Instance Hardness–Based Relevance for Imbalanced Regression

Vitor M. Leitao[†], Juscimara G. Avelino[†], George D. C. Cavalcanti[†], Rafael M. O. Cruz[‡]
[†]Centro de Informática, Universidade Federal de Pernambuco, Recife, Brazil
[‡]École de Technologie Supérieure, University of Quebec, Montreal, Canada
{vml2, jga2, gdcc}@cin.ufpe.br, rafael.menelau-cruz@etsmtl.ca

***Abstract*—Imbalanced regression problems arise when the target variable has an asymmetric distribution, resulting in underrepresented value ranges in the dataset. Traditional approaches for identifying rare instances rely on a relevance function that assigns higher importance to specific regions of the target distribution. However, the effectiveness of imbalance-aware learning methods depends strongly on how relevance is defined. In more complex scenarios, such as bimodal distributions, traditional relevance functions struggle to capture rarity, as they assign fixed relevance values based solely on target values, thereby compromising the distinction between truly rare and normal instances. To address these limitations, this study proposes an Instance Hardness-based relevance function (InHaR) for identifying rare instances in regression problems. Unlike traditional relevance functions, the proposed approach incorporates learning difficulty, allowing rarity to be inferred not only from the target distribution but also from the difficulty of instances for the learning algorithm. This property is particularly important in bimodal scenarios, where rarity cannot be accurately inferred from target values alone. Experimental results demonstrate that the InHaR correctly identifies rare regions under bimodal distributions and, when used to guide resampling strategies such as Random Oversampling (RO) and Gaussian Noise (GN), leads to significant improvements in predictive performance compared to traditional relevance-based approaches. The code, dataset, and further details about the proposed method are publicly available at https://github.com/VitorLeitao/instance-hardness-Imbalanced-regression.**

***Index Terms*—Imbalanced Regression, Relevance Function, Instance Hardness.**

I. Introduction

In machine learning, imbalanced problems occur when the data distribution is asymmetric, so that certain regions of the space of interest contain a significantly smaller number of instances than other, better-represented regions. This scenario has been widely studied in classification tasks [1]–[3], but it also arises in regression problems [4], where the target variable exhibits non-uniform distributions characterized by the presence of rare or hard-to-model values [5]. Imbalanced regression can negatively impact the performance of predictive models, as they tend to favor denser regions of the distribution at the expense of less frequent regions [6].

The identification of rare instances is a fundamental step and has traditionally been addressed in the literature through a relevance function[1], which classifies instances as rare or normal based on the target variable distribution. In this context, rare instances correspond to observations in extreme or low-density regions of the distribution, whereas normal instances comprise the majority of samples associated with frequent target values. This identification is required for applying data resampling strategies such as Random Oversampling (RO) and Gaussian Noise (GN) [8]. However, despite its widespread use, the relevance function proposed by Ribeiro (2011) [7] presents important limitations, particularly in bimodal distributions. In such scenarios, relevance is assigned based on global statistical thresholds derived from the target values, which can lead to an inadequate characterization of rarity. As a consequence, the function fails to distinguish rare from normal samples located in low-density regions between modes, assigning uniformly low or near-zero relevance to these instances and effectively treating them as non-relevant [9], [10]. Figure 1 illustrates this behavior.

In light of this limitation, this work proposes an Instance Hardness-based relevance function (InHaR) for regression problems, grounded in the Instance Hardness (IH) concept [11], which quantifies how difficult an instance is for a learning algorithm to correctly predict. The motivation for this approach is supported by findings in the literature indicating that rare cases are often associated with higher prediction difficulty and tend to be poorly modeled by standard learning algorithms [5], [7]. Leveraging this relationship, the proposed function defines relevance directly in terms of learning difficulty, overcoming limitations of traditional relevance function, particularly their inability to distinguish different degrees of rarity in complex target distributions such as bimodal scenarios.

Unlike relevance functions based exclusively on statistical rarity, Instance Hardness captures the behavior of the learning process itself, allowing instances to be considered relevant even when they are frequent in the data distribution but systematically mispredicted. By moving beyond rarity as a purely distributional concept, this perspective explicitly considers overall predictive performance, leading to a more nuanced characterization of instance importance. As a result, it enables the identification of challenging regions overlooked by traditional relevance functions and allows learning methods to focus on instances that genuinely hinder model performance.

The evaluation of the proposed approach was conducted from three main perspectives: (i) the analysis of the correlation

[1]Throughout this work, the term *traditional relevance function* refers to the relevance function proposed by Ribeiro (2011) [7].

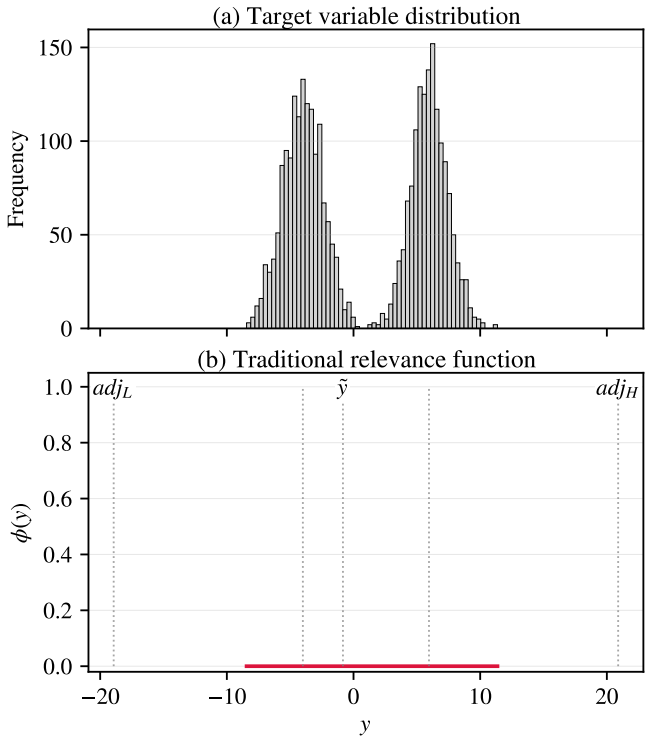

Fig. 1. Bimodal target distribution and the corresponding traditional relevance function. (a) Histogram of the target variable. (b) Tukey-based relevance function $\phi(y)$ [7], whose control points are derived from boxplot statistics. The function is anchored at the median $\tilde{y}$ and at the adjacent lower and upper limits, defined as $adj_L = Q1 - 1.5 \cdot IQR$ and $adj_H = Q3 + 1.5 \cdot IQR$, where $Q1$ and $Q3$ denote the first and third quartiles and $IQR = Q3 - Q1$. Values outside these limits are assigned higher relevance.

between Instance Hardness (IH) and the traditional relevance function (ϕ); (ii) the assessment of the InHaR function in scenarios with bimodal distributions; and (iii) the use of the InHaR in resampling strategies. In the first perspective, we investigated whether there is a significant correlation between the concepts of rarity and instance difficulty, that is, whether instances considered rare, associated with high values of ϕ, also correspond to difficult instances. In the second perspective, the effectiveness of the InHaR function in correctly identifying different degrees of rarity in regressions with bimodal distributions was validated, highlighting its ability to overcome the limitations of the relevance function [7] in these scenarios. Finally, in the third perspective, the use of the InHaR as a criterion in resampling methods was analyzed, replacing relevance functions in the definition of rare and normal instances. The results show that, in approximately 70% of the tests performed, the proposed approach achieved superior performance compared to the traditional approach.

II. Related Work

Imbalance in regression problems has been less explored than in classification, but has recently attracted increasing attention in the literature. This setting arises when certain regions of the target variable distribution are underrepresented, which can compromise a model's ability to accurately predict rare values [5]. Existing approaches to address imbalanced regression are commonly grouped into three categories: regression models, modifications to the learning process, and evaluation metrics [6]. While both individual and ensemble regressors are widely used, their performance under imbalance is often improved by altering the learning process, particularly through preprocessing strategies based on resampling or cost-sensitive learning [7], [8]. Commonly used resampling techniques include Random Oversampling [8], Gaussian Noise [4], SmoteR [12], and its extensions such as SMOGN and SmoteR-Geometric [13], [14]. Finally, traditional global metrics such as MSE may fail to reflect performance in underrepresented regions of the target space, motivating the proposal of evaluation measures specifically tailored to imbalanced regression, including the Squared Error Relevance Area (SERA) [15].

In imbalanced regression, resampling strategies commonly rely on a relevance function, originally proposed in [7], which is used to identify rare instances based on the target variable. The relevance function maps each target value $y \in Y$ to a relevance score $\phi(y) \in [0, 1]$, where higher values indicate greater importance or rarity. This mapping is typically constructed using Piecewise Cubic Hermite Interpolating Polynomials (pchip) over a set of control points, defined either through domain knowledge or automatically [15], [16].

In its automatic formulation, the control points are derived from Tukey's boxplot statistics [17]. Specifically, the relevance function is anchored at the median of the target variable distribution and at the adjacent lower and upper limits, defined as $adj_L = Q1 - 1.5 \cdot IQR$ and $adj_H = Q3 + 1.5 \cdot IQR$, where $Q1$ and $Q3$ are the first and third quartiles and $IQR = Q3 - Q1$. Target values outside these adjacent limits are assigned higher relevance scores, reflecting their lower frequency in the distribution. Given a user-defined relevance threshold t_R, instances are classified as rare or normal, guiding resampling and evaluation procedures.

Despite its widespread adoption, recent studies have highlighted limitations of the relevance function, particularly in multimodal distributions. In such scenarios, distinct regions of the target space may receive similar relevance scores, leading to misclassification of rare and normal instances across different modes [9]. Distance-based alternatives, such as the Distance-Based Relevance Function [10], have been proposed to alleviate this issue by incorporating local density information. Nevertheless, these approaches remain fundamentally dependent on the marginal distribution of the target variable, disregarding the attribute space and the learning difficulty associated with individual instances.

Given this scenario, a gap in the literature exists regarding rarity criteria that do not rely solely on the statistical distribution of the target variable. In this context, this work introduces a relevance criterion grounded in the concept of Instance Hardness [18], which quantifies the difficulty of predicting individual instances. The proposed approach does not assume that statistical rarity and learning difficulty are equivalent; instead, it defines relevance directly in terms of predictive

difficulty, allowing instances that are hard to learn to be treated as rare regardless of their frequency in the target distribution. By shifting the definition of relevance from distributional properties to learning behavior, this work extends relevance-based resampling methods and offers an alternative perspective for addressing imbalanced regression.

III. THE PROPOSED RELEVANCE FUNCTION

The Algorithm 1 presents the complete procedure for the proposed relevance function, called InHaR. The method receives as input a dataset D and a relevance threshold τ, and outputs two disjoint sets containing rare (D_R) and normal (D_N) instances. The core idea is to define rarity based on predictive difficulty, quantified through the Instance Hardness (IH) measure.

Algorithm 1 Instance Hardness-based relevance function (InHaR)

Require: Dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, threshold τ
Ensure: Set of rare instances D_R, set of normal instances D_N
Compute the Instance Hardness values for all instances:
  $IH \leftarrow$ `instance_hardness`(D)
Initialize $D_R \leftarrow \emptyset$, $D_N \leftarrow \emptyset$
for $i = 1$ to n **do**
  if $IH_i \geq \tau$ **then**
    $D_R \leftarrow D_R \cup \{(\mathbf{x}_i, y_i)\}$
  else
    $D_N \leftarrow D_N \cup \{(\mathbf{x}_i, y_i)\}$
  end if
end for
return D_R, D_N

The first step of the algorithm consists of computing the Instance Hardness values for all instances in the dataset. In this work, IH is defined according to the formulation proposed for regression tasks in [18], as shown in Equation 1. Given an instance (x_i, y_i), its difficulty is estimated from the prediction errors produced by a set of regressors $\mathcal{L}$, where $h_j(x_i)$ denotes the output of regressor j for instance x_i. The normalization term is defined as $\gamma = \frac{1}{n}\sum_i y_i^2$, ensuring scale invariance across different datasets.

$$IH_{\mathcal{L}}(x_i, y_i) = 1 - \frac{1}{|\mathcal{L}|}\sum_{j=1}^{|\mathcal{L}|} \exp\left(-\frac{d\big(y_i, h_j(x_i)\big)}{\gamma}\right) \quad (1)$$

As illustrated in Algorithm 1, the computed IH values are then used as a relevance criterion. Since IH assumes continuous values in the interval $[0, 1]$, higher values indicate instances that are more difficult to predict and, consequently, potentially more relevant in imbalanced regression scenarios. To separate rare and normal instances, a relative threshold τ is defined over the IH distribution. Instances with $IH \geq \tau$ are assigned to the rare set D_R, while those with $IH < \tau$ are classified as normal and assigned to D_N.

By relying on predictive difficulty rather than exclusively on the marginal distribution of the target variable, the proposed relevance function incorporates information from the attribute space and the behavior of learning models. This allows the InHaR to be seamlessly integrated into data resampling techniques such as Random Oversampling (RO) and Gaussian Noise (GN), replacing traditional relevance functions while preserving their general workflow. The specific value of the threshold τ is treated as a method parameter and is defined experimentally, as detailed in the methodology section.

IV. EXPERIMENTAL PROTOCOL

Datasets. To evaluate the behavior of the resampling methods guided by the InHaR in different imbalanced regression scenarios, we conducted a comprehensive experimental investigation involving 29 datasets. The main information about the datasets used is available in Table I, where instances whose relevance function ϕ [19] value is greater than 0.7 are considered rare.

Regression Models. To assess the effect of resampling strategies guided by different relevance criteria across diverse learning paradigms, we considered a set of widely adopted regression models with distinct inductive biases: Random Forest Regressor (RF), Bagging Regressor (BG), XGBRegressor (XGB), Support Vector Regressor (SVR), and Multilayer Perceptron Regressor (MLP)[2]. These models were selected because they are commonly used in imbalanced regression studies ([6], [10]) and represent complementary families of learners, including ensemble-based methods, kernel-based models, and neural networks.

Instance Hardness. In the experimental evaluation, Instance Hardness (IH) was used as the sole criterion to estimate instance relevance. Following the definition introduced in Section 3, IH was instantiated using error-based measures computed from the regression models defined in the *Regression Models* subsection. The resulting hardness values were normalized to the range $[0, 1]$, where higher values indicate greater prediction difficulty. A fixed threshold ($\tau = 0.7$) was adopted to distinguish rare from normal instances, and these IH scores were then used to guide the resampling process.

Metrics. To evaluate model performance across datasets, we employed two traditional regression metrics: Mean Absolute Error (MAE) and Mean Squared Error (MSE). These metrics provide a general assessment of predictive accuracy and are independent of any relevance definition. Metrics specifically designed for imbalanced regression, such as SERA [15] or adaptations of the F1-score [20], were not adopted as primary evaluation criteria, as they depend explicitly on a predefined relevance function. Relying on evaluation measures that directly incorporate a relevance function could introduce bias into the assessment, particularly when the relevance definition itself is under analysis, potentially favoring methods aligned

[2]RF, BG, SVR, and MLP were implemented using the `scikit-learn` library (version 1.6.1). XGB was implemented using the `xgboost` library (version 3.1.3).

TABLE I
SUMMARY OF THE 29 DATASETS USED, INCLUDING THE NUMBER OF INSTANCES (N), NUMBER OF ATTRIBUTES (p), AND THE PERCENTAGE OF RARE CASES ($\%r$).

Dataset	N	p	$\%r$	**Dataset**	N	p	$\%r$	**Dataset**	N	p	$\%r$
a1	198	11	14.14	acceleration	1732	14	5.14	cpu_small	8192	12	8.70
a2	198	11	11.11	airfoild	1503	5	10.71	debutanizer	2394	7	10.03
a3	198	11	16.16	analcatdata_apnea3	450	11	22.89	fuel_consumption_country	1764	37	9.30
a7	198	11	13.64	available_power	1802	15	8.71	forestFires	517	12	15.28
abalone	4177	8	16.26	boston	506	13	17.98	heat	7400	11	8.97
cocomo_numeric	60	56	16.67	compactiv	8192	21	8.70	kdd_coil_1	316	18	10.76
concreteStrength	1030	8	5.34	lungcancer_shedden	442	24	5.66	maximal_torque	1802	32	7.16
meta	528	65	20.45	mortgage	1049	15	10.10	pdgfr	79	320	12.66
sensory	576	11	11.98	space_ga	3107	6	5.57	treasury	1049	15	10.39
triazines	186	60	10.75	wine	6497	11	23.44				

with that definition rather than reflecting intrinsic predictive performance.

Since this work proposes a new relevance criterion based on Instance Hardness, relying on evaluation measures that directly incorporate a relevance function could introduce bias in favor of the proposed approach. Therefore, MAE and MSE were used as relevance-independent metrics to ensure a neutral comparison among methods. In addition to global performance, MAE and MSE were computed over subsets of instances identified as rare using two relevance criteria: the traditional relevance function and the proposed InHaR. This analysis allows us to assess model behavior across distinct definitions of rarity while maintaining a consistent, unbiased evaluation protocol.

Evaluation Methods. For the experiments comparing traditional resampling approaches with the InHaR, we employed *repeated k-fold* cross-validation. Specifically, a configuration with 5 folds and 2 repetitions (i.e., 2×5 cross-validation) was adopted.

Resampling Methods. To address data imbalance, two classical resampling methods were used: Random Oversampling (RO) and Introduction of Gaussian Noise (GN). RO consists of the random replication of rare instances, while GN generates new synthetic instances by adding Gaussian noise to the attributes of rare instances, while simultaneously removing part of the most frequent instances. These two strategies were chosen because they presented the best empirical performance in the comparative study conducted by Avelino et al. [6], making them suitable references for evaluating the proposed method.

V. RESULTS AND DISCUSSION

This section presents and discusses the experimental results obtained with the proposed approach. The evaluation is conducted from three complementary perspectives: (i) the relationship between Instance Hardness (IH) and the traditional relevance function (ϕ); (ii) the behavior of the InHaR in datasets with bimodal target distributions; and (iii) the impact of InHaR on resampling strategies and the predictive performance of regression models.

A. Correlation between InHaR and ϕ

Initially, an analysis was conducted to investigate the relationship between the traditional relevance function ϕ [19] and the proposed InHaR, assessing whether instances deemed rare by ϕ also correspond to difficult-to-predict instances. For this purpose, the Pearson correlation coefficient was computed between the relevance values and the corresponding Instance Hardness scores. This analysis is not intended to validate equivalence between the two criteria, but rather to assess the extent to which they capture overlapping or complementary information.

Figure 2 presents the distribution of Pearson correlation coefficients between the relevance function and the InHaR across the 29 evaluated datasets. The correlation values are grouped by absolute magnitude into three categories: strong ($|r| \geq 0.6$), moderate ($0.4 \leq |r| < 0.6$), and weak ($|r| < 0.4$). The predominance of moderate to strong correlations indicates that the two measures are related, but does not imply that they identify the same instances as relevant.

To further analyze this relationship at the instance level, instances were conceptually grouped into four categories based on whether they exhibit high or low values of ϕ and Instance Hardness. This analysis reveals instances that are statistically rare and difficult to predict (high ϕ / high IH), as well as those that are frequent and easy to learn (low ϕ / low IH). Importantly, a substantial proportion of instances fall into the high IH / low ϕ category, corresponding to observations that are not considered rare according to the relevance function, yet are consistently difficult to predict. Conversely, instances identified as rare by ϕ may exhibit low Instance Hardness, indicating that statistical rarity does not necessarily imply learning difficulty.

These results show that statistical rarity and prediction difficulty are not equivalent concepts. While IH and ϕ exhibit partial alignment, the InHaR captures difficult instances that are overlooked by relevance functions based solely on the target distribution, supporting its use in resampling.

B. Applying InHaR to Bimodal Distributions

Bimodal distributions of the target variable pose a well-known challenge for traditional relevance functions in imbalanced regression. In particular, relevance functions based solely on the statistical distribution of the target variable, such as the formulation proposed by Ribeiro [7], assign fixed relevance values according to global distributional properties. As illustrated in Figure 1, when the target distribution exhibits

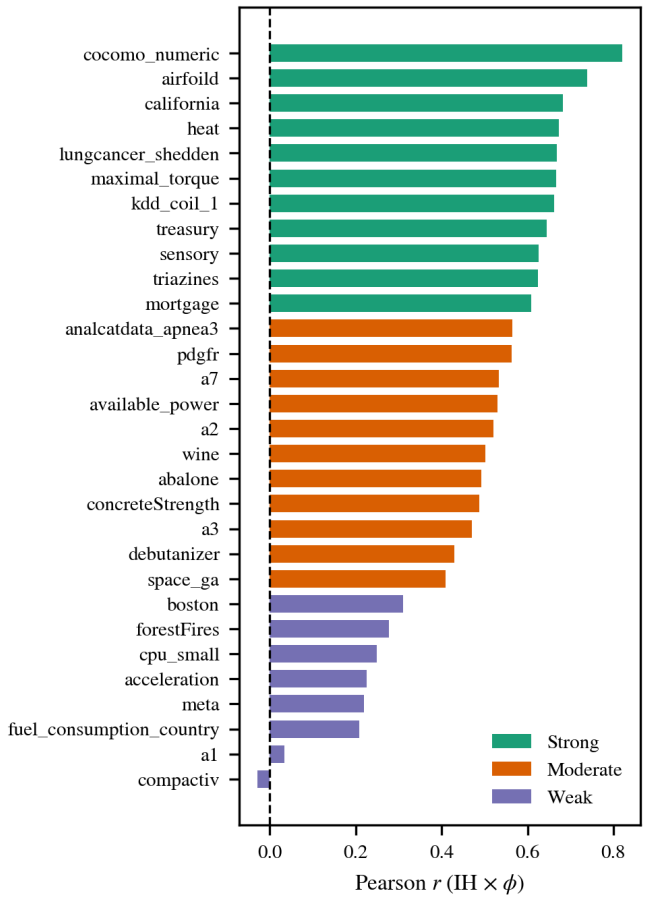

Fig. 2. Distribution of the correlation values between IH and ϕ.

multiple modes, these functions tend to emphasize extreme values while failing to adequately capture low-density regions between modes, leading to an incomplete characterization of rarity [9].

To analyze this scenario in a controlled setting, we generated a synthetic regression dataset with a bimodal target distribution, as illustrated in Figure 1. The dataset consists of two input features sampled from a standard normal distribution, while the target variable is generated from two distinct linear relationships with different offsets and coefficients. This process results in two well-separated modes in the target space, while maintaining overlapping regions in the input space, making the identification of rare or difficult regions non-trivial.

Figure 3 illustrates the behavior of the InHaR in the bimodal scenario. Unlike traditional relevance functions, the proposed approach identifies regions of the target distribution associated with higher prediction difficulty, even when these regions are not located at the extreme tails. Such regions correspond to portions of the bimodal structure that are harder to model and are therefore characterized as rare according to the InHaR criterion. This identification is performed by discretizing the target variable into equidistant bins and computing the median Instance Hardness value within each bin. A global threshold is then defined, and bins whose median IH falls within the top 30% of observed values—corresponding to the 70th percentile of the IH distribution in our experiments—are labeled as rare. This percentile was selected based on preliminary experiments, while a more systematic sensitivity analysis is left for future work. .

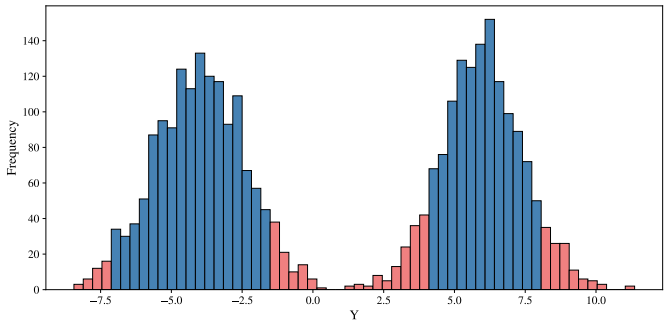

Fig. 3. Identification of rare regions in a bimodal target distribution using the median Instance Hardness (IH) per bin.

Following this analysis, experiments were conducted on this bimodal dataset to assess whether resampling strategies guided by Instance Hardness can improve predictive performance in regions that are systematically hard to learn. In this setting, traditional relevance functions were not considered, as they are unable to meaningfully define rarity in bimodal target distributions.

Table II reports the comparative results obtained using InHaR-based resampling strategies within RO and GN. For both MAE and MSE, InHaR-based resampling achieved three wins out of the five evaluated models. These results indicate that prioritizing hard-to-predict instances during resampling can improve predictive performance even in scenarios where imbalance cannot be characterized solely by target frequency, reinforcing the role of Instance Hardness as an effective relevance criterion for complex target distributions.

TABLE II
COMPARATIVE RESULTS OF WINS IN THE BIMODAL DISTRIBUTION BY METHOD FOR MAE AND MSE.

Model	MAE			MSE		
	None	InHaR-RO	InHaR-GN	None	InHaR-RO	InHaR-GN
BG	0.0593	**0.0592**	0.1132	**0.0124**	0.0126	0.0326
MLP	0.0467	**0.0423**	0.0657	0.0046	**0.0038**	0.0082
RF	**0.0474**	0.0480	0.1015	0.0103	**0.0102**	0.0286
SVR	0.1182	0.1154	**0.0849**	0.1155	0.1095	**0.0508**
XGB	**0.0684**	0.0691	0.1256	**0.0106**	0.0107	0.0326

In a second analysis, the robustness of the proposed method was investigated through controlled variations in the synthetic data generation process, resulting in six distinct datasets. In all cases, the following characteristics were kept fixed: (i) 3,000 instances, (ii) two input variables, and (iii) a bimodal target distribution. The only factor varied was the level of Gaussian noise added to the data, ranging from 0 to 20.

Figure 4 illustrates the behavior of the InHaR under increasing noise levels. As noise increases, the target distribution

gradually loses its clear bimodality and becomes smoother. Despite this change, the InHaR continues to highlight regions associated with higher prediction difficulty, particularly in less represented regions near the extremes of the target space. These results suggest that the proposed approach remains effective at identifying rare regions under moderate noise levels, even as the bimodal structure becomes less pronounced.

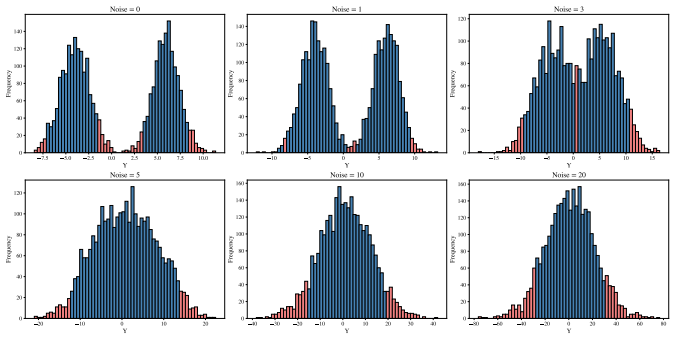

Fig. 4. Behavior of the InHaR under increasing noise levels in a bimodal target distribution.

C. Applying InHaR to Resampling Strategies

In the resampling process, the InHaR is investigated as an alternative criterion for identifying relevant instances. To evaluate the generalization capability of the proposed methodology, an experimental analysis was conducted on 29 datasets (Table I), comparing MAE and MSE across different training strategies for regression models. Specifically, the following were considered: (i) the use of the original dataset, without applying resampling; (ii) resampling guided by the traditional relevance function (ϕ) [4]; (iii) resampling based on the *Distance-Based Relevance Function* (DRF) [10], which takes into account the rarity of observations based on the distance between target values; and (iv) resampling guided by the InHaR, proposed in this work.

Table III presents an overview of the comparative results in terms of the number of wins achieved by each method for both MAE and MSE metrics. This table provides an initial descriptive analysis of the performance tendencies before considering statistical significance.

In this comparison, five distinct models were evaluated on 29 datasets, totaling 145 model–dataset pairs. As shown in Table III, considering the MAE metric, the resampling functions that used the InHaR achieved superior performance in 68 pairs, with 48 associated with InHaR-RO and 20 with InHaR-GN. These results outperformed the other approaches in approximately 50% of the evaluated pairs. For the MSE metric, 58 pairs performed better: 38 related to InHaR-RO and 20 to InHaR-GN, corresponding to about 40% of the tested pairs.

Table IV summarizes the pairwise Wilcoxon signed-rank test results, which were employed due to their suitability for paired comparisons when the normality of performance differences across datasets cannot be assumed. The results show a consistent advantage of the InHaR-based strategy across both RO and GN paradigms. For both MAE and MSE metrics, InHaR-based variants achieve substantially more wins over their respective baselines, with most of these improvements being statistically significant. Compared with the DRF variants, InHaR-based methods also show a higher number of wins, and a considerable portion of these differences is statistically significant. Importantly, when InHaR-based methods incur losses, these are predominantly non-significant, indicating that the proposed strategy rarely leads to statistically meaningful performance degradation. Overall, the results confirm that the InHaR function provides robust and statistically reliable improvements for both RO and GN.

TABLE IV
WILCOXON PAIRWISE COMPARISON (INHAR VS. OTHERS). WINS/LOSSES AND SIGNIFICANT RESULTS ($p < 0.05$) FOR MAE AND MSE.

Comparison	MAE		MSE	
	Win (sig.)	Loss (sig.)	Win (sig.)	Loss (sig.)
InHaR-RO vs RO	**27 (22)**	2 (0)	**20 (13)**	9 (1)
InHaR-RO vs DRF-RO	**17 (11)**	12 (0)	**13 (9)**	16 (1)
InHaR-GN vs GN	**24 (19)**	5 (4)	**20 (19)**	9 (5)
InHaR-GN vs DRF-GN	**23 (15)**	6 (3)	**16 (8)**	13 (8)

1) Performance Analysis on Rare Instances: Model performance is evaluated on subsets of rare instances identified under different relevance criteria. Mean Absolute Error (MAE) and Mean Squared Error (MSE) are computed over instances classified as rare either by the traditional relevance function (ϕ) or by the proposed InHaR, with the goal of analyzing how resampling strategies behave when assessed on the regions prioritized by each criterion.

Table V presents pairwise Wilcoxon comparisons between InHaR-based resampling methods and their respective baselines (RO and GN), considering only instances labeled as rare under each relevance definition. Results are reported as wins and losses across datasets, with statistically significant differences ($p < 0.05$) indicated in parentheses.

When evaluation is restricted to instances identified as rare by the traditional relevance function, InHaR-based resampling methods exhibit fewer wins than losses across both MAE and MSE. This behavior is expected, as resampling is guided by a different relevance criterion. In contrast, when evaluation focuses on instances identified as rare by the InHaR, resampling guided by Instance Hardness consistently achieves more wins across both RO and GN for both error metrics. These results indicate that relevance-based resampling strategies should be evaluated within the regions defined by their own relevance criteria, as assessing one criterion on subsets identified by another may lead to misleading conclusions.

An additional analysis was conducted on instances classified as normal by both relevance criteria. In this scenario, InHaR-based resampling methods exhibit a clear advantage over standard RO and GN, with a large number of statistically significant wins and very few losses. This behavior indicates that, unlike traditional relevance-based resampling, the InHaR does not introduce substantial performance degradation in well-represented and easier regions of the data space.

TABLE III
COMPARATIVE RESULTS OF WINS BY METHOD FOR MAE AND MSE

Model	MAE							MSE						
	None	**RO**	**InHaR-RO**	**DRF-RO**	**GN**	**InHaR-GN**	**DRF-GN**	**None**	**RO**	**InHaR-RO**	**DRF-RO**	**GN**	**InHaR-GN**	**DRF-GN**
BG	5	5	**9**	3	0	6	1	**8**	4	6	5	0	5	1
MLP	4	6	**10**	4	1	4	0	7	5	**8**	4	1	4	0
RF	6	6	**8**	2	0	6	1	6	5	**9**	4	0	5	0
SVR	**11**	1	**11**	4	0	1	1	7	1	5	**9**	2	3	2
XGB	6	6	**10**	4	0	3	0	6	6	**10**	4	0	3	0
Total	32	24	**48**	17	1	20	3	34	21	**38**	26	3	20	3

TABLE V
PAIRWISE COMPARISON OF INHAR-BASED RESAMPLING METHODS AGAINST RO AND GN ON SUBSETS OF INSTANCES DEFINED BY DIFFERENT RELEVANCE CRITERIA. WINS AND LOSSES ACROSS DATASETS ARE REPORTED, WITH STATISTICALLY SIGNIFICANT RESULTS ($p < 0.05$) SHOWN IN PARENTHESES.

Subset definition	Metric	InHaR-RO vs RO	InHaR-GN vs GN
ϕ-rare	MAE	9 (7) / 20 (20)	7 (6) / 22 (22)
ϕ-rare	MSE	9 (7) / 20 (20)	7 (6) / 22 (22)
InHaR-rare	MAE	**15 (9)** / 14 (6)	**16 (9)** / 13 (11)
InHaR-rare	MSE	**15 (9)** / 14 (6)	**16 (9)** / 13 (11)
Normal (ϕ & InHaR)	MAE	**26 (26)** / 3 (2)	**27 (26)** / 2 (2)
Normal (ϕ & InHaR)	MSE	**26 (26)** / 3 (2)	**27 (26)** / 2 (2)

These results highlight an important trade-off in relevance-guided resampling. Traditional relevance functions tend to aggressively oversample statistically rare regions, often at the expense of performance in normal regions. By contrast, the InHaR prioritizes instances based on learning difficulty, resulting in a less aggressive and more balanced resampling strategy that improves performance on difficult cases while preserving accuracy on normal instances.

VI. CONCLUSION

In conclusion, the results show that defining the InHaR function is a viable alternative to traditional distribution-based relevance functions, particularly in bimodal target distributions where static relevance assignments fail to capture local variations in learning difficulty. Using relevance-independent metrics (MAE and MSE), resampling strategies guided by the InHaR consistently improved predictive performance when compared to both the original data and traditional relevance-based resampling. The use of relevance-independent metrics avoids potential bias that may arise when evaluation measures are directly coupled with the relevance definition under analysis, while still indicating that prioritizing hard-to-predict instances can effectively guide learning.

Finally, it is necessary to identify the limitations of the proposed approach. Since it relies on a pool of machine learning models, the computational cost may be high for datasets with a large number of records. Moreover, the choice of the threshold plays an important role in controlling how many instances are prioritized during resampling, potentially affecting model performance. While a fixed threshold was adopted based on preliminary experiments, a systematic analysis of threshold sensitivity was not conducted, leaving open the question of whether a single threshold range is sufficient across datasets or whether dataset-specific tuning is required.

REFERENCES

[1] H. Guo, Y. Li, J. Shang, M. Gu, Y. Huang, and B. Gong, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017.

[2] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[3] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[4] P. Branco, R. P. Ribeiro, and L. Torgo, "Ubl: an r package for utility-based learning," *arXiv preprint arXiv:1604.08079*, 2016.

[5] D. Kowatsch, N. M. Müller, K. Tscharke, P. Sperl, and K. Böttinger, "Imbalance in regression datasets," Fraunhofer AISEC, Garching near Munich, Germany, Tech. Rep.

[6] J. G. Avelino, G. D. C. Cavalcanti, and R. M. O. Cruz, "Resampling strategies for imbalanced regression: a survey and empirical analysis," *Artificial Intelligence Review*, vol. 57, no. 82, 2024.

[7] R. Ribeiro, "Utility-based regression," *Ph. D. dissertation*, 2011.

[8] P. Branco, L. Torgo, and R. P. Ribeiro, "Pre-processing approaches for imbalanced distributions in regression," *Neurocomputing*, vol. 343, pp. 76–99, 2019.

[9] S. Stocksieker and D. Pommeret, "A comprehensive survey on imbalanced regression: Definitions, solutions, and future directions," 2025, hAL open archive, hal-05213741.

[10] D. D. In and H. Kim, "Distance-based relevance function for imbalanced regression," *Stats*, vol. 8, no. 3, p. 53, 2025.

[11] P. Y. A. Paiva, C. C. Moreno, K. Smith-Miles, M. G. Valeriano, and A. C. Lorena, "Relating instance hardness to classification performance in a dataset: a visual approach," *Machine Learning*, vol. 111, no. 8, pp. 3085–3123, 2022.

[12] L. Torgo, R. P. Ribeiro, B. Pfahringer, and P. Branco, "Smote for regression," in *Portuguese conference on artificial intelligence*. Springer, 2013, pp. 378–389.

[13] P. O. Branco, L. Torgo, and R. P. Ribeiro, "Smogn: a pre-processing approach for imbalanced regression," in *First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, vol. 74, 2017, pp. 36–50.

[14] L. Camacho, G. Douzas, and F. Bacao, "Geometric smote for regression," *Expert Systems with Applications*, p. 116387, 2022.

[15] R. P. Ribeiro and N. Moniz, "Imbalanced regression and extreme value prediction," *Machine Learning*, vol. 109, pp. 1803–1835, 2020.

[16] R. L. Dougherty, A. S. Edelman, and J. M. Hyman, "Nonnegativity-, monotonicity-, or convexity-preserving cubic and quintic hermite interpolation," *Mathematics of Computation*, vol. 52, no. 186, pp. 471–494, 1989.

[17] J. Tukey, "Exploratory data analysis, limited prelim. ed," 1970.

[18] G. P. Torquette, V. S. Nunes, P. Y. A. Paiva, and A. C. Lorena, "Instance hardness measures for classification and regression problems," Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos, SP, Brazil, Tech. Rep., 2024, published: 27 February 2024.

[19] L. Torgo, R. P. Ribeiro, and B. Pfahringer, "Utility-based regression," in *Lecture Notes in Computer Science*. Springer, 2009, vol. 5812.

[20] L. Torgo and R. Ribeiro, "Precision and recall for regression," in *International Conference on Discovery Science*. Springer, 2009, pp. 332–346.