Online Self-Distillation for Electric Load Demand Forecasting on Greek Energy Market

1st Maria Tzelepi
Dept. of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
mtzelepi@csd.auth.gr

2nd Alkmini Sapountzaki
Dept. of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
alkminis@csd.auth.gr

3rd Nikitas Maragkos
Dept. of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
maragkosn@csd.auth.gr

4th Anastasios Tefas
Dept. of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
tefas@csd.auth.gr

***Abstract*—Even though Knowledge Distillation (KD) has been extensively studied during the recent years considering classification tasks, the research considering forecasting problems is extremely limited, despite the fact that the underlying reasons for the need of KD, linked with requirement for effective and fast models, are also apparent in such problems. In this work, we propose an online distillation method, named *Online Self-Distillation for Forecasting* (OSDF) for ameliorating the baseline forecasting performance considering the Electric Load Demand Forecasting (ELDF) problem on Greek energy market. The experimental evaluation, considering a typical and a fully realistic setup validates the effectiveness of the proposed OSDF method.**

***Index Terms*—Online distillation, Self-distillation, Energy load demand forecasting, Greek energy market**

I. Introduction

Electric Load Demand Forecasting (ELDF) describes the task of predicting the expected electricity demand by analyzing historical load data. ELDF falls into three categories based on the time-scale. That is, short-term load forecast which concerns forecasting of a few hours up to one-day ahead or a week ahead, mid-term load forecast which concerns a time-period of a week to one year, and the long-term forecast with a time frame of up to several years ahead. In this paper we deal with short-term forecasting.

ELDF constitutes, in general, a challenging task in the energy markets due to the variety of factors affecting the forecasting performance [1], [2]. Load demand forecasting is associated with many critical applications ranging from power system operation and planning to energy trading [3], allowing power companies to achieve an efficient balance between demand and supply, avoiding excess reserve of power generation or power interruptions due to load shedding. The aforementioned reasons dictate the demand for accurate load demand forecasting models. This demand has fueled the research interest over the past years [4].

Motivated by the remarkable accomplishments of DL algorithms in a wide variety of problems, ranging from image classification [5], and image retrieval [6] to financial time series forecasting [7], DL algorithms have been proposed so as to tackle the ELDF task [8]–[10] achieving notable performance.

For instance, in [11] a Deep Belief Network (DBN) embedded with parametric Copula models is proposed to forecast the hourly load of a power grid. This model achieved superior performance as compared to previous approaches in both daily and weekly predictions of hourly granularity. In a very recent work [12], a study for the prediction performance considering the following's day load demand separately for all four seasons of the year has been conducted. An hybrid model of a neural network and an Long Short-Term Memory (LSTM) achieved the best results.

In this work, we deal with the Greek energy market. Surveying the relevant literature, we come across several works of ELDF on the Greek Energy Market. For instance, in [13], the effect of dimensionality reduction methods in the day-ahead forecasting performance of neural networks is investigated. Furthermore, in [14] a fuzzy-based ensemble that uses hybrid DL networks is proposed for load demand prediction of the next week. More specifically, initially, a fuzzy clustering technique creates an ensemble prediction and after that a pipeline of Radial Basis Function Neural Network (RBFNN) transforms the data in order to be fitted in a CNN. Finally, the output of this pipeline goes into an hybrid neural network consisting of an RBF, a CNN and two fully connected layers. Additionally, in [15] a methodology which helps the exploitation of the statistical properties of each time series with main focus the optimization of CNN's hyper-parameters is proposed.

Subsequently, in [16] a more realistic approach of ELDF on Greek energy market is investigated. More specifically, the vast majority of the aforementioned methods make two basic assumptions, considering the ELDF problem. First, it is assumed that any historical load data before the day whose load demand we want to predict are available and can be

979-8-3503-9958-5/22/$31.00 ©2022 IEEE

used. Second, real weather information of the aforementioned day is also considered available. In [16], a more realistic setup is followed, considering an information gap between the prediction day and the past load data, retaining however the assumption regarding the weather information. A strategy for filling the aforementioned information gap is proposed, along with a novel loss function.

Finally, in [17], an evaluation study regarding the optimal input features and an effective model architecture considering the ELDF problem on the Greek energy market is performed. Subsequently, using the optimal features and model, a novel regularization method is proposed, ameliorating the baseline forecasting.

In this paper, we also deal with ELDF task on the Greek energy market, proposing an online self-distillation method for improving the performance of a day-ahead forecasting model. Considering generic classification tasks, *Knowledge Distillation* (KD) [18] has been established as an auspicious technique for training fast and effective models by transferring the knowledge acquired usually from more powerful models. Despite its effectiveness, KD suffers from some shortcomings associated with the complex training pipeline (i.e., computationally demanding and time-consuming procedure), since the so-called teacher model should be trained, and after its convergence, the acquired knowledge is transferred to the so-called student model. Thus, *online* distillation has been arisen in the recent literature as a method from circumventing these flaws, by simplifying the training pipeline to a single-stage, omitting the stage of pre-training the teacher model [19], [20].

Event though, KD has been extensively studied during the recent few years with a wide spectrum of applications [21], [22], the research on distillation considering forecasting problems is very limited, despite the fact that the underlying reasons for the need of KD, linked with requirement for effective and fast models, are apparent in such scenarios too, [16]. Surveying the literature, we first come across a KD-based method for wind power prediction in [23]. The method aims to bridge large (park with bid data) and small-scale (turbine with small data) forecasting by proposing a KD regression approach. Subsequently, considering financial forecasting problems, a KD method that exploits sentiment information as a source of additional supervision throughout the training procedure is proposed in [24].

In this paper, apart from the typical setup described above, regarding the unconstrained availability of previous load data and weather information, we also implement a fully realistic setup where neither weather information is used and also there is an information gap in the past load data. We apply the proposed online distillation method, so as to mine further knowledge about the way the model learns to forecast in order to ameliorate its prediction performance on both the setups and evaluate the forecasting performance.

The rest of the manuscript is structured as follows. First, the proposed online distillation method is presented in Section II. Subsequently, in Section III the experiments performed in order to validate the proposed online distillation method are provided. Finally, some conclusions are drawn in Section IV.

II. Proposed Method

In this paper, we propose a novel online distillation method for ameliorating the performance of a neural model, considering the energy load demand forecasting task.

Generally, KD is established in the idea that, considering a classification problem, it is beneficial to train a model with the so-called soft labels that encode additional information about the way the model learns to generalize, instead of training it with the ground truth targets (hard labels). This additional information can be derived either from the same or another model. Considering, for example, the digit recognition task [25], KD argues that by incorporating the additional knowledge that a digit 9 is similar to another digit 8, the generalization ability of the classifier is enhanced. In this paper, the aforementioned idea is extended to forecasting tasks. Specifically, considering the electric load demand forecasting task, we argue that, since similar input features lead to similar forecasted load values, we can improve the generalization ability of the model by incorporating these similarities to the training process, instead of merely training with the actual load values. That is, by ignoring the similarities of the input features, outliers can significantly contribute to the loss, leading to poor generalization performance.

Therefore, we propose a simple yet effective online distillation method, named *Online Self-Distillation for Forecasting* (OSDF) which is able to acquire further knowledge about the way the model learns to forecast the load values, beyond the true load values, from the forecasting model itself, and also in an online manner. As it is also experimentally validated, it is advantageous to *soften* the ground truth targets, so as to regard the knowledge of the model, that is captured in the model's prediction, to the final prediction. This knowledge encodes the similarities of the input features, leading to better performance, as compared to the conventional training where the model is forced to merely match the ground truth, which may lead to over-fitting. Therefore, in the proposed method, the soft targets are computed as a combination of the *true load values* (ground truth targets) and the *forecasted values*.

More specifically, for an input space $\mathcal{X} \subseteq \Re^D$ and an output space $\mathcal{F} \subseteq \Re^d$, we consider a neural network for load demand forecasting, $\phi(\cdot\,; \mathcal{W}) : \mathcal{X} \rightarrow \mathcal{F}$ with weights $\mathcal{W}$. Considering a given input sample $\mathbf{x}_i,\ i = 1, \cdots, N$, its corresponding output of the network, $\phi(\mathbf{x}_i, \mathcal{W})$, and its ground truth vector $\mathbf{g}_i \in \Re^d$, the soft target $\mathbf{s}_i$ in the proposed training procedure is computed as follows:

$$\mathbf{s}_i = \mathbf{g}_i + \lambda \phi(\mathbf{x}_i, \mathcal{W}), \tag{1}$$

where $\lambda \in (0, 1)$ controls the relative importance of the contributed loss components.

Thus, in this paper, instead of training with the ground truth targets, we propose to train the forecasting model with the soft target $\mathbf{s}_i \in \Re^d$, in order to ameliorate the generalization

ability of the model. We use a common loss function considering forecasting tasks for training the model, i.e., the Mean Absolute Percentage Error (MAPE) loss.

Thus, in the proposed method, the loss, $\mathcal{L}_{osdf}$ is formulated as follows, using the computed soft targets:

$$\begin{aligned}\mathcal{L}_{osdf} &= \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\mathbf{s}_i - \phi(\mathbf{x}_i,\mathcal{W})}{\mathbf{s}_i}\right| \\ &\overset{(1)}{=} \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\mathbf{g}_i + \lambda\phi(\mathbf{x}_i,\mathcal{W}) - \phi(\mathbf{x}_i,\mathcal{W})}{\mathbf{g}_i + \lambda\phi(\mathbf{x}_i,\mathcal{W})}\right| \\ &= \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\mathbf{g}_i - (1-\lambda)\phi(\mathbf{x}_i,\mathcal{W})}{\mathbf{g}_i + \lambda\phi(\mathbf{x}_i,\mathcal{W})}\right|. \end{aligned} \tag{2}$$

Then, the model can be trained using stochastic gradient descent to minimize the loss:

$$\Delta\mathcal{W} = -\eta\frac{\vartheta\mathcal{L}_{osdf}}{\vartheta\mathcal{W}}, \tag{3}$$

where η corresponds to the learning rate, while it is noted that more advanced optimization methods can also be used, such as Adam [26].

We should note that the similarities encoded in the soft targets are dynamically learned during the training procedure which is driven by the ground truth targets. Therefore, it is expected that as the training progresses, more meaningful similarities are revealed, and thus more reliable soft targets are generated, which is also experimentally confirmed. Finally, it should be noted that the proposed online distillation method can also be realized in an offline fashion, however, this comes with additional cost, as we previously described.

III. Experimental Evaluation

A. Dataset

In this work, we propose an online self-distillation method for tackling the ELDF task on the Greek Energy Market. We use historical load data provided by the *Greek Public Power Corporation*. We also use weather information (i.e., temperature) obtained from OpenWeather[1]. We use 6 years of data for the model's training, that is load and temperature data for years 2012-2017, for validation load and temperature data for the year 2018, while for testing we use data for the year 2019.

B. Evaluation Metrics

MAPE is used as evaluation metric. Each experiment is repeated ten times, and we report the mean value of MAPE. Furthermore, we provide the curves of mean MAPE throughout the training epochs. Finally, training time in seconds is also reported.

[1]https://openweathermap.org/

C. Model Architecture and Input Features

In this work, a simple and lightweight MLP model is used since as it is stated in [16] simple models can accomplish competitive performance as compared to more complex ones. The model consisting of four layers, including the input and output layers. Regarding the input features, we use several inputs which are generally used in forecasting models [16], [17]. More specifically, we use the load of previous day, load of the day a week before, and load of the day a month before. Additionally, we use weather information of the previous day, of the day a week before, and of the day a month before, as well as temperature of the Target Day (TD), i.e., the day whose load demand we aim to predict. Finally, we utilize two binary indicators for weekend and holiday in order to assist the model capture the periodic and unordinary temporal characteristics of the load time series, while an indicator of which day of the week is the TD is also utilized. The input features are described in Table I. As it will be explained in the subsequent Section, different input features are utilized based on the setup. The output layer consists of 24 neurons, for each of the 24 hours of the day whose load demand we want to predict. The two intermediate layers consist of 1,000 and 400 neurons.

TABLE I
Description of Input Features

Abbreviation	Dim.	Description
L_d	24	Load of the day that is 1 day before TD
L_w	24	Load of the day that is 7 days before TD
L_m	24	Load of the day that is 28 days before TD
T_d	24	Corresponding temperature for L^d
T_w	24	Corresponding temperature for L^w
T_m	24	Corresponding temperature for L^m
T	24	Corresponding temperature for TD
D	1	Indicator of which day of the week is the TD
W	1	Indicator of TD being weekend
H	1	Indicator of TD being holiday

D. Implementation Details

The OSDF method is implemented using the Pytorch framework. The models are trained with Adam optimizer with an initial learning rate of 0.003. The mini-batch is set to 64 samples. The parameter λ in eq. (1) is set to 0.001. The models are trained on an NVIDIA GeForce GTX 1080 with 8GB of GPU memory for 1,000 epochs.

E. Experimental Setup

Two sets of experiments are performed for evaluating the proposed method. In the first set of experiments, the typical setup followed in the relevant literature is implemented. In this setup, all the previous load data before the TD are available. Furthermore, weather information (i.e., temperature) of TD is also available. In this case, all features presented in Table I are utilized (i.e., 171 features). It should be noted, that most of the methods of current literature make the assumption that the weather information of TD is available since this information can be acquired by solving another problem, known as air temperature forecasting [27]. In the second set of

experiments we implement a fully realistic setup where there is an information gap of 4 days before the TD, while temperature of TD is also unavailable. More specifically, considering the Greek Energy Market, previous week's energy data are being published each Thursday, creating a gap of ranging from four to ten missing days. In this paper, we investigate the scenario where there is a gap of four days. Therefore, in order to move to the energy load demand forecasting of the TD, we have to fill the aforementioned gap. To address this issue, we use a single model to predict the load of missing days, and then proceed to the final prediction of the TD. In this case, we use all the features except for all the temperature information (i.e., 75 features). The two implemented setups are illustrated in Fig. 1.

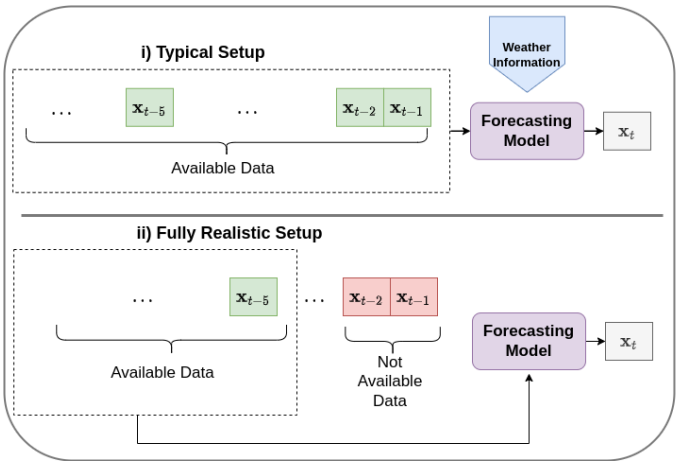

Fig. 1. Description of the two setups regarding the accessibility of previous load data and the weather information. In the typical setup, followed by the literature, both weather information and all previous data are used for the forecasting task. In the second fully realistic setup, the weather information of the target day is not available, and also there is also a gap of four days in the previous load data. Available load data are printed in green, while unavailable load data are printed in red.

F. Experimental Results

In this Section we present the experimental results for evaluating the proposed online distillation method for the ELDF task on the Greek Energy Market. We apply the proposed online distillation method considering the two different setups, and we compare the performance with the baseline training process of training without distillation. Best results are printed in bold.

First, the evaluation results in terms of MAPE considering the typical setup are presented in Table II. As it is shown the proposed method improves the baseline performance. Furthermore, test MAPE throughout the training epochs for the proposed OSDF method against the baseline training without distillation is presented in Fig. 2, where the superiority of the proposed method is illustrated.

Subsequently, in Table III and correspondingly in Fig. 3 the evaluation results for the fully realistic setup are provided. As it can be observed, the proposed OSDF method remarkably ameliorates the baseline forecasting performance of training without distillation. Furthermore, it can be observed that forecasting performance considering the fully realistic scenario is worse as compared to the typical and the partially realistic ones, which is reasonable since in the last case there is the information gap of the previous load data. In the recent literature [16], have been proposed strategies for improving the forecasting performance in such scenario, however we did not proceeded in this direction since this is beyond the scope of this work, which aims to propose and a novel online distillation method.

TABLE II
MAPE (%) FOR THE PROPOSED OSDF METHOD AGAINST THE BASELINE TRAINING WITHOUT DISTILLATION, CONSIDERING THE TYPICAL ELDF SCENARIO.

Method	MAPE (%)
W/o Distillation	1.975
OSDF	**1.964**

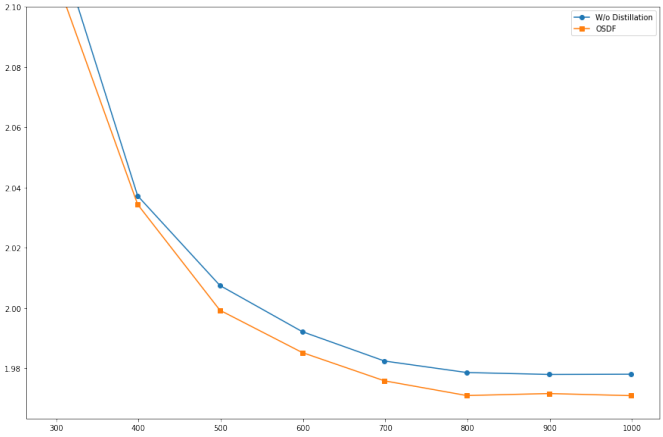

Fig. 2. Test MAPE throughout the training epochs for the proposed OSDF method against the baseline training without distillation, considering the typical ELDF scenario.

TABLE III
MAPE (%) FOR THE PROPOSED OSDF METHOD AGAINST THE BASELINE TRAINING WITHOUT DISTILLATION, CONSIDERING THE FULLY REALISTIC ELDF SCENARIO.

Method	MAPE (%)
W/o Distillation	5.967
OSDF	**5.754**

It should finally be emphasized that the proposed online distillation method achieves to improve the performance without affecting the training and inference time, since the additional knowledge through the soft targets is obtained from the model itself in an online manner. More specifically, 1 training epoch without distillation considering the typical setup (with input of 171 features) requires 0.286 sec, while the proposed distillation method requires additionally 4.81×10^{-5} sec.

IV. CONCLUSIONS

In this paper, we proposed an online distillation method in order to improve the baseline forecasting performance

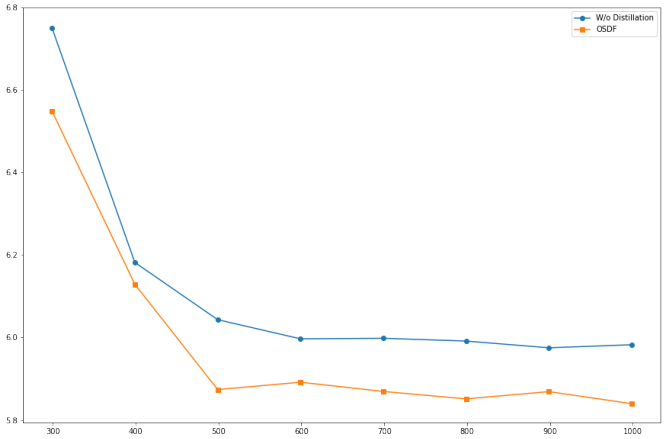

Fig. 3. Test MAPE throughout the training epochs for the proposed OSDF method against the baseline training without distillation, considering the fully realistic ELDF scenario.

considering the Electric Load Demand Forecasting (ELDF) problem on Greek energy market. The proposed OSDF method is able to mine further knowledge beyond the ground truth targets from the model itself and also in an online fashion. The experimental evaluation, considering a typical and a fully realistic setup validated the effectiveness of the proposed method.

ACKNOWLEDGMENT

This work is co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH - CREATE - INNOVATE (project code: T2EDK-03048).

REFERENCES

[1] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *International Journal of Forecasting*, vol. 32, no. 3, pp. 914–938, 2016.

[2] N. Ahmad, Y. Ghadi, M. Adnan, and M. Ali, "Load forecasting techniques for power system: Research challenges and survey," *IEEE Access*, vol. 10, pp. 71 054–71 090, 2022.

[3] M. Jacob, C. Neves, and D. Vukadinović Greetham, *Forecasting and assessing risk of individual electricity peaks*. Springer Nature, 2020.

[4] I. K. Nti, M. Teimeh, O. Nyarko-Boateng, and A. F. Adekoya, "Electricity load forecasting: a systematic review," *Journal of Electrical Systems and Information Technology*, vol. 7, no. 1, pp. 1–19, 2020.

[5] M. Tzelepi and A. Tefas, "Efficient training of lightweight neural networks using online self-acquired knowledge distillation," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[6] ——, "Deep convolutional learning for content based image retrieval," *Neurocomputing*, vol. 275, pp. 2467–2478, 2018.

[7] A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Forecasting stock prices from the limit order book using convolutional neural networks," in *Proceedings of the 2017 IEEE 19th conference on business informatics (CBI)*, vol. 1, 2017, pp. 7–12.

[8] K. Amarasinghe, D. L. Marino, and M. Manic, "Deep neural networks for energy load forecasting," in *Proceedings of the IEEE 26th International Symposium on Industrial Electronics (ISIE)*, 2017, pp. 1483–1488.

[9] W. He, "Load forecasting via deep neural networks," *Procedia Computer Science*, vol. 122, pp. 308–314, 2017.

[10] N. Passalis and A. Tefas, "Global adaptive input normalization for short-term electric load forecasting," in *Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020, pp. 1–8.

[11] Y. He, J. Deng, and H. Li, "Short-term power load forecasting with deep belief network and copula models," in *2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, vol. 1, 2017, pp. 191–194.

[12] S. F. Ahmed *et al.*, "Short-term electrical load demand forecasting based on lstm and rnn deep neural networks," *Mathematical Problems in Engineering*, vol. 2022.

[13] I. P. Panapakidis, T. Perifanis, and A. S. Dagoumas, "The effect of dimensionality reduction methods in short-term load forecasting performance," in *Proceedings of the 2018 15th International Conference on the European Energy Market (EEM)*, 2018, pp. 1–5.

[14] G. Sideratos, A. Ikonomopoulos, and N. D. Hatziargyriou, "A novel fuzzy-based ensemble model for load forecasting using hybrid deep neural networks," *Electric Power Systems Research*, vol. 178, p. 106025, 2020.

[15] N. Andriopoulos, A. Magklaras, A. Birbas, A. Papalexopoulos, C. Valouxis, S. Daskalaki, M. Birbas, E. Housos, and G. P. Papaioannou, "Short term electric load forecasting based on data transformation and statistical machine learning," *Applied Sciences*, vol. 11, no. 1, p. 158, 2021.

[16] N. Maragkos, M. Tzelepi, N. Passalis, A. Adamakos, and A. Tefas, "Electric load demand forecasting on greek energy market using lightweight neural networks," in *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE, 2022, pp. 1–5.

[17] M. Tzelepi and A. Tefas, "Forecasting day-ahead electric load demand on greek energy market," in *Thirteen IEEE International Conference on Information, Intelligence, Systems and Applications (IISA),*. IEEE, 2022.

[18] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[19] M. Tzelepi, N. Passalis, and A. Tefas, "Online subclass knowledge distillation," *Expert Systems with Applications*, vol. 181, p. 115132, 2021.

[20] ——, "Probabilistic online self-distillation," *Neurocomputing*, vol. 493, pp. 592–604, 2022.

[21] B. Pan, Y. Yang, H. Li, Z. Zhao, Y. Zhuang, D. Cai, and X. He, "Macnet: Transferring knowledge from machine comprehension to sequence-to-sequence models," in *Advances in Neural Information Processing Systems*, 2018, pp. 6092–6102.

[22] Z. Li, J. Ye, M. Song, Y. Huang, and Z. Pan, "Online knowledge distillation for efficient pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 740–11 750.

[23] H. Chen, "Knowledge distillation with error-correcting transfer learning for wind power prediction," *arXiv preprint arXiv:2204.00649*, 2022.

[24] G. Panagiotatos, N. Passalis, A. Tsantekidis, and A. Tefas, "Sentiment-aware distillation for bitcoin trend forecasting under partial observability," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3898–3902.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] C. Li, M. Zhao, Y. Liu, and F. Xu, "Air temperature forecasting using traditional and deep learning algorithms," in *2020 7th International Conference on Information Science and Control Engineering (ICISCE)*. IEEE, 2020, pp. 189–194.