IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 August 2022, accepted 7 September 2022, date of publication 16 September 2022,
date of current version 26 September 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3207146*

RESEARCH ARTICLE

# A Sample-Efficient OPF Learning Method Based on Annealing Knowledge Distillation

**ZIHENG DONG[1], KAI HOU[1], (Member, IEEE), ZEYU LIU[1], (Student Member, IEEE), XIAODAN YU[1], HONGJIE JIA[1], (Senior Member, IEEE), AND CHI ZHANG[2,3]**

[1]Key Laboratory of Smart Grid of Ministry of Education, Tianjin University, Tianjin 300072, China
[2]State Grid Jiangsu Electric Power Company Ltd., Nanjing 211102, China
[3]Extra-High Voltage Branch Company, Nanjing 211102, China

Corresponding author: Zeyu Liu (tjulzy@tju.edu.cn)

This work was supported by the National Natural Science Foundation of China (No. 52077150/U2066213) and the Natural Science Foundation of Tianjin City, China (No. 19JCYBJC19200).

**ABSTRACT** To quickly respond to variations in the state of network load demand, a solution using data-driven techniques to predict optimal power flow (OPF) has emerged in recent years. However, most of the existing methods are highly dependent on large data volumes. This limits their application on the newly established or expanded systems. In this regard, this work proposes a sample-efficient OPF learning method to maximize the utilization of limited samples. By decomposing the OPF task before knowledge distillation, deep learning complexity is reduced. Thereafter, knowledge distillation is used to integrate decoupled tasks and improve accuracy in low-data setups. Unsupervised pre-training is introduced to alleviate the demand for labeled data. Additionally, the focal loss function and teacher annealing strategy are adopted to achieve higher accuracy without extra samples. Numerical tests on different systems corroborate the advanced accuracy and training speed over other training methods, especially on fewer-sample occasions.

**INDEX TERMS** Optimal power flow, sample efficiency, annealing knowledge distillation, focal loss function, stacked denoising autoencoder, deep learning.

## I. INTRODUCTION

Optimal power flow (OPF) is the cornerstone of many research areas such as power system security, reliability, and economics. Traditionally, the time scale of OPF is 15 minutes to 1 hour ahead. However, owing to the frequent and uncertain fluctuations of the renewable generations and loads, the OPF needs to be computed more efficiently and even in real-time, to determine the optimal and safe operation strategy [1]. As a result, the efficiency of OPF becomes an urgent issue to be addressed.

Due to the non-convex and non-linear nature of the model, it is difficult to obtain the real-time analytical solution of OPF. The OPF model has undergone the development of linearization and decoupling transformation to reduce the computational burden, such as direct current OPF (DC-OPF) and fast decoupled load flow [2]. Although many advances have been made to simplify the model, the computational efficiency problem is still a bottleneck.

In recent years, deep-learning-based methods have exerted significant efficiency improvement for OPF [3], [4]. It uses a large amount of historical data to approximate the variable relationship and achieve the real-time response. Compared with traditional solvers, the deep learning approach has a computation speed improvement of up to 200 times for DC-OPF and 35 times for alternating current OPF (AC-OPF) [5], [6]. In addition, the deep learning technique provides a feasible solution to address OPF solving in online settings and state combinations. To address the online efficiency problem of OPF learning, several approaches have been studied based on active constraints [7], [8], warm-start points prediction [9], [10], and so on. However, high data requirements of these data-intense methods limit their applications [11].

To reduce the data requirements for training, a hybrid number model-driven approach is adopted to simplify the iterations. Such an approach is no longer simple end-to-end

The associate editor coordinating the review of this manuscript and approving it for publication was Youngjin Kim.

deep learning training, but a way to use training techniques to accelerate the original OPF solving process. The Lagrangian-based reinforcement learning is used in the iterative process to accelerate the convergence to achieve optimality [12]. A trajectory speculation method is proposed to predict and accelerate convergence [13]. To address the large burden of data preparation and storage, an efficient sample generation strategy is presented by compressing the sampling space [14]. Another category of approaches targets fewer samples by leveraging the prior information. Physically informed learning takes advantage of prior information from physical models and avoids the large traditional training datasets [15]. Using constraints as a priori elements, machine learning methods can predict AC-OPF neural networks and Lagrange duality with high fidelity and minimal constraint violations [16], [17]. Similarly, the implementation of pre-classification with active constraints has become a practical solution strategy [18]. Based on the concept of OPF sensitivity, the solutions learned by DNNs and intermediate results are used to accelerate the process of OPF solving [19].

In conclusion, existing deep learning approaches in OPF are either data-intensive or knowledge-demanding. Since the topology or operation is frequently changed in power systems, it is prohibitive to retrain models from scratch and the sample data accumulated in a short time are very limited [20]. Therefore, sample-efficient learning models with high accuracy are well motivated [19].

In this regard, the paper proposes a sample-efficient method for DC-OPF learning, which is suitable for the limited labeled samples. The work is taken on the DC-OPF model because its linear model is more convenient to explore ways for sample efficiency improvement from a theoretical perspective. Specifically, this paper addresses the application of small samples from three perspectives. Firstly, the pre-training strategy is adopted in the stacked denoising autoencoder (SDAE) network. The size of labeled data is reduced by transferring work to the unsupervised pre-training stage. Secondly, the DC-OPF task decomposition strategy and knowledge distillation are combined to reduce the learning complexity. The knowledge distillation learning is improved with a teacher annealing strategy to improve the accuracy. Moreover, the loss function is improved based on focal loss in the training phase to enhance the training effect without adding extra samples. In our work, because the pre-trained results can be reused and the sample size is reduced, the model training speed can be greatly improved. The main contributions of this paper include,

- A sample-efficient method is developed which makes full use of small-scale data. Free of prior knowledge or large dataset, the method enables the easy deployment of deep-learning-based OPF in new system states.
- A method based on DC-OPF task decomposition and knowledge distillation learning is proposed to alleviate the training complexity. The proposed method can be easily extended to different scale systems.
- Novel continuous focal loss (CFL) functions are designed and used to improve the training performance without extra samples. Pre-training and teacher annealing strategies achieve higher accuracy in the small-data regime.

The remaining paper is organized as follows. In section II, challenges of OPF learning are discussed and the scheme of the proposed solution is outlined. Section III details the training process. The overall procedure is described in section IV. Numerical results with the proposed method are shown in section V. Finally, the paper is summarized in section VI.

## II. PROPOSED OPF LEARNING FRAMEWORK

With no consideration of the mapping relationship and data distribution, the conventional learning approaches rely heavily on data volumes, which limits the application. Actually, the variable relationships can be simplified by decoupling the target outputs in separate networks. Therefore, our solution is based on OPF task decomposition and organized in a knowledge distillation framework.

### A. PROBLEM STATEMENT AND CHALLENGES

The OPF determines the most economical generation dispatch while satisfying the load demand and other security constraints. The following optimization formulations are obtained when applying a DC approximation to the traditional AC-OPF.

$$\min \sum_{i\in\Omega_G} \left(c_{2i}P_{Gi}^2 + c_{1i}P_{Gi}\right) \tag{1}$$

$$\begin{cases} P_G - P_D = \sum\limits_{i,j\in\Omega_N} B_{ij}\dfrac{V_{\theta i} - V_{\theta j}}{x} \\ P_{Fk} = B_{ij}\dfrac{V_{\theta i} - V_{\theta j}}{x}, \quad k \in \Omega_{br} \\ P_{Gi,\min} \le P_{Gi} \le P_{Gi,\max}, \quad i \in \Omega_G \\ P_{Fk.\min} \le P_{Fk} \le P_{Fk.\max}, \quad k \in \Omega_{br} \end{cases} \tag{2}$$

where $P_G$ is the power output of $i$th generating unit. $c_{1i}$ and $c_{2i}$ are the generation cost coefficients. $P_D$ is the power demand of the $i$th bus. $P_{Fk}$ is the transmission power of the $k$th branch. $V_{\theta i}$ is the voltage phase angle of the $i$th bus. $\Omega_G$, $\Omega_N$ and $\Omega_{br}$ are the set of generating units, bus, and branches, respectively. $B_{ij}$ is the susceptance of admittance between the $i$th and $j$th bus.

The OPF model contains information about the branch parameters and network topology. The complex model requires a few iterations to reach the optimal solution. It takes a long time to optimize the power flow for a large number of operating states.

The researchers are currently interested in a model-free method based on deep learning, which seeks a function automatically to fit the abstract relationship between power demand and power dispatch. In [4], [5], [6], [7], [8], [9], [10], and [11], load variables are widely used as input features, while power generations and phase angles are considered

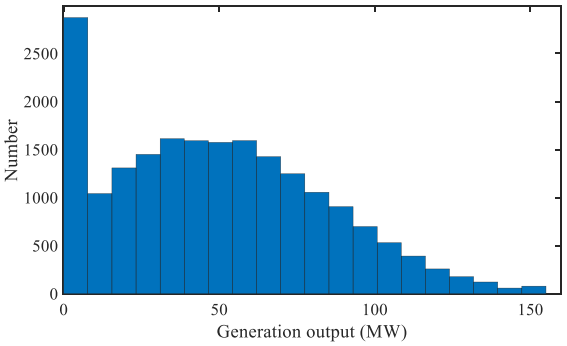


**FIGURE 1.** Example of the class-imbalance problem of the power generation data. The figure shows that the distribution of the 22nd generating unit in RTS-79 system is right skew and thick-tailed.

as output variables. Other unchangeable factors can be contained in network parameters and thus excluded from the input features. The most commonly used loss function in conventional deep learning is the mean square error (MSE) function.

Such a loss function gives the same emphasis on each data and thus the class imbalance of the training data may cause low accuracy, especially with fewer labels. However, the class-imbalance problem is the intrinsic characteristic, as shown in Fig.1. Moreover, different variables and active constraints intensify the training difficulty with fewer samples. In summary, the challenges of using limited samples consist in the training complexity and prediction accuracy.

### B. PROPOSED SOLUTION AND FRAMEWORK

The key idea of the proposed method is to alleviate the training difficulty and improve the sample efficiency. To achieve the first purpose, the DC-OPF can be decoupled to better generalize the variable relationships. To enhance the sample efficiency, a focal loss function may be engaged to give higher importance to the minority class without extra samples. Unsupervised pre-training is also integrated where the training dataset is supplemented by the unlabeled data.

For all the solutions given above, knowledge distillation [21] is introduced here. It involves building a small lightweight model and training it with the supervised information from a larger model. The large and small models are called the Teacher model and Student model, respectively. The supervised information from the output of the Teacher model is called knowledge, and the process by which the student learns the supervised information from the teacher is called Distillation. Knowledge distillation is an ideal candidate to integrate them for the following reasons:

1) Data availability: Only a limited amount of labeled training data is required because the historical data is partially replaced by the predictions from teacher models. It is worth stressing that the proposed knowledge distillation methodology can be combined with unsupervised pre-training approaches.

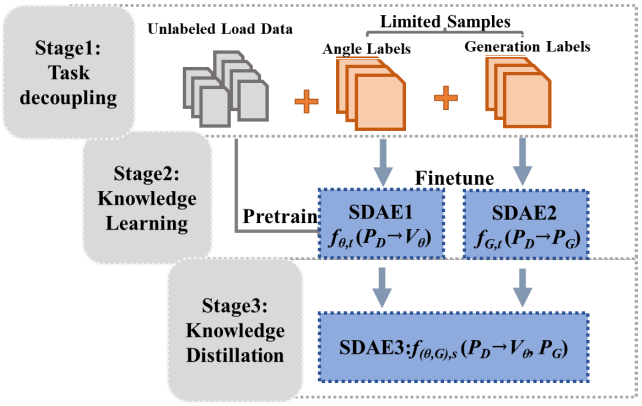


**FIGURE 2.** Scheme of the proposed method.

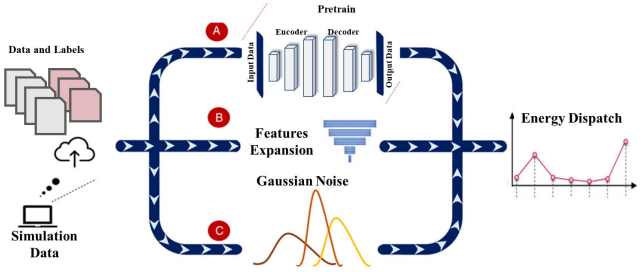


**FIGURE 3.** Key structure of the SDAE network.

2) Input pattern: The OPF problem inputs are consistent in each teacher model and the student model. These properties enable OPF learning to utilize the same pre-training results.
3) Accuracy: The student model can achieve higher accuracy compared with the teacher models [22].

Therefore, the aforementioned solutions are organized in the proposed OPF framework based on knowledge distillation, including three stages.

*Stage 1:* Task decoupling: The DC-OPF task is decoupled according to the types of variables. Unlabeled data is used in the pre-training stage which can be shared in later training.

*Stage 2:* Knowledge learning: Two networks are trained separately to output phase angle and power generation.

*Stage 3:* Knowledge Distillation: The separate models obtained from stage 2 are treated as teacher models. The knowledge is passed to another network (i.e., the student model) to enable it to output phase angle and power generation.

### C. NETWORK ARCHITECTURE

The SDAE network, with fewer hyperparameters, is compatible with the proposed method which combines unsupervised pretraining and finetuning. There are three differences between our network and the SDAE network normally used as shown in Fig.3.

First, for the hidden layer setting, the traditional SDAE network has the smallest hidden layer in the middle, i.e. the bottleneck layer. In contrast, the middle layer is the widest in

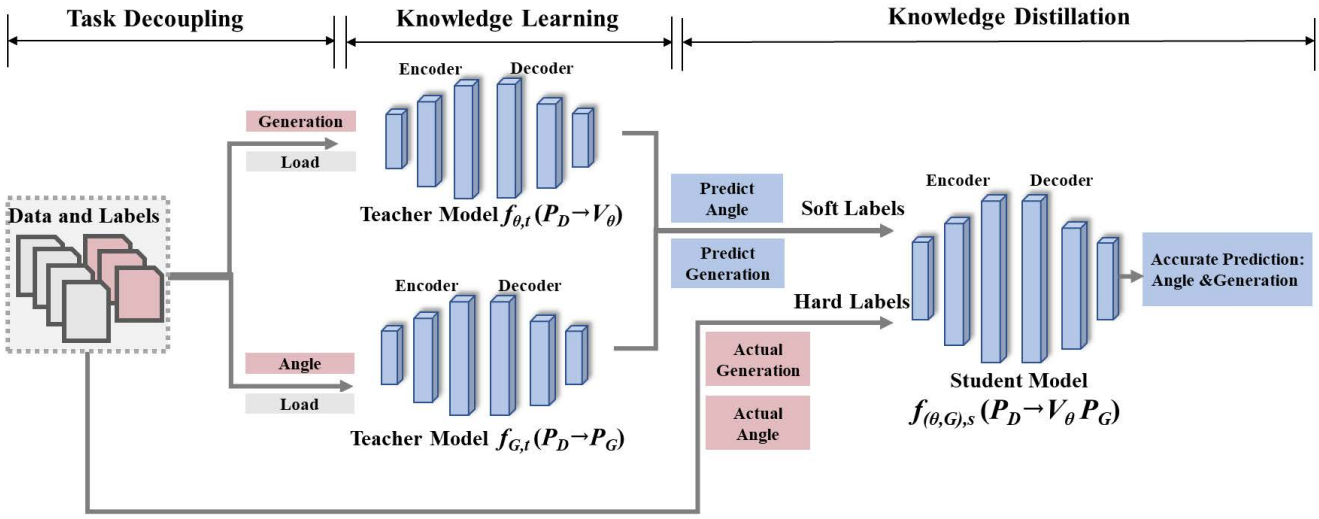


**FIGURE 4.** Label flow in OPF learning.

this paper. Second, the hidden layer is symmetric, combined with coding layers and decoding layers. The coding layer is designed as a gradually widening structure to enhance the feature diversity. Third, given the input features, the random Gaussian noise is used to add noise erosion to the data.

## III. PROPOSED METHOD BASED ON KNOWLEDGE DISTILLATION

### A. TASK DECOUPLING STRATEGY

For such a multi-output problem as OPF, a wider or deeper network is required for accuracy, but the larger size of the network also increases the training burden. Therefore, the main idea of the OPF task decoupling is to train separate models for different variable relationships. The training pressure is no longer confined to a single network by learning the decomposed DC-OPF in paralleled models. The sample flow in training is illustrated in Fig.4.

The variable relationships can be categorized into two types and thus the OPF task is decoupled as follows,

$$\begin{aligned} &decoupled\ task: f_T = \left\{f_{\theta,t}, f_{G,t}\right\} \\ &f_{\theta,t}\left(P_D \rightarrow V_\theta\right): X = \left[P_D\right], \quad Y = \left[V_\theta\right] \\ &f_{G,t}\left(P_D \rightarrow P_G\right): X = \left[P_D\right], \quad Y = \left[P_G\right] \end{aligned} \tag{3}$$

where $f_T$ denotes the task of training teacher models. $f_{\theta,t}$ and $f_{G,t}$ are the teacher models to predict $V_\theta$ and $P_G$, respectively.

The teacher model $f_{\theta,t}(P_D \rightarrow V_\theta)$: learns the knowledge of the voltage angle on each bus. In the training dataset, the voltage angle is regarded as the label.

The teacher model $f_{G,t}(P_D \rightarrow P_G)$: learns the knowledge of the generation dispatch. The actual generation dispatch is treated as the label. The load demand data is the input of both teacher models.

The student model $f_{(\theta,G),s}(P_D \rightarrow V_\theta, P_G)$: mimics the real label data and teacher model predictions. The training dataset also involves the power demand as state data. Note that the labels are not only the actual data but also the prediction from teacher models.

### B. TRAINING TEACHER MODEL BASED ON LIMITED LABELS

By decoupling the DC-OPF, the mapping function of each teacher model is single-variable oriented. The single-task model for predicting phase angle or power generation is based on the SDAE network with two-stage training (i.e. unsupervised pre-training and supervised finetune). Since the model inputs are the same, the pre-training results are shared in two teacher models.

#### 1) PRE-TRAIN

The pre-training of the proposed model involves only unlabeled load data in a task-agnostic way. These data are readily available in the power system. The pre-training is to train most of the parameters with unlabeled samples which are readily accessible. The computational burden is eased for subsequent supervised training.

The unlabelled state data is used in a self-supervised manner based on feature reconstruction. Feature reconstruction means the original feature can be recovered to its initial form after an encoding-decoding process. Pre-training aims to minimize the distance between the original features and their corresponding transformations. The more similar the reconstruction feature is, the more valuable features can be kept by the encoder.

In a traditional SDAE network, the input value is usually erased with random zeros to enable the network with anti-noise ability. However, this random zero strategy is unsuitable for our input vector because it may lose key features. To deal with this, a random gaussian noise strategy is proposed to avoid feature loss.

$$P_{D,noise} = \eta P_D \odot \operatorname{sgn}(r - p) + P_D \tag{4}$$

where $P_{D,noise}$ is the input vector $P_D$ added with random noise. $\eta$ is the noise ratio, which obeys a Gaussian distribution and lies between $+5\%$ and $-5\%$. $r$ is a random vector. $p$ is the probability vector of noise arising. $\mathrm{sgn}(r-p)$ is the noise flag in the Monte-Carlo simulation. $\odot$ represents the element-wise multiplication.

The Gaussian noise strategy is used to add noise to the features, which prevents the distortion of original data and increases the diversity of the input samples.

#### 2) FINETUNE BASED ON FOCAL LOSS

The pre-training is agnostic to phase angles or power generation, but only extracts generic features from the state data. Thus, the results of the pre-training can be shared by the models $f_{\theta,t}(P_D \rightarrow V_\theta)$ and $f_{G,t}(P_D \rightarrow P_G)$. On the basis of pre-training results, network parameters need to be fine-tuned with the generation and angle labels.

The finetuning stage is oriented to minimize the gap between predictions and real values. To make full use of the minority samples, the focal weight is incorporated into the traditional MSE function.

The traditional MSE functions in teacher models are expressed as,

$$L_{\theta t} = \|V_\theta - V_{\theta t}\|_2^2 = \frac{1}{n_\theta}\sum_{i\in\Omega_N}\left(V_{\theta,i} - V_{\theta t,i}\right)^2 \tag{5}$$

$$L_{Gt} = \|P_G - P_{Gt}\|_2^2 = \frac{1}{n_G}\sum_{i\in\Omega_G}\left(P_{G,i} - P_{Gt,i}\right)^2 \tag{6}$$

where $L_{\theta t}$ and $L_{Gt}$ are the loss function of models $f_{\theta,t}(P_D \rightarrow V_\theta)$ and $f_{G,t}(P_D \rightarrow P_G)$, respectively. $V_{\theta,i}$ and $V_{\theta t,i}$ are the $i$th element in the actual value and teacher prediction, respectively. $P_{G,i}$ and $P_{Gt,i}$ are the $j$th element in the target and actual output vector. $||V_{\theta,i} - V_{\theta t,i}||_2^2$ and $||P_{G,i} - P_{Gt,i}||_2^2$ denote the MSE. $n_\theta$ and $n_G$ are the numbers of the bus and generating units, respectively. For such an MSE function, it gives the same weight to the error of each variable $V_{\theta,i}$ or $P_{G,i}$.

To increase the sensitivity and sample efficiency of rare samples, a new loss function is proposed by introducing Focal Loss. It is to address the extreme imbalance between positive and negative samples by supporting some categories with discrete labels such as 0 or 1 [23]. For the learning task in this paper, the label is a continuous value between 0 and 1. Therefore, it is necessary to ensure the previous balanced positive and negative, hard and easy sample properties and to allow it to support the supervision of continuous values. It naturally leads to one of our expanded forms of Focal Loss on continuous labels, which we call Continuous Focal Loss (CFL). CFL functions are expressed as,

$$L_{\theta t} = \frac{1}{n_\theta}\sum_{i\in\Omega_N} a_i p_i \left(V_{\theta,i} - V_{\theta t,i}\right)^2 \tag{7}$$

$$L_{Gt} = \frac{1}{n_G}\sum_{i\in\Omega_G} a_i p_i \left(P_{G,i} - P_{Gt,i}\right)^2 \tag{8}$$

where $a_i$ and $p_i$ are the focal weights. $a_i$ is determined by the target value, while $p_i$ is related to the prediction value. $a_i$ can be expressed as,

$$a_i = \frac{1}{\ln\left(1.1 + P[Y_i]\right)}, \quad Y_i = V_{\theta,1}, V_{\theta,2}, \cdots P_{G,1}, P_{G,2}\cdots \tag{9}$$

where $Y_i$ is the $i$th element in the label vector, which is the angle or generation. $P[Y_i]$ is the proportion of the corresponding category of $Y_i$. The category is obtained by dividing the entire range of values into 20 intervals. The proportion is determined by the number of labels whose values fall into the same interval.

From the perspective of deep learning, the outputs which are often equal to zero or maximum reflect that the features are more distinctive and easier to learn. For these categories, the corresponding parameters $p$ are attributed with lower values. The weight $p$ is obtained from the predicted values after a power operation as follows,

$$\begin{aligned} p_i &= Y_{ti}^r\left(1 - Y_{ti}\right)^r + 0.5, \\ Y_{ti} &= V_{\theta t,1}, V_{\theta t,2}, \cdots P_{Gt,1}, P_{Gt,2}\cdots \end{aligned} \tag{10}$$

where $r$ is set as 1. $Y_{ti}$ is the $i$th value of the teacher model prediction and the subscript $t$ denotes the teacher model. The states whose label is close to 0 or 1 are easy to learn, so the percentage should be smaller.

The gradient descent algorithm is more suitable to minimize the loss function in deep learning models [24]. The gradient descent process can be expressed as,

$$\Delta w_{\theta k}^{(l,\tau)} = \eta\left(\frac{1}{m}\sum_{k=1}^{m}\frac{\partial L_{\theta,t}^{\tau}}{\partial W_{\theta k}^{(l,\tau)}}\right) - \mu \times \Delta w_{\theta k}^{(l,\,\tau-1)} \tag{11}$$

$$\Delta w_{Gk}^{(l,\tau)} = \eta\left(\frac{1}{m}\sum_{k=1}^{m}\frac{\partial L_{G,t}^{\tau}}{\partial W_{Gk}^{(l,\tau)}}\right) - \mu \times \Delta w_{Gk}^{(l,\,\tau-1)} \tag{12}$$

where $w_{\theta k}^{(l,\tau)}$ and $w_{Gk}^{(l,\tau)}$ are the $k$th weights in the $l$th layer after $\tau$th updating. $\eta$ is the learning rate. $m$ is the neural number of the $l$th layer. $\mu$ is the momentum. $L_{\theta,t}^{\tau}$ and $L_{G,t}^{\tau}$ are the loss function. $\Delta w_{\theta k}^{(l,\tau)}$ and $\Delta w_{Gk}^{(l,\tau)}$ are the parameter alterations of $\tau$ th iteration.

### C. TRAINING STUDENT MODEL BASED ON ANNEALING KNOWLEDGE DISTILLATION

This section focuses on proposing the learning method of knowledge distillation [22]. The knowledge distillation learning process for the regression model is presented and the annealing strategy [25] is combined afterward.

Since the pretraining stage is task-agnostic and unrelated to the downstream work, the result of the pre-train stage can be reused for the student model initialization and only fine-tuning is required.

The knowledge-distillation-based finetuning stage aims to approximate the results of existing single-task models, which is achieved by minimizing the gap between the prediction of

the teacher model and the student model.

$$L_{s,t} = L_{\theta s,\theta t} + L_{Gs,Gt} \tag{13}$$

$$L_{\theta s,\theta t} = \frac{1}{n_\theta} \sum_{i \in \Omega_N} a_i p_i \left(V_{\theta t,i} - V_{\theta s,i}\right)^2 \tag{14}$$

$$L_{Gs,Gt} = \frac{1}{n_G} \sum_{i \in \Omega_G} a_i p_i \left(P_{Gt,i} - P_{Gs,i}\right)^2 \tag{15}$$

where $L_{s,t}$ is the loss function, which evaluates the difference between teacher prediction and student prediction. $L_{\theta s,\theta t}$ is the loss function of $V_{\theta t}$ and $V_{\theta s}$. $L_{Gs,Gt}$ is the loss function of $P_{Gt}$ and $P_{Gs}$. $V_{\theta s}$ and $P_{Gs}$ are predictions of the student model. $V_{\theta s,i}$ and $P_{Gs,i}$ are the $i$th angle and $i$th generation predicted by the student model.

The difference between the teacher model and the student model is the same as the calculation of the loss function. Minimizing this difference function is equivalent to training the multitask model with the predicted values of the single-task model as labels. And this is undoubtedly less accurate than using authoritative labels since the predicted values are always not 100% accurate. To cope with this problem, teacher models are regarded as the lower bound and the usage is specified as follows,

$$L = \lambda_\theta L_{\theta s,\theta t} + (1 - \lambda_\theta) L_{\theta s} + \lambda_G L_{Gs,Gt} + (1 - \lambda_G) L_{Gs} \tag{16}$$

where $L$ is a comprehensive loss function, which is combined by $L_{\theta s,\theta t}$, $L_{\theta s}$, $L_{Gs,Gt}$, and $L_{Gs}$. $\lambda_\theta$ and $\lambda_G$ are the weights of teacher models, which are determined after comparison as follows,

$$\lambda_\theta = \begin{cases} 0, & if\ L_{\theta t} > L_{\theta s} \\ 1, & else \end{cases} \tag{17}$$

$$\lambda_G = \begin{cases} 0, & if\ L_{Gt} > L_{Gs} \\ 1, & else \end{cases} \tag{18}$$

where $L_{\theta t}$ and $L_{Gt}$ are obtained by (7) and (8). $L_{\theta s}$ and $L_{Gs}$ are the loss function value of the student model, which can be expressed as,

$$L_{\theta s} = \frac{1}{n_\theta} \sum_{i \in \Omega_N} a_i p_i \left(V_{\theta i} - V_{\theta s,i}\right)^2 \tag{19}$$

$$L_{Gs} = \frac{1}{n_G} \sum_{i \in \Omega_G} a_i p_i \left(P_{Gi} - P_{Gs,i}\right)^2 \tag{20}$$

In (17) and (18), errors of the student and teacher models are compared. If the teacher model outperforms the student model, then the student model learns from teacher models. Otherwise, the student is trained by actual labels.

To avoid the accuracy limitations of teacher models, the teacher annealing approach is adopted in our training. A dynamic annealing weight is introduced in the two-objective loss function, which can be expressed as,

$$\lambda_\theta = \begin{cases} 0. & if\ L_{\theta t} > L_{\theta s} \\ 1 - \dfrac{e}{e_{\max}} & else \end{cases} \tag{21}$$

$$\lambda_G = \begin{cases} 0. & if\ L_{Gt} > L_{Gs} \\ 1 - \dfrac{e}{e_{\max}} & else \end{cases} \tag{22}$$

where $e$ and $e_{\max}$ are the finetune epoch index and the max epoch number, respectively. $\lambda$ is the dynamic annealing weight, which increases linearly with iteration.

The (19) and (20) indicate the knowledge distillation process is divided into two stages. In the early stage, the student model learns from the teacher models $f_{\theta,t}(P_D \rightarrow V_\theta)$ and $f_{G,t}(P_D \rightarrow P_G)$. With the increasing finetune epochs, the student model experiences a gradual transition to supervised learning under target labels.

## IV. ALGORITHM AND FLOWCHART OF THE DATA-DRIVEN OPF

The proposed approach provides a sample-efficient OPF-solving framework to determine the optimal generation dispatch. The overall process is shown in Fig.5.

*Step 1:* Input the historical data or simulation data of the power flow under different system states.

*Step 2:* Select the unlabeled data for SDAE pretraining and the encoder layer parameters are determined.

*Step 3:* The labeled samples are classified into angle labels and generation labels.

*Step 4:* For angle labels, a new network is constructed based on the encoder.

*Step 5:* Finetune the network in step 4 and obtain a teacher network.

*Step 6:* Calculate teacher predictions and focal loss function.

(With the generation labels, the teacher model $f_{G,t}(P_D \rightarrow P_G)$ is trained in parallel so as with steps 4-6.)

*Step 7:* Construct a new network as a student model based on the pre-trained encoder.

*Step 8:* Set the maximum epoch and initialize the current epoch counter.

*Step 9:* Calculate the difference between student and teacher predictions as (13)-(14), as well as the loss function for each variable according to (16).

*Step 10:* Comparing. If the student is more precise, the weight of teacher $\lambda$ is zero. Otherwise, $\lambda$ decreases with finetuning epochs linearly as (21)-(22).

*Step 11:* The weighted sum of the loss function is calculated and used in parameter updating.

*Step 12:* Repeat steps 10-12 until the epoch counter reaches the limitation and the OPF training is finished.

## V. CASE STUDY

Numerical test cases are carried out on the RTS-79 system [26]. The 9-bus [27], 118-bus [28], and southern Brazil power systems [29], [30] are involved to test the scalability of the proposed method. The hardware and software used in the case study include Intel i5-10600KF CPU, 16G RAM, WINDOWS 10, and Python 3.8. The Gurobi toolkit is also involved in benchmark calculation when evaluating accuracy.

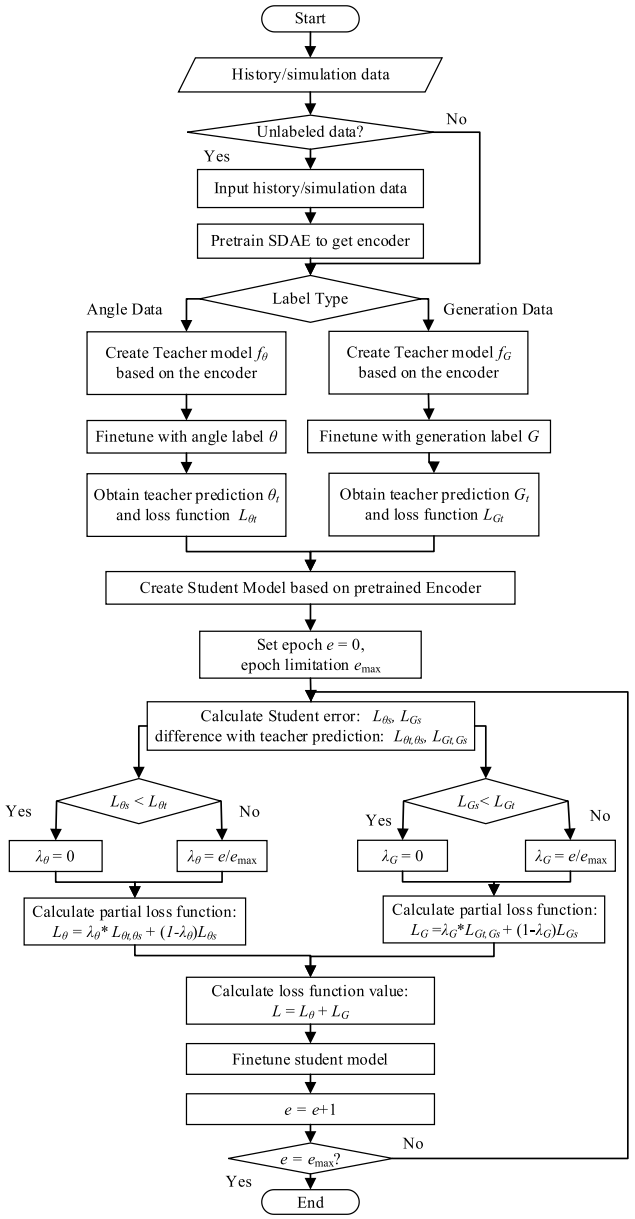


**FIGURE 5.** Flowchart of the proposed method.

The hyper-parameters of the deep learning model in the case study are listed as follows. The number of hidden layers is 3 for each neural network. The teacher encoder size is (150, 200, 250) for each layer and the student model is half the size accordingly. For the pretraining stage, the SGD optimizer is used whose initial learning rate is 0.1 and momentum is 0.8. For the finetune stage, the Adam optimizer is involved with parameter $\beta$ as (0.7, 0.92) and learning rate as 0.0005. The total epochs are 40 for pretraining and 50 for model finetuning. The variance of Gaussian noise is set as 5%. Batch sizes for pre-training and finetune are 256 and 128, respectively.

The accuracy indices are defined as follows,

$$Ac_\theta = \frac{n\left(\left|V'_\theta - V_\theta\right| \le \theta_0\right)}{n} \tag{23}$$

$$Ac_G = \frac{n\left(\left|P'_G - P_G\right| \le G_0\right)}{n} \tag{24}$$

where $V'_\theta$ and $V_\theta$ are the predicted and actual phase angles. $P'_G$ and $P_G$ are the predicted and actual power generation. The judgment thresholds are set to 0.1 rad and 1 MW.

### A. PERFORMANCE OF THE PROPOSED METHOD

A variety of methods in Table 1 are compared with the proposed method (M6) in the RTS-79 system. All the methods are based on the same training dataset whose size is 15000. The depth of the random forest method M1 is 8. M2 directly predicts both phase angle and generation output in the same SDAE network. In M3, the teacher models are trained separately with the MSE loss function. In M4, they are trained with the focal loss function. The knowledge distillation process is integrated into M5 and M6 where teacher networks are obtained via M4.

Various methods in Table 1 are applied to the RTS-79 system. The computational performance is displayed in Table 2. M2, M5, and M6 have the same network structure, and Fig.6 compares relative errors of their predicted node phase angles.

M0 is the method that invokes the Matpower toolkit for solving, and its outcomes are used as the benchmark. As shown in Table 2, the solution time of traditional optimization algorithms is 248.2683 s.

Comparing M1 and M2, the SDAE network is proved to be more effective in predicting OPF. This is because the computational effectiveness of random forests depends greatly on the manual selection of features. Moreover, the training effectiveness of random forests is limited by the size of the output volume. The requirements of tree size and layers in M1 increase accordingly with the output scale. The problem of preferring a large number of parameters is difficult to solve and ultimately detrimental to accuracy.

Results of M2 show that task decomposition enables the network to concentrate on one particular problem. By decomposing the task, interactions between unrelated features can be avoided to occupy parameter resources, so that parameters can work together to achieve an accurate output.

Fig.7 compares the generation results obtained by M3 and M4. The advanced focal loss function is effective in improving the prediction of unbalanced distribution variables. This technique changes the weighting factor of the data difference, allowing the model to notice small sample data without sample data addition. In the RTS-79 system, generation units 23, 24, and 25-30 are always prioritized in generation dispatch due to their low cost. The other generating units operate only in fewer states with high load levels. The generation labels show uneven distribution, but the Focal loss function enhances the attention of the network to the minority data, thus improving the overall effectiveness of the method.

The results of the M5 and M6 in Table 2 show that knowledge distillation can integrate multiple high-precision single-task models while maintaining the same level of

**TABLE 1. Methods in details.**

| Methods | Description | Details | | | |
|---|---|---|---|---|---|
| | | Task decomposition | Focal loss function | Knowledge distillation | Teacher annealing |
| M0 | Model-based benchmark | | | | |
| M1 | Random Forest | | | | |
| M2 | Original SDAE | × | × | × | × |
| M3 | Teacher SDAEs for angle and generation predictions | √ | × | × | × |
| M4 | | √ | √ | × | × |
| M5 | Student SDAE based on knowledge distillation | √ | √ | √ | × |
| M6 | | √ | √ | √ | √ |

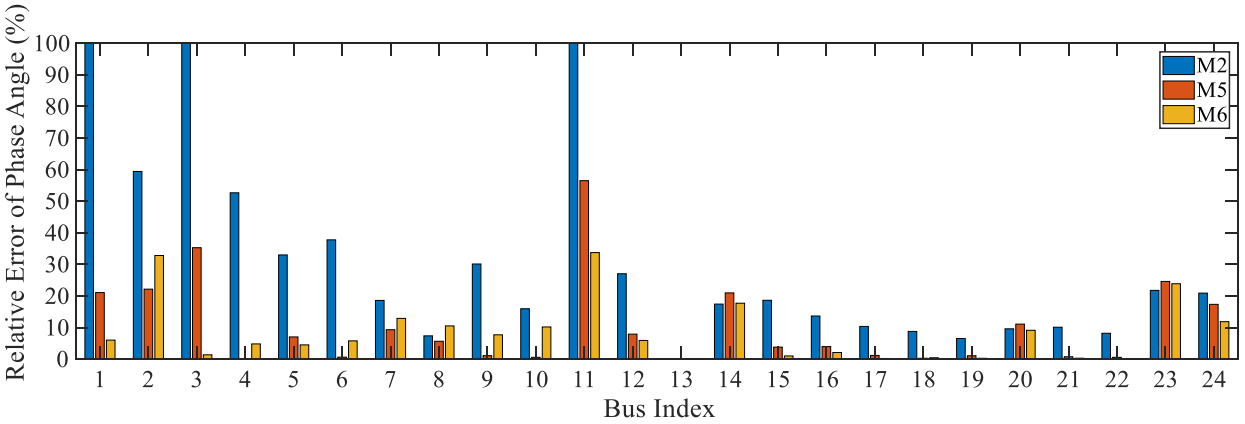


**FIGURE 6. Prediction comparison of the relative error of phase angle in the RTS-79 system.**

**TABLE 2. Accuracy in RTS-79 system.**

| Methods | Time (s) | Accuracy index | |
|---|---|---|---|
| | | $Ac_\theta$ | $Ac_G$ |
| M0 | 248.2683 | 1 | 1 |
| M1 | 0.5774 | 0.67796 | 0.75022 |
| M2 | 0.00292 | 0.78911 | 0.79181 |
| M3 | 0.0568 | 0.86366 | 0.79911 |
| M4 | 0.0658 | 0.85093 | 0.86660 |
| M5 | 0.0588 | 0.86002 | 0.81433 |
| **M6** | **0.0767** | **0.88980** | **0.86815** |

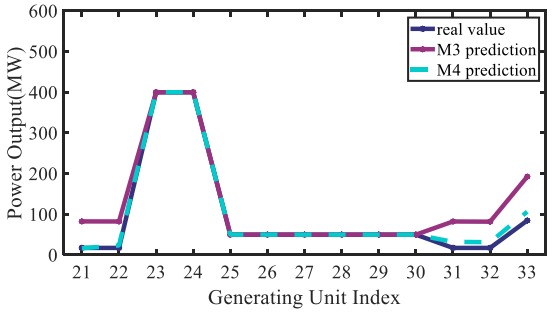


**FIGURE 7. Prediction comparison of M3 and M4.**

accuracy. Fig.6 shows that the teacher annealing strategy can achieve higher accuracy in knowledge distillation.

### *B. FEASIBILITY ON SMALL SAMPLE SIZE*

As shown in Table 3, by reusing the pre-trained model, the proposed training method using the knowledge distillation strategy can achieve high accuracy results in one minute.

Table 3 and Fig.8 present the results of the application of the proposed method on a small sample dataset. It shows that the pre-training strategy helps to improve the accuracy on small sample size. This is because a large amount of unlabeled data can be used to train the shallow layer of the SDAE network in the pre-training phase, thus reducing the learning burden in the supervised phase. The proposed method maintains a higher accuracy level over other methods, demonstrating its feasibility for small sample states. For example, the M2 requires 15000 samples to roughly attain the accuracy that the proposed method (M6) achieves with 200 samples.

### *C. SCALABILITY ANALYSIS*

The proposed method is also applied to the systems with different scales and the results are presented in Table 4. the Brazilian system has 242 nodes with 53 generators, and the specific settings such as line capacity can be found

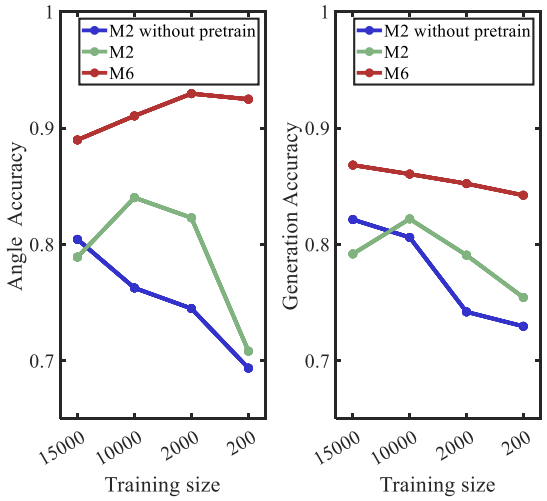


**FIGURE 8.** Comparisons on training data size.

**TABLE 3.** Accuracy improvement of pre-training in the RTS-79 system.

| Size of Training data | Methods | Training Time (s) | Accuracy index | |
|---|---|---|---|---|
| | | | $Ac_\theta$ | $Ac_G$ |
| 15000 | M2 without pretrain | 7.061 | 0.804348 | 0.821355 |
| 15000 | M2 | 65.495 | 0.789115 | 0.791812 |
| 15000 | M6 | 74.228 | 0.889796 | 0.868153 |
| 10000 | M2 without pretrain | 4.872 | 0.762594 | 0.806033 |
| 10000 | M2 | 42.249 | 0.840198 | 0.821935 |
| 10000 | M6 | 49.654 | 0.910454 | 0.860447 |
| 2000 | M2 without pretrain | 0.1117 | 0.744890 | 0.742027 |
| 2000 | M2 | 38.889 | 0.822885 | 0.790783 |
| 2000 | M6 | 48.669 | 0.929577 | 0.852235 |
| 200 | M2 without pretrain | 0.094 | 0.693679 | 0.729629 |
| 200 | M2 | 27.527 | 0.708294 | 0.754456 |
| 200 | M6 | 35.576 | 0.924794 | 0.842256 |

**TABLE 4.** Applying to different systems.

| Test system | Time (s) | | Accuracy index | |
|---|---|---|---|---|
| | Train | Apply | $Ac\theta$ | $Ac_G$ |
| IEEE 9 | 43.03816 | 0.00399 | 0.987467 | 0.999617 |
| RTS 79 | 48.669 | 0.0767 | 0.924794 | 0.842256 |
| IEEE 118 | 198.38320 | 0.057817 | 0.814898 | 0.938225 |
| Southern Brazil | 220.06567 | 0.111701 | 0.875065 | 0.938225 |

in [30]. For each test system, training data contains 200 samples.

It can be seen that the accuracy of the proposed method slightly decreases as the size of the system increases but maintains an acceptable level. Therefore, the proposed method is suitable for power systems with different scales.

## VI. CONCLUSION

This paper proposes a sample-efficient method based on DC-OPF decomposition and knowledge distillation to enable training with limited samples. Numerical results prove the proposed DC-OPF task decomposition can improve the training generalization on limited samples, and the annealing operation in the knowledge distillation can finally enhance the accuracy by 10% for angle and 8% for power generation. Moreover, the accuracy improvement of the proposed method is over 12% which is more significant in lower-data setups. Compared to the plain deep learning method, the proposed method can reduce the sample size by 98.6% and improve the accuracy in phase angle by 12% and generation by 2%. In the future, it will be studied in AC-OPF with consideration of reactive power and voltage magnitude.

## REFERENCES

[1] Z. Liu, K. Hou, H. Jia, J. Zhao, D. Wang, Y. Mu, and L. Zhu, "A Lagrange multiplier based state enumeration reliability assessment for power systems with multiple types of loads and renewable generations," *IEEE Trans. Power Syst.*, vol. 36, no. 4, pp. 3260–3270, Jul. 2021.

[2] M. B. Cain, R. P. O'Neill, and A. Castillo, "History of optimal power flow and formulations: Optimal power flow paper 1," in *Federal Energy Regulatory Commission 1*, 2012, pp. 1–36. [Online]. Available: https://api.semanticscholar.org/CorpusID:109052209

[3] O. A. Alimi, K. Ouahada, and A. M. Abu-Mahfouz, "A review of machine learning approaches to power system security and stability," *IEEE Access*, vol. 8, pp. 113512–113531, 2020.

[4] H. Dharmawardena and G. K. Venayagamoorthy, "A distributed data-driven modelling framework for power flow estimation in power distribution systems," *IET Energy Syst. Integr.*, vol. 3, no. 3, pp. 367–379, Sep. 2021.

[5] T. Zhao, X. Pan, M. Chen, A. Venzke, and S. H. Low, "DeepOPF: A deep neural network approach for DC optimal power flow for ensuring feasibility," in *Proc. IEEE Int. Conf. Smart Grid Commun.*, Tempe, AZ, USA, Nov. 2020, pp. 1–6.

[6] X. Pan, M. Chen, T. Zhao, and S. H. Low, "DeepOPF: A feasibility-optimized deep neural network approach for AC optimal power flow problems," 2020, *arXiv:2007.01002*.

[7] D. Deka and S. Misra, "Learning for DC-OPF: Classifying active sets using neural nets," in *Proc. IEEE Milan Power Tech.*, Jun. 2019, pp. 1–6.

[8] F. Hasan, A. Kargarian, and J. Mohammadi, "Hybrid learning aided inactive constraints filtering algorithm to enhance AC OPF solution time," *IEEE Trans. Ind. Appl.*, vol. 57, no. 2, pp. 1325–1334, Mar. 2021.

[9] K. Baker, "Learning warm-start points for AC optimal power flow," 2019, *arXiv:1905.08860*.

[10] L. Chen and J. E. Tate, "Hot-starting the AC power flow with convolutional neural networks," 2020, *arXiv:2004.09342*.

[11] M. Chatzos, T. W. K. Mak, and P. V. Hentenryck, "Spatial network decomposition for fast and scalable AC-OPF learning," *IEEE Trans. Power Syst.*, vol. 37, no. 4, pp. 2601–2612, Jul. 2022.

[12] Z. Yan and Y. Xu, "Real-time optimal power flow: A Lagrangian based deep reinforcement learning approach," *IEEE Trans. Power Syst.*, vol. 35, no. 4, pp. 3270–3273, Apr. 2020.

[13] D. Biagioni, P. Graf, X. Zhang, A. S. Zamzam, K. Baker, and J. King, "Learning-accelerated ADMM for distributed DC optimal power flow," in *Proc. Amer. Control Conf. (ACC)*, New Orleans, LA, USA May 2021, pp. 1–6.

[14] J. Liu, Z. Yang, J. Zhao, J. Yu, B. Tan, and L. Wenyuan, "Explicit data-driven small-signal stability constrained optimal power flow," *IEEE Trans. Power Syst.*, vol. 37, no. 5, pp. 1–12, Sep. 2021, doi: 10.1109/TPWRS.2021.3135657.

[15] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *J. Comput. Phys.*, vol. 378, pp. 686–707, Feb. 2019.

[16] M. Chatzos, F. Fioretto, T. W. K. Mak, and P. V. Hentenryck, "Highfidelity machine learning approximations of large-scale optimal power flow," 2020, *arXiv:2006.16356*.

[17] F. Fioretto, T. W. Mak, and P. Van Hentenryck, "Predicting AC optimal power flows: Combining deep learning and Lagrangian dual methods," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 630–637.

[18] X. Lei, Z. Yang, J. Yu, J. Zhao, Q. Gao, and H. Yu, "Data-driven optimal power flow: A physics-informed machine learning approach," *IEEE Trans. Power Syst.*, vol. 36, no. 1, pp. 346–354, Jan. 2021.
[19] M. K. Singh, V. Kekatos, and G. B. Giannakis, "Learning to solve the AC-OPF using sensitivity-informed deep neural networks," *IEEE Trans. Power Syst.*, vol. 37, no. 4, pp. 2833–2846, Jul. 2022.
[20] Y. Chen, S. Lakshminarayana, C. Maple, and H. V. Poor, "A meta-learning approach to the optimal power flow problem under topology reconfigurations," *IEEE Open Access J. Power Energy*, vol. 9, pp. 109–120, 2022.
[21] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[22] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.
[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2017, *arXiv:1708.02002*.
[24] S. Suryansh. (2018). *Gradient Descent: All You Need to Know*. [Online]. Available: https://hackernoon.com/gradient-descent-aynk-7cbe95a778da.
[25] A. Jafari, M. Rezagholizadeh, P. Sharma, and A. Ghodsi, "Annealing knowledge distillation," 2021, *arXiv:2104.07163*.
[26] P. M. Subcommittee, "IEEE reliability test system," *IEEE Trans. Power App. Syst.*, vol. PAS-98, no. 6, pp. 2047–2054, Nov. 1979.
[27] R. H. Al-Rubaiy and W. K. Al-Jubor, "Study and simulation of IEEE 9 bus system with UPFC for transient stability analysis," *J. Appl. Sci. Res.*, vol. 12, no. 8, Aug. 2016.
[28] A. Chang and M. Adibi, "Power system dynamic equivalents," *IEEE Trans. Power App. Syst.*, vols. PAS–89, no. 8, pp. 1737–1744, Nov. 1970.
[29] *Brazilian Interconnected Power System Maps*. Accessed: Apr. 23, 2022. [Online]. Available: http://www.ons.org.br/paginas/sobre-o-sin/mapas
[30] *Brazilian Interconnected Power System Data*. Accessed: Apr. 23, 2022. [Online]. Available: https://sites.google.com/view/southernbrazilian/data

**ZIHENG DONG** received the B.S. degree in electrical engineering from Tianjin University, Tianjin, China, in 2020, where she is currently pursuing the M.S. degree with the Electrical Engineering Department. Her research interests include power system reliability and data-driven methods for power systems.

**KAI HOU** (Member, IEEE) received the Ph.D. degree in electrical engineering from Tianjin University, Tianjin, China, in 2016. He is currently an Associate Professor with the Electrical Engineering Department, Tianjin University. His research interests include reliability and risk assessments of power systems, integrated energy systems, and smart grids.

**ZEYU LIU** (Student Member, IEEE) received the B.S. degree from Tianjin University, Tianjin, China, in 2018, where he is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering. His research interests include power system reliability assessment, uncertainty analysis, and integrated energy systems planning and reliability evaluation.

**XIAODAN YU** received the B.S. and M.S. degrees in electrical engineering from Tianjin University, Tianjin, China, in 1996 and 1999, respectively. She is currently an Associate Professor with the Electrical Engineering Department, Tianjin University. Her research interests include power reliability assessment, stability analysis and control, distribution network planning, and integrated energy systems.

**HONGJIE JIA** (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Tianjin University, Tianjin, China, in 1996, 1999, and 2001, respectively. He is currently a Professor with the Electrical Engineering Department, Tianjin University. His research interests include power reliability assessment, stability analysis and control, distribution network planning, and integrated energy systems.

**CHI ZHANG** received the M.S. degree in electrical engineering from Tianjin University, in 2020. He is currently an Engineer at EHV Branch Company, State Grid Jiangsu Electric Power Company Ltd. His research interest includes UHVDC power transmission technologies.