ITBoost: Information-Theoretic Trust for Robust Boosting

Ye Su[1,2], **Longlong Zhao**[1*], **Diego García-Gil**[3*], **Jipeng Guo**[4], **Gangchun Zhang**[1], **Jinxin Chen**[1], **Jinsong Chen**[1]

[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Department of Software Engineering, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada
[4]College of Information Science and Technology, Beijing University of Chemical Technology
{ye.su, ll.zhao, gc.zhang, jx.chen2, js.chen}@siat.ac.cn, djgarcia@ugr.es, guojipeng@buct.edu.cn

Abstract

Gradient boosting remains a strong and widely used method for tabular data learning, but its performance often degrades when training labels are noisy. This behavior is largely related to the way boosting algorithms emphasize samples with large gradients, without explicitly accounting for whether such errors originate from informative hard cases or from unreliable labels. We address this issue by reconsidering how sample reliability is evaluated during boosting. Instead of relying on instantaneous error, we examine the evolution of each sample's residuals across iterations. Based on this insight, we propose Information-Theoretic Trust Boosting (ITBoost), which uses the Minimum Description Length principle to measure the complexity of residual trajectories. Samples whose residual patterns fluctuate in an irregular manner are treated as less trustworthy and are down-weighted during learning. Theoretically, we derive a tighter generalization bound for ITBoost under label noise. Empirical results on various tabular benchmarks indicate that ITBoost provides improved robustness in noisy environments over leading boosting and deep tabular models, while retaining best average performance on clean data.

1 Introduction

Gradient Boosting Decision Tree (GBDT) have emerged as the standard for structured data modeling, delivering state-of-the-art performance across a wide range of machine learning tasks [Bentéjac *et al.*, 2021; Shwartz-Ziv and Armon, 2022]. The effectiveness of these algorithms lies in their iterative learning paradigm, specifically, by sequentially training weak learners to fit the residuals of the ensemble formed by previous learners [Friedman, 2001]. A core driver of this process is the model's continuous effort to address its own prediction errors: the magnitude of the gradient essentially guides where the model focuses its learning attention in the next iteration.

Yet, this focus on prioritizing large gradients stands out as a key flaw in gradient boosting. Label noise often causes sharp, even catastrophic, performance declines [Long and Servedio, 2008; Miao *et al.*, 2015; Einziger *et al.*, 2019]. Mislabeled samples are persistent sources of error, and their large gradients slowly lead the model astray. Figure 1 clearly shows this problem: as noise levels rise from 10% (a) to 30% (b), the decision boundary of a standard GBDT becomes fragmented. Here, the model's rigid reliance on gradient magnitude as a guiding cue leads it to distort its structure to fit noisy samples (marked "X"), a choice that severely undermines generalization. This is not just an edge case: label noise is far from uncommon in practical fields like medicine and finance. As such, GBDT's vulnerability poses a major obstacle to its real-world use [Frénay and Verleysen, 2013; Shi *et al.*, 2024].

Current efforts to tackle this issue rely on either robust loss functions [Huber, 1992; Zhang and Sabuncu, 2018] or sample selection strategies [Ferreira and Figueiredo, 2012; Dubout and Fleuret, 2014; Liu *et al.*, 2020], but these approaches fall short of addressing the core issue. They still rest on the flawed premise that gradient magnitude alone dictates a sample's relevance to the true decision boundary. This presents a challenge that's hard to resolve: algorithms cannot tell apart two types of samples that both yield large gradients—genuinely valuable hard examples sitting close to the decision boundary, and erroneous noisy samples with incorrect labels [Song *et al.*, 2022]. If we blunt all large gradients outright, we end up hampering the model's ability to learn from critical hard examples; if we accept them uncritically, noise will inevitably corrupt the model's learning process.

In response to this dilemma, we question this foundational assumption and advance a new perspective: a sample's gradient reliability should not hinge on its immediate magnitude, but rather on how interpretable its error history is. Hard examples, while tough to learn, produce residual sequences with recognizable structural patterns. By contrast, noisy examples, intrinsically misaligned with the true data distribution, result in far more erratic prediction errors. Drawing on this observation, we introduce Information-Theoretic Trust Boosting (ITBoost). ITBoost assigns a "trust score" to each sample's gradient. Instead of only asking, "How

*Corresponding authors

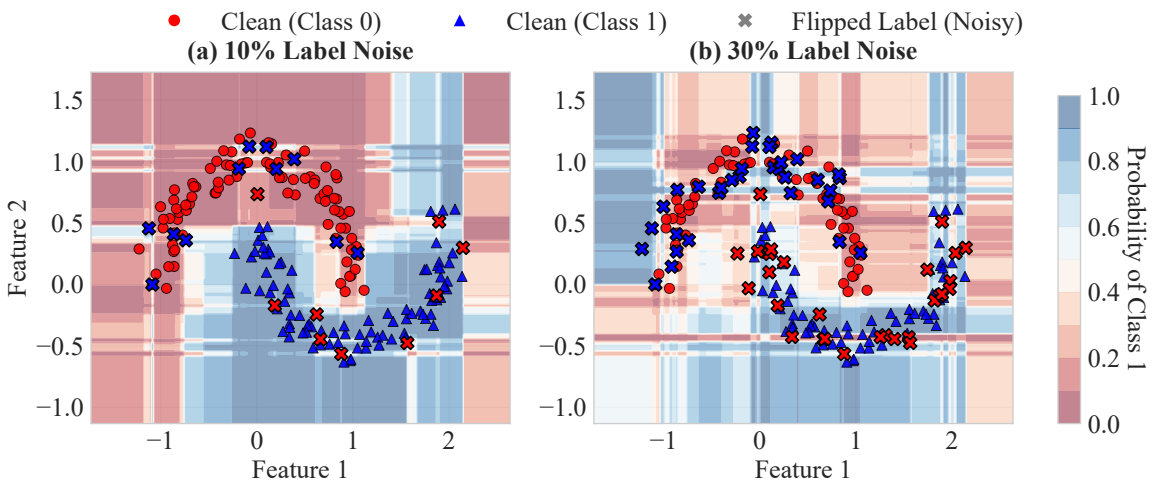

Figure 1: Standard GBDT overfitting noise on data with 10% and 30% label noise.

large is the sample's error?", we shift focus to: "How trustworthy is the historical pattern of its errors?" To formalize this question, we adopt the Minimum Description Length (MDL) principle from information theory [Rissanen, 1978; Grünwald, 2007]. Concretely, we compute the sequential complexity of each sample's residual trajectory online, using the Lempel-Ziv algorithm [Gasieniec *et al.*, 1996] to quantify randomness. A random, uncompressible residual sequence (high complexity) flags the sample as "untrustworthy," prompting us to reduce its weight in the learning process. This new weighting mechanism lets ITBoost effectively de-emphasize noise while still prioritizing genuinely challenging yet informative samples. Our key contributions are summarized below:

- We introduce ITBoost, a robust boosting method that tackles label noise at its root by leveraging an information-theoretic trust framework, an innovative departure from standard boosting.
- We develop an MDL-based sample weighting mechanism; uniquely, this mechanism uses residual time-series characteristics instead of instantaneous values to guide the boosting process.
- We provide theoretical justification for ITBoost, demonstrating that it achieves a tighter generalization error bound than standard GBDT when label noise is present.
- We show that ITBoost delivers optimal performance and robustness, outperforming leading boosting algorithms and modern deep tabular models in noisy environments.

2 Information-Theoretic Boosting

Our work is based on the insight that instantaneous gradient magnitude is an insufficient metric, and often misleading, for assessing a sample's true importance. Indeed, a large gradient may indicate just as readily a valuable, hard-to-learn case as it does a corrupted, noisy counterpart. To clear up this ambiguity, we propose leveraging the full history of a sample's learning trajectory to underpin our assessment of its importance.

2.1 From Gradient Magnitude to Residual History

Let us frame the learning process through a temporal lens. For each sample x_i, the sequence of pseudo-residuals, $\mathbf{g}_i = (g_i^{(1)}, g_i^{(2)}, \ldots, g_i^{(M)})$, constitutes a time series documenting how the model's understanding of that sample evolves over iterations. What we argue here is that the traits of this time series hold far more information about the sample's inherent nature than the magnitude of any single residual element alone.

- **Hard but Clean Samples.** Authentic hard samples exhibit low-complexity residual structures. Specifically, *(i)* boundary samples induce structured sign oscillations as the decision boundary is progressively refined, while *(ii)* persistent hard samples generate stable, monotonic residual sequences due to systematic model underestimation. Although these samples are difficult to learn, both residual patterns are highly compressible and thus have low algorithmic complexity. Guided by the MDL principle, ITBoost correctly identifies them as reliable and assigns higher weights, directing the model toward these valid yet challenging cases.
- **A Noisy Sample.** Samples with corrupted labels conflict with the underlying feature structure. Attempts to fit such data lead to erratic and unpredictable model adjustments, producing residual sequences with chaotic sign flips and no discernible pattern, indicative of high complexity. This behavior clearly distinguishes noisy samples from both the structured oscillations of boundary samples and the stable persistence of hard clean samples.

It is this fundamental distinction in the temporal signature of residual sequences that forms the core insight underpinning our work. An unpredictable, random-like sequence is a strong indicator that the model's errors stem from label cor-

ruption, rather than being an integral part of the structured learning process.

2.2 Quantifying Trust with Algorithmic Complexity

To formalize this intuition, we measure the randomness of a sequence using its algorithmic complexity. The Kolmogorov complexity [Kolmogorov, 1965; Li and Vitányi, 2008], $K(s)$, of a sequence s is the length of the shortest program that generates s. Since $K(s)$ is uncomputable, we employ a practical approximation: the Lempel-Ziv (LZ) complexity, which effectively measures a sequence's compressibility [Wyner and Ziv, 2002; Zozor *et al.*, 2005].

At each iteration m, we consider the binarized residual history for each sample i, $\mathbf{s}_i^{(m)} = (\text{sign}(g_i^{(1)}), \ldots, \text{sign}(g_i^{(m)}))$. This binarization is a crucial design choice, as it intentionally discards the noisy and volatile error magnitude to focus purely on the temporal pattern of the error's direction—a more stable and robust signal of a sample's learning trajectory. We then compute its LZ complexity, $C(\mathbf{s}_i^{(m)})$. We define the information-theoretic trust weight, $w_i^{(m)}$, as the product of the gradient magnitude and a trust term, $\tau_i^{(m)}$:

$$w_i^{(m)} = |g_i^{(m)}| \cdot \tau_i^{(m)}, \tag{1}$$

where the trust term is an exponential penalty based on the normalized complexity of its history:

$$\tau_i^{(m)} = \exp\left(-\bar{C}(\mathbf{s}_i^{(m)})\right). \tag{2}$$

Here, $\bar{C}(\mathbf{s}_i^{(m)})$ is the LZ complexity, normalized to the range $[0, 1]$ across all samples at iteration m. This formulation allows ITBoost to temper the influence of samples that have proven to be historically unreliable.

To compute this normalized complexity, we employ a dynamic Min-Max scaling procedure at each boosting iteration m. For the set of raw LZ complexities $\{C_1, C_2, \ldots, C_N\}$ computed at step m, the normalized complexity for sample i is given by:

$$\bar{C}_i = \frac{C_i - \min(\{C_k\}_{k=1}^N)}{\max(\{C_k\}_{k=1}^N) - \min(\{C_k\}_{k=1}^N)}. \tag{3}$$

This per-iteration, relative normalization is a critical design choice. A key implication is its adaptive response to outliers. It pins the noisy sample's normalized complexity near 1.0, ensuring it receives a significant trust penalty via the exponential term $e^{-\bar{C}_i}$. It compresses the normalized complexities of most clean samples into a narrow band close to 0, thereby shielding them from undue penalization. This dynamic, outlier-driven normalization thus enhances the separation between trustworthy and untrustworthy samples in the trust domain, enabling a more targeted and aggressive suppression of noise.

2.3 Algorithm Summary

The complete ITBoost algorithm is presented in Algorithm 1. The key modification to the standard GBDT framework is the introduction of Step 3, where the information-theoretic trust weights are computed. These weights are then used to guide the training of the weak learner in Step 4, replacing the implicit weighting by gradient magnitude in the standard squared-error loss minimization.

Algorithm 1 The ITBoost Algorithm

Input: Training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, loss $\mathcal{L}(y, F)$, iterations M, learning rate ν.

Initialize $F_0(x) \leftarrow \arg\min_c \sum_{i=1}^N \mathcal{L}(y_i, c)$.
Initialize empty residual histories $s_i \leftarrow$ "" for $i = 1, \ldots, N$.
for $m = 1$ to M **do**
  Step 1. Compute pseudo-residuals:
    $g_i^{(m)} \leftarrow -\left[\frac{\partial \mathcal{L}(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$
  Step 2. Update binarized residual histories:
    $s_i \leftarrow s_i \circ (\text{'1'}$ if $g_i^{(m)} > 0$ else $\text{'0'})$
  Step 3. Compute information-theoretic trust weights:
    Compute LZ complexity for each history: $C_i \leftarrow$ LZ-Complexity(s_i).
    Normalize complexities across samples: $\bar{C}_i \leftarrow$ Normalize$(C_1, \ldots, C_N)$ to $[0, 1]$.
    Compute final weights using Eq. (1): $w_i^{(m)} \leftarrow |g_i^{(m)}| \cdot \exp(-\bar{C}_i)$.
  Step 4. Train weak learner $h_m(x)$ on pseudo-residuals using trust weights:
    $h_m \leftarrow \arg\min_{h \in \mathcal{H}} \sum_{i=1}^N w_i^{(m)} \cdot (g_i^{(m)} - h(x_i))^2$
  Update the ensemble: $F_m(x) \leftarrow F_{m-1}(x) + \nu \cdot h_m(x)$.
end for
Output: $F_M(x)$.

3 Theoretical Analysis

In this section, we provide theoretical justification for the robustness of ITBoost. We demonstrate that our information-theoretic weighting scheme leads to a tighter generalization error bound under a standard model of label noise compared to traditional GBDT.

3.1 Assumptions

Let $\mathcal{F}$ be the function class of the final ensemble, which is the convex hull of the base learner class $\mathcal{H}$. We analyze the generalization error, $\mathbb{E}_{x,y}[\ell(y, F(x))]$, where ℓ is the 0-1 loss. Our analysis will be based on a surrogate loss, the logistic loss $\mathcal{L}(y, F) = \log(1 + \exp(-yF))$, for $y \in \{-1, 1\}$, which is convex and 1-Lipschitz. We adopt a standard model for label noise:

Assumption 1 (Symmetric Label Noise)**.** *The observed labels y_i in the training set $\mathcal{D}$ are generated from the true (clean) labels y_i^* according to an independent process, such that for a noise rate $p \in [0, 1/2)$:*

$$P(y_i \neq y_i^*) = p \quad \textit{and} \quad P(y_i = y_i^*) = 1 - p.$$

Our key insight is that noisy samples will, with high probability, exhibit more complex residual histories. We formalize

this by assuming that the expected complexity of a noisy sample's residual history is higher than that of a clean sample.

Assumption 2 (Complexity Separation)**.** *Let $C(\mathbf{s}_i)$ be the LZ-complexity of a sample's residual history. There exists a gap $\Delta_C > 0$ such that the expected complexity conditioned on the label being noisy versus clean is separated:*

$$\mathbb{E}[C(\mathbf{s}_i)|y_i \neq y_i^*] \geq \mathbb{E}[C(\mathbf{s}_i)|y_i = y_i^*] + \Delta_C. \quad (4)$$

Remark (Scope of Noise Types). It is important to clarify that Assumption 2 is specifically tailored to stochastic or unstructured label noise, the primary type of label noise observed in practical real-world scenarios. In such cases, conflicts between features and their corresponding labels introduce randomness into the model's learning dynamics. A notable limitation, however, emerges when dealing with systemic or adversarial noise. This kind of noise tends to generate stable, low-complexity residual patterns, which bear resemblance to the persistent hard samples discussed in Section 2.1. Notably, without clean validation data or additional external supervision, purely data-driven methods cannot inherently distinguish such structured corruptions from genuine hard examples. Consequently, this category of structured noise falls outside the scope of this work.

This assumption posits that our complexity measure is a meaningful signal for identifying noise, a hypothesis we validate empirically in Section 4. In Eq. (4), we assumed that the expected residual complexity of noisy samples is strictly greater than that of clean samples. While this assumption aligns with intuitive reasoning, it is relatively strong and lacks explicit conditions that guarantee the separation $\Delta_C > 0$. To this end, we reformulate the original assumption into two equivalent, verifiable variants, which also allow for finite-sample probabilistic guarantees.

Assumption 3 (Entropy-Rate or Empirical Separability)**.** *Let each sample i produce a residual-sign sequence $S_i^{(T)} = (s_{i,1}, ..., s_{i,T})$ with normalized complexity $\widehat{C}_T(S_i) \in [0, 1]$. We assume one of the following holds:*

(A) Entropy-rate gap. *Residual sequences for clean and noisy samples can be modeled as two stationary ψ-mixing processes with entropy rates H_{clean} and H_{noisy}, satisfying $H_{noisy} - H_{clean} \geq \Delta_H > 0$. Under the standard Lempel–Ziv consistency theorem [Wyner and Ziv, 2002], the normalized complexity estimator $\widehat{C}_T$ converges exponentially to the true entropy rate:*

$$Pr(|\widehat{C}_T - H| \geq \epsilon) \leq c_1 e^{-c_2 T \epsilon^2}, \quad (5)$$

so for sufficiently large T, we obtain a probabilistic separation $\mathbb{E}[\widehat{C}_T \mid clean] \geq \Delta_c > 0$ with $\Delta_c \simeq \Delta_H - 2\epsilon_T$.

(B) Empirical separability. *For finite samples, if $\widehat{C}_T \in [0, 1]$, let $\bar{C}_{clean}$ and $\bar{C}_{noisy}$ denote the empirical means. Then, by Hoeffding's inequality,*

$$Pr(|\bar{C} - \mathbb{E}\widehat{C}| > \epsilon) \leq 2e^{-2n\epsilon^2}. \quad (6)$$

Hence, whenever the true expectation gap $\Delta_c^\star > 2\epsilon$, the empirical means are separable with probability at least $1 - \delta$, provided $n \geq \frac{1}{2\epsilon^2} \log \frac{2}{\delta}$.

These two formulations replace the untestable Kolmogorov-complexity gap by measurable or statistically verifiable criteria. Empirical separability (B) can be directly checked in experiments, while the entropy-rate formulation (A) connects to the asymptotic theory of universal compression.

3.2 Generalization Bound under Label Noise

With these assumptions, we first establish the statistical properties of our trust weights. The following lemma characterizes the behavior of the trust term τ derived from the complexity random variable.

Lemma 1 (Hoeffding / sub-Gaussian bounds for τ)**.** *Let C be a real-valued random variable with mean $\mu = \mathbb{E}[C]$ and define $\tau := \mathbb{E}[e^{-C}]$. Then:*

1. ***Jensen lower bound:*** *Because $f(x) = e^{-x}$ is convex, $\tau \geq e^{-\mu}$.*
2. ***Bounded-range upper bound:*** *If $C \in [a, b]$ with range $R = b - a$, Hoeffding's lemma gives $\tau \leq \exp(-\mu + R^2/8)$.*
3. ***Sub-Gaussian upper bound:*** *If $C - \mu$ is σ^2-sub-Gaussian, then $\tau \leq \exp(-\mu + \sigma^2/2)$.*

This lemma makes explicit that the expected trust weight τ is tightly constrained by the expected complexity μ. Specifically, an increase in residual complexity brings about an exponential decay in the upper bound of the trust weight—which in turn furnishes the mathematical basis needed to justify the use of complexity scores for down-weighting samples.

Proof. The detailed proof is provided in Appendix A.1 in supplementary material, please. □

Based on Lemma 1, we quantify the relative weighting of noisy versus clean samples.

Proposition 1 (Ratio bound for noisy vs. clean samples)**.** *Let C_{noisy} and C_{clean} denote the residual-complexity variables of noisy and clean samples, with expectations μ_{noisy} and μ_{clean}. Define the complexity gap $\Delta_c = \mu_{noisy} - \mu_{clean} > 0$. Then, if both C's lie in $[a, b]$ or are σ^2-sub-Gaussian:*

$$\frac{\tau_{noisy}}{\tau_{clean}} \leq \exp\left(-\Delta_c + corr\right), \quad (7)$$

where corr is a small correction term related to the variance or range of C.

What this proposition makes clear is that, as long as the complexity gap Δ_c surpasses the correction term (derived from variance), noisy samples will be down-weighted exponentially in comparison to clean ones. This mechanism lays theoretical groundwork for the algorithm's capacity to suppress noise influence, guiding gradient updates to align more closely with the direction dictated by the clean data distribution.

Proof. We provide the full derivation in Appendix A.2 in supplementary material, please. □

Finally, we state our main theoretical result regarding the generalization error.

Theorem 1 (Quantitative Effect of Complexity Separation)**.** *Assume the loss ℓ is L-Lipschitz, weak learners are bounded by B, and the trust-weight mapping is Lipschitz. Under Assumption 3, if $\Delta_c > 0$ and complexities are sub-Gaussian, then the expected empirical risk improvement per boosting iteration satisfies:*

$$\mathbb{E}[R_{t+1}(F) - R_t(F)] \leq -LL_g B\Delta_c + O\left(\frac{1}{\sqrt{n}}\right). \quad (8)$$

Theorem 1 indicates that the generalization risk improvement is explicitly dependent on the complexity gap Δ_c. The term $-LL_g B\Delta_c$ represents a risk reduction derived from the separation between noisy and clean complexities. This suggests that larger separability contributes to more robust convergence, offering a theoretical basis for ITBoost's performance in noisy environments compared to standard GBDT which lacks this complexity-aware regularization.

Proof. The complete proof is deferred to Appendix A.3 in supplementary material, please. □

4 Experiments

In this section, we conduct a comprehensive empirical evaluation to validate the performance of ITBoost. The primary goal of this section is to answer a critical question: Does our information-theoretic trust mechanism, designed for robustness, maintain optimal performance on clean and noise datasets? (Section 4.2 and 4.3). Additionally, we conduct an ablation study for residual binarization (Section 4.4) and visualize the information-theoretic trust mechanism of ITBoost (Section 4.5). Additionally, we provide the complexity and efficiency analysis (Section B) in supplementary material. To ensure statistical significance, all comparative results are validated using Friedman tests (Section C) in supplementary material.

Implementation and Reproducibility. All experiments were conducted locally using Jupyter Notebook in an Anaconda environment on Windows 10 (Version 10.0.19045). The hardware infrastructure utilized an Intel processor (Intel64 Family 6 Model 151 Stepping 2, GenuineIntel) equipped with 20 logical cores. The software stack was built on Python 3.12.4, utilizing the following key libraries: pandas 2.2.2, numpy 1.26.4, scikit-learn 1.7.2, statsmodels 0.14.2, imbalanced-learn 0.14.0, lightgbm 4.3.0, xgboost 2.0.3, catboost 1.2.3, ngboost 0.5.7, and tabpfn 6.0.6.

4.1 Setup

Dataset. We used five different datasets from OpenML repository, which are all publicly available. These datasets include Madelon (2,600 samples, 500 features) [Guyon *et al.*, 2007], Jasmine (2,984 samples, 144 features), Bioresponse (3,751 samples, 1,776 features), Creditcard (284,807 samples, 30 features) [Dal Pozzolo *et al.*, 2015], and the OVA_Uterus (1,545 samples, 10,936 features) [Stiglic and Kokol, 2010]. We employed random undersampling [Hasanin and Khoshgoftaar, 2018; Liu and Tsoumakas, 2020] to address class imbalance in the Creditcard and OVA_Uterus datasets.

Baselines. We compared ITBoost against a comprehensive suite of SOTA tabular learning algorithms. This includes six established boosting methods: AdaBoost (ADA) [Freund and Schapire, 1997], GBDT, XGBoost (XGB) [Chen *et al.*, 2015], LightGBM (LGBM) [Ke *et al.*, 2017], CatBoost (CAT) [Prokhorenkova *et al.*, 2018], and NGBoost (NGB) [Duan *et al.*, 2020]. Additionally, we included **TabPFN** [Hollmann *et al.*, 2025], a recent Transformer-based prior-data fitted network, to benchmark our method against modern deep tabular foundation models. *Note that specialized robust boosting variants (e.g., RobustBoost) are excluded from this comparison due to the lack of maintained open-source implementations, consistent with standard practice in recent literature [Luo* et al.*, 2025].* A detailed discussion of these related works is provided in Appendix D. For a fair and reproducible comparison of the algorithms' intrinsic capabilities, we were configured with "random_state=42" to ensure reproducibility, with all other parameters set to the official recommendation values provided by the respective software packages (e.g., Scikit-learn). The proposed ITBoost algorithm was configured in the same manner, following an official setup consistent with these baselines. We did not optimize the parameters of any algorithms. This approach evaluates the fundamental "out-of-the-box" performance of the underlying architectures, a crucial factor for practical applications.

Evaluation Protocol. We employed a 5-fold stratified cross-validation protocol for all experiments. The performance of each algorithm was reported as the mean and standard deviation across the five folds and the best results were highlighted in **bold**, second best are underlined. The primary evaluation metrics were Accuracy (ACC), F1-Score (F1), Area Under the ROC Curve (AUC), and Logarithmic Loss (Log Loss).

4.2 Overall Performance

Table 1 presents the performance comparison of ITBoost against seven SOTA algorithms. The results revealed a clear performance hierarchy. Among the traditional boosting families, ADA generally yielded the lowest performance, while CAT and LGBM demonstrated strong competitiveness. Notably, the recent Transformer-based TabPFN demonstrated strong performance, frequently securing the second-best results. However, ITBoost outperformed all baselines across nearly all datasets and metrics. The performance gap was particularly striking on the *Madelon* dataset, where ITBoost achieves an accuracy of 92.04% compared to 84.38% for the best GBDT baseline (CAT) and 90.50% for TabPFN. Since Madelon is a synthetic dataset known for its high dimensionality and abundance of uninformative, distracting features (feature noise), this result is highly significant. It suggests that our information-theoretic trust mechanism is not only effective against synthetic label flipping (as we will explore in Section 4.3) but also inherently robust against intrinsic data corruption and spurious correlations. By identifying the chaotic residual patterns caused by these noisy features, ITBoost effectively acts as a dynamic noise filter on the "clean" data.

Dataset	Metric	ADA	GBDT	XGB	LGBM	CAT	NGB	TabPFN	ITBoost
OVA_Uterus	ACC (↑)	0.9073	0.8954	0.9033	0.9194	0.9194	0.8548	0.9237	**0.9438**
	F1 (↑)	0.9077	0.8956	0.9046	0.9201	0.9206	0.8554	0.9236	**0.9453**
	AUC (↑)	0.9667	0.9416	0.9661	**0.9777**	0.9731	0.9050	0.9758	0.9675
	Log Loss (↓)	0.4825	0.6694	0.2642	0.2476	0.2277	0.5822	0.2432	**0.1357**
Madelon	ACC (↑)	0.6135	0.7335	0.8146	0.8196	0.8438	0.7335	0.9050	**0.9204**
	F1 (↑)	0.6122	0.7404	0.8155	0.8203	0.8442	0.7408	0.9050	**0.9200**
	AUC (↑)	0.6600	0.8117	0.8843	0.8935	0.9087	0.8107	**0.9666**	0.9542
	Log Loss (↓)	0.6700	0.5397	0.4522	0.4129	0.4064	0.5506	0.2326	**0.1643**
Jasmine	ACC (↑)	0.7922	0.8090	0.7962	0.8046	0.8090	0.7999	0.8324	**0.9179**
	F1 (↑)	0.8128	0.8201	0.8114	0.8214	0.8278	0.8278	0.8489	**0.9185**
	AUC (↑)	0.8439	0.8574	0.8577	0.8608	0.8679	0.8539	0.8930	**0.9415**
	Log Loss (↓)	0.5481	0.4232	0.5027	0.4419	0.4134	0.4424	0.3780	**0.1814**
Bioresponse	ACC (↑)	0.7582	0.7851	0.7982	0.7984	0.7918	0.7665	0.8062	**0.9174**
	F1 (↑)	0.7773	0.8059	0.8150	0.8165	0.8113	0.7915	0.8209	**0.9238**
	AUC (↑)	0.8219	0.8536	0.8675	0.8719	0.8626	0.8316	0.8776	**0.9567**
	Log Loss (↓)	0.6164	0.4759	0.5460	0.4620	0.4611	0.5168	0.4413	**0.1844**
Creditcard	ACC (↑)	0.9655	0.9634	0.9644	0.9685	0.9705	0.9370	0.9614	**0.9726**
	F1 (↑)	0.9655	0.9631	0.9642	0.9679	0.9701	0.9341	0.9605	**0.9726**
	AUC (↑)	0.9881	0.9901	0.9910	0.9918	**0.9921**	0.9851	0.9919	0.9914
	Log Loss (↓)	0.4850	0.1171	0.1223	0.1260	**0.0969**	0.1647	0.1046	0.1076
Average	ACC (↑)	0.8073	0.8373	0.8553	0.8621	0.8669	0.8183	0.8857	**0.9344**
	F1 (↑)	0.8151	0.8450	0.8621	0.8692	0.8748	0.8299	0.8918	**0.9360**
	AUC (↑)	0.8561	0.8909	0.9133	0.9191	0.9201	0.8773	0.9410	**0.9623**
	Log Loss (↓)	0.5604	0.4451	0.3775	0.3381	0.3211	0.4513	0.2799	**0.1547**

Table 1: Performance comparison on the original binary classification datasets (no synthetic noise added).

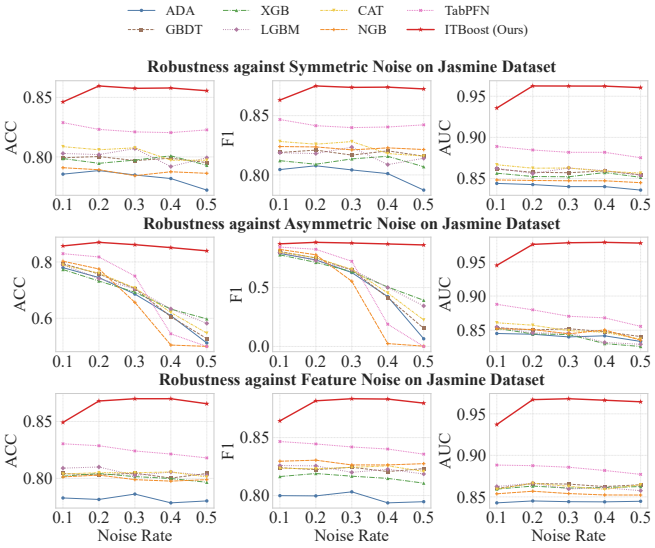

Figure 2: Robustness comparison of boosting algorithms under different types of noise on the Jasmine dataset.

4.3 Robustness Analysis

Figure 2 shows the robustness of leading boosting methods and deep tabular model (TabPFN) under three noise conditions: symmetric label noise, asymmetric label noise, and feature noise on the Jasmine dataset. These results exhibited that, as the noise rate escalates, leading boosting methods exhibit a consistent drop in performance. Notably, while TabPFN outperformed leading boosting methods, it still displayed clear vulnerability, with a more pronounced decline observed under severe asymmetric noise. However, ITBoost maintained consistently high performance even at extreme noise levels, retaining a substantial advantage over leading ensembles and TabPFN. Such empirical evidence provides robust support for our core hypothesis. Leading boosting methods are prone to being misled by spurious correlations: GBDTs rely heavily on gradient magnitude, while TabPFN's fixed priors fail to adapt to instance-specific corruptions. ITBoost's information-theoretic trust mechanism, by contrast, effectively distinguishes the erratic error patterns generated by noisy samples. Through systematically down-weighting the influence of these samples based on residual complexity, ITBoost preserves the model's integrity and its capacity to capture the stable underlying data structure—ultimately achieving optimal robustness and maintaining optimal performance in scenarios where existing SOTA methods struggle or fail.

4.4 Ablation Study

A core design choice underpinning ITBoost's robustness lies in the binarization of residual history: this design intentionally discards error magnitude to focus exclusively on the temporal pattern of error directions. We further validate the necessity of this simplification, whether it is critical for robustness or causes detrimental information loss, through an ablation study. Specifically, we compare ITBoost-Binary with ITBoost-Quantized, a variant that retains magnitude information via four-symbol encoding. On clean Jasmine and Bioresponse datasets (Table 2), the two variants delivered comparable performance; ITBoost-Binary even achieved a slight edge while cutting training time by 20–25%, confirming no compromise to baseline effectiveness. Under label noise conditions (Figure 3), however, ITBoost-Binary outperformed the

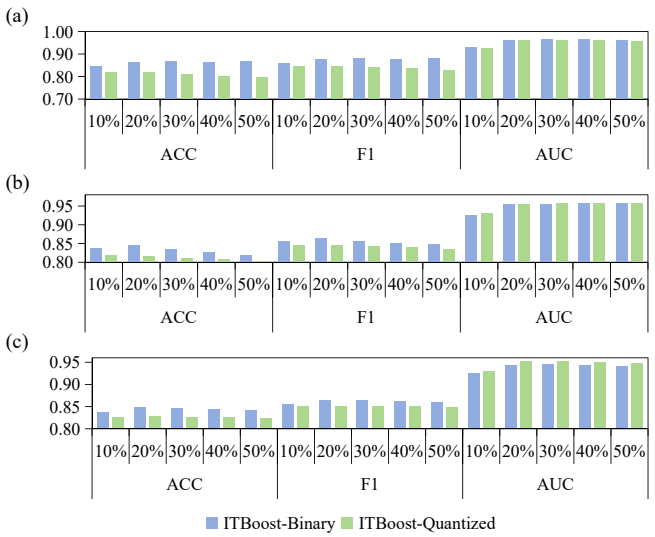

Figure 3: Ablation study for the role of residual binarization in IT-Boost under different types of noise on Jasmine dataset. (a) 10%-50% symmetric noise. (b) 10%-50% asymmetric noise. (c) 10%-50% feature noise.

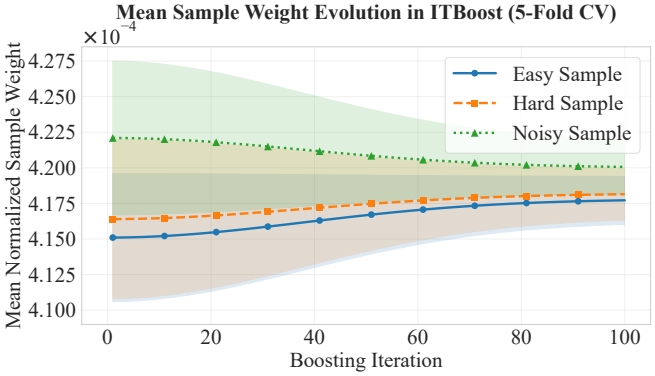

Figure 4: Visualization of information-theoretic trust mechanism for ITBoost: a case study for Jasmine dataset with 20% label noise.

Quantized variant across all metrics. This performance gap widened as noise intensity increased, which confirms that gradient magnitude from mislabeled samples constitutes a spurious signal. By focusing solely on directional history, the binary formulation better captures the chaotic patterns linked to untrustworthy labels. Notably, this advantage fades under feature noise—highlighting that binarization serves as a targeted, effective defense specifically against label noise. In summary, these results demonstrate that binary residual encoding is not a detrimental simplification, but rather a targeted noise-filtering mechanism that underpins ITBoost's robust performance.

4.5 Inside ITBoost: A Visual Case Study

To provide a mechanistic understanding of our framework, Figure 4 visualizes the average weight evolution for representative sample types on the Jasmine dataset with 20% noise. The result revealed theoretically consistent trajectories: the easy sample (blue) showed a gently increasing weight as the model solidifies its confidence, while the hard sample (orange) maintained a consistently higher weight, confirming the model correctly focuses on informative boundary points. The most compelling evidence lied with the noisy sample (green). It began with the highest weight due to its large initial error, but as the boosting iterations proceed and its residual history became demonstrably random, its weight entered a steep and continuous decline. This provides direct empirical validation of our core claim: the information-theoretic trust mechanism successfully identifies the noisy sample's chaotic error pattern as untrustworthy and systematically suppresses its influence, preventing it from dominating the learning process. This case study thus provides direct empirical evidence that our trust mechanism successfully differentiates valuable difficulty from misleading noise, validating the core principle behind ITBoost's robustness.

Dataset	Ablation	ACC	F1	AUC	Log Loss	Time
Jasmine	Binary (Ours)	**0.9179**	**0.9185**	0.9415	**0.1814**	**13.81**
	Quantized	0.9149	0.9163	**0.9507**	0.1922	18.10
Bioresponse	Binary (Ours)	**0.9174**	**0.9238**	**0.9567**	**0.1844**	**102.83**
	Quantized	0.9153	0.9199	0.9545	0.1916	132.87

Table 2: Ablation study for the role of residual binarization in IT-Boost under the clean Jasmine and Bioresponse datasets.

5 Conclusion

In this paper, we proposed ITBoost, a framework that redefines sample importance based on the historical trustworthiness of error sources rather than instantaneous residuals. By leveraging the MDL principle to quantify the complexity of residual trajectories, ITBoost effectively distinguishes informative hard examples from chaotic noise, systematically suppressing the influence of corrupted labels. Our theoretical analysis establishes a tighter generalization bound under label noise, while extensive empirical results demonstrate that ITBoost achieves optimal robustness in noisy environments without compromising performance on clean data. We acknowledge that real-time computation of Lempel–Ziv complexity incurs notable overhead, especially on large-scale datasets. Future work will improve scalability via approximate measures, sliding-window evaluation, and periodic trust updates. More broadly, information-theoretic trust opens several promising directions. Leveraging temporal learning dynamics is not limited to boosting: it can be adapted to deep learning for more robust optimizers, applied to curriculum learning to better identify genuinely difficult samples, and extended to online and continual learning to detect and respond to concept drift. By shifting from a static to a dynamic, history-aware notion of sample importance, this work lays the groundwork for a new class of more adaptive and resilient learning algorithms.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62403043 and 42171323.

References

[Bentéjac *et al.*, 2021] Candice Bentéjac, Anna-Mária Csörgő, and Gonzalo Martínez-Muñoz. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3):1937–1967, 2021.

[Chen *et al.*, 2015] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, et al. Xgboost: extreme gradient boosting. R package version 0.4-2, 2015.

[Dal Pozzolo *et al.*, 2015] Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson, and Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. In *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 159–166. IEEE, 2015.

[Duan *et al.*, 2020] Tony Duan, Avati Anand, Daisy Yi Ding, Khanh K. Thai, Sanjay Basu, Andrew Ng, and Alejandro Schuler. Ngboost: Natural gradient boosting for probabilistic prediction. In *International Conference on Machine Learning*, pages 2690–2700. PMLR, 2020.

[Dubout and Fleuret, 2014] Charles Dubout and François Fleuret. Adaptive sampling for large scale boosting. *The Journal of Machine Learning Research*, 15(1):1431–1453, 2014.

[Einziger *et al.*, 2019] Gil Einziger, Maayan Goldstein, Yaniv Sa'ar, and Itai Segall. Verifying robustness of gradient boosted models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2446–2453, 2019.

[Ferreira and Figueiredo, 2012] A. J. Ferreira and M. A. Figueiredo. Boosting algorithms: A review of methods, theory, and applications. In *Ensemble Machine Learning*, pages 35–85. Springer, 2012.

[Freund and Schapire, 1997] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[Friedman, 2001] Jerome H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[Frénay and Verleysen, 2013] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2013.

[Gasieniec *et al.*, 1996] Leszek Gasieniec, Marek Karpinski, Wojciech Plandowski, and Wojciech Rytter. Efficient algorithms for lempel-ziv encoding. In *Scandinavian Workshop on Algorithm Theory*, pages 392–403, Berlin, Heidelberg, 1996. Springer.

[Grünwald, 2007] Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

[Guyon *et al.*, 2007] Isabelle Guyon, Jiwen Li, Theodor Mader, Patrick A. Pletscher, Georg Schneider, and Markus Uhr. Competitive baseline methods set new standards for the NIPS 2003 feature selection benchmark. *Pattern Recognition Letters*, 28(12):1438–1444, 2007.

[Hasanin and Khoshgoftaar, 2018] Taha Hasanin and Taghi Khoshgoftaar. The effects of random undersampling with simulated class imbalance for big data. In *Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI)*, pages 70–79. IEEE, 2018.

[Hollmann *et al.*, 2025] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

[Huber, 1992] Peter J. Huber. Robust estimation of a location parameter. In *Breakthroughs in Statistics: Methodology and Distribution*, pages 492–518. Springer, New York, NY, 1992.

[Ke *et al.*, 2017] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, et al. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[Kolmogorov, 1965] Andrei N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1):1–7, 1965.

[Li and Vitányi, 2008] Ming Li and Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*, volume 3. Springer, New York, 2008.

[Liu and Tsoumakas, 2020] Boyu Liu and Grigorios Tsoumakas. Dealing with class imbalance in classifier chains via random undersampling. *Knowledge-Based Systems*, 192:105292, 2020.

[Liu *et al.*, 2020] X. Liu, S. Luo, and L. Pan. Robust boosting via self-sampling. *Knowledge-Based Systems*, 193:105424, 2020.

[Long and Servedio, 2008] Philip M. Long and Rocco A. Servedio. Random classification noise defeats all convex potential boosters. In *Proceedings of the 25th International Conference on Machine Learning*, pages 608–615, 2008.

[Luo *et al.*, 2025] J. Luo, Y. Quan, and S. Xu. Robust-gbdt: leveraging robust loss for noisy and imbalanced classification with gbdt. *Knowledge and Information Systems*, pages 1–21, 2025.

[Miao *et al.*, 2015] Qiguang Miao, Yang Cao, Ge Xia, Maoguo Gong, Jianfeng Liu, and Jiankai Song. RBoost: Label noise-robust boosting algorithm based on a nonconvex loss function and the numerically stable base learners. *IEEE Transactions on Neural Networks and Learning Systems*, 27(11):2216–2228, 2015.

[Prokhorenkova *et al.*, 2018] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[Rissanen, 1978] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[Shi *et al.*, 2024] Jun Shi, Ke Zhang, Chao Guo, Yi Yang, Yan Xu, and Jinhui Wu. A survey of label-noise deep learning for medical image analysis. *Medical Image Analysis*, 95:103166, 2024.

[Shwartz-Ziv and Armon, 2022] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

[Song *et al.*, 2022] Hwanjun Song, Minseok Kim, Donghyun Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2022.

[Stiglic and Kokol, 2010] Gregor Stiglic and Peter Kokol. Stability of ranked gene lists in large microarray analysis studies. *BioMed Research International*, 2010:616358, 2010.

[Wyner and Ziv, 2002] Aaron D. Wyner and Jacob Ziv. The sliding-window lempel-ziv algorithm is asymptotically optimal. *Proceedings of the IEEE*, 82(6):872–877, 2002.

[Zhang and Sabuncu, 2018] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[Zozor *et al.*, 2005] Steeve Zozor, Philippe Ravier, and Olivier Buttelli. On lempel–ziv complexity for multidimensional data analysis. *Physica A: Statistical Mechanics and its Applications*, 345(1-2):285–302, 2005.