Appendix of Manuscript
Revisiting Probability Distribution Assumptions for Information Theoretic Feature Selection

Yuan Sun,[1*] **Wei Wang,**[2] **Michael Kirley,**[2] **Xiaodong Li,**[1] **Jeffrey Chan**[1]
[1]RMIT University, Melbourne, Australia
[2]University of Melbourne, Parkville, Australia
[1]{yuan.sun, xiaodong.li, jeffrey.chan}@rmit.edu.au
[2]weiw8@student.unimelb.edu.au, mkirley@unimelb.edu.au

I Proof of Theorems

Consider a Sequential Forward Selection algorithm where a candidate feature X_m is selected such that the Mutual Information (MI) between X_m and label C given the selected feature subset ($\boldsymbol{S}$) is maximized:

$$X_m = \underset{X_m \in \boldsymbol{X} \setminus \boldsymbol{S}}{\arg\max} \big\{ J(X_m) := I(X_m; C|\boldsymbol{S}) \big\}, \tag{A.1}$$

To estimate the high-order conditional MI $I(X_m; C|\boldsymbol{S})$, existing methods often make assumptions on the probability distribution of features to decompose the $I(X_m; C|\boldsymbol{S})$ into a series of low-order MI quantities. Two typically sets of assumptions are: *Feature Independence Distribution* (FID) assumption and *Geometric Mean Distribution* (GMD) assumption.

In Appendix I.A, we prove Theorem 1 under the FID assumption, and Appendix I.B, Theorem 2 under the GMD assumption. Furthermore, in Corollary 1 and 2, we show that existing methods from Table 1 of the manuscript can be derived based on the two assumptions and these methods differ only in the order parameters used.

I.A Proof of Theorem Under FID

First, we re-describe the math notations and the FID assumption. Given selected features $\boldsymbol{S}$ and a candidate feature X_m, let $\boldsymbol{S}_k \subseteq \boldsymbol{S}$ contain k ($0 \leq k \leq |\boldsymbol{S}|$) randomly selected features from $\boldsymbol{S}$. Define a trial feature set $\boldsymbol{T} = X_m \cup \boldsymbol{S}$ and its subset $\boldsymbol{T}_k = X_m \cup \boldsymbol{S}_{k-1}$ where the first feature is X_m and the remaining $k-1$ features are from $\boldsymbol{S}$; $\boldsymbol{T}_0 = \emptyset$.

Feature Independence Distribution (FID) The FID assumption states that $\boldsymbol{S}$ and X_m are independent and class-conditionally independent at order k ($0 \leq k \leq |\boldsymbol{S}|$):

$$p(\boldsymbol{T} \setminus \boldsymbol{T}_k | \boldsymbol{T}_k) \simeq \prod_{X_i \in \boldsymbol{T} \setminus \boldsymbol{T}_k} p(X_i|\boldsymbol{T}_k), \tag{A.2}$$

$$p(\boldsymbol{T} \setminus \boldsymbol{T}_k | \boldsymbol{T}_k, C) \simeq \prod_{X_i \in \boldsymbol{T} \setminus \boldsymbol{T}_k} p(X_i|\boldsymbol{T}_k, C). \tag{A.3}$$

where we use $\simeq$ to denote "asymptotic" equality, in the sense that Eq. (A.2) and (A.3) will become exact when $k = |\boldsymbol{S}|$.

*Corresponding author

Note that FID is a generic case of several feature distribution assumptions used in literature, because it restores these assumptions by simply setting k to different values. When $k = 0$ (i.e., $\boldsymbol{T}_k = \emptyset$), FID becomes the feature independence and class-conditional independence assumptions, that have been used in (Balagani and Phoha 2010; Brown et al. 2012):

$$p(X_m, \boldsymbol{S}) \simeq \prod_{X_i \in \boldsymbol{T}} p(X_i), \tag{A.4}$$

$$p(X_m, \boldsymbol{S}|C) \simeq \prod_{X_i \in \boldsymbol{T}} p(X_i|C). \tag{A.5}$$

When $k = 1$, the following assumptions are restored (Balagani and Phoha 2010; Brown et al. 2012):

$$p(\boldsymbol{S}|X_m) \simeq \prod_{X_i \in \boldsymbol{S}} p(X_i|X_m), \tag{A.6}$$

$$p(\boldsymbol{S}|X_m, C) \simeq \prod_{X_i \in \boldsymbol{S}} p(X_i|X_m, C). \tag{A.7}$$

When $k = 2$, Eq. (A.2) is equivalent to the one used in (Vinh et al. 2016):

$$p(\boldsymbol{S}|X_m) \simeq p(X_j|X_m) \prod_{X_i \in \boldsymbol{S} \setminus X_j} p(X_i|X_m, X_j), \tag{A.8}$$

where X_j is any feature in $\boldsymbol{S}$. Here we have used the probability chain rule $p(\boldsymbol{S}|X_m) = p(X_j|X_m)p(\boldsymbol{S} \setminus X_j|X_m, X_j)$.

Theorem 1. *Under the FID assumption of order k, the objective function $J(X_m)$ in Eq. (A.1) is equivalent to:*

$$J^{k,k}_{\mathrm{FID}}(X_m) \sim \sum_{i=1}^{k} I(X_{t_i}; C|\boldsymbol{T}_{i-1}) + \sum_{X_i \in \boldsymbol{T} \setminus \boldsymbol{T}_k} I(X_i; C|\boldsymbol{T}_k), \tag{A.9}$$

where $\boldsymbol{T}_{i-1} = \{X_{t_1}, \cdots, X_{t_{i-1}}\}$ and $X_{t_1} = X_m$; and $\sim$ denotes "equivalent to". More generally, consider different values of k in Eq. (A.2) and (A.3), denoted as k_1 and k_2 respectively; the objective function $J(X_m)$ is equivalent to:

$$J^{k_1,k_2}_{\mathrm{FID}}(X_m) \sim \sum_{i=1}^{k_1} H(X_{t_i}|\boldsymbol{T}_{i-1}) + \sum_{X_i \in \boldsymbol{T} \setminus \boldsymbol{T}_{k_1}} H(X_i|\boldsymbol{T}_{k_1}) - \sum_{i=1}^{k_2} H(X_{t_i}|\boldsymbol{T}_{i-1}, C) - \sum_{X_i \in \boldsymbol{T} \setminus \boldsymbol{T}_{k_2}} H(X_i|\boldsymbol{T}_{k_2}, C). \tag{A.10}$$

Proof. Based on the probability chain rule,

$$p(X_m, \boldsymbol{S}) = \left(\prod_{i=1}^{k_1} p(X_{t_i}|\boldsymbol{T}_{i-1}) \right) p(\boldsymbol{T} \setminus \boldsymbol{T}_{k_1}|\boldsymbol{T}_{k_1}). \quad \text{(A.11)}$$

Substituting $p(\boldsymbol{T} \setminus \boldsymbol{T}_{k_1}|\boldsymbol{T}_{k_1})$ using Eq. (A.2) yields

$$p(X_m, \boldsymbol{S}) = \prod_{i=1}^{k_1} p(X_{t_i}|\boldsymbol{T}_{i-1}) \prod_{X_i \in \boldsymbol{T} \setminus \boldsymbol{T}_{k_1}} p(X_i|\boldsymbol{T}_{k_1}). \quad \text{(A.12)}$$

Thus,

$$H(X_m, \boldsymbol{S}) = \sum_{i=1}^{k_1} H(X_{t_i}|\boldsymbol{T}_{i-1}) + \sum_{X_i \in \boldsymbol{T} \setminus \boldsymbol{T}_{k_1}} H(X_i|\boldsymbol{T}_{k_1}). \quad \text{(A.13)}$$

Similarly, based on Eq. (A.3), we obtain

$$H(X_m, \boldsymbol{S}|C) = \sum_{i=1}^{k_2} H(X_{t_i}|\boldsymbol{T}_{i-1}, C) + \sum_{X_i \in \boldsymbol{T} \setminus \boldsymbol{T}_{k_2}} H(X_i|\boldsymbol{T}_{k_2}, C). \quad \text{(A.14)}$$

As $I(X_m; C|\boldsymbol{S}) = H(X_m, \boldsymbol{S}) - H(X_m, \boldsymbol{S}|C) - I(\boldsymbol{S}; C)$, and $I(\boldsymbol{S}; C)$ is a constant for a given $\boldsymbol{S}$ and C,

$$I(X_m; C|\boldsymbol{S}) \sim H(X_m, \boldsymbol{S}) - H(X_m, \boldsymbol{S}|C). \quad \text{(A.15)}$$

Substituting Eq. (A.13) and (A.14) into (A.15) yields (6). When $k_1 = k_2 = k$, we assume $\boldsymbol{T}_{k_1} = \boldsymbol{T}_{k_2} = \boldsymbol{T}_k$. Then we have $H(X_{t_i}|\boldsymbol{T}_{i-1}) - H(X_{t_i}|\boldsymbol{T}_{i-1}, C) = I(X_{t_i}; C|\boldsymbol{T}_{i-1})$ and $H(X_i|\boldsymbol{T}_k) - H(X_i|\boldsymbol{T}_k, C) = I(X_i; C|\boldsymbol{T}_{k_2})$. Thus Eq. (6) becomes (A.9). □

In Theorem 1, we have derived a set of methods based on the FID assumption, with the objective function defined by Eq. (A.10). In the following, we show that by varying the values of k_1 and k_2 in Eq. (A.10), we can recover the objective functions of three existing MI based methods.

Corollary 1. *$J^{0,0}_{\text{FID}}$ is equivalent to the Mutual Information Maximization (MIM) criterion (Lewis 1992):*

$$J^{0,0}_{\text{FID}}(X_m) \sim J_{mim}(X_m) := I(X_m; C); \quad \text{(A.16)}$$

$J^{1,0}_{\text{FID}}$ is equivalent to the Mutual Information Feature Selection (MIFS) criterion with $\beta = 1$ (Battiti 1994):

$$J^{1,0}_{\text{FID}}(X_m) \sim J_{mifs}(X_m) := I(X_m; C) - \beta \sum_{X_i \in \boldsymbol{S}} I(X_m; X_i); \quad \text{(A.17)}$$

$J^{1,1}_{\text{FID}}$ is equivalent to the Conditional Informative Feature Extraction (CIFE) criterion (Lin and Tang 2006):

$$J^{1,1}_{\text{FID}}(X_m) \sim J_{cife}(X_m) := I(X_m; C) - \sum_{X_i \in \boldsymbol{S}} I(X_m; X_i) + \sum_{X_i \in \boldsymbol{S}} I(X_m; X_i|C). \quad \text{(A.18)}$$

Proof. When $k_1 = k_2 = k = 0$, $\boldsymbol{T}_k = \emptyset$. Thus

$$J^{0,0}_{\text{FID}}(X_m) \sim \sum_{X_i \in \boldsymbol{T}} I(X_i; C). \quad \text{(A.19)}$$

As $\sum_{X_i \in \boldsymbol{S}} I(X_i; C)$ is a constant for a fixed $\boldsymbol{S}$, $J^{0,0}_{\text{FID}}(X_m) \sim I(X_m; C)$.

When $k_1 = 1$ and $k_2 = 0$, $\boldsymbol{T}_{k_1} = X_m$ and $\boldsymbol{T}_{k_2} = \emptyset$. Thus

$$J^{1,0}_{\text{FID}} \sim H(X_m) + \sum_{X_i \in \boldsymbol{S}} H(X_i|X_m) - \sum_{X_i \in \boldsymbol{T}} H(X_i|C) \quad \text{(A.20)}$$

As $\sum_{X_i \in \boldsymbol{S}} H(X_i|C)$ is a constant for fixed $\boldsymbol{S}$ and C,

$$J^{1,0}_{\text{FID}} \sim \sum_{X_i \in \boldsymbol{S}} H(X_i|X_m) + H(X_m) - H(X_m|C). \quad \text{(A.21)}$$

It is true that $H(X_m) - H(X_m|C) = I(X_m; C)$; and $H(X_i|X_m) = H(X_i) - I(X_m; X_i)$, where $H(X_i)$ is a constant for a fixed X_i. Thus

$$J^{1,0}_{\text{FID}} \sim I(X_m; C) - \sum_{X_i \in \boldsymbol{S}} I(X_i; X_m). \quad \text{(A.22)}$$

When $k_1 = k_2 = 1$, $\boldsymbol{T}_{k_1} = \boldsymbol{T}_{k_2} = X_m$. Thus,

$$J^{1,1}_{\text{FID}}(X_m) = I(X_m; C) + \sum_{X_i \in \boldsymbol{S}} I(X_i; C|X_m). \quad \text{(A.23)}$$

As $I(X_i; C|X_m) = I(X_i; C) - I(X_i; X_m) + I(X_i; X_m|C)$ and $I(X_i; C)$ is a constant for a given X_i and C,

$$J^{1,1}_{\text{FID}}(X_m) \sim I(X_m; C) - \sum_{X_i \in \boldsymbol{S}} I(X_i; X_m) + \sum_{X_i \in \boldsymbol{S}} I(X_i; X_m|C). \quad \text{(A.24)}$$

□

In Corollary 1, we have shown that the MIM, MIFS and CIFE methods are all based on the FID assumption, and differ only in the order k used. We now proceed to the cases of GMD assumption.

I.B Proof of Theorem Under GMD

We first re-describe the math notations and the GMD assumption. A k-combination of a set $\boldsymbol{S}$ is a subset of k distinct elements of $\boldsymbol{S}$. The number of k-combinations of the set $\boldsymbol{S}$ is equal to the binomial coefficient $\binom{|\boldsymbol{S}|}{k} = \frac{|\boldsymbol{S}|!}{k!(|\boldsymbol{S}|-k)!}$. We denote the i_{th} k-combination of $\boldsymbol{S}$ as $\boldsymbol{S}^i_k$ and all the possible k-combinations of $\boldsymbol{S}$ as $\mathbb{S}_k$. Thus $\boldsymbol{S}^i_k \in \mathbb{S}_k$, where $1 \leq i \leq \binom{|\boldsymbol{S}|}{k}$.

Geometric Mean Distribution (GMD) The GMD assumption states that the (class-conditional) probability density function of a candidate feature X_m given the selected features $\boldsymbol{S}$ is equal to the geometric mean of the (class-conditional) probability density function of X_m conditioning on any k ($0 \leq k \leq |\boldsymbol{S}|$) features in $\boldsymbol{S}$:

$$p(X_m|\boldsymbol{S}) \simeq \left(\prod_{\boldsymbol{S}^i_k \in \mathbb{S}_k} p(X_m|\boldsymbol{S}^i_k) \right)^{\frac{1}{\binom{|\boldsymbol{S}|}{k}}}, \quad \text{(A.25)}$$

$$p(X_m|\boldsymbol{S}, C) \simeq \left(\prod_{\boldsymbol{S}^i_k \in \mathbb{S}_k} p(X_m|\boldsymbol{S}^i_k, C) \right)^{\frac{1}{\binom{|\boldsymbol{S}|}{k}}}. \quad \text{(A.26)}$$

Theorem 2. *Under the GMD assumption, the evaluation criterion of a candidate feature X_m, shown in Eq. (A.1), is equivalent to:*

$$J^{k,k}_{\mathrm{GMD}}(X_m) \sim \frac{1}{\binom{|\boldsymbol{S}|}{k}} \sum_{\boldsymbol{S}^i_k \in \mathbb{S}_k} I(X_m; C|\boldsymbol{S}^i_k). \tag{A.27}$$

More generally, we consider different values of k in Eq. (A.25) and (A.26), denoted as k_1 and k_2 respectively. The objective function $J(X_m)$ is equivalent to:

$$\begin{aligned} J^{k_1,k_2}_{\mathrm{GMD}}(X_m) \sim & \frac{1}{\binom{|\boldsymbol{S}|}{k_1}} \sum_{\boldsymbol{S}^i_{k_1} \in \mathbb{S}_{k_1}} H(X_m|\boldsymbol{S}^i_{k_1}) \\ & - \frac{1}{\binom{|\boldsymbol{S}|}{k_2}} \sum_{\boldsymbol{S}^i_{k_2} \in \mathbb{S}_{k_2}} H(X_m|\boldsymbol{S}^i_{k_2}, C). \end{aligned} \tag{A.28}$$

Proof. If Eq. (A.25) holds (with $k = k_1$):

$$H(X_m|\boldsymbol{S}) = \frac{1}{\binom{|\boldsymbol{S}|}{k_1}} \sum_{\boldsymbol{S}^i_{k_1} \in \mathbb{S}_{k_1}} H(X_m|\boldsymbol{S}^i_{k_1}). \tag{A.29}$$

If Eq. (A.26) holds (with $k = k_2$):

$$H(X_m|\boldsymbol{S}, C) = \frac{1}{\binom{|\boldsymbol{S}|}{k_2}} \sum_{\boldsymbol{S}^i_{k_2} \in \mathbb{S}_{k_2}} H(X_m|\boldsymbol{S}^i_{k_2}, C). \tag{A.30}$$

Substituting Eq. (A.29) and (A.30) into $I(X_m; C|\boldsymbol{S}) = H(X_m|\boldsymbol{S}) - H(X_m|\boldsymbol{S}, C)$ yields (A.28). When $k_1 = k_2 = k$, $H(X_m|\boldsymbol{S}^i_k) - H(X_m|\boldsymbol{S}^i_k, C) = I(X_m; C|\boldsymbol{S}^i_k)$. Thus Eq. (A.28) becomes (A.27). □

In Theorem 2, we have derived a set of methods based on the GMD assumption, with the objective function defined by Eq. (A.28). By varying the values of k_1 and k_2, we can recover four existing MI based methods as below.

Corollary 2. *$J^{0,0}_{\mathrm{GMD}}$ is equivalent to the Mutual Information Maximization (MIM) criterion (Lewis 1992):*

$$J^{0,0}_{\mathrm{GMD}}(X_m) \sim J_{mim}(X_m) := I(X_m; C); \tag{A.31}$$

$J^{1,0}_{\mathrm{GMD}}$ is equivalent to the Minimum-Redundancy Maximum-Relevance (MRMR) criterion (Peng, Long, and Ding 2005):

$$J^{1,0}_{\mathrm{GMD}}(X_m) \sim J_{mrmr}(X_m) := I(X_m; C) - \frac{1}{|\boldsymbol{S}|} \sum_{X_i \in \boldsymbol{S}} I(X_m; X_i); \tag{A.32}$$

$J^{1,1}_{\mathrm{GMD}}$ is equivalent to the Joint Mutual Information (JMI) criterion (Yang and Moody 1999; Meyer, Schretter, and Bontempi 2008):

$$\begin{aligned} & J^{1,1}_{\mathrm{GMD}}(X_m) \sim J_{jmi}(X_m) \\ & := I(X_m; C) - \frac{1}{|\boldsymbol{S}|} \sum_{X_i \in \boldsymbol{S}} I(X_m; X_i) + \frac{1}{|\boldsymbol{S}|} \sum_{X_i \in \boldsymbol{S}} I(X_m; X_i|C); \end{aligned} \tag{A.33}$$

$J^{2,1}_{\mathrm{GMD}}$ is equivalent to the Relaxed Minimum-Redundancy Maximum-Relevance (RMRMR) criterion (Vinh et al. 2016):

$$\begin{aligned} & J^{2,1}_{\mathrm{GMD}}(X_m) \sim J_{rmrmr}(X_m) \\ & := I(X_m; C) - \frac{1}{|\boldsymbol{S}|} \sum_{X_i \in \boldsymbol{S}} I(X_m; X_i) + \frac{1}{|\boldsymbol{S}|} \sum_{X_i \in \boldsymbol{S}} I(X_m; X_i|C) \\ & - \frac{1}{|\boldsymbol{S}|(|\boldsymbol{S}|-1)} \sum_{X_i \in \boldsymbol{S}} \sum_{X_j \in \boldsymbol{S} \setminus X_i} I(X_m; X_j|X_i). \end{aligned} \tag{A.34}$$

Proof. When $k_1 = k_2 = k = 0$, $\binom{|\boldsymbol{S}|}{k} = 1$ and $\mathbb{S}_k = \{\emptyset\}$; thus

$$J^{0,0}_{\mathrm{GMD}}(X_m) \sim I(X_m; C). \tag{A.35}$$

When $k_1 = 1$ and $k_2 = 0$, $\binom{|\boldsymbol{S}|}{k_1} = |\boldsymbol{S}|$ and $\mathbb{S}_{k_1} = \boldsymbol{S}$. Thus,

$$J^{1,0}_{\mathrm{GMD}} \sim \frac{1}{|\boldsymbol{S}|} \sum_{X_i \in \boldsymbol{S}} H(X_m|X_i) - H(X_m|C). \tag{A.36}$$

As $H(X_m|X_i) = H(X_m) - I(X_m; X_i)$,

$$J^{1,0}_{\mathrm{GMD}} \sim H(X_m) - \frac{1}{|\boldsymbol{S}|} \sum_{X_i \in \boldsymbol{S}} I(X_m; X_i) - H(X_m|C). \tag{A.37}$$

Substituting $H(X_m) - H(X_m|C)$ with $I(X_m; C)$ yields Eq. (A.32).

When $k_1 = k_2 = k = 1$,

$$J^{1,1}_{\mathrm{GMD}}(X_m) \sim \frac{1}{|\boldsymbol{S}|} \sum_{X_i \in \boldsymbol{S}} I(X_m; C|X_i). \tag{A.38}$$

Substituting $I(X_m; C|X_i)$ with $I(X_m; C) - I(X_m; X_i) + I(X_m; X_i|C)$ yields Eq. (A.33).

When $k_1 = 2$ and $k_2 = 1$,

$$\begin{aligned} J^{2,1}_{\mathrm{GMD}}(X_m) \sim & \frac{2}{|\boldsymbol{S}|(|\boldsymbol{S}|-1)} \sum_{X_i \in \boldsymbol{S}} \sum_{X_j \in \boldsymbol{S}, j>i} H(X_m|X_i, X_j) \\ & - \frac{1}{|\boldsymbol{S}|} \sum_{X_i \in \boldsymbol{S}} H(X_m|X_i, C). \end{aligned} \tag{A.39}$$

As $H(X_m|X_i, X_j) = H(X_m|X_j, X_i)$,

$$\begin{aligned} J^{2,1}_{\mathrm{GMD}}(X_m) \sim & \frac{1}{|\boldsymbol{S}|(|\boldsymbol{S}|-1)} \sum_{X_i \in \boldsymbol{S}} \sum_{X_j \in \boldsymbol{S} \setminus X_i} H(X_m|X_i, X_j) \\ & - \frac{1}{|\boldsymbol{S}|} \sum_{X_i \in \boldsymbol{S}} H(X_m|X_i, C). \end{aligned} \tag{A.40}$$

As $H(X_m|X_i, X_j) = H(X_m|X_i) - I(X_m; X_j|X_i)$,

$$\begin{aligned} J^{2,1}_{\mathrm{GMD}}(X_m) \sim & \frac{1}{|\boldsymbol{S}|} \sum_{X_i \in,} H(X_m|X_i) - \frac{1}{|\boldsymbol{S}|} \sum_{X_i \in \boldsymbol{S}} H(X_m|X_i, C) \\ & - \frac{1}{|\boldsymbol{S}|(|\boldsymbol{S}|-1)} \sum_{X_i \in \boldsymbol{S}} \sum_{X_j \in \boldsymbol{S} \setminus X_i} I(X_m; X_j|X_i). \end{aligned} \tag{A.41}$$

Substituting $H(X_m|X_i) - H(X_m|X_i, C) = I(X_m; C|X_i) = I(X_m; C) - I(X_m; X_i) + I(X_m; X_i|C)$ into Eq. (A.41) yields Eq. (A.34). □

Algorithm 1 Sequential Forward Selection ($\boldsymbol{X}$, C, K)

1: Initialize $\boldsymbol{S} \leftarrow \emptyset$.
2: Initialize $y[m] \leftarrow 0$, for each m from 1 to $|\boldsymbol{X}|$.
3: **while** $|\boldsymbol{S}| < K$ **do**
4: $\quad y_{\max} \leftarrow 0$; $idx \leftarrow -1$.
5: $\quad$ **for** m from 1 to $|\boldsymbol{X}|$ **do**
6: $\quad\quad$ **if** X_m is not in $\boldsymbol{S}$ **then**
7: $\quad\quad\quad y[m] \leftarrow I(X_m; C|\boldsymbol{S})$, or $\tilde{I}(X_m; C|\boldsymbol{S})$.
8: $\quad\quad\quad$ **if** $y[m] > y_{\max}$ **then**
9: $\quad\quad\quad\quad y_{\max} \leftarrow y[m]$.
10: $\quad\quad\quad\quad idx \leftarrow m$.
11: $\quad$ Adding X_{idx} to $|\boldsymbol{S}|$.
return $\boldsymbol{S}$.

In Corollary 2, we have shown that the MIM, MRMR, JMI and RMRMR methods are all based on the GMD assumption, and differ only in the order k used.

II Supplementary Methodology

In this section, we describe the Sequential Forward Selection algorithm for feature selection. *We also provide the C++ and Python source codes of the methods used in this study in the supplementary material (Source_Code.zip file).*

Given a supervised learning task with feature vector $\boldsymbol{X}$ and class label C, the goal of MI or VI based feature selection methods is to search for a subset of K features ($\boldsymbol{S}$) such that the MI or VI between $\boldsymbol{S}$ and C is maximized. Sequential Forward Selection (Kittler 1986; Pudil, Novovičova, and Kittler 1994) is a class of greedy methods that selects a candidate feature X_m at a time such that the MI (or VI) between X_m and C given the selected feature subset ($\boldsymbol{S}$) is maximized. The high-level idea of the Sequential Forward Selection algorithm is shown in Algorithm 1. By replacing the $I(X_m; C|\boldsymbol{S})$ or $\tilde{I}(X_m; C|\boldsymbol{S})$ in line 7 with different approximations derived from the AMD, GMD and FID assumptions, we can obtain various feature selection methods (12 in total) used in the main manuscript.

III Supplementary Experiments

In this section, we provide more details for Section 4 Experiments and additional experiments that compare the proposed methods under *Arithmetic Mean Distribution* (AMD) assumption against existing methods in literature.

III.A Additional Comparison of Probability Distributions

We select the Isolet dataset as an example to further illustrate the effectiveness of each probability distribution assumption in Fig. A1. The convergence plots show that when varying the number of selected features from 1 to 100, the methods based on the AMD and GMD assumption generally generate much lower classification error rates than the methods based on the FID assumption. Furthermore, the convergence curve of the AMD-based method is better than (under) that of the GMD-based method. These results further support our findings in the main manuscript that:

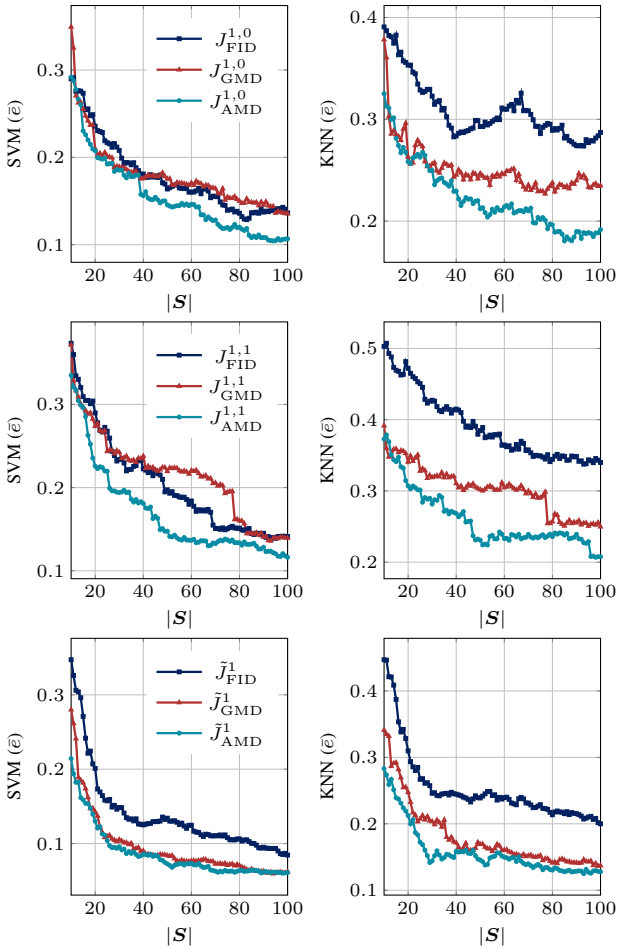

Figure A1: Classification error rates of SVM or KNN on the Isolet dataset when using the features selected by the methods based on the FID, GMD and AMD assumptions. The horizontal axis represents the number of selected features ($|\boldsymbol{S}|$). The vertical axis represents the mean error rates $\bar{e}$.

1. The AMD and GMD assumptions can potentially reduce the estimation bias of the FID assumption;
2. By fixing the unnormalized probability density issue, the AMD assumption can potentially select more informative features than the GMD assumption.

III.B Additional Comparison of Order k in Probability Distributions

In this subsection, we present additional experimental results to illustrate the effects of order k on the performance of feature selection methods. These results include:

1. the plots of classification error rates generated by the methods with different order k (Figure A2).
2. the convergence curve of methods with different order k on a selected dataset (Figure A3).
3. the CPU running time of methods with different order k (Figure A4).

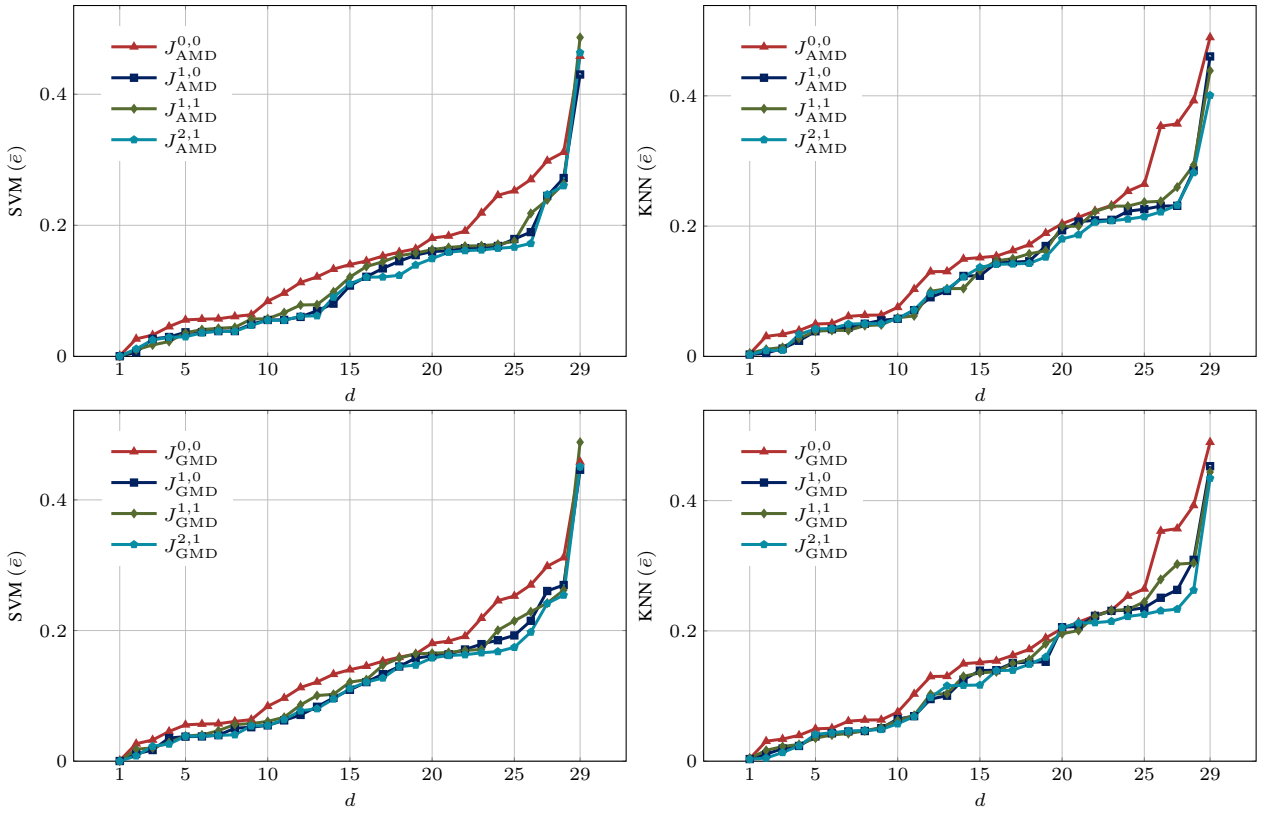

Figure A2: The effects of the number of conditional features (k) on the performance of feature selection methods. The classification errors of $J^{0,0}_{\mathrm{AMD}}$, $J^{1,0}_{\mathrm{AMD}}$, $J^{1,1}_{\mathrm{AMD}}$ and $J^{2,1}_{\mathrm{AMD}}$ are shown in top figures; and those of $J^{0,0}_{\mathrm{GMD}}$, $J^{1,0}_{\mathrm{GMD}}$, $J^{1,1}_{\mathrm{GMD}}$, and $J^{2,1}_{\mathrm{GMD}}$ are shown in bottom figures. The horizontal axis represents the indices of data sets (d), and the vertical axis represents mean classification error rates ($\bar{e}$) generated by KNN or SVM when using the features selected by each method.

Note that to generate the plots in Figure A2 and A4, we sort the classification error rates produced or CPU running time used by each method on the 29 datasets in ascending order for better visualization purpose. Thus the dataset index in these plots may not match the true dataset index in Table 2 of the manuscript. However, the detailed classification error rates of each method on each dataset can be found in Table A1 and A2.

From Figure A2 and A3, we can observe that the performance of AMD and GMD based methods generally improves as k increases from 0 to 2. These results are consistent with our findings in the main manuscript. However, Figure A4 show that the CPU running time of these methods also increases as k increases.

III.C Comparison Between 12 Methods Based on AMD, GMD and FID

In this subsection, we systematically compare the performance of the 12 MI and VI based methods that use the AMD, GMD or FID assumptions with different order k. The detailed SVM and KNN classification error rates for each method on each dataset are shown in Table A1 and A2 respectively. We also rank the 12 methods on each datasets based on their classification array rates and the average rankings are shown in the last row of each table. We observe that the methods based on the AMD assumption are very competitive against the methods based on GMD and FID. Furthermore, the VI-based method $\tilde{J}^{1}_{\mathrm{AMD}}$ and the MI-based method $J^{2,1}_{\mathrm{AMD}}$ achieve overall the best ranking among the 12 methods we have considered in our framework. In the next subsection, we will compare $\tilde{J}^{1}_{\mathrm{AMD}}$ and $J^{2,1}_{\mathrm{AMD}}$ against other competitive methods that are not covered by our framework.

III.D Additional Comparison Between the AMD Based Methods and Other State-of-the-arts

In this subsection, we presented additional experimental results on the comparison between the AMD-based methods ($\tilde{J}^{1}_{\mathrm{AMD}}$ and $J^{2,1}_{\mathrm{AMD}}$) against other competitive methods. The detailed SVM and KNN classification error rates generated by each method are presented in Table A3 and A4 respectively. The last row of the tables lists the average ranking of each method. The results show that the MI and VI based methods under the AMD assumption achieve overall the best and second best average ranking compared to 5 other state-of-the-art methods. It confirms the efficacy of the AMD assumption in terms of selecting informative features for real-world feature selection tasks.

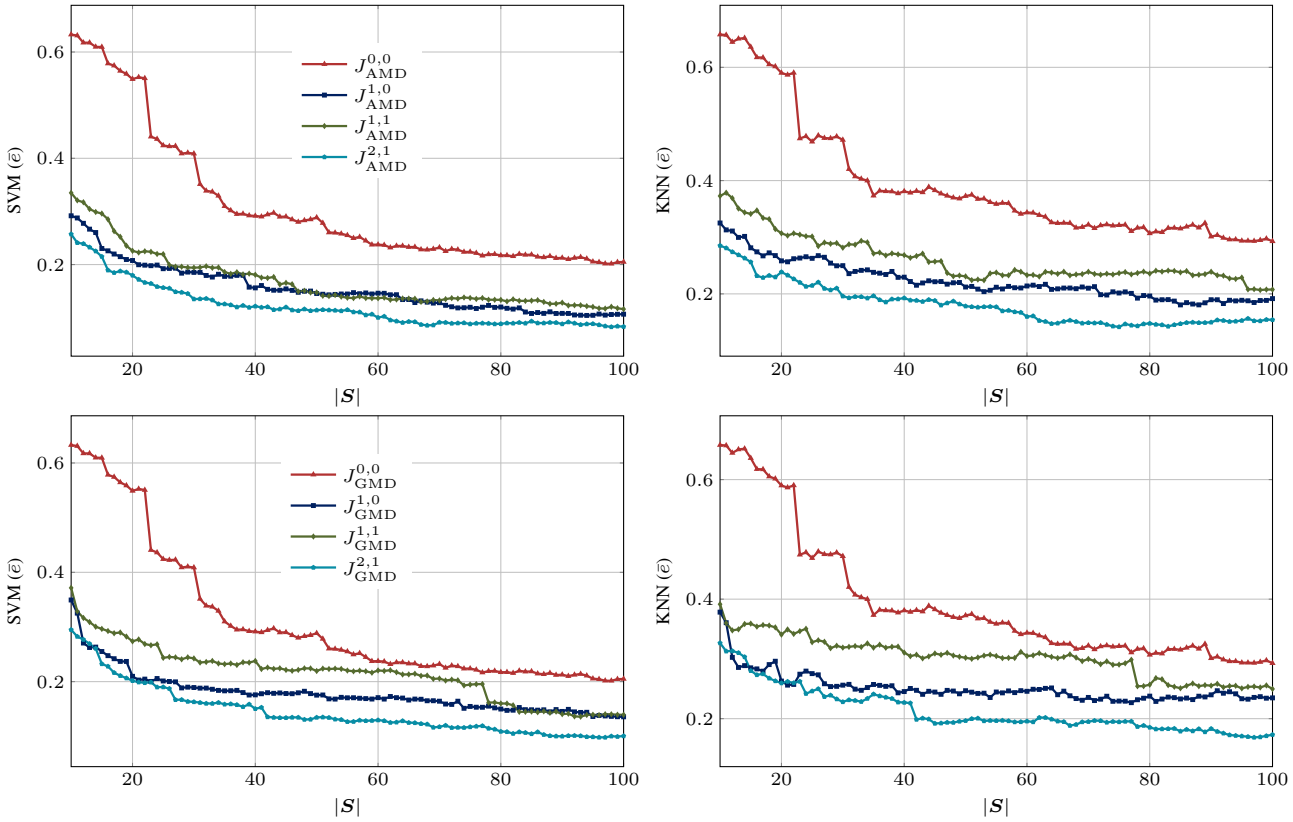

Figure A3: The classification error rates of SVM or KNN when using the subset of features selected by 1) $J^{0,0}_{\mathrm{AMD}}$, $J^{1,0}_{\mathrm{AMD}}$, $J^{1,1}_{\mathrm{AMD}}$, $J^{2,1}_{\mathrm{AMD}}$ (top figures) and 2) $J^{0,0}_{\mathrm{GMD}}$, $J^{1,0}_{\mathrm{GMD}}$, $J^{1,1}_{\mathrm{GMD}}$, $J^{2,1}_{\mathrm{GMD}}$ (bottom figures) on the Isolet dataset. The horizontal axis represents the number of selected features ($|\boldsymbol{S}|$). The vertical axis represents the mean error rates $\bar{e}$ when using the selected features in the classification task.

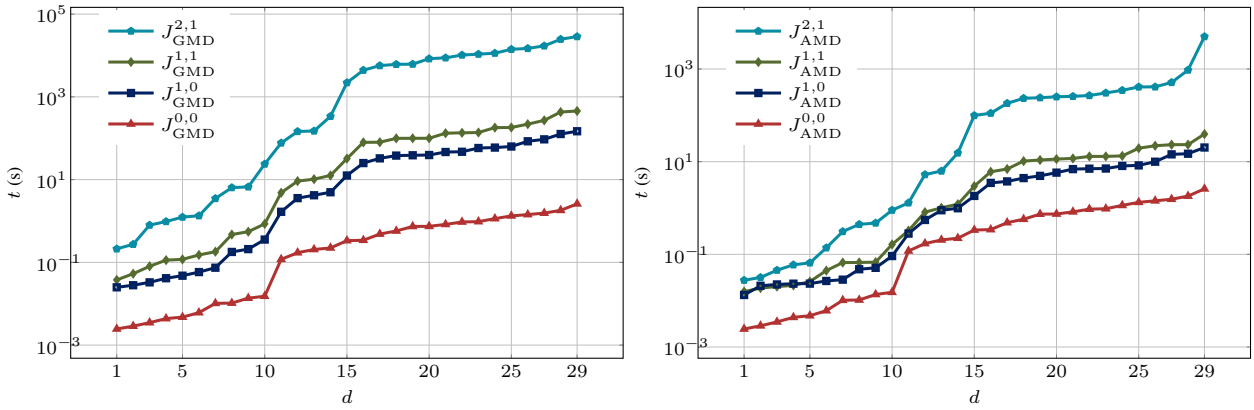

Figure A4: The CPU running time (in seconds) of the methods based on geometric mean and arithmetic mean assumptions with different number of conditional features. The horizontal axis represents the indices of the data sets d, and the vertical axis represents the running time t.

Table A1: The comparison between the feature selection methods based on AMD, GMD and FID in terms of the SVM classification error rates (mean ± std %). The statistically significantly best error rates are highlighted in bold. The last row ($\bar{r}$) computes the average ranking of each method across the datasets.

Data	Methods Based on AMD				Methods Based on GMD				Methods Based on FID			$J^{0,0}s$
	$\tilde{J}^1_{\mathrm{AMD}}$	$J^{2,1}_{\mathrm{AMD}}$	$J^{1,1}_{\mathrm{AMD}}$	$J^{1,0}_{\mathrm{AMD}}$	VMI_{pairwise}	RMRMR	JMI	MRMR	$\tilde{J}^1_{\mathrm{FID}}$	CIFE	MIFS	MIM
d_1	2.00 ± 1.63	2.90 ± 2.23	**1.74 ± 1.73**	2.96 ± 1.93	2.32 ± 2.04	2.28 ± 2.02	1.84 ± 1.71	**1.71 ± 1.55**	2.16 ± 1.89	12.04 ± 4.20	14.55 ± 4.94	2.67 ± 2.56
d_2	3.83 ± 0.77	3.82 ± 0.86	4.09 ± 0.93	**3.71 ± 0.86**	3.79 ± 0.76	3.84 ± 0.79	4.02 ± 0.85	**3.76 ± 0.75**	4.05 ± 0.83	5.05 ± 1.34	4.82 ± 1.14	4.55 ± 1.35
d_3	**7.35 ± 2.50**	9.04 ± 2.90	9.85 ± 3.13	8.05 ± 3.08	7.60 ± 3.36	9.51 ± 3.14	10.05 ± 3.10	8.33 ± 3.25	9.28 ± 4.38	33.94 ± 4.50	21.16 ± 3.30	8.41 ± 3.04
d_4	**1.92 ± 2.10**	2.99 ± 2.83	4.43 ± 3.34	3.86 ± 3.26	2.77 ± 2.65	3.95 ± 3.10	6.09 ± 3.79	5.20 ± 3.76	3.07 ± 2.96	4.76 ± 3.80	11.37 ± 2.95	14.01 ± 11.17
d_5	12.77 ± 4.15	16.14 ± 4.55	15.36 ± 4.37	17.91 ± 4.25	**11.46 ± 3.81**	17.43 ± 4.14	16.53 ± 4.46	19.25 ± 4.50	12.88 ± 4.20	24.18 ± 7.15	21.47 ± 6.04	21.90 ± 5.23
d_6	6.16 ± 1.05	**5.51 ± 1.12**	6.68 ± 1.22	**5.57 ± 1.04**	6.40 ± 1.18	6.36 ± 1.19	6.72 ± 1.20	6.25 ± 1.18	6.62 ± 1.14	5.84 ± 1.07	5.99 ± 1.06	6.36 ± 1.20
d_7	**9.83 ± 4.73**	11.07 ± 4.99	13.71 ± 4.83	10.80 ± 4.37	16.79 ± 5.36	11.11 ± 4.84	12.48 ± 4.87	10.93 ± 4.63	16.13 ± 5.65	41.30 ± 8.06	42.97 ± 7.52	19.12 ± 11.72
d_8	**16.48 ± 2.24**	**16.67 ± 2.12**	**16.30 ± 2.05**	**16.65 ± 2.22**	16.83 ± 2.19	**16.59 ± 2.10**	**16.64 ± 2.17**	**16.37 ± 2.20**	**16.39 ± 2.13**	**16.66 ± 2.19**	17.26 ± 2.18	**16.41 ± 2.15**
d_9	14.86 ± 2.92	12.37 ± 1.74	17.57 ± 2.47	13.42 ± 1.93	15.08 ± 2.43	12.73 ± 1.73	16.96 ± 2.71	13.33 ± 1.95	14.66 ± 2.58	15.70 ± 2.06	**11.80 ± 1.79**	13.31 ± 1.83
d_{10}	**8.58 ± 3.43**	12.09 ± 4.20	16.83 ± 5.38	15.45 ± 4.53	9.41 ± 4.30	14.48 ± 4.65	21.47 ± 5.15	17.92 ± 3.92	13.76 ± 5.54	20.27 ± 5.93	17.78 ± 4.28	31.16 ± 12.86
d_{11}	5.73 ± 0.46	5.52 ± 0.51	5.69 ± 0.51	5.55 ± 0.50	5.71 ± 0.46	**5.46 ± 0.53**	5.75 ± 0.45	**5.49 ± 0.50**	5.70 ± 0.47	5.70 ± 0.53	5.67 ± 0.46	5.73 ± 0.45
d_{12}	**16.52 ± 0.99**	**16.50 ± 1.00**	16.89 ± 1.27	16.85 ± 1.24	16.56 ± 0.98	16.78 ± 1.15	17.10 ± 1.25	17.10 ± 1.30	16.76 ± 1.22	16.59 ± 1.10	16.57 ± 0.97	18.05 ± 1.92
d_{13}	1.85 ± 2.27	1.14 ± 1.46	1.01 ± 1.25	**0.62 ± 1.01**	2.00 ± 1.99	0.82 ± 1.19	2.03 ± 1.66	1.13 ± 1.32	0.88 ± 1.72	3.18 ± 2.90	3.36 ± 2.44	3.25 ± 2.14
d_{14}	4.20 ± 1.59	3.82 ± 1.37	**3.59 ± 1.27**	3.86 ± 1.39	**3.63 ± 1.26**	4.04 ± 1.46	3.88 ± 1.49	3.98 ± 1.50	3.86 ± 1.42	12.20 ± 2.64	6.84 ± 2.33	6.09 ± 2.32
d_{15}	50.40 ± 10.32	46.34 ± 10.17	48.67 ± 11.50	**43.01 ± 10.38**	52.06 ± 9.35	45.10 ± 10.32	48.83 ± 10.77	44.61 ± 9.27	54.06 ± 7.97	55.48 ± 9.15	52.97 ± 8.29	45.82 ± 9.92
d_{16}	3.76 ± 2.18	**3.59 ± 2.02**	4.29 ± 2.08	3.68 ± 1.93	5.09 ± 2.73	3.83 ± 2.22	4.67 ± 2.29	3.79 ± 1.94	5.39 ± 2.73	31.26 ± 5.49	7.65 ± 2.97	11.28 ± 6.31
d_{17}	**21.96 ± 7.79**	24.70 ± 7.51	26.34 ± 6.14	24.49 ± 6.42	33.01 ± 7.48	24.15 ± 6.47	26.19 ± 6.38	26.05 ± 5.99	32.40 ± 9.86	78.22 ± 8.11	47.69 ± 6.45	29.84 ± 6.34
d_{18}	8.78 ± 4.42	**6.20 ± 3.66**	7.84 ± 4.07	6.96 ± 3.34	9.75 ± 4.55	7.71 ± 4.14	8.62 ± 4.68	7.08 ± 3.02	10.99 ± 4.44	18.85 ± 3.29	10.28 ± 4.52	9.66 ± 5.16
d_{19}	15.87 ± 2.67	14.92 ± 2.50	**14.44 ± 2.44**	**14.52 ± 2.36**	**14.55 ± 2.40**	14.71 ± 2.57	14.69 ± 2.42	**14.50 ± 2.37**	14.72 ± 2.29	**14.33 ± 2.43**	14.62 ± 2.43	15.30 ± 2.53
d_{20}	13.41 ± 4.29	17.23 ± 5.23	21.82 ± 7.31	18.96 ± 6.30	**13.28 ± 4.44**	19.75 ± 6.47	22.88 ± 8.36	21.49 ± 7.61	16.45 ± 6.28	26.46 ± 6.19	15.08 ± 4.24	24.58 ± 10.33
d_{21}	24.93 ± 2.99	26.04 ± 3.09	23.89 ± 3.05	27.22 ± 3.56	**23.39 ± 3.13**	25.42 ± 3.40	24.20 ± 3.15	26.97 ± 3.48	24.93 ± 3.09	24.00 ± 3.10	27.45 ± 3.42	25.29 ± 3.25
d_{22}	0.05 ± 0.35	0.05 ± 0.34	**0.00 ± 0.10**	0.02 ± 0.22	0.01 ± 0.17	0.01 ± 0.15	0.01 ± 0.16	0.02 ± 0.21	0.06 ± 0.41	0.18 ± 0.60	**0.00 ± 0.11**	0.10 ± 0.47
d_{23}	16.56 ± 2.20	**16.24 ± 2.39**	16.62 ± 2.06	**16.12 ± 2.19**	16.89 ± 2.35	**16.18 ± 2.17**	16.43 ± 2.27	**16.11 ± 2.33**	**16.37 ± 2.16**	16.52 ± 2.51	16.57 ± 2.15	18.38 ± 3.61
d_{24}	12.16 ± 0.57	**12.11 ± 0.59**	**12.10 ± 0.61**	12.15 ± 0.59	12.14 ± 0.55	**12.09 ± 0.62**	**12.10 ± 0.61**	**12.11 ± 0.61**	12.14 ± 0.60	13.16 ± 0.65	13.10 ± 0.69	12.13 ± 0.61
d_{25}	**8.94 ± 4.65**	13.94 ± 6.11	17.10 ± 6.16	16.52 ± 4.37	13.54 ± 5.74	16.30 ± 5.28	20.05 ± 5.40	18.52 ± 5.40	16.14 ± 6.38	19.34 ± 6.16	28.84 ± 6.36	27.00 ± 6.03
d_{26}	6.54 ± 4.58	6.12 ± 3.19	7.89 ± 3.63	**6.03 ± 3.67**	6.63 ± 3.96	8.04 ± 3.72	10.24 ± 3.08	9.70 ± 3.49	6.96 ± 6.21	26.88 ± 5.60	9.08 ± 4.22	14.55 ± 4.57
d_{27}	2.57 ± 1.06	2.51 ± 1.22	2.27 ± 1.09	2.67 ± 1.23	2.60 ± 1.17	2.62 ± 1.20	2.85 ± 1.17	3.54 ± 1.70	2.79 ± 1.22	**1.65 ± 1.51**	3.74 ± 2.73	5.68 ± 3.47
d_{28}	15.81 ± 0.62	15.89 ± 0.66	**15.76 ± 0.66**	16.00 ± 0.75	**15.78 ± 0.64**	15.81 ± 0.67	**15.78 ± 0.64**	**15.79 ± 0.66**	15.83 ± 0.66	22.25 ± 2.78	27.31 ± 5.90	15.91 ± 0.93
d_{29}	6.01 ± 1.67	**4.74 ± 1.67**	5.76 ± 1.61	**4.91 ± 1.63**	5.78 ± 1.67	5.39 ± 1.65	5.62 ± 1.90	5.04 ± 1.62	**4.62 ± 1.61**	6.22 ± 1.62	5.34 ± 1.71	5.57 ± 1.70
$\bar{r}$	4.38	**3.59**	5.21	4.10	4.72	4.28	6.48	4.66	5.55	8.79	8.45	8.66

Table A2: The comparison between the feature selection methods based on AMD, GMD and FID in terms of the KNN classification error rates (mean $\pm$ std %). The statistically significantly best error rates are highlighted in bold. The last row ($\bar{r}$) computes the average ranking of each method across the datasets.

Data	Methods Based on AMD				Methods Based on GMD				Methods Based on FID			$J^{0,0}s$
	$\tilde{J}^1_{\mathrm{AMD}}$	$J^{2,1}_{\mathrm{AMD}}$	$J^{1,1}_{\mathrm{AMD}}$	$J^{1,0}_{\mathrm{AMD}}$	VMI_{pairwise}	RMRMR	JMI	MRMR	$\tilde{J}^1_{\mathrm{FID}}$	CIFE	MIFS	MIM
d_1	2.58 $\pm$ 1.88	3.38 $\pm$ 2.26	2.78 $\pm$ 1.81	**2.39 $\pm$ 1.81**	4.28 $\pm$ 2.35	**2.35 $\pm$ 1.80**	2.57 $\pm$ 1.74	**2.34 $\pm$ 1.76**	2.47 $\pm$ 1.82	19.27 $\pm$ 5.35	14.08 $\pm$ 4.40	3.40 $\pm$ 2.30
d_2	4.71 $\pm$ 0.88	5.02 $\pm$ 0.92	4.84 $\pm$ 0.95	5.01 $\pm$ 1.12	4.72 $\pm$ 0.92	4.71 $\pm$ 0.91	**4.66 $\pm$ 0.91**	**4.63 $\pm$ 0.88**	4.84 $\pm$ 0.92	7.23 $\pm$ 1.74	6.98 $\pm$ 1.55	5.06 $\pm$ 1.09
d_3	**8.07 $\pm$ 3.31**	9.65 $\pm$ 2.89	10.40 $\pm$ 2.95	9.07 $\pm$ 2.80	9.50 $\pm$ 4.08	9.79 $\pm$ 2.69	10.36 $\pm$ 3.05	9.54 $\pm$ 2.97	11.72 $\pm$ 6.25	37.70 $\pm$ 3.76	25.99 $\pm$ 3.33	10.32 $\pm$ 3.23
d_4	**0.64 $\pm$ 0.70**	0.94 $\pm$ 0.82	1.38 $\pm$ 1.02	1.20 $\pm$ 1.00	0.79 $\pm$ 0.98	1.33 $\pm$ 1.01	2.25 $\pm$ 1.15	1.95 $\pm$ 1.15	1.05 $\pm$ 0.81	1.28 $\pm$ 1.07	4.66 $\pm$ 0.81	6.14 $\pm$ 5.54
d_5	**8.53 $\pm$ 3.27**	10.31 $\pm$ 3.59	10.42 $\pm$ 3.60	12.37 $\pm$ 3.80	9.85 $\pm$ 3.52	11.57 $\pm$ 3.72	13.03 $\pm$ 3.85	13.98 $\pm$ 3.96	9.50 $\pm$ 3.51	12.59 $\pm$ 4.39	20.54 $\pm$ 5.87	17.16 $\pm$ 4.39
d_6	6.82 $\pm$ 1.38	7.14 $\pm$ 1.28	6.22 $\pm$ 1.33	7.09 $\pm$ 1.19	6.61 $\pm$ 1.26	6.86 $\pm$ 1.31	6.25 $\pm$ 1.42	6.91 $\pm$ 1.35	**6.04 $\pm$ 1.35**	7.22 $\pm$ 1.31	7.03 $\pm$ 1.35	6.34 $\pm$ 1.28
d_7	15.52 $\pm$ 5.20	14.13 $\pm$ 5.02	15.00 $\pm$ 5.08	12.35 $\pm$ 4.74	18.87 $\pm$ 6.36	**11.69 $\pm$ 4.49**	13.72 $\pm$ 4.62	12.51 $\pm$ 4.82	17.66 $\pm$ 5.62	49.75 $\pm$ 8.12	49.03 $\pm$ 14.33	14.96 $\pm$ 8.03
d_8	**19.80 $\pm$ 2.53**	21.08 $\pm$ 2.26	**19.96 $\pm$ 2.52**	20.89 $\pm$ 2.35	20.47 $\pm$ 2.30	21.20 $\pm$ 2.28	**19.56 $\pm$ 2.36**	20.67 $\pm$ 2.44	20.16 $\pm$ 2.36	20.03 $\pm$ 2.75	**19.49 $\pm$ 2.31**	20.36 $\pm$ 2.39
d_9	14.14 $\pm$ 2.21	13.67 $\pm$ 1.93	13.14 $\pm$ 2.03	14.37 $\pm$ 2.08	**13.09 $\pm$ 2.23**	13.96 $\pm$ 1.97	13.57 $\pm$ 2.05	13.91 $\pm$ 2.29	13.48 $\pm$ 2.02	13.89 $\pm$ 2.03	13.80 $\pm$ 2.33	**13.00 $\pm$ 2.11**
d_{10}	**15.76 $\pm$ 3.77**	18.03 $\pm$ 3.73	25.98 $\pm$ 4.16	22.29 $\pm$ 3.51	17.95 $\pm$ 4.86	21.26 $\pm$ 3.85	30.25 $\pm$ 3.50	25.09 $\pm$ 2.65	25.27 $\pm$ 5.48	39.08 $\pm$ 4.85	30.85 $\pm$ 3.14	39.27 $\pm$ 10.69
d_{11}	4.48 $\pm$ 0.60	4.97 $\pm$ 1.63	**3.99 $\pm$ 0.82**	5.53 $\pm$ 1.69	4.48 $\pm$ 0.59	4.87 $\pm$ 1.68	4.25 $\pm$ 0.86	5.07 $\pm$ 1.66	4.05 $\pm$ 1.02	4.23 $\pm$ 1.22	6.76 $\pm$ 1.36	4.97 $\pm$ 0.82
d_{12}	14.59 $\pm$ 1.43	**14.28 $\pm$ 1.34**	14.68 $\pm$ 1.51	14.47 $\pm$ 1.39	14.47 $\pm$ 1.31	14.88 $\pm$ 1.79	15.01 $\pm$ 1.63	15.26 $\pm$ 1.74	14.67 $\pm$ 1.72	14.91 $\pm$ 1.35	14.79 $\pm$ 1.19	16.25 $\pm$ 2.58
d_{13}	0.53 $\pm$ 1.03	0.91 $\pm$ 1.31	1.09 $\pm$ 1.47	0.52 $\pm$ 0.96	**0.35 $\pm$ 0.90**	0.48 $\pm$ 1.02	1.66 $\pm$ 1.61	1.07 $\pm$ 1.33	0.54 $\pm$ 1.09	3.68 $\pm$ 3.20	6.93 $\pm$ 5.28	3.07 $\pm$ 1.90
d_{14}	4.91 $\pm$ 1.52	5.03 $\pm$ 1.59	3.87 $\pm$ 1.39	4.17 $\pm$ 1.52	4.54 $\pm$ 1.47	4.37 $\pm$ 1.56	**3.54 $\pm$ 1.45**	3.80 $\pm$ 1.38	4.48 $\pm$ 1.51	14.98 $\pm$ 2.69	10.52 $\pm$ 2.34	6.32 $\pm$ 2.70
d_{15}	45.57 $\pm$ 9.74	**40.08 $\pm$ 10.41**	43.85 $\pm$ 9.87	46.04 $\pm$ 10.04	48.12 $\pm$ 8.33	43.44 $\pm$ 9.74	44.38 $\pm$ 10.24	45.29 $\pm$ 12.13	49.32 $\pm$ 7.94	49.66 $\pm$ 7.85	53.70 $\pm$ 8.35	48.96 $\pm$ 9.84
d_{16}	**5.81 $\pm$ 2.35**	**5.80 $\pm$ 2.32**	5.93 $\pm$ 2.40	**5.77 $\pm$ 2.32**	6.03 $\pm$ 3.35	**5.74 $\pm$ 2.35**	7.03 $\pm$ 2.73	6.47 $\pm$ 2.56	8.02 $\pm$ 3.28	35.74 $\pm$ 5.67	14.82 $\pm$ 3.96	15.16 $\pm$ 5.24
d_{17}	28.23 $\pm$ 7.32	28.22 $\pm$ 6.97	29.38 $\pm$ 6.61	28.56 $\pm$ 6.56	39.70 $\pm$ 7.95	**26.25 $\pm$ 6.83**	30.42 $\pm$ 7.00	30.91 $\pm$ 6.29	40.95 $\pm$ 8.03	73.96 $\pm$ 7.08	52.35 $\pm$ 6.85	35.33 $\pm$ 7.19
d_{18}	17.48 $\pm$ 5.48	**14.19 $\pm$ 4.07**	16.22 $\pm$ 4.35	14.60 $\pm$ 3.69	19.47 $\pm$ 5.01	15.97 $\pm$ 4.82	18.01 $\pm$ 4.58	15.08 $\pm$ 3.32	23.21 $\pm$ 3.56	33.51 $\pm$ 2.90	18.53 $\pm$ 4.72	18.92 $\pm$ 5.14
d_{19}	10.01 $\pm$ 2.20	12.19 $\pm$ 2.59	**9.98 $\pm$ 2.13**	10.05 $\pm$ 1.95	10.37 $\pm$ 2.16	11.65 $\pm$ 2.47	10.29 $\pm$ 1.96	10.05 $\pm$ 1.97	10.22 $\pm$ 2.19	**9.76 $\pm$ 2.22**	10.41 $\pm$ 2.06	13.03 $\pm$ 3.64
d_{20}	**14.86 $\pm$ 5.10**	18.67 $\pm$ 6.69	23.07 $\pm$ 8.56	20.69 $\pm$ 7.75	15.11 $\pm$ 4.92	21.48 $\pm$ 7.87	24.45 $\pm$ 9.86	23.19 $\pm$ 8.91	17.67 $\pm$ 6.45	27.23 $\pm$ 6.01	17.69 $\pm$ 5.07	26.45 $\pm$ 11.97
d_{21}	15.08 $\pm$ 2.65	15.25 $\pm$ 2.74	15.75 $\pm$ 2.75	16.95 $\pm$ 3.20	16.10 $\pm$ 2.56	**13.91 $\pm$ 2.68**	15.62 $\pm$ 2.71	15.19 $\pm$ 2.87	15.30 $\pm$ 2.61	17.40 $\pm$ 2.77	19.52 $\pm$ 4.03	15.39 $\pm$ 2.92
d_{22}	0.43 $\pm$ 1.06	**0.27 $\pm$ 0.81**	0.48 $\pm$ 1.11	**0.29 $\pm$ 0.82**	0.93 $\pm$ 1.55	**0.29 $\pm$ 0.80**	0.46 $\pm$ 1.05	**0.33 $\pm$ 0.85**	0.88 $\pm$ 1.54	7.61 $\pm$ 5.63	2.94 $\pm$ 4.43	0.39 $\pm$ 1.06
d_{23}	**20.23 $\pm$ 2.66**	20.60 $\pm$ 2.48	**19.99 $\pm$ 2.61**	20.96 $\pm$ 2.53	**20.18 $\pm$ 2.64**	20.52 $\pm$ 2.42	**20.07 $\pm$ 2.55**	20.57 $\pm$ 2.50	20.43 $\pm$ 2.75	20.33 $\pm$ 2.61	**20.24 $\pm$ 2.48**	21.35 $\pm$ 2.90
d_{24}	**22.89 $\pm$ 4.82**	23.21 $\pm$ 4.78	23.06 $\pm$ 4.75	23.11 $\pm$ 4.85	22.96 $\pm$ 4.69	23.09 $\pm$ 4.71	23.12 $\pm$ 4.77	23.07 $\pm$ 4.76	23.07 $\pm$ 4.73	23.31 $\pm$ 3.55	24.22 $\pm$ 2.98	23.17 $\pm$ 4.83
d_{25}	**17.53 $\pm$ 4.80**	20.83 $\pm$ 5.47	23.69 $\pm$ 4.07	23.07 $\pm$ 5.50	22.49 $\pm$ 5.22	23.33 $\pm$ 5.13	27.91 $\pm$ 3.80	26.30 $\pm$ 4.56	27.95 $\pm$ 4.72	32.04 $\pm$ 7.78	39.48 $\pm$ 4.00	35.70 $\pm$ 5.25
d_{26}	28.57 $\pm$ 4.68	21.44 $\pm$ 4.23	23.81 $\pm$ 3.91	**19.36 $\pm$ 4.15**	29.06 $\pm$ 4.83	22.54 $\pm$ 3.89	23.29 $\pm$ 3.99	23.55 $\pm$ 3.94	33.55 $\pm$ 4.50	65.87 $\pm$ 6.81	32.16 $\pm$ 4.63	25.35 $\pm$ 4.40
d_{27}	4.30 $\pm$ 1.62	4.21 $\pm$ 1.62	4.72 $\pm$ 1.71	**3.86 $\pm$ 1.53**	4.84 $\pm$ 1.55	4.62 $\pm$ 1.54	5.17 $\pm$ 1.58	4.56 $\pm$ 1.77	7.38 $\pm$ 2.33	10.62 $\pm$ 2.78	15.49 $\pm$ 5.70	7.54 $\pm$ 3.77
d_{28}	22.30 $\pm$ 1.57	**22.18 $\pm$ 1.48**	22.24 $\pm$ 1.50	22.59 $\pm$ 1.33	22.30 $\pm$ 1.57	22.22 $\pm$ 1.49	22.26 $\pm$ 1.53	22.32 $\pm$ 1.51	22.46 $\pm$ 1.53	31.79 $\pm$ 3.55	39.05 $\pm$ 7.21	22.33 $\pm$ 1.56
d_{29}	5.05 $\pm$ 1.65	4.29 $\pm$ 1.54	**3.99 $\pm$ 1.55**	4.38 $\pm$ 1.60	5.12 $\pm$ 1.64	**4.12 $\pm$ 1.51**	**4.02 $\pm$ 1.59**	4.27 $\pm$ 1.50	4.32 $\pm$ 1.58	5.28 $\pm$ 1.75	5.06 $\pm$ 1.83	**3.97 $\pm$ 1.42**
$\bar{r}$	**3.76**	4.62	4.59	5.03	5.28	4.24	5.41	5.10	5.90	9.48	9.48	8.07

Table A3: The comparison between the AMD based methods against other competitive methods in terms of the SVM classification error rates (mean ± std %). The statistically significantly best error rates are highlighted in bold. The last row ($\bar{r}$) computes the average ranking of each method across the datasets.

Data	$\tilde{J}^1_{\mathrm{AMD}}$	$J^{2,1}_{\mathrm{AMD}}$	MRI	$\mathcal{SPEC}_{\mathrm{CMI}}$	CMIM	Trace	SPEC
d_1	2.00 ± 1.63	2.90 ± 2.23	1.92 ± 1.64	3.13 ± 2.20	**1.10 ± 1.73**	2.92 ± 2.31	19.06 ± 5.02
d_2	**3.83 ± 0.77**	**3.82 ± 0.86**	4.00 ± 0.81	4.16 ± 0.94	3.96 ± 0.81	5.37 ± 1.44	5.74 ± 1.50
d_3	**7.35 ± 2.50**	9.04 ± 2.90	10.29 ± 4.03	12.74 ± 3.59	9.28 ± 3.48	20.66 ± 10.06	27.63 ± 8.75
d_4	**1.92 ± 2.10**	2.99 ± 2.83	4.12 ± 3.69	10.62 ± 9.17	2.22 ± 2.98	18.05 ± 14.31	34.86 ± 14.49
d_5	12.77 ± 4.15	16.14 ± 4.55	**10.86 ± 3.87**	15.61 ± 4.79	18.69 ± 4.53	18.20 ± 4.74	19.66 ± 8.21
d_6	6.16 ± 1.05	**5.51 ± 1.12**	**5.50 ± 1.16**	**5.58 ± 1.08**	6.57 ± 1.22	6.37 ± 1.03	6.40 ± 1.13
d_7	**9.83 ± 4.73**	11.07 ± 4.99	18.41 ± 5.53	36.34 ± 7.53	15.95 ± 6.12	28.19 ± 15.20	25.61 ± 9.75
d_8	**16.48 ± 2.24**	16.67 ± 2.12	**16.20 ± 2.38**	**16.37 ± 1.99**	**16.46 ± 2.38**	**16.22 ± 2.13**	16.96 ± 2.30
d_9	14.86 ± 2.92	**12.37 ± 1.74**	17.07 ± 2.34	17.51 ± 2.54	12.99 ± 1.72	12.71 ± 2.00	13.20 ± 2.14
d_{10}	**8.58 ± 3.43**	12.09 ± 4.20	14.16 ± 4.98	24.22 ± 9.44	11.38 ± 4.59	22.33 ± 6.98	46.56 ± 13.45
d_{11}	5.73 ± 0.46	5.52 ± 0.51	5.71 ± 0.49	5.75 ± 0.48	5.70 ± 0.47	**5.39 ± 0.49**	18.20 ± 13.63
d_{12}	**16.52 ± 0.99**	**16.50 ± 1.00**	16.56 ± 1.22	17.50 ± 2.68	16.71 ± 1.09	17.87 ± 1.78	24.21 ± 9.15
d_{13}	1.85 ± 2.27	1.14 ± 1.46	**0.97 ± 1.26**	2.62 ± 1.96	1.28 ± 2.13	2.95 ± 2.14	22.70 ± 10.87
d_{14}	4.20 ± 1.59	**3.82 ± 1.37**	4.27 ± 1.43	5.11 ± 1.65	4.55 ± 1.52	11.53 ± 4.69	9.43 ± 4.20
d_{15}	50.40 ± 10.32	46.34 ± 10.17	53.67 ± 8.01	52.18 ± 9.10	48.92 ± 10.71	**43.33 ± 10.34**	61.40 ± 8.35
d_{16}	3.76 ± 2.18	**3.59 ± 2.02**	4.54 ± 2.58	9.47 ± 6.42	4.86 ± 2.42	11.68 ± 5.67	26.64 ± 9.22
d_{17}	**21.96 ± 7.79**	24.70 ± 7.51	28.47 ± 7.42	30.55 ± 7.51	26.87 ± 6.97	35.36 ± 6.16	65.88 ± 9.51
d_{18}	8.78 ± 4.42	6.20 ± 3.66	9.19 ± 4.28	20.86 ± 10.24	**5.40 ± 3.94**	15.37 ± 11.32	18.54 ± 8.35
d_{19}	15.87 ± 2.67	14.92 ± 2.50	**14.54 ± 2.35**	**14.58 ± 2.40**	**14.71 ± 2.39**	16.42 ± 2.52	19.00 ± 5.28
d_{20}	**13.41 ± 4.29**	17.23 ± 5.23	21.22 ± 6.84	24.00 ± 9.91	14.43 ± 4.80	24.73 ± 10.66	22.21 ± 12.69
d_{21}	24.93 ± 2.99	26.04 ± 3.09	**23.24 ± 3.12**	23.55 ± 2.92	24.63 ± 3.13	24.12 ± 3.17	27.53 ± 3.69
d_{22}	0.05 ± 0.35	0.05 ± 0.34	**0.02 ± 0.19**	**0.01 ± 0.16**	0.04 ± 0.30	0.30 ± 0.88	7.92 ± 10.40
d_{23}	16.56 ± 2.20	**16.24 ± 2.39**	16.59 ± 2.19	16.59 ± 2.21	**16.48 ± 2.24**	18.38 ± 3.40	21.11 ± 2.74
d_{24}	12.16 ± 0.57	12.11 ± 0.59	12.11 ± 0.62	12.09 ± 0.61	**12.05 ± 0.63**	12.43 ± 0.57	19.02 ± 9.61
d_{25}	**8.94 ± 4.65**	13.94 ± 6.11	15.43 ± 5.26	23.32 ± 6.11	12.41 ± 5.76	23.58 ± 6.47	36.27 ± 12.56
d_{26}	6.54 ± 4.58	**6.12 ± 3.19**	6.56 ± 2.89	12.59 ± 5.79	9.22 ± 3.96	14.20 ± 7.09	54.54 ± 8.09
d_{27}	2.57 ± 1.06	2.51 ± 1.22	2.67 ± 1.17	5.09 ± 1.88	**1.81 ± 0.95**	3.79 ± 2.63	9.56 ± 5.06
d_{28}	15.81 ± 0.62	15.89 ± 0.66	**15.79 ± 0.66**	**15.76 ± 0.67**	15.85 ± 0.76	15.91 ± 0.85	17.89 ± 1.36
d_{29}	6.01 ± 1.67	4.74 ± 1.67	5.31 ± 1.80	5.37 ± 1.70	**4.25 ± 1.66**	**4.27 ± 1.67**	5.74 ± 1.68
$\bar{r}$	2.58	**2.34**	2.86	4.24	2.72	4.82	6.52

Table A4: The comparison between the AMD based methods against other competitive methods in terms of the KNN classification error rates (mean ± std %). The statistically significantly best error rates are highlighted in bold. The last row ($\bar{r}$) computes the average ranking of each method across the datasets.

Data	$\tilde{J}^1_{\mathrm{AMD}}$	$J^{2,1}_{\mathrm{AMD}}$	MRI	$\mathcal{SPEC}_{\mathrm{CMI}}$	CMIM	Trace	SPEC
d_1	2.58 ± 1.88	3.38 ± 2.26	4.07 ± 2.73	4.10 ± 2.53	**2.21 ± 1.80**	4.53 ± 2.55	22.36 ± 4.83
d_2	**4.71 ± 0.88**	5.02 ± 0.92	4.90 ± 0.91	4.98 ± 0.99	5.09 ± 0.99	5.78 ± 1.37	6.46 ± 1.47
d_3	**8.07 ± 3.31**	9.65 ± 2.89	10.87 ± 3.07	14.07 ± 3.30	9.89 ± 2.88	24.55 ± 10.04	36.37 ± 9.24
d_4	**0.64 ± 0.70**	0.94 ± 0.82	1.40 ± 1.11	4.43 ± 3.46	0.90 ± 1.08	10.41 ± 9.32	21.23 ± 10.12
d_5	**8.53 ± 3.27**	10.31 ± 3.59	11.05 ± 3.64	12.31 ± 4.03	10.80 ± 3.76	14.16 ± 4.08	29.86 ± 6.36
d_6	6.82 ± 1.38	7.14 ± 1.28	6.55 ± 1.30	6.56 ± 1.36	**6.20 ± 1.29**	6.80 ± 1.33	7.73 ± 1.49
d_7	15.52 ± 5.20	**14.13 ± 5.02**	18.34 ± 5.56	42.10 ± 7.67	17.70 ± 5.63	21.68 ± 8.14	31.80 ± 8.10
d_8	**19.80 ± 2.53**	21.08 ± 2.26	**19.89 ± 2.32**	**19.54 ± 2.32**	20.18 ± 2.31	20.66 ± 2.30	20.20 ± 2.32
d_9	14.14 ± 2.21	13.67 ± 1.93	13.60 ± 2.36	14.00 ± 2.10	**13.33 ± 2.01**	13.92 ± 2.03	13.58 ± 1.91
d_{10}	**15.76 ± 3.77**	18.03 ± 3.73	23.90 ± 3.84	32.38 ± 8.47	19.64 ± 4.05	27.83 ± 4.95	59.45 ± 9.23
d_{11}	4.48 ± 0.60	4.97 ± 1.63	4.02 ± 0.85	**3.93 ± 0.58**	4.96 ± 0.94	5.35 ± 1.30	16.75 ± 13.03
d_{12}	14.59 ± 1.43	**14.28 ± 1.34**	14.56 ± 1.37	15.44 ± 2.96	14.90 ± 1.58	16.22 ± 2.58	22.47 ± 9.43
d_{13}	0.53 ± 1.03	0.91 ± 1.31	**0.27 ± 0.74**	3.06 ± 2.06	0.70 ± 1.16	2.97 ± 1.92	27.17 ± 8.17
d_{14}	4.91 ± 1.52	5.03 ± 1.59	**4.75 ± 1.54**	5.52 ± 1.87	5.36 ± 1.61	13.40 ± 5.00	11.29 ± 4.44
d_{15}	45.57 ± 9.74	**40.08 ± 10.41**	46.99 ± 8.64	47.64 ± 8.47	47.12 ± 9.80	43.45 ± 11.06	63.22 ± 11.26
d_{16}	**5.81 ± 2.35**	**5.80 ± 2.32**	7.34 ± 2.84	14.68 ± 5.68	6.78 ± 2.80	16.94 ± 5.72	28.72 ± 6.16
d_{17}	**28.23 ± 7.32**	**28.22 ± 6.97**	29.10 ± 5.94	33.17 ± 7.06	29.09 ± 6.91	32.29 ± 6.60	70.08 ± 6.36
d_{18}	17.48 ± 5.48	14.19 ± 4.07	20.68 ± 3.75	33.67 ± 9.16	**14.06 ± 4.36**	25.92 ± 11.38	31.27 ± 6.51
d_{19}	**10.01 ± 2.20**	12.19 ± 2.59	**10.07 ± 2.00**	10.31 ± 2.09	10.35 ± 2.21	14.31 ± 2.56	19.23 ± 7.63
d_{20}	**14.86 ± 5.10**	18.67 ± 6.69	22.57 ± 8.22	25.56 ± 11.62	15.59 ± 5.50	27.21 ± 12.25	25.72 ± 14.04
d_{21}	**15.08 ± 2.65**	15.25 ± 2.74	16.03 ± 2.56	16.65 ± 2.55	15.51 ± 2.64	**15.22 ± 2.73**	17.95 ± 4.25
d_{22}	0.43 ± 1.06	0.27 ± 0.81	0.66 ± 1.30	0.48 ± 1.08	0.92 ± 1.60	**0.07 ± 0.42**	10.26 ± 12.58
d_{23}	20.23 ± 2.66	20.60 ± 2.48	**19.90 ± 2.62**	**19.91 ± 2.62**	20.48 ± 2.52	21.14 ± 2.92	22.39 ± 2.89
d_{24}	**22.89 ± 4.82**	23.21 ± 4.78	23.07 ± 4.76	23.02 ± 4.72	23.17 ± 4.83	22.95 ± 3.56	29.75 ± 4.63
d_{25}	**17.53 ± 4.80**	20.83 ± 5.47	25.38 ± 5.36	30.75 ± 4.16	21.23 ± 5.00	34.29 ± 4.24	45.61 ± 8.81
d_{26}	28.57 ± 4.68	21.44 ± 4.23	29.45 ± 4.51	25.26 ± 5.24	39.79 ± 5.00	**20.71 ± 4.28**	58.10 ± 5.03
d_{27}	4.30 ± 1.62	4.21 ± 1.62	4.27 ± 1.46	6.33 ± 2.28	**2.53 ± 1.57**	4.99 ± 2.86	16.06 ± 7.81
d_{28}	22.30 ± 1.57	**22.18 ± 1.48**	22.25 ± 1.49	22.24 ± 1.49	22.22 ± 1.44	**22.21 ± 1.42**	25.75 ± 1.63
d_{29}	5.05 ± 1.65	**4.29 ± 1.54**	**4.08 ± 1.52**	**4.08 ± 1.51**	4.48 ± 1.49	**4.25 ± 1.51**	**4.18 ± 1.38**
$\bar{r}$	**2.41**	2.65	3.00	4.17	3.03	4.62	6.31

References

[Balagani and Phoha 2010] Balagani, K. S., and Phoha, V. V. 2010. On the feature selection criterion based on an approximation of multidimensional mutual information. *IEEE Transactions on Pattern Analysis & Machine Intelligence* (7):1342–1343.

[Battiti 1994] Battiti, R. 1994. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks* 5(4):537–550.

[Brown et al. 2012] Brown, G.; Pocock, A.; Zhao, M.-J.; and Luján, M. 2012. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of Machine Learning Research* 13(Jan):27–66.

[Cover and Thomas 2012] Cover, T. M., and Thomas, J. A. 2012. *Elements of information theory*. John Wiley & Sons.

[Kittler 1986] Kittler, J. 1986. Feature selection and extraction. *Handbook of Pattern Recognition and Image Processing* 59–83.

[Lewis 1992] Lewis, D. D. 1992. Feature selection and feature extraction for text categorization. In *Proceedings of the Workshop on Speech and Natural Language*, 212–217. Association for Computational Linguistics.

[Lin and Tang 2006] Lin, D., and Tang, X. 2006. Conditional infomax learning: an integrated framework for feature extraction and fusion. In *European Conference on Computer Vision*, 68–82. Springer.

[Meyer, Schretter, and Bontempi 2008] Meyer, P. E.; Schretter, C.; and Bontempi, G. 2008. Information-theoretic feature selection in microarray data using variable complementarity. *IEEE Journal of Selected Topics in Signal Processing* 2(3):261–274.

[Peng, Long, and Ding 2005] Peng, H.; Long, F.; and Ding, C. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(8):1226–1238.

[Pudil, Novovičova, and Kittler 1994] Pudil, P.; Novovičova, J.; and Kittler, J. 1994. Floating search methods in feature selection. *Pattern Recognition Letters* 15(11):1119–1125.

[Vinh et al. 2016] Vinh, N. X.; Zhou, S.; Chan, J.; and Bailey, J. 2016. Can high-order dependencies improve mutual information based feature selection? *Pattern Recognition* 53:46–58.

[Yang and Moody 1999] Yang, H. H., and Moody, J. 1999. Data visualization and feature selection: New algorithms for nongaussian data. In *Advances in Neural Information Processing Systems*, 687–693.