Table 5: Main notations and their descriptions.

Notation	Description
• Spaces and Labels	
$\bar{\mathbb{R}} = \{[x^{\mathrm{l}}, x^{\mathrm{r}}] \mid x^{\mathrm{l}}, x^{\mathrm{r}} \in \mathbb{R}, x^{\mathrm{l}} \leq x^{\mathrm{r}}\}$	the set of all real-valued intervals
$\bar{\mathbb{R}}^p = \{([x_1^{\mathrm{l}}, x_1^{\mathrm{r}}], \cdots, [x_p^{\mathrm{l}}, x_p^{\mathrm{r}}])^\top\}$	the set of all p-dimension interval-valued vector
$\bar{\mathcal{X}} \subset \bar{\mathbb{R}}^p$	input (feature) space of LIND problem
$\mathcal{X}_v \subset \mathbb{R}^p, v \in [c]$	single-view input (feature) space
$\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_c \subset \mathbb{R}^p \times \cdots \times \mathbb{R}^p$	multi-view input (feature) space
$\mathcal{Y}$	output (label) space
$[K] = \{1, \cdots, K\}$	$1, \cdots, K$ represent the labels in $\mathcal{Y}$
• Distributions	
$\bar{X} = [X^{\mathrm{l}}, X^{\mathrm{r}}]$	interval-valued random variable
$\bar{\mathbf{X}} = (\bar{X}_1, \cdots, \bar{X}_p)^\top$	interval-valued random vector
$\mathbf{X}^{\mathrm{l}} = (X_1^{\mathrm{l}}, \cdots, X_p^{\mathrm{l}})^\top,$	real-valued random vector
$\mathbf{X}^{\mathrm{r}} = (X_1^{\mathrm{r}}, \cdots, X_p^{\mathrm{r}})^\top$	
$\mathcal{D}^{\mathrm{l}}, \mathcal{D}^{\mathrm{r}}$	distribution of real-valued random vector $\mathbf{X}^{\mathrm{l}}, \mathbf{X}^{\mathrm{r}}$
$\bar{\mathcal{D}}$	interval distribution over $\bar{\mathcal{X}}$
$\mathcal{D}$	multi-view distribution over $\mathcal{X}$
$\bar{S}_{\bar{X}} = \{\bar{\mathbf{x}}_i = (\bar{x}_{i1}, \cdots, \bar{x}_{ip})^\top\}_{i=1}^m$	a sample drawn i.i.d. from $\bar{\mathcal{X}}$
$S_{X^v} = \{\mathbf{x}_i^v = (x_{i1}^v, \cdots, x_{ip}^v)^\top\}_{i=1}^m$	the single-view sample drawn i.i.d. from $\mathcal{X}_v$
$S_X = \{\mathbf{X}_i = (\mathbf{x}_i^1, \cdots, \mathbf{x}_i^c)\}_{i=1}^m$	the multi-view sample drawn i.i.d. from $\mathcal{X}$
• Loss Function ad Function Spaces	
$\ell(\cdot, \cdot)$	loss : $\mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}_+$
$\mathcal{H}$	hypothesis space of the LIND problem
$\mathcal{H}_v$	hypothesis space of v-th view, $v = 1, \cdots, c$
$\mathcal{H}_{\mathrm{co}}$	multi-view hypothesis space
f_v	predict function of $\boldsymbol{h}_v \in \mathcal{H}_v, v = 1, \cdots, c$
f_{co}	predict function of $\boldsymbol{h}_{\mathrm{co}} \in \mathcal{H}_{\mathrm{co}}$
• Risks and Complexities	
$R_{\bar{\mathcal{D}}}(\boldsymbol{h})$	risk of $\boldsymbol{h} \in \mathcal{H}$
$R_{\mathcal{D}}(\boldsymbol{h}_{\mathrm{co}})$	risk of $\boldsymbol{h}_{\mathrm{co}} \in \mathcal{H}_{\mathrm{co}}$
$\widehat{\mathcal{R}}_{\bar{S}_{\bar{X}}}(\mathcal{H})$	empirical Rademacher complexity of $\mathcal{H}$ with respect to the sample $\bar{S}_{\bar{X}}$
$\mathcal{R}_{\bar{S}_{\bar{X}}}(\mathcal{H})$	Rademacher complexity of $\bar{\mathcal{H}}$ with respect to the sample $\bar{S}_{\bar{X}}$
$\mathcal{R}_{S_{X^v}}(\mathcal{H}_v)$	Rademacher complexity of $\mathcal{H}_v$ with respect to the sample S_{X^v}
$\mathcal{R}_{S_X}(\mathcal{H}_{co})$	Rademacher complexity of $\mathcal{H}_{co}$ with respect to the sample S_X

A NOTATIONS

In this section, we summarize important notations in Table 5.

To prove Theorem 2, 3 and Corollary 1, for any $\boldsymbol{h}_v \in \mathcal{H}_v$, we let

$$\begin{array}{rl} \boldsymbol{h}_v(\mathbf{x}_i^v): & \mathcal{X}_v \to \mathbb{R}^K \\ & \mathbf{x}_i^v \to (h_{v1}(\mathbf{x}_i^v), \cdots, h_{vK}(\mathbf{x}_i^v))^\top. \end{array}$$

Without loss of generality, we suppose that $\sum_{k=1}^{K} h_{vk}(\mathbf{x}_i^v) = 1$ and the predict function f_v of $\boldsymbol{h}_v$ is defined as

$$f_v(\mathbf{x}_i^v) = \arg\max_{1 \le k \le K} h_{vk}(\mathbf{x}_i^v).$$

Then, for any $\boldsymbol{h}_{\text{co}} \in \mathcal{H}_{\text{co}}$, we let

$$\begin{aligned} \boldsymbol{h}_{\text{co}}(\mathbf{X}_i): \quad & \mathcal{X} \to \mathbb{R}^K \\ & \mathbf{X}_i = (\mathbf{x}_i^1, \cdots, \mathbf{x}_i^c) \to (h_{\text{co}}^1(\mathbf{X}_i), \cdots, h_{\text{co}}^K(\mathbf{X}_i))^\top, \end{aligned}$$

where $\boldsymbol{h}_{\text{co}}^q(\mathbf{X}_i) = \sum_{v=1}^{c} \mathbf{w}_v^{q\top} \boldsymbol{h}_v(\mathbf{x}_i^v), \mathbf{w}_v^q = (w_{v1}^q, \cdots, w_{vK}^q)^\top$ and without loss of generality, we suppose $\sum_{q=1}^{K} h_{\text{co}}^q(\mathbf{X}_i) = 1$. Therefore, we have $\sup_{\boldsymbol{h}_{\text{co}} \in \mathcal{H}_{\text{co}}} \| \boldsymbol{h}_{\text{co}} \|_\infty \le 1$. The predict function f_{co} of $\boldsymbol{h}_{\text{co}}$ is defined as

$$f_{\text{co}}(\mathbf{X}_i) = \arg\max_{1 \le q \le K} h_{\text{co}}^q(\mathbf{X}_i).$$

B PROOFS

In this appendix, we prove Theorem 1, 2, 3 and Corollary 1 in Section 3.2. To prove Theorem 1, we first give some related definitions and prove the Azuma's Inequality and McDiarmid's Inequality of interval-valued random variables.

Definition 3 (Interval Probability Density Function). *Suppose $X^{\text{l}}, X^{\text{r}}$ are two real-valued random variables and have the same continuous pdf $p_X(x)$. We define $\bar{p}_{\bar{X}}(x)$ as the interval pdf of interval-valued random variable $\bar{X}$, where*

$$\bar{p}_{\bar{X}}(x) = \left[\min_{x \in [X^{\text{l}}, X^{\text{r}}]} p_X(x), \max_{x \in [X^{\text{l}}, X^{\text{r}}]} p_X(x) \right].$$

Let $\bar{\mathbf{X}} = (\bar{X}_1, \cdots, \bar{X}_p)^\top$ be a p-interval-valued random vector and the interval pdf of $\bar{X}_j$ is $\bar{p}_{\bar{X}_j}(x), j \in [p]$. Then, we denote the joint interval pdf of $\bar{\mathbf{X}}$ as

$$\bar{p}_{\bar{\mathbf{X}}}(\mathbf{x}) = \left[\prod_{j=1}^{p} \min_{x_j \in [X_j^{\text{l}}, X_j^{\text{r}}]} p_{X_j}(x_j), \prod_{j=1}^{p} \max_{x_j \in [X_j^{\text{l}}, X_j^{\text{r}}]} p_{X_j}(x_j) \right], \mathbf{x} = (x_1, \cdots, x_p)^\top.$$

Definition 4 (Interval Probability Distribution). *Let $\bar{\mathbf{X}} = (\bar{X}_1, \cdots, \bar{X}_p)^\top$ be a p-interval-valued random vector with the joint interval pdf $\bar{p}_{\bar{\mathbf{X}}}(\mathbf{x})$. Let $\mathbf{X}^{\text{l}} = (X_1^{\text{l}}, \cdots, X_p^{\text{l}})^\top, \mathbf{X}^{\text{r}} = (X_1^{\text{r}}, \cdots, X_p^{\text{r}})^\top$ be two real-valued random vectors following probability distribution $\mathcal{D}^{\text{l}}, \mathcal{D}^{\text{r}}$. We define $\bar{\mathcal{D}}$ as the interval probability distribution of $\bar{\mathbf{X}}$ (denoted as $\bar{\mathbf{X}} \sim \bar{\mathcal{D}}$), if*

$$\bar{\mathcal{D}}(\bar{\mathbb{R}}^p) = \bar{\int} \bar{p}_{\bar{\mathbf{X}}}(\mathbf{x}) d\mathbf{x} = 1,$$

where $\bar{\int} \bar{p}_{\bar{\mathbf{X}}}(\mathbf{x}) d\mathbf{x} = \frac{1}{2} \int d\mathcal{D}^{\text{l}}(\mathbf{x}) + \frac{1}{2} \int d\mathcal{D}^{\text{r}}(\mathbf{x})$. Therefore, $\bar{\mathbf{X}} \sim \bar{\mathcal{D}}$ if and only if $\mathbf{X}^{\text{l}} \sim \mathcal{D}^{\text{l}}$ and $\mathbf{X}^{\text{r}} \sim \mathcal{D}^{\text{r}}$. Then, we denote $\mathbb{P}(\bar{\mathbf{X}} \in \bar{B}) = \bar{\mathcal{D}}(\bar{B})$ as the probability of the event $\{\bar{\mathbf{X}} \in \bar{B}\}$, where $\bar{B} \in \bar{\mathcal{B}}$ and $\bar{\mathcal{B}}$ is the Borel σ-algebra in $\bar{\mathbb{R}}^p$ (Jeffreys, 1998).

Definition 5. *Let $\bar{\mathbf{X}} = (\bar{X}_1, \cdots, \bar{X}_p)^\top$ be a p-interval-valued random vector with the joint interval pdf $\bar{p}_{\bar{\mathbf{X}}}(\mathbf{x})$ and $\mathbf{X}^{\text{l}} = (X_1^{\text{l}}, \cdots, X_p^{\text{l}})^\top \sim \mathcal{D}^{\text{l}}, \mathbf{X}^{\text{r}} = (X_1^{\text{r}}, \cdots, X_p^{\text{r}})^\top \sim \mathcal{D}^{\text{r}}$ are two real-valued random vectors. Then, the probability with respect to the function $g : \bar{\mathcal{X}} \to \mathbb{R}_+$ is defined as:*

$$\mathbb{P}(g(\bar{\mathbf{X}}) \ge \varepsilon) = \frac{1}{2} \int_{\text{A}} d\mathcal{D}^{\text{l}}(\mathbf{x}) + \frac{1}{2} \int_{\text{B}} d\mathcal{D}^{\text{r}}(\mathbf{x}),$$

where $\text{A} = \{\mathbf{X}^{\text{l}} \in \mathbb{R}^p : g(\bar{\mathbf{X}}) \ge \varepsilon\}, \text{B} = \{\mathbf{X}^{\text{r}} \in \mathbb{R}^p : g(\bar{\mathbf{X}}) \ge \varepsilon\}$.

Definition 6 (Independence). *The interval-valued random vectors $\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_n$ are said to be (mutually) independent if and only if the real-valued random vectors $\mathbf{X}_1^{\text{l}}, \cdots, \mathbf{X}_n^{\text{l}}, \mathbf{X}_1^{\text{r}}, \cdots, \mathbf{X}_n^{\text{r}}$ are (mutually) independent. Then, we denote $\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_n$ as i.i.d. interval-valued random vectors if and only if $\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_n$ are independent and have the same interval probability distribution.*

Definition 7. *The empirical Rademacher complexity of $\mathcal{H}$ with respect to $\bar{S}_{\bar{X}}$ is defined as:*

$$\widehat{\mathcal{R}}_{\bar{S}_{\bar{X}}}(\mathcal{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{h}\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}\sigma_{ik}h_k(\bar{\mathbf{x}}_i)\right], \tag{4}$$

where $\boldsymbol{\sigma} = [\sigma_{ik}]_{m\times K}$ is a $m \times K$ matrix, with σ_{ik}s independent random variables drawn from the Rademacher distribution, i.e. $\mathbb{P}(\sigma_{ik} = +1) = \mathbb{P}(\sigma_{ik} = -1) = \frac{1}{2}, i \in [m], k \in [K]$. The Rademacher complexity $\mathcal{R}_{\bar{S}_{\bar{X}}}(\mathcal{H})$ is equal to the interval expectation of $\widehat{\mathcal{R}}_{\bar{S}_{\bar{X}}}(\mathcal{H})$.

Definition 8. *A sequence of $V_1, V_2, \cdots$ is a martingale difference sequence with respect to interval-valued random variables $\bar{X}_1, \bar{X}_2, \cdots$ if for any $i > 0$, V_i is a real-value function of $\bar{X}_1, \cdots, \bar{X}_i$ and $\mathbb{E}_{\bar{\mathcal{D}}}[V_{i+1}|\bar{X}_1, \cdots, \bar{X}_i] = 0$.*

Theorem 4 (Azuma's Inequality of Interval-valued Random Variables)**.** *Let $V_1, V_2, \cdots$ be a martingale difference sequence with respect to the interval-valued random variables $\bar{X}_1, \bar{X}_2, \cdots$ and assume that for any $i > 0$ there is a constant $c_i \geq 0$ and Z_i, which is a real-value function of $\bar{X}_1, \cdots, \bar{X}_{i-1}$, satisfies*

$$Z_i \leq V_i \leq Z_i + c_i.$$

Then for any $\varepsilon > 0$ and $m \in N^+$, the following inequalities hold:

$$\begin{aligned} \mathbb{P}\left[\sum_{i=1}^{m} V_i \geq \varepsilon\right] &\leq \exp\frac{-2\varepsilon^2}{\sum\limits_{i=1}^{m} c_i^2}, \\ \mathbb{P}\left[\sum_{i=1}^{m} V_i \leq -\varepsilon\right] &\leq \exp\frac{-2\varepsilon^2}{\sum\limits_{i=1}^{m} c_i^2}. \end{aligned} \tag{5}$$

Proof. Suppose $\bar{X} = [X^{\mathrm{l}}, X^{\mathrm{r}}]$ is an interval-valued random variable. According to Definition 5, we have

$$\begin{aligned} \mathbb{P}(g(\bar{X}) \geq \varepsilon) &= \frac{1}{2}\left(\int_{\mathrm{A}} e^{-tg(\bar{X})}e^{tg(\bar{X})}d\mathcal{D}^{\mathrm{l}}(x) + \int_{\mathrm{B}} e^{-tg(\bar{X})}e^{tg(\bar{X})}d\mathcal{D}^{\mathrm{r}}(x)\right) \\ &\leq e^{-t\varepsilon}\frac{1}{2}\left(\int_{\mathrm{A}} e^{tg(\bar{X})}d\mathcal{D}^{\mathrm{l}}(x) + \int_{\mathrm{B}} e^{tg(\bar{X})}d\mathcal{D}^{\mathrm{r}}(x)\right) \\ &\leq e^{-t\varepsilon}\mathbb{E}_{\bar{\mathcal{D}}}[e^{tg(\bar{X})}]. \end{aligned}$$

By the convexity of $x \to e^x$, for any $x \in [a, b]$, the following holds:

$$e^{tx} \leq \frac{b-x}{b-a}e^{ta} + \frac{x-a}{b-a}e^{tb}.$$

Thus, using $\mathbb{E}_{\bar{\mathcal{D}}}[V_{i+1}|\bar{X}_1, \cdots, \bar{X}_i] = 0$, then

$$\begin{aligned} \mathbb{E}_{\bar{\mathcal{D}}}[e^{tV_{i+1}}|\bar{X}_1, \cdots, \bar{X}_i] &\leq \mathbb{E}_{\bar{\mathcal{D}}}\left[\frac{Z_{i+1}+c_{i+1}-V_{i+1}}{c_{i+1}}e^{tZ_{i+1}} + \frac{V_{i+1}-Z_{i+1}}{c_{i+1}}e^{t(Z_{i+1}+c_{i+1})}|\bar{X}_1, \cdots, \bar{X}_i\right] \\ &= \frac{Z_{i+1}+c_{i+1}}{c_{i+1}}e^{tZ_{i+1}} + \frac{-Z_{i+1}}{c_{i+1}}e^{t(Z_{i+1}+c_{i+1})} \leq e^{t^2c_{i+1}^2/8}. \end{aligned}$$

Let $S_k = \sum\limits_{i=1}^{k} V_i$. Then, for any $t > 0$, we can write

$$\begin{aligned} \mathbb{P}[S_m \geq \varepsilon] &\leq e^{-t\varepsilon}\mathbb{E}_{\bar{\mathcal{D}}}[e^{tS_m}] \\ &= e^{-t\varepsilon}\mathbb{E}_{\bar{\mathcal{D}}}[e^{tS_{m-1}}\mathbb{E}_{\bar{\mathcal{D}}}[e^{tV_m}|\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_{m-1}]] \\ &\leq e^{-t\varepsilon}\mathbb{E}_{\bar{\mathcal{D}}}[e^{tS_{m-1}}]e^{t^2c_m^2/8}(\text{iterating previous argument}) \\ &\leq e^{-t\varepsilon}e^{t^2\sum\limits_{i=1}^{m}c_i^2/8}(\text{let } t = 4\varepsilon/\sum_{i=1}^{m}c_i^2) = e^{\frac{-2\varepsilon^2}{\sum_{i=1}^{m}c_i^2}}, \end{aligned}$$

the second statement is shown in a similar way. □

Theorem 5 (McDiarmid's Inequality of Interval-valued Random Variables)**.** *Let $\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_m \in \bar{\mathcal{X}} \subset \bar{\mathbb{R}}^p$ be a set of $m \geq 1$ interval-valued random vectors and assume that there exist $c_1, c_2, \cdots, c_m > 0$ such that $f : \bar{\mathcal{X}}^m \to \mathbb{R}$ satisfies the following conditions:*

$$|f(\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_i, \cdots, \bar{\mathbf{X}}_m) - f(\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_i^{\prime}, \cdots, \bar{\mathbf{X}}_m)| \leq c_i,$$

for any $i \in [m]$ and any points $\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_i, \cdots, \bar{\mathbf{X}}_m, \bar{\mathbf{X}}_i^{'} \in \bar{\mathcal{X}}$. Let $f(\bar{S})$ denote $f(\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_m)$, then, for any $\varepsilon > 0$, the following inequalities hold:

$$\begin{aligned} &\mathbb{P}[f(\bar{S}) - \mathbb{E}_{\bar{S}}[f(\bar{S})] \geq \varepsilon] \leq \exp \frac{-2\varepsilon^2}{\sum\limits_{i=1}^{m} c_i^2}, \\ &\mathbb{P}[f(\bar{S}) - \mathbb{E}_{\bar{S}}[f(\bar{S})] \leq -\varepsilon] \leq \exp \frac{-2\varepsilon^2}{\sum\limits_{i=1}^{m} c_i^2}. \end{aligned} \tag{6}$$

Proof. Define a sequence of random variables $V_k, k \in [m]$, as follows:

$$\begin{aligned} &V = f(\bar{S}) - \mathbb{E}_{\bar{S}}[f(\bar{S})], \\ &V_1 = \mathbb{E}_{\bar{S}}[V|\bar{\mathbf{X}}_1] - E_{\bar{S}}[V], \\ &V_k = \mathbb{E}_{\bar{S}}[V|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k] - \mathbb{E}_{\bar{S}}[V|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_{k-1}]. \end{aligned}$$

Note that $V = \sum\limits_{i=1}^{m} V_i$. Furthermore, the interval-valued random vector $\mathbb{E}_{\bar{S}}[V|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k]$ is a function of $\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k$, therefore:

$$\mathbb{E}_{\bar{S}}[\mathbb{E}_{\bar{S}}[V|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k]|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_{k-1}] = \mathbb{E}_{\bar{S}}[V|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_{k-1}],$$

which implies $\mathbb{E}_{\bar{S}}[V_k|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_{k-1}] = 0$. Thus, the sequence $(V_k), k \in [m]$ is a martingale difference sequence. Next, observe that, since $\mathbb{E}_{\bar{S}}[f(\bar{S})]$ is a scalar, V_k can be expressed as follows:

$$V_k = \mathbb{E}_{\bar{S}}[f(\bar{S})|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k] - \mathbb{E}_{\bar{S}}[f(\bar{S})|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_{k-1}].$$

Thus, we can define an upper bound W_k and lower bound U_k for V_k by:

$$\begin{aligned} W_k &= \sup_{\bar{\mathbf{X}}} \mathbb{E}_{\bar{S}}[f(\bar{S})|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k, \bar{\mathbf{X}}] - \mathbb{E}_{\bar{S}}[f(\bar{S})|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k], \\ U_k &= \inf_{\bar{\mathbf{X}}'} \mathbb{E}_{\bar{S}}[f(\bar{S})|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k, \bar{\mathbf{X}}^{'}] - \mathbb{E}_{\bar{S}}[f(\bar{S})|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k], \end{aligned}$$

$$\begin{aligned} W_k - U_k \quad &= \sup_{\bar{\mathbf{X}}, \bar{\mathbf{X}}'} \{\mathbb{E}_{\bar{S}}[f(\bar{S})|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k, \bar{\mathbf{X}}] - \mathbb{E}_{\bar{S}}[f(\bar{S})|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k, \bar{\mathbf{X}}^{'}]\} \\ &\leq \tfrac{1}{2} \sup_{\bar{\mathbf{X}}, \bar{\mathbf{X}}'} \{\mathbb{E}_{(\mathcal{D}^{\mathrm{l}})^{m-k}}[|f(\bar{S}_1) - f(\bar{S}_2)|] + \mathbb{E}_{(\mathcal{D}^{\mathrm{r}})^{m-k}}[|f(\bar{S}_1) - f(\bar{S}_2)|]\} \\ &\leq c_k, \end{aligned}$$

where $\bar{S}_1 = (\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k, \bar{\mathbf{X}}, \bar{\mathbf{X}}_{k+1}, \cdots, \bar{\mathbf{X}}_m), \bar{S}_1 = (\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k, \bar{\mathbf{X}}^{'}, \bar{\mathbf{X}}_{k+1}, \cdots, \bar{\mathbf{X}}_m)$. Thus, $U_k \leq V_k \leq W_k \leq U_k + c_k$. In the view of these inequalities, we can apply Theorem 4 to $V = \sum\limits_{i=1}^{m} V_i$, which yields the result. □

B.1 Proof of Theorem 1

For any sample $\bar{S} = \{\bar{\mathbf{z}}_i = (\bar{\mathbf{x}}_i, y_i)\}_{i=1}^{m} \sim \bar{\mathcal{D}}^m$ and any $\ell \in \mathcal{L}_{\mathcal{H}}$, we denote

$$\Phi(\bar{S}) = \sup_{\ell \in \mathcal{L}_{\mathcal{H}}} \{\mathbb{E}_{\bar{\mathcal{D}}}[\ell(\bar{\mathbf{z}})] - \tfrac{1}{m} \sum_{i=1}^{m} \ell(\bar{\mathbf{z}}_i)\} = \sup_{\ell \in \mathcal{L}_{\mathcal{H}}} \{\mathbb{E}_{\bar{\mathcal{D}}}[\ell(\bar{\mathbf{z}})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]\}.$$

Let $\bar{S}$ and $\bar{S}^{'}$ be two samples differing by exactly one point, say $\bar{\mathbf{z}}_m$ in $\bar{S}$ and $\bar{\mathbf{z}}_m^{'}$ in $\bar{S}^{'}$. Then, since the difference of suprema does not exceed the supremum of the difference, we have

$$\Phi(\bar{S}^{'}) - \Phi(\bar{S}) \leq \sup_{\ell \in \mathcal{L}_{\mathcal{H}}} \{\widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})] - \widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}})]\} \leq \sup_{\ell \in \mathcal{L}_{\mathcal{H}}} \tfrac{\ell(\bar{\mathbf{z}}_m) - \ell(\bar{\mathbf{z}}_m^{'})}{m} \leq \tfrac{C_\ell}{m}.$$

Similarly, we can obtain $\Phi(\bar{S}) - \Phi(\bar{S}^{'}) \leq \frac{C_\ell}{m}$, thus $|\Phi(\bar{S}^{'}) - \Phi(\bar{S})| \leq \frac{C_\ell}{m}$. Based on Definition 2, $\Phi(\bar{S})$ is a function of random variables $\mathbf{X}_i^{\mathrm{l}}$ and $\mathbf{X}_i^{\mathrm{r}}$ and we have

$$\begin{aligned} \mathbb{E}_{\bar{S}'}\{\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})]\} \quad &= \tfrac{1}{2}\{\mathbb{E}_{\mathcal{D}^{\mathrm{l}}}[\tfrac{1}{m} \sum_{i=1}^{m} \ell(\bar{\mathbf{z}}_i^{'})] + \mathbb{E}_{\mathcal{D}^{\mathrm{r}}}[\tfrac{1}{m} \sum_{i=1}^{m} \ell(\bar{\mathbf{z}}_i^{'})]\} \\ &= \tfrac{1}{2}\{\tfrac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\mathcal{D}^{\mathrm{l}}}[\ell(\bar{\mathbf{z}}_i^{'})] + \tfrac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\mathcal{D}^{\mathrm{r}}}[\ell(\bar{\mathbf{z}}_i^{'})]\} \\ &= \tfrac{1}{2}\{\mathbb{E}_{\mathcal{D}^{\mathrm{l}}}[\ell(\bar{\mathbf{z}})] + \mathbb{E}_{\mathcal{D}^{\mathrm{r}}}[\ell(\bar{\mathbf{z}})]\} = \mathbb{E}_{\bar{S}'}[\ell(\bar{\mathbf{z}})]. \end{aligned}$$

Then, by Theorem 5, for any $\delta > 0$, with probability at least $1 - \delta/2$, the following holds:

$$\Phi(\bar{S}) \leq \mathbb{E}_{\bar{S}}[\Phi(\bar{S})] + C_\ell\sqrt{\tfrac{\log(2/\delta)}{2m}},$$
$$\mathbb{E}_{\bar{S}}[\Phi(\bar{S})] = \mathbb{E}_{\bar{S}}[\sup_{\ell\in\mathcal{L}_\mathcal{H}}\{\mathbb{E}_{\bar{S}'}[\ell(\bar{\mathbf{z}})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]\}] = \mathbb{E}_{\bar{S}}[\sup_{\ell\in\mathcal{L}_\mathcal{H}}\mathbb{E}_{\bar{S}'}\{\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]\}].$$

Because

$$\begin{aligned}
&\sup_{\ell\in\mathcal{L}_\mathcal{H}}\mathbb{E}_{\bar{S}'}\{\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]\}\\
=\;&\sup_{\ell\in\mathcal{L}_\mathcal{H}}\tfrac{1}{2}\{\mathbb{E}_{(\mathcal{D}^{\mathrm{l}})^m}[\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]] + \mathbb{E}_{(\mathcal{D}^{\mathrm{r}})^m}[\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]]\}\\
\leq\;&\tfrac{1}{2}\sup_{\ell\in\mathcal{L}_\mathcal{H}}\{\mathbb{E}_{(\mathcal{D}^{\mathrm{l}})^m}[\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]] + \mathbb{E}_{(\mathcal{D}^{\mathrm{r}})^m}[\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]]\}\\
\leq\;&\tfrac{1}{2}\{\mathbb{E}_{(\mathcal{D}^{\mathrm{l}})^m}\sup_{\ell\in\mathcal{L}_\mathcal{H}}[\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]] + \mathbb{E}_{(\mathcal{D}^{\mathrm{r}})^m}\sup_{\ell\in\mathcal{L}_\mathcal{H}}[\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]]\}\\
=\;&\mathbb{E}_{\bar{S}'}\sup_{\ell\in\mathcal{L}_\mathcal{H}}\{\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]\}.
\end{aligned}$$

Then, we have

$$\mathbb{E}_{\bar{S}}[\Phi(\bar{S})] \leq \mathbb{E}_{\bar{S},\bar{S}'}\sup_{\ell\in\mathcal{L}_\mathcal{H}}\{\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]\} = \mathbb{E}_{\bar{S},\bar{S}'}\sup_{\ell\in\mathcal{L}_\mathcal{H}}\{\tfrac{1}{m}\sum_{i=1}^{m}[\ell(\bar{\mathbf{z}}_i^{'}) - \ell(\bar{\mathbf{z}}_i)]\}.$$

We introduce Rademacher variables σ_is, that are uniformly distributed independent random variables taking values in $\{-1, +1\}$,

$$\begin{aligned}
\mathbb{E}_{\bar{S}}[\Phi(\bar{S})] \quad&\leq \mathbb{E}_{\bar{S},\bar{S}'}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{\ell\in\mathcal{L}_\mathcal{H}}\{\tfrac{1}{m}\sum_{i=1}^{m}[\sigma_i\ell(\bar{\mathbf{z}}_i^{'}) - \ell(\bar{\mathbf{z}}_i)]\}(\sup(U+V) \leq \sup U + \sup V)\\
&\leq \mathbb{E}_{\bar{S}'}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{\ell\in\mathcal{L}_\mathcal{H}}\{\tfrac{1}{m}\sum_{i=1}^{m}\sigma_i\ell(\bar{\mathbf{z}}_i^{'})\} + \mathbb{E}_{\bar{S}}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{\ell\in\mathcal{L}_\mathcal{H}}\{\tfrac{1}{m}\sum_{i=1}^{m}-\sigma_i\ell(\bar{\mathbf{z}}_i)\}.
\end{aligned}$$

Because the definition of Rademacher complexity and the fact that the variables σ_i and $-\sigma_i$ are distributed in the same way, then

$$\mathbb{E}_{\bar{S}}[\Phi(\bar{S})] \leq 2\mathbb{E}_{\bar{S}}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{\ell\in\mathcal{L}_\mathcal{H}}\{\tfrac{1}{m}\sum_{i=1}^{m}\sigma_i\ell(\bar{\mathbf{z}}_i)\} = 2\mathcal{R}_{\bar{S}}(\mathcal{L}_\mathcal{H}).$$

Then using δ instead of $\delta/2$, with probability $1 - \delta$, the following holds :

$$\begin{gathered}
\Phi(\bar{S}) \leq 2\mathcal{R}_{\bar{S}}(\mathcal{L}_\mathcal{H}) + C_\ell\sqrt{\tfrac{\log(1/\delta)}{2m}}\\
\mathbb{E}_{\bar{\mathbf{z}}\sim\bar{\mathcal{D}}}[\ell(\bar{\mathbf{z}})] - \tfrac{1}{m}\sum_{i=1}^{m}\ell(\bar{\mathbf{z}}_i) \leq 2\mathcal{R}_{\bar{S}}(\mathcal{L}_\mathcal{H}) + C_\ell\sqrt{\tfrac{\log(1/\delta)}{2m}}.
\end{gathered} \tag{7}$$

We observe that changing one point in $\bar{S}$ changes $\widehat{\mathcal{R}}_{\bar{S}}(\mathcal{L}_\mathcal{H})$ by at most C_ℓ/m. Then, again using Theorem 5, with probability $1 - \delta/2$ the following holds:

$$\mathcal{R}_{\bar{S}}(\mathcal{L}_\mathcal{H}) \leq \widehat{\mathcal{R}}_{\bar{S}}(\mathcal{L}_\mathcal{H}) + C_\ell\sqrt{\tfrac{\log(2/\delta)}{2m}}.$$

Then with probability at least $1 - \delta$:

$$\begin{gathered}
\Phi(\bar{S}) \leq 2\widehat{\mathcal{R}}_{\bar{S}}(\mathcal{L}_\mathcal{H}) + 3C_\ell\sqrt{\tfrac{\log(1/\delta)}{2m}}\\
\mathbb{E}_{\bar{\mathbf{z}}\sim\bar{\mathcal{D}}}[\ell(\bar{\mathbf{z}})] - \tfrac{1}{m}\sum_{i=1}^{m}\ell(\bar{\mathbf{z}}_i) \leq 2\widehat{\mathcal{R}}_{\bar{S}}(\mathcal{L}_\mathcal{H}) + 3C_\ell\sqrt{\tfrac{\log(1/\delta)}{2m}}.
\end{gathered} \tag{8}$$

Next we let,

$$\Psi(\bar{S}) = \inf_{\ell\in\mathcal{L}_\mathcal{H}}\{\mathbb{E}_{\bar{\mathbf{z}}\sim\bar{\mathcal{D}}}[\ell(\bar{\mathbf{z}})] - \tfrac{1}{m}\sum_{i=1}^{m}\ell(\bar{\mathbf{z}}_i)\} = -\sup_{\ell\in\mathcal{L}_\mathcal{H}}\{-\mathbb{E}_{\bar{\mathbf{z}}\sim\bar{\mathcal{D}}}[\ell(\bar{\mathbf{z}})] + \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]\}.$$

In the same way, with probability at least $1-\delta$ the following holds:

$$\begin{aligned}\mathbb{E}_{\bar{\mathbf{z}}\sim\bar{\mathcal{D}}}[\ell(\bar{\mathbf{z}})]-\tfrac{1}{m}\sum_{i=1}^{m}\ell(\bar{\mathbf{z}}_i)\geq -2\mathcal{R}_{\bar{S}}(\mathcal{L}_{\mathcal{H}})-C_\ell\sqrt{\tfrac{\log(1/\delta)}{2m}}\\ \mathbb{E}_{\bar{\mathbf{z}}\sim\bar{\mathcal{D}}}[\ell(\bar{\mathbf{z}})]-\tfrac{1}{m}\sum_{i=1}^{m}\ell(\bar{\mathbf{z}}_i)\geq -2\widehat{\mathcal{R}}_{\bar{S}}(\mathcal{L}_{\mathcal{H}})-3C_\ell\sqrt{\tfrac{\log(2/\delta)}{2m}}.\end{aligned}\tag{9}$$

Since ℓ is Lipschitz continuous, according to Maurer (2016), we have

$$\widehat{\mathcal{R}}_{\bar{S}}(\mathcal{L}_{\mathcal{H}})\leq\sqrt{2}L_\ell\widehat{\mathcal{R}}_{\bar{S}_{\bar{X}}}(\mathcal{H}).\tag{10}$$

Following from Eqs. (7), (8), (9) and for any $\delta>0$, with probability at least $1-\delta$, each of the following holds for all $\ell\in\mathcal{L}_{\mathcal{H}}$:

$$|\mathbb{E}_{\bar{\mathbf{z}}\sim\bar{\mathcal{D}}}[\ell(\bar{\mathbf{z}})]-\tfrac{1}{m}\sum_{i=1}^{m}\ell(\bar{\mathbf{z}}_i)|\leq 2\widehat{\mathcal{R}}_{\bar{S}}(\mathcal{L}_{\mathcal{H}})+3C_\ell\sqrt{\tfrac{\log(2/\delta)}{2m}}.\tag{11}$$

Using $R_{\bar{\mathcal{D}}}(\boldsymbol{h})=\mathbb{E}_{\bar{\mathcal{D}}}[\ell(\boldsymbol{h}(\bar{X}),y)]$ and Eqs. (10) and (11), we have for any $\delta>0$, with probability at least $1-\delta$, each of the following holds for all $\ell\in\mathcal{L}_{\bar{\mathcal{H}}}$:

$$|R_{\bar{\mathcal{D}}}(\boldsymbol{h})-\widehat{R}_{\bar{\mathcal{D}}}(\boldsymbol{h})|\leq 2\sqrt{2}L_\ell\widehat{\mathcal{R}}_{\bar{S}_{\bar{X}}}(\mathcal{H})+3C_\ell\sqrt{\tfrac{\log(2/\delta)}{2m}}.$$

B.2 Proof of Theorem 2

Let $\mathbf{X}=(\mathbf{x}^1,\cdots,\mathbf{x}^c)\in\mathcal{X}$. Without loss of generality, we suppose $\mathrm{err}(f_1)\leq\cdots\leq\mathrm{err}(f_c)$. First, we consider the case where $c=2$. Then, we provide an upper bound on the error rate of f_{co}.

$$\begin{aligned}\mathrm{err}(f_{\mathrm{co}}) &=\mathbb{P}_{\mathcal{D}}(f_{\mathrm{co}}(\mathbf{X})\neq y)\\ &=\mathbb{P}(f_{\mathrm{co}}(\mathbf{X})\neq y|\mathbf{X}\in\mathbf{D}^{C}_{\mathcal{F}}(f_1,f_2))+\mathbb{P}(f_{\mathrm{co}}(\mathbf{X})\neq y|\mathbf{X}\in\mathbf{D}_{\mathcal{F}}(f_1,f_2))\\ &\leq\tfrac{1}{2}[\mathrm{err}(f_1)+\mathrm{err}(f_2)-\mathbb{P}_{\mathcal{D}}(\mathbf{X}\in\mathbf{D}_{\mathcal{F}}(f_1,f_2))]+\mathbb{P}(f_{\mathrm{co}}(\mathbf{X})\neq y|X\in\mathbf{D}_{\mathcal{F}}(f_1,f_2)),\end{aligned}\tag{12}$$

where $\mathbf{D}^{C}_{\mathcal{F}}(f_1,f_2)$ is denoted as the complement set of $\mathbf{D}_{\mathcal{F}}(f_1,f_2)$. According to Eq. (12) and $\mathrm{err}(f_1)\leq\mathrm{err}(f_2)$, if

$$\mathbb{P}(f_{\mathrm{co}}(\mathbf{X})\neq y|\mathbf{X}\in\mathbf{D}_{\mathcal{F}}(f_1,f_2))\leq\tfrac{1}{2}[\mathrm{err}(f_1)-\mathrm{err}(f_2)+\mathbb{P}_{\mathcal{D}}(\mathbf{X}\in\mathbf{D}_{\mathcal{F}}(f_1,f_2))],$$

we have $\mathrm{err}(f_{\mathrm{co}})\leq\mathrm{err}(f_1)$. Next, we consider the case where $c>2$. For $c>2$, we have $\boldsymbol{h}_{\mathrm{co}}\in\mathcal{H}_{\mathrm{co}}$,

$$h^q_{\mathrm{co}}(\mathbf{X})=\sum_{v=1}^{k+1}\mathbf{w}^{q\top}_v\boldsymbol{h}_v(\mathbf{x}^v)=\mathbf{w}^{q\top}_1\boldsymbol{h}_1(\mathbf{x}^1)+\sum_{v=2}^{c}\mathbf{w}^{q\top}_v\boldsymbol{h}_v(\mathbf{x}^v).$$

So exists $\alpha_q\in\mathbb{R}_+$, such that $\sum_{q=1}^{K}\alpha_q\sum_{v=2}^{c}\mathbf{w}^{q\top}_v\boldsymbol{h}_v(\mathbf{x}^v)=1$, then exists

$$\boldsymbol{h}^{c-1}_{\mathrm{co}}\in\mathcal{H}^{c-1}_{\mathrm{co}}(\mathbf{x}^2,\cdots,\mathbf{x}^c),\ \text{where } h^{c-1,q}_{\mathrm{co}}=\alpha_q\sum_{v=2}^{c}\mathbf{w}^{q\top}_v\boldsymbol{h}_v(\mathbf{x}^v).$$

We combine the last $c-1$ views i.e., $\mathbf{X}^{'}=(\mathbf{x}^2,\cdots,\mathbf{x}^c),\mathbf{X}=(\mathbf{x}^1,\mathbf{X}^{'})$. So exists

$$\boldsymbol{h}^{c-1}_{\mathrm{co}}\in\mathcal{H}^{c-1}_{\mathrm{co}}(\mathbf{x}^2,\cdots,\mathbf{x}^c)\subset\mathcal{H}(\mathbf{X}^{'}),\ \text{such that } h^q_{\mathrm{co}}(\mathbf{X})=\mathbf{w}^{q\top}_1\boldsymbol{h}_1(\mathbf{x}^1)+\tfrac{1}{\alpha_q}h^{c-1,q}_{\mathrm{co}}(\mathbf{X}^{'}).$$

Therefore we have $\boldsymbol{h}_{\mathrm{co}}\in\mathcal{H}_{\mathrm{co}}(\mathbf{x}^1,\mathbf{X}^{'})$. Let $f^{c-1}_{\mathrm{co}}(\mathbf{X})=\arg\max_{1\leq q\leq K}h^{c-1,q}_{\mathrm{co}}(\mathbf{X})$ denoted as the predict function of $\boldsymbol{h}^{c-1}_{\mathrm{co}}$. Because the conclusion is true when $c=2$, so exists $M\in(0,1)$, such that

$$\text{if } \mathbb{P}(f_{\mathrm{co}}(\mathbf{X})\neq y|\mathbf{X}\in\mathbf{D}_{\mathcal{F}}(f_1,f^{c-1}_{\mathrm{co}}))\leq M, \text{we have } \mathrm{err}(f_{\mathrm{co}})\leq\mathrm{err}(f_1).$$

Because $\mathbf{D}_{\mathcal{F}}(f_1,f^{c-1}_{\mathrm{co}})\subset\mathbf{D}_{\mathcal{F}}(f_1,\cdots,f_c)$, so

$$\mathbb{P}(f_{\mathrm{co}}(\mathbf{X})\neq y|\mathbf{X}\in\mathbf{D}_{\mathcal{F}}(f_1,f^{c-1}_{\mathrm{co}}))\leq\mathbb{P}(f_{\mathrm{co}}(\mathbf{X})\neq y|\mathbf{X}\in\mathbf{D}_{\mathcal{F}}(f_1,\cdots,f_c)).$$

Therefore, the conclusion is true when $c>2$ which yields the result.

B.3 Proof of Corollary 1

According to Theorem 3.1, 3.2 in Mohri et al. (2012) and Theorem 2 in Maurer (2016), we have

$$|R_{\mathcal{D}}(\boldsymbol{h}_{\text{co}}) - \widehat{R}_{\mathcal{D}}(\boldsymbol{h}_{\text{co}})| \leq 2\sqrt{2}L_{\text{co}}\mathcal{R}_{S_X}(\mathcal{H}_{\text{co}}) + C'_{\ell}\sqrt{\tfrac{\log(1/\delta)}{2m}}. \tag{13}$$

Next, let

$$\mathbf{W} = \left(\mathbf{w}_1^{1\top}, \cdots, \mathbf{w}_c^{1\top}, \cdots, \mathbf{w}_1^{K\top}, \cdots, \mathbf{w}_c^{K\top}\right)^\top,$$

$$\mathbf{H} = \left(\sum_{i=1}^{m}\sigma_{i1}\boldsymbol{h}_1(\mathbf{x}_i^1)^\top, \cdots, \sum_{i=1}^{m}\sigma_{i1}\boldsymbol{h}_c(\mathbf{x}_i^c)^\top, \cdots, \sum_{i=1}^{m}\sigma_{iK}\boldsymbol{h}_1(\mathbf{x}_i^1)^\top, \cdots, \sum_{i=1}^{m}\sigma_{iK}\boldsymbol{h}_c(\mathbf{x}_i^c)^\top\right)^\top.$$

Then, we have

$$\begin{aligned}
\mathcal{R}_{S_X}(\mathcal{H}_{\text{co}}) \quad &= \tfrac{1}{m}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}[\sup_{\boldsymbol{h}_{\text{co}}\in\mathcal{H}_{\text{co}}} \sum_{i=1}^{m}\sum_{q=1}^{K}\sigma_{iq}h_{\text{co}}^{q}(\mathbf{X}_i)] \\
&= \tfrac{1}{m}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}[\sup_{\boldsymbol{h}_j\in\mathcal{H}_v, ||\mathbf{W}||_2\leq\Lambda} \sum_{i=1}^{m}\sum_{q=1}^{K}\sigma_{iq}\sum_{v=1}^{c}\mathbf{w}_v^{q\top}\boldsymbol{h}_v(\mathbf{x}_i^v)] \\
&= \tfrac{1}{m}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}[\sup_{\boldsymbol{h}_v\in\mathcal{H}_v, ||\mathbf{W}||_2\leq\Lambda} \langle\mathbf{W},\mathbf{H}\rangle] \\
&\leq \tfrac{1}{m}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}[\sup_{\boldsymbol{h}_v\in\mathcal{H}_v, ||\mathbf{W}||_2\leq\Lambda} ||\mathbf{W}||_2||\mathbf{H}||_2](\text{using Cauchy-Schwarz inequality}) \\
&\leq \tfrac{\Lambda}{m}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}[\sup_{\boldsymbol{h}_v\in\mathcal{H}_v} [\sum_{v=1}^{c}\sum_{q=1}^{K}||\sum_{i=1}^{m}\sigma_{iq}\boldsymbol{h}_v(\mathbf{x}_i^v)||_2^2]^{\frac{1}{2}}] \\
&(\text{using Jensen's inequality and } i \neq j \Rightarrow \mathbb{E}_{\vec{\sigma}}[\sigma_{ip}\sigma_{jp}] = 0) \\
&\leq \tfrac{\Lambda}{m}\left[\mathbb{E}_{\mathcal{D}}[\sup_{\boldsymbol{h}_j\in\mathcal{H}_j} K\sum_{i=1}^{m}\sum_{v=1}^{c}||\boldsymbol{h}_v(\mathbf{x}_i^v)||_2^2]\right]^{\frac{1}{2}} \\
&\leq \tfrac{\Lambda}{m}\sqrt{Kcm} = \sqrt{\tfrac{Kc\Lambda^2}{m}}.
\end{aligned}$$

Then, we yield the final result

$$|R_{\mathcal{D}}(\boldsymbol{h}_{\text{co}}) - \widehat{R}_{\mathcal{D}}(\boldsymbol{h}_{\text{co}})| \leq 2L_{\text{co}}\sqrt{\tfrac{2Kc\Lambda^2}{m}} + C'_{\ell}\sqrt{\tfrac{\log(1/\delta)}{2m}}. \tag{14}$$

B.4 Proof of Theorem 3

Because $\sum\limits_{q=1}^{K}\sum\limits_{v=1}^{c}\sum\limits_{k=1}^{K} w_{vk}^{q}h_{vk}(\mathbf{x}_i^v) = 1$ and for any $v \in [c], k \in [K], 0 \leq h_{vk}(\mathbf{x}_i^v) \leq 1$, so $\sum\limits_{q=1}^{K}\sum\limits_{v=1}^{c}\sum\limits_{k=1}^{K} w_{vk}^{q} \leq 1$. Then,

$$\begin{aligned}
\mathcal{R}_{S_X}(\mathcal{H}_{\text{co}}) \quad &= \tfrac{1}{m}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}[\sup_{\boldsymbol{h}_{\text{co}}\in\mathcal{H}_{\text{co}}} \sum_{i=1}^{m}\sum_{q=1}^{K}\sigma_{iq}h_{\text{co}}^{q}(\mathbf{X}_i)] \\
&= \tfrac{1}{m}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}[\sup_{\boldsymbol{h}_v\in\mathcal{H}_v, ||\mathbf{W}||_2\leq\Lambda} \sum_{i=1}^{m}\sum_{q=1}^{K}\sigma_{iq}\sum_{v=1}^{c}\mathbf{w}_v^{q\top}\boldsymbol{h}_v(\mathbf{x}_i^v)] \\
&= \tfrac{1}{m}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}[\sup_{\boldsymbol{h}_v\in\mathcal{H}_v, ||\mathbf{W}||_2\leq\Lambda} \sum_{v=1}^{c}\sum_{q=1}^{K}\sum_{k=1}^{K} w_{vk}^{q}\sum_{i=1}^{m}\sigma_{iq}h_{vk}(\mathbf{x}_i^v)] \\
&\leq \tfrac{1}{m}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}[\sup_{\boldsymbol{h}_v\in\mathcal{H}_v} \max_{v\in[c],q\in[K]} \sum_{i=1}^{m}\sum_{k=1}^{K}\sigma_{ik}h_{vk}(\mathbf{x}_i^v)] \\
&\leq \max_{v\in[c]}\mathcal{R}_{S_{X^v}}(\mathcal{H}_v) \\
&= \min_{v\in[c]}\mathcal{R}_{S_{X^v}}(\mathcal{H}_v) + \max_{v\in[c]}\mathcal{R}_{S_{X^v}}(\mathcal{H}_v) - \min_{v\in[c]}\mathcal{R}_{S_{X^v}}(\mathcal{H}_v)
\end{aligned}$$

C Membership Function-based Method

In this section, we give further details of the membership function-based method to extract multi-view information from interval-valued data.

First, we introduce two types of fuzzy number and four different defuzzification methods used to construct the membership function-based method. The first type of fuzzy number called triangular fuzzy number. A triangular fuzzy number $\widetilde{x}$ can be characterized by $\mathrm{Tr}(a_1, b_1, a_2)$ and the membership function is shown as follows:

$$\mu_{\widetilde{x}}(t) = \begin{cases} 0, & t < a_1 \\ \dfrac{t-a_1}{b_1-a_1}, & a_1 \le t < b_1 \\ \dfrac{t-a_2}{b_1-a_2}, & b_1 \le t < a_2 \\ 0, & t \ge a_2. \end{cases}$$

Gaussian fuzzy number is the second type of fuzzy number. A Gaussian fuzzy number $\widetilde{x}$ can be characterized by $\mathrm{Ga}(c, \delta_1, \delta_2)$ and the membership function is given in the following equation:

$$\mu_{\widetilde{x}}(t) = \begin{cases} \exp(-(t-c)/2\delta_1)^2, & t < c \\ \exp(-(t-c)/2\delta_2)^2, & t \ge c. \end{cases}$$

Next, we introduce the four different defuzzification methods.

MOM. The first method is called *Mean/Middle of Maxima* (MOM) (Oussalah, 2002) which is widely-used due to its calculation simplicity. MOM is defined as:

$$\mathrm{MOM}(\widetilde{x}) = \mathrm{Mean}(t = \arg\max\nolimits_t \mu_{\widetilde{x}}(t)). \tag{15}$$

COG. *The Centre of Gravity* (COG) (Oussalah, 2002) is another widely-used defuzzification method. The definitions of COG for discrete and continuous membership functions are shown as follows:

$$\mathrm{COG}(\widetilde{x}) = \frac{\sum t\mu_{\widetilde{x}}(t)}{\sum \mu_{\widetilde{x}}(t)}(\text{discrete}) = \frac{\int t\mu_{\widetilde{x}}(t)dt}{\int \mu_{\widetilde{x}}(t)dt}(\text{continuous}). \tag{16}$$

ALC. The third approach, called *averaging level cuts* (ALC) (Oussalah, 2002), is defined as the flat averaging of all midpoints of the α-cuts.

$$\mathrm{ALC}(\widetilde{x}) = \tfrac{1}{2}\int_0^1 (\widetilde{x}_\alpha^L + \widetilde{x}_\alpha^U)d\alpha. \tag{17}$$

VAL. The final method is called *value of a fuzzy number* (VAL) (Delgado et al., 1998) which uses α-levels as weighting factors in averaging the α-cut midpoints. VAL is defined as :

$$\mathrm{VAL}(\widetilde{x}) = \int_0^1 \alpha(\widetilde{x}_\alpha^L + \widetilde{x}_\alpha^U)d\alpha. \tag{18}$$

We denote $D = \{(\bar{\mathbf{x}}_i, y_i)\}_{i=1}^m$ as the interval-valued dataset, where $\bar{\mathbf{x}}_i = (\bar{x}_{i1}, \cdots, \bar{x}_{ip})^\top \in \bar{\mathbb{R}}^p, y_i \in [K]$. Then, the construction process of the membership function-based method is introduced. We divide this method into two parts. In the first part, we use two functions $F_1(\cdot;\beta), F_2(\cdot;\beta)$ to transfer a interval-valued feature to a triangular fuzzy number and a Gaussian fuzzy number respectively. $F_1(\cdot;\beta), F_2(\cdot;\beta)$ are defined as:

$$\begin{gathered} F_1(\bar{x}_{ij};\beta) = \mathrm{Tr}(x_{ij}^{\mathrm{l}}, \beta x_{ij}^{\mathrm{l}} + (1-\beta)x_{ij}^{\mathrm{r}}, x_{ij}^{\mathrm{r}}), \\ F_2(\bar{x}_{ij};\beta) = \mathrm{Ga}(\beta x_{ij}^{\mathrm{l}} + (1-\beta)x_{ij}^{\mathrm{r}}, S_{1j}, S_{2j}), \\ S_{1j} = \sqrt{\mathrm{Var}(A_j)}, S_{2j} = \sqrt{\mathrm{Var}(B_j)}, \\ A_j = \{x_{ij}^{\mathrm{l}} : i \in [m], (\bar{\mathbf{x}}_i, y_i) \in D\}, B_j = \{x_{ij}^{\mathrm{r}} : i \in [m], (\bar{\mathbf{x}}_i, y_i) \in D\}, j \in [p], \end{gathered}$$

where $\beta \in [0,1]$ is a hyperparameter to control the shape of the membership function, $\mathrm{Var}(\cdot)$ is used to find the variance of the set. Using the above process, one interval-valued feature $\bar{\mathbf{x}}_i$ can be transferred into two fuzzy-valued features $\widetilde{\mathbf{x}}_i^1 = (\widetilde{x}_{i1}^1, \cdots, \widetilde{x}_{ip}^1)^\top$ and $\widetilde{\mathbf{x}}_i^2 = (\widetilde{x}_{i1}^2, \cdots, \widetilde{x}_{ip}^2)^\top$, where

$$\widetilde{\mathbf{x}}_i^\tau = \mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta) = (F_\tau(\bar{x}_{i1};\beta), \cdots, F_\tau(\bar{x}_{ip};\beta))^\top, \tau = 1, 2.$$

Table 6: Hyperparameters for the proposed method and four baselines

Algorithm	Basic classifier	Hyperparameters	Ranges
DF-SVM		regularization parameter, kernel type, shape parameter β	$\{0.1, 0.2, \cdots, 1, 2, \cdots, 10\}$, {'linear', 'poly', 'rbf'}, $\{0, 0.1, \cdots, 1\}$
DF-MLP		learning rate, shape parameter β	$\{0.001, 0.01, 0.1\}$, $\{0, 0.1, \cdots, 1\}$
Mv-IIE-2, Mv-IIE-3	SVM	regularization parameter, kernel type	$\{0.1, 0.2, \cdots, 1, 2, \cdots, 10\}$, {'linear', 'poly', 'rbf'}
	RF	min samples leaf, the number of trees	$\{1, \cdots, 10\}$, $\{5, 10, \cdots, 100\}$
	Net	learning rate	$\{0.001, 0.01, 0.1\}$
Mv-IIE	same above	same above, shape parameter β	same above, $\{0, 0.1, \cdots, 1\}$

In the second part, we use the four defuzzification methods to transfer the two fuzzy-valued features $\widetilde{\mathbf{x}}_i^1, \widetilde{\mathbf{x}}_i^2$ into eight crisp-valued features

$$\mathrm{MOM} \circ \mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta), \mathrm{COG} \circ \mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta), \mathrm{ALC} \circ \mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta), \mathrm{VAL} \circ \mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta), \tau = 1, 2.$$

According to Eq. (15), we find that $\mathrm{MOM} \circ \mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta) = \mathrm{MOM} \circ \mathbf{F}_2(\bar{\mathbf{x}}_i;\beta)$. Therefore, we can use the aforementioned membership function-based method to extract multi-view information, which contains seven parts: $\mathrm{MOM} \circ \mathbf{F}_1(\bar{\mathbf{x}}_i;\beta)$ and $\mathrm{COG} \circ \mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta), \mathrm{ALC} \circ \mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta), \mathrm{VAL} \circ \mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta), \tau = 1, 2$. We denote $\mathcal{T} = \{\mathbf{T}_v(\cdot;\beta)\}_{v=1}^7$ as a set of transfer functions constructed by using the membership function-based method, where

$$\mathbf{T}_1 = \mathrm{MOM} \circ \mathbf{F}_1, \mathbf{T}_2 = \mathrm{COG} \circ \mathbf{F}_1, \mathbf{T}_3 = \mathrm{COG} \circ \mathbf{F}_2, \mathbf{T}_4 = \mathrm{ALC} \circ \mathbf{F}_1,$$
$$\mathbf{T}_5 = \mathrm{ALC} \circ \mathbf{F}_2, \mathbf{T}_6 = \mathrm{VAL} \circ \mathbf{F}_1, \mathbf{T}_7 = \mathrm{VAL} \circ \mathbf{F}_2.$$

By applying the aforementioned transfer functions to extract crisp-valued information from the interval-valued data, one interval-valued feature $\bar{\mathbf{x}}_i$ can be transferred into seven different parts $\mathbf{X}_i^{\mathrm{Mv}} = (\mathbf{x}_i^1, \cdots, \mathbf{x}_i^7)$, where for any $i \in [m], v \in [7], \mathbf{x}_i^v = \mathbf{T}_v(\bar{\mathbf{x}}_i;\beta), \mathbf{T}_v \in \mathcal{T}$.

D EXPERIMENTAL DETAILS

In this section, the experiment details of all the baselines and our approach on both synthetic and real-world datasets are given. Moreover, the experiment details of the INPP framework are given. We implement the model with PyTorch 1.9.0. All experiments are conducted on a NVIDIA Quadro GV100 GPU with 32 GB memory.

Synthetic Datasets: For **DF-MLP** and **Mv-IIE** with basic classifier C_3, Adam (Kingma & Ba, 2015) is used as the optimization algorithm with momentum $= 0.9$, weight decay $= 0.0001$, and cross-entropy loss is used as the category label prediction loss. We set epochs equal to 200 and the mini-batch size equal to 200 for all datasets. The network structure of the basic classifier C_3 is a two-layer network with ReLU and Dropout in all the layers ($100 \times 100 \times \#\text{classes}$). For each algorithm on each dataset, we randomly divide each dataset into a training set (60%), a validation set (20%) and a test set (20%). First, we select the hyperparameters that can obtain the highest classification accuracy on the validation set. The hyperparameters that need to be selected are shown in Table 6. Then, the selected optimal hyperparameters are used to test the performance of each algorithm on the test set. In addition, the validation set is also used to select the candidate views of our proposed framework. We repeat the entire experiment process 10 times. Thus, the final results are shown in the form of "mean± standard deviation". Classification accuracy is used to evaluate the performance of the proposed model. The definition of classification accuracy is shown as follows:

$$\mathrm{Accuracy} = \frac{|\bar{\mathbf{x}} \in \bar{\mathcal{X}} : f(\bar{\mathbf{x}}) = \arg\min_{k \in [1,K]} h_k(\bar{\mathbf{x}})|}{|\bar{\mathbf{x}} \in \bar{\mathcal{X}}|},$$

where $f(\bar{\mathbf{x}})$ is the ground truth label of $\bar{\mathbf{x}}$, while $\boldsymbol{h}(\bar{\mathbf{x}}) = (h_1(\bar{\mathbf{x}}), \cdots, h_K(\bar{\mathbf{x}}))^\top$ is the label predicted by the presented algorithms and the baselines.

Real-world Datasets: The experiment details of the proposed method and the four baselines are basically the same as the synthetic datasets. We note that the mushroom dataset is an imbalanced dataset which means that each category contains a different number of instances. Therefore, we

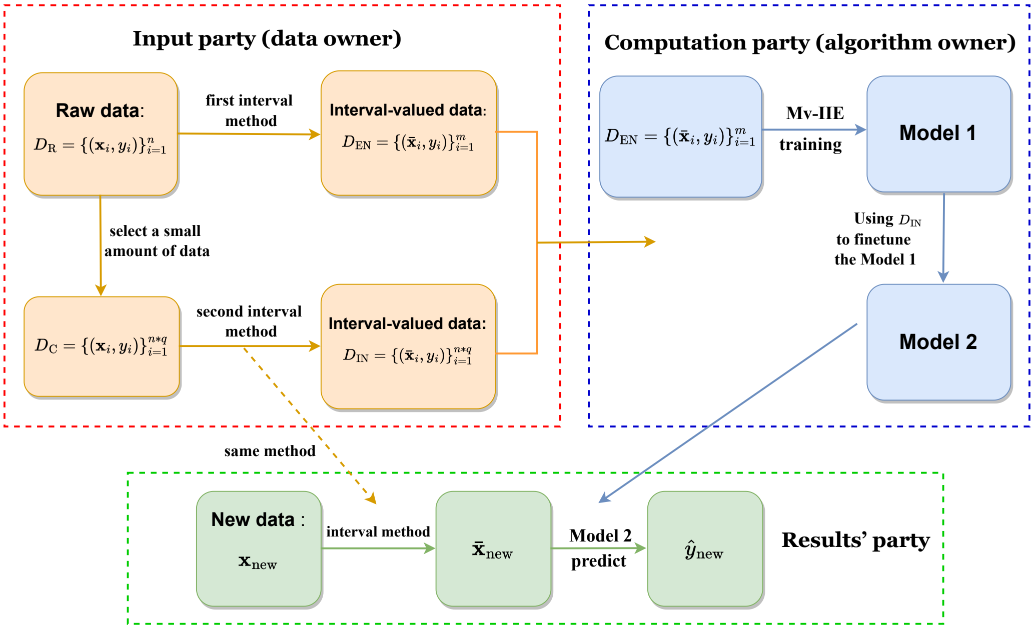
Input party (data owner)
Raw data:
$D_{\mathrm{R}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$
first interval method
Interval-valued data:
$D_{\mathrm{EN}} = \{(\bar{\mathbf{x}}_i, y_i)\}_{i=1}^{m}$
select a small amount of data
$D_{\mathrm{C}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n*q}$
second interval method
Interval-valued data:
$D_{\mathrm{IN}} = \{(\bar{\mathbf{x}}_i, y_i)\}_{i=1}^{n*q}$
Computation party (algorithm owner)
$D_{\mathrm{EN}} = \{(\bar{\mathbf{x}}_i, y_i)\}_{i=1}^{m}$
Mv-IIE training
Model 1
Using D_{IN} to finetune the Model 1
Model 2
same method
New data : $\mathbf{x}_{\mathrm{new}}$
interval method
$\bar{\mathbf{x}}_{\mathrm{new}}$
Model 2 predict
$\hat{y}_{\mathrm{new}}$
Results' party

Figure 4: **INPP** framework: The input party (denoted in **orange**) applies two interval methods to transfer the raw data into two interval-valued datasets. The computation party (denoted in **blue**) uses D_{EN} to train Model 1 by applying Mv-IIE framework and D_{IN} is used to fine-tune Model 1 to obtain Model 2. The results' party (denoted in **green**) uses Model 2 for new data prediction.

preprocess this dataset using a random oversampling technique (KMeansSMOTE (Last et al., 2017)) and use balanced accuracy (Brodersen et al., 2010) instead of ordinary classification accuracy to compare model performance on the mushroom dataset. The definition of balanced accuracy is shown as follows:

$$\text{Balanced Accuracy} = \tfrac{1}{K}\sum_{k=1}^{K}(\text{Recall of } k\text{-th class}),$$
$$\text{Recall} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}),$$

where TP is true positive, TN is true negative, FP is false positive and FN is false negative. After the process of the random oversampling technique, the data of each category in the mushroom dataset is expanded to 30. In addition, the Wilcoxon rank-sum test results of the method, which obtains the best performance, compared to the other methods are given on real-world datasets.

INPP Framework: The structure of INPP framework is shown in Figure 4. We randomly divide the original dataset (letter recognition dataset) into a raw dataset from the data owner(s) (70%) and a new dataset (30%) from the results' party. We choose $L = 6, T = 15$ and set $q = 0.20, 0.30, 0.50$. From Table 3, Mv-IIE with SVM-rbf (SVM with radial basis kernel function) achieve best outcomes on the second synthetic dataset. Therefore, we use SVM-rbf as the basic classifier of Mv-IIE in this experiment. The experimental details of Mv-IIE are the same as the aforementioned. The experiment details of the four well-known machine learning methods on the original dataset are the same as the experiment details of the four baselines on the synthetic datasets.

E Details of The Two Real-world Datasets Descriptions

In this section, we briefly introduce the two real-world datasets used in the experiments.

Mushroom Dataset : The first dataset is extracted from https://www.mykoweb.com/CAF/, which contains 248 instances in 17 fungi species categories. There are five interval-valued variables: the pileus cap width Pw, the stipe length Sl, the stipe thickness St, the spores major axis length Sma, and the spores minor axis length Smi. Some instances of the mushroom dataset are shown in Table 1. The goal of our experiment on this dataset is to predict the species category of the California mushroom using five interval-valued features.

Table 7: Some Instances of the Weather Dataset

Local times	T	P0	P	U	Td	Y
31/12/2021	[10.6, 13.3]	[757.8, 760.3]	[759.4, 762.1]	[81, 93]	[9.4, 11.1]	1
24/12/2021	[4.4, 12.2]	[757.3, 762.1]	[759.0, 763.6]	[40, 61]	[-5.0, 1.7]	0
23/12/2021	[-1.1, 5.0]	[763.4, 768.2]	[762.2, 769.9]	[38, 55]	[-10.0, 5.0]	0
22/12/2021	[2.8, 10.6]	[752.5, 761.6]	[754.0, 763.2]	[34, 93]	[-9.4, 2.2]	1

Table 8: Experiment results (accuracy±standard deviation of accuracies) of the ablation study on the synthetic and real-world datasets. The bold value represents the highest accuracy in each column.

Algorithms	Basic classifier	First synthetic dataset	Second synthetic dataset	Mushroom dataset	Weather dataset
view 1	C_1	97.97% ±0.80%	94.22% ±2.05%	76.81% ±3.07%	96.94% ±0.96%
	C_2	97.85% ±0.92%	91.27% ±2.31%	82.29% ±5.26%	97.03% ±0.68%
	C_3	98.12% ±0.66%	92.21% ±1.76%	77.56% ±3.36%	96.80% ±1.25%
view 2	C_1	96.50% ±0.56%	94.26% ±1.99%	76.66% ±3.83%	97.12% ±0.74%
	C_2	95.10% ±1.10%	91.81% ±1.87%	83.35% ±5.06%	96.83% ±0.97%
	C_3	96.47% ±0.68%	92.45% ±1.85%	79.62% ±4.15%	96.83% ±0.96%
view 3	C_1	95.20% ±0.56%	94.41% ±2.05%	76.55% ±3.25%	97.01% ±0.94%
	C_2	94.00% ±0.81%	91.67% ±2.28%	82.44% ±4.65%	96.69% ±0.99%
	C_3	94.30% ±0.91%	91.52% ±2.62%	79.62% ±3.32%	96.78% ±1.12%
view 4	C_1	97.82% ±0.61%	94.26% ±2.10%	76.67% ±3.86%	97.12% ±0.98%
	C_2	97.13% ±1.04%	90.88% ±2.98%	82.45% ±5.26%	96.72% ±1.20%
	C_3	97.62% ±0.93%	92.21% ±2.15%	79.39% ±3.32%	96.76% ±0.98%
view 5	C_1	97.97% ±0.80%	94.17% ±1.87%	75.07% ±3.18%	96.96% ±0.89%
	C_2	97.50% ±0.78%	91.08% ±2.49%	82.70% ±4.88%	96.96% ±0.69%
	C_3	98.12% ±0.66%	90.49% ±2.19%	71.38% ±5.94%	96.42% ±1.03%
view 6	C_1	98.00% ±0.80%	94.17% ±2.13%	77.12% ±3.14%	97.01% ±0.87%
	C_2	98.05% ±0.76%	90.54% ±2.11%	82.78% ±5.08%	96.94% ±0.70%
	C_3	98.12% ±0.66%	92.55% ±1.65%	77.90% ±4.48%	96.58% ±1.05%
view 7	C_1	97.95% ±0.77%	94.36% ±2.02%	76.81% ±3.07%	97.05% ±0.75%
	C_2	97.38% ±1.03%	90.54% ±2.07%	82.89% ±5.13%	97.03% ±0.68%
	C_3	98.12% ±0.66%	91.57% ±2.33%	75.13% ±4.82%	96.96% ±1.04%
Mv-IIE	C_1	$\mathbf{98.25}\% \pm \mathbf{0.69}\%$	$\mathbf{94.66}\% \pm \mathbf{1.81}\%$	81.75% ±4.09%	$\mathbf{97.26}\% \pm \mathbf{0.81}\%$
	C_2	97.13% ±1.18%	90.49% ±2.51%	$\mathbf{83.69}\% \pm \mathbf{3.39}\%$	96.46% ±0.73%
	C_3	98.05% ±0.72%	86.96% ±1.35%	82.67% ±3.14%	96.62% ±1.20%

Weather Dataset : The second dataset is the meteorological data of Washington (from January 1, 2016 to December 31, 2021), provided by the 'Reliable Prognosis' site (https://rp5.ru/), which contains 2191 instances. Each instance in this dataset is the meteorological data for one day in Washington, which is described by five interval-valued variables (air temperature T, atmospheric pressure at weather station level $P0$, atmospheric pressure reduced to main sea level P, humidity U and dew-point temperature Td) and one category variable (Precipitation or not: $0 \equiv$ No Precipitation, $1 \equiv$ Precipitation). Some instances of this dataset are shown in Table 7. We aim to use the five interval-valued features for precipitation prediction.

F Ablation Study

This section shows the ablation study of the Mv-IIE framework. We present the predictions on all single-view features using the three basic classifiers on both synthetic and real-world datasets. All results are illustrate in Table 8. This verifies the proposed framework's superiority and rationality in addressing interval-valued data classification problems.