MICRO-LEARNING FOR LEARNING-HARD PROBLEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning systems increasingly face high complexity data whose nonlinear structure, noise, imbalance, or limited sample size thwart conventional models. We formalize this difficulty through the notion of Learning Hard Problems (LH-Ps), tasks that (i) defeat the vast majority of models, yet (ii) admit at least one high-quality solution if the relevant domain knowledge is appropriately incorporated during training. To address LH-Ps we introduce Micro-Learning (MiL), a principled framework that constructs traininglets: small, knowledge-fused subsets of the training data with demonstrably low complexity—and infers a deterministic local model for each that collectively form a global predictor. We prove that the decision version of optimal traininglet selection is NP-complete, establishing a strong theoretical foundation for MiL. MiL dramatically reduces overfitting risk by eliminating irrelevant or noisy samples, while retaining interpretability and reproducibility through deterministic optimization in a Reproducing Kernel Hilbert Space. Experiments in benchmark domains, from music information retrieval to medical proteomics, show that MiL solves LH-Ps successfully and outperforms deep learning and classical baselines, especially on imbalanced or small-sample datasets, with negligible overfitting. Beyond an effective algorithm, our work provides (i) the first formal definition and characterization of LH-Ps, (ii) a Learning-Hard Index (LHI) to quantify task difficulty pre-training, and (iii) theoretical guarantees on traininglet optimality and complexity. Together, these contributions enrich learning theory and offer a path to ethical AI.

1 INTRODUCTION

Modern AI increasingly confronts *high-complexity data* whose non-linearity, noise, imbalance, or scarcity overwhelm conventional models. We argue that many such tasks belong to a distinct, under-studied class we call *Learning-Hard Problems (LH-Ps)*. An LH-P is characterized by two simultaneous properties:

1. *Near-universal failure* – almost every model in a broad hypothesis space delivers mediocre or poor performance;
2. *Latent solvability* – there exists at least one model that can achieve high-quality results once appropriate domain knowledge is fused into the training process, i.e., a good performance certificate exists and is verifiable.

Definition 1 (Learning-Hard Problem (LH-P)). *Let $\mathcal{X}, \mathcal{Y}, \mathcal{P}, \mathbb{H}$ be as above. For each $h \in \mathbb{H}$, let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a loss, and define the generalization risk $R(h) = \mathbb{E}_{(x,y)\sim\mathcal{P}}\big[L\big(h(x), y\big)\big]$. Assume a* family of knowledge-injection operators $\mathcal{K} = \{\varphi_\kappa : \mathcal{X} \to \mathcal{X}\}_\kappa$. *A supervised task is an LH-P with respect to $(\mathbb{H}, \mathcal{K})$ if there exist constants $0 < \tau \ll \tau^\star$ satisfying*

(C1) Near-universal failure: $\min_{h\in\mathbb{H}} R(h) \;\geq\; \tau^\star$,
(C2) Latent solvability: $\exists\, \kappa \in \mathcal{K},\ h^\star \in \mathbb{H}$ *s.t.* $R\big(h^\star \circ \varphi_\kappa\big) \;\leq\; \tau$.

Here $(h^\star \circ \varphi_\kappa)(x) = h^\star\big(\varphi_\kappa(x)\big)$; the operator φ_κ is a fixed, and knowledge-fusion preprocessing map that can be a label-aware projection or a re-sampling operator that corrects distribution shift for both training and test data.

Interpretation. C1 states that *all* vanilla models drawn from $\mathbb{H}$ incur high risk, whereas C2 guarantees the *existence of a verifiable certificate of solvability*: some pair $(h^\star, \kappa)$ achieves low risk once appropriate knowledge is injected.

Importantly, Definition 1 is existential; it does *not* assert that standard training procedures can efficiently discover $(h^\star, \kappa)$. In fact, as we prove in the *supplemental*, the corresponding decision problem: determining if a satisfactory traininglet exists is NP-complete (*Theorem 1 in Section 4*).

LH-Ps are pervasive: polyphonic music tagging (Fuhrmann & Herrera, 2010), speech-emotion recognition on tiny corpora (Haq et al., 2008), protein-mass-spectrometry diagnosis (Han et al., 2023), and heavily imbalanced COVID-19 triage all score high on our Learning-Hard Index (LHI) measure. (*Section 3*).

Figure 1 illustrates the core challenge of LH-Ps using the IRMAS dataset. While a solution path may exist conceptually (a), the raw data appears as an inseparable swirl when visualized with t-SNE (b). Feature selection offers little improvement (c), demonstrating that simple dimensionality reduction is insufficient. However, when label information is fused into the embedding process (d), the classes become clearly distinct. This reveals the central thesis: the problem is not a lack of signal, but the failure of standard methods to leverage domain knowledge, directly motivating our Micro-Learning (MiL) approach.

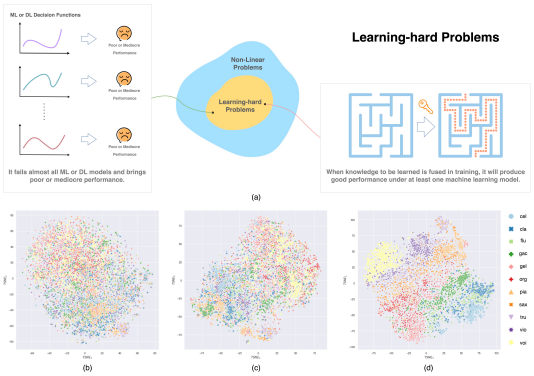

Figure 1: (a) A conceptual diagram of LH-Ps, where a viable solution path (gold) exists but is difficult for standard learners to find. (b-d) t-SNE visualizations of the IRMAS dataset: (b) The raw data shows no clear class structure. (c) Applying feature selection provides minimal separation. (d) Label-aware t-SNE exposes clear class clusters, demonstrating latent solvability.

Why Deep Learning Falters?
Deep networks excel at feature learning, yet their nested nonlinear decision function $G(x) = f_{\text{softmax}}\big(g_L \circ \cdots \circ g_1(x)\big)$ acts like an extremely high-order polynomial. Small input perturbations can thus cause large output swings, producing *over-fitting*, poor reproducibility, and limited explainability (Samek et al., 2017; Li et al., 2019). Although techniques such as sharpness-aware minimisation (Foret et al., 2020) help, they neither expand the hypothesis space nor inject external knowledge, so deep models remain *case-specific* fixes and cannot solve LH-Ps.

For concreteness, consider the IRMAS music-tagging benchmark visualized in Fig.1. Even SOTA architectures struggle: the carefully engineered convolutional network of Han et al. (2017) reaches only 60.2% micro-F1, while the more sophisticated multitask CNN with onset-group auxiliary classification proposed by Yu et al. (2020) climbs to 68.5 %, still far from acceptable for a modern MIR system. These results exemplify the central pathology of LH-Ps: deep models, however refined, search within a fixed hypothesis space and cannot exploit latent domain structure.

Micro-Learning (MiL). We introduce *Micro-Learning* (MiL), a principled framework for LH-Ps. For every query point or small query batch: MiL extracts an instance-specific traininglet: a micro-sized, knowledge-fused subset of the training set D. This per-query data selection echoes the spirit of curriculum learning (Bengio, 2009), but it is performed online and individually rather than via a global easy-to-hard schedule.

On this traininglet we fit a deterministic, explainable learner—e.g., an SVM or variants in a Reproducing Kernel Hilbert Space, yielding a custom prediction function unique to that query. By bypassing the usual global-to-local generalization step, MiL supplies *an over-fitting-resistant, interpretable, and reproducible* predictor; the ensemble of these locals forms the global decision function. Unlike local-SVM ensembles (Aha, 1997; Tappen et al., 2001) or meta-learning kernels (Snell et al., 2017), MiL fuses domain knowledge before model induction, a step we show is critical for LH-Ps.

Contributions. (1) *Theory:* We formalize LH-Ps, introduce the Learning-Hard Index (LHI), and prove the decision problem of optimal traininglet selection is NP-complete (yet tractable via our precision heuristic). (2) *Algorithm:* We present Micro-Learning (MiL), an overfitting-resistant, ex-

plainable, reproducible framework. (3) *Guarantees:* MiL contracts the train–test total-variation distance and reduces local Rademacher complexity (Koltchinskii, 2006). (4) *Empirics:* On music, speech, health, omics, and COVID-19 benchmarks, MiL successfully solves LH-Ps that defeat baselines, with negligible overfitting.

2 Related works

Kernel methods fail on LHPs due to poor scalability ($\mathcal{O}(n^2)$ storage) and their tendency to amplify noise in high-variance data (Vapnik, 2000; Yu et al., 2020).

Deep networks excel at automatic feature extraction but their nested nonlinearities are sensitive to perturbations, causing over-fitting, weak reproducibility and limited interpretability (Samek et al., 2017; Li et al., 2019). Techniques such as sharpness-aware minimization (SAM) (Foret et al., 2020) and refined capacity measures (Galanti et al., 2023; Jiang et al., 2019; Zhang et al., 2021; Ramasinghe et al., 2023) reduce—but do not eliminate—these issues, and they neither enlarge the hypothesis space nor inject task-level knowledge, leaving many LH-Ps unsolved.

Local and test-time learners. Gradient-based meta-learning such as MAML (Finn, 2017), and local model-builders like LIME (Marco, 2016), MAPLE (Gregory, 2018), and T3A (Iwasawa et al., 2021) rely only on observed features, leaving them vulnerable to LH-P failure modes. Similarly, test-time adaptation methods (e.g., Tent (Wang et al., 2021)) fine-tune on target batches but cannot escape the original hypothesis space or fuse the external knowledge required to solve LH-Ps.

3 Diagnosing Learning-Hardness

To efficiently diagnose LH-Ps without relying on their expensive formal definition, we introduce the Learning-Hard Index (LHI), a metric that quantifies a dataset's inherent complexity. This section also establishes the theoretical complexity of solving LH-Ps.

Learning-Hard Index (LHI)

LHI is a scalar in [0, 1] that quantifies the intrinsic complexity of a dataset. Assuming nominally clean labels, we classify a task as an LH-P whenever its training-set $LHI \geq 0.80$). In general, a higher LHI indicates greater learning difficulty and therefore a higher likelihood that the task is learning-hard.

In contrast to measures such as Rademacher complexity, which assess the richness of a model class, LHI is entirely data-centric. Because it can be computed before any training begins, LHI serves as a lightweight, model-agnostic score for comparing datasets and for deciding whether specialized methods, such as our Micro-Learning (MiL) framework introduced later, are warranted.

Learning-Hard Index (LHI). Let $X = \{(x_i, y_i)\}_{i=1}^m$ be a labeled dataset. We first obtain a locality-preserving embedding $X_r = f_{\text{dm}}(X)$ (e.g., tSNE or UMAP van der Maaten (2008); McInnes et al. (2018)), then group X_r with Θ (e.g., k-means), producing pseudolabels $y_{\text{p}i}$ and producing the set $X_p = \{(x_i, y_{\text{p}i})\}_{i=1}^m$.The LHI is

$$\text{LHI}(X) = 1 - \text{AMI}(X_r, X_p) = 1 - \frac{\text{MI}(X_r, X_p) - \mathbb{E}[\text{MI}(X_r, X_p)]}{\frac{1}{2}(H(X_r) + H(X_p)) - \mathbb{E}[\text{MI}(X_r, X_p)]}. \tag{1}$$

where MI denotes mutual information and $H(\cdot)$ is Shannon entropy. Because AMI rewards embeddings that preserve local neighborhoods, it serves as a strong basis for the LHI. Local nonlinear dimension reduction maps such as t-SNE, known for maintaining data locality Han et al. (2022), yield an LHI that faithfully reflects intrinsic task difficulty. In contrast, global linear projections (e.g., PCA) blur minority manifolds and deflate the LHI, so we use locality-preserving f_{dm} instead.

Thresholding $\text{LHI}(X) \geq 0.80$—i.e., when the embedding retains $\leq 20\%$ of neighborhood mutual information—reliably flags learning-hard tasks that demand specialized training (e.g., Micro-Learning) to achieve acceptable accuracy.

LH-P Datasets

We evaluate LHI on five benchmarks spanning music, speech, health, and medicine: IRMAS (Fuhrmann & Herrera, 2010), CASIA (Li et al., 2016), SAVEE (Haq et al., 2008), Ovarian (Han et al., 2023), and a curated COVID-19 triage dataset. Table 1 summarizes key statistics; Although COVID19 falls slightly below the 0.80 threshold, we include it as a quasi-hard control to test MiL's sensitivity to difficulty. More data details can be found in *supplemental G.*

Table 1: Datasets of learning-hard problems

Dataset	(n,p)	Imbalance/rate	Classes	LHI
IRMAS	(6705,518)	N	11	90.7%
CASIA	(1200,54)	N	6	87.3%
SAVEE	(480,54)	N	7	85.4%
COVID-19	(128,48)	Y (57.03%)	3	78.5%
Ovarian	(266,20531)	Y (98.50%)	2	97.6%

3.1 Theoretical Complexity of LH-Ps

The LHI identifies *when* conventional training fails, but not *how* to succeed. Our core insight is that an LH-P can often be solved on a judiciously chosen customized small *subset* of the training set—later called a *traininglet*. We formalize this idea using *local Rademacher complexity* (Bartlett & Mendelson, 2002).

Local Rademacher complexity. For a sample $S = \{z_1, \ldots, z_n\}$ and a function class $\mathcal{F}$, define the radius-r neighborhood $\mathcal{F}_r(f) = \{g \in \mathcal{F} : \|g - f\| \le r\}$. Its local Rademacher complexity is

$$\mathcal{R}_n\big(\mathcal{F}_r(f)\big) \;=\; \mathbb{E}_{S,\sigma}\Big[\sup_{g\in\mathcal{F}_r(f)} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\, g(z_i)\Big], \tag{2}$$

where each σ_i is an independent Rademacher variable. Smaller $\mathcal{R}_n$ implies tighter generalization bounds in the neighborhood of f.

Prop.1 (*proof in supplemental A*) provides the theoretical grounding for our approach, stating that every LH-P contains a 'sweet-spot' model within a region of minimal overfitting risk (i.e., minimal local Rademacher complexity). Our Micro-Learning (MiL) framework is designed to systematically find this low-capacity region.

Proposition 1 (Low-capacity witness). For any LH-P with hypothesis class $\mathbb{H}$ and any radius $r > 0$, there exists a model $f^\star \in \mathbb{H}$ such that $\mathcal{R}_n\big(\mathcal{F}_r(f^\star)\big) \;=\; \inf_{f\in\mathbb{H}} \mathcal{R}_n\big(\mathcal{F}_r(f)\big)$, meaning $f^\star$ minimizes the local Rademacher complexity over $\mathbb{H}$. Consequently, $f^\star$ and every model within its r-ball neighborhood enjoy the tightest generalization bound available in the entire hypothesis space.

Why Prop. 1 matters. Even though $\mathbb{H}$ is inflated by noise and nonlinearity, Prop. 1 guarantees at least one "sweet-spot" region where overfitting risk is minimal. The practical challenge is to reach that region without exhaustively searching $\mathbb{H}$.

Standing on Prop. 1, Prop. 2 (*proof in supplemental B*) shows that for any given test point, a model trained on a suitably crafted traininglet is more likely to match the ideal Bayes prediction than any model trained on the full dataset. This strategy provides a practical path to realizing the low-capacity "sweet spot" guaranteed by Proposition 1.

Proposition 2 (Traininglet sufficiency). For any test point p, there exist a traininglet S_pS and $\Theta_p \in \mathbb{H}$ such that the classifier trained only on this traininglet $h_{\Theta_p,S_p} \in \mathbb{H}$ satisfies

$$\Pr\big[h_{\Theta_p,S_p}(p) = f_{\text{Bayes}}(p)\big] \;>\; \sup_{\Theta\in\mathbb{H}} \Pr\big[h_{\Theta,S}(p) = f_{\text{Bayes}}(p)\big]. \tag{3}$$

Here $h_{\Theta,S'}$ is the model obtained by fitting hypothesis Θ on dataset S', and f_{Bayes} denotes the Bayes-optimal classifier. Hence, isolating the low-capacity traininglet S_p and training locally yields a predictor whose Bayes-matching probability strictly exceeds that of every full-data model—exactly the strategy embodied in our MiL framework.

4 Overfitting-resistant Micro-Learning (MiL)

From LHI Diagnosis to MiL. In LHPs, the training distribution is generally noisy or mismatched with the query distribution, so global empirical risk minimization, including deep networks, tends to underperform, and typically undergo severe overfitting.

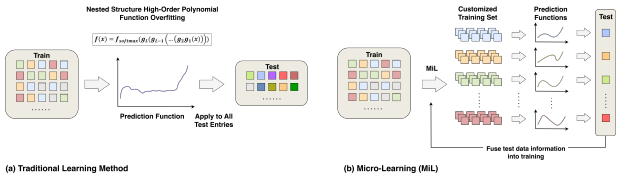

Figure 2: Comparisons of the principles of MiL and traditional ML. MiL produces a local prediction function for each test point or batch, while traditional ML relies on a global prediction function that must generalize across all queries.

Section 3 quantified this difficulty with the LHI and proved that every LH-P contains a provably low-capacity "sweet-spot" region (Propositions 1–2). We now realize those guarantees by introducing Overfitting-Free Micro-Learning (MiL).

Key Idea of MiL. This MiL model rejects the traditional 'one-size-fits-all' paradigm, training a single deeply nested decision function $G(x)$ on all data and hoping it generalizes. For every query point x', MiL constructs on the fly a *traininglet* $\mathcal{T}_{x'}$: a compact, knowledge-fused subset of the training data that sits inside the low-capacity "sweet-spot" guaranteed by Propositions 1–2. Knowledge fusion is realized by intersecting multi-metric neighborhoods, enforcing full-label coverage, merges complementary meta-traininglets, and pruning points tied to noisy "bad" samples to yield a compact, high-quality traininglet for each query (Section 4.2 for details)

MiL then fits a deterministic RKHS model (e.g., SVM or variants) on $\mathcal{T}_{x'}$, called a *learning-let*, producing a query-specific classifier $f_{x'}$. This traininglet-selection + RKHS-fitting pipeline yields an *Overfitting-resistant, reproducible, and interpretable* predictor for each query, converting the existential guarantee of Proposition 2 into an operational algorithm. Unlike traditional models, MiL can generate a tailored decision function for each query point rather than generalizing a whole function to various query points in the testing. Figure 2 compares MiL with traditional ML.

4.1 Traininglet: Definition and Theory

Definition 2 (Traininglet). *Let $X = \{(x_i, y_i)\}_{i=1}^n$ be the labeled training set and $\mathcal{Y} = \{1, \ldots, k\}$ its* label set. *Denote by* $\mathrm{LHI}(\cdot) \in [0, 1]$ *the Learning-Hard Index. For a query point x', the* traininglet

$$\mathcal{T}_{x'} = \arg\min_{T \subseteq X} \big(\mathrm{LHI}(T),\ |T|\big) \quad \textit{s.t.} \quad \mathcal{Y} \subseteq \{y_i :\ x_i \in T\},$$

minimizes the pair $\big(\mathrm{LHI}(T), |T|\big)$ lexicographically, first the lowest LHI, then the smallest size.

Theorems 1 and 2, stated next, establish (i) traininglet decision problem (TRAININGLET-DEC) is NP-complete, implying the NP-hardness of finding the optimal traininglets, and (ii) the guaranteed existence of a low-capacity "sweet-spot" solution for every LH-P; detailed proofs are provided in *supplemental*.

Theorem 1 (TRAININGLET-DEC is NP-complete). *Given a labeled set $X = \{(x_i, y_i)\}_{i=1}^n$, a set of required labels $\mathcal{Y}$, a budget $b \le n$, and an LHI bound $\ell \in [0, 1]$, the problem of deciding*

$$\exists\, T \subseteq X :\ |T| \le b,\ \mathrm{LHI}(T) \le \ell,\ \mathcal{Y} \subseteq \{y_i : x_i \in T\} \qquad \text{(TRAININGLET-DEC)}$$

is NP*-complete, assuming* $\mathrm{LHI}(\cdot)$ *is computable in polynomial time.*

Before we move to Theorem 2 we need one technical fact. A sample is called *σ-noisy* if replacing its label by a fresh dummy label increases AMI by at least $\sigma > 0$. Removing such a point always lowers the overall LHI of a dataset X. Lemma 1 formalizes this monotonicity and is the key step used to build the low-complexity traininglets of Theorem 2.

Lemma 1 (Removable-Noise Monotonicity; proof in *supplemental D*). *Let $Z \subseteq X$ be labeled data and $z = (x_z, y_z) \in Z$. If substituting a fresh dummy label $\perp$ for y_z increases* AMI *by at least $\sigma > 0$, then* $\mathrm{LHI}(Z \setminus \{z\}) \ \le\ \mathrm{LHI}(Z) - \sigma$.

For any LH-P and any batch of test queries we can always pick traininglets whose intrinsic complexity is strictly reduced, guaranteeing a move into the low-capacity regime promised by learning theory.

Theorem 2 (Existence of Low-Complexity Traininglets)**.** *Let X be the training set of any learning-hard problem (LH-P) and let $x'_1, \ldots, x'_s$ be an arbitrary query batch. Then there exist traininglets $\mathcal{T}_{x'_1}, \ldots, \mathcal{T}_{x'_s} \subseteq X$ with $\min_j \mathrm{LHI}\big(\mathcal{T}_{x'_j}\big) < \mathrm{LHI}(X)$. Moreover, if X contains a σ-noisy point in the sense of Lemma 1 (so $\sigma > 0$), the inequality sharpens to $\min_j \mathrm{LHI}\big(\mathcal{T}_{x'_j}\big) \leq \mathrm{LHI}(X) - \sigma$. proof in supplemental E.*

4.2 MiL Traininglet construction

Since finding the lexicographically optimal traininglet is NP-hard, we introduce two practical heuristics: naive traininglet construction (NTC) and precision traininglet construction (PTC). Both heuristics serve the same purpose: to isolate a compact, label-complete subset whose LHI lands in the low-capacity "sweet-spot" promised by Theorem 2.

NTC. NTC, the basis for our "Naive MiL" variant, creates a traininglet by intersecting small metric balls (e.g., Euclidean and correlation) and is effective primarily on large, clean training datasets.

PTC. PTC is our robust heuristic for creating high-quality, low-LHI traininglets, especially for challenging data. The process unfolds in four stages: (1) Probing Learning to find optimal neighborhood radii using a label-aware score; (2) Training Sanitization to prune noisy samples; (3) Meta-Traininglet Fusion to merge four complementary subsets for broad coverage; and (4) Precision Pruning to remove any final outliers. The resulting pipeline, which we call the full MiL framework, consistently isolates distribution-matched traininglets that are effective on datasets ranging from tiny and imbalanced to large-scale corpora.

Naïve traininglet construction (NTC). NTC builds a traininglet for query x'_i by intersecting several small metric balls so that retained points are simultaneously close to x'_i in multiple geometric views of the data. Formally, $\mathcal{T}_{x'_i} = \bigcap_{j=1}^{m} \Big\{ x \in X : \; d_j(x, x'_i) < \varepsilon_j \Big\}, \qquad m \geq 2$, where d_1 and d_2 are typically Euclidean distance and Pearson correlation; a third view such as Wasserstein (images/audio) or cosine distance (sparse text) can be added when beneficial.

Label rebalancing. If the neighborhood $\mathcal{N}_\varepsilon(x'_i)$ lacks any label o, we append the nearest sample of that label: $\mathcal{N}'_\varepsilon(x'_i) = \mathcal{N}_\varepsilon(x'_i) \cup \Big\{ \arg\min_{x \in \mathcal{S}_o} d_j(x, x'_i) \Big\}$.

Limitations. NTC presumes a large, clean dataset; the fixed radii ε_j in equation **??** are rarely optimal, and noise within a ball can raise LHI even after rebalancing via equation **??**. It either remains unknown how to select ε_j for a batch of query points. These issues motivate the more robust Precision Traininglet Construction (PTC) introduced next.

4.2.1 Precision Traininglet Construction (PTC.)

PTC operationalizes the guarantee of Theorem 2, reliably identifying the low-capacity "sweet-spot" for any query batch. Its four-stage pipeline—Probing Learning, Training Sanitization, Meta-traininglet Fusion, and Precision Pruning—is detailed in Algorithm 1 (*Supplemental*).

1. Probing learning. We estimate the optimal neighborhood radius k and batch size z (queries processed jointly) by a Monte-Carlo search: over $M{=}5 - 30$ random 80/20 splits of training data X. We evaluate every (k, z) on Naive-MiL to maximize a target *D-index* Han et al. (2023), finally, we average the resulting D-index, and retain only non-dominated pairs.

Specifically, we random-split X into an 80% *train-train* subset X_{tr} and a 20% *train-test* subset X_{te}. Across a bounded grid of (k, z) pairs, Naive-MiL predicts the labels of X_{va} from X_{tr}. We select the pair $(k^\star, z^\star)$ that maximizes D (D-index) in each search. $(k^\star, z^\star) = \arg\max_{k,z} D_{\text{Naive-MiL}}\big(X_{\mathrm{tr}}, X_{\mathrm{te}}, k, z\big)$. The D-index, an interpretable ML assessment score bounded by $(0, 2]$—is defined for a K-class problem as $D = \frac{1}{K} \sum_{i=1}^{K} \Big[\log_2\big(1 + \alpha_i\big) + \log_2\Big(1 + \frac{s_i + p_i}{2}\Big) \Big]$, where α_i, s_i, and p_i denote the accuracy, sensitivity, and specificity per class, respectively.

By maximizing the D-index—a measure of class discrimination—this probing step replaces the heuristic radii of Naive Traininglet Construction (NTC) with a data-driven calibration tuned for maximum predictive utility.

2. Training sanitisation. Running Naive-MiL with $(k^\star, z^\star)$ on training data X yields a deterministic prediction $\widehat{y}_i$ for every sample (x_i, y_i). This partitions X into correctly and incorrectly predicted subsets ("good guys" and "bad guys"):

$$\mathcal{G} = \{x_i \in X \mid \widehat{y}_i = y_i\}, \qquad \mathcal{B} = \{x_i \in X \mid \widehat{y}_i \neq y_i\}. \tag{4}$$

Noise pruning. For each $x_b \in \mathcal{B}$ we remove both the error point and its ϵ-ball neighbors $\mathcal{N}_\epsilon(x_b)$: $X^{\text{clean}} = X \setminus \Big(\mathcal{B} \cup \bigcup_{x_b \in \mathcal{B}} \mathcal{N}_\epsilon(x_b)\Big)$. By Lemma 1, deleting each ϵ-ball lowers the LHI by at least $\sigma > 0$; hence $\text{LHI}\big(X^{\text{clean}}\big) \leq \text{LHI}(X) - \sigma$, moving the data toward the low-capacity "sweet-spot" required by Theorem 2. The sanitization process prunes 18–41% of the training data across our five benchmarks, reducing the LHI by 6–20%. To prevent data loss for rare classes, a *minority-class safeguard* re-introduces the nearest 'good' instance ($\mathcal{G}$) for any class that is fully eliminated. This yields a lean, noise-free, and label-complete dataset for the subsequent PTC steps.

3. Meta-traininglet fusion. For every query x_i' we fuse four meta-trainginlets $\mathcal{T}_{x_i'}^{(j)}, j = 1, 2, 3, 4$ into a single, label-complete union: $\mathcal{U}_{x_i'} = \bigcup_{j=1}^{4} \mathcal{T}_{x_i'}^{(j)}$. This union (i) contains every class, (ii) is at most $3k^{'} + |\mathcal{G}|$ points, and (iii) lowers LHI, and provide a compact, well-balanced basis for PTC.

The 1^{st} meta-traininglet $\mathcal{T}_{x_i'}^{(1)}$ is *a local ball* capturing geometric proximity. It is created using NTC with the optimal neighbor size k' in the cleaned training data: $\mathcal{T}_{x_i'}^{(1)} = \text{NTC}\big(x_i', k', X^{\text{clean}}\big)$.

The 2^{nd} and 3^{rd} meta-trainginlets, $\mathcal{T}_{x_i'}^{(2)}$ and $\mathcal{T}_{x_i'}^{(3)}$, are *1-hop and 2-hop transfers*, injecting first-order semantic context and adding broader manifold structure, respectively. They are generated by performing nearest-neighbor search (NNS) on $\mathcal{G}$ (the set of "good guys" from training sanitization) to obtain each point's first- and second-closest neighbors, $\mathcal{N}_1(x_i')$ and $\mathcal{N}_2(x_i')$, and then merging their trainginlets:

$$\mathcal{T}_{x_i'}^{(2)} = \bigcup_{x_i' \in \mathcal{G}} \mathcal{T}_{\mathcal{N}_1(x_i)}, \qquad \mathcal{T}_{x_i'}^{(3)} = \bigcup_{x_i' \in \mathcal{G}} \mathcal{T}_{\mathcal{N}_2(x_i')} \tag{5}$$

The 4^{th} meta-traininglet $\mathcal{T}_{x_i'}^{(4)}$ is *a random anchor* plugging residual topology gaps. It is formed by randomly selecting a "good guy" $x_g \in \mathcal{G}$ and combining it with its traininglet, $\mathcal{T}_{x_i'}^{(4)} = \mathcal{T}_{x_g}$.

4. Precision pruning. Remove any point within a neighbor radius of B to obtain the *precision PTC* $\mathcal{T}_{x_i'}^{\text{PTC}} = \mathcal{U}_{x_i'} \setminus \bigcup_{b \in B} \mathcal{N}(b)$. This last cut shrinks LHI by removing additional noise or outliers.

Why PTC works. *Stage 1* aligns neighborhoods with labels; by Prop. 1, it lands in an r-ball of minimal local Rademacher radius, and Prop. 2 guarantees that the resulting traininglet outperforms any fulldata model, tightening the generalization bound; *Stage 2* excises high-entropy samples and their neighbors, lowering the empirical VC dimension; *Stage 3* re-establishes full label coverage, guaranteeing a non-empty feasible ball (Thm. 2); *Stage 4* removes residual outlier anchors, tightening the generalization bound to $\mathcal{O}\big(1/\sqrt{|T|}\big)$, where T is the final traininglet size.

Complexity. Let $n{=}|X|$, p features, and $k^\star \ll n$. Probing learning requires $\mathcal{O}(Mn^2)$ distance evaluations for M Monte-Carlo draws ($M \leq 30$). After training-sanitization, each meta-traininglet is at most $k^\star{+}|G|$ points. Beyond being a mere heuristic, PTC is theoretically grounded: Proposition 3 (*proof in suppl. F*) proves that it strictly reduces the total-variation distance between the training and test distributions. This provides the statistical alignment essential for reliable generalization on LH-Ps.

Proposition 3 (PTC contracts the training–test gap). Let P_{tr} and P_{te} denote the training and test distributions of an LH-P. After applying PTC within the MiL pipeline, the resulting distribution P_{PTC} satisfies the strict total variation contraction, i.e., $P_{\text{tr}} \xrightarrow{\text{PTC}} P_{\text{te}}$. $\big\|P_{\text{PTC}} - P_{\text{te}}\big\|_{\text{TV}} < \big\|P_{\text{tr}} - P_{\text{te}}\big\|_{\text{TV}}$.

Explainability and reproducibility of MiL MiL's local learner (learning-let) is mainly a multiclass SVM trained on each traininglet (Vapnik, 2000). This choice ensures high reproducibility due to SVM's deterministic convex optimization and is well-suited for the small sample size of trainginlets. For multiclassification, we use a One-vs-One (OvO) approach, which is robust

Table 2: Performance of MiL on five benchmarks

Dataset	Dindex	Acc	Sen	Prec	F1
IRMAS	1.8162	0.8431	0.8387	0.8449	0.8431
CASIA	1.7949	0.8283	0.8314	0.8297	0.8283
SAVEE	1.7015	0.7625	0.7365	0.7458	0.7625
COVID19	1.9424	0.9544	0.9644	0.9632	0.9544
Ovarian	1.7939	0.9815	1.0000	0.9811	0.9815

to class imbalance. The final decision is a majority vote over pairwise classifiers of the form: $f_{ik}(x) = \sum_j \alpha_j^{ik} y_j^{ik} K(x_j^{ik}, x) + b_{ik}$. This approach makes the influence of the traininglet's support vectors directly interpretable.

SVM–micro–CNN–let. We endow MiL with translation-aware representation learning—yet keep the determinism of large-margin theory—by replacing every SVM "node" with an *SVM–micro–CNN–let*: a compact stack of 3×3 convolutions, batch normalisation, and ReLU layers followed by global average pooling that produces a learned feature map $\phi_{\boldsymbol{\theta}} \colon \mathbb{R}^{d\times d} \to \mathbb{R}^m$. A linear SVM head, trained with the multiclass hinge objective, then yields the decision surface $f_c(x) = \mathbf{w}_c^\top \phi_{\boldsymbol{\theta}}(x) + b_c$; the parameters $(\mathbf{w}_c, b_c)$ are updated jointly with $\boldsymbol{\theta}$ via stochastic gradient descent. Because the optimisation is convex in $(\mathbf{w}, b)$ for fixed $\boldsymbol{\theta}$, SVM–micro–CNN–lets retain the reproducibility guarantees and RKHS interpretability of classical SVMs while gaining the expressive, translation-equivariant power of CNNs—disentangling complex local patterns even from small traininglets and advancing explainable, robust learning.

MiL Complexity. MiL's complexity model offers a practical alternative to prohibitive $O(n^3)$ kernel SVMs. It incurs a significant, one-time $O(Mn^2)$ preprocessing cost for PTC. At inference, a shared *SVM–micro–CNN–let* extracts features, allowing a local linear SVM to be solved efficiently ($O(m^3)$) on a constant-size traininglet m. This amortizes to a total inference time of $O(n)$ that is embarrassingly parallel. Memory complexity is similarly reduced from $O(n^2)$ for a full Gram matrix to $O(np)$ for the features. This two-phase design trades a significant, one-time $O(n^2)$ preprocessing cost for highly efficient $O(n)$ inference, making MiL a practical framework for solving large-scale LH-Ps where traditional $O(n^3)$ methods are infeasible.

5 Results: Mastering LH-Ps

Benchmarks and Baselines. We evaluate MiL's performance across the five benchmark datasets summarized in Table 1. These datasets, spanning domains from music to medical proteomics, were specifically chosen as representative LH-Ps, a fact corroborated by their high LHI scores. The COVID-19 dataset, with an LHI of 78.5%, falls just below our 80% threshold. However, we include it as a crucial case study: its extremely small sample size (n=128) presents a different but equally potent learning challenge that thwarts standard models. It therefore serves as an important test of MiL's robustness on sample-starved LH-Ps.

MiL is compared with *15 baselines* chosen to cover the three dominant paradigms for small or noisy data: *(i) Classical non-parametrics* — SVM, Random Forest, Extra-Trees, Naïve Bayes, 2-layer DNN; *(ii) Mainstream static DL* — CNN, LSTM, GRU, Bi-LSTM, Bi-GRU; *(iii) Hybrid/capsule refinements* — Conv-LSTM, Conv-GRU, Conv-BiLSTM, Conv-BiGRU, CapsNet (Vapnik, 2000; Geurts et al., 2006). (LeCun et al., 2015; Sabour et al., 2017; Cho et al., 2014; Hochreiter & Schmidhuber, 1997).. Online TTA methods such as Tent and T3A are omitted: they assume large, stationary target batches and fixed feature extractors, assumptions that fail in LH-Ps where queries are single and highly shifted.

Hyper-parameters are tuned by nested grid search. We report mean over five repeated 5-fold CV runs (IRMAS, CASIA, COVID-19, Ovarian) and a single 10-fold CV (SAVEE), following established practice on small-sample speech corpora.

Performance and statistical validation. Table 2 reports MiL's D-index, accuracy, sensitivity, precision, and F1 across five benchmarks. MiL surpasses both classical ML and advanced DL baselines on every dataset. A one-tailed U-test on the *Metric-integrated Lift* (MiL, the mean of 7 classification measures) shows $\mathrm{Med}(\mathrm{MiL}_{\text{ours}}) = 0.97\,[0.95, 1.00]$, $\mathrm{Med}(\mathrm{MiL}_{\text{bestDL}}) =$

$0.77\,[0.76, 0.84]$, $U = 23$, $p = 1.6 \times 10^{-2}$, $\delta_{\text{Cliff}} = 0.84$. Repeating the test over all 35 raw metric values yields $U = 1082$, $p < 2 \times 10^{-8}$, $\delta_{\text{Cliff}} \approx 0.77\,(P(\text{ours} > \text{DL}) \approx 0.89)$. MiL therefore outperforms every convolutional, recurrent, and capsule DL models, providing concise, effect-size-centered evidence of its architectural superiority.

Similarly, a battery of 25 Bonferroni-adjusted Mann-Whitney tests Dunn (1961) (α=0.01) establishes complete stochastic dominance of MiL over every general ML baselines on all five datasets. Even under extreme imbalance (Ovarian), where rivals lose specificity, MiL maintains a solid decision boundary, confirming its knowledge-fused traininglets dominate LH-P landscapes and deliver superior results statistically (*see suppl. J*).

PTC ablation studies also strongly support the key design choices of its four-stage pipeline—probing-learning, sanitisation, meta-fusion, and precision pruning—showing that each stage contributes a statistically significant, complementary slice of the overall performance gain (*see suppl. I*).

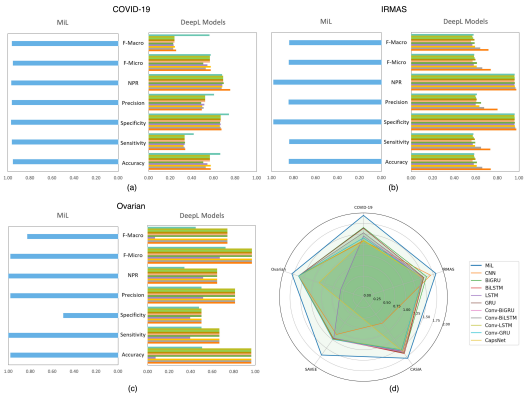

Figure 3: Comparison of MiL with eleven DL models using traditional metrics **(a–c)** on COVID-19, IRMAS, and Ovarian datasets, and d-index values **(d)** across all five datasets.

Fig. 3 contrasts MiL with 11 DL baselines on three benchmarks (*suppl. H*). Even in its naïve form on IRMAS and Ovarian, MiL tops every DL model—and the gap widens on small-sample tasks like COVID-19 and Ovarian, where deep nets falter. Beyond raw accuracy, MiL adds what the DL stack cannot: deterministic training, transparent decision boundaries, and resistance to overfitting.

Traininglet visualization. Figure 4 illustrates the traininglets of the COVID-19, IRMAS, and Ovarian datasets for a single entry and batch. It's interesting to note that the high-quality, customized traininglets curated for each query demonstrates exceptional separability, also validated by their small LHI values. A quick k-means check confirms the drop in LHI from 0.79 to 0.26 (down to 0.05 for the Ovarian case), illustrating how MiL engineers a far simpler sub-distribution around every query. The traininglets for the other data in the supplemental.

6 Discussion and Conclusion

We formalized LH-Ps - tasks that defeat most learners yet become solvable once domain knowledge is fused, and addressed them with MiL. MiL builds tiny, label-complete *traininglets* guided by a data-centric LHI; optimal selection is NP-complete, and our analysis supplies capacity and distribution-shift guarantees that ground its instance-specific SVM–micro–CNN architecture. Empirically, MiL outperforms 15 classical and deep baselines on five demanding benchmarks, rescuing specificity where rivals collapse while remaining fully reproducible and interpretable. MiL's naive runtime is $\mathcal{O}(n^3 + n^2 p)$—costly for large n, but FPGA/GPU acceleration can offset this and speed PTC without losing SVM determinism. Severe noise or imbalance may call for pre-denoising or resampling, yet MiL still provides a principled route to robust AI on scarce, skewed data and enables hardware-efficient, noise-aware extensions. All code and data: `https://anonymous.4open.science/r/iclr26-anon-code-9DB6/`.

Figure 4: Traininglet visualization of MiL on COVID-19 (a, b), IRMAS (c, d), and Ovarian (e, f). (a), (c), and (e) show a single test entry; (b), (d), and (f) show a test batch.

REFERENCES

David W. Aha. Editorial: Special issue on lazy learning. *Artificial Intelligence Review*, 11(1–5): 7–10, 1997. doi: 10.1023/A:1006665904461.

P. Bartlett and S Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002. URL http://www.jmlr.org/papers/v3/bartlett02a.html.

et. al Bengio. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pp. 41–48. ACM, 2009. URL https://dl.acm.org/doi/10.1145/1553374.1553380.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation, 2014. URL http://arxiv.org/abs/1406.1078.

Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961. doi: 10.1080/01621459.1961.10482090.

et al. Finn. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135. PMLR, 2017. URL https://proceedings.mlr.press/v70/finn17a.html.

Pierre Foret, Ariel Kleiner, H. Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. 2020. URL https://www.semanticscholar.org/paper/Sharpness-Aware-Minimization-for-Efficiently-Foret-Kleiner/a2cd073b57be74453315202989228cb4122270a.

Ferdinand Fuhrmann and P. Herrera. Polyphonic instrument recognition for exploring semantic similarities in music. 2010. URL https://www.semanticscholar.org/paper/Polyphonic-Instrument-Recognition-for-exploring-in-Fuhrmann-Herrera/d38cb0117b04c5cc0fee686bf1acd21b9e5239f0.

Tomer Galanti, Mengjia Xu, Liane Galanti, and Tomaso Poggio. Norm-based generalization bounds for compositionally sparse neural networks, 2023. URL http://arxiv.org/abs/2301.12033.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. 63:3–42, 2006. doi: 10.1007/s10994-006-6226-1.

et al. Gregory. Model agnostic supervised local explanations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 2520–2529, 2018.

Henry Han, Wentian Li, Jiacun Wang, Guoxiong Qin, and Xia Qin. Enhance explainability of manifold learning. *Neurocomputing*, 500:877–895, 2022. doi: 10.1016/j.neucom.2022.05.119.

Henry Han, Yi Wu, Jiacun Wang, and Ashley Han. Interpretable machine learning assessment. 561, 2023. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.126891. URL https://doi.org/10.1016/j.neucom.2023.126891.

Yoonchang Han, Jaehun Kim, Kyogu Lee, Yoonchang Han, Jaehun Kim, and Kyogu Lee. Deep convolutional neural networks for predominant instrument recognition in polyphonic music. 25 (1):208–221, 2017. ISSN 2329-9290. doi: 10.1109/TASLP.2016.2632307. URL https://doi.org/10.1109/TASLP.2016.2632307.

Sana Haq, Philip Jackson, and J. Edge. Audio-visual feature selection and reduction for emotion classification. pp. 185–190, 2008.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. 9(8):1735–1780, 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Yusuke Iwasawa, Daiki Kimura, Yuto Yamada, and Shinichi Nakajima. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pp. 4499–4510, 2021. URL https://proceedings.neurips.cc/paper/2021/file/b026390ae2f861b1599237b40af16b3a-Paper.pdf.

Yiding Jiang, Behnam Neyshabur, H. Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. 2019. URL https://www.semanticscholar.org/paper/Fantastic-Generalization-Measures-and-Where-to-Find-Jiang-Neyshabur/8f18c9da3d1763723c6ef8c3734d74db005d0cff.

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6):2593–2656, 2006.

Yann LeCun, Y. Bengio, and Geoffrey Hinton. Deep learning. 521:436–44, 2015. doi: 10.1038/nature14539.

Haidong Li, Jiongcheng Li, Xiaoming Guan, Binghao Liang, Yuting Lai, and Xinglong Luo. Research on overfitting of deep learning. In *2019 15th International Conference on Computational Intelligence and Security (CIS)*, pp. 78–81, 2019. doi: 10.1109/CIS.2019.00025. URL https://ieeexplore.ieee.org/abstract/document/9023664.

Yong-Feng Li, Su-Yuan Zhao, Bo-Hao Wang, Guan-Ying Liu, and Shan Yu. Casia natural emotional audio-visual database. In *2016 7th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–6. IEEE, 2016.

et al. Marco. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1135–1144. ACM, 2016.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. In *Proceedings of the ICML Workshop on Dimension Reduction*, pp. 1–9, 2018. arXiv:1802.03426.

Sameera Ramasinghe, Lachlan Ewen Macdonald, Moshiur Farazi, Hemanth Saratchandran, and Simon Lucey. How much does initialization affect generalization? In *Proceedings of the 40th International Conference on Machine Learning*, pp. 28637–28655. PMLR, 2023.

Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules, 2017. URL http://arxiv.org/abs/1710.09829.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, 2017. URL https://arxiv.org/abs/1708.08296v1.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4077–4087, 2017. URL https://proceedings.neurips.cc/paper/2017/file/cb8da6767469bfa3f30688b5e23dcae5-Paper.pdf.

Michael F. Tappen, Edward H. Adelson, and William T. Freeman. Estimating intrinsic component images using non-linear regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1992–1999, 2001. doi: 10.1109/CVPR.2001.990594.

G van der Maaten, Hinton. Visualizing data using t-sne. *JMLR*, 9:2579–2605, 2008. doi: 10.1145/1143844.1143878.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000. ISBN 978-1-4419-3160-3 978-1-4757-3264-1. doi: 10.1007/978-1-4757-3264-1. URL http://link.springer.com/10.1007/978-1-4757-3264-1.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=uXl3bK3IiI.

Dongyan Yu, Huiping Duan, Jun Fang, and Bing Zeng. Predominant instrument recognition based on deep neural network with auxiliary classification. 28:852–861, 2020. ISSN 2329-9290, 2329-9304. doi: 10.1109/TASLP.2020.2971419. URL https://ieeexplore.ieee.org/document/8979336/.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. 64(3):107–115, 2021. ISSN 0001-0782, 1557-7317. doi: 10.1145/3446776. URL https://dl.acm.org/doi/10.1145/3446776.