Deep learning for combating energy theft:
A hybrid long-short term memory approach

Authors
Hong-Xin Gao, MSc (Royal Holloway, 2021)
Stefanie Kuenzel, Power System Group, Department of Electronic Engineering, Royal Holloway
Xiao-Yu Zhang, Power System Group, Department of Electronic Engineering, Royal Holloway

Abstract

Energy theft is a critical issue for power system operators. Billions of pounds are lost every year due to this. In a conventional power grid, energy theft is difficult to detect due to limited communication and data transition. Smart meters, along with big data mining technology, lead to significant technological innovation in the field of energy theft detection. This article describes a *convolutional long short-term memory* (ConvLSTM) based energy theft detection model to identify electricity thieves. The novel ConvLSTM network is a variant of the long short-term memory (LSTM) model, originally used for spatial-time-series prediction. The goal of our work is to verify that it can also handle classification issues effectively. Experiments show improved efficiency of model deployment in realistic scenarios while reducing training time. The final experimental results also show that the proposed ConvLSTM model exhibits good robustness. Compared to the multilayer perceptron (MLP) and the convolutional neural network & long short-term memory (CNN-LSTM), it has a more established performance metric and generalization capability. [a]

[a] This article is published online by Computer Weekly as part of the 2022 Royal Holloway information security thesis series `https://www.computerweekly.com/ehandbook/Royal-Holloway-Deep-learning-for-countering-energy-theft`. It is based on an MSc dissertation written as part of the MSc in Information Security at the ISG, Royal Holloway, University of London. The full thesis is published on the ISG's website at `https://www.royalholloway.ac.uk/research-and-teaching/departments-and-schools/information-security/research/explore-our-research/isg-technical-reports/`.

What is energy theft?

We may have experienced the theft of our mobile phones or wallets, but we will probably rarely be aware of energy theft. This is because energy theft does not usually result in immediate property damage to the individual using the energy, but it does always undermine the legal rights of the normal energy user. Energy theft is considered a crime in many countries, and it usually refers to the intentional avoidance of paying legitimate bills by energy users. Energy theft is also highly concealed and fraudulent, for instance, some individual electricity consumers or businesses falsify electricity records and defraud electricity suppliers through intentional tampering and illegal wiring. Energy theft not only disrupts a fair market for electricity consumption, but also causes significant financial losses to utility companies. As a result, electricity theft is a criminal offence in many countries. According to a Northeast Group LLC report in 2017, the power supply sector worldwide loses about US$96 billion a year due to non-technical loss (NTL), which include electricity theft and fraud. For example, malicious bitcoin mining is a currently popular form of electricity theft. At present, electricity theft is one of the major causes of NTL in the grid. One of the reasons for the rapid growth of advanced metering infrastructure (AMI) over the past few years is to reduce the NTL caused by electricity theft. With the rapid development of technology, the methods of electricity theft have diversified from the traditional malicious tampering with physical metering equipment to today's remote computer intrusions. Figure 1 shows examples of attacker models of energy thefts.

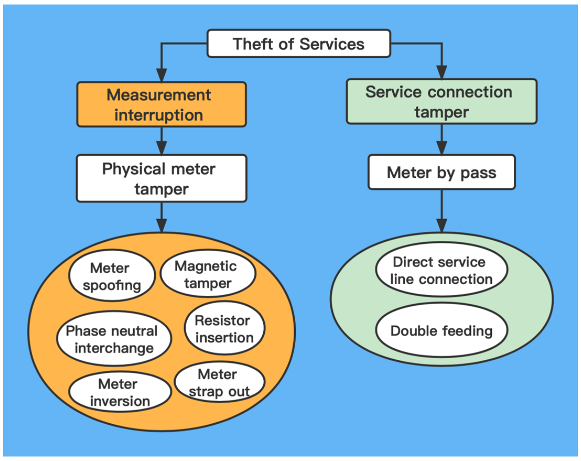


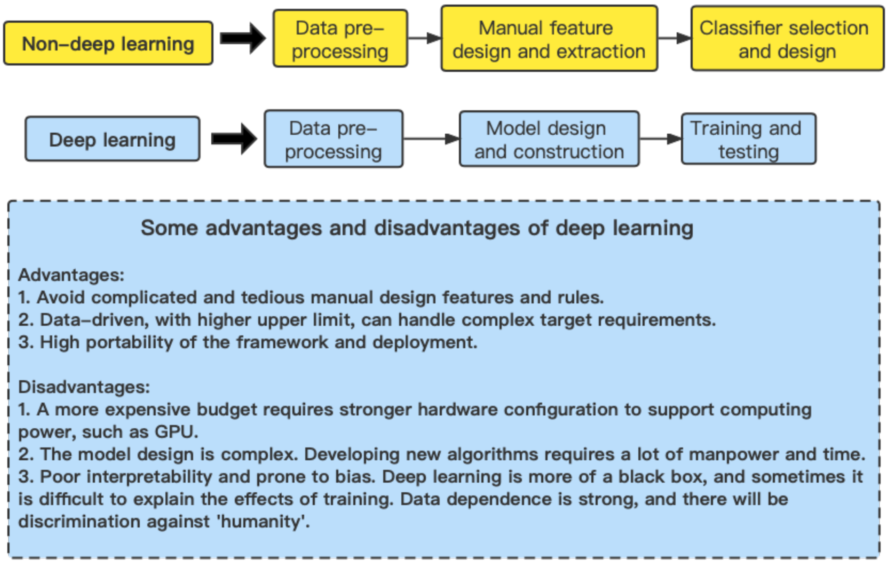
Figure 1: Different attack modes of energy theft - adapted from *Networked Energy Services. (2020, June 17). Energy Theft and Fraud Reduction. N&S.*

Why deep learning in energy theft detection?

Due to the complexity of power theft methods, it is difficult to achieve effectiveness with manual detection alone. Machine learning-based energy theft detection (ETD) models are also being deployed in smart grids and can be used to assist in the automatic detection and identification of electricity thieves. In general, machine learning uses algorithms and knowledge involving interdisciplinary methods such as statistics, probability theory, computing and mathematics, to analyse and process data. New data are then classified or predicted by computing and learning the features and patterns inherent in the data. In simple terms, the goal of machine learning is to enable models to observe, think, compute, and identify, as humans do, thereby assisting and replacing manual methods. Predictive modelling is an important concept in machine learning, with the benefit of allowing for the development of models that make the most accurate predictions, rather than focusing on explaining why a prediction was made.

Research in the field of machine learning in ETD can be based on both classical machine learning (non-deep learning) and deep learning. Research issues can be divided into classification recognition and regression prediction. In the niche area, non-technical loss (NTL)-based ETD approaches can be divided into three categories, data-oriented, network-oriented and hybrid models that combine both. Network-oriented approaches are usually based on localised automated meter reading and sensors, for instance, based on the concept of smart substations and network voltage sensitivity to identify theft states. This approach generally suffers from performance-based issues such as high-cost equipment and personnel training, with some inflexibility in future deployments and paradigm shifts. The data-oriented approach is based on machine learning techniques and is conducted through supervised and unsupervised learning. This approach allows models to be built from substantial data collection, reducing the development and operational costs of complex power equipment components. The hybrid model can show better performance through the combination of these two approaches. For example, the combination of power observation devices and machine learning algorithms can effectively reduce false positives. However, the balance between performance and cost is an issue that needs to be considered in the future.

The development of the Internet of Things (IoT) has also enabled smart grids to benefit from the era of big data, which also allows data-oriented approaches to be used to their advantage. Non-deep learning usually involves a lot of manual feature extraction and is inefficient when dealing with complex issues. As a result, deep learning is used to deal with more complex scenarios. Deep learning can automatically extract features and optimise their *weights* by building models - the weight is the parameter that is varied to adjust the expected computational outcome, and the training process is essentially an update of the weights. It requires more data to support better generalization of the model, which also allows it to uncover potential patterns and trends between electricity consumption data.


Non–deep learning
Data pre–processing
Manual feature design and extraction
Classifier selection and design
Deep learning
Data pre–processing
Model design and construction
Training and testing

Some advantages and disadvantages of deep learning

Advantages:
1. Avoid complicated and tedious manual design features and rules.
2. Data–driven, with higher upper limit, can handle complex target requirements.
3. High portability of the framework and deployment.

Disadvantages:
1. A more expensive budget requires stronger hardware configuration to support computing power, such as GPU.
2. The model design is complex. Developing new algorithms requires a lot of manpower and time.
3. Poor interpretability and prone to bias. Deep learning is more of a black box, and sometimes it is difficult to explain the effects of training. Data dependence is strong, and there will be discrimination against 'humanity'.

Figure 2: Some advantages and disadvantages of deep learning.

Figure 2 illustrates some of the differences, advantages, and disadvantages of non-deep versus deep learning. Current ETD models in deep learning can be based on different branches such as multilayer perceptron (MLP), convolutional neural network (CNN), and long short-term memory (LSTM). MLP is a prototype of a classical neural network and is the most fundamental model for deep learning. CNN was first proposed for use in the field of computer vision (for example, image recognition), and LSTM can be applied to natural language processing (for example, text translation).

ConvLSTM with batch normalization

We briefly introduce the ConvLSTM model used in our project, which is a variant of the LSTM and a modification of the currently popular CNN-LSTM.

A convolutional neural network (CNN) consists of three main parts, the convolutional layer, the pooling layer, and the fully connected layer. The convolutional layer is responsible for extracting local features in 2-dimensional (2D) electricity data. Individual user data in the raw dataset is a 1-dimensional time series containing daily electricity consumption data. This can be divided by time (e.g., quarterly division) into 2D matrix grids, each 2D grid containing the number of points of daily electricity data × the number of days in each quarter. CNN is commonly used in computer vision, such as image recognition, and the electricity data observations in a 2D grid can be considered as pixel points in an image. The pooling layer allows for more efficient dimensionality reduction which reduces the number of operations and also effectively avoids overfitting. The fully connected layer consists of neurons in a typical neural network and connects the nodes in the front and back layers. It can be thought of as a classification layer where the data processed by the convolutional and pooling layers is fed and the expected results are obtained by means of a specific activation function.

A long short-term memory (LSTM) network is a special recurrent neural network that memorizes and propagates key information from the initial to the final stages of the network. Its time-recursive structural features can be applied to forecasting time-series input tasks, e.g., stock price prediction and text translation. Electricity consumption data is temporally sequential, and LSTM captures long-term dependencies between data and learns the correlation of electricity consumption across different time segments.

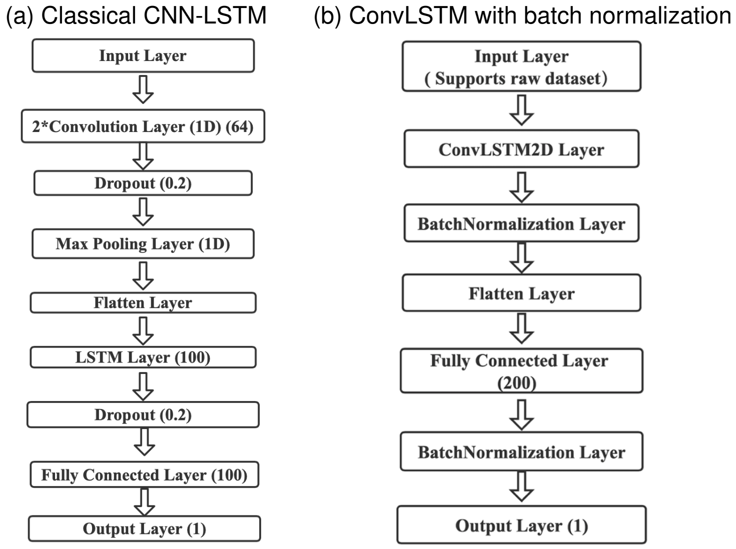


Figure 3: CNN-LSTM and ConvLSTM

Figure 3(a) shows the layers in a classical CNN-LSTM stack.

Unlike the structure of CNN-LSTM stack, in a convolutional long short-term memory (ConvLSTM) model, an LSTM replaces the pooling layer and allows for discovery of deeper relationships between time-series data. ConvLSTM uses convolutional computation in the fully connected layer, and the parameters learned in this way can be used to capture the underlying spatial features. Electricity consumption data with time-series properties can be fitted with sequential input in ConvLSTM. ConvLSTM has been proposed to solve regression issues using its temporal memory properties, and it is mainly used for forecasting with multidimensional time-series properties and spatial expansion.

Figure 3(b) shows the structure of a ConvLSTM model, which is clearer and more concise than a classical CNN-LSTM after embedding batch normalization. Batch normalization is a data transformation that is needed before model training. Here it is directly embedded into the model, which enables the proposed ETD model to support raw format electricity. Other advantages are as follows:

1. It can accelerate convergence and improve model accuracy.
2. It can be used to solve the exploding gradient problem (which can lead to unstable networks).
3. It eliminates a separate transformation link and improvev training efficiency.

Potential advantages of ConvLSTM over CNN-LSTM are:

1. Simpler architecture with fewer layers required.
2. Instead of CNN and LSTM detecting spatial and temporal patterns respectively, Spatio-temporal patterns can be detected simultaneously
3. LSTM can replace the pooling layer, reducing local information loss and facilitating the capture of long-term dependencies.

The project

The goal of this project is to build a novel ETD model based on ConvLSTM and to verify its accuracy and robustness in identifying electricity thieves, and to show that it outperforms the MLP and CNN-LSTM models.

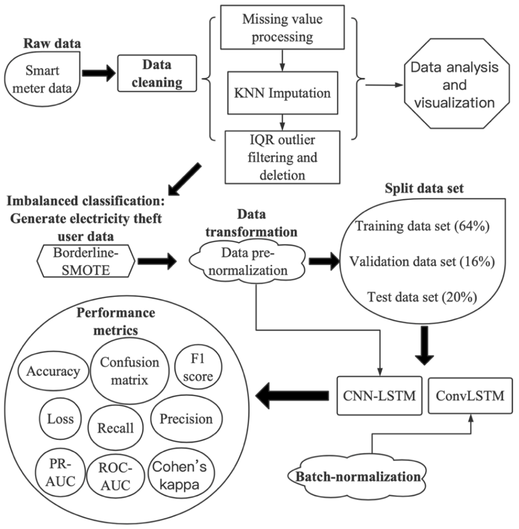


Figure 4: Overview flow

The overall system

The dataset contains two categories: normal users and electricity thieves. Essentially this is dealing with a binary classification problem by means of supervised learning. The key processes are shown in Figure 4.

1. **Data cleaning**, including removing and filling missing values with a k-nearest neighbors (KNN) imputation technique and filtering outliers with interquartile range (IQR).
2. **Visualization** of the dataset by, for example, power curves and Pearson correlation coefficients, to initially check for potential trends and correlations between the data.
3. **Generate** more realistic data on thieves by borderline-SMOTE (which is a technique for dealing with imbalanced sample sizes).
4. **Data transformation**. MLP and CNN-LSTM do normalization before the data is fed into the model, and ConvLSTM embeds batch normalization directly in the model.
5. **Splitting** the training (64%), validation (16%) and test (20%) datasets for each model. Maintain the best model by optimizing feature extraction, hyperparameters, functions etc.
6. **Compare and analyse** the differences in robustness and model convergence efficiency between the ConvLSTM and baseline models using a more comprehensive performance metric.

The setup

In this project, all experiments are based on Python (Version 3.7.6) programming, in which the deep learning framework is based on TensorFlow (Version 2.4.0). The hardware platform is a laptop computer, processor configuration is 2.6 GHz 6-core Intel Core I7, graphics configuration is AMD Radeon Pro 5300M 4 GB. Meanwhile, with the support of Free cloud GPU, 30GB RAM, 8 CPUs.

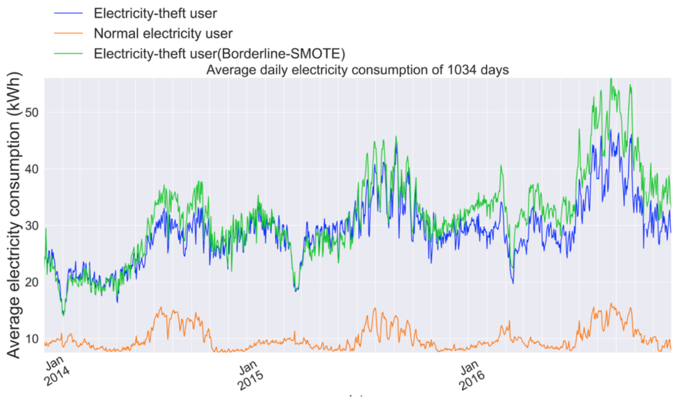


Figure 5: Comparison of electricity thieves produced by Borderline-SMOTE

The raw dataset

The dataset selected for this paper was obtained from real electricity consumption data published by the State Grid Corporation of China (SGCC)[1]. The dataset is shown in Table 1 and contains the daily electricity consumption in kilowatt-hour (kWh) of 42,372 customers between 1 January 2014 and 31 October 2016 (1034 days).

Table 1: Raw data status

SGCC data set		
Description	Quantity	Class tag
Normal users	38757	0
Electricity thieves	3615	1
Total	42372	/

Total Number of Data	Amount of Missing Values	Amount of Zero Values
43812648	11233528	5788603

Imbalanced classification sorting

An imbalanced sample size between the two classes in the dataset can lead to a lack of objective judgement and erroneous bias in the model, for example, the model can be sensitive to minority classes. This can be addressed uisng a Synthetic Minority Oversampling Technique (SMOTE). Here there is a possibility of overlap between the minority and majority classes in the raw dataset or statistical observations of electricity data and SMOTE may confuse the two classes, resulting in inaccurate classification data being produced. However, the variant technique Borderline-SMOTE will classify observations in this minority class as noise points when the data adjacent to the minority class are all in the majority class and ignore them when generating the data. It is equivalent to creating boundaries in the vicinity of some outliers, which is more conducive to the accuracy of the generated data. Figure 5 shows the comparison between the thieves generated by borderline-SMOTE and the original data, with the trend matching the real theft user status.

[1] Zheng, Z., Yang, Y., Niu, X., Dai, H. N., & Zhou, Y. (2017). Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. IEEE Transactions on Industrial Informatics, 14(4), 1606-1615.

Hyperparameter

To ensure the reliability and authenticity of the evaluation, the project maximized the consistency between the baseline and the proposed model in terms of both the activation function and the main hyperparameter optimization. More specifically, to further validate the superiority of ConvLSTM ontology in mining time-series depth features, we intentionally kept the CNN-LSTM training parameters the same as the ConvLSTM. We refer the interested reader to the full dissertation for the details of the optimizer and other algorithms used in the experiment.

Result

For model evaluation, we used a more comprehensive performance metric on the test dataset (20%). This is shown in Table 2:

Table 2: Metrics comparison

	Accuracy	Loss	Precision	Recall	F1-Score	Cohen's kappa	ROC-AUC	PR-AUC
MLP	0.964	0.162	0.969	0.960	0.964	0.927	0.985	0.983
CNN-LSTM	0.977	0.131	0.979	0.976	0.978	0.955	0.991	0.992
ConvLSTM	0.984	0.089	0.984	0.985	0.984	0.969	0.993	0.991

All three proposed models performed well in terms of performance metrics. However, ConvLSTM performed best in the full range of metrics. A brief explanation of the metrics is as follows:

The *loss* of a model is the opposite of *accuracy*, with smaller values representing better model performance. *Precision* and *recall* represent the classification accuracy of electricity thieves and actual electricity thieves, respectively. *F1-score* captures the trends in precision and recall. *Cohen's kappa* can be used as a means of judging the strength of the model's classification predictions - the kappa value is more of a measure comparing the observed accuracy with the expected accuracy, and the larger the value the better the performance of the model.

ROC-AUC is the area under the receiver operating characteristic (ROC) curve. The ROC curve visualizes True Positive (electricity thieves are correctly classified) and False Positive (normal users are incorrectly predicted as electricity thieves) in a trade-off manner, with the area under the curve (AUC) visually reflecting the classification capacity expressed by ROC curve. Like the ROC-AUC, *Precision-recall (PR)-AUC* has a score between 0 and 1, with the higher the value, the larger the area of the PR curve. Both are more often applied in the form of threshold intrinsic trade-offs for the perfect evaluation of models.

To better demonstrate the above conclusions, it is necessary to further analyse the convergence process of the model. Convergence refers to a dynamic tendency for a model to reach an optimal state, a point where accuracy and loss functions can reach an optimum. A high convergence efficiency also means that the model has fewer iterations and is more robust. Figure 6 gives a visual analysis of the models' performance in the training and validation datasets..

Both CNN-LSTM and ConvLSTM can reach the optimal model state before 90 epochs, while MLP needs around 150 epochs to reach it. In addition, MLP and CNN-LSTM reach smooth convergence around 60 epochs and 40 epochs, respectively, while ConvLSTM reaches the model convergence state well around 20 epochs. It is worth noting that ConvLSTM performs better than the other two models in terms of noise control throughout the curve fluctuation state, which is also due to the application of batch normalization in the model. Combined with the above analysis, ConvLSTM outperforms MLP and CNN-LSTM in terms of model convergence efficiency and model generalization ability, showing the robustness of the model and prediction.

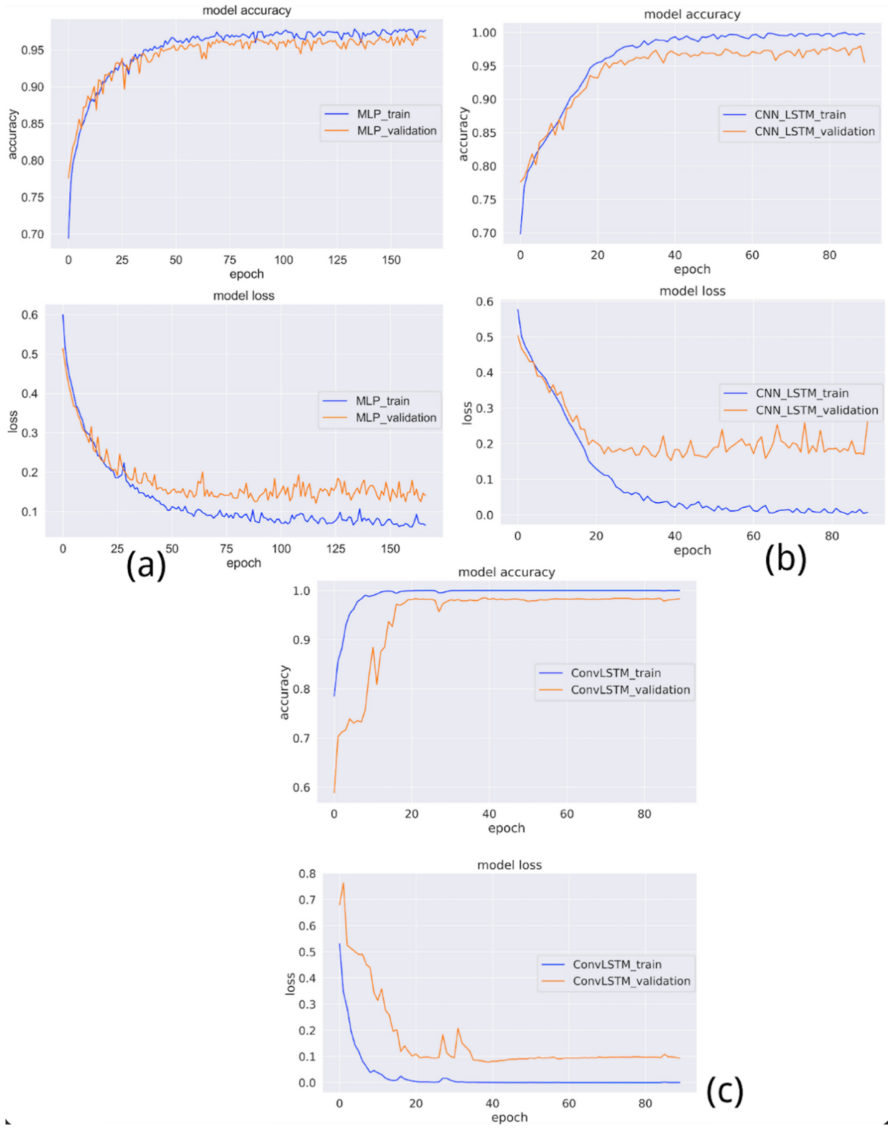


Figure 6: History of the models. (a) MLP (b) CNN-LSTM (c) ConvLSTM

Conclusion

In this project, three ETD models based on deep learning were successfully constructed, and all showed good results in identifying electricity thieves. The proposed ConvLSTM with the addition of batch normalization not only outperforms baseline models in terms of model structure and performance metrics but also outperforms in terms of model robustness and convergence efficiency. The extension of ConvLSTM to multidimensional electricity consumption data is the next step we will consider in our work. We can add multiple features to the dataset, such as weather and geographic location, to better match the input of multi-dimensional tensor. This is also a way for the ETD model to incorporate objective factors to detect potential electricity thieves.

Biographies

Hong-Xin Gao received a B.B.A. degree and a B. Eng. degree from the Shandong Agricultural University, China in 2011, and an M.Sc. with distinction in Information Security from Royal Holloway, University of London, in 2021. He worked for a state-owned electric power enterprise in China from November 2011 to May 2019. During this time, he worked in production, bidding, and project management positions, mainly responsible for project operations for power smart terminals. His current research interests include deep learning techniques for anomaly detection in smart grid, analytics and processing in big data.

Stefanie Kuenzel received an M.Eng. and a Ph.D. degree from Imperial College, London in 2010 and 2014, respectively. She is currently the Head of the Power Systems Group and a Senior Lecturer with the Department of Electronic Engineering, Royal Holloway, University of London. Her current research interests include renewable generation and transmission, including HVDC as well as Smart Meters.

Xiao-Yu Zhang received a B. Eng. degree from the North China Electric Power University, Beijing, China in 2016 and an M.Sc. with distinction in Electrical Power System from University of Birmingham, UK, in 2017. He is currently pursuing a Ph.D. degree in Electrical Engineering from the Royal Holloway, University of London. His research interests include deep learning technology & data analytics in smart grids, smart grid privacy and security, and demand-side management.

Series editor: S.- L. Ng