Neural Networks for Authenticating Integrated Circuits Based on Intrinsic Nonlinearity

Sudarsan Sadasivuni*, Sanjeev Tannirkulam Chandrasekaran*, Akshay Jayaraj† and Arindam Sanyal*
*Electrical Engineering Department, University at Buffalo, Buffalo, NY.
†Intel Corporation, Folsom, CA 95630, USA. Email: {ssadasiv, arindams}@buffalo.edu

***Abstract*—This work presents a machine learning approach to identify integrated circuits based on intrinsic nonlinearity arising out of random variations introduced during device fabrication. The random variations ensure that each integrated circuit has a distinct nonlinearity signature which can be analyzed by a machine learning model to distinguish between chips fabricated from the same mask. We have analyzed multiple samples of two analog-to-digital converters (ADCs) - a continuous-time $\Delta\Sigma$ oversampled ADC and a discrete-time nyquist ADC. The two ADCs have different dominant nonlinearity contributors - inter-symbol interference for the oversampled ADC and static mismatch for the nyquist ADC. A 3-layer artificial neural network can identify the different sample chips for each ADC with a worst-case mean accuracy of 95.97%.**

I. Introduction

Physical Unclonable Functions (PUFs) have become popular over the last decade as low power solutions for hardware authentication and secret key generation. Entropy source for Si PUF is random variations introduced during fabrication of integrated circuits. When interrogated with same challenge vector, different PUFs fabricated from the same mask produce orthogonal responses which are ideally unpredictable. Popular PUFs leverage mismatches in delays between nominally identical paths [1], [2], or differences in threshold voltages between transistors [3], [4] to generate a unique "fingerprint". Similar to PUFs, integrated circuits, such as data converters, experience random variations introduced during fabrication. Hence, in theory, random variations in an integrated circuit can be extracted to form a unique identifier without requiring a dedicated PUF circuit. The work in [5] advanced this premise by digitizing mismatch between unit elements of digital-to-analog converter (DAC) in a $\Delta\Sigma$ modulator. The digitized mismatch data is used to derive a weak PUF with high uniqueness. However, limitation of the technique in [5] is that it requires a high resolution auxiliary mismatch estimation circuit which adds to the design overhead.

We propose a machine learning based approach in which artificial neural networks (ANNs) are used to extract a unique identifier from integral nonlinearity (INL) data of an analog-to-digital converter (ADC). The proposed approach does not require any additional circuit for calculating INL, and thus, has no overhead. We have validated our proposed approach on two different data converters - 1) a 12-bit continuous-time (CT) ring oscillator based $\Delta\Sigma$ ADC [6], and 2) an 11-bit discrete-time (DT) successive approximation register (SAR) ADC [7]. Both ADCs are fabricated in 65nm CMOS process. For the VCO ADC, the dominant source of nonlinearity is inter-symbol interference (ISI) error, whereas in SAR the dominant source of nonlinearity is static mismatch in the capacitive DAC. For the two classes of ADCs, we have INL data for multiple chips measured under different voltage and temperature conditions. For each class of ADC, machine learning models can distinguish between the different chips with a mean accuracy >95%. Section II describes how the data set is created for applying the machine learning models, while Section III presents the validation results from different machine learning models. Finally, the conclusion is brought up in Section IV.

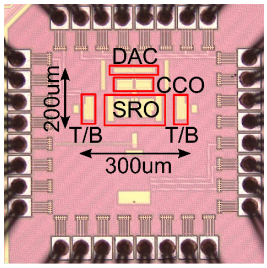
II. Proposed Machine Learning Based Identification Technique

A. Overview of ADCs Used

As mentioned in Section I, we have selected an oversampled, CT ADC and a nyquist, DT ADC with different dominant nonlinearity sources to validate our claims. Static element mismatch in the VCO ADC is high-pass shaped by intrinsic data weighted averaging (DWA) and ISI error, arising primarily due to unequal rise and fall times in the DAC, is the dominant source of nonlinearity. While the mismatch estimation technique of [5] can be applied to extract mismatch in the SAR ADC and derive a weak PUF, the estimation technique cannot be easily applied to the VCO ADC to estimate ISI errors. However, machine learning models can be applied directly to INL data from both the ADCs without requiring any additional signal processing steps.

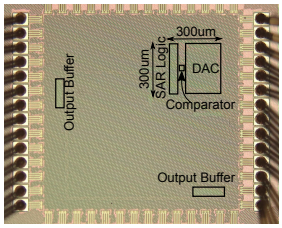
Fig. 1 shows the VCO ADC die photograph and summarizes the ADC performance. The ADC has an SNDR of 70.2dB and SFDR of 81dB, and the SNDR is not limited by distortion tones due to ISI error. Fig. 2 shows SAR ADC die photograph and summarizes its performance. The SAR ADC has an SNDR of 63.5dB and SFDR of 70dB. Similar to VCO ADC, distortion tones arising out of capacitor mismatch do not limit the SNDR. INL data from the VCO ADC is recorded from 5 chips over voltage and temperature corners ranging from 0.9-1.1V and 0-50C respectively, while INL data from the SAR ADC is recorded from 6 chips over voltage and temperature corners of 0.9-1.2V and 0-50C respectively. While the INL data is available for all the VCO chips for all voltage and temperature corners, chips 4 and 5 for SAR ADC did not have measurement data available for all the corners.

978-1-7281-8058-8/20/$31.00 ©2020 IEEE



Supply(V)	1
Power(mW)	0.1
Area(mm^2)	0.06
F$_s$(MHz)	52
BW(MHz)	2.3
SNDR(dB)	70.2
SFDR(dB)	81
FoM$_w$(fJ/step)	8.6

Fig. 1: VCO ADC die photo and performance summary [6]



Supply(V)	1
Power(μW)	9
Area(mm^2)	0.12
F$_s$(MHz)	1.28
BW(MHz)	0.64
SNDR(dB)	63.5
SFDR(dB)	70
FoM$_w$(fJ/step)	5.6

Fig. 2: SAR ADC die photo and performance summary [7]

Fig. 3(a) and (b) show INL data for chip 2 and chip 3 of VCO ADC respectively. The INL data under nominal condition (1V, 30C) is highlighted in thick, black line for both the chips. Fig. 4(a) and (b) show INL data for chip 1 and chip 3 of SAR ADC respectively. INL data under nominal condition (1V, 30C) is highlighted in thick, black line for both chips.

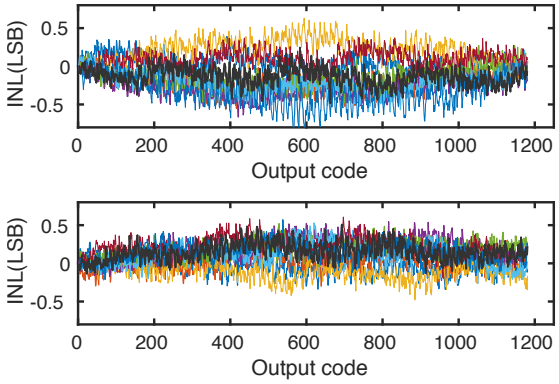


Fig. 3: Measured INL for 2 VCO ADC chips across VT

B. Dataset Generation

The motivation of this work is to classify the different ADC chips from their nonlinearity signature. We first calculated Lee-distance [8] between the different chips by quantizing the INL and DNL values to 12-bit numbers. Lee-distance is similar to hamming distance but is calculated on a $q-$ary alphabet ($q \geq 2$) instead of binary alphabet. Fig. 5(a) and (b) show the normalized inter and intra Lee-distances for the VCO and SAR ADCs respectively, with inter distances being calculated between INL/DNL values from different chips of the same ADC and intra distances being calculated between INL/DNL values from the same ADC chips measured at different VT

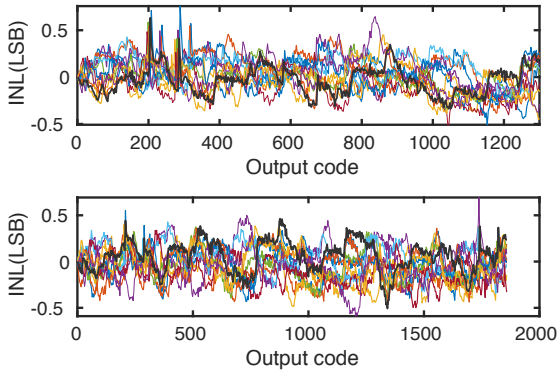


Fig. 4: Measured INL for 2 SAR ADC chips across VT

conditions. The inter and intra Lee-distances overlap for both SAR and VCO ADCs, thus making classification based on Lee-distance alone impossible. Hence, we decided to employ machine learning models which non-linearly project the INL data to higher dimension and allows classification of different chips as will be shown in the following sections.

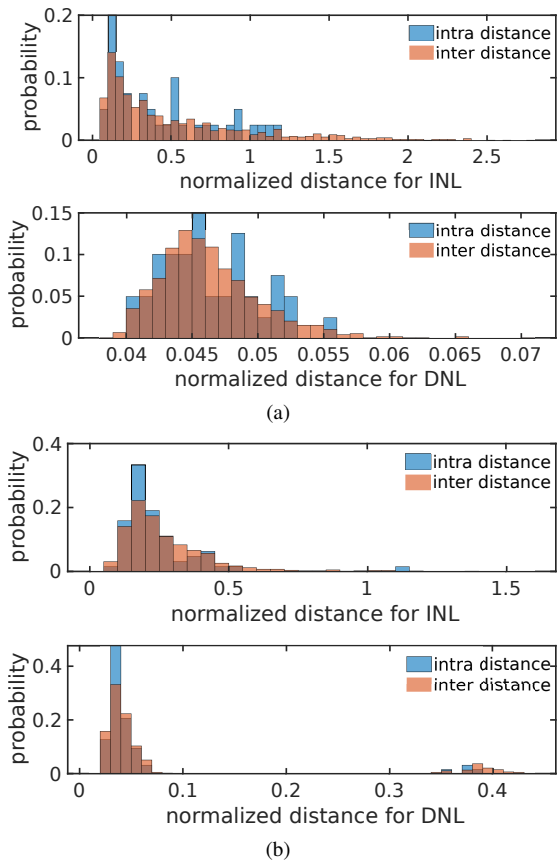


Fig. 5: Normalized inter and intra Lee-distances between INL and DNL for a) VCO b) SAR chips

We have 45 INL curves for the 5 VCO ADC chips and 69 INL curves for the 6 SAR ADC chips. While these data points are usually enough for ADC testing and characterization, the data set is still small for training a machine learning model. To increase the number of data points, we synthesized

additional INL curves for each chip by adding a small error with standard deviation of 0.01 to 200 randomly selected points on the original INL curve. The standard deviation is selected to ensure that synthesis of data does not corrupt the class boundaries, and is verified by performing 2-sample Kolmogorov-Smirnov (KS) test [9]. The 2-sample KS test is used to check the null hypothesis that the two sequences under test are from the same distribution. The result of KS test is '1' if it rejects the null hypothesis at 5% significance level, i.e, if the two sequences under test are from different distributions, and '0' otherwise. For all the synthesized INL data, KS test returns '0' when compared against INL data from the same class of chip, and returns '1' when compared against INL data from the other chip classes. After adding synthesized data, there are 135 INL curves corresponding to 5 chip classes for the VCO ADC, and 207 INL curves corresponding to 6 chip classes for the SAR ADC. The number of synthesized data points is selected to ensure that the machine learning model does not over-fit during training. We tried increasing the synthesized data further, but it results in over fitting of the model. To avoid this, we selected 135 INL curves for 5 chips of VCO ADC and 207 INL curves for 6 chips of SAR ADC.

III. Chip Identification Results

We used three different machine learning models - K-Nearest Neighbor (k-NN), 2-layer artificial neural network (ANN) and 3-layer artificial neural network (ANN) to classify the different chips for the two ADCs. K-NN is an unsupervised machine learning model and used as our baseline model, while the ANNs are supervised machine learning models. For the k-NN, we used 10-NN classifier with euclidean distance and squared inverse distance weight metrics. For the ANNs, we used tanh activation function for hidden layers and softmax activation function for output layer. The 2-layer ANN has 100 neurons in the first hidden layer and 50 neurons in the second hidden layer, while the 3-layer ANN has 100 neurons in the first hidden layer, 50 neurons in the second hidden layer and 25 neurons in the third hidden layer. For the ANN models, we used 70% of the data for training, 30% of the data for testing. The testing of the model was performed on data by excluding all synthesized data used as validation dataset. All the 3 models are trained and validated separately for the SAR and VCO ADC chips. The number of neurons in the different layers for the ANN are chosen to maximize accuracy of classifying various chips. As an example, Table I shows the average classification accuracy and f-1 score for different 3 layer ANN architectures for both SAR and ADC chips. For both SAR and VCO ADCs, 100-50-25 architecture has the highest average accuracy and f-1 score.

Table II compares classification performance of the ANNs and k-NN on the validation dataset from VCO and SAR ADC classes. The ANN models were simulated 50 times, with new training, test and validation sets selected randomly each time. For each chip, Table II shows mean accuracy and f-1 score for ANN models as well as the standard deviations. The 3-layer ANN has the best average accuracy and f-1 score for both VCO and SAR ADCs, and has 6% more average accuracy than the baseline k-NN (k=10) model.

TABLE I: Classification results for different 3-layer ANNs

	VCO			SAR	
	Acc	**f-1**		**Acc**	**f-1**
100-90-80	0.95	0.89	**100-50-25**	0.99	0.97
100-75-50	0.96	0.90	**300-200-100**	0.97	0.90
100-50-25	0.98	0.91	**100-90-80**	0.96	0.91
300-200-100	0.95	0.88	**100-75-50**	0.98	0.94
800-200-50	0.97	0.92	**400-200-100**	0.97	0.91
800-300-100	0.96	0.91	**500-300-150**	0.97	0.91

Table III shows the worst-case confusion matrix out of the 50 trials for VCO ADC classification using 3-layer ANN, as well as f-1 score for each class. The true positives are shaded in gray and occur along diagonal of the confusion matrix. Chip 1 has the lowest f-1 score for the worst-case classifier performance, and has 7 false negatives, i.e, the classifier predicts 3 instances of chip 1 as chip 2 and 4 instances of chip 1 as chip 3. While both false positives and false negatives reduce classifier performance, from security point of view false negative is better than false positive. Table IV shows the worst case confusion matrix for SAR ADC class. Chip 4 has the lowest f-1 score and has 5 false positives. However, the high number of false positives is due to the reason that for the worst-case confusion matrix shown in Table IV, the training set did not have enough samples of chip 4 to perform classification with high accuracy on the validation set, since the chips 4 and 5 have less VT data points than the other SAR chips.

The feature selection technique is important to develop any reliable machine learning model since it improves performance to classify data and reduces run time to validate the model. To get insight into the machine learning algorithms used for classifying the ADC chips, we computed feature importance scores for the INL features for the SAR ADC. The feature importance scores are computed using neighborhood component analysis (NCA) which measures the average leave-one-out classification error, and shown in Fig. 6. The high feature importance scores correspond to INL values due to mismatch in the LSB capacitors in the SAR DAC. Mismatch in the n-th capacitor leads to a jump in the INL value when the n-th bit in the SAR output makes a transition. Since different chips have different mismatch, their INL signatures will be different and are used by the machine learning model for classification even though the mismatch value is not large enough to degrade SNDR of the ADC. Higher mismatch in LSB capacitors compared to MSB capacitors is due to the fact that the MSB capacitors are placed in a common-centroid fashion, but the LSB capacitors are grouped together at one end of the DAC [10].

To verify the feature importance scores, we used 253 features with the highest feature importance scores to test performance of the 3-layer ANN on the validation dataset. The average accuracy and f-1 score with only the important features are 96.17% and 0.93 respectively, compared to 98.7% and 0.96 when all features are considered, which verifies that the important features alone are enough to classify the ADC chips with high accuracy.

TABLE II: Comparison of performance of different machine learning models for ADC classification

	VCO ADC					
	2-layer ANN		3-layer ANN		k-NN	
	accuracy(%)	f1-score	accuracy(%)	f1-score	accuracy(%)	f1-score
Chip 1	97.09 ± 4.00	0.9021 ± 0.09	95.97 ± 4.00	0.9052 ± 0.05	90.28	0.7132
Chip 2	97.10 ± 2.96	0.9011 ± 0.32	97.56 ± 4.20	0.9398 ± 0.05	90.12	0.7618
Chip 3	97.23 ± 2.90	0.9144 ± 0.08	97.59 ± 4.00	0.9308 ± 0.08	89.12	0.8612
Chip 4	98.05 ± 2.88	0.9274 ± 0.07	98.78 ± 3.04	0.9457 ± 0.07	94.30	0.8131
Chip 5	96.90 ± 4.17	0.9309 ± 0.07	98.81 ± 2.90	0.9450 ± 0.07	94.58	0.8572
Average	97.27 ± 1.60	0.9152 ± 0.08	97.80 ± 2.45	0.9347 ± 0.06	91.28	0.8013
	SAR ADC					
Chip 1	98.13 ± 1.94	0.9427 ± 0.06	97.97 ± 3.28	0.9336 ± 0.08	88.28	0.7412
Chip 2	98.25 ± 1.94	0.9521 ± 0.05	98.23 ± 1.94	0.9400 ± 0.07	91.20	0.8124
Chip 3	97.19 ± 2.63	0.9200 ± 0.08	98.80 ± 1.76	0.9679 ± 0.05	88.60	0.8615
Chip 4	99.31 ± 1.49	0.9624 ± 0.08	99.40 ± 1.35	0.9667 ± 0.08	99.8	0.8762
Chip 5	99.52 ± 1.52	0.9667 ± 0.11	99.90 ± 3.1	0.9933 ± 0.02	92.13	0.8912
Chip 6	99.40 ± 1.35	0.9920 ± 0.02	97.83 ± 2.04	0.9703 ± 0.03	92.50	0.8912
Average	98.63 ± 0.90	0.9560 ± 0.03	98.70 ± 1.21	0.9620 ± 0.04	92.01	0.8456

TABLE III: Worst case confusion matrix for VCO

		Actual					
	chip	**1**	**2**	**3**	**4**	**5**	**f1-score**
Predicted	**1**	10	0	0	0	0	0.7407
	2	3	9	0	3	0	0.7500
	3	4	0	13	0	0	0.8667
	4	0	0	0	9	0	0.8571
	5	0	0	0	0	16	1.0000

TABLE IV: Worst case confusion matrix for SAR

		Actual						
	chip	**1**	**2**	**3**	**4**	**5**	**6**	**f1-score**
Predicted	**1**	18	0	0	0	0	1	0.9730
	2	0	16	0	0	0	3	0.9143
	3	0	0	16	0	0	0	0.8889
	4	0	0	5	7	0	0	0.7778
	5	0	0	0	0	7	0	1.0000
	6	0	0	0	0	0	32	0.9412

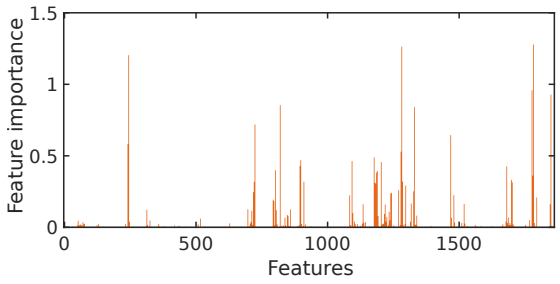


Fig. 6: Feature importance scores for SAR ADC

IV. Conclusion

This work has presented a machine learning approach which can use intrinsic non-linearity of ADCs to identify different samples with high accuracy. The proposed approach is validated by performing classifications on two different ADC types -1) CT over-sampled ADC with ISI error as dominant non-linearity source, and 2) DT Nyquist ADC with static element mismatch as dominant non-linearity source. We have shown that a 3-layer ANN can identify different samples within each ADC class with mean accuracy >95% without increasing design overhead. The proposed technique can be an alternative to weak PUFs for chip authentication and opens up a new research direction for chip verification using machine learning.

References

[1] J. W. Lee, D. Lim, B. Gassend, G. E. Suh, M. Van Dijk, and S. Devadas, "A technique to build a secret key in integrated circuits for identification and authentication applications," in *IEEE Symposium on VLSI Circuits*, 2004, pp. 176–179.

[2] G. E. Suh and S. Devadas, "Physical unclonable functions for device authentication and secret key generation," in *Proceedings of the 44th annual design automation conference*. ACM, 2007, pp. 9–14.

[3] A. B. Alvarez, W. Zhao, and M. Alioto, "Static physically unclonable functions for secure chip identification with 1.9–5.8% native bit instability at 0.6–1 V and 15 fJ/bit in 65 nm," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 3, pp. 763–775, 2016.

[4] X. Xi, H. Zhuang, N. Sun, and M. Orshansky, "Strong subthreshold current array PUF with 2^{65} challenge-response pairs resilient to machine learning attacks in 130nm CMOS," in *IEEE Symposium on VLSI Circuits*, 2017, pp. C268–C269.

[5] A. Herkle, J. Becker, and M. Ortmanns, "Exploiting Weak PUFs From Data Converter Nonlinearity – Eg, A Multibit CT $\Delta\Sigma$ Modulator," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, no. 7, pp. 994–1004, 2016.

[6] A. Jayaraj, A. Das, S. Arcot, and A. Sanyal, "8.6fJ/step VCO-Based CT 2nd-Order $\Delta\Sigma$ ADC," in *IEEE Asian Solid State Circuits Conference (A-SSCC)*, 2019, pp. 197–200.

[7] A. Jayaraj, S. T. Chandrasekaran, A. Ganesh, I. Banerjee, and A. Sanyal, "Maximum Likelihood Estimation Based SAR ADC," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 66, no. 8, pp. 1311–1315, 2019.

[8] B. Bose, B. Broeg, Y. Kwon, and Y. Ashir, "Lee distance and topological properties of k-ary n-cubes," *IEEE Transactions on Computers*, vol. 44, no. 8, pp. 1021–1030, 1995.

[9] F. J. Massey Jr, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[10] L. Chen, A. Sanyal, J. Ma, and N. Sun, "A 24-μW 11-bit 1-MS/s SAR ADC with a bidirectional single-side switching technique," in *IEEE European Solid State Circuits Conference (ESSCIRC)*, 2014, pp. 219–222.