# Lightweight Spectrum Prediction Based on Knowledge Distillation

*Runmeng CHENG [1, 2], Jianzhao ZHANG [2], Junquan DENG [2], Yanping ZHU [1]*

[1] Nanjing University of Information Science and Technology, Nanjing, China
[2] The Sixty-Third Research Institute, National University of Defense Technology, Nanjing, China

20211249676@nuist.edu.cn, jianzhao63s@nudt.edu.cn, jqdeng@nudt.edu.cn, 001520@nuist.edu.cn

Submitted May 26, 2023 / Accepted August 20, 2023 / Online first October 3, 2023

**Abstract.** *To address the challenges of increasing complexity and larger number of training samples required for high-accuracy spectrum prediction, we propose a novel lightweight model, leveraging a temporal convolutional network (TCN) and knowledge distillation. First, the prediction accuracy of TCN is enhanced via a self-transfer method. Then, we design a two-branch network which can extract the spectrum features efficiently. By employing knowledge distillation, we transfer the knowledge from TCN to the two-branch network, resulting in improved accuracy for spectrum prediction of the lightweight network. Experimental results show that the proposed model can improve accuracy by 19.5% compared to the widely-used LSTM model with sufficient historical data and reduces 71.1% parameters to be trained. Furthermore, the prediction accuracy is improved by 17.9% compared to Gated Recurrent Units (GRU) in the scenarios with scarce historical data.*

## Keywords

Spectrum prediction, knowledge distillation, temporal convolutional network, lightweight networks, few-shot learning

## 1. Introduction

In the heterogeneous spectrum sharing networks, spectrum prediction enables the secondary users (SUs) to learn the spectrum usage patterns of the primary users (PUs), so the spectrum holes can be identified proactively and the optimal spectrum access strategy can be obtained by the SUs on-line. This reduces the communication delay and improves the channel throughput. In recent years, a number of spectrum prediction models have been proposed, including moving average [1], [2], hidden Markov model, Bayesian reasoning [3], multilayer perceptron [4] and recurrent network [5], etc. However, most existing studies assumed that the training samples were abundant and unabridged. In practice, when the spectrum usages of PUs or the electromagnetic environment changes, it is challenging to ensure the pre-trained models perform well due to the lack of site-specific historical data. At the same time, existing models usually come with high complexity to attain high predicting accuracy, and is difficult to be implemented in mobile edge devices with limited storage and computing resources.

Considering the uncertainty of data collection behaviors by mobile devices and dynamic changes of the spectrum environment in wireless communication [6], researchers have paid attention to the scenarios with incomplete or insufficient data. Literature [7] considered the incompleteness of historical data and converted the two-dimensional spectrum prediction problem into a matrix completion problem. Ding et al. [8] analyzed the impact of abnormal data on the rank distribution of the spectrum matrix, and developed an optimization method to address the problem via matrix recovery theory. Based on the work in [8], a robust online spectrum prediction framework (ROSP) was proposed in [9], and a joint optimization algorithm for matrix completion and matrix recovery was designed and tested with real spectrum data. In [10], a new spectrum predicting approach for the electromagnetic countermeasure environment was proposed, in which transfer learning was adopted to tackle the issue of sparse spectrum data. Furthermore, the similarity of the spectrum data among different scenarios was measured and the applicability of the single frequency point training model to other bands could be obtained. This innovative approach offered a fresh perspective on the spectrum prediction problem. In [11], the generative adversarial network (GAN) and deep transfer learning was combined to construct a spectrum prediction model suitable for different frequency bands. However, it required a large amount of similar data of the target domain for prediction, and the network structure was complex and difficult for training. Therefore, in [12], the transfer learning and meta-learning was combined for the spectrum prediction, for which the meta-learning was adopted to learn experiences gained from similar spectrum prediction tasks. This approach resulted in improved adaptability in terms of cross-band prediction when compared to the solution proposed in [11]. In summary, a few schemes have been proposed to address the spectrum prediction problems with relatively scarce historical data, yet

DOI: 10.13164/re.2023.0469

the complexity of the model and the difficulty in the actual deployment of the model have not been fully considered. To account for this problem, the knowledge distillation approach is investigated in this study to construct a lightweight prediction model.

Knowledge distillation was initially introduced by Hinton et al. [13] to transfer knowledge from larger models to smaller ones. This technique enables the simpler models to achieve similar performances compared with the complex teacher models, while with reduced model parameters. Since its introduction, various variants and network architectures have been developed for different practical applications [14]. For example, a prediction model was proposed in [15] for the furnace temperature based on the transfer learning and knowledge distillation, which used a generative adversarial loss to facilitate the transfer process and established a distillation network based on multitask learning to address the high delay of the deep transfer network. A recent study [16] presented a hybrid mode to predict the remaining service life of aircraft engines, which employed two knowledge distillations. The first distillation is heterogeneous, which can compress the model and accelerate the training, while the second isomorphic distillation aims to improve the model prediction accuracy without changing the model structure. In [17], knowledge distillation was explored in the context of time series classification. The results verified that employing knowledge distillation techniques improved the performance of small-scale convolutional networks on multiple datasets, while also reducing computational costs and storage requirements. The above methods demonstrate that knowledge distillation can significantly improve the prediction accuracy and reduce the model complexity in diverse applications.

In this paper, a knowledge distillation model is proposed to address the problems of limited historical data and high complexity in the spectrum prediction. We design a framework combining knowledge distillation and temporal convolutional network (TCN) to construct a lightweight and fast prediction model, which can achieve accurate prediction even with a small number of available samples. The main contributions are summarized as follows:

- A novel prediction model based on TCN has been designed for spectrum prediction. We use knowledge distillation method to transfer the knowledge learned from a large model to a small model, which can accelerate the spectrum prediction process while attains high prediction accuracies.
- We use self-transfer learning scheme to update the parameters of TCN by freezing certain layers and updating only some of the weights. This approach avoids the problem of information omission in the hidden layer caused by a high hole factor. Additionally, we construct a two-branch network that uses the low-dimensional representation of the data for both reconstruction and prediction, making the proposed model more efficient and robust.
- We conduct extensive experiments with real spectrum data, results demonstrate that the proposed method performs well in scenarios with sufficient and scarce historical data. Moreover, this approach offers a simplified implementation and deployment, even in resource-constrained environments

The rest of this paper is organized as follows. Section 2 presents the spectrum prediction definition and the principle of knowledge distillation. Section 3 proposes a TCN teacher model optimization method based on self-transfer and constructs a two-branch network student model. Section 4 gives the results and discussion of experimental simulations. Finally, we conclude the whole paper in Sec. 5.

## 2. System Model

### 2.1 Problem Statement

We consider a typical centralized cognitive radio (CR) system with a fusion center (FC) and $K$ sparsely distributed SUs. As shown in Fig. 1, the SUs continuously monitor $N$ frequency bands and send the received signal strength (RSS) to FC [11], [18], FC collects the RSSs form spectrum data $\mathbf{D} = \{\mathbf{f}_1,\mathbf{f}_2,\ldots,\mathbf{f}_n,\ldots,\mathbf{f}_N\}$. To facilitate the model for time series prediction [19], we transform the spectrum data $\mathbf{f}_n = \{x_1,x_2,\ldots,x_t,\ \ x_T\}$ by using a sliding window with a length of $c + 1$. We define the first $c$ samples as an input vector $\mathbf{s}_t = \{x_{t-c+1}, x_{t-c+2},\ldots,x_t\}$ and the last sample as the ground truth $y_t = x_{t+1}$, which be predicted. We then have a transformed dataset $\mathbf{f}_n = \{(\mathbf{s}_t, y_t)\}^{m}_{t=1}$ at frequency point $n$, where $m = T - c$. The goal of spectrum prediction is to train a model to predict $y_t$ given the observed spectrum data $\mathbf{s}_t$. Concisely, the following formula is used

$$\hat{y}_t = \arg\max p(y_t \mid \mathbf{s}_t) \tag{1}$$

where $\hat{y}_t$ is the expected output, $y_t$ is the target output, and the predicted value $\hat{y}_t$ should be close to the target value $y_t$.

### 2.2 Knowledge Distillation Model

Knowledge distillation is a deep learning method that transfers knowledge from a complex and highly accurate teacher model to a simpler student model with relatively less performance loss [20]. The teacher model is trained on historical spectrum data and is capable of capturing features that the student model may have difficulty in learning. To extract higher-level features and facilitate enhanced learning for the student model, we use the intermediate layer output of the teacher model as the input for the student model. Two loss functions are to be minimized, namely soft loss and hard loss, to enable the student model to learn the knowledge of the teacher model.

We use the soft loss to measure the difference between the outputs of the teacher model and the student

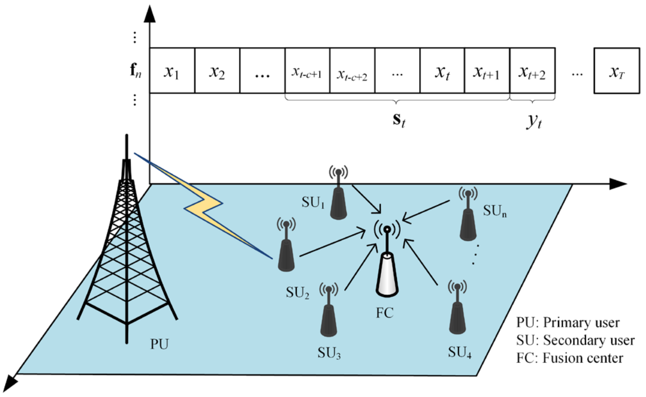


**Fig. 1.** Source of data and division.

model. The goal of soft loss is to maintain the knowledge of teacher model. The output of the teacher model is defined as a soft label, which is also taken as the target variable for the student model. The formula for soft loss is defined as

$$L_{\text{soft}} = \frac{1}{m}\sum_{t=1}^{m}(f^{\text{S}}(\mathbf{s}_t') - f^{\text{T}}(\mathbf{s}_t))^2 \tag{2}$$

where $f^{\text{T}}(\cdot)$ is the teacher model, $f^{\text{S}}(\cdot)$ is the student model, $\mathbf{s}_t$ is the input vector, and $\mathbf{s}_t'$ is the intermediate output of the teacher model which will be discussed in Sec. 3.1.

The hard loss is incorporated to evaluate the prediction accuracy of the student model by measuring the difference between the output of the student model and the ground truth. The ground truth $y_t$, namely the hard label, is used as the target variable of the student model. Then the hard loss is can be formulated as

$$L_{\text{hard}} = \frac{1}{m}\sum_{t=1}^{m}(f^{\text{S}}(\mathbf{s}_t') - y_t)^2. \tag{3}$$

By minimizing both the soft loss and hard loss functions, the student model can achieve high predictive accuracy while reduce computational costs. This approach improves operational efficiency and enhances the adaptability of the models to diverse environments with limited resources.

# 3. Spectrum Prediction Based on Knowledge Distillation

## 3.1 A Framework for Knowledge Distillation Based on Intermediate Layer

The proposed knowledge distillation framework is illustrated in Fig. 2. We implement an intermediate layer on the teacher model to provide more critical data features for the student model. At the same time, knowledge distillation constrains the output of the student model from both the soft and the hard loss functions to improve the prediction performance of the student model. Additionally, we use a factor to balance the generalization ability of learning new data and preserving existing knowledge. If the teacher model is more powerful than the student model, a larger balance factor can be selected to ensure that the student model can learn more existing knowledge. In contrast, a smaller balance factor is enough if both the teacher model and the student model have advanced skills. Then the proposed model parameters are updated by minimizing the following function

$$L_{\text{total}} = \alpha L_{\text{soft}} + (1-\alpha)L_{\text{hard}} \tag{4}$$

where $\alpha$ is the balancing factor following in [0, 1]. Due to the higher complexity and accuracy compared to the student model, we set $\alpha = 0.6$.

## 3.2 Construction of the Teacher Model TCN and the Self-transfer Optimization

Traditional recurrent networks (RNNs) perform well in time series prediction for their inherent memory capabilities. However, in practice, there is an inevitable internal design problem, i.e., only one time step can be handled at a time. This serial computation design results in high memory consumption during training. To address the issue, TCN is taken as the teacher model for distillation in this study, which allows parallel computation and the training process can be accelerated consequently.

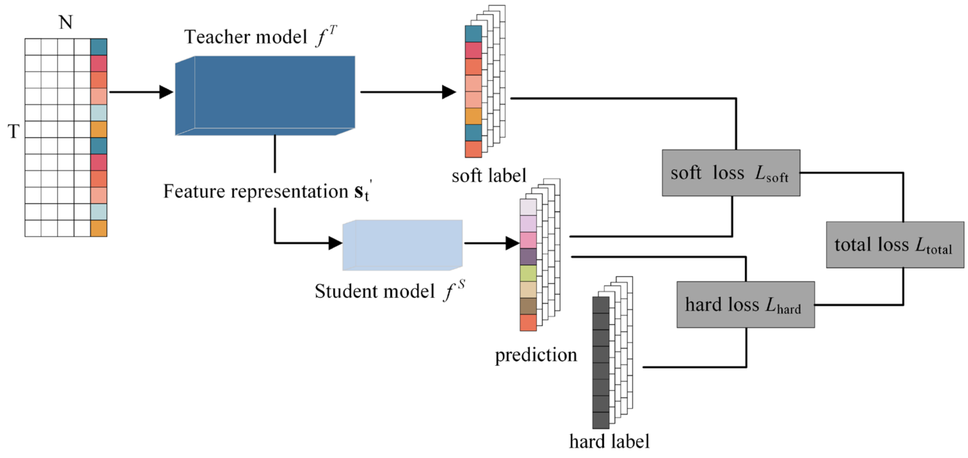

N
T
Teacher model $f^T$
soft label
Feature representation $\mathbf{s}_t'$
soft loss $L_{\text{soft}}$
total loss $L_{\text{total}}$
Student model $f^S$
prediction
hard loss $L_{\text{hard}}$
hard label

**Fig. 2.** The structure of the proposed model.

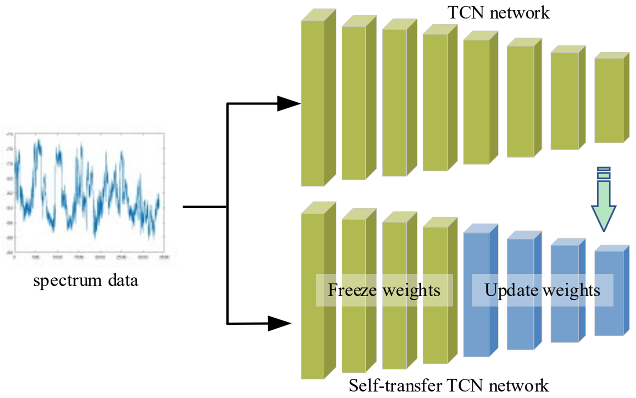


**Fig. 3.** Parameter optimization of teacher model based on self-transfer.

The TCN model is composed of causal convolution layers, dilated causal convolution layers, and residual blocks. The unidirectional nature of causal convolution ensures that future spectral values are predicted based solely on the historical momentary spectrum data. Dilated causal convolution can expand the receptive field of convolutional layers without increasing the number of parameters. This is achieved by introducing a dilation factor that controls the interval size in the convolutional kernel. The dilation rate of each layer increases exponentially by a factor of 2, enabling the higher-level convolutional kernels to capture longer dependencies.

Each layer of the TCN subsamples the input sequence and compresses it into a shorter representation. However, as the numbers of TCN layers increase, the hole factor also increases, potentially causing the loss of vital information in the hidden layers [21]. To improve TCN prediction accuracy and to capture the long-term dependencies better in the sequence [22], we adopted a method of freezing the shallow weights and retraining to update the unfrozen weights. As shown in Fig. 3, we use historical spectrum data training the original TCN model. In order to protect hidden layers of important information, we freeze some of the TCN network layers and update the remaining layers, then we get a more accurate prediction model TCN as the teacher model. More details on hyperparameters are described in Sec. 4.3.

## 3.3 Two-branch Network Student Model

In this paper, we construct a two-branch network that extracts the spectrum data features completely. The architecture is depicted in Fig. 4, it consists of two branches, namely the Encoder-Reconstruction branch $f_{\mathrm{rec}}$ and the Encoder-Prediction branch $f_{\mathrm{pre}}$. These branches serve different purposes.

The Encoder-Reconstruction branch compresses the data into a coded form, which is then reconstructed into an output that closely resembles the original data. This allows the model to learn a low-dimensional representation. The branch comprises an encoder with parameter $\theta_1$ and a reconstructor with parameter $\theta_2$, and it takes the input data $\mathbf{s}_t'$ and outputs $f_{\mathrm{rec}}(\mathbf{s}_t';\theta_1,\theta_2)$. The reconstruction loss is defined as follows

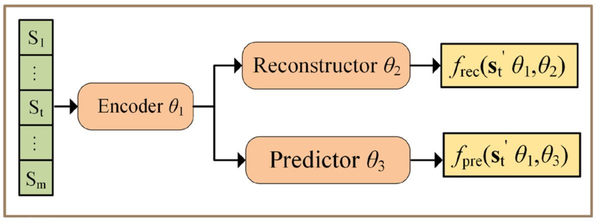


**Fig. 4.** Two-branch student model.

$$L_{\mathrm{rec}} = \frac{1}{m}\sum_{t=1}^{m}(f_{\mathrm{rec}}(\mathbf{s}_t';\theta_1,\theta_2)-\mathbf{s}_t)^2. \tag{5}$$

The Encoder-Prediction branch shares the same encoder as in Encoder-Reconstruction branch and predicts the output values. It consists of an encoder with parameter $\theta_1$ and a predictor with parameter $\theta_3$. This branch receives the data $\mathbf{s}_t'$ and predicts the future spectrum values $f_{\mathrm{pre}}(\mathbf{s}_t;\theta_1,\theta_3)$. We define the prediction loss as

$$L_{\mathrm{pre}} = \frac{1}{m}\sum_{t=1}^{m}(f_{\mathrm{pre}}(\mathbf{s}_t';\theta_1,\theta_3)-y_t)^2. \tag{6}$$

Since this model contains two branches, the total loss of the model should be a weighted combination of two loss functions. Hence, equation (2) can be rewritten as follows

$$L_{\mathrm{hard}} = L_{\mathrm{rec}} + L_{\mathrm{pre}}. \tag{7}$$

The Encoder-Reconstruction branch plays a crucial role in learning the inherent representation, while reducing the noise and redundancy in the input data. By sharing the same encoder, both the reconstructor and the predictor can leverage a low-dimensional data representation, thereby improving the training efficiency and generalization ability of the model.

## 3.4 The Overall TCN-KD Procedure

**Alg. 1.** TCN-KD Training Procedure

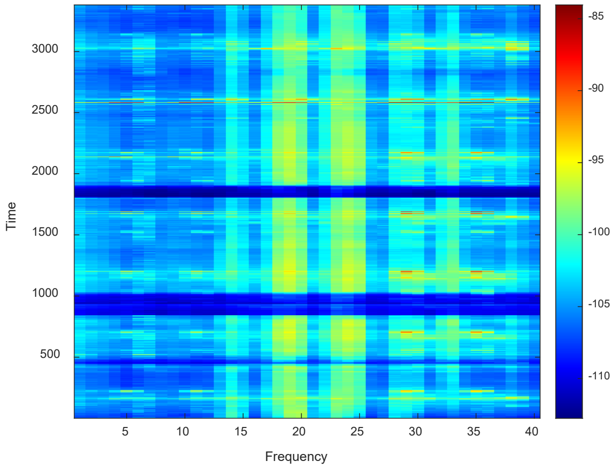
```
input: Spectrum data $\mathbf{D}=\{\mathbf{f}_n\}_{n=1}^{N}$ with $\mathbf{f}_n=\{(\mathbf{s}_t',y_t)\}_{t=1}^{m}$
output: Model $f^{\mathrm{S}}$ with parameters $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$
Initialize $\theta_1$, $\theta_2$ and $\theta_3$
for $n\in\{1,2,...,N\}$ do
  while not converged do
    for $t\in\{1,2,...,m\}$ do
      Obtain $\mathbf{s}_t, y_t \in D$
      Pass $\mathbf{s}_t$ through $f^{\mathrm{T}}$ to obtain $\mathbf{s}_t'$, $f^{\mathrm{T}}(\mathbf{s}_t)$
      Pass $\mathbf{s}_t'$ through $f^{\mathrm{S}}$ to obtain $f_{\mathrm{rec}}(\mathbf{s}_t';\theta_1,\theta_2), f_{\mathrm{pre}}(\mathbf{s}_t';\theta_1,\theta_3)$
      Compute $L_{\mathrm{soft}}$ by (2)
      Compute $L_{\mathrm{hard}}$ by (7)
      Compute $L_{\mathrm{total}}$ by (4)
    end for
  end while
  Update $\theta_1$, $\theta_2$ and $\theta_3$ by minimizing $L_{\mathrm{total}}$
end for
return $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$
```

We refer to the knowledge distillation network of optimized teacher and student models as TCN-KD. The whole training process of TCN-KD is summarized in Algorithm 1. The teacher model receives normalized and fixed-length spectrum data, generating intermediate layer features and soft labels. The student model utilizes middle layer features to generate predicted values. By minimizing both the soft loss and hard loss, the student model reduces the discrepancy with the teacher model. In the testing process of TCN-KD, the input data is passed through the Encoder-Prediction branch of the trained model to obtain the predicted results.

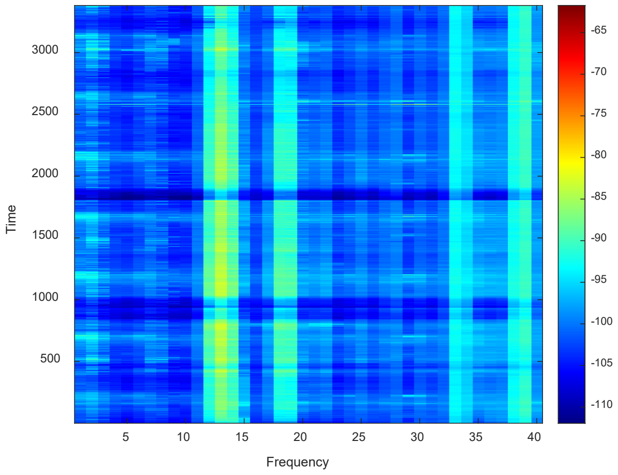
# 4. Experiments and Analysis

## 4.1 Dataset Description and Preliminary Analysis

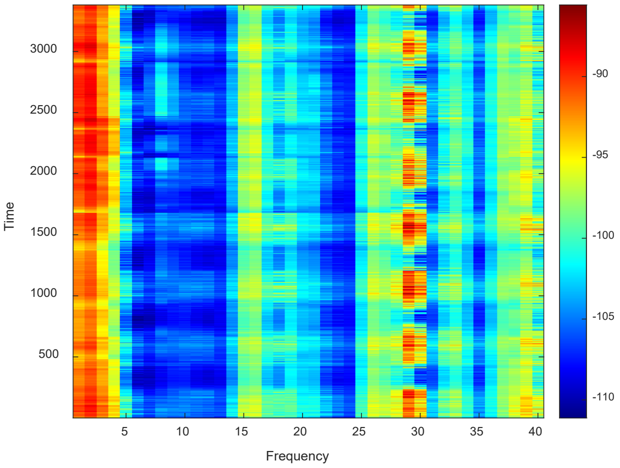
The adopted spectrum data in the experiments was derived from an open-source dataset from RWTH Aachen University in Germany [23]. We selected the dataset located on the roof of a residential area in Maastricht. The dataset consists of four sub-bands with central frequencies of 770 MHz, 2250 MHz, 3750 MHz, and 5150 MHz. Each sub-band is with a bandwidth of 1500 MHz, a frequency resolution of 200 kHz, and a temporal resolution of 1.8 s. For this data, the power spectral density (PSD) value of 1000 time slots are available every half an hour. However, processing these data requires a significant amount of memory and computing capacity, which is difficult for typical computers to process. To solve this problem, we use a weighted average of 100 consecutive values. As the PSD value in the dataset are in dBm/200 kHz, a weighted average cannot be applied directly. Therefore, we first convert PSD values to linear from with mW units, and convert the computed average values to the original logarithmic form to get the final average PSD values. We select spectrum data for four widely-used cellular services as shown in Fig. 5. These services include the GSM 1800 uplink band, GSM 1800 downlink band, GSM 900 uplink band, and GSM 900 downlink band, which were collected over a duration of seven days.



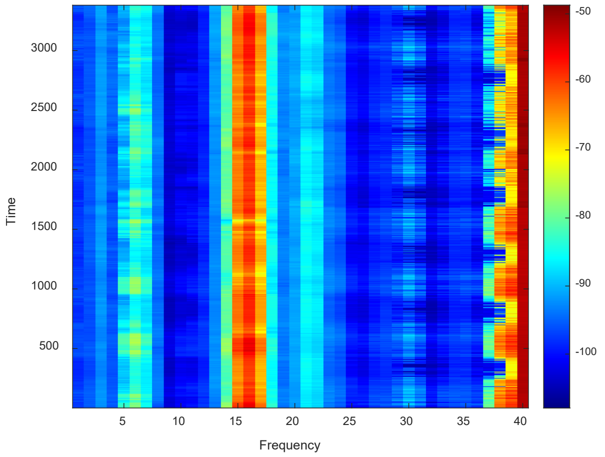
(a) GSM900UL



(b) GSM900DL



(c) GSM1800UL



(d) GSM1800DL

**Fig. 5.** Service activity in different frequency bands.

To meet the input requirements of TCN, it is necessary to normalize the data and divide it into time windows. The normalization step is to scale the original data into the range of [0,1]. Autocorrelation coefficient reflects the degree of correlation between time series at different times [24]. The higher the autocorrelation coefficient, the more significant the correlation between the sequences, and more likely spectrum predictions can be made from the observed spectrum data [25]. Generally, when the autocorrelation coefficient exceeds 0.8, it indicates a high level of correlation. The utilized autocorrelation coefficient is given by

$$\rho_{\mathrm{t,t+\Delta t}} = \frac{\mathrm{cov}(x_t, x_{t+\Delta t})}{\sigma x_t \sigma x_{t+\Delta t}} \tag{8}$$

where cov(·) denotes the covariance, and $\sigma$ denotes the sample standard deviation.

The calculation result is shown in Fig. 6, it can be seen that the autocorrelation coefficient gradually decreases with the lag time step. We also compare and evaluate different window sizes, using the same dataset and model, trying various sliding window sizes, and analyzing their impact on prediction performance using RMSE. Table 1 demonstrates that when the sliding window size is 20, the RMSE is the smallest. Moreover, this window size ensures that the autocorrelation coefficient remains larger than 0.8. Therefore, we set the size of sliding window is 20.

## 4.2 Evaluation Metrics

In this study, we use two metrics for performance evaluation, i.e., the Root Mean Square Error and the Mean Absolute Error (MAE).

RMSE represents the square root of the mean difference between the predicted value and the ground truth, which is computed as

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{9}$$

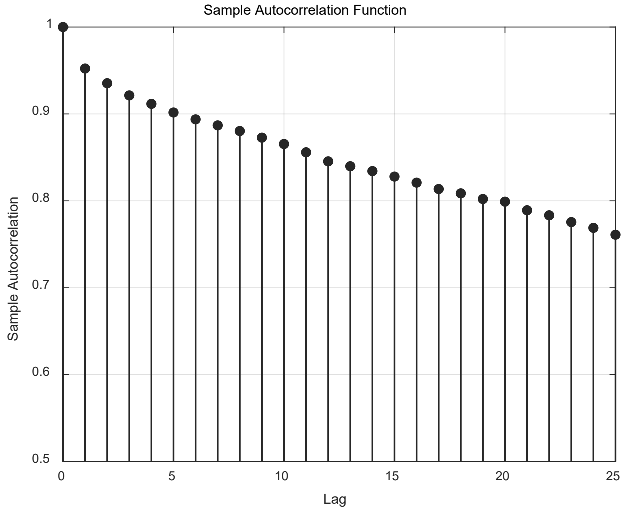


**Fig. 6.** Autocorrelation coefficient.

| Sliding window | 5 | 10 | 15 | 20 | 25 |
|---|---|---|---|---|---|
| RMSE | 0.85775 | 0.8553 | 0.8206 | 0.8119 | 0.8224 |

**Tab. 1.** The average RMSE of different size of sliding windows.

where $m$ is the number of the predicted data, $y$ is the value of the ground truth, and $\hat{y}$ is the value of the predicted data.

MAE calculates the average of the absolute values of the deviations of the predicted values from the ground truth, which can avoid the problem of error offset. The computation of MAE is

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|. \tag{10}$$

## 4.3 Teacher Model Optimization Experiments Based on Self-transfer

In our previous setup, we set the window width as 20, so the hole factor of the TCN can be $d = [1,2,4,8,16]$. The proposed TCN architecture includes an input layer, a one-dimensional convolutional layer, five residual blocks, a fully connected layer, and a final regression output layer. Each convolutional layer has 64 convolutional kernels with a size of 3. Based on the self-transfer TCN optimization network structure mentioned in the previous section, we continue to use RMSprop as the backpropagation optimizer, the epoch value is set as 50, and the learning rate dropped from 0.01 to 0.001.

After multiple experiments and averaging the results, it was determined that freezing the first 27 layers minimizes the prediction error. Therefore, we selected this model as the teacher model in the following experiments. Figure 7 shows the performance comparison of the optimized TCN with the classical time series prediction model [26]. According to the formula (9)–(10), a good model should have small RMSE and MAE. The results indicate that the self-transfer TCN outperforms the original TCN, with a decrease of 5.8% and 4.2% in the evaluation metrics RMSE and MAE, respectively. Moreover, the TCN demonstrates significant advantages in predicting the spectrum data. Compared to the LSTM and support vector regression (SVR), the optimized TCN reduces RMSE by 20.9% and 19.3%, respectively. These findings suggest that TCN is suitable for the spectrum prediction problems, and the self-transfer TCN model can further improve the network prediction accuracy.

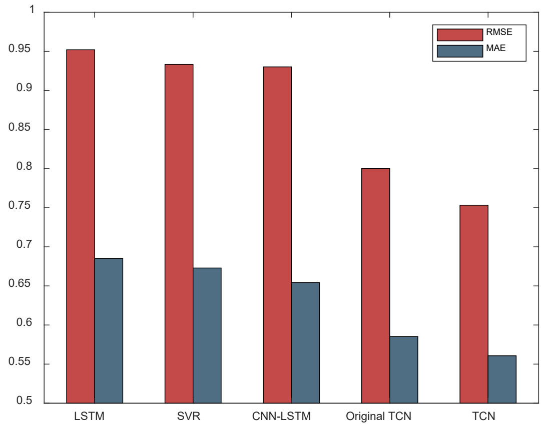


**Fig. 7.** Comparison of teacher model optimization experiment.

## 4.4 Performance of Knowledge Distillation Learning for Spectrum Prediction

Here we perform knowledge distillation simulation experiments to train the lightweight student model, using the self-transfer TCN as teacher model. For the student model, we choose two-branch network, the parameters of which were set as follows. The Encoder is a feed-forward neural network with three hidden layers of different dimensions, which has 20, 64, 32 dimensions separately. The ReLU activation function is applied after each hidden layer. The Reconstructor has two linear layers with 32 and 20 neurons respectively, while the Predictor has two linear layers with 16 and 1 neurons. For notation purpose, our proposed knowledge distillation model is termed as TCN-KD.

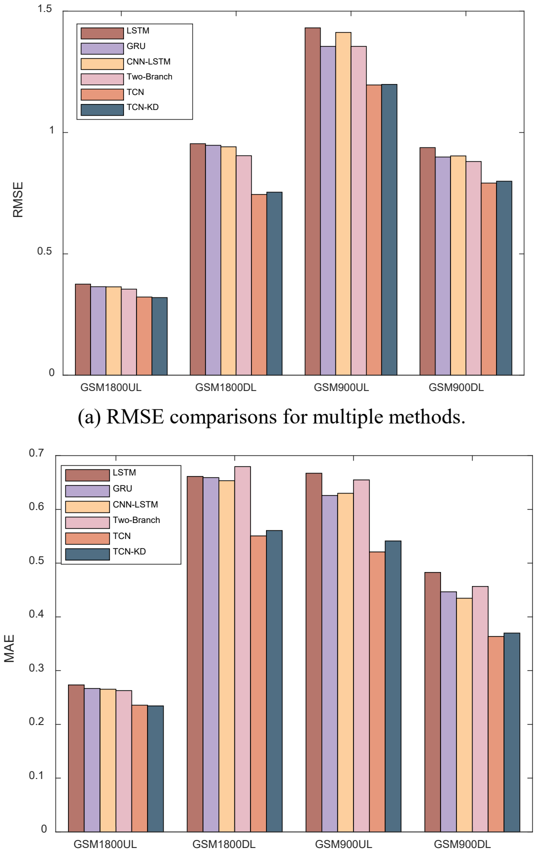
To evaluate the performance of TCN-KD, we conduct experiment in the four spectrum bands, GSM 900 UL, GSM 900 DL, GSM 1800 UL, and GSM 1800 DL, using LSTM, GRU, CNN-LSTM, two-branch network, and TCN-KD. Two-branch network is student model without the help of teacher and TCN-KD is the model with the knowledge of pre-trained teacher. Tables 2 and 3 show the RMSE and MAE results for each model in the experiments. It can be seen that in four different frequency bands, the proposed model is smaller than the RMSE and MAE of others except TCN teacher model. At the same time, TCN-KD has significantly fewer parameters to be trained than the other models. Compared to the two-branch network, TCN-KD has only 2,816 more parameters, but the prediction performance has been significantly improved. This shows that with the help of the teacher model, the student model can obtain higher prediction accuracy at the cost of smaller scale. To gain more insight, we present the RMSEs and MAEs of different models in Fig. 8. Notably, the two-branch network that we proposed has a relatively low RMSE but a higher MAE. This is due to the presence of extreme outliers between predicted and actual values and MAE is sensitive to outliers. However, when we incorporated knowledge distillation for our TCN-KD model, it shows higher accuracy and more stable performance. Although TCN-KD is not as effective as the teacher model TCN, the difference in prediction error is negligible, and the number of parameters to train is significantly lower, indicating that the knowledge distillation step helps improving TCN-KD's prediction performance while maintaining model simplicity.

| RMSE | 1800UL | 1800DL | 900UL | 900DL | Params |
|---|---|---|---|---|---|
| LSTM | 0.3752 | 0.9539 | 1.4313 | 0.9378 | 40 901 |
| GRU | 0.3647 | 0.9474 | 1.3551 | 0.8991 | 30 701 |
| CNN-LSTM | 0.3642 | 0.9412 | 1.4124 | 0.9036 | 53 493 |
| TCN | 0.3222 | 0.7448 | 1.1960 | 0.7917 | 141 633 |
| Two-Branch | 0.3546 | 0.9045 | 1.3551 | 0.8805 | 5 685 |
| TCN-KD | **0.3199** | **0.7542** | **1.1983** | **0.7994** | **8 501** |

**Tab. 2.** RMSEs and numbers of different models in four bands.

| MAE | 1800UL | 1800DL | 900UL | 900DL | Params |
|---|---|---|---|---|---|
| LSTM | 0.2734 | 0.6613 | 0.6673 | 0.4827 | 40 901 |
| GRU | 0.2668 | 0.6591 | 0.6259 | 0.4467 | 30 701 |
| CNN-LSTM | 0.2654 | 0.6534 | 0.6300 | 0.4347 | 53 493 |
| TCN | 0.2357 | 0.5507 | 0.5309 | 0.3636 | 141 633 |
| Two-Branch | 0.2628 | 0.6796 | 0.6549 | 0.4566 | 5 685 |
| TCN-KD | **0.2344** | **0.5607** | **0.5413** | **0.3699** | **8 501** |

**Tab. 3.** MAEs and numbers of different models in four bands.



(a) RMSE comparisons for multiple methods.

(b) MAE comparisons for multiple methods

**Fig. 8.** Comparison of knowledge distillation experiment.

In order to analyze the timeliness performance of the proposed model, we record the total time required for each model from training to prediction. The purpose was to evaluate the efficiency and applicability of the models [27]. In the experiment, we initialize the learning rate to 0.01 and employ an early stopping mechanism. We use RMSprop as the optimizer for 50 epoch training and set the batch size to 12. Results are given in Tab. 4, we can find that the training time of the model usually increases proportionally with the number of training parameters. Due to the large number of trainable parameters in the TCN model, it requires more time to attain higher prediction accuracy. LSTM training is time-consuming due to its serial computing nature. GRU requires less time compared to LSTM because it omits the output gate in its computation process. Utilizing knowledge distillation, TCN-KD achieves better training accuracy and less training time than LSTM and GRU. Given limited resources, TCN-KD is more suitable for practical deployment.

| Model | LSTM | GRU | CNN-LSTM | TCN | TCN-KD |
|---|---|---|---|---|---|
| Times [s] | 44.05 | 31.46 | 28.49 | 39.06 | 19.56 |

**Tab. 4.** Time cost for multiple methods training.

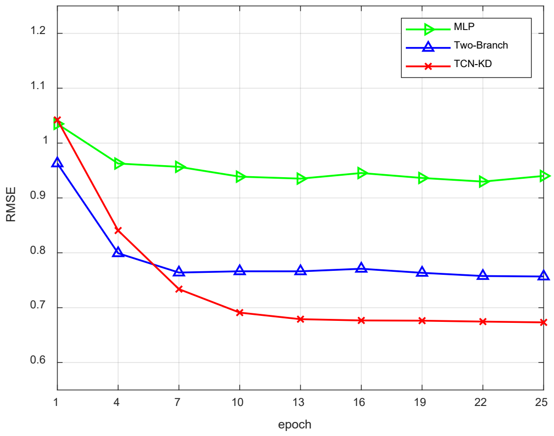


**Fig. 9.** Performance of three prediction model involved.

For further analysis, the performance of MLP two-branch network and the proposed model TCN-KD are compared in Fig. 9. MLP is the two-branch network without reconstruction loss, two-branch network is the TCN-KD without teacher. Although the prediction accuracy of MLP is not high, with the help of reconstruction loss, the accuracy is improved a lot. Meanwhile, with the help of pre-trained teacher model, the prediction accuracy of two-branch neural network is further improved.

## 4.5 Performance in Small Sample Scenarios for Spectrum Prediction

By conducting comparative experiments between TCN-KD and other networks, we find that the proposed spectrum prediction framework based on the knowledge distillation has certain advantages regarding prediction time and accuracy, particularly when there are sufficient target samples. Further experimental verification is necessary to determine whether TCN-KD can maintain good prediction performance with limited data. In this experiment, we change the input data period from 7 days to 2 days and tested various networks on the four spectrum bands. The RMSE was used to evaluate the prediction accuracy. As shown in Fig. 10, the analysis of errors in each frequency band shows that TCN-KD still performs well even with a limited amount of data. For example, in the GSM900UL band, the proposed algorithm reduces RMSE by 24.2% compared to CNN-LSTM and by 12.3% compared to the two-branch network. This is because knowledge distillation enables the student model to acquire comprehensive knowledge from the teacher model, which helps it achieve better generalization performance with limited data. In addition, we observe that CNN-LSTM performs poorly with a small amount of data. This can be attributed to the model complexity and large number of parameters in CNN-LSTM, making it challenging to effectively train with limited data. Consequently, CNN-LSTM tends to overfit and its generalization ability decreases.

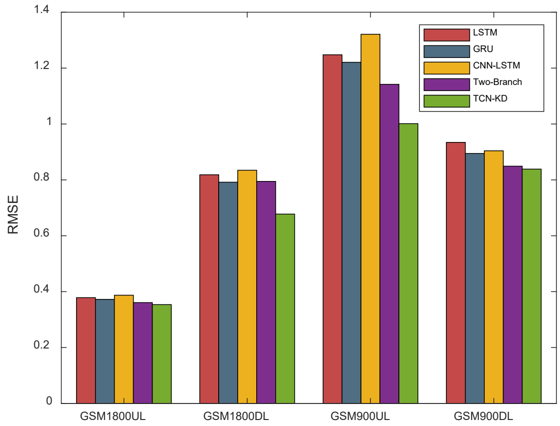


**Fig. 10.** The performance of different models under small sample scenarios in four bands.

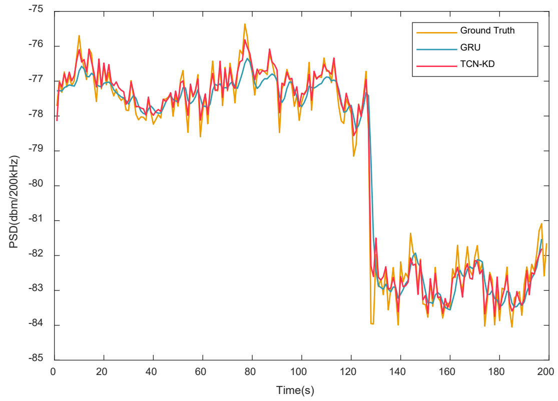


**Fig. 11.** Comparison of the spectrum prediction results.

For visualization purpose, Figure 11 shows the series of some predicted values in the experiment, where the ground truth, predicted values for the GRU and TCN-KD models are plotted. To present a clearer comparison, we selected the GRU with relatively better prediction performance in the algorithm comparison. The results demonstrate that TCN-KD, aided by knowledge distillation, can more effectively capture the intricate details in spectral data compared to the GRU model. Conversely, the GRU model exhibits poor fitting to the actual values.

# 5. Conclusions

In this study, a fast spectrum prediction model termed TCN-KD has been proposed to address the spectrum prediction problem in the real spectrum environment with limited available samples. The knowledge distilled from the teacher model was used to guide the training process of the student model and allow the student model to learn more efficiently and quickly. The complexity and the sample requirements during training of the model was reduced, making it applicable to spectrum prediction tasks. Experiments results with both sufficient and limited data show that TCN-KD can improve prediction accuracy by 19.5% compared to LSTM prediction in sufficient historical data and reduces 71.1% parameters to be trained at the same time. Furthermore, the accuracy can be improved by 17.9% compared to GRU in the scenarios with scarce usable historical data.

In our future work, we will primarily concentrate on expanding the time-domain of spectrum prediction to joint time frequency domains. Such an approach may gather more comprehensive spectrum information, enabling us to make more accurate predictions. Further, the spectrum prediction results will be utilized in dynamic spectrum access to minimize system delay and maximize throughput capacity.

# Acknowledgments

This study was supported by National Natural Science Foundation of China (No. 62131005, No. 62231012).

# References

[1] ELTHOLTH, A. Forward backward autoregressive spectrum prediction scheme. In *2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS)*. Cairns (Australia), 2015, p. 1–5. DOI: 10.1109/icspcs.2015.7391770

[2] MOSAVAT-JAHROMI, H., LI, Y., CAI, L., et al. Prediction and modeling of spectrum occupancy for dynamic spectrum access systems. *IEEE Transactions on Cognitive Communications and Networking*, 2021, vol. 7, no. 3, p. 715–728. DOI: 10.1109/tccn.2020.3048105

[3] CHEN, X., YANG, J., DING, G. Minimum Bayesian risk based robust spectrum prediction in the presence of sensing errors. *IEEE Access*, 2010, vol. 6, p. 29611–29625. DOI: 10.1109/ACCESS.2018.2836940

[4] YIN, L., YIN, S., HONG, W., et al. Spectrum behavior learning in cognitive radio based on artificial neural network. In *2011-MILCOM 2011 Military Communications Conference*. Baltimore (USA), 2011, p. 25–30. DOI: 10.1109/MILCOM.2011.6127671

[5] WANG, X., PENG, T., ZUO, P., et al. Spectrum prediction method for ISM bands based on LSTM. In *2020 5th International Conference on Computer and Communication Systems (ICCCS)*. Shanghai (China), 2020, p. 580–584. DOI: 10.1109/ICCCS49078.2020.9118535

[6] MIAO, J., LI, Y., JING, X., et al. Spectrum sensing based on adversarial transfer learning. *IET Communications*, 2020, vol. 16, no. 17, p. 2059–2069. DOI: 10.1049/cmu2.12459

[7] DING, G., WANG, J., WU, Q., et al. Joint spectral-temporal spectrum prediction from incomplete historical observations. In *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. Atlanta (USA), 2014, p. 1325–1329. DOI: 10.1109/GlobalSIP.2014.7032338

[8] DING, G., ZHAI, S., CHEN, X., et al. Robust spectral-temporal two-dimensional spectrum prediction. In *Machine Learning and Intelligent Communications: First International Conference. MLICOM.* Shanghai (China), 2017, p. 393–401. DOI: 10.1007/978-3-319-52730-7_40

[9] DING, G., WU, F., WU, Q., et al. Robust online spectrum prediction with incomplete and corrupted historical observations. *IEEE Transactions on Vehicular Technology*, 2017, vol. 66, no. 9, p. 8022–8036. DOI: 10.1109/TVT.2017.2693384

[10] LIN, F., CHEN, J., SUN, J., et al. Cross-band spectrum prediction based on deep transfer learning. *China Communications*, 2020, vol. 17, no. 2, p. 66–80. DOI: 10.23919/JCC.2020.02.006

[11] LIN, F., CHEN, J., DING, G., et al. Spectrum prediction based on GAN and deep transfer learning: A cross-band data augmentation framework. *China Communications*, 2021, vol. 18, no. 1, p. 18–32. DOI: 10.23919/JCC.2021.01.002

[12] PENG, C., ZHANG, M., HU, W., et al. Cross-band spectrum prediction algorithm based on Transfer Learning and Meta Learning. In *7th International Conference on Computer and Communications (ICCC)*. Chengdu (China), 2021, p. 2303–2307. DOI: 10.1109/ICCC54389.2021.9674444

[13] HINTON, G., VINYALS, O., DEAN, J. Distilling the knowledge in a neural network. *Computer Science*, 2015, vol. 14, no. 7, p. 38 to 49. DOI: 10.48550/arXiv.1503.02531

[14] SHAO, R., LIU, Y., ZHANG, W., et al. A survey of knowledge distillation on deep learning (in Chinese). *Chinese Journal of Computers*, 2020, vol. 45, no. 8, p. 1638–1673. DOI: 10.11897/SP.J.1016.2022.01638

[15] ZHAI, N., ZHOU, X., LI, S., et al. Prediction method of furnace temperature based on transfer learning and knowledge distillation (in Chinese). *Computer Integrated Manufacturing Systems,* 2022, vol. 28, no. 6, p. 1860–1869. DOI: 10.13196/j.cims.2022.06.024

[16] ZHANG, X., LIU, Y. Remaining useful life prediction of aero-engine based on knowledge distillation compression hybrid model (in Chinese). 15 pages. [Online] Cited at 2022-10-12. Available at: http://kns.cnki.net/kcms/detail/11.5946.TP.20221011.1557.024.html

[17] AY, E., DEVANNE, M., WEBER, J., et al. A study of knowledge distillation in fully convolutional network for time series classification. In *2022 International Joint Conference on Neural Networks (IJCNN)*. Padua (Italy), 2022, p. 1–8. DOI: 10.1109/IJCNN55064.2022.9892915

[18] ZHANG, Y., HU, G., CAI, Y. Proactive spectrum monitoring with spectrum monitoring data transmission in dynamic spectrum sharing network: Joint design of precoding and antenna selection. *IET Communications*, 2021, vol. 15, no. 18, p. 2265–2274. DOI: 10.1049/cmu2.12260

[19] YU, L., CHEN, J., DING, G. Spectrum prediction via long short term memory. In *2017 3rd IEEE International Conference on Computer and Communications (ICCC).* Chengdu (China), 2017, p. 643–647. DOI: 10.1109/COMPCOMM.2017.8322623

[20] ZHANG, T., ZHANG, Y., CAO, W., et al. Less is more: Fast multivariate time series forecasting with light sampling-oriented MLP structures. *Computer Science*, 2022, p. 1–11. DOI: 10.48550/arXiv.2207.01186

[21] HARELL, A., MAKONIN, S., BAJIC, I. A causal neural network for power disaggregation from the complex power signal. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Brighton (UK), 2019, p. 8335–8339. DOI: 10.1109/ICASSP.2019.8682543

[22] CHANG, S., LI, B., SIMKO, G., et al. Temporal modeling using dilated convolution and gating for voice-activity-detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Calgary (Canada), 2018, p. 5549 to 5553. DOI: 10.1109/ICASSP.2018.8461921

[23] WELLENS, M. *Empirical Modelling of Spectrum Use and Evaluation of Adaptive Spectrum Sensing in Dynamic Spectrum Access Networks*. [Online] Cited 2010-05-14. Available at: https://publications.rwth-aachen.de/record/51779/files/3248.pdf

[24] LI, X., LIU, Z., CHEN, G., et al. Deep learning for spectrum prediction from spatial-temporal-spectral data. *IEEE Communications Letters,* 2021, vol. 25, no. 4, p. 1216–1220. DOI: 10.1109/LCOMM.2020.3045205

[25] SUN, J., SHEN, L., DING, G., et al. Predictability analysis of spectrum state evolution: Performance bounds and real-world data analytics. *IEEE Access*, 2017, vol. 5, no. 10, p. 22760–22774. DOI: 10.1109/ACCESS.2017.2766076

[26] LAZCANO, A., HERRERA, P., MONGE, M. A combined model based on recurrent neural networks and graph convolutional networks for financial time series forecasting. *Mathematics*, 2023, vol. 11, no. 1, p. 1–21. DOI: 10.3390/math11010224

[27] YIM, J., JOO, D., BAE, J., et al. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* Honolulu (USA), 2017, p. 4133–4141. DOI: 10.1109/CVPR.2017.754

# About the Authors ...

**Runmeng CHENG** is a M.S. student at the School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, Nanjing, China. Her research interests include smart spectrum management.

**Jianzhao ZHANG** (corresponding author) received the Ph.D. degree in Communication Engineering from the PLA University of Science and Technology, Nanjing, China, in 2012. He is currently an Associate Researcher in the Sixty-Third Research Institute, National University of Defense Technology, Nanjing, China. His research interests include spectrum environment cognition and smart spectrum management.

**Junquan DENG** received the B.Eng. degree in Automation Engineering from Tsinghua University, Beijing, China, in 2011. He obtained his M.Sc. degree in Computer Science from the National University of Defense Technology (NUDT), Changsha, China, in 2013, and completed the Ph.D. degree in Information Theory from Aalto University, Aalto, Finland, in 2018. He is an Associate Research Fellow with the Sixty-Third Research Institute, National University of Defense Technology, Nanjing, China. His research interests include device-to-device communication, millimeter-wave communication, mobile relaying in 5G cellular networks, and machine learning with wireless network data.

**Yanping ZHU** received the Ph.D. degree in Information and Communication Engineering from Nanjing University of Science and Technology, China, in 2014. She also worked as an academic visitor in University of Sheffield, UK, in 2018. She is currently working as an associate professor at Nanjing University of Information Science and Technology, Nanjing, China. Her research interests include emotion recognition algorithm based on EEG brainwave signals, UWB localization and close-range microwave imaging.