A Hybrid ConvLSTM-based Anomaly Detection Approach for Combating Energy Theft

Hong-Xin Gao, *Student Member, IEEE,* Stefanie Kuenzel, *Senior Member, IEEE,*
and Xiao-Yu Zhang, *Member, IEEE*

***Abstract*—In a conventional power grid, energy theft is difficult to detect due to limited communication and data transition. The smart meter along with big data mining technology leads to significant technological innovation in the field of energy theft detection. This paper proposes a convolutional long short-term memory (ConvLSTM) based energy theft detection (ETD) model to identify electricity theft users. In this work, electricity consumption data is reshaped quarterly into a two-dimensional matrix and used as the sequential input to the ConvLSTM. The convolutional neural network (CNN) embedded into the long short-term memory (LSTM) can better learn the features of the data on different quarters, months, weeks, and days. Besides, the proposed model incorporates batch normalization. This technique allows the proposed ETD model to support raw format electricity consumption data input, reducing training time and increasing the efficiency of model deployment. The result of the case study shows that the proposed ConvLSTM model exhibits good robustness. It outperforms the multilayer perceptron (MLP) and CNN-LSTM in terms of performance metrics and model generalization capability. Moreover, the result also demonstrates that K-fold cross-validation can improve the ETD prediction accuracy.**

***Index Terms*—energy theft, smart grid, binary classification, ConvLSTM, deep learning.**

NOMENCLATURE

AMI	Advanced metering infrastructure
AMR	Automated meter reading
AUC	Area under the curve
CNN	Convolutional neural network
ConvLSTM	Convolutional long short-term memory
ETD	Energy theft detection
IQR	Inter-quartile range
KNN	K-nearest neighbors
LSTM	Long short-term memory
MLP	Multilayer perceptron
NTL	Non-technical losses
PCC	Pearson correlation coefficients
PR	Precision-recall
ReLU	Rectified linear unit
RNN	Recurrent neural network
ROC	Receiver operating characteristic
SMOTE	Synthetic minority oversampling technique
SVM	Support vector machine

I. INTRODUCTION

THE smart grid has made great progress as a mainstream trend in the current development of electricity networks. It effectively combines the electricity consumption of grid service users with intelligent communication and monitoring, enabling an evolution from automated meter reading (AMR) to advanced metering infrastructure (AMI) [1]. AMI is a critical part of the smart grid layout, integrating intelligent measurement, collection, storage, and energy data analysis [2]. It also marks a shift towards intelligent and digital communication between utility companies and electricity consumers. As the core equipment of the smart grid, smart meters not only provide precise and synchronised measurement and collection from end-users and provide efficient data guarantee for AMI intelligent analysis [3]. As the information and communication modules in smart meters continue to be integrated and iterated, energy theft through physical approach is becoming more advanced and covert. For instance, attacks go beyond traditional meter tampering by exploiting system vulnerabilities to manipulate meter readings and execute cyber-attacks [4]. Energy theft is a serious social hazard, which can be illustrated as illegal electricity customers using utility's energy in breach of contract or manipulating their meter reading to avoid paying the bill [4]. This problem causes huge financial losses to utility companies, seriously infringes on the legal rights of normal electricity users and disrupts the fair market environment for electricity consumption. Electricity theft has become one of the main causes of non-technical losses (NTL) in smart grids. According to the recent report, the global electricity supply sector loses approximately US$25 billion annually to nontechnical losses, including theft, fraud, etc. For example, in India, annual losses due to electricity theft are approximately US$4.5 billion [5]. The 2020s will be a critical period for the development of smart grids globally, with North America, Western Europe, and other countries already on the path to building smart grids. However, this is still an area of massive investment for emerging markets. The report states that the 50 emerging markets will invest over US$40.7 billion in the coming years [6]. This is certainly a signal to drive the global AMI layout's development and implies a national commitment in terms of smart grid NTL governance.

This work, as part of the PrivIoT project, has been supported by the PETRAS National Centre of Excellence for IoT Systems Cybersecurity, which has been funded by the UK EPSRC under grant number EP/S035362/1.

H. X. Gao is with the Department of Information Security, Royal Holloway, University of London, Egham Hill, Egham TW20 0EX, U.K. (e-mail: hongxin.gao.2020@live.rhul.ac.uk).

S. Kuenzel is with the Department of Electronic Engineering, Royal Holloway, University of London, Egham Hill, Egham TW20 0EX, U.K. (e-mail: Stefanie.Kuenzel@rhul.ac.uk)

X. Y. Zhang is with School of Artificial Intelligence, Anhui University, Hefei 230601, Anhui, China; and he was with the Department of Electronic Engineering, Royal Holloway, University of London, Egham Hill, Egham TW20 0EX, U.K. (e-mail: xiaoyu.zhang@rhul.ac.uk)

NTL-based approaches to ETD can be divided into three categories [7], data-oriented, network-oriented and hybrid models. Network-oriented approaches are usually based on localised AMR, and sensor ideas, e.g., [8], [9] propose smart substations and network voltage sensitivity concepts to identify theft states. Performance is based on high-cost equipment and staff training, among other aspects, and there is a certain inflexibility in future deployment and modality transformation. The data-oriented approach is based on machine learning techniques and proceeds through supervised and unsupervised learning [7]. Jamaica is one of the first countries to use machine learning to combat energy theft. Its public service company identified electricity theft customers around 2017 by deploying an AMI-based machine learning model [10]. Machine learning allows for feature engineering and modelling electricity consumption data, which also implies optimisation of manual detection methods. Currently, classical machine learning and deep learning are two data-oriented approaches to electricity theft detection. In classical machine learning for binary classification issues, for instance, support vector machine (SVM) and extreme gradient boosting (XGBoost) classifiers for ETD are supervised learning methods [11], [12]. They are based on labelling electricity consumption samples to identify theft customers using electricity consumption data features. Classical decision trees and random forests also show good identification accuracy in ETD [13].

With the development of AMI and advances in neural networks, deep learning-based ETD methods has also been introduced to combat energy theft. The advantage of AMI is that it provides a large amount of data support and a comprehensive approach to energy monitoring which caters to the characteristics of deep learning that requires substantial data support to prevent over-fitting models with optimal generalisation [14]. Neural networks as a concept in deep learning have been widely used in computer vision, speech recognition and anomaly detection [15]. Neural networks are robust to noise in the input data and mapping functions and can even support learning and prediction in the presence of missing values. Also, neural networks do not make strong assumptions about the mapping function and can easily learn linear and non-linear relationships [16]. In addition to this, deep learning allows for automatic extraction in terms of feature engineering compared to classical machine learning. It is oriented toward data and demand pattern concepts, and cross-domain techniques can be well applied to ETD problems [14]. For example, convolutional neural network (CNN) is currently the mainstream neural network for processing image classification and computer vision [17], which can be good for automatic feature extraction and global optimisation. Long short-term memory (LSTM), as a variant of recurrent neural network (RNN), controls the transmission state by gating the state and can capture the relationship between time series effectively [18]. More importantly, LSTM solves the problem of gradient disappearance and gradient explosion problems during training long sequences [19]. In [13], CNN and LSTM are analysed separately for comparison and outperformed classical machine learning techniques. In [20], [21], [22], hybrid deep learning models showed better feature extraction and model structure expansion in ETD. LSTM has successfully solved a number of power system isssues, such as load forecasting [23], energy dis-aggregation [24], etc.

[22] built an ETD model in a classical CNN-LSTM stack. However, the CNN used for feature extraction in the front segment is restricted to one-dimensional data as input. [20], [21] are based on the CNN-LSTM with improvements such as expanding and deepening the convolution layers and data augmentation to optimise feature extraction and support 2D electricity consumption data input. However, this method is limited by the architecture of the underlying model, the CNN layer feature extraction is not fully embedded in the whole and can only be fed into the LSTM after feature extraction. Front-end input limits the scaling of multi-dimensional data and can affect the model to extract deeper and more subtle anomalous features. In terms of model optimisation, [20], [21], [22] all incorporate a dropout layer, which can effectively prevent the over-fitting of the model. Rectified linear unit (ReLU) is used as an activation function, which only activates positive values. In the backpropagation process, each unit calculates its weight based on the loss values emanating from the upper layer [25]. The optimiser Adam is also applied in [20], which adaptively adjusts the learning rate.

Regarding imbalanced data, the classical synthetic minority oversampling technique (SMOTE) is used in [20]. The traditional SMOTE runs the risk of overlapping samples from a few classes and thus over-fitting the model. In the development of deep learning, the novel ConvLSTM is proposed to be applied to predict spatio-temporal sequences for regression issues [26]. Compared to CNN-LSTM, it can optimise the excessive redundancy in temporal data. Importantly ConvLSTM uses LSTM instead of pooling layers in CNN to reduce the loss of detailed local information and capture long-term dependencies in sequences. In human behaviour recognition, ConvLSTM has better recognition accuracy and false alarm rates when dealing with multiple classification issues [27].

In this paper, our main contribution is adopting the ConvLSTM architecture on the proposed ETD model. It supports fully connected layers for convolutional computation, replacing the matrix multiplication of traditional CNN-LSTM stacked attributes. This allows CNN feature extraction to be embedded throughout the model, allowing better extraction of local electricity usage features. On the other hand, it also facilitates the LSTM further to capture the deeper periodicity of the electricity consumption data. In addition, a batch-normalisation technique is added to the proposed model. It supports raw electricity consumption data input, eliminating the need for tedious and time-consuming data pre-transformation. We use an improved dataset balancing method, borderline-SMOTE, which generates more realistic data on electricity theft users for imbalanced datasets. Furthermore, it can reduce the overlapping of data and prevent over-fitting of the model.

The remaining sections of the paper are presented below. Section II presents the overall approach and introduces the proposed ConvLSTM model incorporating the data transformation component by comparing it with CNN-LSTM. Section III presents the key steps throughout the experiment. Section

IV assesses and compares the baseline and proposed models through a more comprehensive set of metrics. In Section V, we conclude with the results of the proposed model for the continuation of future work.

II. Methodology

A. Overall system

The approach of this paper is to build a novel ETD model based on ConvLSTM and verify its accuracy and robustness in identifying electricity theft users by simultaneous comparison and full validation with MLP and CNN-LSTM. The dataset contains two categories of normal users and electricity theft users and essentially deals with the binary classification issue by employing supervised learning. The key processes are shown in Fig. 1. The whole process is divided into three main steps, which are data pre-processing; model training; and model evaluation. The pre-processing step contains four sub-steps: 1) Data cleaning, including removing and filling missing values with k-nearest neighbours (KNN) imputation and filtering outliers with inter-quartile range (IQR). 2) Visualise the dataset by power curves and Pearson correlation coefficients (PCC) to initially check for potential trends and correlations between the data. 3) Train-test-splitting. The dataset is split into 64% training, 16% validation and 20% test data. 4) Solving imbalanced classification issues in training and validation datasets. The borderline-SMOTE sampling technique generates more realistic theft user data (for training and validation datasets only). The second step contains five sub-steps: 1) Data transformation and reshaping. The batch-normalisation technique is embedded in ConvLSTM and CNN-LSTM models without additional transformation steps. The raw electricity consumption data can be used directly in training and testing. 2) The optimal model is trained and selected using 10-fold, 5-fold cross-validation and no cross-validation. 3) Tuning the hyperparameters to find the optimal model parameters and save the model which performs the best with training data. In the final step evaluation, performance metrics are employed to evaluate the proposed ConvLSTM ETD model.

B. Data Transformation and Reshaping

This paper uses a novel batch-normalisation technique for all three models. This technique makes the training of deep neural networks more efficient. As stated by Ioffe and Szegedy, the batch-normalisation method can accelerate deep network training by reducing internal co-variate shift [28]. It can be embedded directly behind each input layer in the normalisation model without needing separate data transformation. In practice, this saves the time and number of training sessions consumed by changes in the fed data. More importantly, batch normalisation has little effect on the initial weights and model hyperparameter changes while reducing generalisation errors [28]. The main equations are as follows, with γ and β being the initial training parameters and m and j being the sample and batch values, respectively.

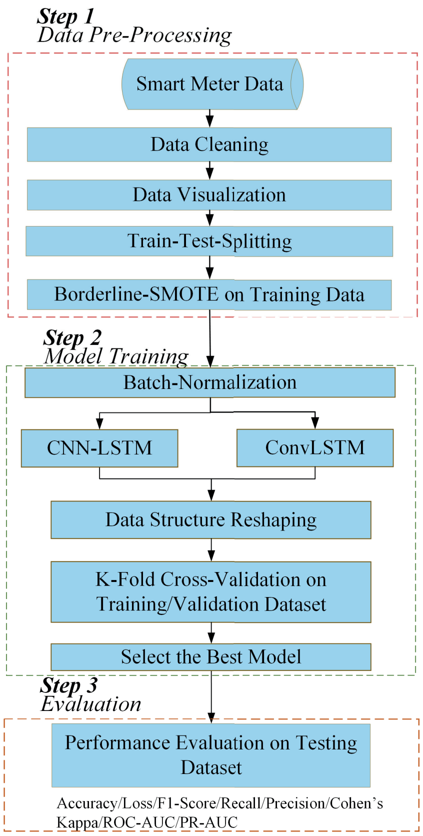


Fig. 1. Flowchart of the proposed ConvLSTM-based energy theft model with detailed data processing, model training, and evaluation process.

$$E_x = \frac{1}{m}\sum_{i=1}^{j}\mu\frac{(i)}{B} \tag{1}$$

$$\mathrm{Var}_x = \left(\frac{m}{m-1}\right)\frac{1}{m}\sum_{i=1}^{j}\sigma\frac{2(i)}{B} \tag{2}$$

$$y = \frac{\gamma}{\sqrt{Var_x+\epsilon}}x + \left(\beta + \frac{\gamma E_x}{\sqrt{Var_x+\epsilon}}\right) \tag{3}$$

Prior to feeding the one-dimensional electricity consumption data into the model, the data requires reshaping to match the model input shape. In the CNN-LSTM model, a two-dimensional matrix dividing the electricity consumption data into $3\times3\times1$ by quarter is used, which is essentially a stacked structure of CNN and LSTM supporting three-dimensional tensor inputs, in this paper i.e., user quarterly electricity consumption data, number of quarters and number of samples.

To ensure consistency in model evaluation, the proposed model reshapes the electricity consumption data into three dimensions, i.e., samples, time steps and features, in the same way as the CNN-LSTM model and according to quarters. However, the novel ConvLSTM 2D layer [29] supports 5D tensor inputs, i.e., samples, time, rows, columns, and channels, which can also be interpreted as the time step being decomposed into rows × columns of image data points. In this paper, Time is the number of quarters, Columns is the quarterly electricity consumption data, and the time step is divided into quarterly quantities × quarterly data, i.e., each sub-series contains a sequence of three months of electricity consumption data. Rows and Channels are one because each user's raw electricity consumption data is one-dimensional, and there are no additional raw features.

C. Stacking of Classical CNN and LSTM

The CNN consists of three main parts, the convolutional layer, the pooling layer, and the fully connected layer [15], [17]. In this paper, it can be elaborated as the convolutional layer is responsible for extracting local features in 2D electricity data, i.e., features are extracted in sliding window mode within each subsequence (2D matrix segmentation according to quarter time). The pooling layer allows for more efficient dimension reduction than the convolution layer, which reduces the number of operations and effectively avoids overfitting. In the classical CNN-LSTM structure, CNN layers can be encapsulated in Time-Distributed layers [30], and the extracted features are flattened for use in LSTM.

LSTM network is an RNN trained by time backpropagation [18]. It has a unique formulation that avoids the problem that other RNNs cannot be trained and scaled, while it also overcomes the problem of vanishing and exploding gradients. Truncated backpropagation through time (TBPTT) [31] is a key concept in the training LSTM model. Unlike neurons, the memory blocks of LSTM networks contain states and outputs and are connected in layers [19]. Each block has three gates: a forget gate, an input gate, and an output gate [31]. In this paper, the CNN-LSTM model can be used to control whether they are triggered or not by a sigmoid [25] activation function that produces an output between 0 and 1 in binary classification. The classical LSTM equations [18] can be derived as follows, the input gate is Eqn. (4), the forget gate is Eqns. (5) and (6), and the output gate and mainline generation output are Eqns. (7) and (8), respectively. Where the hidden state and cell state are h_t and C_t, respectively. f_t, i_t and o_t are the activation vectors of forget gate, input gate and output gate, respectively. σ denotes the sigmoid function, $\circ$ denotes the Hadamard product, and b is the bias vector parameter.

$$i_t = \sigma\left(W_{xi}xt + W_{hi}h_{t-1} + W_{ci} \circ C_{t-1} + b_i\right) \tag{4}$$

$$f_t = \sigma\left(W_{xf}xt + W_{hf}h_{t-1} + W_{cf} \circ C_{t-1} + b_f\right) \tag{5}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh\left(W_{xc}xt + W_{hc}h_{t-1} + b_c\right) \tag{6}$$

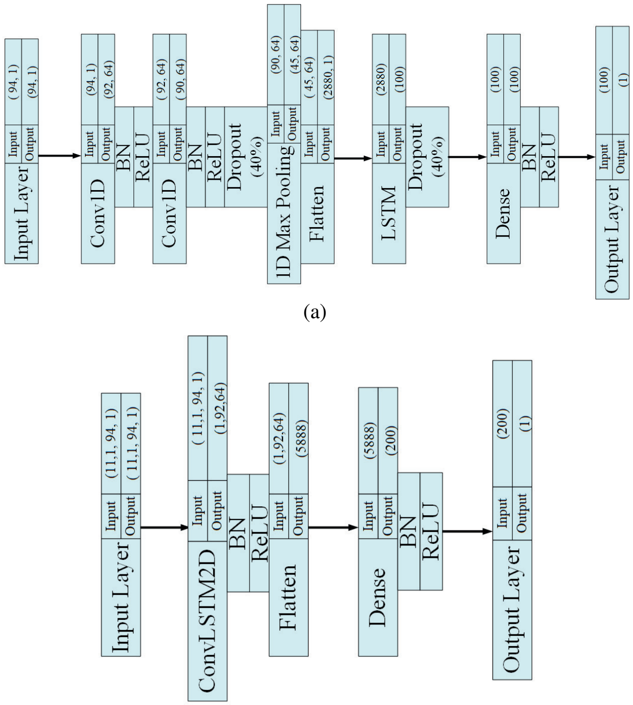
$$o_t = \sigma\left(W_{xo}xt + W_{ho}h_{t-1} + W_{co} \circ C_t + b_o\right) \tag{7}$$

$$h_t = o_t \circ \tanh\left(C_t\right) \tag{8}$$

A sliding window is a method to transform time series into supervised learning [17]. A maximum pooling layer is added after two consecutive CNN layers, and a 40% dropout layer is added after each CNN and LSTM. The dropout layer [32] uses adaptive regularisation to prevent the model's overfitting and ensure that the model is in an optimal state. A comparison of the architecture of CNN-LSTM and the proposed model is shown in Fig. 2.

D. Proposed ConvLSTM Model

Unlike the structure of the CNN-LSTM stack, the LSTM replaces the pooling layer in ConvLSTM and discovers deeper relationships between time-series data through the embedding structure. ConvLSTM uses convolution computation in the fully connected layer, which means that ConvLSTM replaces the matrix multiplication of each gate in the LSTM cell with convolution operations. In other words, the parameters learned are convolution kernel weights and can be used to capture the underlying spatial features by performing convolution operations in multi-dimensional data. 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 [26], and it is mainly used for forecasting with multi-dimensional time-series properties and spatial expansion. As shown in Fig. 3, the network structure of the ConvLSTM is a variant of the LSTM with feedforward features of input transformations and recursive transformations implemented by convolution [29].



(a)

(b)

Fig. 2. Architecture of (a) 1D CNN-LSTM ETD model; (b) ConvLSTM ETD model.

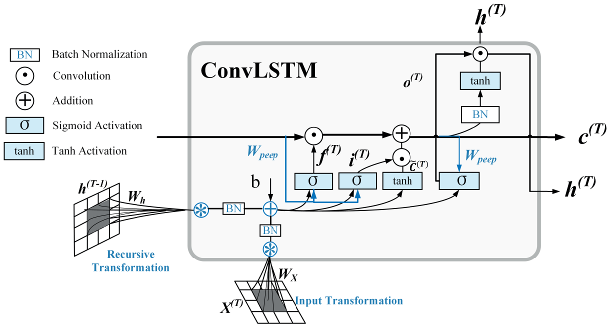


Fig. 3. Block-diagram of the ConvLSTM block with the operations inside the block.

In the proposed model, each user's electricity consumption data can be divided into 2D image $M \times N$ in terms of quarters, i.e., quarterly electricity consumption data × the number of quarters. Each image $X \in R^{P \times M \times N}$, which has one pixel P, and then the observed feature value can be as follows, with R being the domain of the observed data [26].

2D data can be seen as the spatial dimension of the image, and features are captured mainly by the convolutional layer. 3D has an extra-temporal dimension, and the temporal features are captured by the LSTM, as shown in Fig.4. While the internal structure of the ConvLSTM, C and X represents the unit output, H is the hidden state, and the example uses the structure convolution operation present and new values as shown in Fig. 5.

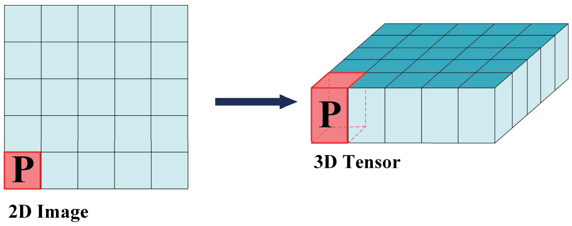


Fig. 4. Transformation of 2D image to 3D tensor.

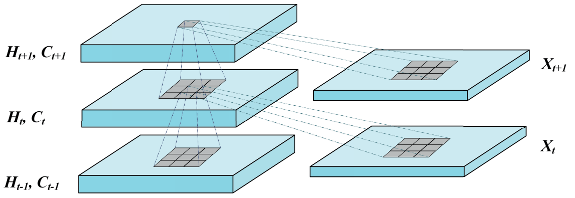


Fig. 5. The internal structure of ConvLSTM.

In contrast to CNN-LSTM, H and X in ConvLSTM use convolutional operations instead of parametric matrix multiplication, and the learned parameters are the weights of the convolution kernel. * denotes the convolutional operation, which can have multiple convolvtional filters. The main equations can be derived as follows [26]:

$$i_t = \sigma\left(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i\right) \tag{9}$$

$$f_t = \sigma\left(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f\right) \tag{10}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh\left(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c\right) \tag{11}$$

$$o_t = \sigma\left(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o\right) \tag{12}$$

$$H_t = o_t \circ \tanh\left(C_t\right) \tag{13}$$

Even without the multi-channel features, the proposed model can still dig deeper than the baseline model in the convolutional operation into the temporal features between the number of electricity consumption data. As shown in Fig. 2 (b), the structure of the proposed model becomes clearer and more concise after embedding batch-normalisation. A point worth noting is that the output still needs to be flattened to a long vector before the dense layer can be interpreted.

TABLE I
RAW DATA STATUS

SGCC dataset				
Description	**Quantity**	**Class Tag**	**Duration**	**Days**
Normal users	38757	0		
Theft users	3615	1	1/1/ 2014 to 31/10/2016	1034
Total users	42372	/		
Total Data	**Missing Values**		**Zero Values**	
43812648	11233528		5788603	

III. Experimental Setup

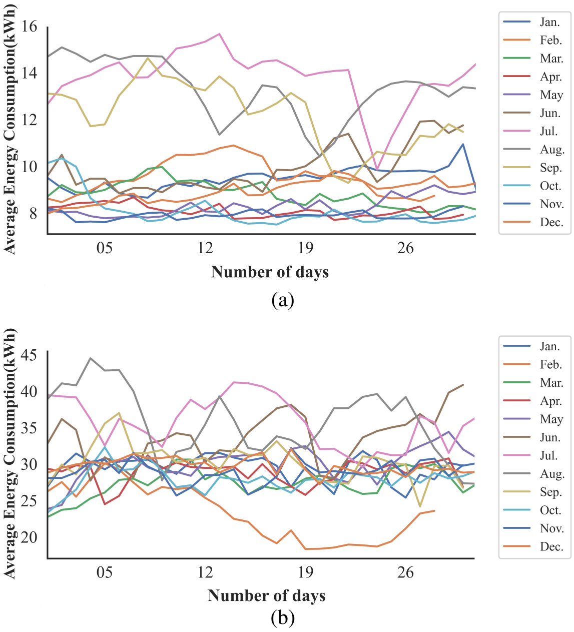
In this paper, all experiments are based on Python (Version 3.7.6) programming, in which the deep learning framework is based on TensorFlow (Version 2.4.0). The hardware platform is a laptop computer, the processor is 2.6 GHz 6-core Intel Core I7, and the graphics are AMD Radeon Pro 5300M 4 GB. Meanwhile,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

A. Data Description

1) **Preview of Raw Dataset**: The dataset selected for this paper is obtained from real electricity consumption data published by the State Grid Corporation of China (SGCC) [21]. The dataset contains the daily electricity consumption in kilowatt-hour (kWh) of 42,372 customers between 1 January 2014 and 31 October 2016 (1034 days). 38,757 of these customers are normal electricity users (labelled 0), and 3,615 are customers who have been identified as electricity thieves (labelled 1). An example of the dataset is presented in Table I.

2) **Electricity Consumption Data Visualization**: The anomalous manifestations of electricity theft are not only present on the surface of the data, but the underlying patterns and trends are equally characteristic. While machine learning can replace manual detection of potential electricity theft, we also need to communicate with the data, which is what data analysis is all about. As Fig. 6 shows, electricity usage data for normal customers tends to be more stable and less volatile in months other than summer, with July, August and September showing significantly stronger usage and fluctuations than other months. However, the data for electricity theft customers appears unusually chaotic, with a very large drop in December and an overall trend that does not conform to natural patterns.

In [21], data from four weeks can be extracted for further analysis, for example, plotting the data between the two types of users again and plotting PCC. These methods show correlations and potential regularities between the data in each of the two categories of users. The two categories of users can first be plotted again weekly, as shown in Fig.7. Normally, normal users show good cyclical and seasonal patterns, but electricity theft users continue to have mixed chaotic electricity usage characteristics. Thus, annual, quarterly, monthly, weekly, and daily electricity usage characteristics can be used as a benchmark for extracting features. Fig.8 clearly shows that the data correlation of normal users is much stronger than electricity theft users. The correlation coefficient for electricity theft customers does not exceed a maximum of 0.3 and has a negative correlation. However, the correlation coefficient

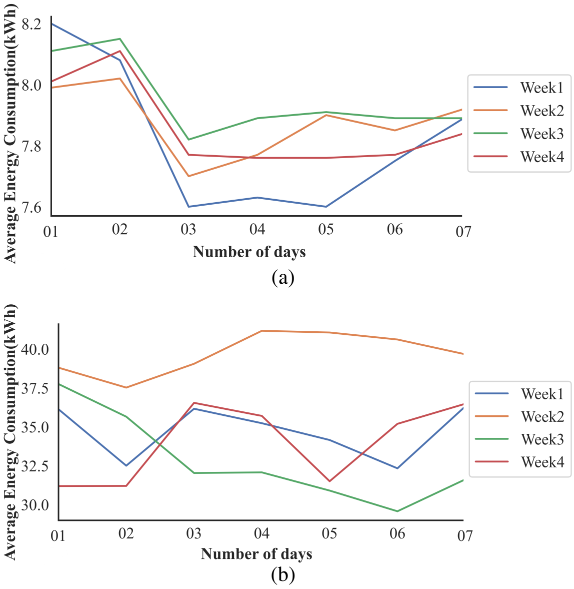


(a)

(b)

Fig. 6. Average monthly electricity consumption in 2015. (a) normal energy users (b) energy theft users.

for normal customers is generally higher than 0.8 and shows a strong positive correlation. In the PCC, values above 0.5 or below -0.5 represent a relatively significant correlation. Positive values closer to 1 indicate a stronger direct correlation. A negative value closer to -1 represents a strong indirect correlation [33].



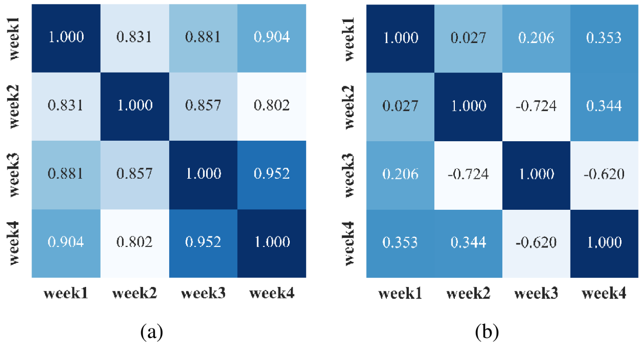
(a)

(b)

Fig. 7. Average daily electricity consumption every four weeks. (a) normal energy users (b) energy theft users.

B. Data Preparation

1) **Missing Data Filtering and Imputation**: This paper sets a rejection baseline of 3%, i.e., users with more than one month of missing data will be removed. The minimum threshold of missing data is also used for monthly data to retain certain characteristics of the raw data. The dataset used for this paper involves continuous missing data. Therefore the KNN imputation is a more efficient method. Its algorithm is based on similarity and relies on a distance metric, the default of which is the Euclidean distance metric. KNN imputation is effective for handling missing values in continuous and ordered data, and its imputation accuracy and reduction of statistical errors are typically better than 1NN (e.g., two neighbouring data) [34]. The main point is that the imputed values are the actual values that occur rather than the constructed values, which also allows for better preservation of the original data structure. In this paper, number of nearest neighbors k is selected as 5.



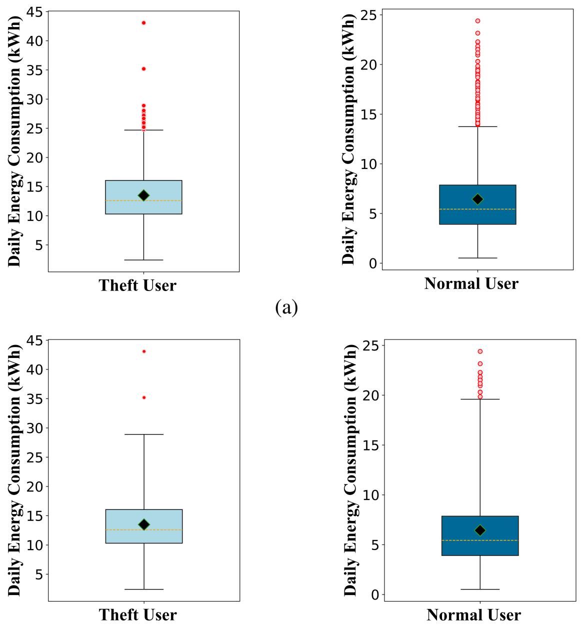
(a)

(b)

Fig. 8. Pearson's correlation coefficient (PCC). (a) normal energy users (b) energy theft users.

2) **Outlier Processing:** To maintain a true state of the electricity consumption data, we use Boxplots to screen for outliers. It uses the quartiles of the data to identify outliers among them. The boxplot shows the distribution of data based on a summary of five numbers (minimum, first quartile (Q1), median, third quartile (Q3), and maximum), and the maximum value is Q3 + 1.5×IQR and the minimum value Q1 - 1.5×IQR. However, due to the characteristic nature of electricity consumption data, the statistics can be conducted in such a way as to create 'false' outliers. Q3 + 1.5×IQR/Q1 -1.5×IQR can be defined as a minor or moderate outlier, and Q3 + 3×IQR/Q1 -3×IQR as an extreme outlier [35]. By screening the minor and extreme outliers for all users over 1034 days, a partial sample is shown in Fig.9.

3) **Imbalanced Classification Sorting**: Borderline-SMOTE is an improvement on SMOTE. For example, the overlap between the minority and majority classes in the raw dataset or statistical observations of electricity data is possible. SMOTE may confuse the two classes of data, resulting in inaccurate classification data. However, 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 [36]. It is equivalent to creating boundaries in the vicinity of some outliers, which is more conducive to the accuracy of the generated data. Fig.10 compares the theft users generated by borderline-SMOTE and the original data (for the training and validation datasets only), with the trend matching the real theft user's status.



(a)

(b)

Fig. 9. Sample comparison of outlier screening. (a) minor outliers over 1034 days (1.5×IQR) (b) extreme outliers over 1034 days (3×IQR).

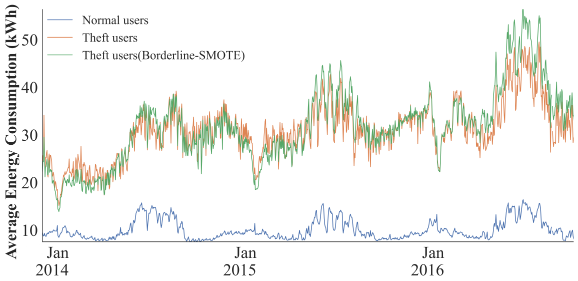


Fig. 10. Comparison of energy theft users generated by borderline- SMOTE.

C. Hyperparameter

To ensure the reliability and authenticity of the experimental evaluation, the activation functions and main hyperparameters of the three models are maximally kept consistent. More specifically, to further validate the superiority of the ConvLSTM ontology in mining time-series depth features, we intentionally align the training parameters of the CNN-LSTM with the ConvLSTM, as shown in Table II. In this paper, the output layer activation function is sigmoid [25], which is more stable for scaled data. Sigmoid is suitable for this paper's binary classification prediction output (0 represents the label of normal users, 1 represents the label of electricity theft users) because it exists between 0 and 1.

$$S(x) = \frac{1}{1+e^{-x}} \tag{14}$$

The other activation function is ReLU [25], which trains the model faster, ensures near-global weight optimisation, and increases the non-linearity of the network. It is more conducive to backpropagation and avoids gradient explosion or vanishing issues.

TABLE II
HYPERPARAMETER SELECTION

Hyperparameter	MLP	CNN-LSTM	Proposed ConvLSTM
Output layer activation	Sigmoid	Sigmoid	Sigmoid
Other activation	ReLU	ReLU	ReLU
Batch size	250	250	64
Epochs/Early stopping	400/30	400/30	400/30
Loss	Binary_crossentropy	Binary_crossentropy	Binary_crossentropy
Optimizer	Adam	Adam	Adam
Learning rate	0.001	0.001	0.001
Dropout	0.8	0.4	N/A

$$f(x_i) = \max(0, x_i) \tag{15}$$

The optimiser is Adam (initial learning rate 0.001) [37], an optimisation algorithm that can replace the classical stochastic gradient descent method to update the network weights in the training data iteratively. In short, Adam can adaptively adjust each network's learning rate. Furthermore, the logarithmic loss is also the first to deal with binary classification issues, namely the binary_crossentropy in Keras. Binary cross-entropy compares each predicted probability with the actual category output, which can be either 0 or 1. It then calculates a score that penalises the probability based on the distance from the expected value. To ensure training and modelling efficiency, 'Early Stopping' is applied to the model so that the network could be better generalised. The dropout layer is only applied to MLP and CNN-LSTM, as the proposed ConvLSTM model did not show overfitting performance during the experiments.

D. Validation Strategy

This paper selects the optimal model during the training and validation process and is applied it to a separate dataset for testing. Unlike the traditional training-to-test process, this approach increases the test results' stability, optimally, and reliability. K-fold cross-validation is applied to the validation process, where the training-validation dataset is divided into K copies, with K-1 copies used as training data and one copy as validation data on a rotating basis [38].

To further ensure the impartiality of the test results, three validation strategies are used simultaneously for the three models, i. e. the classical 10-fold and 5-fold cross-validation and no cross-validation. A total of nine test results are used for the final comparison and analysis.

IV. RESULT

A. Metrics Description

For model evaluation, comprehensive performance metrics were used on the test dataset (20%) to validate the accuracy, reliability, and robustness of the proposed model in identifying electricity theft users.

1) **Loss (Binary cross-entropy/Log loss)**: The loss function can be used to judge the predicted outcome of a classification model, i. e. the difference between the predicted value and the actual value. The loss function in binary classification is the binary cross-entropy, where y_i represents 0 or 1 in the label. The larger the prediction deviation, the higher the log-loss value. The equation is shown below:

TABLE III
TYPICAL BINARY CLASSIFICATION CONFUSION MATRIX.

		True Class	
		Negative (normal user)	Positive (theft user)
Predicted Class	Negative (normal user)	TN	FN
	Positive (theft user)	FP	TP

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\left(p\left(y_i\right)\right) + \left(1-y_i\right)\cdot \log\left(1-p\left(y_i\right)\right) \quad (16)$$

2) **Confusion Matrix**:The confusion matrix is a technique used to summarise the performance of a classification algorithm by providing a more intuitive indication of the correctness and error types of the model [39]. It also reduces false positives and increases true positives to ensure that the model efficiently fits a real-world usage scenario. TP (True Positive) refers to the number of theft users correctly classified in this paper. TN (True Negative) is the number of normal users correctly classified. FP (False Positive) is the number of normal users incorrectly predicted as theft users. FN (False Negative) is the number of theft users incorrectly predicted as normal users. A typical confusion matrix is shown in Table III.

Accuracy is a straightforward and meaningful metric in a state where the number of data set classes is balanced. It refers to the frequency of correct predictions out of all predictions made by the model.

$$\text{Accuracy } = \frac{TN+TP}{TN+FN+TP+FP} \quad (17)$$

Precision indicates the ability of the model to correctly predict positives from all positive predictions, while recall indicates the ability of the model to correctly predict positives from actual positive samples, i.e., representing the classification accuracy of theft users and actual theft users respectively.

$$\text{Precision } = \frac{TP}{TP+FP} \quad (18)$$

$$\text{Recall } = \frac{TP}{TP+FN} \quad (19)$$

F1-score gives equal weight to precision and recall and captures trends in them. It is an important measure of a classification model in the presence of false positives and false negatives.

$$\text{F1 -score } = 2 \times \frac{\text{Precision } \times \text{ Recall}}{\text{Precision } + \text{ Recall}} \quad (20)$$

Cohen’s kappa can be used to judge the strength of the model’s classification predictions. The Kappa value CK is a metric for comparing the observed accuracy with the expected accuracy.

$$\text{CK } = \frac{2\times(TP\times TN - FN\times FP)}{(TP+FP)(FP+TN)(TP+FN)(FN+TN)} \quad (21)$$

The receiver operating characteristics (ROC) curve is composed of TP and FP, and the area under the ROC curve (ROC-AUC) is the area of the ROC curve and FP. The larger the ROC-AUC, the better the classification ability of the model. The PR curve consists of precision and recall and is evaluated similarly to the ROC-AUC. The ROC-AUC is independent of threshold selection and reflects the characteristics of the model, while the PR-AUC can be considered as the average of the precision calculated for each recall threshold. The focus of the PR curve on the minority class makes it an effective diagnostic for imbalanced binary classification models.

B. Case Study

In this section, the performance of the proposed ConvLSTM-based ETD method is evaluated by comparing with related deep learning based ETD models (MLP and CNN-LSTM models). Three cross validation methods as introduced in III.D are applied for each model. The ETD performance is shown in Tables IV. From the table, the proposed CovnLSTM model outperforms other models in prediction. Especially, the ConvLSTM with 10-fold cross-validation shows the highest values of almost all metrics, expect for Loss. The result demonstrates that K-fold cross validation can optimize the deep neural network model and reduces the overfitting at the same time. It is also noticed that the benchmark model, MLP, shows the worst performance in all cases. MLP model has the simplest structure without the memory cell to store the historical information. Hence, MLP is inefficient in time-series tasks such as the ETD in this paper.

The PR and ROC curves for the proposed ConvLSTM ETD method and other related ETD methods are plotted in Figs. 11 and 12. PR curve is a plot of the precision (y-axis) and the recall (x-axis) for different probability thresholds and a model with perfect skill is depicted as a point at a coordinate of (1,1), a no-skill classifier will be a horizontal line with value 0.5 on the plot (blue with dashes in Fig. 11). In turns of the ROC curve, it plots of False Positive Rate vs. True Positive Rate, a point in the top left of the plot indicates a perfect prediction, while the no skill classifier is presented as a diagonal line (blue with dashes in Fig. 12). The ConvLSTM with 10-fold cross-validation has the best performance in terms of both ROC and PR curves. It has the largest AUC (a value close to 1), which indicates the performance of ConvLSTM is very close to the perfect point. It is also observed that CNN-LSTM with 10-fold cross-validation has similar performance as ConvLSTM model on the ROC and PR curves. In TP and FP recognition, it can achieve prediction accuracies of 47.37% and 1.63%, which is close to the performance of ConvLSTM with 10-fold. This indicates that CNN-LSTM with 10-fold also has a strong ability to identify and discriminate against normal users but is not as good as ConvLSTM in predicting electricity theft users.

Fig. 13 shows the Loss and Accuracy of the three models by taking 10-fold validation strategy as examples. ConvLSTM performs the best in generalisation ability and convergence efficiency. As shown in Fig. 13, its generalisation gap is around 0.1. In addition, ConvLSTM only requires around 60 epochs to converge to the best model state, while MLP and CNN require 150 and 125 epochs, respectively. MLP and CNN-LSTM reach smooth convergence at around 100 and 80 epochs, respectively. Furthermore, ConvLSTM still outperforms the other two models in noise control without

TABLE IV
ENERGY THEFT DETECTION PERFORMANCE

Models	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1-score**	**CK**	**ROC-AUC**	**PR-AUC**
MLP with 10-Fold	0.915	0.336	0.95	0.876	0.911	0.83	0.959	0.963
CNN-LSTM with 10-Fold	0.957	0.204	0.967	0.947	0.957	0.915	0.976	0.982
ConvLSTM with 10-Fold	**0.966**	0.23	**0.984**	**0.948**	**0.966**	**0.932**	**0.977**	**0.98**
MLP with 5-Fold	0.896	0.315	0.949	0.838	0.89	0.7921	0.95	0.957
CNN-LSTM with 5-Fold	0.944	**0.201**	0.972	0.914	0.942	0.887	0.972	0.978
ConvLSTM with 5-Fold	0.958	0.288	0.975	0.947	0.957	0.916	0.974	0.974
MLP without K-Fold	0.891	0.363	0.92	0.856	0.887	0.792	0.944	0.947
CNN-LSTM without K-Fold	0.941	0.202	0.958	0.922	0.94	0.969	0.969	0.974
ConvLSTM without K-Fold	0.942	0.221	0.972	0.911	0.94	0.884	0.974	0.978

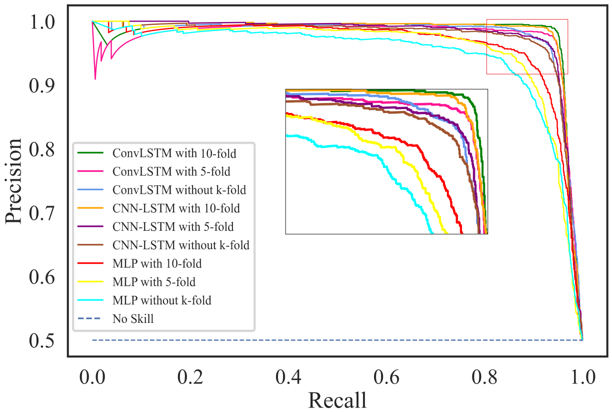


Fig. 11. PR curves for the proposed ConvLSTM ETD method and other related ETD methods.

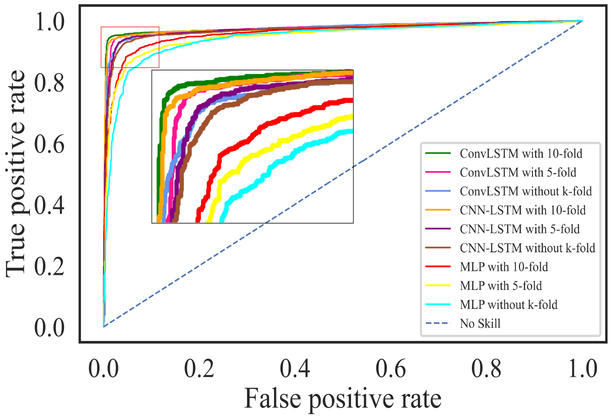


Fig. 12. ROC curves for the proposed ConvLSTM ETD method and other related ETD methods.

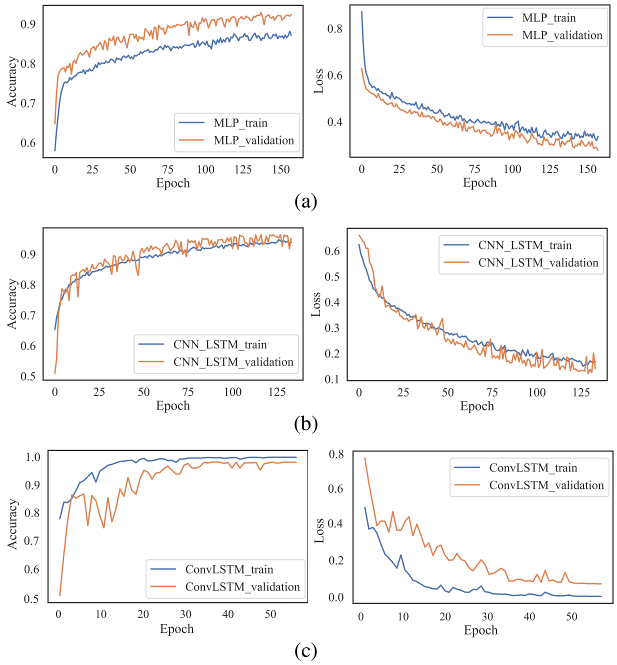


Fig. 13. Accuracy and loss of the proposed ConvLSTM-based ETD method and related ETD models (a) MLP ETD model; (b) CNN-LSTM ETD model; and (c) proposed ConvLSTM ETD model.

dropout. This result demonstrates that ConvLSTM has strong predictive robustness, and its model structure can effectively avoid over-fitting.

V. CONCLUSION

In this paper, a hybrid ConvLSTM ETD method is proposed. The proposed method combines ConvLSTM and a batch-normalisation to improve the flexibility of the training and testing phases. Moreover, the ETD model utilizes borderline-SMOTE to generate synthetic energy theft samples to better solve the imbalanced classification problem. From the simulation, the ConvLSTM shows excellent identification of electricity theft users under all three validation strategies and has significant advantages in model robustness, convergence efficiency and generalisation ability. This paper also demonstrates that the model with 10-fold cross-validation outperforms the models with the 5-fold cross-validation and no-cross-validation methods.

The extension of multi-dimensional electricity usage data is an issue to consider in future work. Multiple features can be added to the dataset to better match the input of a multi-dimensional tensor, e.g., weather, geographical location, etc. This is also a way for the ETD model to incorporate objective factors to detect potential electricity thieves.

ACKNOWLEDGMENT

The authors would like to acknowledge the funding by the PETRAS National Centre of Excellence for IoT Systems Cybersecurity, which has been funded by the UK EPSRC under grant number EP/S035362/1.For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Data Availability

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

References

[1] X. Fang, S. Misra, G. Xue, and D. Yang, “Smart grid — the new and improved power grid: A survey,” *IEEE Communications Surveys & Tutorials*, vol. 14, no. 4, pp. 944–980, 2012.

[2] K. Yu, M. Arifuzzaman, Z. Wen, D. Zhang, and T. Sato, “A key management scheme for secure communications of information centric advanced metering infrastructure in smart grid,” *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 8, pp. 2072–2085, 2015.

[3] L. Peretto, “The role of measurements in the smart grid era,” *IEEE Instrumentation & Measurement Magazine*, vol. 13, no. 3, pp. 22–25, 2010.

[4] Z. Yan and H. Wen, “Performance analysis of electricity theft detection for the smart grid: An overview,” *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–28, 2022.

[5] S. S. S. R. Depuru, L. Wang, and V. Devabhaktuni, “Electricity theft: Overview, issues, prevention and a smart meter based approach to control theft,” *Energy policy*, vol. 39, no. 2, pp. 1007–1015, 2011.

[6] C. Gokhale-Welch and L. Beshilas, “Smart grids in emerging markets-private sector perspectives,” National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2020.

[7] G. M. Messinis and N. D. Hatziargyriou, “Review of non-technical loss detection methods,” *Electric Power Systems Research*, vol. 158, pp. 250–266, 2018.

[8] S. Weckx, C. Gonzalez, J. Tant, T. De Rybel, and J. Driesen, “Parameter identification of unknown radial grids for theft detection,” in *2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, 2012, pp. 1–6.

[9] P. Kadurek, J. Blom, J. F. G. Cobben, and W. L. Kling, “Theft detection and smart metering practices and expectations in the netherlands,” in *2010 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT Europe)*, 2010, pp. 1–6.

[10] E. S. M. A. Program, *Energy analytics for development: big data for energy access, energy efficiency, and renewable energy*. World Bank, 2017.

[11] B. Coma-Puig, J. Carmona, R. Gavaldà, S. Alcoverro, and V. Martin, “Fraud detection in energy consumption: A supervised approach,” in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016, pp. 120–129.

[12] Z. Yan and H. Wen, “Electricity theft detection base on extreme gradient boosting in ami,” *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–9, 2021.

[13] R. R. Bhat, R. D. Trevizan, R. Sengupta, X. Li, and A. Bretas, “Identifying nontechnical power loss via spatial and temporal deep learning,” in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016, pp. 272–279.

[14] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[15] M. Khanafer and S. Shirmohammadi, “Applied ai in instrumentation and measurement: The deep learning revolution,” *IEEE Instrumentation & Measurement Magazine*, vol. 23, no. 6, pp. 10–17, 2020.

[16] G. Dorffner, “Neural networks for time series processing,” in *Neural network world*. Citeseer, 1996.

[17] S. Albawi, T. A. Mohammed, and S. Al-Zawi, “Understanding of a convolutional neural network,” in *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1–6.

[18] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in *International conference on machine learning*. PMLR, 2013, pp. 1310–1318.

[20] M. N. Hasan, R. N. Toma, A.-A. Nahid, M. M. Islam, and J.-M. Kim, “Electricity theft detection in smart grid systems: A cnn-lstm based approach,” *Energies*, vol. 12, no. 17, p. 3310, 2019.

[21] Z. Zheng, Y. Yang, X. Niu, H.-N. Dai, and Y. Zhou, “Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids,” *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1606–1615, 2018.

[22] R. U. Madhure, R. Raman, and S. K. Singh, “Cnn-lstm based electricity theft detector in advanced metering infrastructure,” in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, pp. 1–6.

[23] X. Zhang, S. Kuenzel, N. Colombo, and C. Watkins, “Hybrid short-term load forecasting method based on empirical wavelet transform and bidirectional long short-term memory neural networks,” *Journal of Modern Power Systems and Clean Energy*, 02 2022.

[24] X. Zhang, C. Watkins, and S. Kuenzel, “Multi-quantile recurrent neural network for feeder-level probabilistic energy disaggregation considering roof-top solar energy,” *Engineering Applications of Artificial Intelligence*, vol. 110, p. 104707, 04 2022.

[25] S. Sharma, S. Sharma, and A. Athaiya, “Activation functions in neural networks,” *towards data science*, vol. 6, no. 12, pp. 310–316, 2017.

[26] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, “Convolutional lstm network: A machine learning approach for precipitation nowcasting,” *Advances in neural information processing systems*, vol. 28, 2015.

[27] J. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, “Deep convolutional neural networks on multichannel time series for human activity recognition,” in *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[28] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[29] T. T. Verlekar and A. Bernardino, “Video based fire detection using xception and conv-lstm,” in *International Symposium on Visual Computing*. Springer, 2020, pp. 277–285.

[30] R. Mutegeki and D. S. Han, “A cnn-lstm approach to human activity recognition,” in *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 2020, pp. 362–366.

[31] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional lstm and other neural network architectures,” *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[32] S. Wager, S. Wang, and P. S. Liang, “Dropout training as adaptive regularization,” *Advances in neural information processing systems*, vol. 26, 2013.

[33] A. Richardson, “Nonparametric statistics for non-statisticians: A step-by-step approach by gregory w. corder, dale i. foreman,” 2010.

[34] M. Kuhn, K. Johnson, *et al.*, *Applied predictive modeling*. Springer, 2013, vol. 26.

[35] R. Dawson, “How significant is a boxplot outlier?” *Journal of Statistics Education*, vol. 19, no. 2, 2011.

[36] H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-smote: a new over-sampling method in imbalanced data sets learning,” in *International conference on intelligent computing*. Springer, 2005, pp. 878–887.

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

[38] J. D. Rodriguez, A. Perez, and J. A. Lozano, “Sensitivity analysis of k-fold cross validation in prediction error estimation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 569–575, 2010.

[39] P. Branco, L. Torgo, and R. P. Ribeiro, “A survey of predictive modeling on imbalanced domains,” *ACM Computing Surveys (CSUR)*, vol. 49, no. 2, pp. 1–50, 2016.

Hong-Xin Gao received the B.B.A. degree and the B. Eng. degree from the Shandong Agricultural University, Shandong, China in 2011. The M.S. degree with distinction in Information Security from the Royal Holloway, University of London, London, U.K., 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 area is deep learning for anomaly detection in advanced smart grids.

Stefanie Kuenzel (S'11-M'14-SM'19) received the M.Eng.and Ph.D. degrees from Imperial College London, London,U.K., 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, London, U.K.. Her current research interests include renewable generation and transmission, including HVDC as well as Smart Meters.

Xiao-Yu Zhang received the B. Eng. degree from the North China Electric Power University, Beijing, China in 2016. The M.S.degree with distinction in electrical power system from University of Birmingham, Birmingham, U.K., in 2017, and the Ph.D. degree in electrical engineering from the Royal Holloway, University of London, London, U.K., in 2022. He is currently a lecturer in School of Artificial Intelligence, Anhui University, Hefei, China. His research interests include deep learning technology & data analytics in smart grids, smart grid privacy and security, and, demand-side management.