Proceedings of Machine Learning Research 266:1–23, 2025 Conformal and Probabilistic Prediction with Applications

DUEn: An Ensemble Framework Enhanced by Distribution-Free Uncertainty for Regression

Songlin Du SODU@STUDENT.UNIMELB.EDU.AU
The University of Melbourne

Ling Luo LING.LUO@UNIMELB.EDU.AU
The University of Melbourne

Ilia Nouretdinov I.R.NOURETDINOV@CS.RHUL.AC.UK
Royal Holloway, University of London

Uwe Aickelin UWE.AICKELIN@UNIMELB.EDU.AU
The University of Melbourne

Editor: Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson and Henrik Boström

Abstract

The main objective of ensemble learning is to aggregate multiple models to better capture complex data distributions. Various ensemble techniques, including bagging and boosting, have been investigated and widely embraced in both research and practical applications. In this work, we enhance ensemble learning by incorporating distribution-free uncertainty inspired by conformal prediction. Conformal prediction allows us to quantify any model's uncertainty rigorously with valid coverage guarantees under lenient assumptions of the data distribution. We propose a novel ensemble learning framework called **D**istribution-Free **U**ncertainty-Aware **En**semble Framework (DUEn) for regression tasks which uses the information from distribution-free uncertainty in the form of intervals to benefit final point predictions and makes outputs more accurate and robust. Moreover, we propose a weighted interval agreement approach that aggregates base learners considering the degrees of uncertainty of their predictions. Experiments conducted on multiple data sets from different domains illustrate that DUEn is capable of enhancing the accuracy of regression by effectively using data while considering each base learner's distribution-free uncertainty.

Keywords: Ensemble methods; Conformal prediction; Uncertainty quantification; Fuzzy logic

1. Introduction

Ensemble learning, based on the idea of group decision making, trains multiple machine learning models (a.k.a. base learners) and combines their outputs leveraging each model's strengths since a single model is unlikely to capture the underlying structure of the data accurately (Mienye and Sun, 2022; Shahhosseini et al., 2022). Decisions made by these base learners are under the presence of uncertainty, and inaccurate quantification of such uncertainty can lead to consequential model failures (Amodei et al., 2016; Liu et al., 2019).

The main target of this work is to enhance the performance of ensemble methods via considering the uncertainty of each base learner. In machine learning, two widely accepted types of uncertainties are *aleatoric* (data) uncertainty and *epistemic* (model) uncertainty (Der Kiureghian and Ditlevsen, 2009; Hüllermeier and Waegeman, 2021). Aleatoric uncertainty is an inherent property of data from its generating mechanism, which is irreducible.

On the other hand, epistemic uncertainty refers to limited model cognition due to a lack of knowledge or the complexity, which can be mitigated by collecting more data or adjusting models (Zhou et al., 2022). The objective of uncertainty quantification is to estimate both aleatoric and epistemic uncertainties appropriately (Sullivan, 2015). Conformal Prediction (CP) (a.k.a. conformal inference), a user-friendly framework for generating statistically rigorous uncertainty quantification with reliable coverage, has been gradually gaining attention in recent years (Vovk et al., 2005). Unlike Gaussian processes which measure uncertainty based on the assumption of a certain distribution family (e.g. Gaussian) to capture the aleatoric uncertainty (Liu et al., 2019), CP is able to quantify distribution-free uncertainty under a lenient assumption of data distribution called exchangeability and is flexible to be established on any pre-trained model (Angelopoulos and Bates, 2021). In this work, we use Inductive Conformal Prediction (ICP) instead, because CP suffers from heavy computational cost, which overcomes this drawback with the cost of informational efficiency (Papadopoulos et al., 2002). Therefore, how to better use this distribution-free uncertainty to improve the performance of ensemble methods is our primary focus. The Interval Agreement Approach (IAA) was designed to combine experts' opinions, where their opinions are presented in the form of intervals (Wagner et al., 2014). IAA is able to model interval-valued data by using fuzzy sets, and gives a more accurate and robust result by defuzzification (Maadi et al., 2020).

Inspired by these advances, we propose **D**istribution-Free **U**ncertainty-Aware **En**semble Framework (DUEn), a novel ensemble framework enhanced by CP, which can leverage distribution-free uncertainty. Unlike conventional ensemble methods which generate weights to aggregate point predictions from base learners or train a second-level learner, DUEn converts the ensemble task to reasoning from distribution-free uncertainty intervals of individual base learners. We also propose a Weighted Interval Agreement Approach (WIAA) that distinguishes the importance of base learners by their uncertainty degree. Additionally, DUEn uses training data more effectively. Although DUEn only uses part of the training data to fit base learners due to the characteristic of ICP, it outperforms ensemble methods trained on more data.

Overall, the *contributions* of this study are listed below:

- We propose a distribution-free uncertainty based ensemble framework DUEn for regression problems, where the final point predictions are improved by using distribution-free uncertainty of each base learner. Meanwhile, this framework is flexible to be built on **any** machine learning models.
- We propose a new aggregation method, named WIAA, considering the uncertainty intervals further to enhance the performance of our framework.
- We conduct experiments to demonstrate that our proposed framework performs well on 9 out of 10 data sets and analyse the influence of various significance levels on prediction performance of the proposed model. We use two case studies on data sets of different sizes to manifest the advantages of our proposed framework.

The structure of this paper is given in the following manner; Section 2 summarises different combination methods in ensemble learning area and related works on CP framework

and IAA method. The technical details of ICP, IAA and our proposed framework are presented in Section 3 and 4. In Section 5, we introduce the experimental setup and provide results and analysis. Section 6 contains the conclusion of our findings and provides future research directions.

2. Related Work

2.1. Ensemble Learning and Uncertainty

Ensemble methods, especially deep ensembles, are prominent learning approaches that combine predictions made by a pool of learners (Dietterich, 2000; Lakshminarayanan et al., 2017). By combining predictions, ensemble methods improve the overall performance and achieve high robustness compared to a single model (Brown et al., 2005; Shahhosseini et al., 2022) and show clear advantages across various benchmarks (Ovadia et al., 2019). Conventional ensemble methods try to compute optimal weights by optimisation models to combine the predictions or train a second learning algorithm to predict the responses (Large et al., 2019). Mentch (Mentch and Hooker, 2016) proposed an approach to derive a model's asymptotic distribution. There exists some work in uncertainty estimation and decomposition of multi-model ensembles by Monte Carlo methods or Bayesian approaches (Sanderson, 2018). Liu et al. (2019) proposed an ensemble method which uses the uncertainty from Gaussian process to mitigate the prediction bias and a calibration function to mitigate the distribution bias. Our proposed framework is distinct from previous ensemble methods in that our method combines distribution-free uncertainty intervals of base learners instead of working on point predictions.

2.2. Conformal Prediction

Uncertainty quantification aims to accurately estimate aleatoric and epistemic uncertainties (Abdar et al., 2021). Conformal prediction is a distribution-free uncertainty quantification method since it does not require any assumption of the data distribution except exchangeability (Angelopoulos and Bates, 2021; Shafer and Vovk, 2008; Nouretdinov et al., 2001). It is worth noting that uncertainty intervals quantified by CP cannot distinguish two types of uncertainties as they are defined over observable outcomes (Stanton et al., 2023). One of the most important elements in CP is the score function which is designed to measure similarity. In this paper, we employ two residual-based nonconformity score functions (Lei et al., 2018; Papadopoulos et al., 2002). More details of other score functions and how to choose them are provided in the review (Kato et al., 2023). There is a series of studies on aggregating multiple intervals into a single interval (Gasparin and Ramdas, 2024; Linusson et al., 2020). CP has received a lot of attention recently and is widely used in different areas, such as Bayesian optimisation (Stanton et al., 2023) and decision making system (Cresswell et al., 2024).

2.3. Interval Agreement Approach

Interval aggregation functions have been explored and applied to decision making tasks successfully in recent years (Bentkowska and Pekala, 2018). Wagner et al. (2014) proposed a method to convert intervals to fuzzy sets coined interval agreement approach (IAA).

Considering the good performance of the IAA method to address intervals and uncertainty, Maadi et al. (2020) applied the IAA method to ensemble learning for classification problems. They presented a new interval modelling method by bagging strategy to build prediction intervals. Then, IAA was used to combine these interval values to generate fuzzy sets and to make the final prediction.

3. Preliminaries

This paper falls squarely within ensemble learning in the context of regression problems, therefore, the final outputs are point estimations. Let $\mathcal{X} \in \mathbb{R}^d$ denote the feature space with d dimensions and $\mathcal{Y} \in \mathbb{R}$ be the corresponding label space. Let $D = (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ be the training data set, where $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ and these observations are generated from an unknown distribution exchangeably. The ICP and IAA methods are described in the following two subsections.

3.1. Inductive Conformal Prediction

In general, provided a machine learning model $f(.)$ and a significance level $\epsilon \in (0, 1)$, ICP constructs a prediction interval with a probability guarantee $(1 - \epsilon)\%$ that the true label is included in the interval. ICP divides D into two disjoint sets: training set proper $D_{Tr} = \{x_1, ..., x_{n-m}\}$, calibration set $D_{Calib} = \{x_{n-m+1}, ..., x_n\}$. Then, we fit the model $f(.)$ on D_{Tr}. Next, a score function (SF) s is chosen to calculate the scores $S = s(f(x_i), y_i)$ for observations $(x_i, y_i) \in D_{Calib}$. In this paper, we use two residual-based score functions Eq. 1 and Eq. 2 (Kato et al., 2023; Lei et al., 2018; Papadopoulos et al., 2002):

$$\alpha_i = |y_i - f(x_i)|, \qquad (1) \qquad\qquad \alpha_i = \frac{|y_i - f(x_i)|}{\sigma_i}, \qquad (2)$$

where $f(x_i)$ is the estimation made by a model and $\sigma_i = e^{\mu_i}$ (μ_i is the estimation of the value $|y_i - f(x_i)|$). The score function Eq. 2 is an extension that accounts for non-constant residual variance. The score functions can be replaced by any advanced function, like Conformalised Quantile Regression (Romano et al., 2019). After this step, A quantile threshold $\hat{q} = Q(\frac{(m+1)(1-\epsilon)}{m})$ is computed. For a given test point x^*, the prediction interval $I = [f(x^*) - S(\hat{q}), f(x^*) + S(\hat{q})]$ is formed by the estimation $f(x^*)$ and the ranked score $S(\hat{q})$.

3.2. Interval Agreement Approach

IAA provides a method to generate fuzzy sets from interval-valued data and is applied to reason a point prediction from uncertainty intervals by defuzzification. Consider an interval $I = [I_l, I_r]$ which has left boundary l and right boundary r, now suppose we have a set of intervals $\widetilde{I} = \{I_1, ..., I_n\}$. In the IAA method, a Type-1 Fuzzy Set (T1 FS) named $\bar{I}$ can be created from intervals so that it represents the agreement among them. The membership function of $\bar{I}$ is defined as follows:

$$\mu_{\bar{I}} = \sum_{i=1}^{n} b_i / (\cup_{j_1=1}^{n-i+1} \cup_{j_2=j_1+1}^{n-i+2} \cdots \cup_{j_i=j_{i-1}+1}^{n} (I_{j_1} \cap ... \cap I_{j_i})) \qquad (3)$$

In Eq. 3, $b_i = {}^{i}/_{n}$ represents the degree of membership. Additionally, the symbol '/' signifies the assignment of a specific degree of membership, rather than indicating division. In the IAA, the degree of membership means the number of intervals overlapped at one point. In an extreme case where all intervals overlap, b_i is equal to 1. To simplify, the membership function of $\bar{I}$ for a point x can be rewritten as Eq. 4. As stated in Eq. 4, a degree of membership for point x is determined by its number of occurrences within each interval divided by the total number of intervals.

$$\mu_{\bar{I}}(x) = \frac{\sum_{i-1}^{n} \mu_{I_i}(x)}{n} \tag{4}$$

where

$$\mu_{\bar{I}_i}(x) = \begin{cases} 1, & l_{I_i} \leq x \leq r_{I_i} \\ 0, & else \end{cases} \tag{5}$$

4. Proposed framkwork: DUEn

Our Distribution-Free Uncertainty-Aware Ensemble (DUEn) framework enhances ensemble learning by CP, leveraging the knowledge of distribution-free uncertainty from finite samples. The workflow of our proposed method DUEn is presented in Fig. 1. We also propose the Weighted Interval Agreement Approach (WIAA), which makes the final result mainly rely on base learners with less uncertainty. This novel combination of ICP and IAA, albeit straightforward, leads to a distribution-free uncertainty based ensemble learning framework that yields better performance compared to simple combination of point predictions. The **theoretical analysis** of how DUEn framework uses the uncertainty from each base learner is given in Appendix A.

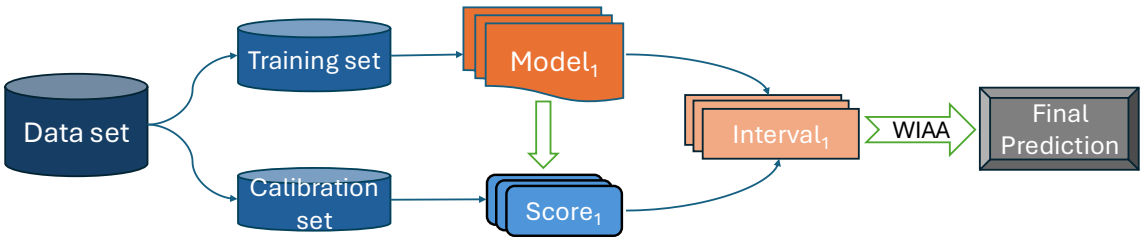


Figure 1: The workflow of DUEn. At first, we use the training set to train base models and then apply trained models to the calibration set to compute the nonconformity scores. Next, we infer the prediction interval for each base model based on the point prediction and the corresponding score. Finally, WIAA is used to reason the final point prediction from intervals.

4.1. DUEn - IAA

To begin with, the DUEn framework requires K base learners for ensemble learning. Since ICP approach can be established on any machine learning algorithm, different models like random forest, support vector machine, neural network etc., can be incorporated as base learners. At this stage, the uncertainty of each base learner is quantified by ICP which is rigorous with finite samples. Therefore, there are K prediction intervals $I_1, ..., I_K$ so far for each test point x^*.

We firstly use the IAA method to generate T1 FS. Using prediction intervals from previous step and the IAA function Eq. 4, the output of this procedure is a final T1 FS (FTFS) by aggregating all intervals. The FTFS is presented as Eq. 6, which is shown as a list of tuples while each element in the list indicates regions of change over the membership function. l and r represent the left boundary and the right boundary respectively. h stands for the height which presents the agreement of all intervals. v means the number of sections divided by intervals. A toy example is provided in Fig. 2(a).

$$R_i = [(R_{il}, R_{ir}), R_{ih}]; FTFS = R_1, ..., R_v \tag{6}$$

Finally, the centroid of FTFS is reasoned and used to determine the final prediction of the test point. The final point prediction is calculated by Eq. 7. The whole process of this algorithm is given in Algorithm 1.

$$C(FTFS) = \frac{\sum_{i=0}^{v}(R_{ih} \times R_{il}) + (R_{ih} \times R_{ir})}{\sum_{i=0}^{v} 2(R_{ih})} \tag{7}$$

Algorithm 1 DUEn - IAA

Input: Training set D, a test point x^*, K base learners $f_k(\centerdot)_{k=1}^{K}$, a significance level ϵ, score function s

Output: Point prediction y_{pred}

1: Split D into two disjoint parts D_{Tr} and D_{Calib}.
2: **for** each learner k in K **do**
3: Train base learner $f_k(\centerdot)$ on D_{Tr}.
4: **for** each point x_i in calibration set D_{Calib} **do**
5: Compute its corresponding point estimation by learner $f_k(x_i)$.
6: Based on the score function s (Eq. 1 or Eq. 2), calculate the nonconformity score.
7: **end for**
8: Rank all scores and select one according to the significance level ϵ and the size of calibration set.
9: Compute the point estimation $f_k(x^*)$ for test point x^*.
10: Generate prediction interval I_k from $f_k(x^*)$ and selected nonconformity score.
11: **end for**
12: Generate FTFS from K prediction intervals (Eq. 4 and Eq. 6).
13: Reason the centroid from FTFS as our final prediction y_{pred} (Eq. 7).
14: **return** y_{pred}

4.2. DUEn - WIAA

In Algorithm 1, the IAA method treats each prediction interval as equally important without considering the uncertainty degree which contains. This limitation motivates us to propose the algorithm DUEn-WIAA. In Algorithm 2, we add two types of weights, model weight $\mathcal{W}^p$ and uncertainty weight $\mathcal{W}^u$, to distinguish the importance of base learners by their accuracy and uncertainty. As for the model weights, there are multiple methods to compute optimal weights for base learners, i.e. cross-validation and stacking. In this work, we employ Generalised Ensemble Method (GEM) on D_{Tr} to compute the model weights, using an optimisation function to determine the global optimal weights. As to uncertainty weights, we use the lengths of prediction intervals to represent the uncertainties contained by intervals. The length of prediction interval is also one significant criterion of the predictive efficiency of CP framework, which shows the degree of uncertainty. Therefore, the core of this algorithm is to use the information from the uncertainty to adjust the importance of base models and select the base learner on which the final results mainly depend. Models that provide less uncertainty should get higher weights during the aggregation. The working procedure of DUEn-WIAA is presented in Algorithm 2. The steps of using uncertainty information are explained as follows.

After getting K intervals, different intervals represent different degrees of uncertainty. Thus, they are not equally important. Considering this, we compute weights $w_1^u, ..., w_K^u$ for intervals to distinguish their importance via their lengths $l_1, ..., l_K$. Since wider interval indicates higher uncertainty, we compute the multiplicative inverse of lengths $1/l_1, ..., 1/l_K$ at first. Then, each weight is calculated by softmax function as shown in Eq. 8.

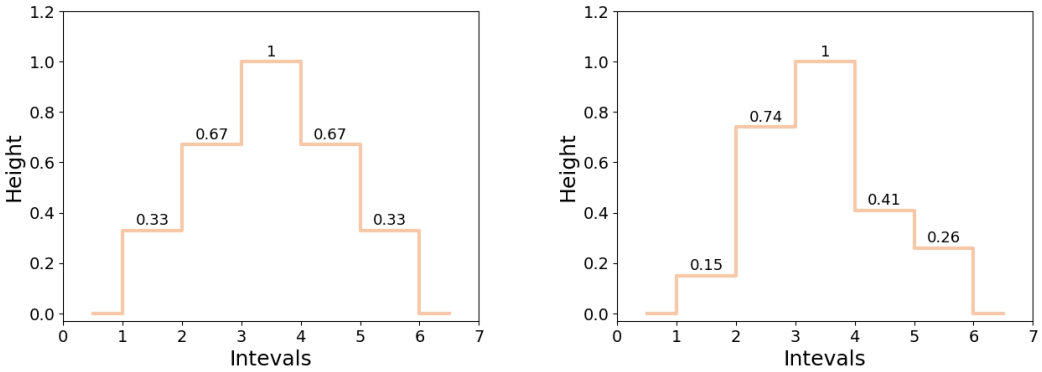
$$w_i^u = \frac{1/{l_i}^2}{\sum_1^K 1/{l_i}^2} \tag{8}$$

Unlike Eq. 4, our proposed method WIAA adjusts the definition of $\mu_I(x)$ by using the model weights and uncertainty weights. Fig. 2 illustrates the difference between IAA and WIAA. Considering two groups of intervals {[1, 4], [2, 5], [3, 6]} and {[1, 5], [2, 4], [3, 6]}, IAA provides the same output for these two examples since it does not consider the length of different uncertainty intervals, which shows symmetric height in Fig. 2(*a*). However, the weighted version overcomes this problem and the height of each section adapts with the degree of uncertainty. The formula of weighted membership function 9 is given below, where $\gamma \in (0, 1)$ is a hyper-parameter to balance between these two weights.

$$\mu_I(x) = \sum_{i-1}^{n} (\gamma * w_i^p + (1 - \gamma) * w_i^u) * \mu_{\bar{I}_i}(x) \tag{9}$$

5. Experiments

This section presents the details of experimental setup, containing data sets, base learners selection, hyperparameter selection, and experimental results. The experiments focus on comparing the prediction results generated by our proposed framework to other ensemble methods which combines point predictions.



(a) IAA (b) WIAA

Figure 2: This is an illustration of the difference between IAA and WIAA. For simplicity, we set γ to 0 to show the influence of uncertainty weights. Consider the first instance {[1, 4], [2, 5], [3, 6]}, IAA and WIAA will provide the same result, as presented in Fig. (a) because the lengths of all intervals are the same. In such case, there is no difference between IAA and WIAA. For the second instance {[1, 5], [2, 4], [3, 6]}, the lengths of intervals are different, which leads to an asymmetric plot as Fig. (b).

Algorithm 2 DUEn - WIAA

Input: Training set D, test point x^*, K base learners $f_k(.)_{k=1}^K$, a significance level ϵ, score function s, optimal model weights $\mathcal{W}^p$, γ

Output: Point prediction y_{pred}

1: Compute lines 1 to 11 of Algorithm 1.
2: **for** each prediction interval I_k **do**
3: Compute weight w_k^u (Eq. 8).
4: **end for**
5: Generate FTFS from K prediction intervals and two weights (Eq. 6 and Eq. 9).
6: Reason the centroid from FTFS as our final prediction y_{pred}.
7: **return** y_{pred}

5.1. Experimental Setup

Data sets. To evaluate the performance and generalisability of the proposed models, we conduct experiments on 10 public data sets from University of California at Irvin (UCI) Machine Learning Repository (Dua and Graff, 2017). Minimal pre-processing procedures, like encoding categorical features, have been applied on the selected data sets. For the data splitting step, our proposed methods are different from traditional methods because of the characteristics of ICP. For traditional point-combination ensemble methods, 80% of the data set is used as training set to select hyper-parameters and train the models, and the rest 20% is used as test set. Our proposed DUEn has to split the dataset into three parts, which are training, calibration and test sets. For fairness, 20% is used as test set, while 80% is divided into training and calibration sets. As for the choice of the size of calibration set, there are no special rules to enforce (Kato et al., 2023). We decided to use 60% - 75% of the whole data set for training process and hyper-parameter selection, and the rest 20% - 5% as calibration set to capture the uncertainty. Since different sets were used for hyper-parameter selection, there could be small variance in the final hyper-parameters selected from grid search. The details of these data sets and splitting strategies are described in Table 1.

Table 1: Statistics of 10 UCI data sets. The last column 'Size' means the percentage of the overall data set for calibration set.

Data set	No. Instances	No. of Features		Size
		Categorical	Numerical	
Servo	167	2	2	20%
Liver Disorder(LD)	345	0	5	20%
Graduate(Grad)	400	2	5	20%
Breast Cancer(BC)	569	1	29	20%
Heart Failure(HF)	299	6	5	20%
Abalone(Aba)	4177	1	7	10%
Wine	4898	2	10	10%
Power Plant(PP)	9567	0	4	5%
Appliances(App)	19735	1	26	5%
Conductivity(Cond)	21263	2	79	5%

Contenders. We compare the performance of the proposed DUEn framework using different combination approaches (IAA or WIAA) and two different score functions with four point combination ensemble methods which use weights or second-level learners to aggregate point predictions from base learners. The four benchmark ensemble methods are: (1) Basic Ensemble Method (BEM): this is a benchmark method using the average value from base learners; (2) Generalised Ensemble Method (GEM): the second benchmark uses nonlinear convex optimisation function to find the global optimal weights to combine results from

base learners; (3) Stacking method with different second-level learners: linear regression and random forest, as the third and fourth benchmarks.

Implementation details and evaluation metrics. We employ XGBoost, random forest, lasso and support vector machine as base learners. For the selection of hyper-parameters, we used grid search to find the best combination of hyper-parameters, in the comprehensive hyper-parameters space. The searching space is provided in the Appendix B. Since the score function Eq.2 requires another model to estimate σ_i, we use the same model as the base learner for simplicity. As for the significance level (ϵ) and trade-off parameter (γ) in our proposed framework, the searching spaces are from 0.05 to 0.95 and 0 to 1 with step 0.01 respectively. As this paper focuses on regression problems, we use Mean Absolute Error ($\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$) and Root Mean Squared Error ($\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$) as evaluation metrics to compare the performance of our models to others. The entire process is repeated 50 times and we report the mean and standard deviation of MAE and RMSE.

5.2. Experimental Results

Comparison of DUEn variants. Table 2 shows MAE results of our proposed framework DUEn with IAA and WIAA, using two different score functions. We conduct a Mann Whitney U test to identify the significant difference ($p \leq 0.05$) between IAA and WIAA. With the exception of LD (0.12) and HF (0.07), there are significant differences in 8 out of 10 data sets. The results indicate that using score function Eq.2 slightly improves the performance of proposed framework for all data sets as the intervals given by score function Eq.1 can be sacrificed in some cases, like when data noise is heteroscedastic. Hence, we only present the DUEn with score function Eq.2 in the following result tables.

Table 2: The MAE results (mean± std) of proposed framework with different score functions. The results with $*$ represent that there is significant difference between IAA and WIAA.

MAE	score function Eq.1		score function Eq.2	
	IAA	WIAA	IAA	WIAA
Servo	0.31±0.06	0.27±0.06*	0.30±0.07	**0.25±0.06***
LD	2.33±0.17	2.29±0.16	2.32±0.17	**2.27±0.16**
Grad	4.92±0.38	4.49±0.32*	4.86±0.38	**4.45±0.31***
BC	9.06±0.64	5.38±0.48*	9.02±0.64	**5.31±0.46***
HF	0.54±0.08	0.53±0.08	0.53±0.08	**0.51±0.08**
Aba	2.68±0.08	2.55±0.07*	2.65±0.08	**2.52±0.08***
Wine	3.68±0.07	2.49±0.05*	3.66±0.07	**2.48±0.05***
PP	2.68±0.05	2.24±0.03*	2.62±0.05	**2.20±0.04***
App	4.06±0.08	3.81±0.07*	4.00±0.08	**3.75±0.08***
Cond	8.22±0.10	6.18±0.10*	7.74±0.09	**6.09±0.10***

Table 3: The MAE for proposed algorithms and benchmarks. The best result on each dataset is indicated in boldface type. The improvements are calculated between the results of DUEn-WIAA and the best base learners or contenders. All results (mean ± std) on the same row are in the same unit, omitting leading zeros.

MAE	Base learners				Contenders				DUEn		Improvement
	XGB	RF	SVM	Lasso	Stack_lr	Stack_rf	BEM	GEM	IAA	WIAA	
Servo	0.30±0.06	0.28±0.05	0.34±0.07	0.92±0.07	0.27±0.06	0.27±0.06	0.38±0.06	0.27±0.06	0.30±0.07	**0.25±0.06**	+7.41%
LD	2.59±0.19	2.37±0.16	2.34±0.17	2.42±0.17	2.37±0.16	2.51±0.21	2.35±0.16	2.36±0.16	2.32±0.17	**2.27±0.16**	+2.99%
Grad	5.95±0.48	4.72±0.32	6.21±0.40	4.59±0.32	4.57±0.32	5.22±0.34	5.02±0.38	4.57±0.33	4.86±0.38	**4.45±0.31**	+2.63%
BC	13.25±0.73	5.41±0.43	44.73±5.49	8.93±0.74	5.46±0.42	5.66±0.42	14.73±1.00	5.42±0.43	9.02±0.64	**5.31±0.46**	+1.85%
HF	0.67±0.07	0.56±0.08	0.58±0.07	0.56±0.08	0.56±0.08	0.60±0.08	0.56±0.08	0.56±0.08	0.53±0.08	**0.51±0.08**	+8.93%
Aba	3.57±0.15	2.89±0.10	4.04±0.13	2.75±0.10	2.59±0.08	2.71±0.09	2.78±0.08	2.75±0.10	2.65±0.08	**2.52±0.08**	+2.70%
Wine	**2.45±0.06**	3.77±0.08	7.38±0.13	7.00±0.12	**2.45±0.06**	2.53±0.05	4.65±0.09	**2.45±0.06**	3.66±0.07	2.48±0.05	-1.22%
PP	2.22±0.03	2.79±0.05	5.05±0.09	3.62±0.05	2.22±0.03	2.31±0.03	3.01±0.05	2.22±0.03	2.62±0.05	**2.20±0.04**	+0.90%
App	4.02±0.07	4.57±0.07	4.75±0.11	5.29±0.07	3.97±0.06	4.00±0.06	4.12±0.07	3.93±0.07	4.00±0.08	**3.75±0.08**	+4.58%
Cond	6.13±0.09	7.79±0.10	17.26±0.16	13.36±0.09	6.13±0.09	6.21±0.09	9.78±0.09	6.14±0.09	7.74±0.09	**6.09±0.10**	+0.65%

Comparison of DUEn and benchmark ensemble methods. Table 3 shows the results of MAE of proposed DUEn framework and their competitors, which consists of 4 base learners and 4 ensemble methods. Since neither DUEn-IAA nor BEM assigns weights to base learners, it is obvious that DUEn-IAA outperforms BEM on 10 data sets. However, it only has advantages on two data sets (LD and HF) among all contenders and base learners. After further considering the degree of uncertainty, DUEn-WIAA pays more attention to base learners which provide less uncertainty given the same significance level and surpasses all competitors on 9 out of 10 data sets. As to the rest one (Wine) data set, ensemble methods do not improve on the base learners while our method DUEn-WIAA only shows a slight decrease in performance. Especially on Servo, HF and App, the improvements are 7.41%, 8.93% and 4.58% respectively. Table 4 demonstrates the results of RMSE values, which displays a different pattern compared to MAE values. The reason is that RMSE is more sensitive to outliers, therefore, the improvements on small-scale data sets (the first five rows) are not as pronounced as they are on MAE values. As the size of data set increases, the proposed method has a slight negative effect (less than 1%) except PP, which indicates that DUEn-WIAA is somewhat sensitive to outliers.

5.3. Sensitivity Analysis

In our framework, there is two main hyper-parameters related to reasoning a point prediction from intervals, which are the significance level ϵ that controls the length of prediction interval and trade-off parameter γ between two kinds of weights $\mathcal{W}^u$ and $\mathcal{W}^p$. The results of Conductivity and Graduate are presented in this subsection and the rest results are given in the Appendix C.

Coverage rate $(1-\epsilon)$. For ϵ, since WIAA also has another parameter γ, we set γ to 0 to analyse the influence of uncertainty weight $\mathcal{W}^u$. As presented in Fig. 3, generally, when the coverage rate increases, the intervals generated by ICP become wider and gain more information from distribution-free uncertainty, which presents in the way of decreasing MAE for IAA and WIAA. For the IAA method, it is obvious that the MAE value

Table 4: The RMSE for proposed algorithms and benchmarks. The meaning of each column is consistent with those in the Table 3 above.

RMSE	Base learners				Contenders				DUEn		Improvement
	XGB	RF	SVM	Lasso	Stack_lr	Stack_rf	BEM	GEM	IAA	WIAA	
Servo	0.54±0.04	0.55±0.04	0.63±0.08	1.13±0.02	0.53±0.04	0.59±0.06	0.60±0.05	0.53±0.05	0.56±0.06	**0.50±0.06**	+5.66%
LD	3.35±0.12	2.99±0.11	3.10±0.14	3.10±0.12	3.01±0.11	3.22±0.16	3.01±0.12	3.00±0.12	2.99±0.13	**2.93±0.12**	+2.01%
Grad	7.73±0.06	6.57±0.03	7.79±0.04	6.37±0.04	6.37±0.04	7.20±0.04	6.73±0.04	6.38±0.04	6.63±0.05	**6.25±0.04**	+1.88%
BC	17.71±0.37	7.62±0.20	47.42±5.25	12.28±0.40	7.62±0.20	8.18±0.15	17.17±0.06	7.62±0.20	11.86±0.28	**7.61±0.21**	+0.13%
HF	1.10±0.07	0.97±0.08	0.97±0.10	0.97±0.10	0.96±0.10	1.02±0.08	0.96±0.10	0.96±0.10	0.95±0.10	**0.93±0.10**	+3.13%
Aba	5.76±0.20	5.07±0.17	5.75±0.10	4.87±0.14	**4.60±0.15**	4.87±0.16	4.76±0.15	4.74±0.15	4.69±0.15	4.64±0.14	-0.87%
Wine	3.28±0.16	4.88±0.17	9.45±0.43	8.76±0.32	**3.28±0.16**	3.41±0.14	5.88±0.18	**3.28±0.16**	4.75±0.17	3.30±0.15	-0.61%
PP	3.12±0.21	3.71±0.20	6.42±0.21	4.55±0.13	3.11±0.22	3.23±0.23	3.87±0.17	3.13±0.21	3.48±0.21	**3.10±0.22**	+0.32%
App	7.54±0.46	8.53±0.56	10.64±0.99	9.36±0.74	**7.52±0.47**	7.80±0.44	8.45±0.73	**7.52±0.48**	8.43±0.79	7.58±0.58	-0.79%
Cond	10.04±0.61	12.04±0.50	23.20±0.44	17.66±0.39	**10.01±0.57**	10.30±0.59	13.52±0.25	**10.01±0.57**	11.47±0.33	10.03±0.42	-0.20%

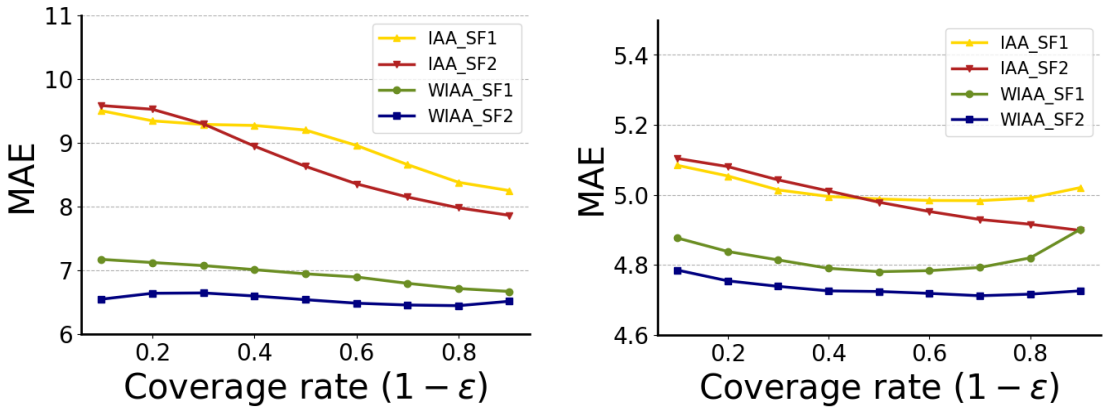
drops dramatically as coverage rate increases, which illustrates that our model derives more information from and benefits from uncertainty. Regarding WIAA, this method assigns huge advantages to models with less uncertainty. Therefore, the MAE stays low despite a low coverage rate. About the choice of coverage rate, we recommend 0.8 to capture the uncertainty information.

Trade-off parameter γ. Fig. 4(a) (Conductivity) exhibits more variation in MAE across different hyper-parameter values. Fig. 4(b) (Graduate), in contrast, shows a more stable performance with less sensitivity to γ and coverage rate changes. However, the darker regions (representing lower values) are predominantly concentrated in the center-right side of both plots, which means high coverage rate (0.8) and a moderate level of γ (0.5). Given the role of the parameter γ in balancing the influence of two weights within the model, its selection is critical for optimizing performance.

5.4. Case Studies

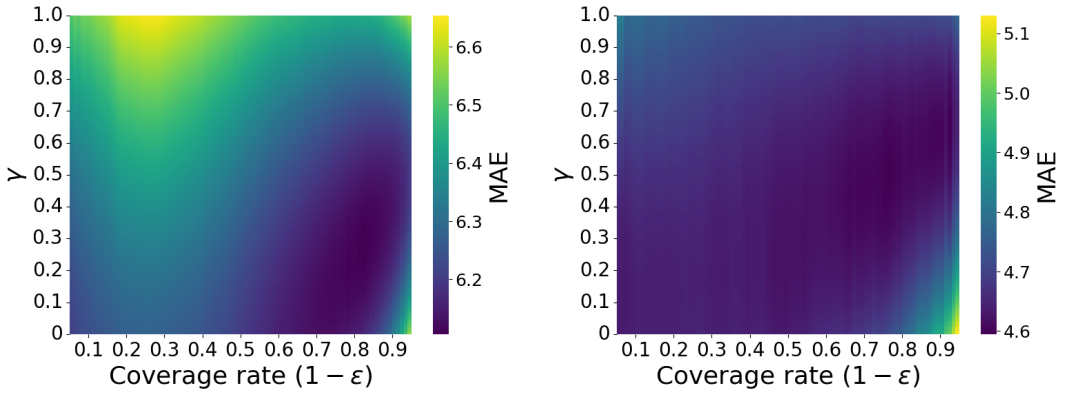
In this part, we conduct two case studies on data sets with different sizes: Heart Failure and Appliances.

Case study 1: Heart Failure. For small-scale data set (number of instances: 299, number of features: 11), the base learners are likely to overfit and capture the noise in the training data if they are designed to be complicated. Consequently, the performance of ensemble methods can also be influenced by this phenomenon. Fig. 5(a) illustrates that the performance of base learners (RF and SVM) are better when using just 60% data to train. The manifestations in all ensemble models are lower MAE values but higher standard deviation. Our proposed method DUEn-WIAA with score function Eq.2 outperforms these models no matter they are trained on 60% or 80%, which means that it is not necessary to use 80% data to train base learners and we can effectively use part of data within it to gain more information. Furthermore, we test the effect of different sized calibration sets (decrease from 40% to 10% in the step of 10%) on our proposed method. Fig. 5(b) shows that, on small-scale data set, the performance of WIAA remains quite stable with varying sizes of calibration set, especially score function Eq.2.



(*a*) Conductivity (*b*) Graduate

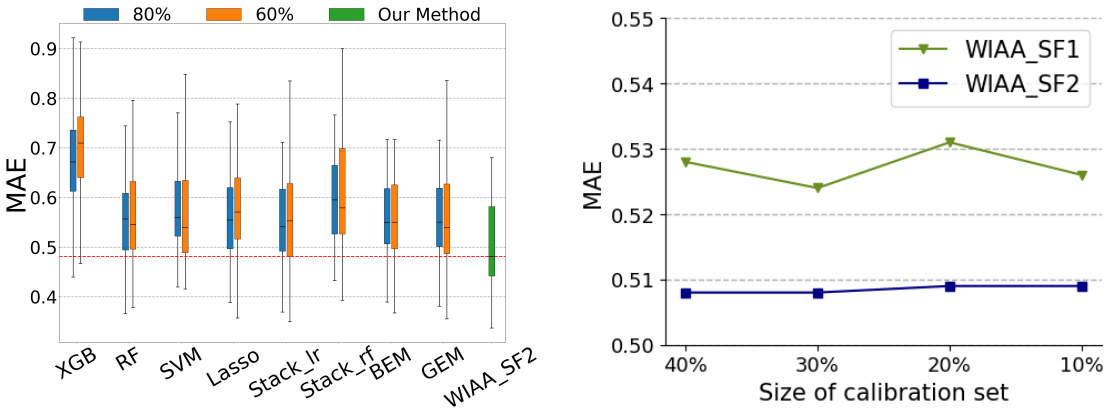
Figure 3: Influence of different coverage rates on proposed framework with different score functions, where the horizontal axis and vertical axis indicate the coverage rate $(1-\epsilon)$ and MAE values respectively. 'SF1' and 'SF2' in the label means different score functions Eq. 1 and Eq. 2 respectively.



(*a*) Conductivity (*b*) Graduate

Figure 4: Influence of γ on proposed DUEn-WIAA.

Case study 2: Appliances. In the case of relatively large data sets with many features (number of instances: 19735, number of features: 27), the base learners underfit and are unlikely to capture the distribution of data if they are not complicated enough. Therefore, the performance of base learners and ensemble methods can be insignificantly effected by such situation. To further illustrate the effectiveness of using distribution-free uncertainty to reason the final result, we conducted an experiment to verify the impact of using 5% less data on training base learners and ensemble methods. Fig. 6(*a*) showcases that the performances of base learners or ensemble methods being compared are slightly affected by less training data. The results of RF, SVM and Lasso are not given since their MAE values are significantly higher. However, although our proposed method has lost some information



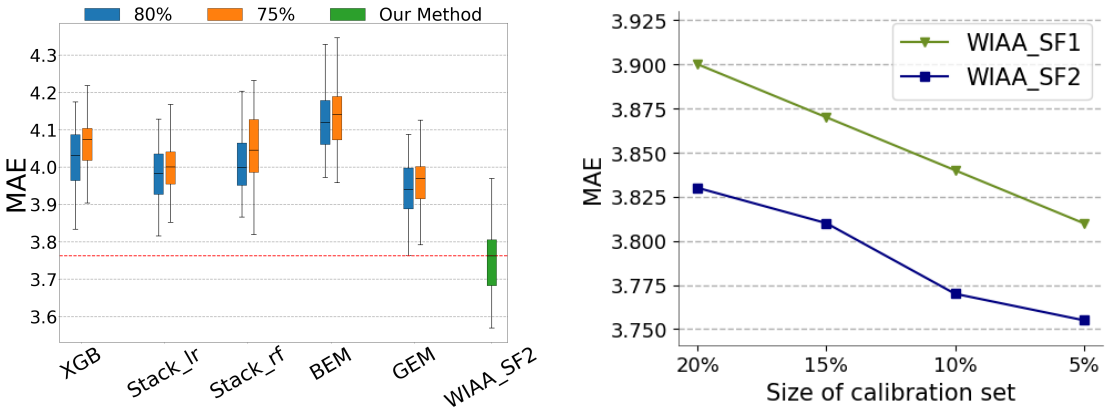
(*a*) Impact of less training data (*b*) Varying calibration sets

Figure 5: Case study of Heart Failure data set. The left figure shows the influence of less training data on base learners and competitors. The boxplots with blue color represent that they are trained on 80% of the whole data set. Orange colored boxplots are models trained on 60%. Our proposed method WIAA with score function Eq.2 are shown in green color. The right figure demonstrates the influence of the size of calibration set on two score functions. The size of calibration set ranges from 40% to 10%.

while training, it has gained more by using the rest 5% data as calibration set to capture the uncertainty. Afterwards, we explored the effect of different sizes of the calibration set, decreasing from 20% to 5% in the step of 5%. Fig. 6(*b*) shows that the MAE values decreases when the size of calibration set becomes smaller but the trend gradually flattens out.

6. Conclusion and Future Work

Conclusion. In this work, we have initiated the application of the CP framework to ensemble learning with the help of IAA, focusing on regression tasks. The proposed framework DUEn converts the ensemble tasks from combining point predictions to combining uncertainty intervals, aiming to take advantage of considering distribution-free uncertainty. Moreover, we proposed a new interval agreement method WIAA, which further uses the distribution-free uncertainty to distinguish base learners. The experimental results on the 9 data sets illustrate the advantages of taking distribution-free uncertainty into account while making predictions. In the case that base learners are complicated to data, our proposed method can reduce the impact of overfitting on base learners to some extend since our proposed framework only uses part of the training set to fit base learners. As for comparatively large data set, although proposed model loses knowledge while training on less data, it gains more by effectively utilising the rest to quantify uncertainty and deriving more information from it. In summary, our work highlights the potential of applying CP to ensemble learning, emphasising the importance of leveraging each base learner's distribution-free uncertainty



(*a*) Impact of less training data (*b*) Varying calibration sets

Figure 6: Case study of Appliances data sets. Two figures show the same experiments as case study 1. In Fig. 6(*a*), orange colored boxplots are models trained on 75%. In Fig. 6(*b*), the size of calibration set ranges from 20% to 5%.

in regression tasks. The proposed methods offer a principled way to integrate uncertainty estimation into ensemble learning, laying a solid foundation for the development of more robust predictive models.

Prospects. Our work paves the way for numerous intriguing avenues for future exploration. Firstly, we have used the same significance level for different base models in this work. However, it would be interesting to conduct experiments for different significance values for various models. In addition, our proposed DUEn can be extended to deep ensembles as the flexibility of CP. It is also worth exploring if this proposed framework can be extended to include experts' opinions as an additional source of information.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv:2107.07511*, 2021.

Urszula Bentkowska and Barbara Pekala. Diverse classes of interval-valued aggregation functions in medical diagnosis support. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications: 17th International Conference, IPMU 2018, Cádiz, Spain, June 11-15, 2018, Proceedings, Part III 17*, pages 391–403. Springer, 2018.

Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: a survey and categorisation. *Information fusion*, 6(1):5–20, 2005.

Jesse C Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. Conformal prediction sets improve human decision making. *arXiv preprint arXiv:2401.13744*, 2024.

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

Dheeru Dua and Casey Graff. Uci machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Matteo Gasparin and Aaditya Ramdas. Merging uncertainty sets via majority vote. *arXiv preprint arXiv:2401.09379*, 2024.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Yuko Kato, David MJ Tax, and Marco Loog. A review of nonconformity measures for conformal prediction in regression. *Conformal and Probabilistic Prediction with Applications*, pages 369–383, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

James Large, Jason Lines, and Anthony Bagnall. A probabilistic classifier ensemble weighting scheme based on cross-validated accuracy estimates. *Data mining and knowledge discovery*, 33(6):1674–1709, 2019.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Henrik Linusson, Ulf Johansson, and Henrik Boström. Efficient conformal predictor ensembles. *Neurocomputing*, 397:266–278, 2020.

Jeremiah Liu, John Paisley, Marianthi-Anna Kioumourtzoglou, and Brent Coull. Accurate uncertainty estimation and decomposition in ensemble learning. *Advances in neural information processing systems*, 32, 2019.

Mansoureh Maadi, Uwe Aickelin, and Hadi Akbarzadeh Khorshidi. An interval-based aggregation approach based on bagging and interval agreement approach in ensemble learning. In *2020 IEEE Symposium Series on Computational Intelligence*, pages 692–699. IEEE, 2020.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(26):1–41, 2016.

Ibomoiye Domor Mienye and Yanxia Sun. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access*, 10:99129–99149, 2022.

Ilia Nouretdinov, Thomas Melluish, and Volodya Vovk. Ridge regression confidence machine. In *ICML*, pages 385–392. Citeseer, 2001.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pages 345–356. Springer, 2002.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Benjamin Mark Sanderson. Uncertainty quantification in multi-model ensembles. In *Oxford Research Encyclopedia of Climate Science*. 2018.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Mohsen Shahhosseini, Guiping Hu, and Hieu Pham. Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. *Machine Learning with Applications*, 7:100251, 2022.

Samuel Stanton, Wesley Maddox, and Andrew Gordon Wilson. Bayesian optimization with conformal prediction sets. In *International Conference on Artificial Intelligence and Statistics*, pages 959–986. PMLR, 2023.

Timothy John Sullivan. *Introduction to uncertainty quantification*, volume 63. Springer, 2015.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Christian Wagner, Simon Miller, Jonathan M Garibaldi, Derek T Anderson, and Timothy C Havens. From interval-valued data to general type-2 fuzzy sets. *IEEE Transactions on Fuzzy Systems*, 23(2):248–269, 2014.

Xinlei Zhou, Han Liu, Farhad Pourpanah, Tieyong Zeng, and Xizhao Wang. A survey on epistemic (model) uncertainty in supervised learning: Recent advances and applications. *Neurocomputing*, 489:449–465, 2022.

Appendix A. Theoretical Analysis

In this section, we provide theoretical analysis for the proposed DUEn framework. We present how the distribution-free uncertainty of each model influences the final output, which highlights the role of uncertainty in mitigating each model's bias on target.

Ensemble by Averaging. Given a set of base learners $\{f_k(.)\}_{k=1}^K$, a classic ensemble model (averaging) assumes the form:

$$F_{AVE}(x) = \frac{1}{K}\sum_{k=1}^{K} f_k(x). \tag{10}$$

Assume that the true regression function each model needs to learn is $h(x)$, $f_k(x)$ can be rewritten to $f_k(x) = h(x) + err_k(x)$. Therefore, it is easy to get that the MAE of a single base learner is:

$$\overline{err}(f) = \frac{1}{K}\sum_{k=1}^{K}\int |err_k(x)|p(x)dx,$$

where x is sampled from the data distribution $p(x)$. Similarly, the MAE of F_{AVE} can be written as:

$$err(F_{AVE}) = \int |\frac{1}{K}\sum_{k=1}^{K} err_k(x)|p(x)dx.$$

It is obvious that:

$$err(F_{AVE}) \leq \overline{err}(f),$$

which means that the ensemble error is not larger than the error of a single base learner.

Following this, the theoretical analysis of DUEn based on the IAA method is given below.

Assumption 1 (Sampling stability Lei et al. (2018)) *For a sufficiently large sample size n, $\mathbb{P}\big(\|f_k(x) - \tilde{f}(x)\|_\infty \geq \eta_k\big) \leq \rho_k$, for some sequences satisfying $\eta_k = o(1)$, $\rho_k = o(1)$ as $k \to \infty$, and some function $\tilde{f}(x)$ which $\tilde{f}(x)$ does not need to be close to the true regression function $h(x)$. We only need $f_k(x)$ to concentrate around $\tilde{f}(x)$.*

Theorem 1 *If $I_1, ..., I_K$ are conformal intervals with valid marginal coverage guarantee $(1-\epsilon)$, and assume there are no overlapping endpoints among intervals for simplicity, the MAE of DUEn $err(F_{Ours})$ is:*

$$err(F_{Ours}) \leq \frac{1}{K}\sum_{k=1}^{K}\int |err_k(x) + \frac{K-1}{K+1}S_k(q)|p(x)dx \tag{11}$$

Proof. Let the function $G(.)$ represent the agreement of endpoints of the interval I_k over all intervals $\{I_k\}_{k=1}^K$. Based on Eq. 7, the centroid $F_{Ours}(x^*)$ can be generalised as:

$$F_{Ours}(x) = \frac{1}{K}\sum_{k=1}^{K}\frac{G(I_{kl}) * I_{kl} + G(I_{kr}) * I_{kr}}{G(I_{kl}) + G(I_{kr})}.$$

Then, I_{kl} and I_{kr} can be replaced by $f_k(x^*) - S_k(q)$ and $f_k(x^*) + S_k(q)$ respectively according to the Assumption 1. Therefore, $F_{Ours}(x)$ can be simplified as:

$$F_{Ours}(x) = \frac{1}{K}\sum_{k=1}^{K}(f_k(x) - \frac{G(I_{kl}) - G(I_{kr})}{G(I_{kl}) + G(I_{kr})}S_k(q)),$$

where $0 \leq |G(I_{kl}) - G(I_{kr})| \leq K - 1$ and $2 \leq G(I_{kl}) + G(I_{kr}) \leq 2K$. It is obvious that the positivity or negativity of the coefficient $\frac{G(I_{kl})-G(I_{kr})}{G(I_{kl})+G(I_{kr})}$ is based on the left and right of the model's point estimation in the most agreed part among all intervals (e.g. $[3, 4]$ in the example in Fig. 2). Therefore, each model's prediction is adjusted to the most agreed part among all prediction intervals by its distribution-free uncertainty. Following this, the MAE of DUEn framework is:

$$\begin{aligned}
err(F_{Ours}) = &\int |\frac{1}{K}\sum_{k=1}^{K} f_k(x) - h(x) \quad - \\
&\qquad \frac{G(I_{kl}) - G(I_{kr})}{G(I_{kl}) + G(I_{kr})}S_k(q)|p(x)dx \\
= &\int |\frac{1}{K}\sum_{k=1}^{K} err_k(x) \quad - \\
&\qquad \frac{G(I_{kl}) - G(I_{kr})}{G(I_{kl}) + G(I_{kr})}S_k(q)|p(x)dx \\
\leq &\frac{1}{K}\sum_{k=1}^{K}\int |err_k(x) + \frac{K-1}{K+1}S_k(q)|p(x)dx.
\end{aligned}$$

For the two extreme cases that there is no overlapping between intervals ($\bigcap_{k=1}^{K} I_k = \emptyset$) and all intervals are the same, our method is equivalent to ensemble by averaging. As for our second algorithm 2, the weights assigned to intervals (base learners) make the coefficient $\frac{G(I_{kl})-G(I_{kr})}{G(I_{kl})+G(I_{kr})}$ more effective. In Algorithm 2, higher weights are assigned to more precise models, which means the difference between $G(I_{kl})$ and $G(I_{kr})$ is small, so the point estimations are not adjusted too much by uncertainty information. On the contrary, for weak base learners, the point estimations made by these models will be shifted more to the most agreed part by all prediction intervals.

Appendix B. Experimental Setting

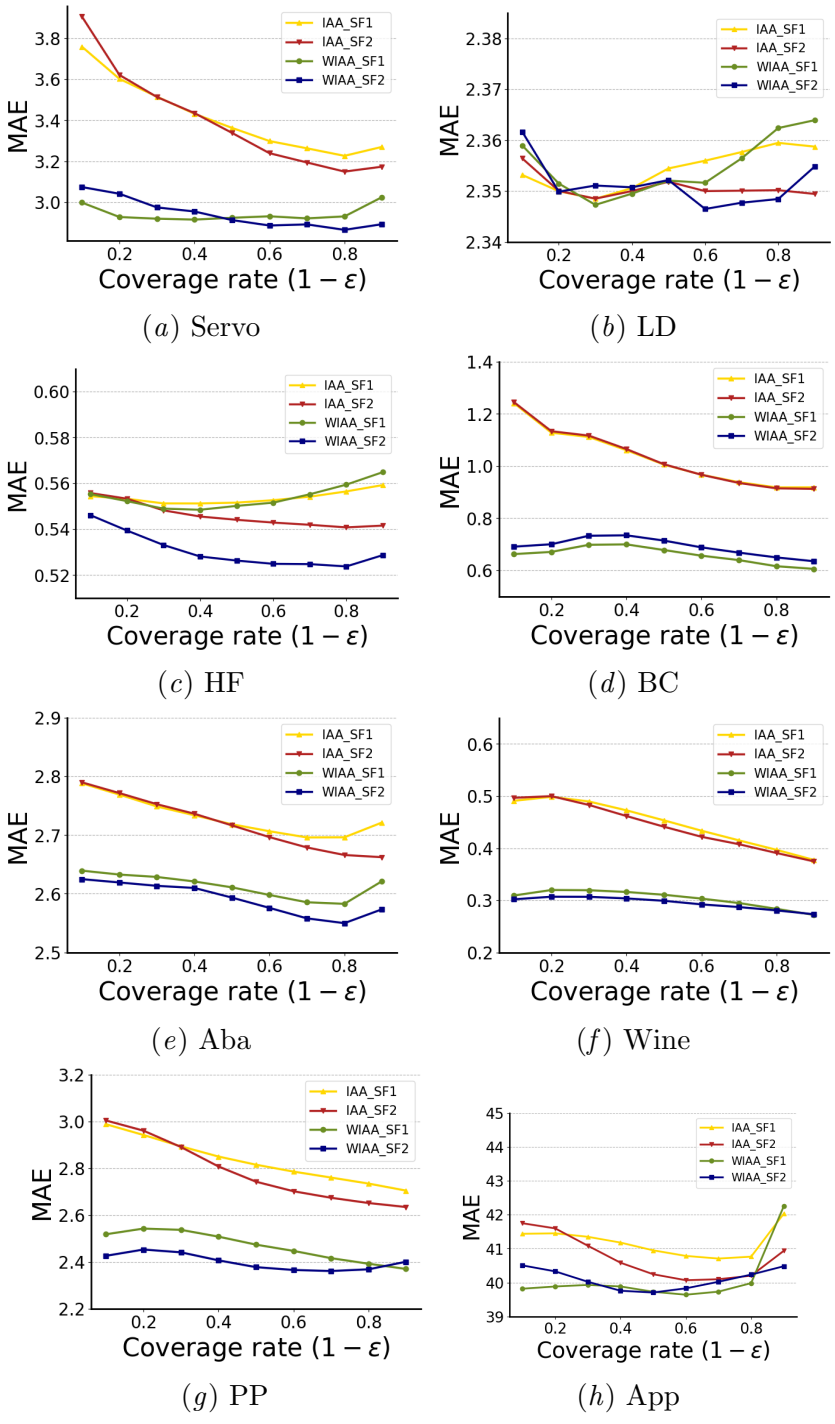
Table 5 demonstrates the searching spaces of hyper-parameters of base learners.

Table 5: Base learners and hyper-parameter spaces

Base learner	Hyper-parameter	Scope
XGBoost	'max_depth'	[2, 4, 6]
	'n_estimators'	[100, 500, 1000]
	'colsample_bytree'	[0.2, 0.6, 0.8]
	'min_child_weight'	[3, 5, 7]
	'gamma'	[0.3, 0.5, 0.7]
	'subsample'	[0.4, 0.6, 0.8]
Random Forest	'n_estimators'	[100, 300, 500]
	'max_features'	['sqrt', 'log2']
	'min_samples_split'	[2, 4, 8]
	'max_depth'	[4, 5, 6, 7, 8]
SVM	'C'	[1, 5, 10]
	'degree'	[3, 8]
	'coef0'	[0.01, 10, 0.5]
	'gamma'	['auto', 'scale']
Lasso	'alpha'	[0.001, 0.01, 0.05, 0.1, 0.5, 1, 5]

Appendix C. Experimental Results

Coverage rate ($1-\epsilon$). Fig. 7 provide the results of the sensitivity analysis of coverage rate. As shown in the figures, a common choice of coverage rate is around 0.8 except Appliances data set. For LD data set, although the lowest MAE value is given while coverage rate is 0.6, the MAE value is just slightly better than at a coverage rate of 0.8.



(a) Servo (b) LD

(c) HF (d) BC

(e) Aba (f) Wine

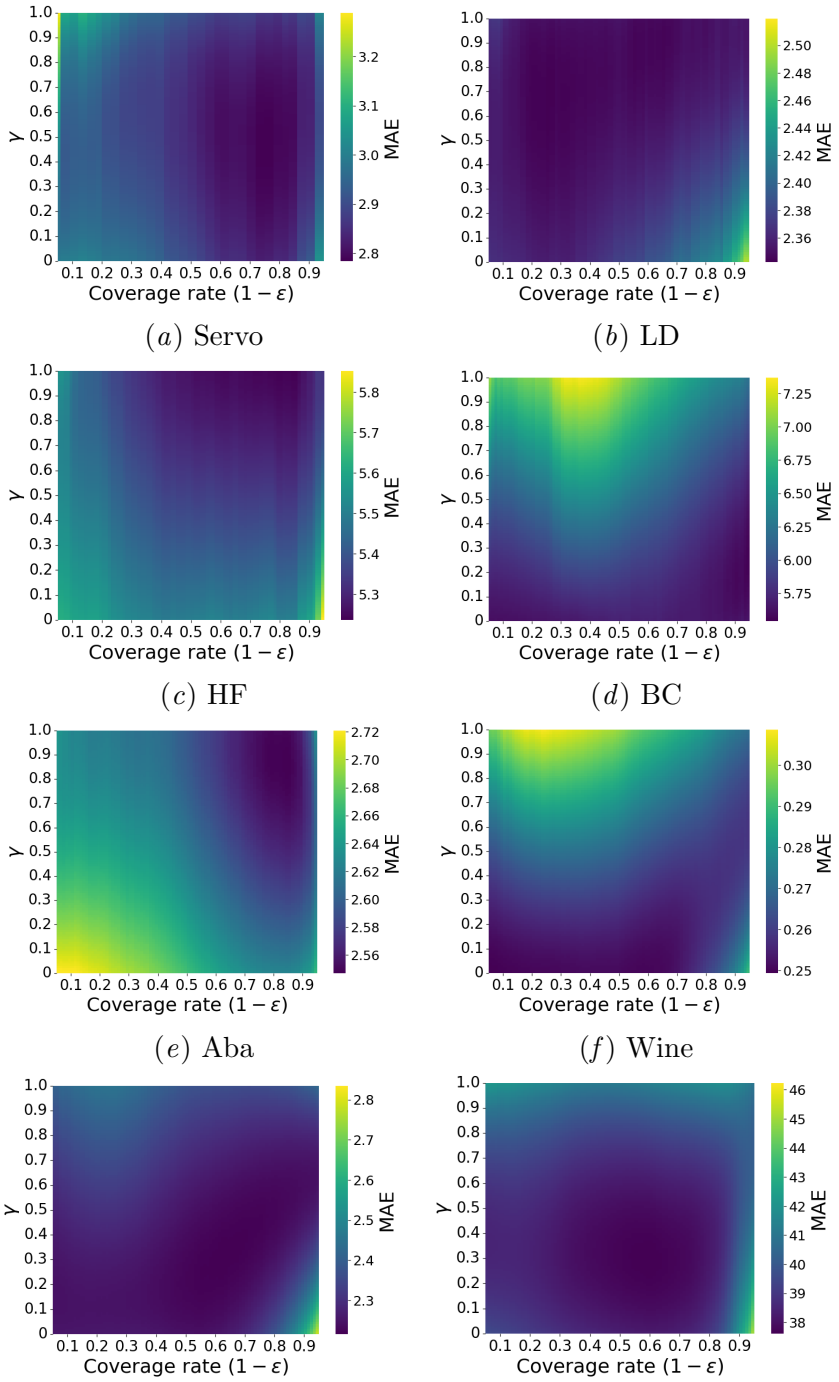
(g) PP (h) App

Figure 7: Influence of different coverage rates on proposed framework with different score functios, where the horizontal axis and vertical axis indicate the coverage rate $(1-\epsilon)$ and MAE values respectively.

Trade-off parameter. Fig. 8 demonstrates the heatmaps of the test datasets. There is no universally optimal combination of coverage rate and γ across all datasets. For Servo, App and PP, a good value is around a moderate level of γ (0.5). For LD, HF and Aba, it is better to choose a high value of γ (0.9). However, for the BC and Wine data sets, it is better to select a low value of γ (0.1). Lower values of γ and higher coverage rates tend to reduce the MAE for most datasets, but the sensitivity of MAE to these parameters varies depending on the dataset. This implies that the optimal trade-off between these parameters is dataset-specific.

Appendix D. Code Link

Here is the anonymous link of the code which contains the details of the main algorithm: `https://anonymous.4open.science/r/CEL-48EC`, taking one data set (HF) as an example.



(a) Servo (b) LD

(c) HF (d) BC

(e) Aba (f) Wine

(g) PP (h) App

Figure 8: Influence of γ on proposed DUEn-WIAA.