POLITECNICO DI TORINO
Repository ISTITUZIONALE

Data-Efficient Cross-Generation Adaptation for Microcontroller Performance Screening

Original
Data-Efficient Cross-Generation Adaptation for Microcontroller Performance Screening / Bellarmino, N., Vancini, A., Kilian, T., Cantoro, R.. - (In corso di stampa). (39th IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems Rome (ITA) September 30th - October 2nd, 2026).

Availability:
This version is available at: 11583/3016155 since: 2026-09-30T11:26:38Z

Publisher:
IEEE

Published
DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©9999 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

01 October 2026

Data-Efficient Cross-Generation Adaptation for Microcontroller Performance Screening

Nicolò Bellarmino*, Alessandro Vancini*, Tobias Kilian†, Riccardo Cantoro*,

*Politecnico di Torino Torino, Italy
†Infineon Technologies AG Munich, Germany

***Abstract*—Microcontroller (MCU) performance screening requires accurate estimation of the maximum operating frequency ($F_{\max}$) under worst-case voltage and temperature conditions, but early production stages provide only a limited number of labeled characterization samples. Historical measurements from previous MCU generations are often available, yet their reuse is hindered by cross-generation shifts in Speed Monitor (SMON) features and target-frequency distributions.**

This paper presents an empirical study of dataset-level adaptation protocols for low-data $F_{\max}$ prediction across related MCU generations. Two incremental training schemes are compared: *mixed incremental training*, where legacy and target samples are interleaved during model updates, and *baseline-first adaptation*, where legacy datasets initialize the predictor and target samples are progressively introduced. Per-product normalization and target-prioritized sample weighting are used to improve cross-dataset compatibility.

Experiments on a target MCU product with three legacy industrial datasets show that naive mixed training is unstable in the low-label regime, whereas baseline-first adaptation yields smoother convergence and improved target-domain sample efficiency. Results further indicate that, in this setting, training protocol and source-target data handling have a larger impact than increased model complexity.

***Index Terms*—Fmax, Speed Monitors, Ring Oscillators, Speed Binning, Machine Learning, Device Testing, Manufacturing, Transfer Learning, Linear Models**

I. Introduction

Performance screening is a critical stage in the manufacturing flow of automotive-grade microcontrollers (MCUs), where devices must satisfy worst-case operating frequency specifications under extreme voltage and temperature conditions. The characterization of the maximum operating frequency ($F_{\max}$) relies on time-consuming functional tests performed on a limited subset of devices, creating a severe labeling bottleneck during early production stages. Reducing characterization effort while preserving screening accuracy is therefore essential for both test-cost containment and time-to-market reduction.

Recent studies have shown that machine learning (ML) models can predict $F_{\max}$ from indirect on-chip measurements, such as Speed Monitors (SMONs), enabling alternate-test strategies for MCU performance screening [1], [2], [3], [4], [5], [6]. Prior work has shown that lightweight regression pipelines can achieve high accuracy with limited labeled data [7], and that active learning and semi-supervised approaches can further improve sample efficiency [8]. However, these studies have largely focused on single-product or single-generation settings.

In industrial practice, newly introduced MCU generations are typically characterized by scarce labeled data, whereas substantial historical measurements from previous generations are often available. Reusing such legacy data is not straightforward, since cross-generation shifts affect both SMON feature distributions and target $F_{\max}$ values. Under these conditions, naive data pooling may lead to unstable adaptation or negative transfer.

This paper investigates dataset-level adaptation protocols for low-data MCU performance screening under cross-generation distribution shift. The analysis is performed on a fixed and well-established lightweight pipeline and focuses on how legacy and target data should be incorporated during training, rather than on modifying the existing predictive model itself.

In particular, two incremental training schemes are compared: (i) *mixed incremental training*, in which legacy and target samples are interleaved during model updates, and (ii) *baseline-first adaptation*, in which legacy datasets are first used to initialize a source model and target samples are then progressively introduced.

The main contributions of this paper are as follows:

- An empirical evaluation of dataset-level adaptation protocols for cross-generation MCU performance screening in the low-label regime.
- A comparison between naive cross-generation mixing and baseline-first adaptation, showing that structured legacy-data reuse improves target-domain sample efficiency.
- An assessment of lightweight linear pipelines under cross-generation data shift, highlighting the importance of training protocol and source-target data handling relative to increased model complexity.

II. Background and Related Work

The prediction of circuit performance parameters from indirect measurements has a long tradition in *alternate test* methodologies [9], [10], [11]. More broadly, machine learning has become increasingly relevant in semiconductor test and reliability analysis, where predictive models are used to extract information from low-cost measurements and manufacturing data in order to reduce test cost while preserving product quality [12], [13]. In the digital domain, machine learning

techniques have been successfully applied to MCUs performance screening.

MCU performance screening aims to verify compliance with worst-case operating-frequency specifications under extreme voltage and temperature conditions. The target value considered in this work is the maximum operating frequency,

$$F_{\max} = \min_{p \in \mathcal{P}} F_p,$$

where $\mathcal{P}$ denotes the set of functional patterns and F_p the maximum frequency sustained under pattern p [1], [14]. The acquisition of $F_{\max}$ labels requires time-intensive board-level functional testing and is therefore limited to a relatively small subset of devices during early production stages [1], [8].

Recent works have used SMONs, i.e., on-chip ring oscillators whose oscillation frequencies correlate with device speed, as low-cost predictors for alternate-test-based screening [11], [15], [16].

Previous works have shown that supervised models can predict $F_{\max}$ with high accuracy from SMON-derived features [7], [8], [17], [18], [19], with lightweight linear pipelines—occasionally combined with polynomial feature transformations—emerging as the strongest performing approach in this application domain.

SMONs are distributed across the die to capture process variability, including Within-Die effects [20], and provide a structured tabular feature set with low acquisition overhead.

Prior studies have also investigated data-efficiency improvements through active learning and semi-supervised strategies [3], [21], but these works remain largely focused on single-generation settings or highly-parametrized neural networks. Cross-domain prediction under limited target labels has also been studied in adjacent areas such as intelligent fault diagnosis and analog-circuit diagnosis, where transfer learning and domain adaptation are used to mitigate source-target mismatch [22], [23]. In MCU performance screening, preliminary evidence on inter-product transfer has been reported in [24], but the effect of training protocol and source-target discrepancy remains insufficiently characterized.

The contribution is therefore placed at the dataset level: the predictive pipeline is kept fixed, while the study investigates how legacy and target data should be structured and incorporated during training under cross-generation shift.

III. Methodology

Let $\mathcal{D}^{(A)} = \{(\mathbf{x}_i^{(A)}, y_i^{(A)})\}_{i=1}^{N_A}$ denote the target dataset and $\{\mathcal{D}^{(k)}\}_{k=1}^{K}$ the legacy datasets. Each sample consists of SMON measurements $\mathbf{x} \in \mathbb{R}^d$ and target value $y = F_{\max}$. The objective is to learn a predictor $f : \mathbb{R}^d \to \mathbb{R}$ that minimizes the prediction error on $\mathcal{D}^{(A)}$ under limited target-label availability. In the considered setting, source and target distributions generally differ, i.e., $p_k(\mathbf{x}, y) \neq p_A(\mathbf{x}, y)$, due to cross-generation shifts in both SMON features and target-frequency distributions. As shown in Fig. 1, global scaling reveals substantial geometric separation between product generations. Naive cross-generation pooling may therefore be suboptimal.

To address this issue, the study adopts a dataset-level adaptation protocol based on two components:

1) **Per-product normalization**, used to mitigate scale and offset differences across datasets while preserving within-product structure;
2) **Controlled target adaptation**, in which legacy data are incorporated according to a predefined incremental training scheme and progressively complemented with labeled target samples.

Two incremental training schemes are considered. In *mixed incremental training* (Strategy 1), legacy and target samples are interleaved during the incremental process, with training set update

$$\mathcal{T}_m = \mathcal{T}_{m-1} \cup \{(\mathbf{x}_j, y_j)\}.$$

In *baseline-first adaptation* (Strategy 2), a source baseline is first constructed from the legacy datasets,

$$\mathcal{T}_0 = \bigcup_{k=1}^{K} \mathcal{D}^{(k)},$$

and target samples are progressively added,

$$\mathcal{T}_m = \mathcal{T}_0 \cup \{(\mathbf{x}_i^{(A)}, y_i^{(A)})\}_{i=1}^{m}.$$

The regression pipeline is re-estimated at each step on the expanded training set. Strategy 1 directly mixes source and target data throughout training, whereas Strategy 2 uses legacy data to initialize the fit and target samples to progressively adapt the predictor toward the target generation.

To mitigate cross-generation scale differences, features and labels are normalized independently within each dataset:

$$\tilde{\mathbf{x}} = \frac{\mathbf{x} - \boldsymbol{\mu}^{(k)}}{\boldsymbol{\sigma}^{(k)}},$$

where $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\sigma}^{(k)}$ are the feature-wise mean and standard deviation estimated on the corresponding product-specific training data. In multi-dataset experiments, only the SMON features common to the considered source-target combination are retained, and a product-identity encoding $\mathbf{e}_k$ is appended to the normalized feature vector to capture residual product-specific offsets. To bias the fit toward the target dataset, target samples receive weight $w_i = 1$, whereas legacy samples receive weight $w_i = \alpha$, with $0 < \alpha \ll 1$. Using Ridge regression, the parameter vector is estimated as

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{\beta}} \sum_i w_i \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda \|\boldsymbol{\beta}\|_2^2.$$

Under this formulation, legacy samples contribute to the global fit and help stabilize the regression in the low-label regime, while target samples dominate the adaptation as their number increases.

To reduce target-label acquisition effort, a diversity-based active learning rule is used to select target samples from the unlabeled pool:

$$\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{U}_m} \min_{\mathbf{z} \in \mathcal{S}_m} \|\tilde{\mathbf{x}} - \tilde{\mathbf{z}}\|_2,$$

where $\mathcal{S}_m$ and $\mathcal{U}_m$ denote the labeled and unlabeled target sets, respectively. Active learning is applied uniformly to all experiments, including the single-generation baseline, so that comparisons across adaptation strategies are not biased by differences in target-sample acquisition policy.

IV. Experimental Setup

A. Datasets and Protocol

Four real-world industrial datasets are considered. Dataset A is the target MCU generation, whereas datasets B, C, and D are legacy generations. The four datasets correspond to related MCU products measured with comparable screening flows and shared SMON semantics, although they differ in design and technological conditions. This setting reflects a realistic industrial scenario in which feature commonality is preserved only on a restricted shared subset. In particular, dataset A comprises approximately 130 SMON features, whereas the legacy datasets contain between 130 and 650 features. Each sample consists of SMON measurements and target value $F_{\max}$, defined as the minimum operating frequency observed across the considered functional patterns.

After data cleaning, including the removal of incomplete measurements and underrepresented wafers, the final dataset sizes are: A (412 samples), B (1642), C (1244), and D (1030). In multi-dataset experiments, only the features common to the considered source-target combination are retained. When A is the target domain, the shared feature space is therefore defined by the SMON subset available in A.

For each experiment, dataset A is split into training and test sets using an 80/20 partition stratified at wafer level. All reported metrics are computed on the test portion of A, so the evaluation quantifies adaptation performance on a single target generation. Results are averaged over five random seeds.

The reference predictive pipeline consists of Ridge regression with ℓ_2 regularization and a PCA transformation with 32 retained components. Additional models are evaluated in selected experiments, including Polynomial Ridge regression [8], Random Forest regression [25], and TabPFN [17]. Hyperparameters are selected through grid-based validation (5-fold grid search). These models are reported only in selected configurations to contextualize the behavior of the reference linear Ridge-based adaptation pipeline.

B. Evaluation Metrics

Learning curves are used to assess data efficiency. Let m denote the number of labeled target samples from dataset A used during training. For each value of m, the model is retrained and evaluated on the fixed test split of A. Under Strategy 2, all selected legacy samples are included from the beginning, while m target samples are progressively added.

We evaluated the first training size at which stable attainment is achieved. We define the stability point $m^\star$ as the first training size after which the learning curve remains at or below the final single-generation baseline reference. This criterion is intended to capture practical early-deployment readiness.

Performance is primarily measured using the normalized Root Mean Squared Error (nRMSE) [26],

$$\mathrm{nRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}}{\bar{y}},$$

where $\bar{y}$ is the mean target value on the test set. Mean Absolute Percentage Error (MAPE) and coefficient of determination (R^2) are also reported at selected operating points. Since the baseline PCA pipeline requires at least 32 labeled target samples to estimate the retained principal components reliably, area-under-curve summaries are interpreted only over comparable training ranges.

The value of α (described in Section III) was set to 10^{-5} based on repeated preliminary experiments on the full training setup, using a candidate set of $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. The selected value provided the most stable trade-off between effective target adaptation and regularizing contribution from legacy samples.

To quantify cross-generation discrepancy, several complementary measures are computed on the shared feature space before and after per-product normalization. First, the silhouette score is used to evaluate whether samples form product-specific clusters. Second, domain separability is assessed by training a logistic classifier to discriminate between product domains and summarizing its performance through

$$\mathrm{AUROC}_{\mathrm{dist}} = |2 \cdot \mathrm{AUROC} - 1|,$$

so that 0 corresponds to indistinguishable domains and 1 to perfect separability. Finally, Euclidean and Mahalanobis distances between dataset centroids are computed to capture residual domain mismatch in terms of mean displacement and covariance-aware geometry.

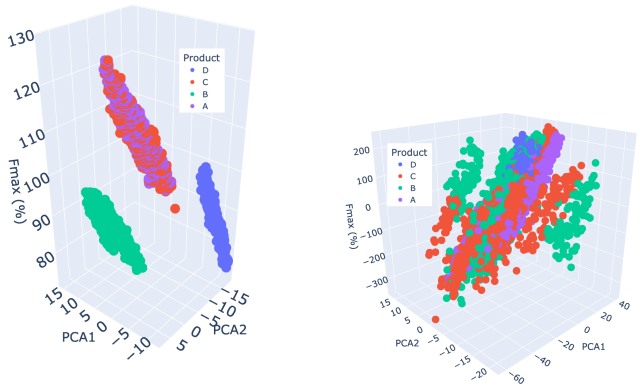


Fig. 1: PCA visualization of the shared feature space. Left: after global scaling, the datasets occupy distinct regions. Right: after per-product scaling, scale and offset differences are reduced, although residual source-target discrepancy remains visible.

V. Results and Discussion

A. Single-Generation Baseline and Cross-Generation Discrepancy

The single-generation baseline is obtained by training the Ridge pipeline exclusively on dataset A. With the full target training set, the model achieves $R^2 \approx 0.95$ and nRMSE $\approx 1.5\%$ (Table I), providing the reference performance level for the target generation.

TABLE I: Comparison of baseline and proposed strategies across different training sizes.

Model	Train	R^2	MAPE	nRMSE	nMAE	AUC
Baseline (A)						
Ridge	20	-	-	-	-	-
	32	**0.92**	**1.47%**	**1.86%**	**1.48%**	**-**
	100	0.95	1.20%	1.52%	1.20%	1.05
	330	0.95	1.20%	1.53%	1.20%	4.60
S1 (A+B+C+D)						
Ridge	20	0.90	1.76%	2.23%	1.77%	-
	32	**0.95**	**1.28%**	**1.62%**	**1.27%**	-
	100	0.94	1.29%	1.69%	1.29%	1.06
	330	0.95	1.19%	1.52%	1.19%	4.62
P. Ridge	20	0.91	1.65%	2.06%	1.63%	-
	32	**0.94**	**1.45%**	**1.81%**	**1.44%**	-
	100	0.94	1.44%	1.79%	1.44%	1.16
	330	0.95	1.22%	1.53%	1.21%	4.78
S2 (A+B+C+D)						
Ridge	20	0.94	1.40%	1.76%	1.40%	-
	32	**0.94**	**1.29%**	**1.65%**	**1.29%**	-
	100	0.95	1.21%	1.57%	1.21%	0.96
	330	0.95	1.19%	1.52%	1.19%	4.64
P. Ridge	20	0.91	1.64%	2.08%	1.63%	-
	32	**0.91**	**1.65%**	**2.10%**	**1.64%**	-
	100	0.95	1.33%	1.67%	1.33%	1.09
	330	0.95	1.22%	1.52%	1.21%	4.81
TabPFN	20	0.94	1.36%	1.69%	1.37%	-
	32	**0.94**	**1.33%**	**1.65%**	**1.32%**	-
	100	0.95	1.23%	1.55%	1.22%	0.97
	330	0.95	1.21%	1.52%	1.20%	4.63
RF	20	0.88	1.99%	2.46%	2.01%	-
	32	**0.91**	**1.69%**	**2.07%**	**1.70%**	-
	100	0.94	1.40%	1.71%	1.40%	1.14
	330	0.95	1.26%	1.57%	1.26%	5.00
S2 (A+B+C)						
Ridge	20	0.93	1.41%	1.77%	1.41%	-
	32	**0.95**	**1.26%**	**1.57%**	**1.26%**	-
	100	0.95	1.21%	1.51%	1.21%	0.92
	330	0.95	1.22%	1.51%	1.22%	4.47
P. Ridge	20	0.88	1.81%	2.33%	1.79%	-
	32	**0.91**	**1.74%**	**2.11%**	**1.72%**	-
	100	0.95	1.27%	1.56%	1.26%	1.11
	330	0.96	1.20%	1.47%	1.20%	4.64

The corresponding learning curve progressively approaches its final low-error regime as more labeled target samples become available. In practice, the curve becomes substantially flatter after approximately 80–100 labeled samples, whereas at 32 samples the baseline still exhibits nRMSE = 1.86%. Since the adopted PCA pipeline retains 32 components, the baseline cannot be evaluated below that threshold.

TABLE II: Pairwise domain separability (AUROC) using linear logistic regression.

Pair	Raw		Per-Product Scaled	
	$\mathrm{AUROC}_\star$	$\mathrm{AUROC}_{\mathrm{dist}}$	$\mathrm{AUROC}_\star$	$\mathrm{AUROC}_{\mathrm{dist}}$
A vs B	1.00	1.00	0.73	0.47
A vs C	0.63	0.26	0.72	0.43
A vs D	1.00	1.00	0.79	0.57
B vs C	1.00	1.00	0.72	0.44
B vs D	1.00	1.00	0.78	0.56
C vs D	1.00	1.00	0.77	0.55

Cross-generation discrepancy is substantial before normalization. Under raw global scaling, the silhouette score is 0.284 and the mean $\mathrm{AUROC}_{\mathrm{dist}}$ reaches 0.877, with five out of six dataset pairs attaining $\mathrm{AUROC}_{\mathrm{dist}} = 1.0$ (perfect separability, see Table II). After per-product normalization, the silhouette score decreases to -0.015 and the mean $\mathrm{AUROC}_{\mathrm{dist}}$ to 0.502 (Table III), indicating that a large component of the observed mismatch is related to scale and offset effects. However, residual separability remains non-zero, motivating controlled adaptation rather than naive cross-generation pooling (Fig. 1).

TABLE III: Cross-generation domain discrepancy before and after per-product normalization.

Metric	**Raw (Global)**	**Per-Product Scaled**
Silhouette score	0.284	-0.015
Mean $\mathrm{AUROC}_{\mathrm{dist}}$	0.877	0.502
Max $\mathrm{AUROC}_{\mathrm{dist}}$	1.000	0.574
Min $\mathrm{AUROC}_{\mathrm{dist}}$	0.261	0.432

B. Incremental Training Strategies

Strategy 1 interleaves legacy and target samples throughout the incremental process. Despite per-product scaling and target-focused weighting, the resulting learning curves exhibit strong oscillations and irregular behavior, as shown in Fig. 2(a). Early-stage accuracy may temporarily improve relative to the single-generation baseline, e.g., nRMSE = 1.62% at 32 samples for A+B+C+D, but these local improvements are not accompanied by smooth or reliable convergence. This behavior is consistent with the residual source-target mismatch observed after normalization and indicates that naive cross-generation mixing is not a stable adaptation protocol in the low-label regime.

Strategy 2 follows a different protocol: the initial predictor is built from the selected legacy datasets, and labeled target samples from dataset A are added only afterward. This yields consistently smoother learning curves and a more regular adaptation trajectory than Strategy 1. At 32 labeled target samples, the single-generation baseline reaches nRMSE = 1.86%, whereas Strategy 2 reduces the error to 1.65% when combining A+B+C+D. More importantly, the curves show that baseline-first adaptation approaches the final low-error region

of the single-generation baseline earlier and with much lower variability than both the baseline itself and Strategy 1.

In this setting, the practical objective is not merely to minimize the final error at full target-data availability, but to approach plateau-level screening performance as early as possible during target-data acquisition. Under this perspective, Strategy 2 is preferable because it reaches the low-error region earlier and with substantially lower variability than naive cross-generation mixing.

To assess source composition, we removed dataset D from the legacy pool. This choice is motivated by the discrepancy analysis: after per-product normalization, the A–D pair exhibits the largest residual separability ($\mathrm{AUROC}_{\mathrm{dist}} = 0.57$), making D the most distant source domain relative to the target generation. Consistently, the A+B+C configuration slightly improves early-stage performance, reaching 1.77% at 20 labeled target samples and 1.57% at 32 samples. This indicates that source selection matters: more legacy data do not necessarily improve adaptation if residual mismatch remains significant. Legacy-data reuse is beneficial when source-target compatibility is sufficiently controlled.

C. Model Comparison and Practical Implications

Figure 2(b) compares the best Ridge-based Strategy 2 configuration with TabPFN and Random Forest. Across the considered low-label regime, Ridge remains competitive with, and in some regions superior to, these more complex alternatives. In particular, Ridge with Strategy 2 and source pool A+B+C reaches the low-error region early and maintains a stable trajectory close to the final baseline plateau. TabPFN shows comparable behavior, whereas Random Forest converges more slowly and remains less competitive in the early low-sample regime. As stated in Section IV-A, these models are reported only in selected configurations as comparative references. While TabPFN attains competitive accuracy even with few labeled samples, its substantially higher memory footprint and inference cost make it less attractive than lightweight linear pipelines for industrial deployment. The same consideration applies, more broadly, to benchmark models whose predictive benefit does not compensate for their increased computational overhead in the intended application scenario.

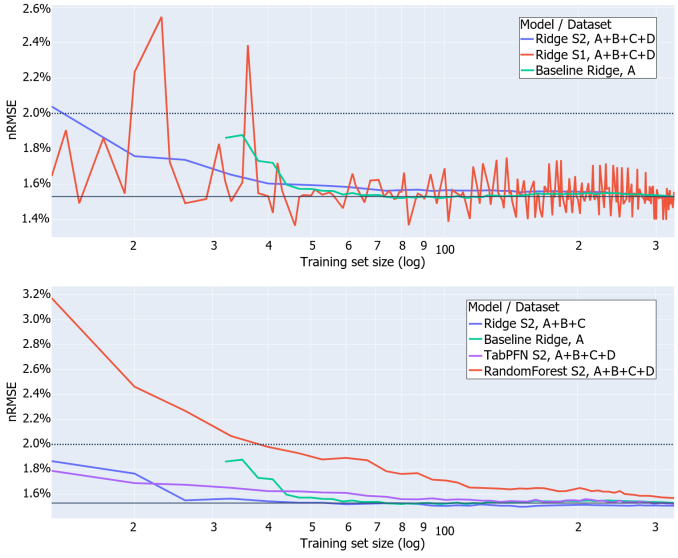
From a practical standpoint, the main advantage of baseline-first adaptation lies in the early adaptation stage rather than in final full-data accuracy. The learning curves show that this protocol approaches plateau-level screening performance with fewer labeled target samples and with a much smoother trajectory than naive cross-generation mixing. Under the adopted stable-attainment criterion, the best source-selected Strategy 2 configuration (A+B+C) reaches the final single-generation baseline-error region at 47 labeled target samples, whereas the baseline reaches that reference only at the end of the 330-sample training process (Table IV). Since target labeling requires time-intensive characterization [1], [8], this corresponds, under the measurement-cost assumptions adopted in this work, to reducing the required labeled target set from 330 to about 47 samples, i.e., from approximately 165 hours to 23.5 hours of characterization effort, for a net saving of about 141.5 hours during early deployment.

TABLE IV: Stable-attainment point with respect to the final single-generation baseline reference (nRMSE = 1.53%).

Configuration	Stable-attainment point $m^\star$
Baseline Ridge, A	330
Ridge S1, A+B+C+D	not reached
Ridge S2, A+B+C+D	289
Ridge S2, A+B+C	**47**
TabPFN S2, A+B+C+D	289
Random Forest S2, A+B+C+D	not reached

However, this benefit depends strongly on source composition. In particular, the A+B+C configuration is the practically relevant one: under the adopted stable-attainment criterion, it reaches the final single-generation baseline-error region substantially earlier than the baseline itself, whereas the A+B+C+D configuration does not provide the same early-attainment advantage. Therefore, the main practical implication is not simply that adding more legacy data helps, but that appropriately selected legacy datasets can reduce the target-label effort required to reach stable deployment-level accuracy. Under the measurement-cost assumptions adopted in this work, this earlier stable attainment can be interpreted as a substantial reduction in characterization effort during early deployment. This estimate should be understood as an engineering interpretation of the observed learning-curve behavior, rather than as a universal cost model.

Fig. 2: Comparison of transfer strategies and model complexity (mean over 5 runs). (a), upper) Strategy 2 yields smoother convergence than Strategy 1 for the A+B+C+D source pool. (b), lower) Under Strategy 2, lightweight Ridge remains competitive with more complex alternatives. The baseline curve starts at 32 target samples because the adopted PCA pipeline retains 32 components.



VI. Conclusion and Future Work

This work investigated dataset-level adaptation protocols for MCU performance screening under cross-generation distribution shift, with the goal of improving target-domain sample efficiency for new product generations.

The reported results show that naive multi-generation data mixing leads to unstable behavior in the low-label regime, whereas a baseline-first adaptation protocol enables smoother and more effective reuse of legacy datasets. However, the benefit of transfer is not monotonic in the amount of legacy data: the experiments show that source selection is critical, since residual source-target mismatch can offset the expected gain from simply enlarging the source pool. When combined with per-product normalization and appropriate source selection, baseline-first adaptation reaches the target low-error region earlier than the single-generation baseline, thereby improving practical sample efficiency under target-label scarcity.

The present study remains centered on adaptation to a single target generation. Future work will therefore extend the analysis to multi-target evaluation settings, investigate automated source selection based on quantitative domain-discrepancy measures, and assess adaptive sampling mechanisms that explicitly account for residual source-target mismatch.

References

[1] R. Cantoro et al., "Machine Learning based Performance Prediction of Microcontrollers using Speed Monitors," in *IEEE International Test Conference (ITC)*, 2020.

[2] N. Bellarmino et al., "Grouped Feature Selection for SMONs Placement in MCU Performance Screening," in *2025 IEEE 26th Latin American Test Symposium (LATS)*, IEEE, 2025.

[3] N. Bellarmino et al., "Deep learning strategies for labeling and accuracy optimization in microcontroller performance screening," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2024.

[4] T. Kilian et al., "An efficient high-volume production performance screening using on-chip ring oscillators," in *2023 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)*, 2023.

[5] T. Kilian et al., "A scalable design flow for performance monitors using functional path ring oscillators," in *2021 IEEE International Test Conference (ITC)*, 2021.

[6] F. Angione et al., "Test, reliability and functional safety trends for automotive system-on-chip," in *2022 IEEE European Test Symposium (ETS)*, 2022.

[7] N. Bellarmino et al., "Exploiting active learning for microcontroller performance prediction," in *IEEE European Test Symposium (ETS)*, 2021.

[8] N. Bellarmino et al., "A Multi-Label Active Learning Framework for Microcontroller Performance Screening," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2023.

[9] H.-G. Stratigopoulos et al., "Nonlinear decision boundaries for testing analog circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2005.

[10] H.-G. Stratigopoulos et al., "Defect filter for alternate rf test," in *2010 15th IEEE European Test Symposium*, 2010.

[11] H. Ayari et al., "Making predictive analog/rf alternate test strategy independent of training set size," in *IEEE International Test Conference (ITC)*, 2012.

[12] H. Amrouch et al., "Special session: Machine learning for semiconductor test and reliability," in *Proceedings of the 2021 IEEE 39th VLSI Test Symposium (VTS)*, Invited paper, 2021.

[13] J. Liaperdos et al., "Machine learning in alternate testing of integrated circuits," in *Machine Learning Paradigms*, ser. Learning and Analytics in Intelligent Systems, G. A. Tsihrintzis et al., Eds., Cham: Springer, 2019.

[14] R. McLaughlin et al., "Automated Debug of Speed Path Failures Using Functional Tests," in *27th IEEE VLSI Test Symposium*, 2009.

[15] J. Brockman et al., "Predictive subset testing: Optimizing ic parametric performance testing for quality, cost, and yield," *IEEE Transactions on Semiconductor Manufacturing*, 1989.

[16] P. Variyam et al., "Prediction of analog performance parameters using fast transient testing," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2002.

[17] N. Bellarmino et al., " In-Context Learning for Microcontroller Performance Screening Using Tabular Foundation Models," in *2025 28th Euromicro Conference on Digital System Design (DSD)*, Los Alamitos, CA, USA: IEEE Computer Society, Sep. 2025.

[18] N. Bellarmino et al., "Feature Selection for Cost Reduction In MCU Performance Screening," in *IEEE 24th Latin American Test Symposium (LATS)*, 2023.

[19] N. Bellarmino et al., "Cosmo: Compressed sensing for models and logging optimization in mcu performance screening," *IEEE Transactions on Computers*, 2024.

[20] S. Asai, Ed., *VLSI Design and Test for Systems Dependability*. Springer Japan, 2019.

[21] N. Bellarmino et al., "Semi-Supervised Deep Learning for Microcontroller Performance Screening," in *IEEE European Test Symposium (ETS)*, 2023.

[22] C. Wang et al., "A comprehensive survey on domain adaptation for intelligent fault diagnosis," *Knowledge-Based Systems*, 2025.

[23] W. Chen et al., "A transfer learning method for fault diagnosis of analog circuit using deep subdomain adaptation network," in *Proceedings of the 2023 Prognostics and Health Management Conference (PHM-Paris)*, 2023.

[24] N. Bellarmino et al., "Enabling Inter-Product Transfer Learning on MCU Performance Screening," in *IEEE Asian Test Symposium (ATS)*, 2023.

[25] J. Zeng et al., "On correlating structural tests with functional tests for speed binning of high performance design," in *IEEE International Test Conference (ITC)*, 2004.

[26] N. Bellarmino et al., "Microcontroller Performance Screening: Optimizing the Characterization in the Presence of Anomalous and Noisy Data," in *IEEE International Symposium on On-Line Testing and Robust System (IOLTS)*, 2022.