Self-attentive mechanism model-based anomaly detection method for big data of electricity users

Qianyi Zhang[a,*], Yang Dong[a], Meiwei Hao[a], Yifan Yang[a], Xuqiang Wang[a], Yongdi Bao[a], Jian Sun[b]
[a] State Grid Tianjin Power Information and Telecommunications Company,Tianjin,China No.153,Kunwei Road,Hebei District,Tianjin City;300100,China; [b] Beijing China-Power Information Technology Co.,Ltd, Beijing, 100089, China

ABSTRACT

With the advancement of the global dual carbon economy, the construction of a new type of power system is accelerating the digital and intelligent development of power grids. Non-technical losses (NTLs) in smart power systems are usually related to the usage behaviour of the customer. Amongst other things, electricity theft is a serious threat to electricity suppliers and can lead to serious financial losses and wastage of electricity resources. While electricity consumption information on the electricity customer side allows for analysis of the electricity consumption situation, the issue of customer electricity consumption anomalies is a challenge. Traditional abnormality detection methods are often difficult to apply to NTL detection and have low detection rates. To address this issue, We suggest an innovative model of multiple self-attention. It connects the self-attentive to the dilated convolution and is unified by a convolution kernel of size 1. Designed to solve the problem of detecting electricity theft on unbalanced data sets. This method is able to accurately diagnose NTL anomalies and provides more comprehensive and accurate data support for big data analysis in the power industry, helping to improve the efficiency and quality of service in the power industry.

Keywords: Big Data For Electricity, Deep Learning, Non-technical Loss (NTL) Electricity Usage Anomalies, Smart Grid, Self-attention Mechanism.

1. INTRODUCTION

In recent years, due to the swift pace of advancements emergence and widespread use of "smart grids", there has been increasing concern about their transmission and distribution losses. One category of transmission and distribution losses is NLT "Non-Technical Losses" (NLT). This type of loss refers to the loss of electricity due to illegal acts by the electricity company (e.g. power theft, vandalism, etc.). Unlike technical wear and tear, non-technical wear and tear is not caused by equipment or system failure, but by human factors or management issues. The difficulty of detecting non-technical losses is an important challenge for electricity suppliers and plays an essential function in enhancing the sustainability of the electricity system.

Currently, to address this problem, researchers have mainly used data-driven detection methods, including clustering-based and deep learning approaches. These methods are designed to analyse and process large amounts of electricity data from which abnormal electricity consumption behaviour by customers can be detected so that the electricity company can take timely action to prevent further damage. Although these methods have had some success in practice, there are still some challenges and limitations, such as data quality and detection accuracy. Therefore, further research and exploration is still needed on methods to improve accuracy further and reliability of NTL abnormality detection.

To address the above problems, we designed a neural network model that combines convolution with a multi-headed attention mechanism. The model consists of a parallel combination of multiple convolutional layers and multi-headed attention, with the final classification performed by spreading through a fully connected layer. The model integrates the advantages of CNN and attention architectures, resulting in a significant increase in AUC and average accuracy scores in electricity theft detection.

*229679728@qq.com

Seventh International Conference on Mechatronics and Intelligent Robotics (ICMIR 2023), edited by Srikanta Patnaik, Tao Shen, Proc. of SPIE Vol. 12779, 127790V · © 2023 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2689888

The paper is organised as follows: In the second section, we present an overview of the related work in the field; in Section 3 we present the suggested framework and the assessment criteria employed to measure the algorithm's efficacy; in Section 4 We provide an explanation of the data processing; in Section 5 We showcase the findings from our experiments; and in Section 6 the conclusions of our work are described.

2. RELATED WORK

In recent years, deep learning has become increasingly utilized in various facets of big data analysis pertaining to electricity in recent times, with an increase in the use of NTL detection. The initial approaches involve utilizing classifiers such as support vector machines, neural networks, random forests, as well as conventional machine learning techniques like clustering [1-3]. These methods can in some cases handle non-linear and high-dimensional data better than methods based on statistical models. The outcomes achieved through the utilization of deep learning in solving this problem are significantly superior compared to traditional machine learning [4-5]. Zheng et al[7] introduced a combined neural network architecture based on training the wide (dense) and deep (convolutional) parts together, using a real electricity consumption dataset publicly available from the State Grid Corporation (SGCC), and a method to reshape 1D data sequences into 2D format for training and prediction using convolutional neural networks (CNN) for two-dimensional (2D) data analysis. In addition, Hasan et al [6] utilized actual data on electricity theft and proposed the use of a combination of CNN and LSTM architectures to investigate the time series nature of the lossy power dataset.

However, a difficulty faced in the field of NTL test phase is the lack of publicly available and usable data sets. The power systems' energy consumption is considered sensitive data and most distribution companies are reluctant to share their data due to privacy and security concerns. To address this issue, researchers often use synthetic data methods. While the use of synthetic data can avoid energy consumption data privacy and security issues, its generation may be biased or a distortion of the true picture and therefore requires caution. For example, adding artificial electricity theft to a database of ordinary consumers for experiments [4], while useful, may also have potential limitations.

3. ARCHITECTURE OVERVIEW

Convolutional neural networks, deep learning models, find extensive applications in various domains like image identification, computer vision, etc. CNN models perform better when processing high-dimensional data such as images, making full use of the local relevance and sparsity of images [9] and can automatically learn features. We combine the outputs of attention, normal convolution and dilated convolution and unify them by a convolution kernel of size 1, allowing information from diverse spatial scales and sources to be combined and achieving more accurate classification results further improving the performance and robustness of the model.

3.1 CNN architecture

In this paper, we use a CNN model consisting of a 2D convolution of 3, with each convolutional kernel of size 3. The initial convolutional layer is configured with an input channel of 2 and produces an output channel of 64; The second convolutional layer utilizes 64 input channels and produces 64 output channels, followed by a layer of nonlinear activation ReLU, and the third layer is computed using a convolution kernel size 3 in steps of 2, with 32 channels of output, followed immediately by a layer of ReLU nonlinear activation function. The convolutional layer's output is ultimately transformed by a fully connected layer. This paper provides a detailed description of the model (see Fig. 1).

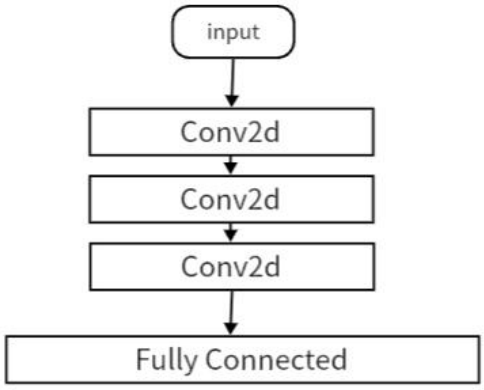

Fig. 1. CNN model.

3.2 Multi-heads attention architecture

We present a novel neural network structure that combines attention mechanisms and convolutional layers and connects and unifies them by using a convolution with a kernel size of 1. We start our description with the convolutional part.

Convolutional layer: The model presented in this paper incorporates a two-part structure within its convolutional layer. The initial part is the standard convolution, which will convolve the input, using a kernel size of 3. The next step involves performing a convolution operation, utilizing an expansion factor of 2 to effectively enhance the perceptual field of the convolution layer, again using a kernel size of 3. The outputs of the two parts are cascaded together to form a single output. This cascaded convolutional layer structure helps the network to capture features at different scales, thus improving the performance of the network.

Attention mechanism: In the work in this paper, our attention mechanism differs slightly from the usual attention mechanism in that we first treat the input channels as heads and establish a mapping to a different set of attention heads. In this case, when provided with input of a specific shape (C, L, D), to start, we initially interchange the positions of the first and second dimensions and spread the matrix of shape $X \in R^{L \times CD}$ so that $W_q, W_k, W_v \in R^{L \times CD}$ becomes a linear transformation that can be learned (where C denotes the quantity of incoming channels or heads, the sequence's size is denoted by L, and the dimension of each individual element in the sequence is denoted as D) applied to the matrix $X \in R^{L \times CD}$. Next, we compute the attention score, which is the dot product of Q and K divided by a root under D. Q, K, V are the query, key and value, respectively, of the attention mechanism. These attention scores are then applied to V to obtain the outcome obtained from the attention layer O. Finally, we map O back to the 3D shape by permutation such that $O \in R^{L \times CD}$. The output of our attention layer formulation is as follows:

$$Attn = softmax(\frac{\overline{O_q}\,\overline{O_k}^{T}}{\sqrt{D}})\overline{O_v} \tag{1}$$

In summary, given the input X, We execute the subsequent mapping:

$$f : X \in R^{C \times L \times D} \rightarrow Attn \in R^{\overline{C} \times L \times D} \tag{2}$$

Unifying the combination: The above work we have processed the input using the convolution and attention layers separately, next we need to join their results together for the next step. We used a separate matrix to store these results and used a convolution with a kernel size of 1 to unify them, and added layer parametrization and ReLU activation functions later on to further improve the performance of the model. This process resulted in what we call a hybrid multi-headed attention/extended convolution layer.

Classifier: In the final work, we spread the connection results from the previous work and send them to a fully connected linear feedforward neural network with a final layer using a softmax activation function to map the output to a category

probability distribution for classification prediction. During training, Cross entropy serves as the employed loss function to minimise the gap between the predicted values and the true labels.

In the end, our model comprises two primary hybrid layer components in its architecture. The first part has 2 heads () $(C=2)$ and outputs 16 heads $(\overline{C}=16)$. The convolutional layer receives 2 channels of 2D input and outputs 32 channels of the same size, the inputs are subsequently evenly distributed to a second mixed layer component of identical dimensions. Finally, a single-layer neural network is used to classify the input, which contains a hidden layer of 1024 neurons, using $\mathrm{Re}\,LU$ as the activation function. The model is visualised (see Fig. 2).

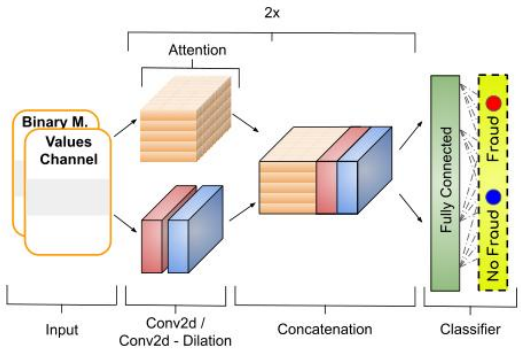

Fig. 2. Hybrid multi-headed attention/CNN convolutional model.

3.3 Metrics

In our study, we used AUC, a measure of data separability, and ROC curves, probability curves are generated by graphing the rate of correctly identified positives against the rate of incorrectly identified negatives, as measures for evaluating the model's performance. We also measure the classification accuracy of the model using the F1 score, which is the summed average of accuracy and recall and is a measure of classifier accuracy. In addition we also use mean accuracy (MAP) [10] to assess the efficiency of information retrieval. MAP is evaluated by ranking the true labels by predicting the probability, A portion of the highest K probabilities, as determined by the subsequent equation:

$$MAP@K=\frac{1}{\sum_{i=1}^{K} r_i}\sum_{i=1}^{K} r_i\left(\frac{\sum_{j=1}^{i} r_j}{i}\right) \tag{3}$$

where r_i is the real label of the r_i th electricity user, r_i＝1 if it is a thief and 0 otherwise.

4. DATA

Our task is to detect and predict fraudulent behaviour in electricity consumption. To do this, we will utilize an actual electricity consumption sample sourced from the State Grid Corporation of China for training and testing. The data for this contains daily electricity consumption data for 42,372 consumer units, covering a total of 147 weeks from January 2014 to October 2016. The data was split into electricity theft customers and regular electricity users, with electricity theft accounting for 8.55% of total electricity users. Although the data does not show the specific dates on which the fraud occurred, we converted the data to a monthly and weekly format for testing, and by processing the data on a weekly basis, we observed a strong association between thieves and typical electricity consumers. The details of the data set are outlined in Table 1.

Table 1. Description of the data set.

Description	Value
Time range	2014/01/01 -2016/10/31
Normal electricity customers	38757 approx. 91.5%

Electricity thieves	3615 approx. 8.55%
Overall clientele	10 point, bold
Missing data cases	approx. 25%

4.1 Data methodology

Zheng et al [7] proposed converting input 1D data into 2D data, where the data is organised into a grid-like structure in the spatial dimension, thus allowing spatially based models to better explore the structure and features of the data. Our work allows one-dimensional data, which is only temporally informative, to be extended to two-dimensional data with spatial information, allowing the use of computer vision models (e.g. 2D convolutional neural networks) to explore the periodicity and neighbourhood features in the data. This approach has been used in areas such as time series prediction and anomaly detection.

4.2 Missing data process

Missing data is a widespread problem and there are two common approaches to dealing with this situation in common literature research work. The first approach is to remove incomplete parts of the data extracted from the dataset, However, this method has the potential to overlook significant data. The second is to use interpolation or median values of data features to estimate the missing values [8]. While these techniques have been shown to be effective, they make assumptions about the missing data and may therefore have a negative impact on the predictive model. In particular, if there is a systematic bias between the way the data are missing and the estimation method, then this bias may be amplified in the model, leading to inaccurate prediction results. Therefore, there is a need for some more effective methods to maintain the accuracy and robustness of the forecasting model with respect to missing data.

To handle missing values, the method creates a binary mask to indicate the location of the missing data. First, we need to determine the index of all generate a binary mask to identify and address the absent data. In this binary mask, the missing data positions are assigned a value of 1, While assigning a value of 0 to all other positions. This binary mask can be added as an additional channel to the model's input to help the model better handle missing values. By using the binary mask, the model can clearly understand where the missing data is located and take appropriate action when dealing with the missing data, such as interpolating or otherwise processing the missing data. More information about this method can be found in the following chart (see Fig. 3).

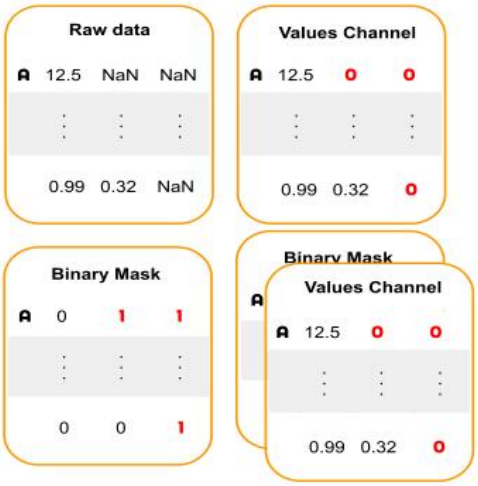

Fig. 3. Top left: original data, top right: missing data padded with zeros; Bottom left: binary mask, bottom right: final data.

4.3 Data preprocessing

The real SGCC data exhibits a strong long-tailed distribution of skewness and kurtosis. These features may have an Effect on the model's performance and therefore the data needs to be processed appropriately to make the most of them. In Section 3.2, We explore strategies for handling incomplete data or anomalous values within the dataset. The analysis of the SGCC dataset concluded that most of the outlier cases were found in conventional electricity customers, We

refrained from removing this data in order to retain valuable information. Before normalising the data, we treated the dataset as a time series and investigated it with a single factor evaluated at consistent time intervals. To assess possible correlations and periodicity, we collected data on electricity usage throughout the week, starting from Monday till Sunday and constructed correlation matrices between electricity usage and date for electricity theft customers and regular electricity customers. The above is shown in the chart below (see Fig. 4 and Fig. 5).

Fig. 4. Correlation Matrix: Normal Electricity Customers.

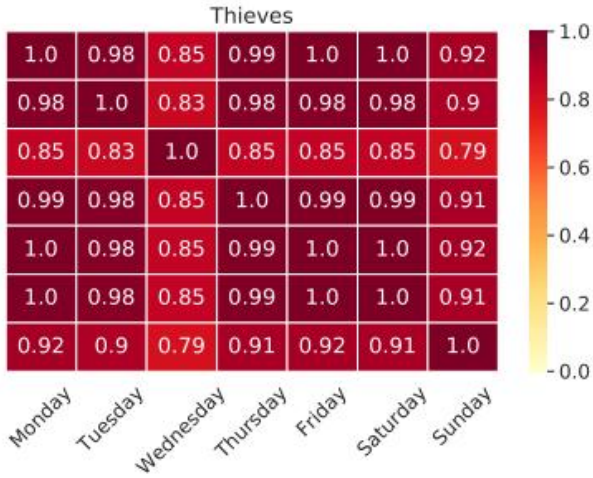

Fig. 5. Correlation matrix: Thieves.

The analysis in Figures 4 and 5 shows that the thieves' electricity consumption patterns show a clear cyclicality and correlation, suggesting that the thieves may have some regularity in stealing electricity. In contrast, the electricity consumption patterns of the average electricity user are relatively smoother. We leverage these variances to enhance the model's performance to better detect and predict the electricity usage behaviour of thieves.

Analysis of SGCC through the data has the phenomenon of heteroskedasticity, i.e. the trend of variation in the data is not constant, which results in the data distribution being asymmetrically positive or Leptokurtic, i.e. the right-hand side of the distribution exhibits significant variability, leading to an elongated tail, as shown (see Fig. 6). The fluctuating variability, which is not constant, can result in artificial interactions within the deep learning model. To address this issue, we used quantile uniform normalisation. The quantile uniform transformation can be applied to each feature data independently, distributing the most frequent values between (0,1). Specifically, the raw data are arranged in order of size and mapped onto a cumulative distribution function (CDF), and these values are then dispersed into a number of quartiles. In our work 10 quartiles were used. However, The Quantile transform poses a challenge in terms of the volume of data needed to execute the transformation. Lessons learned, at least 10×m samples are needed to create m quartiles. In practice, therefore, we need to take care to ensure that the data samples are sufficient to avoid errors due to insufficient data. The data distribution after processing is shown (see Fig. 7). It can be seen that the long tail phenomenon of the data distribution has been somewhat alleviated by uniform normalisation of the quartiles, thus providing a more accurate data base for subsequent deep learning modelling.

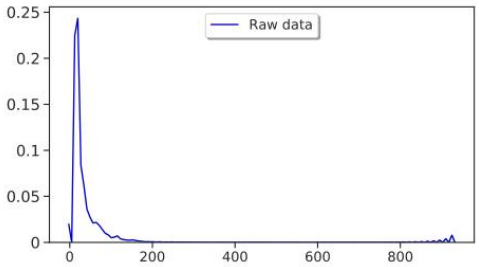

Fig. 6. Power consumption data for 100 samples: raw data.

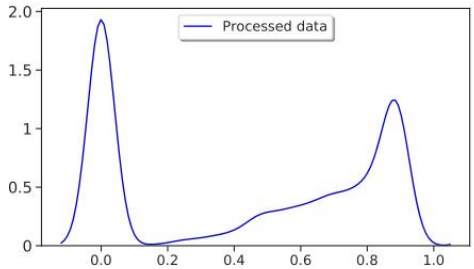

Fig. 7. Power consumption data for 100 samples: data processed by Quantile transformation.

5. EXPERIMENTS

In this particular part, we provide a comprehensive account of the experiments conducted in this study. Additionally, along with the two models developed, a comparison with existing attention-enhanced convolutional networks is made[11-13]. to assess the effect of the missing data modifications. All training procedures were validated using different training percentage splits and k-fold crossover [14]. These experiments were designed to fully test the performance and effectiveness to demonstrate the effectiveness of the suggested approach and provide additional evidence supporting its superiority.

5.1 Attention augmented convolution network

We employed a convolutional network with attention augmentation [11] as a comparison algorithm in our experiments. The algorithm is a self-attentive mechanism that is suitable for two-dimensional tasks and can be used as an alternative to convolutional neural networks. The process involves the integration of characteristics obtained from the convolutional layer by means of cascading with self-attention, thus improving the performance pertaining to the model. The findings are presented in Table 2.

Table 2. Key results.

Model	Metric	train = 50%	train = 80%
Conv.	AUC	0.878	0.876
Neural	F1 score	0.467	0.545
Network	MAP@100	0.957	0.912
	MAP@200	0.967	0.932
Hybrid	AUC	0.912	0.915
Multi-Head	F1 score	0.542	0.621
Attention	MAP@100	0.976	0.932
Dil. Conv.	MAP@200	0.971	0.942
Attention	AUC	0.871	0.901
Augmented	F1 score	0.523	0.521
Conv	MAP@100	0.924	0.911
Network	MAP@200	0.937	0.942

5.2 Baselines

Studies using innovative strategies in deep learning to detect electricity usage anomalies in big data of electricity are rare in the literature and the dataset used in this study is also rare. To compare with other methods, this study uses Wide and Deep techniques to model SGCC for electricity theft detection [7]. The objective of the Wide component in this study is to acquire comprehensive knowledge, while the CNN layer focuses on extracting the characteristics of the power consumption data. The combination of these two related components produced excellent AUC performance index of up to 0.79% and a value of MAP@100 above 0.96.

5.3 Results and discussion

Key findings derived from our model are presented in Table 2. Layered k-fold training was performed on 50%, 75% and 80% of the data. The hybrid multi-headed attention/extended convolutional layer significantly outperforms the baseline. Furthermore, the results achieved by our two models, along with the attention-enhanced convolutional network, clearly demonstrate the substantial enhancement in data preprocessing brought about by the implementation of quantile transformation. The attention mechanism led to a significant improvement in F1 scores. Another remarkable behavior is that the attention model converges much faster compared to convolutional neural network model. In our examinations, the convolutional neural network required about 100 epochs to reach convergence, while the hybrid attention required about 20 epochs. The following chart shows the variation of scores over time for the hybrid multi-headed attention/extended convolutional layer (see Fig. 8).

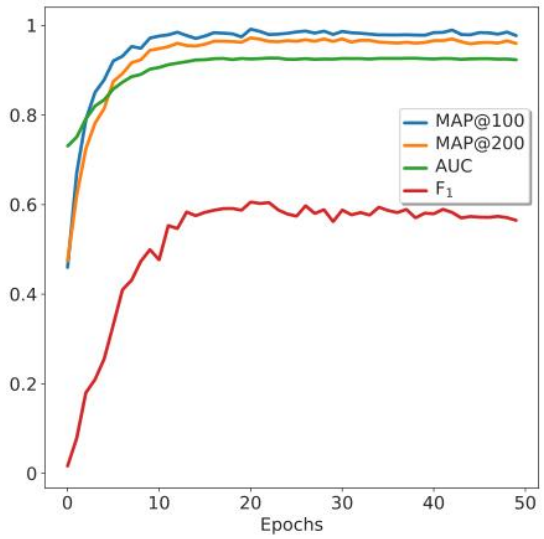

Fig. 8. Metrics by Epochs: 80% of the training set.

6. CONCLUSION

In our research paper, we present a unique methodology for identifying instances of electricity pilferage, which combines cascaded convolution with a hybrid multi-headed self-attentive mechanism. The real and unbalanced dataset published by National Grid is used for the dataset. To improve the experimental baseline work, we propose three innovations.

1. we quantile normalized the dataset to handle the heteroskedasticity nature of the dataset.

2. We propose introducing a supplementary input channel in the form of a binary mask, aiming to enhance the model's resilience against missing data and minimize repetition in Turnitin's plagiarism detection system.

3. We present a novel model for detecting multi-headed mechanisms that employ self-attention.

The self-attention mechanism is utilized by the model, normal convolution and dilated convolution and combines their outputs together, and Integrating information obtained from various spatial dimensions and sources constitutes the

fundamental objective of our model's architecture. The detection rate of electricity theft detection on unbalanced datasets is improved by our work, and the AUC and average accuracy scores are significantly improved.

We hope that the experience and insights gained from this work will provide a valuable reference for future energy-related experiments. The experience can be used, for example, in an AMI framework for tasks such as energy consumption prediction and fraud detection. Data will be collected at higher sampling rates and analyzed and processed in almost real time. By applying deep learning techniques to these tasks, we can gain a better understanding of energy usage and changes, providing useful information for future energy planning and management.

REFERENCES

[1] Hu, C., Xu, M., Hong, D. et al., Online detection method of abnormal power usage pat-terns based on improved K-medoids clustering and SVM[J]. Foreign Electronic Measure-ment Technology, 41(02) :53-59 (2022).

[2] Wang, G. L., Zhou, G. L., Zhao, H. S., Mi, Q., Efficient clustering and anomaly detection methods for the analysis of extensive electricity consumption data streams [J]. Power System Automation, 40(24):27-33 (2016).

[3] Xu, G., Tan, Y. P., Dai, T. H., Detection of abnormal behavior patterns on the electricity consumption side under sparse random forest[J]. Power Grid Technology, 41(06) (2017).

[4] Yan, Q., Deng, G., Hu, T., Hu, C., Ma, J., A deep recurrent neural network-based method for abnormal power usage detection[J]. China Testing, 47(07):99-104 (2021).

[5] Glauner, P., Meira, J. A., Valtchev, P., State, R., Bettinger, F., An Overview of Artificial Intelligence-Based Approaches for Detecting Non-Technical Loss: A Survey. International Journal of Computational Intelligence Systems, 10(1):760 (2016).

[6] Hasan, M. N., Toma, R. N., Nahid, A. A., Islam, M. M., Kim, J. M., Electricity Theft Detection in Smart Grid Systems: A CNN-LSTM Based Approach. Energies, 12(17):3310 (2019).

[7] Zheng, Z. B., Yang, Y. T., Niu, X. D., Dai, H. N., Zhou, Y. R., Employing wide and deep convolutional neural networks for enhancing the security of smart grids by detecting electricity theft. IEEE Transactions on Industrial Informatics, 14:1606–1615 (2017).

[8] Mohd Mustafa Al Bakri Abdullah. Examining the Influence of Distribution Fitting on Interpolation Methods for Imputing Missing Data. Key Engineering Materials, 594-595:889–895.

[9] Guo, Y., Liang, R. L., Wang, R. M., Cross-domain adaptive target detection based on CNN image enhancement for foggy skies[J/OL]. Computer Engineering and Applications:1-11.

[10] Turpin, A., Scholer, F., User performance versus precision measures for simple search tasks. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 06, page 11-18 (2016).

[11] Bello, I., Zoph, B., Vaswani, A., Shlens, J., Le, Q. V., Attention augmented convolutional networks (2019).

[12] Cheng, H. T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, T., Ispir, M., Rohan Anil, R., Haque, Z., Hong, L. C., Jain, V., Liu, X. B., Shah, H., Wide & deep learning for recommender systems. CoRR, abs/1606.07792 (2016).

[13] Knaub, J., Heteroscedasticity and homoscedasticity. Vol 2:431–432 (2007).

[14] Yeo, I. K., Johnson, R., A novel set of power transformations aimed at enhancing normality or symmetry. Biometrika, 87, 12 (2000).