Exploiting Active Learning for Microcontroller Performance Prediction

Exploiting Active Learning for Microcontroller Performance Prediction

Nicolò Bellarmino*, Riccardo Cantoro*, Martin Huch†, Tobias Kilian†‡,
Raffaele Martone*, Ulf Schlichtmann‡ and Giovanni Squillero*

*Politecnico di Torino, Torino, Italy
†Infineon Technologies AG, Munich, Germany
‡Technical University of Munich, Munich, Germany

***Abstract*—Speed monitors provide on-chip measurements of the the performance of integrated circuits. In recent years, they have been extensively used to predict $F_{\max}$ of microcontrollers for speed binning and performance screening during production test. However, while the use of machine learning is getting increasingly popular, the models may become significantly inaccurate if not trained on the appropriate devices. Previous research has demonstrated how to predict performance from speed-monitor data using corner-lot wafers. We show how to extend this approach to select the best corner-lot wafers to label when preparing the training set, thus significantly reducing the time and cost required for the process.**

***Index Terms*—Performance Screening, Fmax, Speed Monitors, Machine Learning, Active Learning**

I. Introduction

Microcontrollers (MCUs) are important components for many products, including many safety-critical products such as automotive and aerospace electronics. Manufacturers must assure their customers that the delivered devices can fulfill the specifications of the datasheet. To meet dependability standards, MCUs undergo several different tests; and one of those tests is performance screening. The performance $F_{\max}$ is the maximal frequency at which the device can correctly execute all its tasks. The desired level of confidence for $F_{\max}$ may vary, and devices employed in safety-critical applications require much higher effort in performance screening than devices for the low-cost consumer market.

Several different approaches exist to identify the $F_{\max}$ of a circuit; the most common are based on structural patterns (e.g., for transition delay testing or path delay testing), functional patterns, and speed monitors [1]–[5]. To implement speed monitors, different kinds of ring oscillators are inserted on-chip at different locations to catch various performance variations [6], [7]. Prediction algorithms can be used to characterize performance-related physical parameters of the chip using speed monitors; they can be based on simple data analysis (e.g., correlations between two quantities) or on more complex machine learning (ML), where algorithms are fed with multiple "features" [3], [4], [7], [8]. A model is trained on *labeled* devices, that is, devices with known $F_{\max}$, and it is later used to predict the $F_{\max}$ of new, *unlabeled* devices. Since the effects of manufacturing defects may not be observable using speed monitors, different methods to identify such outliers, or additional features, may be required [9]–[11].

Wafer-specific relations between $F_{\max}$ and features are known to exist [12]. Thus, if only chips from a single wafer — or from very similar wafers — are used for the initial performance characterization, there is a significant risk that the model will fail on new data. The term "dataset shift" [13]–[15] is sometimes used when the joint distribution of inputs and outputs differs between training and test, while in biological applications these phenomena may be referred to as "batch effect" [16]. Moreover, the effects of manufacturing defects are not easily observable using speed monitors and negatively influence the prediction [3], [8]; in this case, additional features are needed, or methods to identify outliers [9]–[11].

In this paper, we propose a methodology to identify the most useful wafers for model characterization, by observing only the speed-monitor values. Considering how expensive it is to collect functional $F_{\max}$ values for labeling, our methodology can impact significantly on the cost of IC production. The key contribution of this work is a methodology based on unsupervised active learning able to select an effective dataset for training the initial ML models; in our experiments, it reduced by more than one half the size of training.

The rest of the paper is organized as follows. Section II presents related work. Section III presents the generalization problem in ML and active learning as a means to improve it. The proposed methodology is presented in Section IV, and the results gathered in our experiments are presented in Section V. Section VI concludes the paper.

II. Related work

The correlation between structural and functional $F_{\max}$ was investigated in several works. An approach using complex ML algorithms was first presented in [3], while previous works were only considering single parameters [1], [2]. The work identified, among various algorithms applied on a dataset of 60 devices, the *Gaussian Process* [17] as the most effective way to learn from different sets of features (100 flip-flops, 100 transition fault patterns, and 112 testable paths). The work also showed that outliers negatively influence the prediction, and should be identified and removed from the training set; a *conformity check* was implemented for this purpose.

Authors are listed in alphabetical order.

In a related work, the same research group compared two lots of devices of the same product but with different packages, composed of 79 and 74 devices, respectively [4]. According to authors, the models trained on one lot were not able to predict devices belonging to the other lot, and this might be caused by the use of different packages or by the limited number of samples in the dataset.

In a more recent work, a different research group correlated the functional $F_{\max}$ with low-cost embedded sensors to accurately measure the slack of a selected number of paths [7]. The method was tested on 300 OpenSPARCT2 SoC devices equipped with 25 sensors, using transition level simulations of two process variation scenarios with both features and labels coming from the wafer sort.

In [8], we correlated the values of 27 speed monitors coming from wafer sort to functional $F_{\max}$ measured on more than 4,000 packaged devices extracted from 26 corner-lot wafers; training devices were randomly selected among the available ones.

III. (Lack of) Generalizability in ML models

The term *generalizability* is broadly used to describe a model's ability to make accurate predictions on new, unseen data. Attaining it is the ultimate ambition of ML scholars, but it is not always fulfilled.

The best known cause of a lack of generalizability is *overfitting*: when a model instead of *learning* is simply *memorizing* the training set, the error on new data will be significantly higher than on the training data [18]. Another cause of the lack of generalizability is *dataset shift*: when new data differ systematically from the one used during training, performance are naturally impaired.

Dealing with the generalization problem requires analyzing new data and understanding when they are not modeled properly. And, if the problem can be attributed to dataset shift, selecting how to re-train the model. In the following, we describe the concept of dataset shift, which strongly affects the generalization in our models, and active learning that allows us to cope with this problem.

A. Dataset shift

The hypothesis of independent and identically distributed data (IID) often does not hold in a real-world application, and this can lead to sub-optimal models. Dataset shift manifests itself in different forms: the first is the *covariate shift*, when the features considered in the problem, i.e., the independent variables, have different distributions in the test and training set: $p_{\text{train}}(x) \neq p_{\text{test}}(x)$. The second form is the *prior probability shift*, when the dependent variable distributions (i.e., the label distributions) between test and training set are different: $p_{\text{train}}(x|y) = p_{\text{test}}(x|y)$ and $p_{\text{train}}(y) \neq p_{\text{test}}(y)$.

The most common causes are bias in samples selection, that can also occur in cross-validation split schemes without noticing [13], and non-stationary environments that occur when training and test sets environments are different due to temporal or spatial change [14]. Technological processes, such as those we are dealing with, consist of several steps carried out in different environments and influenced by time-dependent parameters: they are therefore non-stationary environments, where the data measurements depend on the time at which the measurement is done. This may lead to time-dependent distributions of the training data causing the dataset shift: time drift is, therefore, to be taken into account during the construction of a model.

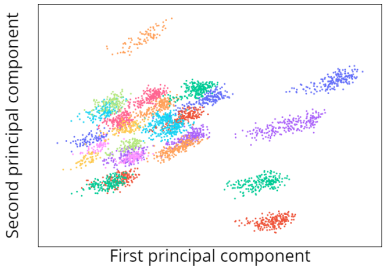

Fig. 1. PCA-based two-dimensional representation of the corner-lot devices used for the initial model's characterization, colored by wafer.

In the current application, the dataset shift cause can be the fact that we have samples from different well-distinguished wafers, and some of them can be underrepresented in a classical random train-test split; new clusters are likely to appear in different positions during production. Fig. 1 shows an example of this phenomenon: devices belonging to the same wafer are plot using the same color.

B. Active learning

In some domains, annotated data is hard and expensive to obtain, such as the cost of the process or the time needed. Active learning (AL) aims at easing the data collection process by automatically deciding which instances an annotator should label to train an algorithm as quickly and effectively as possible. The general settings for an AL framework provide a model trained on a small amount of data and the learning of a function that decides which data points must be inserted in the training set in order to improve the model score.

AL is based on the fact that an ML model trained with a small amount of carefully-chosen data can perform as well as the same model trained with a large random dataset [19], if not better. AL can be applied to classical well-known machine learning models such as SVM to significantly improve classification accuracy with a small amount of labeled data.

Applications such as Remote Sensing Image Classification [20], Text Classification [21], Optical Networks [22] and certainly ours too, need a representative dataset to learn from: training sets created with random labeling can lead to noisy or redundant data, slowing down the learning process. Several methods exist and can be used to select the best samples that have to be labeled, in order to improve classification score. Active Learning performs as well as Domain Adaptation Transfer Learning approaches [22], and can be an alternative to this technique if obtaining samples from other sources is expensive.

IV. Proposed Methodology

The method we proposed in [8] is used as the baseline for the active learning approach. The method was based on state-of-the-art regression algorithms, trained using speed-monitor values as features, and performing the fine-tuning of the models' hyper-parameters using cross-validation. All corner-lot wafers were used to build the model, in a random train-test split; no wafer-level information was taken into account by the method. The root mean squared error (RMSE) of the predictions using the models can be used as the reference error.

In the proposed methodology, wafer-level information is derived and used in an active learning flow. Such an aggregate information is given by the unsupervised strategy used to select the target wafer to include in the training set — we will refer to it as the *selection strategy* [23] of the active learning. Selection strategies compare samples in the current training set with samples in new wafers. Three strategies are analyzed in this work: *Euclidean distance* (ED), *Local Outlier Factor* (LOF), and *Query by Committee* (QBC).

The Euclidean distance $d(W, w_n)$ between wafers in a training set W and a new wafer $w_n \notin W$ is computed as

$$d(W, w_n) = \min_{\forall i \in W} (c_{w_i} - c_{w_n}) \quad (1)$$

where c_{w_i} is the centroid of a generic wafer w_i, calculated as the mean vector in w_i. ED values are computed on the normalized features. Given a set of new wafers, the one with the highest ED is selected.

The *Local Outlier Factor* algorithm [24] measures the local deviation of density of a given sample with respect to its neighbors. It is "local" in that the anomaly score depends on how isolated the object is with respect to the surrounding neighborhood. By comparing the local density of a sample to the local densities of its neighbors, one can identify samples that have a substantially lower density than their neighbors. In our case, the anomaly score LOF(w_n) is computed as the mean value of all anomaly scores LOF($x_i \in w_n$) of the samples in the new wafer w_n. Outliers have higher LOF values, thus the wafer with the highest LOF is selected.

The *Query-by-Committee* method [25] constructs a "committee" of (two or more) models based on the statistics of the current training set. The QBC measures the degree of disagreement among the committee members. In our problem, unlabeled wafers are processed by each of the members and the wafer having the maximum variance in predictions, that is the wafer on which they most disagree, is selected. In the experimental results, we used Decision Tree Regressor, rbf-based Support Vector Regression (SVR), Gaussian Process Regressor, and Polynomial Ridge as the committee members.

Experimental analyses based on the Pearson correlation or the Spearman correlation show that the proposed selection strategies correlate with the error computed using the $F_{\max}$ regression model on the labeled devices.

The proposed flow can be summarized as follows:

1) Given a set of labeled wafers W^L and a set of unlabeled wafers W^U, we identify candidate wafers by means of a selection strategy model and a threshold τ; for example, using LOF, we identify the set $W^N \subseteq W^U$ where each wafer is selected as follows: $w_i^U \in W^N \iff LOF(w_i^U) > \tau$
2) If the set of candidate wafers W^N is not empty, the active learning strategy can be applied: wafers in W^N are ranked according to the selection strategy (LOF in the example). One wafer at a time is extracted from W^N according to the ranking and added to a set of selected wafers W^S. The selection strategy model is retrained using $W^L \cup W^S$ and the remaining wafers in W^N are checked again with the new model: wafers below the threshold are removed from W^N. The extraction/retrain process continues until W^N is empty.
3) Wafers in W^S are labeled and a new $F_{\max}$ model is developed.

One corner-lot wafer has to be included in the initial dataset of the selection strategy model, e.g., the wafer in the production corner-lot.

V. Experimental Evaluation

A. Experimental setup

The proposed methodology was applied to a dataset of 3,616 labeled samples from 26 corner-lot wafers; for each wafer, the number of devices labeled ranged from 46 to 204, with an average of 139. Each sample is described by 27 features, one for each speed monitor. Labels related to functional $F_{\max}$ were collected following the process described in [8], as well as outliers were removed from each wafer as described there.

B. Model selection for active learning

We compared three regression algorithms – Ridge regression using polynomial features, Support Vector Regression (SVR), and Random Forest (RF) – to study the generalization problem in our domain. We used a random approach where $F_{\max}$ models were trained using a subset of the available wafers. At each step, the models are tested on the remaining wafers. Fig 2 shows that the error – normalized RMSE (nRMSE) in this case – is high in the first steps and progressively decreases when adding other wafers. In the last step, SVR and RF are slightly better than Ridge; however, Ridge presents a better generalization in the intermediate steps and the lowest area under the curve, showing how more complex models tend to do overfitting if the amount of data is not sufficient. We used Ridge in the remaining experiments.

We evaluated the three proposed selection strategies. The iterative process used to compare regression algorithms was repeated using active learning with ED, LOF, and QBC, instead of random selection, using each of the 26 wafers in the initial step. The results are reported in Table I. As expected, the models converge way faster than random; ED is the worst strategy, but the fastest to implement, while LOF and QBC have similar profiles. According to the results, we can obtain good accuracy (below 2% of nRMSE) with 13 wafers using a random approach, 11 using ED, 8 using QBC, and only 6

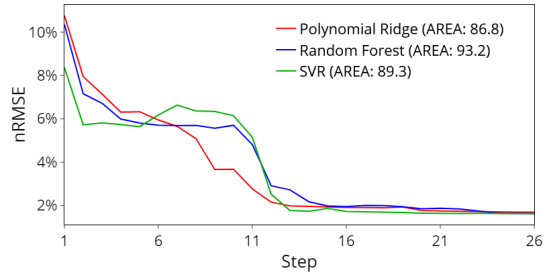

Fig. 2. nRMSE obtained with different regression algorithms, using a random strategy. The curves represent the mean value of the errors depending on all possible initial wafer choices.

TABLE I
POLYNOMIAL RIDGE REGRESSION MEAN AND VARIANCE NRMSE DURING ACTIVE LEARNING USING ALL POSSIBLE INITIAL WAFER CHOICES

Step	nRMSE mean [%]				nRMSE variance [%]			
	Rnd	ED	LOF	QBC	Rnd	ED	LOF	QBC
1	10.76	10.76	10.76	10.76	3.29	3.29	3.29	3.29
2	7.95	4.84	4.95	4.92	2.96	2.68	2.71	2.54
3	7.15	3.39	2.80	2.83	2.98	0.70	0.57	0.63
4	6.31	2.56	2.26	2.58	2.62	0.54	0.23	0.46
5	6.33	3.09	2.12	2.41	2.58	0.97	0.16	0.53
6	5.95	2.41	1.99	2.15	2.45	0.47	0.16	0.21
7	5.65	2.26	1.93	2.02	2.25	0.28	0.09	0.13
8	5.09	2.11	1.89	1.92	1.85	0.27	0.06	0.08
9	3.66	1.99	1.89	1.91	1.26	0.16	0.07	0.09
10	3.68	2.01	1.87	1.87	1.31	0.11	0.06	0.08
11	2.77	1.99	1.83	1.84	0.48	0.09	0.05	0.06
12	2.16	1.94	1.82	1.82	0.11	0.07	0.06	0.06
13	1.98	1.89	1.80	1.78	0.06	0.05	0.05	0.03
...								
26	1.70	1.70	1.70	1.70	0.00	0.00	0.00	0.00

wafers using LOF, which is also the metric with the lowest nRMSE variance.

The results obtained on the $F_{\max}$ models at each step of the active learning were correlated to the unsupervised metric and reported in Table II. The four metrics we used – mutual information, Pearson, Spearman, and Kendall [26] – show very strong correlations between nRMSE and each of ED, LOF, and QBC; also in this case, LOF seems to be the best model to use for the active learning. A graphical view of the correlation can be observed in Fig. 3 (mean values are reported).

TABLE II
CORRELATION BETWEEN SELECTION STRATEGIES AND NRMSE ON THE $F_{\max}$ RIDGE REGRESSION MODELS WITH POLYNOMIAL FEATURES

Selection strategy	Mutual information	Pearson correlation	Spearman correlation	Kendall correlation
Max. ED	0.928	0.708	0.927	0.782
Max. LOF	0.889	0.828	0.967	0.849
Max. QBC	0.898	0.526	0.973	0.854

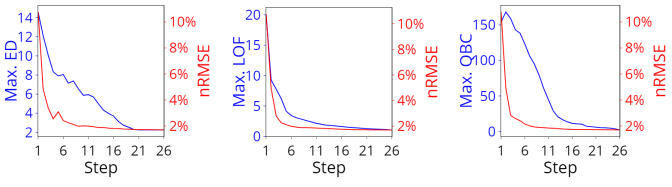

Fig. 3. Correlation between unsupervised scores (EQ, LOF, QBC) for active learning and nRMSE using polynomial Ridge regression.

VI. CONCLUSIONS

We presented an innovate work based on active learning for the performance screening of microcontrollers. Three unsupervised algorithms that highly correlate with the accuracy on the $F_{\max}$ regression models used for production screening were successfully tested on a set of corner-lot wafers. Using the proposed unsupervised algorithms, we observed a significant speedup in the development of the $F_{\max}$ models, given an accurate selection of the most promising devices for the training. In our experiments, we shown that a comparable accuracy can be reached with just a fraction of the dataset.

REFERENCES

[1] B. D. Cory *et al.*, "Speed binning with path delay test in 150-nm technology," *IEEE Design Test of Computers*, 2003.

[2] J. Zeng *et al.*, "On correlating structural tests with functional tests for speed binning of high performance design," in *2004 ITC*, 2004.

[3] J. Chen *et al.*, "Data learning techniques and methodology for Fmax prediction," in *2009 ITC*, 2009.

[4] J. Chen *et al.*, "Selecting the most relevant structural Fmax for system Fmax correlation," in *2010 VTS*, 2010.

[5] S. Mu *et al.*, "Statistical Framework and Built-In Self-Speed-Binning System for Speed Binning Using On-Chip Ring Oscillators," *IEEE VLSI*, 2016.

[6] T. Chan *et al.*, "Synthesis and Analysis of Design-Dependent Ring Oscillator (DDRO) Performance Monitors," *IEEE VLSI*, 2014.

[7] M. Sadi *et al.*, "SoC Speed Binning Using Machine Learning and On-Chip Slack Sensors," *IEEE TCAD*, 2017.

[8] R. Cantoro *et al.*, "Machine Learning based Performance Prediction of Microcontrollers using Speed Monitors," in *2020 ITC*, 2020.

[9] S. H. Wu *et al.*, "A Study of Outlier Analysis Techniques for Delay Testing," in *2008 ITC*, 2008.

[10] D. Sinwar *et al.*, "Outlier detection from multidimensional space using multilayer perceptron, RBF networks and pattern clustering techniques," in *2015 ICACEA*, 2015.

[11] M. Shintani *et al.*, "Artificial Neural Network Based Test Escape Screening Using Generative Model," in *2018 ITC*, 2018.

[12] W. Zhang *et al.*, "Automatic clustering of wafer spatial signatures," in *2013 DAC*, 2013.

[13] J. G. Moreno-Torres *et al.*, "Study on the Impact of Partition-Induced Dataset Shift on k-Fold Cross-Validation," *IEEE Transactions on Neural Networks and Learning Systems*, 2012.

[14] H. Raza *et al.*, "Dataset Shift Detection in Non-stationary Environments Using EWMA Charts," in *2013 IEEE SMC*, 2013.

[15] D. Tuia *et al.*, "Dataset shift adaptation with active queries," in *2011 Joint Urban Remote Sensing Event*, 2011.

[16] H. Li *et al.*, "A New Approach to Batch Effect Removal Based on Distribution Matching in Latent Space," in *2019 BIBM*, 2019.

[17] C. E. Rasmussen *et al.*, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[18] P. Barbiero *et al.*, *Modeling Generalization in Machine Learning: A Methodological and Computational Study*, 2020.

[19] S. Wang *et al.*, "Pool-based active learning based on incremental decision tree," in *2010 ICMLC*, 2010.

[20] D. Tuia *et al.*, "A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification," *IEEE JSTSP*, 2011.

[21] S. Tong *et al.*, "Support Vector Machine Active Learning With Applications To Text Classification," *JMLR*, Dec. 2001.

[22] D. Azzimonti *et al.*, "Comparison of domain adaptation and active learning techniques for quality of transmission estimation with small-sized training datasets [invited]," *IEEE OSA*, 2021.

[23] S. Larguech *et al.*, "Evaluation of indirect measurement selection strategies in the context of analog/RF alternate testing," in *LATW*, 2014.

[24] M. M. Breunig *et al.*, "LOF: Identifying Density-Based Local Outliers," in *ACM SIGMOD*, Dallas, Texas, USA, 2000.

[25] H. S. Seung *et al.*, "Query by Committee," in *ACM COLT*, Pittsburgh, Pennsylvania, USA, 1992.

[26] A. G. Sefidmazgi *et al.*, "Correlation analysis as a dependency measures for inferring of time-lagged gene regulatory network," in *IKT*, 2016.