On LSTM-Based Behavioral Modeling of Radio-Frequency Power Amplifiers with a Small Training Dataset

Ryoki Yamaguchi*, Satoshi Miyata*, Suehiro Shimauchi*, Eiji Mochida†, and Seiji Fujiwara†
* Kanazawa Institute of Technology, Japan
E-mail: c6500850@st.kanazawa-it.ac.jp
† Sumitomo Electric Industries, LTD., Japan

***Abstract*—Accurate modeling of radio-frequency power amplifier (PA) behavior is important for analyzing PA characteristics, developing novel PAs, and implementing digital predistortion (DPD) systems effectively, ultimately contributing to enhanced wireless communication performance. In this study, we examined the performance of neural-network-based PA modeling, particularly in the case of long short-term memory (LSTM)-based modeling, using a relatively small training dataset. The models constructed using various configurations and learning conditions were compared to evaluate performance differences. We found that PA behavior can be modeled more accurately when residual learning and parallel LSTM structures are incorporated. In addition, incorporating both the instantaneous and smoothed amplitudes of the input signal as additional model inputs, alongside the original input signal to the modeled PA, improved performance.**

I. Introduction

As mobile services grow and become more important, improving wireless communication technology has become more crucial. To expand wireless network service areas, radio-frequency (RF) power amplifiers (PAs) are required to manage high-frequency signals while delivering high-output power levels. This requirement causes PAs to operate in a nonlinear region, which potentially compromises the communication quality. Therefore, digital predistortion (DPD) techniques have been developed to compensate for nonlinear distortions by preprocessing PA input signals. DPDs can be achieved by constructing an inverse model of the corresponding target PA. Modeling the forward behavioral characteristics of PAs is also important for analyzing the PA characteristics, developing novel PAs, or verifying the performance of DPD systems.

Memory polynomial (MP)-based PA modeling approaches, such as the generalized memory polynomial (GMP) [1], have been widely employed in both forward and inverse PA modeling. On the other hand, recent advancements in artificial intelligence (AI) technology have introduced various neural network (NN)-based DPDs and behavioral models.

Most NN-based models are built using either convolutional NNs (CNNs) or recurrent NNs (RNNs). CNN-based models [2]–[6] can efficiently capture the nonlinearity of PAs using fewer parameters than fully connected NN (FCNN)-based models [7], [8]. To enable CNN-based models to capture nonlinearity with memory effects [9], it is essential to explicitly include relevant past signal samples as inputs. RNN-based models can be realized in several structures, including FCNN with recurrent loops [10], [11], long short-term memory (LSTM) [11]–[15], bidirectional LSTM (BiLSTM) [11], [16], [17], gated recurrent units (GRU) [11], and just-another-network (JANET) [18]. RNN-based models do not necessarily require explicit past signal samples as input, as they can inherently handle time sequences through their recurrent structures. Models, including both CNN and RNN, can also be constructed [19].

The model structure can be further improved by considering other perspectives. Residual learning approach [6], [20], [21] often improves model learning performance by employing skip connections to bypass specific layers within the models. The model input can be structured in a multichannel form, which includes not only the real and imaginary parts of the complex-valued IQ signal as the target PA input but also incorporates additional features, such as the amplitude and phase of the IQ signal [18], ambient temperature [5], input power back-off (IPBO) [5], or transient smoothed power [14].

In this study, we investigate the accuracy of models under limited training data conditions, with a focus on evaluating the LSTM-based behavioral models for Doherty PAs [22]. NN-based models often exhibit higher accuracy than MP-based models. However, longer training signal data are generally required for NN-based model training than for MP-based model training. For instance, while 10,000 samples are used for an MP-based model training case [23], approximately 300,000 samples are used for an NN-based model training case [24]. We evaluate the LSTM-based models built with various network configurations and learning conditions using a training dataset of approximately 4,000 samples at 122.88 MHz sampling rate. This corresponds to a duration of approximately 33 μs, which is half the length of a single symbol in the LTE signals.

II. LSTM-Based Behavioral Model of RF Power Amplifiers

Here, we focus on LSTM-based PA behavioral models as the primary subject of our study.

979-8-3315-7206-8/25/$31.00

A. Long Short-Term Memory (LSTM)

LSTM [25] was introduced to process time-series data. The configuration of a LSTM layer is illustrated in Fig. 1. At discrete time index t, the input to the LSTM is denoted by $\boldsymbol{x}_t$, which can be in multichannel form. The hidden and cell states are represented by $\boldsymbol{h}_t$ and $\boldsymbol{c}_t$, respectively. The forgetting gate, storage cell, input gate, and output gate are denoted by f, g, i, and o, respectively. The LSTM controls each gate based on the current input $\boldsymbol{x}_t$ and previous hidden state $\boldsymbol{h}_{t-1}$, updates the cell state $\boldsymbol{c}_t$, and generates a new hidden state $\boldsymbol{h}_t$. The hidden state $\boldsymbol{h}_t$ contains information regarding past input signals, which can be valuable for modeling systems with memory effects.

B. LSTM-Based PA Model Configuration

We consider an LSTM-based PA behavioral model, as illustrated in Fig. 2. The primary inputs for the model are the real part $x_r(t)$ and the imaginary part $x_i(t)$ of the complex-valued IQ signal $x(t)$. The model outputs two channel signals, $v_r(t)$ and $v_i(t)$, as the real and imaginary parts of a complex signal, respectively. The output signals $v_r(t)$ and $v_i(t)$ can directly be estimates of the real and imaginary parts of a target PA outputs, y_r and y_i,

$$\hat{y}_r(t) = v_r(t), \quad \hat{y}_i = v_i(t). \tag{1}$$

The skip connections in Fig. 2 allow inputs to bypass directly to outputs, which are utilized as an option in residual learning. The residual learning refers to modeling only the difference between the input and output signals. Here, the predicted real part $\hat{y}_r(t)$ and imaginary part $\hat{y}_i(t)$ of the PA output signal are generated as

$$\hat{y}_r(t) = v_r(t) + x_r(t), \quad \hat{y}_i(t) = v_i(t) + x_i(t). \tag{2}$$

In addition, optional inputs can be added, such as instantaneous amplitude $|x(t)|$ or smoothed amplitude $\overline{|x(t)|}$. The instantaneous amplitude $|x(t)|$ is obtained as

$$|x(t)| = \sqrt{x_r(t)^2 + x_i(t)^2}. \tag{3}$$

By referring to [14], the smoothed amplitude $\overline{|x(t)|}$ can be calculated as

$$\overline{|x(t)|} = \alpha * \overline{|x(t-1)|} + (1-\alpha)|x(t)|, \tag{4}$$

where the smoothing factor α indicates a time constant. We set $\alpha = 0.25$, which is significantly smaller than that used in [14], to prioritize the modeling of the PA transistor response time over that of the transient temperature characteristics associated with the signal power level changes examined in [14].

Figs. 3 and 4 shows the detailed structures of the LSTM-based model. Both figures use a multi-channel signal $\boldsymbol{x}_t$ as input and output complex signals $v_r(t)$ and $v_i(t)$, where the multichannel signal $\boldsymbol{x}_t$ can include the instantaneous amplitude $|x(t)|$ and/or the smoothed amplitude $\overline{|x(t)|}$. The structure shown in Fig. 3 predicts the real part $v_r(t)$ and the imaginary part $v_i(t)$ from the shared hidden state $\boldsymbol{h}_t$ of a single LSTM through the parallel fully-connected (FC) layers. In contrast, the structure shown in Fig. 4 has two LSTM layers in parallel and predicts the real part $v_r(t)$ from the hidden state $\boldsymbol{h}_t^r$ of the first LSTM layer and predicts the imaginary part $v_i(t)$ from $\boldsymbol{h}_t^i$ of the second LSTM layer, respectively. The number of hidden units in each of the two LSTM layers shown in Fig. 4, can be smaller than that in the single LSTM layer in Fig. 3, as each of the two LSTM layers in Fig. 4 independently predicts the real part $v_r(t)$ or the imaginary part $v_i(t)$. Thus, each FC layer in Fig. 4 may also have fewer parameters than the FC layer in Fig. 3.

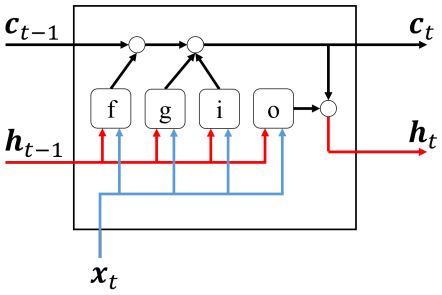

Fig. 1. Configuration of LSTM layer.

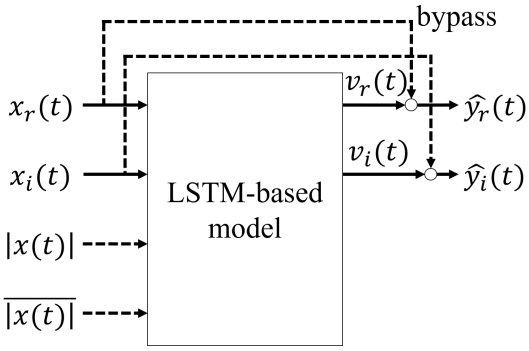

Fig. 2. Configuration of LSTM-based PA behavioral model, where the dashed lines indicate optional skip connections for residual learning.

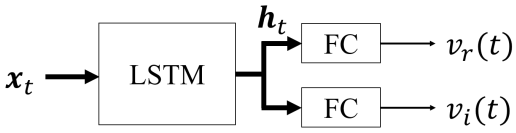

Fig. 3. Single LSTM layer structure.

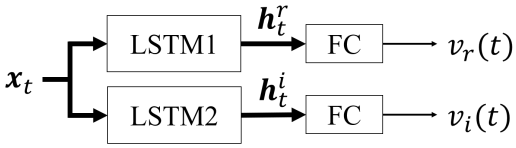

Fig. 4. Parallel dual LSTM layer structure.

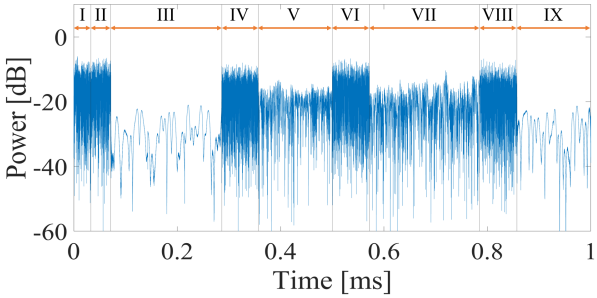

Fig. 5. Input signal used in learning (I) and testing (II–IX).

III. Simulations

Here, we examine the performance of several models built under various network configurations and learning conditions using a small training dataset of approximately 4,000 samples. First, we confirm the effectiveness of the combination of residual learning and dual LSTM layer structure. Subsequently, we compare the results obtained when the model input consists of a two-channel signal with the real and imaginary parts of the IQ signal and when the absolute value and/or the smoothed absolute value of the IQ signal are combined with this two-channel signal. Finally, we evaluate the generalization performance of the models across various initial learning rates using a test dataset comprising multiple sections, some of which have properties distinct from those of the training data.

A. Simulation Conditions

The target PA was a prototyped Doherty-type PA [22] consisting of 17 W GaN FET as a main amplifier and 34 W GaN FET as a peak amplifier. The signal shown in Fig. 5, which is the first 1 ms duration of LTE 20MHz E-UTRA Test Model 2a (E-TM2a) [26], was input to the Doherty PA at gain level of +33 dBm, and its output was measured at sampling rate of 122.88 MHz. As shown in Fig. 5, the input signal is separated in sections I–IX, in which the signal has different characteristics. The signal in section I, approximately 4,000 samples in length (33 μs at a 122.88 MHz sampling rate), was used for training. The number of epochs was set as 90. The learning rate η was initially set to an initial value η_0, which was reduced by 0.5 per 10 epochs. The signals in sections II–XI were used for testing. For evaluation, the normalized mean square error (NMSE) was calculated as

$$\mathrm{NMSE} = 10\log_{10}\frac{\sum|y-\hat{y}|^2}{\sum|y|^2} \quad [\mathrm{dB}], \tag{5}$$

where y is the complex-valued output of the target PA and $\hat{y}$ is the predicted output. The NMSE values were evaluated after averaging the results of the five training trials for each condition.

B. Residual Learning and LSTM-Layer Structure

The NMSE performances with and without residual learning in single and dual LSTM layer structures are shown in Table I and Fig. 6, where $\eta_0 = 0.01$. The single LSTM layer in Fig. 3 had 70 hidden units while each of the dual LSTM layers in Fig. 4 had 35 hidden units. These results were obtained by inputting a complex IQ signal only. Table I shows the NMSEs obtained from the models trained for 90 epochs. This table shows that a significant reduction in the NMSE of approximately 9 [dB] was achieved with residual learning. Fig. 6 shows the convergence of the NMSE, with markers plotted every 10 epochs. The results indicate that the dual LSTM structure also contributed to the improvement in the NMSE convergence speed, although the residual learning contributed significantly.

TABLE I
NMSEs obtained from different model structures.

	NMSE [dB]	
	w/o Residual learning	w/ Residual learning
Single LSTM	−25.30	−35.01
Dual LSTM	−25.90	**−35.04**

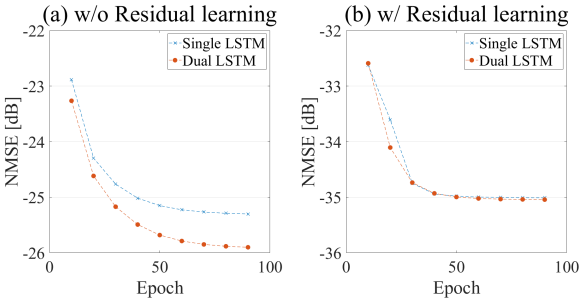

Fig. 6. NMSE convergence.

C. Input Extension

For the model with residual learning and the dual LSTM structure, we examined the effect of optional inputs on the model performance. The results obtained with different input options are shown in Table II, where we compared four cases: $\boldsymbol{x}_t = [x(t)]$, $\boldsymbol{x}_t = [x(t), |x(t)|]$, $\boldsymbol{x}_t = [x(t), \overline{|x(t)|}]$, and $\boldsymbol{x}_t = [x(t), |x(t)|, \overline{|x(t)|}]$. All the models with different input options were obtained by iterative training of 90 epochs with the initial learning rate $\eta_0 = 0.01$. Table II shows that the NMSE was the lowest when both instantaneous and smoothed amplitudes were added to the input.

D. Generalization Performances

From the above results, we focused on the model that employs residual learning and dual LSTM structures with input $\boldsymbol{x}_t = [x(t), |x(t)|, \overline{|x(t)|}]$. To evaluate the generalization performance of the model, the NMSE performance for individual sections I–XI of the test signal depicted in Fig. 5 was assessed by varying the initial learning rate η_0 to values of 0.002, 0.005, 0.01, 0.02, 0.05, and 0.1. The results are shown in Table III. Note that section I was used for training and 'total' means the NMSE for whole the sections II–IX, but excluding section I. For sections II, IV, VI, and VIII, where the input signal level was relatively high, the NMSE values were

TABLE II
NMSEs OBTAINED WITH DIFFERENT INPUT OPTIONS.

LSTM input	NMSE [dB]
$\boldsymbol{x}_t = [x(t)]$	-35.04
$\boldsymbol{x}_t = [x(t), \lvert x(t)\rvert]$	-35.00
$\boldsymbol{x}_t = [x(t), \overline{\lvert x(t)\rvert}]$	-34.65
$\boldsymbol{x}_t = [x(t), \lvert x(t)\rvert, \overline{\lvert x(t)\rvert}]$	$\mathbf{-35.34}$

TABLE III
NMSEs OF INDIVIDUAL SIGNAL SECTIONS WITH DIFFERENT η_0.

		η_0					
		0.002	0.005	0.01	0.02	0.05	0.1
NMSE [dB]	I	-32.75	-38.21	-40.65	-41.62	$\mathbf{-41.94}$	-41.81
	II	-31.97	-37.23	-39.63	-40.68	$\mathbf{-41.05}$	-40.85
	III	-24.91	$\mathbf{-26.81}$	-25.64	-24.35	-23.68	-23.46
	IV	-31.27	-36.56	-39.03	-40.11	-40.53	$\mathbf{-40.58}$
	V	-32.03	-35.49	$\mathbf{-35.51}$	-34.96	-34.55	-34.36
	VI	-32.15	-37.35	-39.70	-40.65	$\mathbf{-41.04}$	-40.97
	VII	-31.56	-35.66	$\mathbf{-35.99}$	-35.36	-34.99	-34.73
	VIII	-31.55	-36.79	-39.18	-40.15	$\mathbf{-40.49}$	-40.42
	IX	-25.17	$\mathbf{-27.54}$	-26.25	-24.84	-24.13	-23.88
	total	-30.93	-34.94	$\mathbf{-35.34}$	-34.72	-34.29	-34.08

reduced when using larger values of η_0, such as 0.05 and 0.1. For sections III, V, VII, and IX, where the input signal level was relative low, the NMSE values were reduced when using smaller values of η_0, such as 0.005 and 0.01. These results suggest that the initial learning rate should be set according to the input signal level, i.e., using a higher value for high signal levels and a lower value for low signal levels. Fig. 7 shows an example of the spectral prediction error for the test signal in section VI among the five training trials with $\eta_0 = 0.05$. In the figure, the four spectra represent the PA input, physically measured PA output, predicted PA output, and error between the measured and predicted outputs, respectively, where the NMSE reached -41.25 dB. In summary, the NMSE values for inputs with higher signal levels tended to be lower, resulting in a higher modeling accuracy. In practical service operations, the signal level rises as the number of multiplexed signal channels increases. Consequently, the results of this experiment hold significant relevance in scenarios involving high-level signal inputs.

IV. CONCLISIONS

In this study, we examined NN-based PA modeling, particularly LSTM-based modeling, using a relatively small training dataset of about 4,000 samples. Simulation results demonstrated that incorporating residual learning and parallel LSTM structures significantly reduced the NMSE. Additionally, incorporating both the signal amplitude and smoothed amplitude as supplementary inputs, alongside the original input signal to the modeled PA, can also improve the performance. By changing the value of the initial learning rate for model learning, we found that a higher initial learning rate improved the modeling accuracy for inputs with high-level signal amplitudes, whereas a lower initial learning rate rather improved the modeling accuracy for inputs with low-level signal amplitudes.

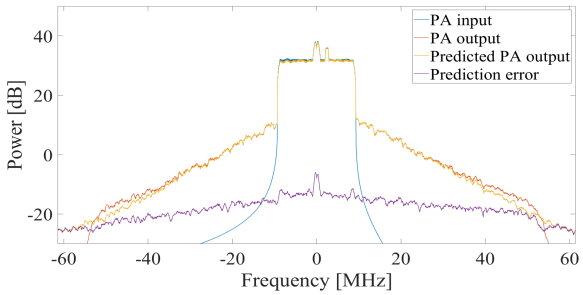

Fig. 7. Example of spectral prediction error (section VI with $\eta_0 = 0.05$).

REFERENCES

[1] 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, 2006.

[2] P. Jaraut, A. Abdelhafiz, H. Chenini, *et al.*, "Augmented Convolutional Neural Network for Behavioral Modeling and Digital Predistortion of Concurrent Multiband Power Amplifiers," *IEEE Transactions on Microwave Theory and Techniques*, vol. 69, no. 9, pp. 4142–4156, 2021.

[3] X. Hu, Z. Liu, X. Yu, *et al.*, "Convolutional Neural Network for Behavioral Modeling and Predistortion of Wideband Power Amplifiers," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 8, pp. 3923–3937, 2022.

[4] Z. Liu, X. Hu, L. Xu, W. Wang, and F. M. Ghannouchi, "Low Computational Complexity Digital Predistortion Based on Convolutional Neural Network for Wideband Power Amplifiers," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 3, pp. 1702–1706, 2022.

[5] A. Motaqi, M. Helaoui, N. Boulejfen, W. Chen, and F. M. Ghannouchi, "Artificial Intelligence-Based Power-Temperature Inclusive Digital Predistortion," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 12, pp. 13 872–13 880, 2022.

[6] Y. Wang, C. Yu, Z. Xu, *et al.*, "Residual Real-Valued Time-Delay Convolutional Neural Network for Behavioral Modeling and Digital Predistortion," in *2024 International Conference on Microwave and Millimeter Wave Technology (ICMMT)*, vol. 1, 2024, pp. 1–3.

[7] M. Rawat, K. Rawat, and F. M. Ghannouchi, "Adaptive Digital Predistortion of Wireless Power Amplifiers/Transmitters Using Dynamic Real-Valued Focused Time-Delay Line Neural Networks," *IEEE Transactions*

on Microwave Theory and Techniques, vol. 58, no. 1, pp. 95–104, 2010.

[8] 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.

[9] J. Pedro, P. Tomé, T. Cunha, *et al.*, "A Review of Memory Effects in AlGaN/GaN HEMT Based RF PAs," in *2021 IEEE MTT-S International Wireless Symposium (IWS)*, 2021, pp. 1–3.

[10] S. Yan, C. Zhang, and Q.-J. Zhang, "Recurrent neural network technique for behavioral modeling of power amplifier with memory effects," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 25, no. 4, pp. 289–298, 2015.

[11] Z. He and F. Tong, "Residual RNN Models With Pruning for Digital Predistortion of RF Power Amplifiers," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 9, pp. 9735–9750, 2022.

[12] D. Phartiyal and M. Rawat, "LSTM-Deep Neural Networks based Predistortion Linearizer for High Power Amplifiers," in *2019 National Conference on Communications (NCC)*, 2019, pp. 1–5.

[13] M. Geng, G. Crupi, and J. Cai, "Accurate and Effective Nonlinear Behavioral Modeling of a 10-W GaN HEMT Based on LSTM Neural Networks," *IEEE Access*, vol. 11, pp. 27 267–27 279, 2023.

[14] S. Miyata, S. Shimauchi, S. Fujiwara, E. Mochida, T. Kawasaki, and M. Tanomura, "Transient Temperature Variation-Dependent RF Power Amplifier Model Based on LSTM," in *2023 Asia-Pacific Microwave Conference (APMC)*, 2023, pp. 390–392.

[15] Z. H. Chang, Y. Zhang, and H. C. Chen, "Dynamic LSTM for 5G Signal Power Amplifier Behavioral Model," *IEEE Microwave and Wireless Technology Letters*, vol. 34, no. 6, pp. 683–686, 2024.

[16] R. Kudupudi, F. L. Pour, D. S. Ha, S. S. Ha, and K. Ramezanpour, "Analysis of Deep Learning Models Towards High Performance Digital Predistortion for RF Power Amplifiers," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2022, pp. 1309–1313.

[17] R. Su, J. Wang, G. Xu, and T. Liu, "Attention mechanism based bidirectional LSTM model for broadband power amplifier linearization," *Electronics Letters*, vol. 59, no. 14, e12869, 2023.

[18] T. Kobal, Y. Li, X. Wang, and A. Zhu, "Digital Predistortion of RF Power Amplifiers With Phase-Gated Recurrent Neural Networks," *IEEE Transactions on Microwave Theory and Techniques*, vol. 70, no. 6, pp. 3291–3299, 2022.

[19] R. Li, Z. Yao, Y. Wang, *et al.*, "Behavioral modeling of power amplifiers leveraging multi-channel convolutional long short-term deep neural network," *IEEE Transactions on Vehicular Technology*, vol. 74, no. 7, pp. 11 456–11 460, 2025.

[20] 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*, 2020, pp. 01–06.

[21] Y. Suo, W. Qiao, C. Jiang, B. Zhang, and F. Liu, "A Residual-Fitting Modeling Method for Digital Predistortion of Broadband Power Amplifiers," *IEEE Microwave and Wireless Components Letters*, vol. 32, no. 9, pp. 1115–1118, 2022.

[22] N. Watanabe, J. Wong, A. Grebennikov, and A. Akiyama, "High-Power High-Efficiency Asymmetric Doherty Amplifiers for Base Station Applications," in *2018 Asia-Pacific Microwave Conference (APMC)*, 2018, pp. 989–991.

[23] A. Abdelhafiz, A. Kwan, O. Hammi, and F. M. Ghannouchi, "Digital Predistortion of LTE-A Power Amplifiers Using Compressed-Sampling-Based Unstructured Pruning of Volterra Series," *IEEE Transactions on Microwave Theory and Techniques*, vol. 62, no. 11, pp. 2583–2593, 2014.

[24] Y. Wu, A. Li, M. Beikmirza, *et al.*, "MP-DPD: Low-Complexity Mixed-Precision Neural Networks for Energy-Efficient Digital Predistortion of Wideband Power Amplifiers," *IEEE Microwave and Wireless Technology Letters*, vol. 34, no. 6, pp. 817–820, 2024.

[25] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) conformance testing," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.141, Jul. 2020, Version 16.6.0.