Walsh-domain Neural Network for Power Amplifier Behavioral Modelling and Digital Predistortion

Cel Thys*‡, Rodney Martinez Alonso*, Antoine Lhomel†, Maxandre Fellmann†,
Nathalie Deltimple†, Francois Rivet†, Sofie Pollin*

*WaveCoRE, Department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven, Belgium
*Univ. Bordeaux, CNRS, Bordeaux INP, IMS, UMR 5218, F-33400 Talence, France
‡email: cel.thys@kuleuven.be

Abstract—**This paper investigates the use of Neural Network (NN) nonlinear modelling for Power Amplifier (PA) linearization in the Walsh-Hadamard transceiver architecture. This novel architecture has recently been proposed for ultra-high bandwidth systems to reduce the transceiver power consumption by extensive parallelization of the digital baseband hardware. The parallelization is achieved by replacing two-dimensional quadrature modulation with multi-dimensional Walsh-Hadamard modulation. The open research question for this architecture is whether conventional baseband signal processing algorithms can be similarly parallelized while retaining their performance. A key baseband algorithm, digital predistortion using NN models for PA linearization, will be adapted to the parallel Walsh architecture. A straighforward parallelization of the state-of-the-art NN architecture is extended with a cross-domain Knowledge Distillation pre-training method to achieve linearization performance on par with the quadrature implementation. This result paves the way for the entire baseband processing chain to be adapted into ultra-high bandwidth, low-power Walsh transceivers.**

Index Terms—**digital predistortion (DPD), power amplifier (PA), time-delay neural network (TDNN), Walsh-Hadamard transform**

I. INTRODUCTION

Future wireless networks should evolve beyond human-centric wideband demand, to serve billions of devices simultaneously, while concurrently supporting high-throughput machine-to-machine communication applications like chip-to-chip ultra-high-speed communications [1]. However, spectrum availability and management have become the main bottleneck for the future development of wireless networks. These factors have pushed the development of one of the key enabling technologies for 6G and beyond, ultra-wideband sub-terahertz (sub-THz) communications.

One of the main challenges to enable THz communications is that the transceivers are more affected by non-linearities across the ultra-wideband frequency range [2]. For instance, for an ultra-wideband system with 10 GHz of bandwidth, high-speed data converters with a Nyquist sampling rate higher than 20 Giga-samples-per-second (GSPS) are required. However, state-of-the-art converters for ultrawideband applications underperform in terms of their Walden and Schreier figures of merit [3], [4]. These figures of merit respectively consider the trade-off between sampling rate, power consumption, and effective number of bits, and between sampling rate, power consumption and the signal-to-noise and distortion ratio.

Therefore, the converter´s linearity and energy efficiency for ultrawideband wireless communications is quite limited [5].

To overcome the hardware limitations imposed by the large sampling rate requirements for ultrawideband communications in near-THz spectrum, the Walsh-Hadamard transceiver architecture has been recently proposed [6], [7]. This transceiver architecture splits the ultrawideband communication system into multiple parallel undersampling digital chains. Each chain modulates a distinct Walsh-Hadamard basis function with a certain *sequency*, similar to OFDM subcarriers with their distinct frequencies. Since each chain has a low digital sampling rate, the overall system power consumption is much lower than an equal bandwidth quadrature system [7]. The architecture can be used with any arbitrary waveform, using Walsh-Hadamard series approximation [6]. The Walsh-Hadamard transformation allows mapping time-domain symbols into a multidimensional orthogonal constellation space. Therefore, an N-order Walsh transformation will enable an N-dimensional constellation mapping and the same number of parallel processing chains in the receiver, reducing each ADC's sampling rate by a factor N [8].

Although the Walsh transformation allows relaxing the requirements in terms of energy, linearity and performance of the analog-to-digital converters, it requires entirely new implementations of baseband algorithms, such as digital predistortion (DPD) to compensate power amplifier non-linearities. Indeed, high bandwidth systems require the use of wideband power amplifiers (PA), which have an inherent tradeoff between power efficiency and linearity. The operation region with higher input power will be more power efficient but also exhibit more nonlinear behavior. To improve power efficiency without sacrificing linearity, many have proposed digital predistortion linearization techniques [9]–[17]. In DPD, the communication signal is sent through an inverse behavioral model of the PA (the predistorter) before going to the PA itself, such that the predistorter cancels out the PA's nonlinear distortion.

One class of DPD algorithms uses the linear-in-parameters Volterra series [9]–[12] to implement the predistorter model. The parameters of these models can be identified using a single least-squares fit or iteratively using algorithms such as recursive least squares or least mean squares [9], [11]. Parallelizing Volterra-based DPD has also been investigated,

for instance in [12] authors propose a separate DPD model for each subcarrier in an OFDM system. A Volterra predistorter model in *N*-dimensional parallel Walsh domain has been presented for the first time by authors in [11].

One of the challenges faced by traditional Volterra predistorters is that some predistorted waveforms can have a larger PAPR compared to the signal itself, particularly for OFDM systems [16]. Another limitation of Volterra models is the high correlation between polynomial bases of higher orders, thus limiting the scaling of Volterra models to high model complexities [17]. To overcome these challenges, deep neural network models have been widely investigated for DPD [13]–[17].

The basic neural network (NN) model used for PA behavioral modelling and distortion is the Real-Valued Time-Delay Neural Network (RVTDNN) [13]. In this model, the complex baseband signal is split into real and imaginary components which are fed to a feedforward NN, together with a vector of previous samples. The outputs of the NN are the real and imaginary components of the predistorted waveform, to be sent to the RF frontend. The NN model parameters are identified iteratively using methods based on stochastic gradient descent. The performance of the basic RVTDNN was improved in [14], augmenting it with additional inputs containing envelope terms (e.g. amplitude or amplitude squared). Authors in [16] have investigated a type of TDNN with delayed and advanced input samples, and found that it reduces the PAPR of the predistorted waveform compared to Volterra predistorters. Wu et al. [15] incorporated a skip connection between inputs and outputs of the predistorter neural network, this forces the neural network layers to model additive nonlinearity. They named the resulting model the Residual Real-valued TDNN (R2TDNN).

To the best of our knowledge, the adaptation of R2TDNN models to N-dimensional parallel Walsh domain and the trade-off between the performance indicators of the DPD and the neural network complexity has not been investigated before. Indeed, our preliminary research showed that a DPD based on R2TDNN extended to N-dimensional Walsh domain is not capable of fully linearizing the response of ultrawideband power amplifiers. Although this approach improves the EVM, the spectral regrowth of the PA is not significantly reduced. We hypothesize that during the training for N dimensions, the neural network optimizer loses track of essential features of the time-frequency response of the amplifier as the amplitude and phase components are transformed into a higher dimensional space (i.e., Walsh domain). We propose pre-training of the Walsh-domain neural network predistorter, using the predistortion waveform generated by a large IQ-domain model. This process is known as Knowledge Distillation [18], wherein the IQ-DPD model is the teacher and the Walsh-DPD model the student that learns to mimic the teacher's output. Multiple authors [19], [20] have similarly used offline pre-training with an ideal predistortion waveform acquired using computationally expensive iterative learning control [21].

The main novelty of this paper is an adaptation of R2TDNN DPD models for an N-dimensional parallel Walsh-Hadamard

communication system, without sacrificing the linearization performance. More specifically, the R2TDNN model will be adapted to the Walsh domain by replacing the IQ inputs and outputs with N-dimensional vectors, similar to the process followed in [11] for Volterra models. To precondition the Walsh-domain predistorter towards specific features relevant to time domain DPD, we propose Knowledge Distillation offline pre-training with an IQ-domain DPD acting as teacher model.

The outline of this paper continues as follows. In Section II we detail the proposed neural network for PA behavioral modeling and DPD in the Walsh sequence domain. In Section III we provide our research findings and provide a numeric analysis demonstrating that the forward modelling and linearization performance of the proposed approach is on par with quadrature-based R2TDNN. Finally, our research findings and conclusions are summarized in section IV.

II. PROPOSED WALSH-DOMAIN NEURAL NETWORK

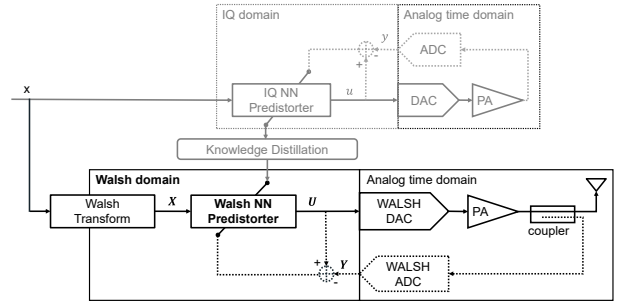The proposed approach to DPD linearization of PAs in Walsh domain is shown in Fig. 1.

Fig. 1: Rationale of proposed Walsh-based Neural Network system enhanced by off-line knowledge distillation pre-training

In this diagram, complex IQ signals are shown in lower case x, while Walsh-domain vectors are shown in bold upper case $\mathbf{X}$. The transmit IQ waveform $x(n)$ is transformed to Walsh-domain, predistorted using a Neural Network model, and then sent towards the PA. The IQ-domain predistorter, a R2TDNN, is drawn in grey since it is only used for offline pre-training of the Walsh-domain predistorter. This process called cross-domain Knowledge Distillation (KD) uses an ideal predistortion waveform generated by the R2TDNN as pre-training data for the Walsh-domain Neural Network (WDNN). This allows the WDNN predistorter to model accurately the ideal predistortion behavior, without specific adaptations in the model architecture.

Whereas conventional DPD operates on IQ signals, the proposed DPD operates on Walsh-domain vectors $\mathbf{X}$, which can be obtained from the respective IQ signal x by means of the Walsh transform

$$\mathbf{X}(i) = \sum_{n=1}^{N} x(n)W_i^N(n) \text{ for } i = 1...N \qquad (1)$$

where $W_i^N(n)$ represents the Walsh-Hadamard basis function with Walsh order N and sequency i [6], [11]. These basis

functions are binary, orthogonal waveforms that decompose the signal $x(n)$ into it's sequency spectrum $\mathbf{X}(i)$. By letting the predistorter operate in the sequency domain, we parallelize the DPD operation by a factor equal to the Walsh order N. The algorithm proposed in [11] exploits this architecture to adapt a block-based Volterra model as predistorter, while our work will investigate the use of a WDNN predistorter.
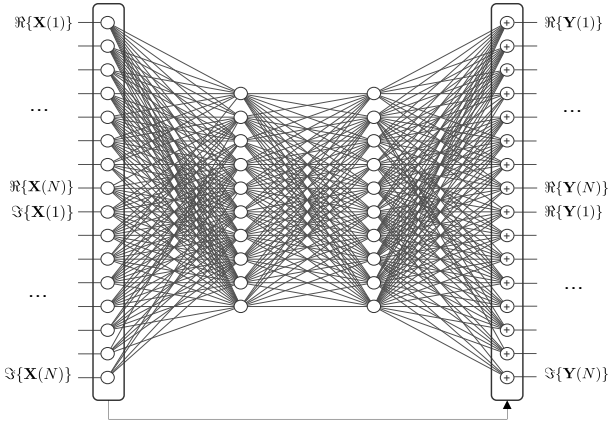

Fig. 2: Proposed Walsh-domain Neural Network

In conventional neural network predistorters, the model architecture is specifically engineered for PA behavioral modelling. This includes a delay line at the NN input containing memory terms [13], further input features representing envelope terms $|x(n)|^k$ [14] and possibly a skip connection between input and output [15]. In our proposed WDDN architecture, the input delay line is replaced with a vector of Walsh coefficients representing the Walsh transform of the complex input. The conventional I and Q outputs are replaced by a vector output of size $2 * N$, corresponding with the sequency domain version of I and Q outputs. The complete model is shown in Fig. 2, including a residual connection between input and output. Although no PA-specific envelope or memory term features are input to the model, it can still function as an accurate PA behavioral model due to our Knowledge Distillation pre-training method.

The neural network behavioral model is trained using the Adam optimizer [22], normalized mean square error (NMSE) loss (3) and maximum 10000 epochs of training. Training data is grouped into training-validation-test sets according to a 60-25-15% split. Hyperparameters of the neural network model are selected using a grid search, where training is stopped early if after 50 epochs there is no improvement in validation loss. We investigated neural networks with hyperbolic tangent and ReLU activations, and similar to [17] we consistently found the tanh outperformed the ReLU activation.

By setting a constraint on the overall number of network weights, we can search for optimal models within a certain model complexity range. We use as model complexity metric the number of Floating-Point-Operations-per-Second (FLOPS) where k, n, f_{symb}, I, O represent the number of hidden layers, amount of neurons per layers, symbol rate and Input / Output

feature size respectively:

$$FLOPS = [2 * n * (I + k * n + O) + O] * f_{symb}. \quad (2)$$

In this equation f_{symb} is 20 Giga-symbols-per-second for IQ-based R2TDNN (1GHz signal bandwidth, 10 times over-sampled). Walsh-based WDNN will have an N times larger I and O, but also N times smaller f_{symb}, so overall similar FLOP numbers as IQ-based neural networks.

The NMSE is used as loss function for NN training as well as the metric to evaluate forward PA modelling accuracy on the test dataset. In (3), $y(n)$ represents the true PA output and $\hat{y}(n)$ represents the NN model estimated output:

$$NMSE_{dB} = 10 \cdot log_{10} \left\{ \frac{\frac{1}{N} \sum_{n=1}^{N} |\hat{y}(n) - y(n)|^2}{\frac{1}{N} \sum_{n=1}^{N} |y(n)|^2} \right\}. \quad (3)$$

To compare NN based predistorters, we use the error vector magnitude (EVM) and adjacent channel leakage ratio (ACLR) as performance metrics. Definitions are given below in (4) and (5), where $x(n)$ is the original transmit signal, $y(n)$ is the output of the combined DPD-PA system and P_{adj}, P_{main} represent adjacent channel and main channel radiated power respectively:

$$EVM_{\%} = 100 * \sqrt{\frac{\frac{1}{N} \sum_{n=1}^{N} |y(n) - x(n)|^2}{\frac{1}{N} \sum_{n=1}^{N} |x(n)|^2}}, \quad (4)$$

$$ACLR = 10 * log10(\frac{P_{adj}}{P_{main}}). \quad (5)$$

For the PA and DPD modelling we considered a waveform with 1 GHz bandwidth and 11.49dB PAPR as input stimulus to a custom-designed CMOS PA operating in D-band. The resulting output waveform is obtained using Cadence post-layout simulation, including the PA memory effects. The custom PA in question has 11 dBm saturation power and 15 dB nominal gain.

III. RESULTS

In this section we compare the Walsh-based WDNN and IQ-based R2TDNN architectures, with the Walsh-domain results always reported using Walsh order $N = 64$. First, we evaluate forward PA modelling accuracy in terms of NMSE. The accuracy/complexity tradeoff in Fig. 3, shows the scaling trend for both models is very similar. Indeed, the difference in accuracy between Walsh and IQ models from 16 TFLOPS onwards is negligible. For lower TFLOP numbers there is a performance gap due to the reduced capacity of the WDNN to model essential time-domain features.

For WDNN-based DPD this issue will be solved by using cross-domain Knowledge Distillation (KD) pre-training, which we validate for the simulated 1 GHz bandwidth waveform.
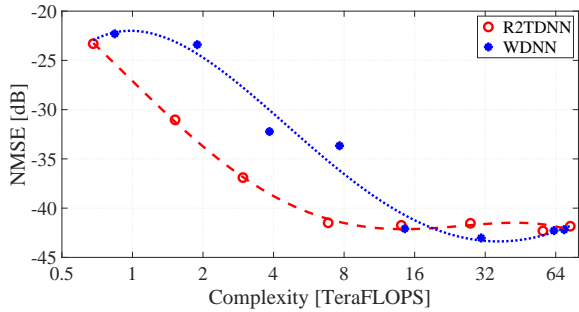
Fig. 3: PA modelling accuracy (NMSE) dependence on model complexity.

Fig. 4 shows the simulated amplitude and phase response of the linearized PA, while Fig. 5 shows the spectrum before and after DPD.
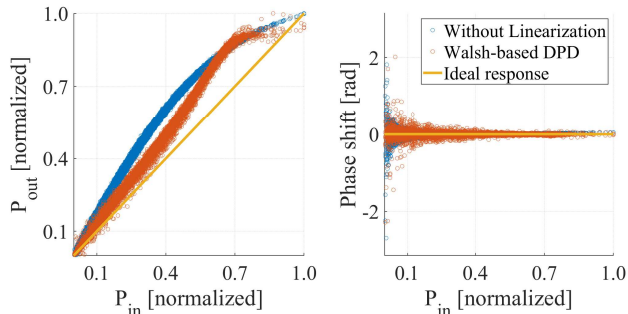

Fig. 4: AM-AM and AM-PM response after applying proposed Walsh-based DPD algorithm
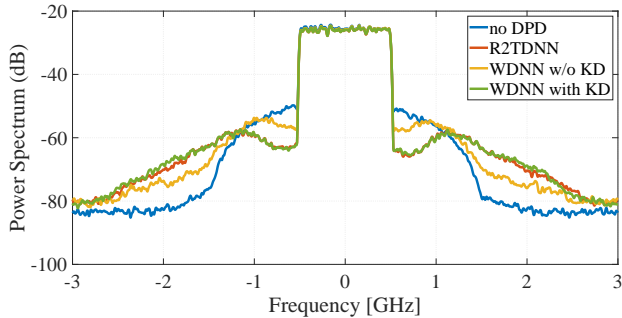

Fig. 5: Spectrum at output of PA after applying considered DPD approaches

Fig. 5 visualizes how the proposed KD pre-training approach improves the Walsh-based DPD. It improves ACLR by approximately 3dB, achieving comparable linearization performance as IQ-DPD, with a difference of less than 0.08dB in ACLR. We refer to Figs. 6 and 7 for a quantitative overview of the linearization performance. Here we show EVM (4) and ACLR (5) evolution as a function of growing model computational complexity (2). In each figure, the baseline PA performance without linearization is shown as a solid line.

The trend on Fig. 7 clearly shows ACLR improvements up to the highest considered complexity, indicating the frequency domain behavior of the PA (the memory effect) requires a higher model complexity. The WDNN-based DPD achieves up to 1.98 dB ACLR improvement after linearization, for model complexity between 16-64 TFLOPS. This is only slightly worse compared to the most complex IQ DPD model. Fig. 6
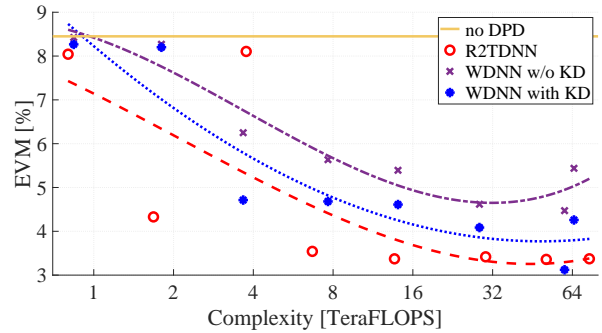

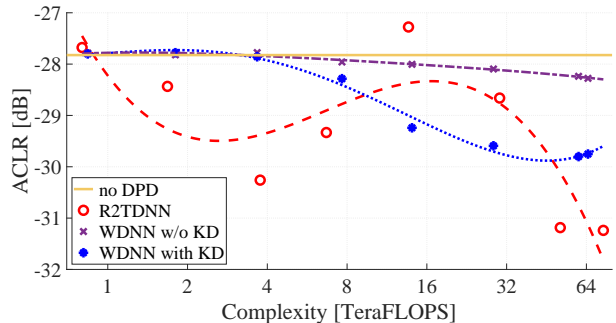Fig. 6: EVM - complexity tradeoff for investigated DPD methods

Fig. 7: ACLR - complexity tradeoff for investigated DPD methods

illustrates that the EVM for both architectures reaches a comparable level for higher model complexities, with a difference between best models of just 0.24%. This demonstrates our Walsh-based DPD model reproduces the time-domain features of the ideal Knowledge-Distillation waveform.

To summarize we show the best achieved PA modelling accuracy and DPD performance in Table I. The table shows Walsh-domain neural network predistortion achieves similar performance as the quadrature equivalent. As the DPD performance metrics are similar, our model allows exploiting the benefits of Walsh architecture for reducing the ADC sampling rate and power consumption for ultrawideband communications.

TABLE I: NN model results summary

System	NMSE [dB]	ACLR [dB]	EVM [%]
PA (no DPD)	/	-27.82	8.45
R2TDNN	-42.31	-31.24	3.36
WDNN	-43.03	-29.80	3.12

IV. CONCLUSION

An adaptation of TDNN models towards Walsh transceiver architectures was proposed, reducing the power consumption of the analog - digital conversion. An offline pre-training method based on knowledge distillation for domain adaptation improves the spectral mask of the proposed approach and achieves similar linearization performance as traditional quadrature architectures. Simulation results confirm these findings. The proposed adaptation of neural network DPD algorithms paves the way for other signal-processing tasks to be incorporated in the same way into the novel Walsh transceiver architecture.

REFERENCES

[1] T. S. Rappaport, Y. Xing, O. Kanhere, S. Ju, A. Madanayake, S. Mandal, A. Alkhateeb, and G. C. Trichopoulos, "Wireless Communications and Applications Above 100 GHz: Opportunities and Challenges for 6G and Beyond," *IEEE Access*, vol. 7, pp. 78 729–78 757, 2019.

[2] B. Che, X. Li, Z. Chen, and Q. He, "A low-complexity transceiver design for terahertz communication based on deep learning," in *2022 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2022, pp. 1159–1164.

[3] B. Murmann, "ADC Performance Survey 1997-2023," [Online]. Available: https://github.com/bmurmann/ADC-survey.

[4] B. Murmann, M. Verhelst, and Y. Manoli, *Analog-to-Information Conversion*. Cham: Springer International Publishing, 2020, pp. 275–292.

[5] G. Manganaro, "An introduction to high sample rate nyquist analog-to-digital converters," *IEEE Open Journal of the Solid-State Circuits Society*, vol. 2, pp. 82–102, 2022.

[6] N. Bouassida, F. Rivet, Y. Deval, D. Duperray, and A. Cathelin, "A concurrent transmitter in cmos 28nm fdsoi technology based on walsh sequences generator," in *2016 14th IEEE International New Circuits and Systems Conference (NEWCAS)*, 2016, pp. 1–4.

[7] P. Ferrer, F. Rivet, H. Lapuyade, and Y. Deval, "A walsh-based arbitrary waveform generator for 5g applications in 28nm fd-soi cmos technology," *IEEE Access*, vol. 11, pp. 117 434–117 442, 2023.

[8] C. Dehos, J.-L. M. Gulfo, D. Lachartre, D. Belot, P. Courouve, and M. Pelissier, "D-band channel aggregation receiver architecture based on if analog processing using digital wavelet," in *2022 29th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, 2022, pp. 1–4.

[9] C. Eun and E. Powers, "A new Volterra predistorter based on the indirect learning architecture," *IEEE Transactions on Signal Processing*, vol. 45, no. 1, pp. 223–227, Jan. 1997.

[10] D. Morgan, Z. Ma, J. Kim, M. Zierdt, and J. Pastalan, "A Generalized Memory Polynomial Model for Digital Predistortion of RF Power Amplifiers," *IEEE Transactions on Signal Processing*, vol. 54, no. 10, pp. 3852–3860, Oct. 2006.

[11] M. Fellmann, F. Rivet, N. Deltimple, E. Kerhervé, H. Lapuyade, and Y. Deval, "A block-based lms using the walsh transform for digital predistortion of power amplifiers," *IEEE Transactions on Communications*, vol. 71, no. 10, pp. 6074–6087, 2023.

[12] A. Brihuega, L. Anttila, and M. Valkama, "Frequency-Domain Digital Predistortion for OFDM," *IEEE Microwave and Wireless Components Letters*, vol. 31, no. 6, pp. 816–818, Jun. 2021.

[13] T. Liu, S. Boumaiza, and F. Ghannouchi, "Dynamic behavioral modeling of 3G power amplifiers using real-valued time-delay neural networks," *IEEE Transactions on Microwave Theory and Techniques*, vol. 52, no. 3, pp. 1025–1033, Mar. 2004.

[14] D. Wang, M. Aziz, M. Helaoui, and F. M. Ghannouchi, "Augmented Real-Valued Time-Delay Neural Network for Compensation of Distortions and Impairments in Wireless Transmitters," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 1, pp. 242–254, Jan. 2019.

[15] Y. Wu, U. Gustavsson, A. G. i. Amat, and H. Wymeersch, "Residual Neural Networks for Digital Predistortion," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, Dec. 2020, pp. 01–06.

[16] T. Gotthans, G. Baudoin, and A. Mbaye, "Digital predistortion with advance/delay neural network and comparison with volterra derived models," in *2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, 2014, pp. 811–815.

[17] R. Hongyo, Y. Egashira, T. M. Hone, and K. Yamaguchi, "Deep neural network-based digital predistorter for doherty power amplifiers," *IEEE Microwave and Wireless Components Letters*, vol. 29, no. 2, pp. 146–148, 2019.

[18] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[19] M. Mengozzi, G. P. Gibiino, A. M. Angelotti, C. Florian, and A. Santarelli, "Beam-dependent active array linearization by global feature-based machine learning," *IEEE Microwave and Wireless Technology Letters*, vol. 33, no. 6, pp. 895–898, 2023.

[20] A. Fischer-Bühner, L. Anttila, M. D. Gomony, and M. Valkama, "Phase-normalized neural network for linearization of rf power amplifiers," *IEEE Microwave and Wireless Technology Letters*, vol. 33, no. 9, pp. 1357–1360, 2023.

[21] J. Chani-Cahuana, P. N. Landin, C. Fager, and T. Eriksson, "Iterative learning control for rf power amplifier linearization," *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 9, pp. 2778–2789, 2016.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.