust exploration. In *International Conference on Learning Representations*, 2021.

[Kingma *et al.*, 2014] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*. 2014.

[Kuang *et al.*, 2023] Huafeng Kuang, Hong Liu, YONGJIAN WU, Shin'ichi Satoh, and Rongrong Ji. Improving adversarial robustness via information bottleneck distillation. In

Thirty-seventh Conference on Neural Information Processing Systems, 2023.

[Liu *et al.*, 2023] Dugang Liu, Pengxiang Cheng, Hong Zhu, Xing Tang, Yanyu Chen, Xiaoting Wang, Weike Pan, Zhong Ming, and Xiuqiang He. Diwift: Discovering instance-wise influential features for tabular data. In *Proceedings of the ACM Web Conference 2023*, pages 1673–1682, 2023.

[Louizos *et al.*, 2018] Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through l_0 regularization. In *International Conference on Learning Representations*, 2018.

[Ma *et al.*, 2018] Fenglong Ma, Ravi Chitta, Jing Zhou, Quanzeng You, Tingting Sun, Jianfeng Gao, and Junchi Gao. Health-atm: A deep architecture for multifaceted patient health record representation and risk prediction. In *Proceedings of the 2018 SIAM International Conference on Data Mining (SDM)*, pages 261–269. SIAM, 2018.

[Masoomi *et al.*, 2020] Aria Masoomi, Chieh Wu, Tingting Zhao, Zifeng Wang, Peter Castaldi, and Jennifer Dy. Instance-wise feature grouping. *Advances in Neural Information Processing Systems*, 33:13374–13386, 2020.

[Mnih and Rezende, 2016] Andriy Mnih and Danilo Rezende. Variational inference for monte carlo objectives. In *International Conference on Machine Learning*, pages 2188–2196. PMLR, 2016.

[Mohamed and Jimenez Rezende, 2015] Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. *Advances in neural information processing systems*, 28, 2015.

[Molnar, 2020] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[Oosterhuis *et al.*, 2025] Harrie Oosterhuis, Lijun Lyu, and Avishek Anand. Local feature selection without label or feature leakage for interpretable machine learning predictions. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2025.

[Peng *et al.*, 2023] Hanyu Peng, Guanhua Fang, and Ping Li. Copula for instance-wise feature selection and rank. In *Uncertainty in Artificial Intelligence*, pages 1651–1661. PMLR, 2023.

[Si *et al.*, 2024] Jacob Yoke Hong Si, Wendy Yusi Cheng, Michael Cooper, and Rahul Krishnan. Interpretabnet: Distilling predictive signals from tabular data by salient feature interpretation. In *Forty-first International Conference on Machine Learning*, 2024.

[Sokar *et al.*, 2022] Ghada Sokar, Zahra Atashgahi, Mykola Pechenizkiy, and Decebal Constantin Mocanu. Where to pay attention in sparse training for feature selection? *Advances in Neural Information Processing Systems*, 35:1627–1642, 2022.

[Sristi *et al.*, 2024] Ram Dyuthi Sristi, Ofir Lindenbaum, Shira Lifshitz, Maria Lavzin, Jackie Schiller, Gal Mishne, and Hadas Benisty. Contextual feature selection with conditional stochastic gates. In *Forty-first International Conference on Machine Learning*, 2024.

[Tibshirani, 1996] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[Tishby *et al.*, 2000] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[Ucar *et al.*, 2021] Talip Ucar, Ehsan Hajiramezanali, and Lindsay Edwards. Subtab: Subsetting features of tabular data for self-supervised representation learning. *Advances in Neural Information Processing Systems*, 34:18853–18865, 2021.

[Williams, 1992] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

[Xiao *et al.*, 2017] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. https://github.com/zalandoresearch/fashion-mnist, 2017. Accessed: 2024-11-22.

[Yamada *et al.*, 2020] Yutaro Yamada, Ofir Lindenbaum, Sahand Negahban, and Yuval Kluger. Feature selection using stochastic gates. In *International conference on machine learning*, pages 10648–10659. PMLR, 2020.

[Yang *et al.*, 2022] Junchen Yang, Ofir Lindenbaum, and Yuval Kluger. Locally sparse neural networks for tabular biomedical data. In *International Conference on Machine Learning*, pages 25123–25153. PMLR, 2022.

[Yoon *et al.*, 2018] Jinsung Yoon, James Jordon, and Mihaela Van der Schaar. Invase: Instance-wise variable selection using neural networks. In *International conference on learning representations*, 2018.