A Deep Learning Method to Predict Fading Channel in Multi-Antenna Systems

Wei Jiang*† and Hans D. Schotten†*
*German Research Center for Artificial Intelligence (DFKI)
Trippstadter Street 122, Kaiserslautern, 67663 Germany
†Technische Universität Kaiserslautern
Building 11, Paul-Ehrlich Street, Kaiserslautern, 67663 Germany

***Abstract*—Channel state information (CSI) plays a vital role in adaptive transmission systems, which adapt their transmission parameters to instantaneous channel conditions. However, the CSI tends to become outdated due to the rapid channel variation caused by multi-path fading. The inaccuracy of outdated CSI imposes a severe impact on the performance of a wide range of wireless systems, highlighting the significance of channel prediction that can combat the outdated CSI effectively. The aim of this paper is to propose a novel predictor, leveraging the strong time-series prediction capability of deep learning, where a deep recurrent neural network incorporating long short-term memory or gated recurrent unit is applied. Performance evaluation is carried out upon multi-antenna fading channels, and the numerical results in terms of prediction accuracy unveil that deep learning can bring a notable performance gain compared with the conventional predictors built on shallow neural networks.**

I. Introduction

With the aid of channel state information (CSI) at the transmitter, a wireless system is able to adaptively choose its transmission parameters, such as the transmit power, constellation size, and coding rate, to achieve great performance. In frequency-division duplex systems, the CSI is obtained through estimating pilot signals at the receiver and then fed back to the transmitter. But the CSI is often inaccurate since it becomes outdated at the time of using the selected parameters to transmit. Despite the avoidance of the feedback delay in time-division duplex systems, the processing delay still raises inaccuracy, especially in high mobility or high frequency band scenarios. It has been extensively recognized that the outdated CSI has an overwhelming impact on the performance of a wide variety of adaptive transmission systems, such as massive multiple-input multiple-output (MIMO) [1], multi-user scheduling [2], interference alignment [3], transmit antenna selection [4], opportunistic relaying [5], resource allocation [6], mobility management [7], and physical layer security [8].

To alleviate the effect of outdated CSI, a large number of algorithms and protocols have been proposed by means of passively compensating for the performance loss with a cost of scarce wireless resources [9] or aiming to achieve merely part of the full potential under the assumption of imperfect CSI [10]. In contrast, an alternative technique referred to as channel prediction [11] provides an efficient solution that improves the accuracy of CSI directly without spending extra radio resources. Its key idea is to forecast future CSI in advance with a time span that can counteract the experienced delay. Through modeling a wireless channel into a set of radio propagation parameters, two statistical prediction approaches - autoregressive [12] and parametric model [13] - have been proposed. However, the modelling is fossilized, leading to a gap between these models and real channels, and - in addition - the parameter estimation process relying on complex algorithms such as MUSIC and ESPRIT [14] is tedious, destroying its applicability in practical systems [15]. As a data-driven technique, neural networks [16] can avoid this estimation process and therefore attracted the interest from researchers. Taking advantage of its capability on time-series prediction [17], a recurrent neural network (RNN) was first proposed to predict single-antenna flat-fading channels in [18] and was further extended to MIMO channels in [19], followed by [20] that designed a frequency-domain RNN predictor for frequency-selective MIMO channels. Besides, the authors of [21] argued for a real-valued RNN to implement a multi-step predictor and further verified its effectiveness in a multiple transmit antenna selection system [22]. More details about the impact of outdated CSI, existing model-based and data-driven predictive methods can refer to [23].

By far the aforementioned prediction approaches are still limited to *shallow* neural networks, to the best knowledge of the authors, deep learning (DL) is still untouched in this field. This paper aims to propose a novel predictor leveraging the strong time-series prediction capability of a deep recurrent neural network, which incorporates long short-term memory (LSTM) or gated recurrent unit (GRU). Performance assessment in terms of prediction accuracy is carried out in multi-antenna fading channels, taking into account a number of factors such as additive noise, mobility (measured by the Doppler shift), MIMO scale, activation function, and number of hidden neurons. Some representative numerical results are illustrated, unveiling that DL can notably improve prediction accuracy in comparison with the conventional predictors built on shallow recurrent neural networks. The rest of this paper is organized as follows: Section II introduces a deep recurrent

*This work was supported by German Federal Ministry of Education and Research (BMBF) through TACNET4.0 project (Grant no. *KIS15GTI007*) and KICK project (Grant no. *16KIS1105*).

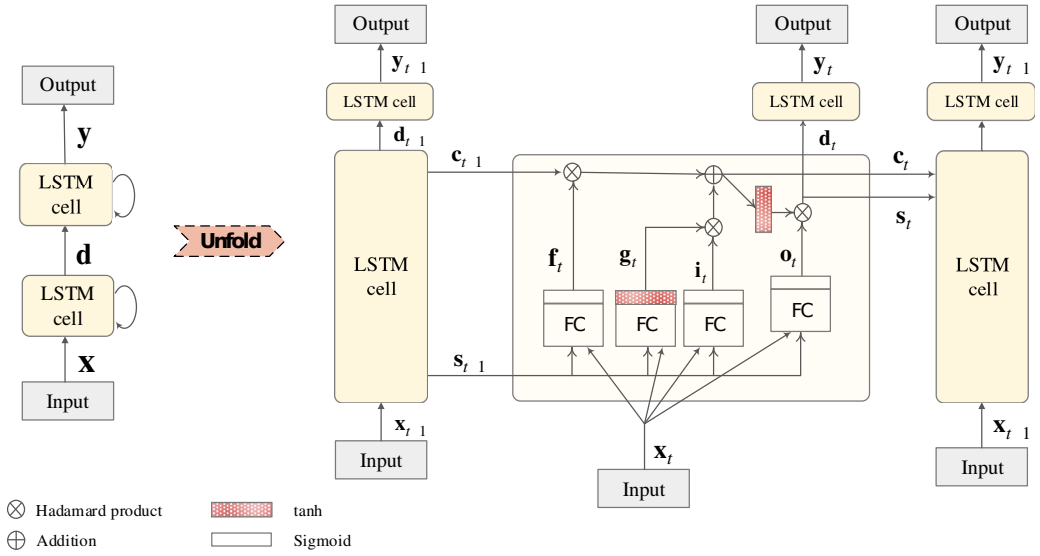


Fig. 1. Schematic diagram of an example two-hidden-layer deep recurrent network with LSTM memory cells.

network incorporating LSTM or GRU. Section III details the simulation configuration and presents the numerical results. Finally, Section IV closes this paper with our remarks.

II. Deep Recurrent Neural Networks

A recurrent neural network is a class of artificial neural networks that can well process sequences of input data by storing historical information in its internal state, exhibiting great potential in time-series prediction [17]. However, it suffers from the problems of gradient exploding or vanishing during its training process, where the back-propagated error signals either tend to be very large, leading to oscillating weights and biases, or tend to zero that implies a prohibitively long training time. To this end, Hochreiter and Schmidhuber proposed the LSTM model in 1997 in their pioneer work of [24]. Though its history is not long, LSTM already broke technological records in machine translation and speech recognition, demonstrated by many unprecedented services with great commercial success, such as Google Translate and Apple iPhone Siri.

The key innovation to enlarge LSTM's capability to deal with long-term dependencies is the introduction of a memory cell and multiplicative gates that regulate the pattern of information flow. The *input* gate protects the memory contents stored in the cell from perturbation by irrelevant inputs, the *forget* gate decides how much the historical information remains in the cell, and the *output* gate controls the extent to which the memory information applied to generate the output. Unlike a feed-forward network where activation flow only in one direction from the input to the output layer, each LSTM memory cell has a feedback loop that allows self-recurrent connection. A deep recurrent network can be built by stacking multiple LSTM memory cells. Without loss of generality, Fig.1 shows an example deep network that consists of an input layer, two hidden layers, and an output layer.

At an arbitrary time step, as shown in the left part of Fig.1, a data vector $\mathbf{x}$ is fed into the first cell as its input. Together with the self-recurrent unit feeding back from the previous time step, an intermediate unit $\mathbf{d}$ is generated and forwarded to the second cell, which gets the output $\mathbf{y}$. Unrolling the network through time, as illustrated in the right part of Fig.1, a memory cell holds two internal states - the short-term state $\mathbf{s}_{t-1}$ and the long-term state $\mathbf{c}_{t-1}$ - at time step $t-1$. Traversing the cell from the left to the right, $\mathbf{c}_{t-1}$ first throws away some memories at the forget gate, integrates new contents selected by the input gate, and then sends out as the current long-term state $\mathbf{c}_t$. Looking inside the first cell as an example, $\mathbf{x}_t$ and $\mathbf{s}_{t-1}$ are fed into four different fully connected (FC) layers, generating three gate vectors:

$$\mathbf{f}_t = \sigma_g\left(\mathbf{W}_f\mathbf{x}_t + \mathbf{U}_f\mathbf{s}_{t-1} + \mathbf{b}_f\right), \tag{1}$$

$$\mathbf{i}_t = \sigma_g\left(\mathbf{W}_i\mathbf{x}_t + \mathbf{U}_i\mathbf{s}_{t-1} + \mathbf{b}_i\right), \tag{2}$$

$$\mathbf{o}_t = \sigma_g\left(\mathbf{W}_o\mathbf{x}_t + \mathbf{U}_o\mathbf{s}_{t-1} + \mathbf{b}_o\right), \tag{3}$$

where $\mathbf{W}$ and $\mathbf{U}$ are weight matrices for the FC layers, $\mathbf{b}$ stands for bias vectors, the subscripts f, i, and o associate with the forget, input, and output gate, respectively, and σ_g represents the *sigmoid* activation function, defining by

$$\sigma_g(x) = \frac{1}{1+e^{-x}}. \tag{4}$$

Dropping a few old memories and adding something new, $\mathbf{c}_{t-1}$ is transformed into the current long-term state

$$\mathbf{c}_t = \mathbf{f}_t \otimes \mathbf{c}_{t-1} + \mathbf{i}_t \otimes \sigma_h\left(\mathbf{W}_g\mathbf{x}_t + \mathbf{U}_g\mathbf{s}_{t-1} + \mathbf{b}_g\right), \tag{5}$$

where $\otimes$ denotes the Hadamard product (element-wise multiplication) for matrices, and σ_h is the *hyperbolic tangent* function denoted by $\tanh$, i.e.,

$$\sigma_h(x) = \frac{e^{2x}-1}{e^{2x}+1}. \tag{6}$$

In addition, $\mathbf{c}_t$ passes through the $\tanh$ function and is further filtered by the output gate to produce the short-term memory, as well as the intermediate unit, i.e.,

$$\mathbf{s}_t = \mathbf{d}_t = \mathbf{o}_t \otimes \sigma_h\left(\mathbf{c}_t\right). \tag{7}$$

Since the emergence of LSTM, its original architecture continues to evolve. Cho *et al.* [25] proposed a simplified version with fewer parameters in 2014, known as GRU, which exhibits even better performance over LSTM on certain smaller data sets. In a GRU cell, the short- and long-term states are merged into a single one, and a new gate marked by $\mathbf{z}_t$ is used to replace the forget and input gates, namely,

$$\mathbf{z}_t = \sigma_g\left(\mathbf{W}_z\mathbf{x}_t + \mathbf{U}_z\mathbf{s}_{t-1} + \mathbf{b}_z\right). \tag{8}$$

The output gate is removed, but an internal state $\mathbf{r}_t$ is applied:

$$\mathbf{r}_t = \sigma_g\left(\mathbf{W}_r\mathbf{x}_t + \mathbf{U}_r\mathbf{s}_{t-1} + \mathbf{b}_r\right). \tag{9}$$

Likewise, the hidden state at the previous time step transverses the cell, drops some old memories, and loads some new information, resulting in the current state that is given by

$$\begin{aligned}\mathbf{s}_t =&(1-\mathbf{z}_t)\otimes\mathbf{s}_{t-1}\\ &+\mathbf{z}_t\otimes\sigma_h\left(\mathbf{W}_s\mathbf{x}_t+\mathbf{U}_s(\mathbf{r}_t\otimes\mathbf{s}_{t-1})+\mathbf{b}_s\right).\end{aligned} \tag{10}$$

III. Performance Evaluation

A performance comparison between the proposed DL predictor and the previous predictors based on shallow recurrent neural networks is conducted. Some representative numerical results in terms of prediction accuracy are illustrated in this section. We apply independent and identically distributed (*i.i.d*) flat-fading MIMO channels having 4 transmit antennas and a single receive antenna as the baseline for comparing the performance. Each subchannel follows the Rayleigh distribution with an average power gain of 0dB, where its channel gain h is zero-mean circularly-symmetric complex Gaussian random variable with a variance of 1, i.e., $h\sim\mathcal{CN}(0,1)$. The maximal Doppler frequency shift is set to f_d=100Hz, emulating fast time-varying environment. To build a data set, continuous-time channel responses are sampled with a rate of f_s=1KHz, satisfying the assumption of flat fading. The data set contains a series of 10^4 consecutive CSI $\{\mathbf{H}[t]\,|t{=}1,2,\ldots,10000\}$, with a time interval of T_s=1ms. In our simulation, 75% of the data is allocated for training and the remaining 25% is testing data.

A conventional RNN predictor, denoted by *RNN* in the legend of the figures, is actually a 3-layer recurrent network consisting of an input layer, an output layer, and a hidden layer. If the hidden layer is replaced with an LSTM or GRU memory cell, the network is transformed into a shallow LSTM or GRU network, respectively, marked by *LSTM* and *GRU* in the legend. The default number of hidden neurons is 30, and $\tanh$ is applied as the default activation function at each hidden neuron. Regarding the proposed predictor, a two-hidden-layer recurrent network with a GRU memory cell at each layer is employed as a representative setup of deep learning, notated by *DL* in the legend. For a fair comparison, the total number of hidden neurons for two hidden layers is also 30. A training process starts from an initial state where all weights and biases are randomly selected. At time step t, recalling Fig.1, the input of the memory cell is $\mathbf{x}_t = \mathbf{H}[t]$ and the output is its D-step ahead prediction, i.e., $\mathbf{y}_t = \hat{\mathbf{H}}[t+D]$. Mean squared error (MSE), a metric for measuring prediction accuracy, is chosen as the cost function during the training phase. It is defined as

$$\text{MSE} = \frac{1}{T}\sum_{t=1}^{T}\left\|\hat{\mathbf{H}}[t+D]-\mathbf{H}[t+D]\right\|^2, \tag{11}$$

where T is the total number of channel samples used for evaluation, $\hat{\mathbf{H}}[t+D]$ denotes the predicted channel matrix at time t, $\mathbf{H}[t+D]$ stands for its actual value, and $\|\cdot\|$ notates the Frobenius norm of a matrix. Using the *batch* training method, a batch of 256 samples is fed into the network at each epoch, the resultant outputs are compared with the desired values and the error signals are propagated back through the network to update the weights and biases by training algorithms such as the Adam optimizer used in our simulation. The training process is iteratively carried out until the network reaches a

TABLE I
Simulation Parameters

Parameters	Values
Sampling rate	$f_s = 1$KHz
Maximum Doppler shift	$f_d = 100$Hz
Channel model	Rayleigh fading
Doppler spectrum	Jakes's model
(default) MIMO scale	4×1
Batch size	256
Dataset size	10^4
Training algorithm	Adam optimizer
Cost function	MSE
(default) Prediction length	1ms
(default) Actuation function	tanh
(default) Hidden neurons	30

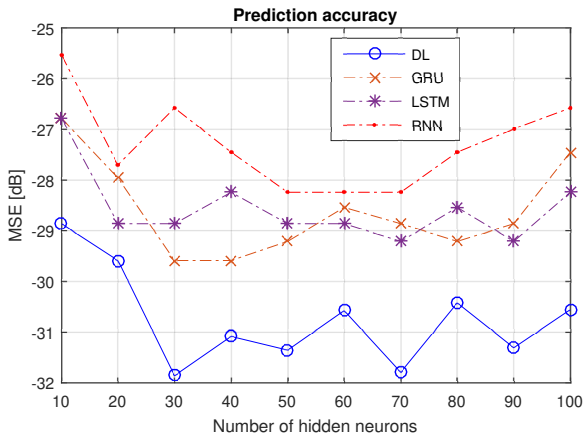
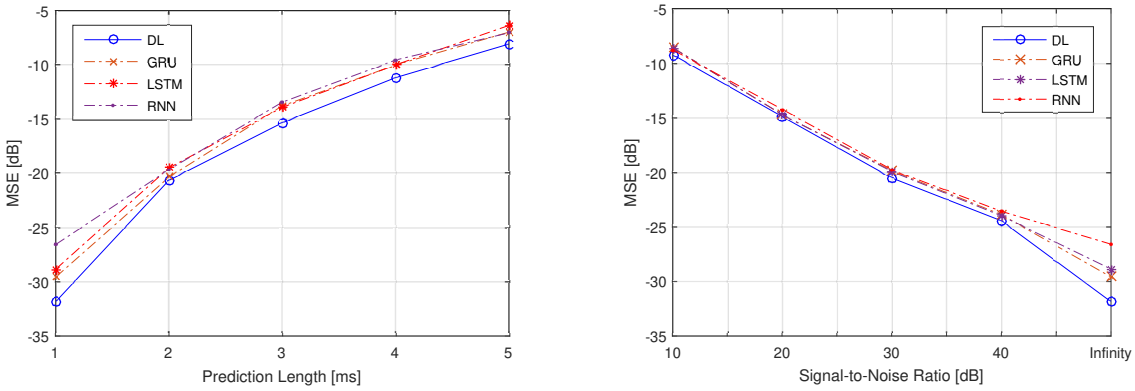
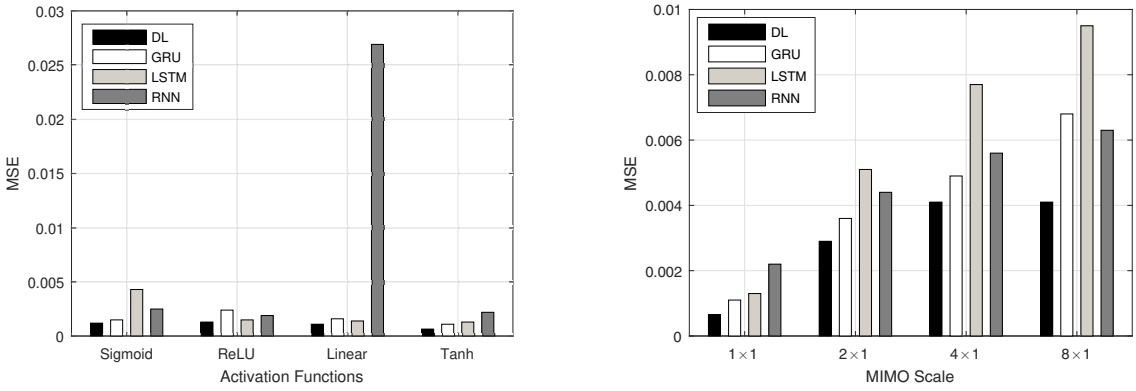


Fig. 2. The results of prediction accuracy for four different predictors.



(a) (b)

Fig. 3. Performance comparisons in terms of: (a) prediction time steps and (b) the strength of noise.



(a) (b)

Fig. 4. Performance comparisons with respect to: (a) activation function and (b) the size of MIMO.

certain convergence condition. Once it completes, the trained network can be employed to predict a future CSI.

The prediction accuracy of the DL predictor as a function of the number of hidden neurons is evaluated, in comparison with those of three conventional predictors. Starting from 10 hidden neurons, as shown in Fig.2, the best result achieved by the conventional predictors is MSE=0.0021. The proposed predictor reduces the MSE to 0.0013 under the same number of hidden neurons. In other words, deep learning can improve the accuracy of channel prediction remarkably by shifting part of hidden neurons to an extra layer, implying comparable computational complexity, rather than adding more neurons with higher complexity. Note that Fig.2 and the following Fig.3 display MSE values in decibels (dB) for a clear illustration, calculating from MSE_{dB}=10 $\log_{10}(MSE)$. With the growth of hidden neurons, the performance of the predictors improves at first because the network capability is strengthened. However, after a turn point, at which the network capability saturates, the occurrence of data overfitting worsens performance. For instance, the minimal MSE of the DL predictor happens when the total number of hidden neurons is 30. As shown in Fig.2, the optimal results of RNN, LSTM, and GRU are 0.0015, 0.0012, and 0.0011, respectively, while deep learning lowers such value to 0.0007. Hence, deep learning remarkably outperforms shallow neural networks, regardless of different number of hidden neurons.

The MSE results of the predictors in terms of different prediction lengths are illustrated in Fig.3a. To begin with one-step ahead mode, i.e., D=1, which corresponds to a prediction length of 1ms as the sampling rate is f_s = 1KHz. The proposed predictor clearly outperforms three conventional ones, with performance gains of 3∼5dB. Increasing D from 1 up to 5 incrementally, the results on longer ranges from 1ms to 5ms are obtained. The prediction accuracy becomes worse with the growing prediction length, because channel's temporal correlation weakens. Under different prediction lengths, deep learning can receive a gain of 1∼2dB compared with others. It is worth emphasizing again that the benefit is achieved under the identical number of hidden neurons, implying comparable complexity. In addition, the effect of noise is observed and illustrated in Fig.3b. The horizonal axis of the figure is the signal-to-noise ratio (SNR) of the channel samples, where the rightmost "infinity" corresponds to an extremely large SNR, i.e., noiseless or noise-free. In noisy channels, three conventional predictors have approximately the same (bad) MSE results. It implies that an effort to improve accuracy against

noise through using an LSTM or GRU cell to enhance the capability of a shallow neural network is invalid. In contrast, deep learning performs well and set a clear performance border with others, exhibiting a gain of nearly $1\mathrm{dB}$.

The performance comparison with respect to different activation functions and different MIMO scales are also carried out. As illustrated in Fig.4a, the recurrent networks can collectively achieve their best performance by using $\tanh$, which is the default activation function in our simulation. In particular, the prediction accuracy of the conventional RNN with the linear activation function is obviously weak, because it cannot deal with nonlinearity well. In this case, LSTM and GRU boosts the performance with an order of magnitude, dropping the MSE value from over 0.025 to less than 0.0025. As we can observe from the figure, deep learning has the best performance under all kinds of activation functions. Looking finally at the impact of MIMO scale, as illustrated in Fig.4b, the performance degrades with the increasing number of antennas. That is because a neural network has to process more dimensions of the input data, bringing extra burden in both the training and predicting. Comparing fairly under the same total number of hidden neurons, a DL-based predictor can achieve clear improvement either in a smaller or in a larger MIMO scale, surpassing all shallow neural networks.

IV. Conclusions

In this paper, a novel channel predictor empowered by a deep recurrent neural network incorporating long short-term memory or gated recurrent unit, were proposed for multi-path fading channels in wireless multi-antenna systems. Performance assessment taking into account a number of affecting factors such as prediction range, number of hidden neurons, additive noise, activation function, and MIMO scale, were conducted, in comparison with three conventional predictors based on shallow recurrent neural networks. The numerical results in terms of mean squared error revealed that deep learning can effectively improve the accuracy in the prediction of MIMO fading channels, with remarkable performance gains. The positive outcomes reported from this paper will encourage a further exploration of deep learning not only in the field of fading channel prediction but also other aspects of wireless communications.

References

[1] K. T. Truong and R. W. Heath, "Effects of channel aging in Massive MIMO systems," *Journal of Commun. and Net.*, vol. 15, no. 4, pp. 338–351, Sep. 2013.

[2] Q. Wang *et al.*, "Multi-user and single-user throughputs for downlink MIMO channels with outdated channel state information," *IEEE Wireless Commun. Lett.*, vol. 3, no. 3, pp. 321–324, 2014.

[3] P. Aquilina and T. Ratnarajah, "Performance analysis of IA techniques in the MIMO IBC with imperfect CSI," *IEEE Trans. Commun.*, vol. 63, no. 4, pp. 1259–1270, 2015.

[4] X. Yu *et al.*, "Unified performance analysis of transmit antenna selection with OSTBC and imperfect CSI over Nakagami-m fading channels," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 494–508, 2017.

[5] J. L. Vicario *et al.*, "Opportunistic relay selection with outdated CSI: Outage probability and diversity analysis," *IEEE Trans. Wireless Commun.*, vol. 8, no. 6, pp. 2872–2876, Jun. 2009.

[6] Z. Wang *et al.*, "Resource allocation in OFDMA networks with imperfect channel state information," *IEEE Commun. Lett.*, vol. 18, no. 9, pp. 1611–1614, 2014.

[7] Y. Teng *et al.*, "Effect of outdated CSI on handover decisions in dense networks," *IEEE Commun. Lett.*, vol. 21, no. 10, pp. 2238–2241, 2017.

[8] A. Hyadi *et al.*, "An overview of physical layer security in wireless communication systems with CSIT uncertainty," *IEEE Access*, vol. 4, pp. 6121–6132, 2016.

[9] W. Jiang *et al.*, "A robust opportunistic relaying strategy for co-operative wireless communications," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2642–2655, Apr. 2016.

[10] D. J. Love *et al.*, "An overview of limited feedback in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1341–1365, 2008.

[11] A. Duel-Hallen, "Fading channel prediction for mobile radio adaptive transmission systems," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2299–2313, Dec. 2007.

[12] A. Duel-Hallen *et al.*, "Long-range prediction of fading signals," *IEEE Signal Process. Mag.*, vol. 17, no. 3, pp. 62–75, May 2000.

[13] R. O. Adeogun *et al.*, "Extrapolation of MIMO mobile-to-mobile wireless channels using parametric-model-based prediction," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4487–4498, 2014.

[14] W. Gardner, "Simplification of MUSIC and ESPRIT by exploitation of cyclostationarity," *Proceedings of the IEEE*, vol. 76, no. 7, pp. 845–847, Jul. 1988.

[15] W. Jiang *et al.*, "A comparison of wireless channel predictors: Artificial Intelligence versus Kalman filter," in *Proc. of IEEE Intl. Commu. Conf. (ICC)*, Shanghai, China, May 2019.

[16] W. Jiang and H. D. Schotten, "Neural network-based fading channel prediction: A comprehensive overview," *IEEE Access*, vol. 7, pp. 118 112–118 124, Aug. 2019.

[17] J. Connor *et al.*, "Recurrent neural networks and robust time series prediction," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 240–254, Mar. 1994.

[18] W. Liu *et al.*, "Recurrent neural network based narrowband channel prediction," in *Proc. IEEE Vehicular Tech. Conf. (VTC)*, Melbourne, Australia, May 2006.

[19] K. T. Truong and R. W. Heath, "Fading channel prediction based on combination of complex-valued neural networks and chirp Z-transform," *IEEE Trans. Neural Netw.*, vol. 25, no. 9, pp. 1686–1695, Sep. 2014.

[20] W. Jiang and H. D. Schotten, "Recurrent neural network-based frequency-domain channel prediction for wideband communications," in *Proc. IEEE Vehicular Tech. Conf. (VTC)*, Kuala Lumpur, Malaysia, Apr. 2019.

[21] W. Jiang *et al.*, "Multi-antenna fading channel prediction empowered by artificial intelligence," in *Proc. IEEE Vehicular Tech. Conf. (VTC)*, Chicago, USA, Aug. 2018.

[22] W. Jiang and H. Schotten, "Neural network-based channel prediction and its performance in multi-antenna systems," in *Proc. IEEE Vehicular Tech. Conf. (VTC)*, Chicago, USA, Aug. 2018.

[23] W. Jiang *et al.*, "Neural network based wireless channel prediction," in *Machine Learning for Future Wireless Communications*, F. L. Luo, Ed. United Kindom: John Wiley&Sons and IEEE Press, Dec. 2019, ch. 16.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Dec. 1997.

[25] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *preprint arXiv:1406.1078*, Jun. 2014.