Sparsely Gated Mixture of Experts Neural Network For Linearization of RF Power Amplifiers

Arne Fischer-Bühner, *Graduate Student Member, IEEE*, Alberto Brihuega, Lauri Anttila, *Member, IEEE*, Matias Turunen, Vishnu Unnikrishnan, *Member, IEEE*, Manil Dev Gomony, *Member, IEEE*, and Mikko Valkama, *Fellow, IEEE*

***Abstract*— This article presents a piecewise neural network (NN) with dynamic sparsity for modeling and linearization of radio frequency (RF) power amplifiers (PAs). A mixture of experts NN (MENN) approach is employed to combine several smaller real-valued time-delay NNs (RVTDNNs) by means of a gating NN. Furthermore, we complement the MENN framework with top-K sparse gating, such that only a subset of experts is activated during each sample inference, reducing the computational complexity at run time. An end-to-end training approach is presented, to optimize the gating alongside with specializing the expert NNs, enabling the experts to collaborate. We experimentally investigate the scaleability of the proposed model in terms of modeling accuracy and linearization performance, as well as run time and model complexity, using RF measurements with two different gallium-nitride Doherty PAs at 1.8 and 3.5 GHz, respectively. Our experiments confirm a significant reduction in run-time complexity due to the sparse gating, with only a small penalty on accuracy, linearization capability and scaleability. Furthermore, the proposed approach is shown to offer favorable complexity-performance trade-offs, outperforming the existing state-of-the-art.**

***Index Terms*— Behavioral modeling, digital predistortion, linearization, mixture of experts neural network (MENN), power efficiency, radio frequency (RF) power amplifiers (PAs), time-delay neural network (NN).**

I. Introduction

THE radio frequency (RF) power amplifier (PA) has been, and continues to be, a major bottleneck for power efficiency and linear transmission in wireless communication systems [1]. Modern, spectrally efficient waveforms with a nonconstant envelope require RF PAs to operate in the nonlinear region in order to be efficient. Although the PA linearity-to-efficiency trade-off has been revisited many times throughout past decades [2], the continuing trends in 5G new radio (NR) and 6G toward higher and higher transmission rates, wider bandwidths, and high peak-to-average power ratio (PAPR) waveforms demand research on improved solutions to the linearity versus efficiency challenge [1], [3].

When operating with high PAPR waveforms, the average output power of the PA has to be significantly backed off from saturation to avoid signal degradation, which compromises the efficiency. Advanced PA architectures and technologies have been developed to improve the power efficiency of PAs, such as the Doherty PA [4] and the load-modulated-balanced PA [5], [6]. Furthermore, gallium nitride (GaN)-based PAs are promising significant efficiency gains, however, at the cost of severe short- and long-term memory effects [7], [8]. Nevertheless, efficient operation of the PA inevitably causes nonlinear distortion, while today's wideband waveforms also excite significant dynamic distortion. Thus, given the stringent linearity requirements defined in, e.g., the 5G NR specifications [9], compensation techniques are necessary to ensure linear amplification, wherein digital predistortion (DPD) is the most capable and established approach [2]. The fundamental concept is to induce artificial nonlinear distortion, inverse to the distortion imposed by the PA, in order to cancel out the nonlinear and dynamic effects of the PA and thereon achieve linearized RF transmission.

Accurate and flexible models describing the inverse nonlinear behavior of PAs are essential for DPD, with Volterra-based polynomial models, such as the generalized memory polynomial (GMP) [10], being commonly used. However, global polynomials like the GMP can have limited capability to accurately describe the distinct behavior of contemporary PAs. Furthermore, scaling the accuracy of global polynomial models by adding more terms is known to have limited potential [11]. Inspired by above, piecewise models have been proposed to partition the modeling space into sub-regions, which can then be modeled locally using, e.g., GMPs. The vector switched GMP (VS-GMP) model was proposed in [11], where multiple GMP models are switched based on the input signal amplitude, yielding a hard partitioning of the modeling space. The decomposed piecewise GMP (DPW-GMP) was

Manuscript received 19 July 2023; revised 18 October 2023; accepted 16 November 2023. This work was supported in part by the European Union's Horizon 2020 Research and Innovation Program through the Marie Skłodowska-Curie under Grant 860921; and in part by the Academy of Finland under Grant 319994, Grant 338224, Grant 332361, Grant 351235, and Grant 345654. This article is an extended version from the 2022 IEEE MTT-S International Microwave Symposium (IMS-2022) [29]. *(Corresponding author: Arne Fischer-Bühner.)*

Arne Fischer-Bühner is with Nokia Bell Labs, 2018 Antwerp, Belgium, and also with the Department of Electrical Engineering, Tampere University, 33720 Tampere, Finland (e-mail: arne.fischer@nokia.com).

Alberto Brihuega is with Nokia Mobile Networks, 90650 Oulu, Finland.

Lauri Anttila, Matias Turunen, Vishnu Unnikrishnan, and Mikko Valkama are with the Department of Electrical Engineering, Tampere University, 33720 Tampere, Finland.

Manil Dev Gomony is with Nokia Bell Labs, 2018 Antwerp, Belgium, and also with the Department of Electrical Engineering, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMTT.2023.3341616.

Digital Object Identifier 10.1109/TMTT.2023.3341616

reported in [12], where the input to the model is decomposed into multiple subsignals by threshold, with a separate GMP model applied to each. However, hard partitioning may lead to discontinuity when combining the sub-models. In [13], the mixture of experts (ME) framework was utilized for soft-partitioning based on a probabilistic scheme that allows multiple local GMPs to be overlapped, and thus improves continuity amongst the different submodels. However, those piecewise models still rely on Volterra-based polynomials, and the underlying problem of a limited scalability remains.

Due to their excellent nonlinear modeling capability and high generality, artificial neural networks (NNs) have been considered as an alternative for PA modeling and DPD [14], [15], [16], [17], [18], [19], [20], [21], [22]. While the complexity of NNs in terms of parameter count and training effort is usually large, they allow scaling the modeling capabilities further, as NNs do not suffer from ill-conditioning for increased model sizes. Furthermore, NNs are more general models, allowing to map more sophisticated nonlinear dynamic behavior. In DPD context, although complex-valued NNs are possible in principle, they are considered expensive to train [23]. Thus, the real-valued time-delay NN (RVTDNN) was proposed for PA behavioral modeling and DPD in [15], [16], and [17], where the in-phase and quadrature-phase (I/Q) components of the baseband signals are fed separately, in parallel, to a single real-valued NN. This approach has then been adopted for DPD in several scenarios [14], [18], [19]. The work in [24] addresses the high generality and complexity of the RVTDNN by restricting the NN to only physically meaningful contributions. As an alternative, the vector decomposition-based RVTDNN was proposed in [20] which operates on the signal's envelope only, and recovers the phase information at a later stage, reducing the overall complexity. However, the involved linear phase recovery limits the overall modeling capability. Furthermore, models based on recurrent NNs have been proposed with particular aim on improving the modeling of memory effects in [21], [22], and [25], however, typically at the cost of increased training and convergence time [18]. Additionally, generalized NN models for coping with different transmission configurations have been explored, e.g., in [26], [27], and [28]. In [28], a dynamically gated NN was proposed to switch and combine various NN submodels based on the transmission configuration. The gating mechanism has conceptual similarity to the methods proposed in this article, however, the aim in [28] is to diversify the NN model for different transmission configurations instead of pushing for enhanced linearization capabilities and improved performance-complexity trade-offs in a given transmit scenario.

In [29], aiming to provide a more flexible neural-network architecture, the MENN was initially introduced for behavioral modeling, inspired by the ME framework. The ME is a probabilistic framework for soft-combining of several specialized expert models, originally applied in the context of behavioral PA modeling and DPD in [13]. A gating unit was employed to provide Gaussian distributed probabilities, which served as a basis for forming a joint model output from several polynomial model experts. The work in [29] applied the ME concept

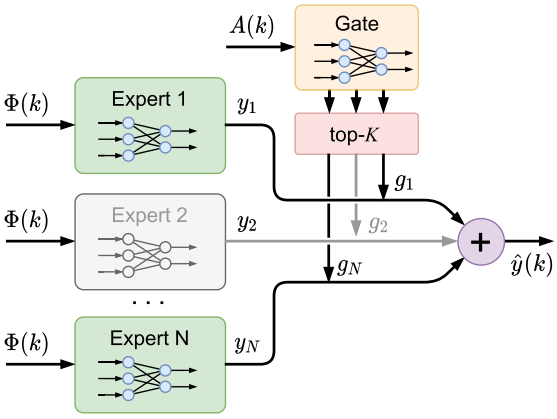


Fig. 1. Illustration of the proposed MENN, with multiple smaller size expert NNs, which are combined by sparse gating based on the input envelope $A(k)$.

of [13] to RVTDNNs [17]. NN experts were used in place of the polynomial model experts and NN-enabled gating was introduced, resulting in non-Gaussian distributed probabilities for combining of the experts. Furthermore, the original training approach of [13], where the training was altered between updating the gate and the experts, was replaced by an end-to-end NN training approach, where partitioning and gating are obtained alongside with training the experts.

In this article, we extend our initial work in [29] and develop the MENN approach further by introducing sparse top-K gating to the MENN processing system as illustrated in Fig. 1. Sparse gating has originally been applied in the context of large-scale NNs as a way reduce the computational cost [30], [31]. The top-K gating allows to dynamically select the K highest rated experts while disabling the remaining experts of the MENN, thus reducing the processing workload during the inference phase, i.e., the run-time complexity. We translate this approach to the DPD context and demonstrate potential to achieve a high degree of linearity with reasonable run-time complexity.

The main technical contributions and novelty of the article can be summarized as follows.

1) We revisit the MENN of [29] and provide additional theoretical foundation while also demonstrating the applicability of MENN in the context of PA linearization.
2) We develop the MENN further and propose the K-MENN approach, which adds top-K sparse gating to reduce the model's processing complexity during inference by dynamically disabling experts.
3) An end-to-end training scheme is proposed, with additions specific to the K-MENN with sparse gating. Further, we describe specific modifications to enable end-to-end training for top-1 selective gating, where only a single expert is selected for each sample inference
4) We consider two different measures of complexity, run-time and model complexity, for assessing the performance-complexity trade-offs of the proposed MENN and K-MENN models, assuming an applicable hardware implementation.

5) The MENN and K-MENN are evaluated in the context of PA behavioral modeling using measured data from a GaN Doherty PA, to showcase the capabilities of the MENN and the newly proposed K-MENN, while also discussing the trade-off of run-time complexity, model accuracy, and model parameter count. We also compare the K-MENN and MENN to the RVTDNN and various state-of-the-art piecewise polynomial models with respect to their complexity.
6) We provide extensive RF measurement results using two different GaN Doherty PAs, to assess and demonstrate the MENN and K-MENN in the context of PA linearization. The linearization performances of the NN models as well as those of the various reference methods are evaluated and compared.

The remainder of the article is organized as follows. The proposed MENN and K-MENN structures and the underlying methods are described in Section II. Next, in Section III, the end-to-end training schemes are detailed for the different considered MENN variants. Measures of model complexity are introduced, assessed and discussed in Section IV. Section V presents the experimental results for behavioral modeling and PA linearization. Finally, Section VI concludes the article.

II. Proposed MENNs

Let $x(k)$ and $\hat{y}(k)$ represent the complex-valued model input and output samples with index k, respectively, with $\hat{y}(k)$ referring to an approximation of the PA output signal $y(k)$. For Volterra-based polynomial models, such as the GMP [10], the nonlinear modeling capability is provided through a set of nonlinear regressors $\tilde{\Phi}(k) \in \mathbb{C}^{1\times W}$, corresponding to $x(k)$, which allows writing the model output as a linear combination of a set of nonlinear basis functions with complex-valued weights $\tilde{\boldsymbol{a}} \in \mathbb{C}^{W\times 1}$ as

$$\hat{y}(k) = \tilde{\Phi}(k)\tilde{\boldsymbol{a}}. \tag{1}$$

These type of models are linear-in-parameters, thus facilitating parameter identification using linear regression techniques, such as ordinary least squares, or gradient-based techniques like the least-mean-squares algorithm. The modeling capabilities are generally limited by the available nonlinear regressors, while increasing the number of regressors makes the associated least-squares problem easily ill-conditioned [32].

A. Real-Valued Time-Delay Neural Network

In contrast to linear-in-parameter models, NN models are inherently nonlinear, i.e., the linear weighting of input regressors is replaced by a nonlinear transformation Ξ performed by a network χ of nonlinear nodes and a set of weights $\boldsymbol{a}$, expressed as

$$\hat{\boldsymbol{y}}(k) = \Xi_{\chi}(\Phi(k)|\boldsymbol{a}). \tag{2}$$

Here, we consider a feed-forward, fully connected RVTDNN with additional augmented terms, as described for example in [17]. The network, shown in Fig. 2, generally consists of an input layer, one or multiple hidden layers (HLs) with nonlinear nodes, and a linear output layer.

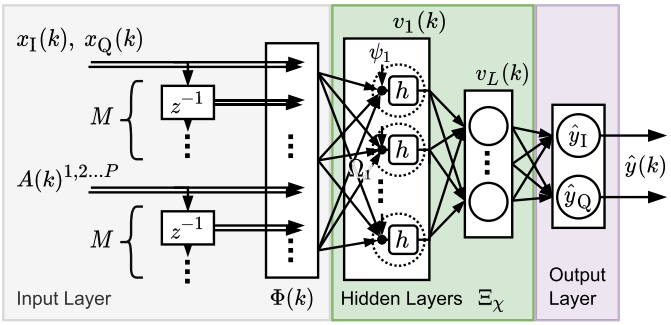


Fig. 2. Block diagram of a RVTDNN. A delay line provides the current and past I/Q and envelope samples to a set of non-linear HLs. The two outputs form the complex-valued model output.

At the NN input layer, similar to linear-in-parameter models, a set of regressors, $\Phi(k)$, is provided. The input layer comprises the current input sample, separated into real-valued I and Q components as $x(k) = x_{\mathrm{I}}(k) + jx_{\mathrm{Q}}(k)$. To enable the NN to also model dynamic effects originating from wideband waveforms, past samples $x_{\mathrm{I}}(k-m), x_{\mathrm{Q}}(k-m)$ with delays $m = 1, 2, \ldots, M$ are provided as additional parallel inputs. Since the PA-induced nonlinearity occurs at RF and thus primarily depends on the input envelope $A(k) = |x(k)|$, the instantaneous and delayed versions of the envelope, as well as powers thereof, are also provided to the NN, expressed as $A(k-m)^p$ with $p = 1, 2, \ldots, P$. Although the NN is inherently capable of creating a nonlinear mapping, these pth order envelope terms have been reported to improve the accuracy for smaller NN sizes by providing an additional nonlinear reservoir to the NN input [17]. Consequently, the input layer provides b_0 input features, expressed formally as

$$\begin{aligned}\Phi(k) = \big[&x_{\mathrm{I}}(k), x_{\mathrm{Q}}(k), \ldots, x_{\mathrm{I}}(k-M), x_{\mathrm{Q}}(k-M),\\ &A(k), \ldots, A(k)^p, \ldots, A(k-M), \ldots, A(k-M)^p\big].\end{aligned} \tag{3}$$

The input layer is followed by L fully connected HLs with b_l nodes. The outputs $\boldsymbol{v}_l(k) \in \mathbb{R}^{1\times b_l}$ of each layer become the inputs of the following one and $\boldsymbol{v}_0(k) = \Phi(k)$. At each node of a layer, nonlinear activation $h(\cdot)$ is applied to a linear combination of the preceding layer's outputs $\boldsymbol{v}_{l-1}$ using the weights $\Omega_l \in \mathbb{R}^{b_{l-1}\times b_l}$, offset by a bias vector $\boldsymbol{\psi}_l \in \mathbb{R}^{1\times b_l}$. Thus, the output vector of layer l becomes

$$\boldsymbol{v}_l(k) = h(\boldsymbol{v}_{l-1}(k)\Omega_l + \boldsymbol{\psi}_l). \tag{4}$$

For the nonlinear activation $h(\cdot)$, we use the Sigmoid function, which is applied element-wise at each node with combined input $z = \boldsymbol{v}_{l-1}(k)\Omega_l$, defined as

$$h(z) = 1/(1+\exp(-z)) \tag{5}$$

while other nonlinear activations can also be considered [20], [33]. The output of the network is formed by a linear output layer with two nodes, written formally as

$$\big[\hat{y}_{\mathrm{I}}(k), \hat{y}_{\mathrm{Q}}(k)\big] = \boldsymbol{v}_L(k)\Omega_{L+1} + \boldsymbol{\psi}_{L+1}. \tag{6}$$

The two outputs are then interpreted as the I and Q components of the complex-valued model output $\hat{y}(k) = \hat{y}_{\mathrm{I}}(k) + \hat{y}_{\mathrm{Q}}(k)$.

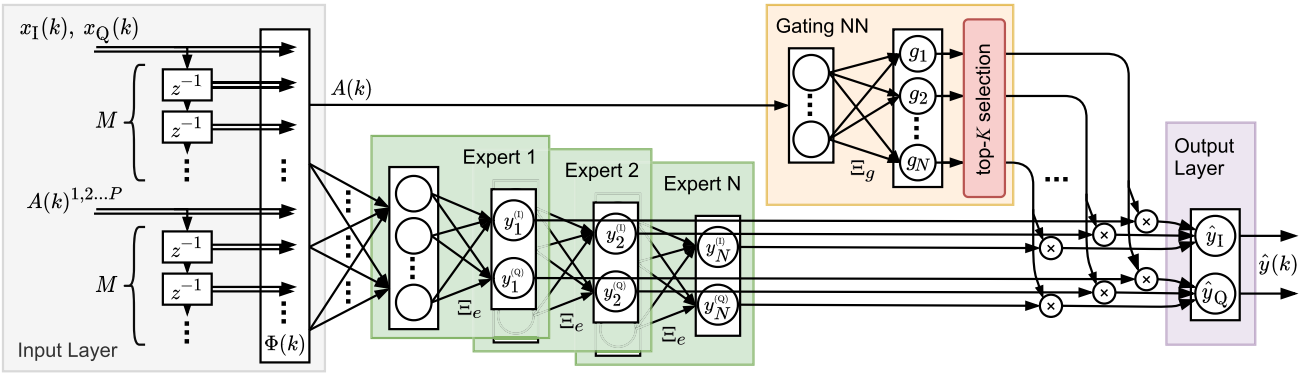


Fig. 3. Block diagram illustrating the proposed MENN processing structure with top-K sparse gating. N real-valued NN experts share a time-delayed input layer and are combined by a gating NN based on the input envelope $A(k)$. The model output is provided as decomposed I and Q components.

B. Proposed MENN

ME is a framework for realizing the divide-and-conquer principle by partitioning the problem space and combining a number of N specialized experts based on a probabilistic scheme [34]. A soft-partitioning unit, referred to as the gate, provides probabilities for the most suitable combination of experts. In its generic form, an ME model with input vector $\boldsymbol{x}$ and output vector $\boldsymbol{y}$ reads [34], [35], [36]

$$P(\boldsymbol{y}|\boldsymbol{x}) = \sum_{n=1}^{N} g_n(\boldsymbol{x}, \boldsymbol{v}) P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{a}_n) \tag{7}$$

where $P(\boldsymbol{y}|\boldsymbol{x})$ is the probability of observing $\boldsymbol{y}$ given $\boldsymbol{x}$. Furthermore, $P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{a}_n)$ is the probability for $\boldsymbol{y}$ of the nth expert associated with a parameter vector $\boldsymbol{a}_n$, and $g_n(\boldsymbol{x}, \boldsymbol{v})$ is the associated weights provided by the gating function with parameters $\boldsymbol{v}$. The gating can generally adopt many shapes, with the so-called mixture model being most commonly adopted. In a mixture model, the gate weights are defined as probabilities, imposing thus the following constraints: $g_n \geq 0$, $n = 1, .., N$ and $\sum_{n=1}^{N} g_n = 1$.

Since for PA modeling and linearization a regression problem needs to be solved, the expectation of the probabilistic framework in (7) can be used as the model prediction $\hat{\boldsymbol{y}}$ [34]. Consequently, the model output becomes a weighted sum of the expert's expectations $\boldsymbol{y}_n = E(P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{a}_n))$, expressed as

$$\hat{\boldsymbol{y}} = \sum_{n=1}^{N} g_n(\boldsymbol{x}, \boldsymbol{v})\, \hat{\boldsymbol{y}}_n(\boldsymbol{x}, \boldsymbol{a}_n). \tag{8}$$

In [13], the ME framework was adopted for PA behavioral modeling and DPD using GMP experts. As the PA nonlinearity acts on the envelope of the transmit signal, the soft-partitioning of the input space is based on the amplitude of the input signal $A(k) = |x(k)|$, leading to

$$\hat{y}(k) = \sum_{n=1}^{N} g_n(A(k), \boldsymbol{v})\, \tilde{\Phi}(k)\tilde{\boldsymbol{a}}_n. \tag{9}$$

A Gaussian mixture model [35] was assumed for the gating variables $g_n(A(k), \boldsymbol{v})$ and expectation maximization (EM) was employed to iteratively converge the model and find a probabilistic partitioning by alternating between updating the gate parameters and the expert parameters.

Inspired by the ME framework and the works in [13], [34], [37], and [38], we propose the MENN, depicted in Fig. 3, as a ME model using RVTDNN experts. Each expert realizes a stand-alone RVTDNN Ξ_e as formulated in Section II-A. The experts share the same inputs $\Phi(k)$ and structure, however, using an individual set of parameters a_n per expert. Thus, the individual expert output becomes

$$\hat{\boldsymbol{y}}_n(k) = \Xi_e(\Phi(k)|\,\boldsymbol{a}_n). \tag{10}$$

In place of the commonly adopted Gaussian-distributed gating probabilities, we instead employ NN-enabled gating, similar to in [37] and [38]. Thus, the soft-partitioning gate can basically adopt any arbitrary distribution. To comply with the probabilistic interpretation of the gating weights, softmax normalization is applied in-line with the ME formulation in [34], expressed as

$$g_n(\boldsymbol{x}, \boldsymbol{v}) = \frac{\exp(\beta_n(\boldsymbol{x}, \boldsymbol{v}))}{\sum_{m=1}^{N} \exp\left(\beta_m(\boldsymbol{x}, \boldsymbol{v})\right)} \tag{11}$$

where $\beta_n(\boldsymbol{x}, \boldsymbol{v})$ denote the nonlinear kernels. Moreover, we restrict the gating probabilities to be a function of the instantaneous input envelope, similar to other piecewise DPD models. Thus, β_n becomes an arbitrary nonlinear function of $A(k)$, realized by a smaller gating NN Ξ_g with one input, one HL with *Sigmoid* activation, and N outputs. This can thus be written as

$$\beta_n(k) = \Xi_g^{(n)}(A(k)|\,\boldsymbol{v}) \tag{12}$$

where $\Xi_g^{(n)}$ is the nth output of the gating NN.

Finally, the output of the aggregate MENN reads

$$\hat{\boldsymbol{y}}(k) = \sum_{n=1}^{N} g_n(A(k), \boldsymbol{v})\, \Xi_e(\Phi(k)|\,\boldsymbol{a}_n). \tag{13}$$

Furthermore, similar to the RVTDNN, the outputs $\hat{\boldsymbol{y}}(k) = [\hat{y}_I(k), \hat{y}_Q(k)]$ of the NN form the complex-valued output signal $\hat{y}(k) = \hat{y}_I(k) + \hat{y}_Q(k)$.

C. Proposed Top-K Gated MENN (K-MENN)

To extend further the above MENN concept and our initial work in [29], we next introduce top-K sparse gating to the MENN framework, with particular emphasis on reducing the

run-time complexity. Such sparse gating has been applied earlier to reduce the computational cost of very large-scale NN models in the context of natural language processing, e.g., in [30] and [31]. Here, we formulate gating schemes that are suitable for efficient, throughput-oriented NN inference in the DPD context. To this end, with the top-K gated ME, only a limited number of $K < N$ expert NNs are allowed to participate in producing an output $\hat{y}(k)$. As highlighted in Fig. 3, we enforce the desired sparsity by placing a top-K selector at the output of the gating NN, which selects, on a per sample basis, the K experts which have the highest rating $g_n(k)$ as provided by the gating NN. Thus, the sparse gating provides $g'_n(k)$ with

$$g'_n(A(k), \boldsymbol{v}, K) = \frac{\text{top-}K_n(\boldsymbol{z})\, g_n(A(k), \boldsymbol{v})}{\sum_{m=1}^{N} \text{top-}K_m(\boldsymbol{z})\, g_m(A(k), \boldsymbol{v})} \tag{14}$$

where $\boldsymbol{z} = [g_1(A(k), \boldsymbol{v}), \ldots, g_N(A(k), \boldsymbol{v})]$ and

$$\text{top-}K_n(\boldsymbol{z}) = \begin{cases} 1 & \text{if } z_n \text{ amongst the } K \text{ largest elements in } \boldsymbol{z} \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

In (14), subsequent to the top-K selection, the gate weights are normalized such that they reflect the relative importance of each expert, while the weights of nonselected experts are set to zero. The selected experts are still weighted respective to the corresponding probabilities after top-K selection, and thus the model output of the MENN with top-K sparse gating reads

$$\hat{\boldsymbol{y}}(k) = \sum_{n=1}^{N} g'_n(A(k), \boldsymbol{v}, K)\, \Xi_e(\Phi(k)|\boldsymbol{a}_n). \tag{16}$$

The choice of K is a trade-off between accuracy and efficiency, that will be further assessed and evaluated in subsequent sections. Based on the sparsity of the gating, we can save considerable amount of computations during NN inference, as Ξ_e only needs to be processed for K active experts per sample inference. Thus, a small K can significantly reduce the inference effort, which we refer to as run-time complexity in the remainder of the article. However, while a small K means that less experts need to be processed, also fewer experts collaborate in providing an output, with top-1 gating being an extreme special case where only one expert is selected at a time and thus the ME becomes a switched NN model [31]. Additionally, sparse gating requires modifications to the training procedure, especially in the top-1 case, where the gating becomes non-differentiable. Such necessary modifications to still allow for end-to-end training are described in the following.

III. End-to-End NN Training

While linear-in-parameter models, such as GMP-based piecewise models, can be identified using a closed-form least-squares solution [39], NNs require iterative optimization techniques that use gradient estimates to converge the model parameters. In this work, we use the adaptive moment estimation (*Adam*) optimizer, which is an extension to the stochastic gradient descent algorithm, maintaining a per-parameter learning rate based on the first and second moments of the gradients [40]. Learning is pursued in batches, i.e., block-wise processing of training data. The parameters of the net are updated according to gradients which are averaged over the batch of B samples. The gradients are computed with respect to minimizing an objective function, given as the batch-wise mean squared error (MSE) of the NN output, i.e.,

$$\mathcal{L}_{\text{MSE}} = \frac{1}{B} \sum_{k \in \text{batch}} \Big(\big(\hat{y}_{\text{I}}(k) - y_{\text{I}}(k)\big)^2 + \big(\hat{y}_{\text{Q}}(k) - y_{\text{Q}}(k)\big)^2 \Big). \tag{17}$$

As in [29], we optimize the parameters of the aggregate MENN using a joint training approach including both the gating and the experts such that the gating is found alongside with the expert's specialization, while minimizing the model error. Thus, we avoid the iterative EM procedure of alternating between updating the experts and the gate used in [13], but instead use an end-to-end training scheme which enables collaboration among the NN experts. It is noted that the described end-to-end training approach is a general method and can as well be applied, e.g., with ME-GMP-based methods and models.

In the following, we describe a number of techniques which can help to improve the training and convergence of the MENN, and especially those of the K-MENN. Not all optimizations are applicable in all cases, and thus Fig. 4 provides an overview of the training procedure, depending on the configuration.

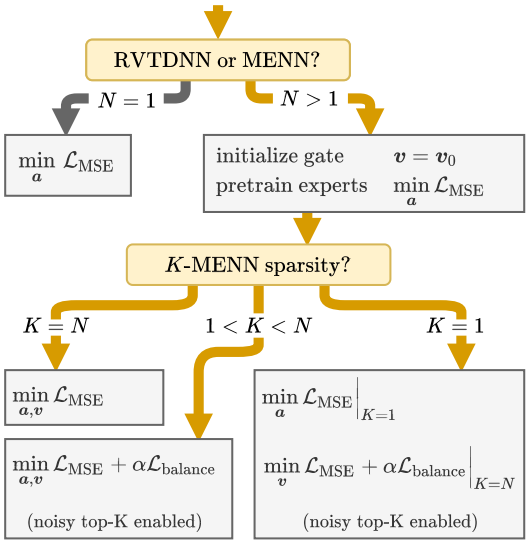


Fig. 4. Conceptual illustration of the training procedures for RVTDNN, MENN, and K-MENN, depending on the NN type, and the K-MENN sparsity factor K. The experts' parameters are contained in $\boldsymbol{a}$ while $\boldsymbol{v}$ specifies the parameters of the gate.

A. Gate Initialization and Expert Pretraining

In our initial experiments, direct gradient-based training was observed not to diversify the experts in a reliable manner, yielding thus a suboptimal partitioning and performance. Instead, greater accuracy can be achieved when providing guidance to the gradient-based training process. In [41], an additional loss is introduced to encourage diversity of the gating output. However, in our RF experiments, we observed

that such loss is difficult to balance and may bias the training unnecessarily. Instead, reliable training can be achieved by initializing the gate parameters with a predefined partitioning, and pretraining the experts accordingly. The joint training then starts from a well-defined initial setting, leading to proper partitioning and expert collaboration. To this end, we initialize the gate NN with parameters $\boldsymbol{v} = \boldsymbol{v}_0$, such that the gating approximates a Gaussian-based initial partitioning given by

$$\beta_{n,0} \approx -\frac{1}{2}\left(\frac{x-\mu_{n,0}}{\sigma_0}\right)^2 \tag{18}$$

where we distribute the starting conditions $\mu_{i,0}$ for the different experts evenly across the input amplitude space, while σ_0 is chosen equally for all the experts.

B. Top-K Gated MENN-Specific Training Modifications

In principle, the training of the K-MENN follows the same end-to-end approach as MENN, however, the top-K selection causes the experts to only train on the data they were selected for, with some undesired implications on the optimization process. We thus adopt two modifications, conceptually similar to what was done in [30] in the natural language processing context, to tailor the training appropriately. Specifically, the following two problems were observed and addressed.

1) First, we observed that the selective training of the experts tends to keep an existing partitioning, and due to the strongly nonlinear behavior of NNs, the experts have only a limited ability to extrapolate the behavior into adjacent regions. This, in turn, compromises the possibility to optimize the initial partitioning. To overcome this limitation, we introduce an additional noise component to the top-K gating rule during the training, which allows the experts to also learn from data which they are not activated for. Thus, in (14), we substitute $\boldsymbol{z}$ with a noisy $\tilde{\boldsymbol{z}}$ with

$$\tilde{\boldsymbol{z}} = \boldsymbol{z} + \mathcal{N}\left(0, \sigma^2_{\text{top-}K}\right) \tag{19}$$

where $\mathcal{N}(0, \sigma^2_{\text{top-}K})$ denotes Gaussian distributed noise with a variance $\sigma^2_{\text{top-}K}$ and zero mean. The variance of the noise component is chosen such that it causes a small variation to the top-K selection. The noise is, however, not injected to the actual weighting of the experts, but only affects the top-K selection. This unlocks the missing continuity between the active regions and improves the optimization of the partitioning, while not limiting the converged accuracy. After training, the noise component can be removed, as it is of no use for the actual inference phase.

2) Second, the gating introduces a problem regarding the balancing of experts. During training, some experts which may receive a higher rating by the gate will get to train with more data, causing them to converge faster. Such effect reinforces itself to the point where some experts can eventually overpower and entirely suppress others. This can be counteracted through regularization, by adding a balancing loss component $\mathcal{L}_{\text{balance}}$ to the optimization objective. This is expressed formally as

$$\mathcal{L}_{\text{K-MENN}} = \mathcal{L}_{\text{MSE}} + \alpha\, \mathcal{L}_{\text{balance}} \tag{20}$$

where α is a small-valued parameter to tune the loss term. The added loss term directs the gating to balance the experts' importance by also minimizing

$$\mathcal{L}_{\text{balance}} = \text{var}(\gamma_1, \ldots, \gamma_N). \tag{21}$$

The importance measure γ_n of an expert n is defined as the mean of the expert's rating provided by the gate, more specifically as

$$\gamma_n = \sum_{\text{k}\in\text{batch}} g_n(k) \tag{22}$$

and is evaluated per training batch.

The additional balancing loss objective only affects the training of the gating NN. It can straightforwardly be included in the end-to-end optimization, by merging the gradients for the two loss objectives.

C. Top-1 MENN Training

Using top-1 gating, only the single highest-rated expert is allowed to be active during each inference. Consequently, the model becomes a switched model, conceptually similar to the vector-switched model in [11], where several GMP models are multiplexed based on the input signal envelope $A(k)$. Top-1 gating is especially attractive, since the gating decision simplifies to hard switching, which can be easily implemented by amplitude thresholding during the inference. Furthermore, top-1 gating results in the highest degree of sparsity for a given expert count N.

However, end-to-end training of such switched model is not straightforward since, intuitively, at least two experts need to be compared to train the gate alongside with the model. The hard switching prevents the back-propagation algorithm from calculating the gradients for the gating NN part. As reported in [31], we can nonetheless optimize the gating, if considering probability for an expert n generating the output y in line with the formulation in (11). Thus, we optimize the gate parameters $\boldsymbol{v}$ regarding $\mathcal{L}_{\text{MSE}}$ by considering the nonswitched performance as formulated in (13). Furthermore, the gate training is regularized with L_{balance} as in (20) to prevent the experts from being suppressed during optimization. The different optimization objectives do not break the joint end-to-end training of gating and experts. Instead, the different objectives are incorporated when computing the gradients for the joint parameter update, with marginal overhead to the overall training complexity. It is also important to note that the expert's parameters $\boldsymbol{a}_n$ are still optimized to minimize the L_{MSE} of the actual model output with top-1 gating, as formulated in (16). Additional gating noise during training, as in (19), is required also for the top-1 gating case to ensure the continuity between the switched regions and to enable the optimization thereof.

D. DPD Model Training

For applying and evaluating different NN models for DPD purposes, a modified indirect learning approach (ILA) [42] is used for training, as depicted in Fig. 5. The principle of indirect learning is to identify the predistorter parameters by modeling the PA input signal as a function of the PA output.

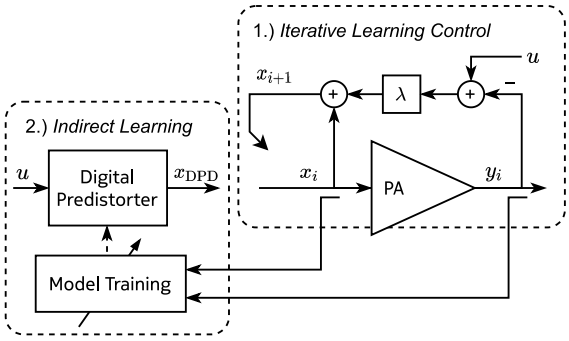


Fig. 5. Illustration of the DPD training scheme. First, a suitable DPD PA excitation is found using ILC. In a second step, the model is trained via ILA, using the ILC optimized PA input and the measured PA output signals.

However, since the PA is excited differently when applying the identified DPD, the initial measurement used for identification may not be representative of the inverse PA behavior. Thus, the identification process needs to be repeated for several iterations until a stable DPD performance is found.

Since our aim is to evaluate and compare the different models, any performance implications of the ILA training procedure should be kept at a minimum. Hence, we take an alternative approach to ensure similar conditions when training the different models. Specifically, instead of iterating the ILA scheme, we first seek a suitable excitation through iterative learning control (ILC) [43], which significantly reduces the evaluation time for each NN model, given the considerable effort for retraining the models in each ILA iteration. ILC is a model-less iterative method to find a suitable input that excites the desired PA output. The transmit signal is repeatedly modified by injecting the output error after retransmission, with opposite phase and weighted by a step-size $\lambda \in (0, 1)$. The process is repeated until the PA output matches the desired output with sufficiently low error. We then use the optimized PA input signal to acquire training data for our models and perform indirect learning. For indirect learning, the polynomial and NN models are trained on the reverse PA behavior, i.e., the ILC optimized input after i iterations $x_i(k)$, and the measured PA output $y_i(k)$ (divided by the target gain) are used as ground-truth data for the model's output $\hat{y}(k)$ and input $x(k)$ for training purposes, respectively. For the actual DPD operation, the trained predistorter operates on the desired transmit signal $u(k)$ and produces the predistorted PA input $x_{\mathrm{DPD}}(k)$.

We apply the described strategy for initial training and evaluation of the models in a static laboratory environment. However, for the actual field application, the models require adaptation to track the potential changes in the PA behavior. Suitable online adaptation schemes, which avoid full retraining of models, are reported, e.g., in [14] and [25]. One straightforward approach builds on updating only the linear output layer of the model. For the MENN, this would translate to adapting the linear output layer of each of the experts, with respect to a fixed gating. If the gating also needs adaptation, then the same end-to-end training approach described previously can be applied. The discussed top-K gating-related training modifications only have limited impact on the fine-tuning and can be omitted.

IV. Model and Run-Time Complexities

We next address the important aspect of processing complexity related to the proposed methods and the corresponding reference schemes. Throughout the rest of this article, we report to two different measures of complexity for the DPD methods: the model complexity and the run-time complexity. In addition to introducing these metrics, we additionally discuss their relation to the actual processing complexity for NN inference and NN training.

The *model complexity* reflects the overall size of the model. We measure the model complexity $\mathcal{C}$ in terms of real-valued parameter count, as this maps directly to the memory cost for storing the model. The real-valued parameter count of a RVTDNN with L fully connected HLs and two outputs can be expressed as

$$\mathcal{C}_{\mathrm{RVTDNN}} = \left(\sum_{l=1}^{L} b_{l-1} \times (b_l + 1)\right) + 2(b_L + 1) \tag{23}$$

where b_l specifies the node count of layer l, for $l \in 1, \ldots, L$, while b_0 is the number of the NN inputs. For the MENN, we additionally have the gating NN, thus the corresponding total complexity reads as

$$\mathcal{C}_{\mathrm{MENN}} = N \times \mathcal{C}(\Xi_{\mathrm{e}}) + \mathcal{C}\big(\Xi_{\mathrm{g}}\big). \tag{24}$$

In this article, we consider MENNs with a single HL for both the gate and the expert NNs, thus

$$\mathcal{C}_{\mathrm{MENN}} = N(b_1 + 1)(b_0 + 2) + (c_1 + 1)(N + 1) \tag{25}$$

with b_1 and c_1 representing the sizes of the HLs of the expert and gate NNs, respectively.

Then, with the *run-time complexity*, we refer to the computational complexity when really processing the model in real-time. One way to assess the run-time complexity of NN models is to quantify the total amount of operations (OPs) required for inference of a single sample, i.e., to count the required multiplications, additions, and nonlinear OPs. Table I shows and compares these for the RVTDNN, MENN, and K-MENN, each with an equal model complexity of $\mathcal{C} =$ 2k parameters. We find that for RVTDNN and MENN, the total OP count proportionally corresponds to their respective model complexities. However, with top-K sparse gating, only a subset of the K-MENN contributes to forming the model output. Consequently, only the active parts of the K-MENN need to be processed and a suitable implementation will spare computation of the inactive experts. This is reflected by a lower amount of OPs stated for the K-MENN in Table I. Given the large complexity of the NN layers, we have neglected the computational effort for providing the NN inputs.

To better relate the run-time complexity to the model complexity $\mathcal{C}$, we introduce the active real-valued parameter count $\mathcal{R}$. For the RVTDNN and MENN, the run-time complexity equals the model complexity since all coefficients are always applied. In the K-MENN case, however, we consider only the

TABLE I
COMPLEXITY COMPARISON OF EQUALLY SIZED RVTDNN, MENN, AND K-MENN. ALL MODELS USE $b_0 = 32$ INPUTS, AND b_1 IS CHOSEN TO MATCH $\mathcal{C} = 2000$ PARAMETERS

Model complexity		*Run-time complexity*		*Training complexity*, $B = 2048$ samples			
NN Model	Total param. count $\mathcal{C}$	Active param. count $\mathcal{R}$	Inference (OPs/sample)	Forward pass (OPs/sample)	Backward pass (OPs/sample)	Param. update (OPs/batch)	Training Step (OPs/batch)
RVTDNN ($b_2 = 10$)	2000	2000 (100%)	3940 (100%)	3940	4958	26 k	18247 k (100%)
MENN (N = 4)	2000	2000 (100%)	3933 (100%)	3933	4249	26 k	16782 k (92%)
K-MENN ($N = 4$, $K = 2$)	2000	1032 (52%)	2022 (51%)	2022	2203	26 k	8679 k (48%)
K-MENN ($N = 4$, $K = 1$)	2000	548 (27%)	1067 (27%)	3933	1180	26 k	10466 k (57%)

coefficient count of the K simultaneously active experts, and the run-time complexity can be expressed as

$$\begin{aligned}\mathcal{R}_{\text{K-MENN}} &= K \times \mathcal{C}(\Xi_{\text{e}}) + \mathcal{C}\big(\Xi_{\text{g}}\big) \\ &= K(b_1+1)(b_0+2) + (c_1+1)(N+1). \end{aligned} \tag{26}$$

For reasonably sized models, the active parameter count $\mathcal{R}$ behaves proportionally to the total OPs, thus being a reasonable measure for the NN run-time complexity.

The corresponding polynomial models, utilized for reference purposes in the upcoming experiments, have as many complex-valued parameters as the model has regressors. For comparison purposes, this is mapped to the model complexity by considering each complex-valued parameter as two real-valued ones. The run-time complexity of polynomial models is, however, strongly linked with the implementation. Considerable complexity arises from creating the model regressors, however, the polynomials may, in parts, be mapped to LUT-powered structures, making it nontrivial to rate the run-time complexity without a specific implementation assumption. In the polynomial model cases, we thus only report the model complexity numbers in the following experiments.

To complete the discussion on processing complexity, the *training complexity* of the different NN models is explored in the following. In our experiments, RVTDNN and K-MENN showed a similar convergence speed and we found that similar settings for the amounts of training samples, epochs, batch size, and learning rate are suitable. Comparing the training complexity thus narrows down to comparing a single NN training step. Therein, a batch of B samples is propagated through the model (forward pass, FP), after which trainable parameters are updated according to gradients minimizing the MSE of the NN output. These gradients are computed by means of the back-propagation algorithm (backward pass, BP). For a reasonable batch size B, forward and backward pass clearly dominate the overall training complexity as their cost scales proportional with B, whereas the model parameters are updated only once per training batch. Fortunately, the sparsity aspect of K-MENN also benefits the training complexity since in both backward and forward pass, computations can be skipped for the disabled experts, whereas the actual parameter update affects all model parameters, irrespective of the sparse gating. The additional balancing loss and gating noise component proposed for the K-MENN in Section III-B both have a negligible effect on the training complexity. However, as discussed in Section III-C, the top-1 gated K-MENN requires evaluation of all experts during the forward pass to properly train the gating, which compromises the gains from sparse gating for the fully switched case. We report the exact absolute and relative training complexities in Table I for the different NN models with $\mathcal{C} = 2000$ parameters.

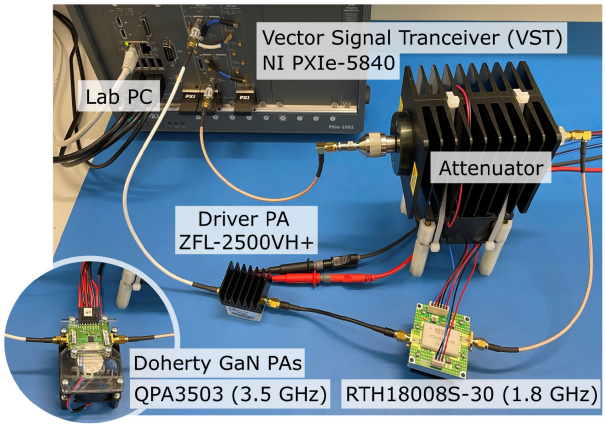


Fig. 6. Graphical illustration of the RF measurement setup and PA modules used in the experiments.

V. RF MEASUREMENT EXPERIMENTS

In this section, different variants of the proposed MENN and K-MENN models are evaluated and compared with the RVTDNN as well as the GMP, and different piecewise polynomial models. We first present forward modeling results based on RF measurements with a GaN Doherty PA operating at the 1.8 GHz band. Subsequently, we apply the compared models for the actual linearization of two different Doherty GaN PAs, operating at the 1.8 GHz and the 3.5 GHz bands, respectively. The performance of each model is evaluated in terms of the adjacent channel leakage ratio (ACLR), the normalized mean squared error (NMSE), and the error vector magnitude (EVM), with respect to the model's run-time complexity and the parameter count.

A. Experimental Setup and Configuration

In our RF experiments, as also illustrated graphically in Fig. 6, we utilized the *NI PXIe-5840* vector signal transceiver (VST) for analog signal generation and up-conversion to RF, in the transmit path, as well as for the downconversion and digitization of the PA output in the observation path. Two different PA units were used in the experiments. For the forward modeling experiments as well as the first set of DPD experiments, we used the *RTH18008S-30* by *RFHIC*, which is a two-stage GaN Doherty PA operating at 1.842 GHz center frequency with static biasing. A linear, high-gain driver

amplifier, *ZFL-2500+*, was used to preamplify the signals to reach the desired output power. In the second set of DPD experiments, we used the *QPA3503* by *Qorvo*, which also is a two-stage GaN Doherty PA, while operating at 3.5 GHz. This PA was used without any additional driver amplifier.

Before feeding the amplified signals to the observation receiver path, sufficient attenuation was provided to match the VST's input dynamic range. Furthermore, we conducted each measurement five times and applied time-domain averaging on the time aligned signals in order to reduce the system's internal noise floor. The transmit path baseband I and Q samples were generated on a host PC running MATLAB, which was also used for controlling the VST transmission and data recording as well as for evaluation of the DPD performance.

The NNs were trained offline with *TensorFlow*, using 600 training epochs on 120k samples of training data. The samples were processed in batches of 2048 samples for each parameter update step. The global learning rate, which serves as an upper limit for the individual parameter update rates in *Adam*, was set to an initial value of $\mu_0 = 0.2$, while then allowed to decay with

$$\mu_p = \mu_0 \times 0.998^p \tag{27}$$

for each batch update step p. For the exponential decay rates of the first and 2nd moment mean estimates, we applied $\beta_1 = 0.98$, and $\beta_2 = 0.999$, respectively. For training the K-MENN, necessary modifications described in Sections III-B and III-C were additionally applied. The additional noise component in (19) to establish continuity between regions during training was applied with $\sigma_{\text{top-}K} = 0.005$ while the regularity balancing term in (20) was weighted with $\alpha = 0.001$. For the actual evaluation or testing of the learned models, separately generated data with 50k samples was used. In the case of forward modeling, training data and validation data are the recorded I/Q input and output samples for the tested PA. For the DPD linearization experiments, ground-truth data are generated using the ILC method detailed in Section III-D.

In the upcoming experiments, we evaluate different models at different sizes. For the NNs, the choices of the input-layer hyperparameters P and M are specific to the experiment. A suitable configuration was determined initially in each case, using a RVTDNN with 2000 parameters as reference or baseline solution. We state the respective choices at the beginning of each corresponding section. Then, to vary the NN sizes, the amount of nodes in the first layer b_1 was altered. For the RVTDNNs with two HLs, the size of the second HL was fixed at $b_2 = 10$ nodes. For the MENN, the HL of the gating NN always uses ten nodes. For each NN, we repeated the training and measurement five times, and averaged the results to mitigate the impact of randomly initialized parameters during training.

To scale up the polynomial models, an exhaustive search on suitable GMP basis function configurations was performed and the best performing cases were selected for comparison. For the piecewise models, we restricted our exploration to up to six partitioning regions. The hard partitioning needed for the VS- and DPW-GMP models was determined by K-means clustering as described in [11]. The soft-partitions of the ME-GMP, are found by means of the EM algorithm [13], with an upper limit of 25 iterations. The polynomials were identified and validated with the same dataset as the NNs. We use the MATLAB built-in least squares solver to identify the coefficients of the polynomials. Ridge regularization [44] was additionally applied to stabilize the estimated polynomial models with high-order GMP terms.

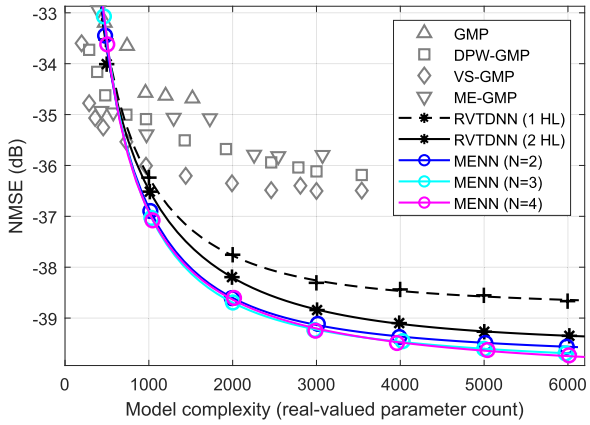


Fig. 7. Measured forward modeling NMSEs of the various GMP-based polynomial models and NN models for the GaN Doherty PA at 1.8 GHz running a 100 MHz 5G NR OFDM signal at +40.7 dBm output power.

Models are reported and compared with respect to their numbers of real-valued parameters, as a measure of complexity, following the analysis principles of Section IV. The complex-valued coefficients of the polynomials are properly mapped to the corresponding real-valued parameter count for fair comparison.

B. Forward Modeling Experiments at 1.8 GHz

As the first set of experimental results, we present the forward modeling performance for the measured behavior of the *RTH18008S-30* operating at 1.8 GHz while including also a driver amplifier *ZFL-2500+*. We applied a 5G NR compliant 256 QAM modulated orthogonal frequency division multiplexing (OFDM) waveform with 100 MHz channel bandwidth and a subcarrier spacing of 30 kHz. The PAPR of the waveform was limited to 8 dB using iterative clipping and filtering. The baseband sampling frequency for processing the signals was chosen as 614.4 Msamples/s to account for the bandwidth expansion due to the PA nonlinearity. The PA was operated close to saturation, with an average RF output power of +40.7 dBm.

To assess the modeling capabilities toward large parameter counts, we sweep the sizes of the NN and polynomial models by adding more nodes to the NNs, or adding more basis-functions in case of the polynomials. Furthermore, for this experiment, the input layer was configured to use first and third order envelope inputs and an input delay line with a delay depth of $M = 5$, resulting in a total of $b_0 = 24$ NN inputs.

Fig. 7 reports the forward modeling NMSEs for the MENN as well as the various reference methods as a function of the real-valued parameter count. No sparse gating is yet applied, in the context of Fig. 7. We can observe that the NNs can

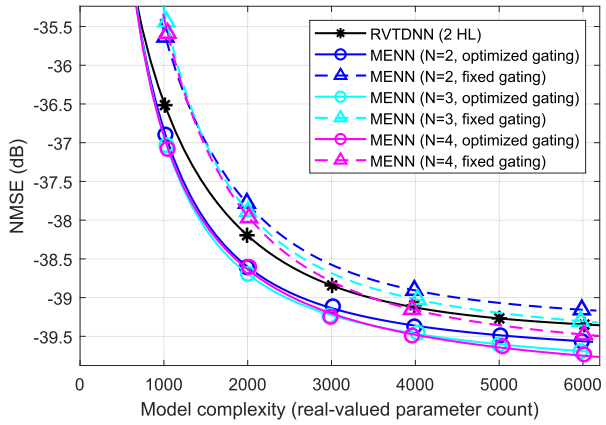


Fig. 8. Measured MENN forward modeling NMSEs with an initially fixed as well as end-to-end optimized gatings, for the GaN Doherty PA at 1.8 GHz running a 100 MHz 5G NR OFDM signal at +40.7 dBm output power.

be scaled to outperform the polynomial models. While polynomial models reach a lower bound for NMSE when adding more basis functions, NNs continue to successfully scale their performance toward lower modeling errors if provided with more nodes in the HL. Furthermore, differences among the considered NN models are apparent. The RVTDNN benefits from having a second HL, especially when scaled toward higher parameter counts. The single HL MENN outperforms the single HL RVTDNN, and performs slightly better if compared to the RVTDNN with two HLs. The MENN was also tested with two, three, and four experts, which all resulted in similar NMSE accuracies with negligible differences, as can be observed through Fig. 7. Thus, in this forward modeling scenario, the number of experts does not essentially impact the modeling accuracy. Furthermore, in our earlier experiments in [29], we observed that adding a third HL to the RVTDNN, or adding more HLs to the MENN, does not yield any notable performance gains.

Additionally, in Fig. 8, the modeling accuracy of the end-to-end optimized MENN is compared with a fixed gating case, where the initial MENN gating was not optimized during training. Clearly, the optimized gating advances the accuracy with respect to the initial segmentation. The differences are more prominent for smaller sized NNs and are as large as 1 dB in the 1000–2000 parameter range.

Next, we apply the proposed sparsely gated K-MENN approach to the same GaN PA forward modeling scenario. The modeling accuracies for the different processing variants are depicted in Fig. 9(a)–(c). For each graph, the number of experts was kept constant while the number of active experts was varied from $K = N$ down to $K = 1$. Additionally, to highlight the differences between the model and run-time complexities, both are plotted and interconnected horizontally. Thus, two horizontally connected data points refer to the same measurement result and NN configuration, but depict the difference in the model and run-time complexities due to the sparse gating. As can be observed, limiting the number of active experts for a fixed MENN realization degrades the accuracy of the K-MENN, compared to the respective nonsparse MENN case of $K = N$. Especially $K = 1$ is already accompanied by a considerable loss of accuracy, if considering only the model complexity. Nevertheless, considering the run-time complexity of the sparsely gated K-MENN, the performance can be pushed below the accuracy-complexity curve of the nonswitched NNs. Naturally, a greater ratio of active versus total number of experts improves the run-time complexity and accuracy trade-off, however, at the cost of an increased total model size. Consequently, $K = 1$ cases can offer the lowest modeling NMSE with the least run-time complexity, however, with a comparably larger model complexity.

Regarding the total number of experts, it can be observed that a higher value for the expert count N allows for a more fine-grained choice on the ratio of the active to total number of experts and thus on the trade-off between the model and the run-time complexities. Also, a larger complexity difference is possible, allowing for high accuracy at lower run-time complexity. Moreover, and importantly, these experimental results indicate that a K-MENN with more experts scales better toward higher parameter counts.

In Fig. 10, the end-to-end optimized gating is illustrated considering the example case of a K-MENN with 2000 total parameters and varied sparsity of $K = 4$, 2, and 1. The upper graphs show the combined outputs together with the individual expert contributions, which are weighted according to the gate weights g'_n. The respective outputs of the gate NNs are shown in the bottom row subfigures, as functions of the input envelope. It is important to note that the gating depends only on the instantaneous input envelope $A(k)$, thus it does not have any memory behavior and appear as continuous lines. Additionally, due to the soft-max normalization, they are bound to a range from zero to one while adding up to one as a whole. In the leftmost case, shown in Fig. 10(a), all experts are active. Thus, the individual expert NNs will strongly collaborate in forming the output, thanks to the end-to-end training. However, none of the experts could work standalone. In the middle case, shown in Fig. 10(b), only two experts were allowed to be active. It is clearly visible and observable that the lowest rated experts are being suppressed due to the top-2 selection. A corresponding example with top-1 selection is then depicted in the rightmost graphs, shown in Fig. 10(c). The gate weights after top-1 selection equal either one or zero, and the hard switching is clearly observable. The soft probability rating, which is the basis for the top-1 selection, is shown in addition to the top-1 selected weights.

Finally, it is noted that we have compared the end-to-end optimized gating with the nonoptimized initial gating also for the K-MENNs, and the respective performance differences are in the order of 0.5 dB for the 1000–2000 parameter range.

C. Digital Predistortion Experiments at 1.8 GHz

Next, we apply the proposed K-MENN approach for linearizing the *RTH18008S-30* PA running the same 100 MHz 5G NR waveform at 1.8 GHz as in the previous forward modeling experiments. Compared to the previous experiments, we now operate the PA with slightly reduced output power, namely +38.9 dBm. This corresponds to an approximate back-off of 9 dB relative to the PA's saturation level, implying thus a

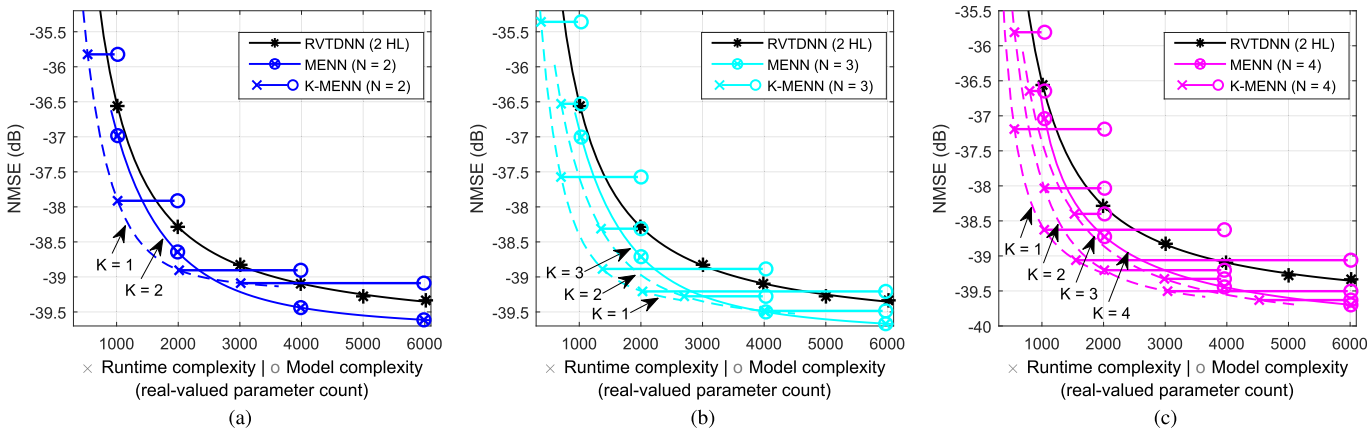


Fig. 9. Measured forward modeling NMSEs of the different MENN and K-MENN configurations with varied K for the GaN Doherty PA at 1.8 GHz being excited with 100 MHz 5G NR OFDM signal at +40.7 dBm output power. For each K-MENN realization, the run-time and model complexities are plotted and connected with a horizontal line. Curves illustrate the interpolated trends for a specific sparsity K. (a) K-MENN with $N = 2$ total experts. (b) K-MENN with $N = 3$ total experts. (c) K-MENN with $N = 4$ total experts.

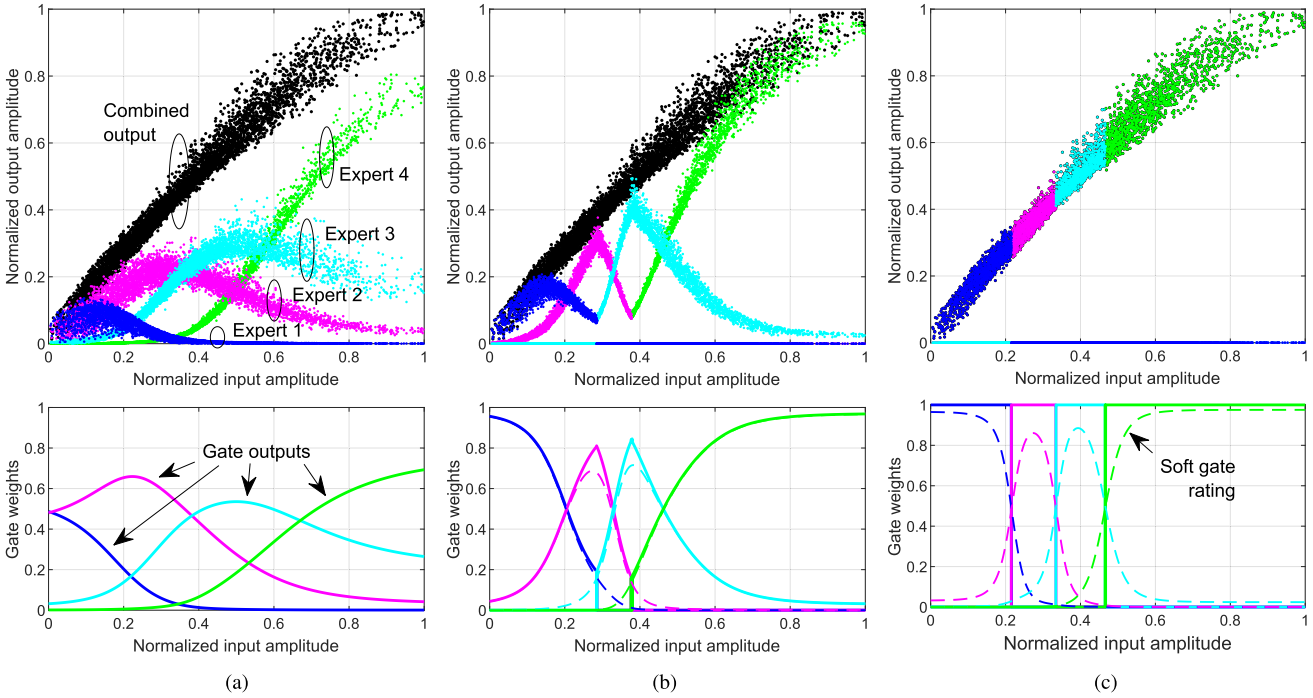


Fig. 10. Example (K-)MENN modeling results for the 1.8 GHz GaN Doherty PA (100 MHz 5G NR OFDM, +40.7 dBm output power) using different (sparse) gating configurations for the four experts. The upper graphs show the amplitude contributions of the individual experts, together with the combined model output. The bottom graphs depict and illustrate the corresponding weights provided at the output of the gating NN. (a) $N = 4$ experts, all active ($K = 4$). (b) $N = 4$ experts, top-2 gating ($K = 2$). (c) $N = 4$ experts, top-1 gating ($K = 1$).

feasible operation point for employing DPD given the earlier stated PAPR characteristics of the digital waveform. Additionally, we also evaluate the models using a carrier aggregation type of waveform, consisting of three 40 MHz OFDM carriers, corresponding to a total aggregate bandwidth of 120 MHz, while having an overall PAPR of 8 dB.

For comparison purposes, we again apply and compare the RVTDNN, MENN, K-MENN, as well as various GMP-based piecewise models. Specifically, when it comes to MENN and K-MENN methods, we consider configurations with $N = 2$ and $N = 4$ experts while assessing the performance with $K \leq N$. The NN node counts are configured such that the baseline MENN cases have matching model complexities to the reference RVTDNN cases, while then also K-MENN cases with reduced run-time complexities are assessed. Additionally, the polynomial-based systems are parameterized using a relatively large number of basis functions, and the respective pareto-best configurations are selected.

The measured ACLRs are reported in Fig. 11, in (a) for the 100 MHz test waveform case and in (b) for 120 MHz carrier aggregation waveform case. To depict the respective run-time and model complexities of the K-MENN, the ACLR results are plotted for both complexities and interconnected horizontally. The overall results are consistent with the findings observed for behavioral modeling. The figures report effective linearization for each of the models. Furthermore, NNs tend to outper-

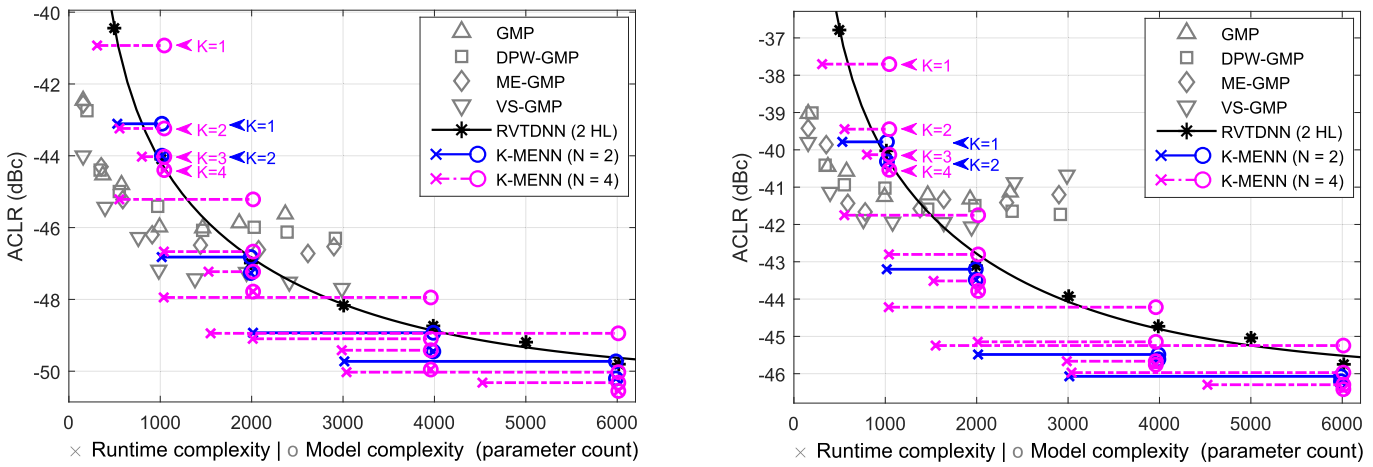


Fig. 11. Measured linearization performance of various DPD models for the 1.8 GHz GaN Doherty PA using two different OFDM waveforms at an output power of +38.9 dBm. K-MENN with $N = 2$ and $N = 4$ experts and varied $K \leq N$ are considered. For each measured K-MENN realization, the model complexity as well as the run-time complexity with respect to the sparse top-K gating are shown and horizontally interconnected. (a) 100 MHz OFDM. (b) 120 MHz (3×40 MHz) OFDM.

TABLE II

SUMMARY OF THE RF MEASUREMENT RESULTS FOR THE LINEARIZATION OF THE 1.8 GHZ GAN DOHERTY PA OPERATING A 100 MHZ OFDM SIGNAL WITH AN OUTPUT POWER OF +38.9 DBM

DPD Model	N	K	hidden layer size $b_{1,2}$	Model Comp. $\mathcal{C}$ (param. count)	Run-time Comp. $\mathcal{R}$ (param. count)	NMSE (dB)	ACLR (dB)	EVM (%)	shown in Fig.12
GMP	(1)	(1)	-	1464	-	-39.0	-46.0	1.71	*)
DPW-GMP	3	(3)	-	1464	-	-39.4	-46.1	1.70	
ME-GMP	5	(5)	-	1440	-	-39.6	-46.5	1.67	
VS-GMP	5	(1)	-	1380	-	-39.7	-47.4	1.67	*)
RVTDNN	(1)	(1)	56,10	1992	1992	-40.3	-47.0	1.65	*)
RVTDNN	(1)	(1)	85,10	3007	3007	-41.2	-48.2	1.64	
MENN	4	4	18	2016	2016	-40.9	-47.8	1.61	*)
K-MENN	4	2	36	3960	2012	-41.8	-49.1	1.60	*)
K-MENN	4	1	36	3960	1038	-41.0	-47.9	1.61	*)
K-MENN	4	1	55	6012	1551	-41.7	-49.0	1.60	

form polynomial models above a certain model size, while polynomial models reach a lower bound for the ACLR. With the proposed K-MENN approach, however, very low ACLRs can be reached at a competitive complexity. As a concrete example, in the range of 2000 active parameters, only the sparsely gated K-MENNs with $K = 2$ or $K = 1$ are competitive with the polynomial models. The K-MENN enables a 50% reduction in run-time complexity compared to an equally performing RVTDNN with the same model complexity. Even better ACLR, or lower run-time complexity, can be achieved if the model complexity is increased by a factor of 2× (100%). As the $K = 1$ setting offers the largest variance of model and run-time complexity, the best linearity results with lowest run-time complexity could be achieved with $K = 1$, $N = 4$.

Different from the forward modeling results, the RVTDNN and MENN need to be scaled toward a high complexity to provide a significant linearization advantage over the polynomial models. Generally, the linearization capabilities of polynomials versus NNs are specific to the PA, its operation point, and the properties of the applied signals. Typically, NNs tend to outperform polynomials when strong nonlinear memory behavior is present. This can be observed when applying a waveform with larger bandwidth, as shown in Fig. 11(b), where a 120 MHz wide OFDM waveform, consisting of three 40 MHz carriers was applied to the same PA. With the larger bandwidth, the NNs tend to outperform the polynomial models also at smaller parameter counts. However, the achievable linearization is also slightly degraded, due to the excitation of strong nonlinear memory behavior, and a comparably lower in-band power density due to the widened bandwidth.

Results of selected model configurations are further detailed and summarized in Table II, for the 100 MHz measurement case. For the polynomial-based systems, only the respective best performing cases are reported, while for the NNs, only configurations with a run-time complexity at around 2000 parameters are detailed. The differences in model accuracy map to the ACLR differences as reported in the table. In addition to the obtained linearized ACLRs, also NMSE and EVM figures are listed. The reported EVM results follow the NMSE and ACLR trends, however, irrespective of the model, the EVM results are bound to around 1.6% which is due to a minimum EVM introduced by the digital PAPR reduction.

Representative power spectra of the linearized PA outputs are shown in Fig. 12, corresponding to the results reported in Table II. Furthermore, Fig. 13 shows the measured AM–AM and AM–PM behavior, with and without K-MENN

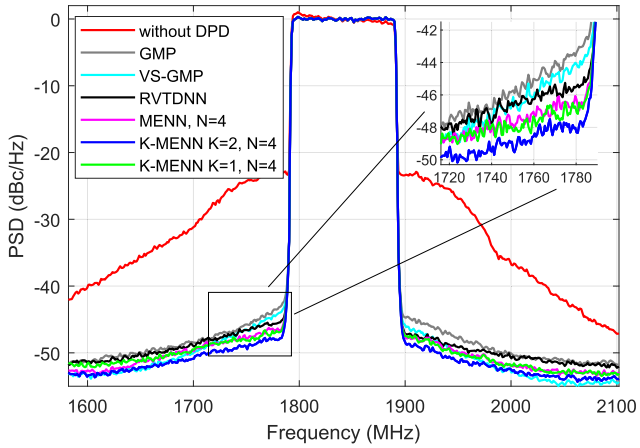


Fig. 12. Measured PA output power spectra of the 1.8 GHz GaN Doherty PA (100 MHz OFDM waveform, +38.9 dBm output power) without and with DPD, using selected models reported in Table II.

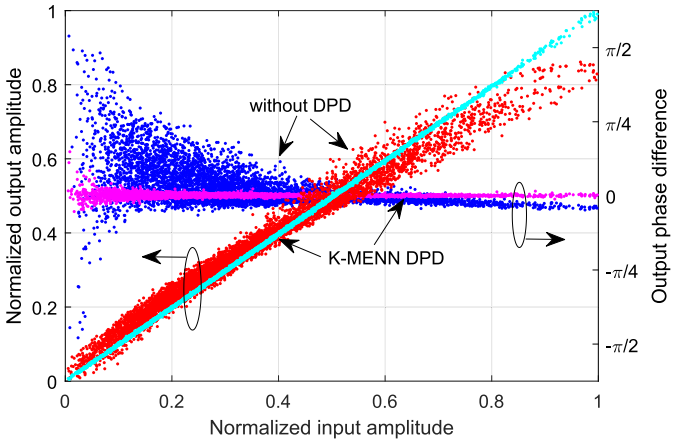


Fig. 13. Measured AM–AM and AM–PM characteristics of the 1.8 GHz GaN Doherty PA with an output power of +38.9 dBm running a 100 MHz OFDM waveform. The linearized output has −41.0 dB NMSE and is achieved using a K-MENN with $N=4$, $K=2$, and a model complexity of 3960 model parameters.

linearization. As can be observed, the NN DPD successfully compensates the strong nonlinear and dynamic distortion effects of the PA, affecting both the phase and the amplitude.

D. Digital Predistortion Experiments at 3.5 GHz

In the second linearization experiment, we apply DPD to the *QPA3503* operating at 3.5 GHz. We use a wide, noncontiguous multicarrier waveform, consisting of five 20 MHz OFDM component carriers with 15 kHz subcarrier spacing and 256 QAM as the subcarrier data modulation. The active component carriers are placed at offsets of [−70, −30, +10, +30, +70] MHz, thus spanning 160 MHz of total linear bandwidth. The signal is oversampled by approximately 5× using a sampling rate of 798.72 Msamp/s. Again, we limit the PAPR of the composite waveform to 8 dB by means of iterative clipping and filtering. The PA is operated with an output power of +35.6 dBm such that clipping free amplification of the waveform is feasible with DPD.

A slightly larger NN input memory depth, namely $M = 7$, is selected for this experiment and additional first-

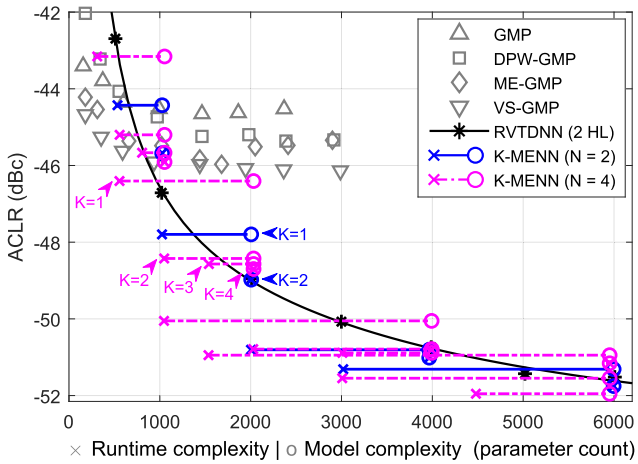


Fig. 14. Measured DPD linearization results for the 3.5 GHz GaN Doherty PA (160 MHz noncontiguous carrier aggregation OFDM waveform, +35.6 dBm output power). K-MENN with $N=2$ and $N=4$ experts and varied $K \leq N$ are considered. For each measured K-MENN realization, the model complexity as well as the run-time complexity with respect to the sparse top-K gating are shown and horizontally interconnected.

and third-order envelope inputs are used to augment NN models, resulting in a total of $b_0 = 32$ NN inputs. As in the first DPD experiment, the NN HLs are parameterized to match the model-complexities of RVTDNN and K-MENN in order to allow for a direct and fair comparison. K-MENN variants with $N = 2$ and $N = 4$ experts are considered, and K-MENN sparsity with $K \leq N$ is assessed. The basis function configurations of the polynomial cases are also reevaluated for the different PA, compared to the 1.8 GHz measurements.

The achieved ACLRs using the different DPD models are reported in Fig. 14. For measuring the ACLR, the average power densities of each 20 MHz wide channel, including the in-between channels of the multicarrier waveform, were considered and the worst case ACLR is always reported. In this scenario, as can be observed, the NNs tend to significantly outperform the polynomial reference methods. The sparsely gated K-MENN allows for low ACLRs, also at low run-time complexity. Thus, very low ACLR is achievable with modest run-time complexity due to the sparse gating, e.g., the range of 1000–2000 parameters. For ACLR levels below −50 dBc, the run-time complexity of the K-MENN is 50%–66% lower compared to equally performing RVTDNN, while the model complexity of K-MENN is increased by 1×–1.5× (0%–50%).

Detailed results for selected configurations are reported in Table III, stating also the obtained NMSE and EVM figures. The corresponding example PA output spectral densities are shown in Fig. 15. The spectrum view reveals that the worst ACLR is found for the in-between adjacent channels of the active carriers. Especially for these in-between adjacent channels, the NNs allow for a significantly better ACLR compared to the polynomial models, while the ACLR differences in outer channels are observed to be smaller. Finally, the scatter graph in Fig. 16 shows the AM–AM and AM–PM profiles of the distortion and the linearized case for an example K-MENN DPD with $K=2$ and a run-time complexity of approximately

TABLE III

SUMMARY OF THE RF MEASUREMENT RESULTS FOR THE LINEARIZATION OF THE 3.5 GHz GaN DOHERTY PA OPERATING A NONCONTIGUOUS 160 MHz SIGNAL WITH AN OUTPUT POWER OF +35.6 dBm

DPD Model	N	K	hidden layer config. $b_{1,2}$	Model Comp. $\mathcal{C}$ (param. count)	Run-time Comp. $\mathcal{R}$ (param. count)	NMSE (dB)	ACLR (dB)	EVM (%)	shown in Fig. 15
GMP	(1)	(1)	-	1464	-	-40.2	-44.6	1.77	*)
DPW-GMP	3	(3)	-	1464	-	-40.9	-45.2	1.74	
ME-GMP	5	(5)	-	1680	-	-41.2	-46.0	1.73	
VS-GMP	6	(1)	-	1994	-	-41.3	-46.1	1.73	*)
RVTDNN	(1)	(1)	46, 10	2010	2010	-42.2	-49.0	1.74	*)
MENN	4	(4)	14	2032	2032	-42.0	-48.7	1.74	*)
MENN	4	(4)	28	3992	3992	-43.7	-50.8	1.72	
K-MENN	4	2	28	3992	2082	-43.5	-50.8	1.71	*)
K-MENN	4	1	28	3992	1046	-42.9	-50.1	1.71	*)
K-MENN	4	1	42	5952	1536	-43.5	-50.9	1.72	

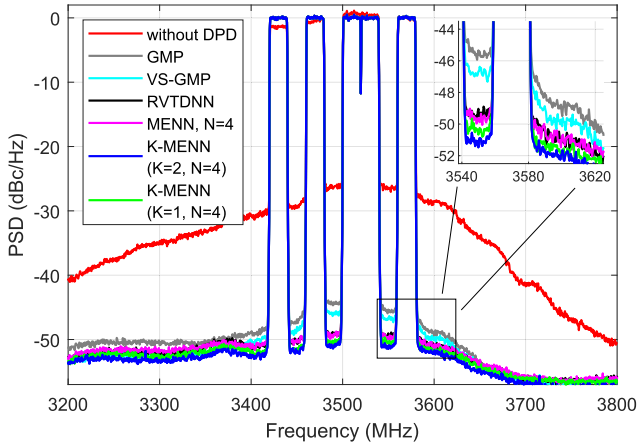


Fig. 15. Measured output power spectra for the 3.5 GHz GaN Doherty PA (160 MHz noncontiguous carrier aggregation OFDM waveform, +35.6 dBm output power) without and with DPD using selected models with around 2000 parameters.

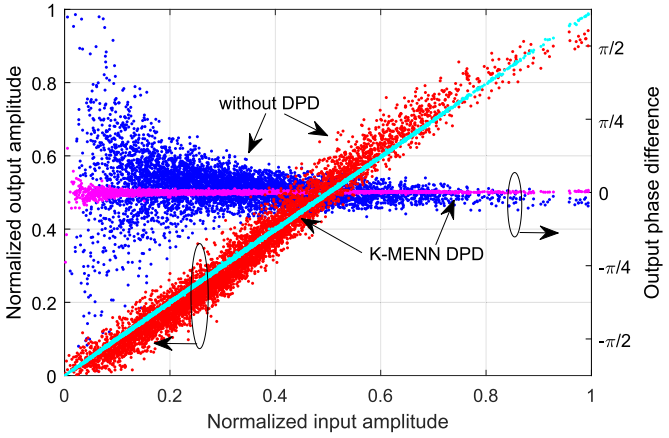


Fig. 16. Measured AM–AM and AM–PM characteristics of the 3.5 GHz GaN Doherty PA with an output power of +35.6 dBm, running a noncontiguous 160 MHz carrier aggregation waveform, consisting of five 20 MHz carriers. The linearized output has −42.9 dB NMSE and is achieved using a K-MENN with $N=4$, $K=1$, and model complexity of 3992 model parameters.

2000 parameters. The high linearization performance is clearly visible in the figure.

VI. CONCLUSION

In this article, we proposed the MENN approach with top-K sparse gating for behavioral modeling and digital predistortion of RF PAs. The K-MENN, such as the MENN, adopts the core idea of the MEs principle, by providing a set of NN experts which are combined by means of a NN gating unit. Combining the soft gating with a top-K selector introduces dynamic sparsity to the model, allowing a new trade-off between model size and run-time complexity without compromising DPD performance. With the proposed end-to-end training scheme, the gating and expert NNs are jointly optimized, enabling the experts to collaborate. Appropriate modifications of the training scheme were discussed, which allow training of the K-MENN at different levels of sparsity, including also the extreme case of a fully switched NN model. An analysis of model, run-time and training complexities was presented, which showed that, assuming a suitable processing platform, in addition to a reduction of run-time complexity, also the training complexity can be reduced when using top-K sparse gating.

The K-MENN was evaluated experimentally using two different PA units, one at 1.8 GHz and another one at 3.5 GHz, and challenging 5G NR type OFDM waveforms. The measured results confirm that the proposed K-MENN is a highly capable solution for both PA forward modeling and actual DPD-based linearization. The sparse gating allows achieving the high modeling capability of larger NN models, at a substantially lower run-time complexity. Our DPD experiments showed potential for run-time complexity reduction by about 50% for very low ACLR levels.

In our future work, we plan to investigate suitable digital processing platforms to support the sparse gating. We furthermore expect that MENN concept can adopt other NN topologies proposed in the literature, as to further decrease the complexity. Also, many degrees of freedom have not yet been investigated, such as combining experts with different capabilities or applying MENN in a more dynamic scenario, where the possibility to add or remove specific experts may help to adapt the NN to new operating conditions.

REFERENCES

[1] C. Fager, T. Eriksson, F. Barradas, K. Hausmair, T. Cunha, and J. C. Pedro, "Linearity and efficiency in 5G transmitters: New techniques for analyzing efficiency, linearity, and linearization in a 5G active antenna transmitter context," *IEEE Microw. Mag.*, vol. 20, no. 5, pp. 35–49, May 2019.

[2] A. Katz, J. Wood, and D. Chokola, "The evolution of PA linearization: From classic feedforward and feedback through analog and digital predistortion," *IEEE Microw. Mag.*, vol. 17, no. 2, pp. 32–40, Feb. 2016.

[3] U. Gustavsson et al., "Implementation challenges and opportunities in beyond-5G and 6G communication," *IEEE J. Microw.*, vol. 1, no. 1, pp. 86–100, Jan. 2021.

[4] V. Camarchia, M. Pirola, R. Quaglia, S. Jee, Y. Cho, and B. Kim, "The Doherty power amplifier: Review of recent solutions and trends," *IEEE Trans. Microw. Theory Techn.*, vol. 63, no. 2, pp. 559–571, Feb. 2015.

[5] R. Quaglia and S. Cripps, "A load modulated balanced amplifier for telecom applications," *IEEE Trans. Microw. Theory Techn.*, vol. 66, no. 3, pp. 1328–1338, Mar. 2018.

[6] J. Pang, C. Chu, Y. Li, and A. Zhu, "Broadband RF-input continuous-mode load-modulated balanced power amplifier with input phase adjustment," *IEEE Trans. Microw. Theory Techn.*, vol. 68, no. 10, pp. 4466–4478, Oct. 2020.

[7] Z. Popovic, "Amping up the PA for 5G: Efficient GaN power amplifiers with dynamic supplies," *IEEE Microw. Mag.*, vol. 18, no. 3, pp. 137–149, May 2017.

[8] S. Amin, P. Händel, and D. Rönnow, "Digital predistortion of single and concurrent dual-band radio frequency GaN amplifiers with strong nonlinear memory effects," *IEEE Trans. Microw. Theory Techn.*, vol. 65, no. 7, pp. 2453–2464, Jul. 2017.

[9] *NR Base Station (BS) Radio Transmission and Reception*, document TS 38.104, Version 18.1.0, (Release 18), 3GPP, Apr. 2023.

[10] D. R. Morgan, Z. Ma, J. Kim, M. G. Zierdt, and J. Pastalan, "A generalized memory polynomial model for digital predistortion of RF power amplifiers," *IEEE Trans. Signal Process.*, vol. 54, no. 10, pp. 3852–3860, Oct. 2006.

[11] S. Afsardoost, T. Eriksson, and C. Fager, "Digital predistortion using a vector-switched model," *IEEE Trans. Microw. Theory Techn.*, vol. 60, no. 4, pp. 1166–1174, Apr. 2012.

[12] A. Zhu, P. J. Draxler, C. Hsia, T. J. Brazil, D. F. Kimball, and P. M. Asbeck, "Digital predistortion for envelope-tracking power amplifiers using decomposed piecewise Volterra series," *IEEE Trans. Microw. Theory Techn.*, vol. 56, no. 10, pp. 2237–2247, Oct. 2008.

[13] A. Brihuega, M. Abdelaziz, L. Anttila, Y. Li, A. Zhu, and M. Valkama, "Mixture of experts approach for piecewise modeling and linearization of RF power amplifiers," *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 1, pp. 380–391, Jan. 2022.

[14] E. Guillena, W. Li, G. Montoro, R. Quaglia, and P. L. Gilabert, "Reconfigurable DPD based on ANNs for wideband load modulated balanced amplifiers under dynamic operation from 1.8 to 2.4 GHz," *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 1, pp. 453–465, Jan. 2022.

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

[16] 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 Trans. Microw. Theory Techn.*, vol. 58, no. 1, pp. 95–104, Jan. 2010.

[17] 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 Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 242–254, Jan. 2019.

[18] F. Mkadem and S. Boumaiza, "Physically inspired neural network model for RF power amplifier behavioral modeling and digital predistortion," *IEEE Trans. Microw. Theory Techn.*, vol. 59, no. 4, pp. 913–923, Apr. 2011.

[19] A. Brihuega, L. Anttila, and M. Valkama, "Neural-network-based digital predistortion for active antenna arrays under load modulation," *IEEE Microw. Wireless Compon. Lett.*, vol. 30, no. 8, pp. 843–846, Aug. 2020.

[20] Y. Zhang, Y. Li, F. Liu, and A. Zhu, "Vector decomposition based time-delay neural network behavioral model for digital predistortion of RF power amplifiers," *IEEE Access*, vol. 7, pp. 91559–91568, 2019.

[21] T. Kobal, Y. Li, X. Wang, and A. Zhu, "Digital predistortion of RF power amplifiers with phase-gated recurrent neural networks," *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 6, pp. 3291–3299, Jun. 2022.

[22] H. Li, Y. Zhang, G. Li, and F. Liu, "Vector decomposed long short-term memory model for behavioral modeling and digital predistortion for wideband RF power amplifiers," *IEEE Access*, vol. 8, pp. 63780–63789, 2020.

[23] H. Leung and S. Haykin, "The complex backpropagation algorithm," *IEEE Trans. Signal Process.*, vol. 39, no. 9, pp. 2101–2104, Sep. 1991.

[24] E. G. Lima, T. R. Cunha, and J. C. Pedro, "A physically meaningful neural network behavioral model for wireless transmitters exhibiting PM-AM/PM-PM distortions," *IEEE Trans. Microw. Theory Techn.*, vol. 59, no. 12, pp. 3512–3521, Dec. 2011.

[25] T. Kobal and A. Zhu, "Digital predistortion of RF power amplifiers with decomposed vector rotation-based recurrent neural networks," *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 11, pp. 4900–4909, Nov. 2022.

[26] Y. Yu, J. Cai, X.-W. Zhu, P. Chen, and C. Yu, "Self-sensing digital predistortion of RF power amplifiers for 6G intelligent radio," *IEEE Microw. Wireless Compon. Lett.*, vol. 32, no. 5, pp. 475–478, May 2022.

[27] X. Hu et al., "Behavioral model with multiple states based on deep neural network for power amplifiers," *IEEE Microw. Wireless Compon. Lett.*, vol. 32, no. 11, pp. 1363–1366, Nov. 2022.

[28] C. Jiang, G. Yang, R. Han, J. Tan, and F. Liu, "Gated dynamic neural network model for digital predistortion of RF power amplifiers with varying transmission configurations," *IEEE Trans. Microw. Theory Techn.*, vol. 71, no. 8, pp. 3605–3616, Feb. 2023.

[29] A. Fischer-Bühner, A. Brihuega, L. Anttila, M. D. Gomony, and M. Valkama, "Mixture of experts neural network for modeling of power amplifiers," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Jun. 2022, pp. 510–513.

[30] N. Shazeer et al., "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," 2017, *arXiv:1701.06538*.

[31] W. Fedus, B. Zoph, and N. Shazeer, "Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res.*, vol. 23, pp. 1–39, Apr. 2021.

[32] A. Barry, W. Li, J. A. Becerra, and P. L. Gilabert, "Comparison of feature selection techniques for power amplifier behavioral modeling and digital predistortion linearization," *Sensors*, vol. 21, no. 17, p. 5772, Aug. 2021.

[33] R. Hongyo, Y. Egashira, T. M. Hone, and K. Yamaguchi, "Deep neural network-based digital predistorter for Doherty power amplifiers," *IEEE Microw. Wireless Compon. Lett.*, vol. 29, no. 2, pp. 146–148, Feb. 2019.

[34] S. E. Yuksel, J. N. Wilson, and P. D. Gader, "Twenty years of mixture of experts," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 8, pp. 1177–1193, Aug. 2012.

[35] L. Xu, M. Jordan, and G. E. Hinton, "An alternative model for mixtures of experts," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 7, G. Tesauro, D. Touretzky, and T. Leen, Eds. Cambridge, MA, USA: MIT Press, Jan. 1994, pp. 633–640.

[36] C. A. M. Lima, A. L. V. Coelho, and F. J. Von Zuben, "Hybridizing mixtures of experts with support vector machines: Investigation into nonlinear dynamic systems identification," *Inf. Sci.*, vol. 177, no. 10, pp. 2049–2074, May 2007.

[37] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, Mar. 1991.

[38] D. Eigen, M. Ranzato, and I. Sutskever, "Learning factored representations in a deep mixture of experts," 2013, *arXiv:1312.4314*.

[39] P. L. Gilabert, R. N. Braithwaite, and G. Montoro, "Beyond the Moore–Penrose inverse: Strategies for the estimation of digital predistortion linearization parameters," *IEEE Microw. Mag.*, vol. 21, no. 12, pp. 34–46, Dec. 2020.

[40] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, Dec. 2014, pp. 1–15.

[41] E. Bengio, P.-L. Bacon, J. Pineau, and D. Precup, "Conditional computation in neural networks for faster models," 2015, *arXiv:1511.06297*.

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

[43] J. Chani-Cahuana, P. N. Landin, C. Fager, and T. Eriksson, "Iterative learning control for RF power amplifier linearization," *IEEE Trans. Microw. Theory Techn.*, vol. 64, no. 9, pp. 2778–2789, Sep. 2016.

[44] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, Feb. 1970.

Arne Fischer-Bühner (Graduate Student Member, IEEE) received the Diplom-Ingenieur (Dipl-Ing.) degree in electrical engineering from the Technical University Dresden, Dresden, Germany, in March 2021. He is currently pursuing the Doctorate degree at Nokia Bell Labs, Antwerp, Belgium, and Tampere University, Tampere, Finland, with Marie-Curie funded industrial.

His current research interests include AI/ML for modeling and compensation of nonlinear impairments in radio frequency (RF) transmitters, DFE signal processing, and related processing architectures and systems.

Alberto Brihuega received the B.Sc. and M.Sc. degrees in telecommunications engineering from Universidad Politécnica de Madrid, Madrid, Spain, in 2015 and 2017, respectively, and the D.Sc. (Tech.) degree (Hons.) in computing and electrical engineering from Tampere University, Tampere, Finland, in 2022.

He is currently a Senior DPD Engineer with Nokia Mobile Networks, Oulu, Finland. His current research interests include statistical and adaptive digital signal processing for compensation of hardware impairments in radio frequency (RF) transceivers, RF algorithms, and RF device modeling.

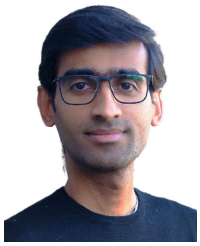
Lauri Anttila (Member, IEEE) received the M.Sc. and D.Sc. degrees (Hons.) in electrical engineering from the Tampere University of Technology (TUT), Tampere, Finland, in 2004 and 2011, respectively.

Since 2016, he has been a University Researcher with the Department of Electrical Engineering, Tampere University (formerly TUT). From 2016 to 2017, he was a Visiting Research Fellow with the Department of Electronics and Nanoengineering, Aalto University, Helsinki, Finland. He has coauthored over 100 refereed articles and three book chapters. His current research interests include radio communications and signal processing, with a focus on the radio implementation challenges in systems such as 5G, full-duplex radio, and large-scale antenna systems.

Matias Turunen is a Researcher and a Laboratory Specialist at the Department of Electrical Engineering, Tampere University (TAU), Tampere, Finland. His current research interests include digital predistortion methods, in-band full-duplex radios with an emphasis on analog radio frequency (RF) cancellation, OFDM radar, and 5G new radio systems.

Vishnu Unnikrishnan (Member, IEEE) received the M.Sc. degree in electrical engineering and the Ph.D. degree in integrated circuits and systems from Linköping University, Linköping, Sweden, in 2012 and 2016, respectively.

He is currently an Assistant Professor at the Department of Electrical Engineering, Tampere University, Tampere, Finland. From 2004 to 2009, he was with Bosch, Bangalore, India. From 2017 to 2021, he was a Post-Doctoral Researcher at the Department of Electronics and Nanoengineering, Aalto University, Espoo, Finland. His current research interests include energy-efficient high-performance integrated circuits and systems, radio/wireline transceivers, and digital implementation/enhancement of analog/mixed-signal functions in integrated circuits.



Manil Dev Gomony (Member, IEEE) received the master's degree in electrical engineering from Linköping University, Linköping, Sweden, in 2010, and the Ph.D. degree in electrical engineering from the Eindhoven University of Technology, Eindhoven, The Netherlands, in 2015.

He is currently a Senior Researcher at Nokia Bell Labs, Antwerp, Belgium, and an Assistant Professor at the Department of Electrical Engineering at Eindhoven University of Technology. His current research interests include the various aspects of low power digital hardware design starting from architecture to circuit-level, covering different processor architectures, and memory systems.

Mikko Valkama (Fellow, IEEE) received the M.Sc. (Tech.) and D.Sc. (Tech.) degrees (Hons.) from Tampere University of Technology, Tampere, Finland, in 2000 and 2001, respectively.

In 2003, he was with the Communications Systems and Signal Processing Institute, SDSU, San Diego, CA, USA, as a Visiting Research Fellow. Currently, he is a Full Professor and the Head of the Unit of Electrical Engineering, Tampere University, Tampere. His general research interests include radio communications, radio localization, and radio-based sensing, with particular emphasis on 5G and 6G mobile radio networks.