Fraud detection on power grids while transitioning to smart meters by leveraging multi-resolution consumption data

Pablo Massaferro, J. Matías Di Martino, and Alicia Fernández

***Abstract*—The technological upgrade of power utilities to smart metering is a process that can take several years. Meanwhile, smart meters coexist with previous generations of digital and electromechanical power meters. While the smart meters provide high-resolution power measurements, electromechanical meters are typically read by an operator once a month. The coexistence of these two technologies poses the challenge of monitoring non-technical losses (NTL) and fraud where some customers' consumption is sampled every 15 minutes, while others are sampled once a month. In addition, since companies already have years of monthly historical consumption, it is natural to reflect how the past data can be leveraged to predict and improve NTL on smart grids. This work addresses both problems by proposing a multi-resolution deep learning architecture capable of simultaneously training and predicting input consumption curves sampled 1 a month or every 15 minutes. The proposed algorithms are tested on an extensive data set of users with and without fraudulent behaviors collected from the Uruguayan utility company UTE and on a public access data set with synthetic fraud. Results show that the multi-resolution architecture performs better than algorithms trained for a specific type of meters (i.e., for a particular resolution).**

***Index Terms*—Non-technical losses, electricity theft, automatic fraud detection, deep learning, multi-resolution, smart meters.**

I. Introduction

TECHNOLOGICAL updates are a big challenge for utility companies. Migration to new technologies in the energy area is a costly and time-consuming process, but it must be done to ensure the efficiency and competitiveness of services. In particular, migrating to an Advanced Metering Infrastructure (AMI) is a process that can take several years. These changes involve a large investment and a plan to replace the measurement equipment throughout the distribution network. Given the magnitude of this investment, companies make these changes in stages. This implies that electricity consumption data from different sources for different customers coexist in companies. For example, while some clients report their monthly consumption, others can be remotely monitored with readings every 15 minutes. An additional challenge is that for those customers for whom the meter was recently updated, their historical consumption (e.g., the last three years) has a period of coarser resolution (before the new meter was installed) and a period of higher resolution.

This work was supported in part by the Uruguayan utility company UTE and by the postgraduate academic commission of Universidad de la República.

A classic problem in the analysis of electricity consumption data is detecting non-technical losses (NTL), an issue of great interest given the economic impact on energy distribution companies [1], [2]. There are numerous proposals in the literature to address this problem for smart meter infrastructures [3]–[6] and the previous generations of meters [7]–[9]. However, to the best of our knowledge, no solutions have been proposed to address NTL in the context of infrastructures with coexisting technologies (leading to a multi-resolution problem across and within customers). The present work addresses this problem and discusses a solution based on deep learning architectures to combine power consumption data of different temporal resolutions while leveraging historical and contemporary ground truth data.

II. Related Works

The emergence of smart meters created new opportunities for data analysis in the electricity sector. Applications to load forecasting, management, and analysis have advanced tremendously, powered by novel machine learning tools. A review of the use of smart meter data in these topics was presented by Wang et al. [10]. Within the analysis of electricity consumption, NTL is a very challenging problem with a very active academic and industry community. The exponential increase in data related to AMI creates new challenges and opportunities. A detailed review on the different approaches to the NTL problem up to 2018 was presented by Messinis et al. [1] and divides the approaches into three large groups: data-oriented, network topology-oriented, and hybrids.

To control energy losses, companies carry out inspections of the meters installed in customers' residences. These activities generate high-quality labeled data allowing the development of detection strategies based on supervised learning. However, access to this data is restricted, and therefore numerous studies limit their analysis to fraud simulations [5], [11]–[14]. A public data set with real fraud and daily consumption data was made available by State Grid Corporation of China [3]. However, this data set

does not have the inspection dates, so there is no way to segment the valid data time intervals (those before inspections). Complex deep learning architectures show excellent results on this data, exceeding 90% precision [3], [6], [15]. This performance is over-optimistic when we compare it with the performance reported by distribution companies when novel inspections are performed based on the prediction of classification algorithms, with precisions ranging from 15% to 47% [16], [17]. This performance gap illustrates how important it is to have access to real data and proper segmentation (prior to the in-site inspections) to guide algorithm training. When the inspection date is unavailable and all the signal is fed into training, algorithms can learn changes in behavior triggered by the company inspection instead of the patterns of fraud we attempt to detect.

In [1] the classical supervised learning approach is described, where expert features are extracted from data to train a classification algorithm. Within the classical approach, the use of algorithms such as SVM, random forest (RF), extreme gradient boosting (XGB), and neural networks stand out [2], [7], [18]. However, in recent years data-oriented approaches with the use of deep learning have shown very good results. In [19] it is shown how data-oriented feature extraction approaches perform better than manual feature extraction for NTL. In recent years deep learning architectures have been used showing promising results on smart meter data. A variety of deep learning solutions have been proposed based on convolutional neural networks (CNN), [3], [20], long short-term memory (LSTM) layers, and recurrent neural networks [16], [17], [21].

Performance metrics play a crucial role when designing, comparing, and evaluating machine learning models. The most used performance measure to assess supervised classification methods is the accuracy (Acc) or the error rate (1-Acc), which are adequate when both classes have similar prevalence levels and the costs of the errors are similar. In NTL, these hypotheses are not fulfilled; there is a high imbalance, which must be considered when selecting the models and in the final evaluation of the test sets. Many works use Acc and detection rate (DR) without considering the imbalance. Among works that consider the problem of classes imbalance [2], [9], [22] measures such as the F1 score, the area under the ROC, and the area under the precision-recall curve are recognized as appropriate. However, there is no consensus on which is the single optimal evaluation metric [22]. In [18] it is proposed to use the economic return of the inspection activities as a performance metric, which takes into account the imbalance and the fact that the fraudulent samples can differently impact the economic return. This work compares models designed with a threshold that maximizes the F1 measure [9] and others that maximize economic return. This last method requires available information to estimate the return per sample.

Given the variability in the uses of electrical energy and the possible forms of fraud, this is a problem of non-separable classes. No matter how powerful an algorithm is, the information from consumption curves has limited performance. There have been several proposals to include more information in decision-making. Recently Hu et al. proposed a recurrent neural network architecture with three inputs: individual consumption, substation consumption, and ambient temperature [16]. This method is compared to a wide and deep network [3] showing superior performance. While there are some approaches to including new information in deep learning architectures [4], [17], [19], [23], there have been no contributions to the NTL literature to use multi-resolution measurements and address the periods where measurement technologies coexist. These periods can last for years; thus, the use of this information is of notorious practical application to power distribution companies. A significant number of publications use deep learning architectures to extract features of daily resolution consumption curves either because it is the resolution of their data [3], [15], [16] or to reduce the dimensions of the input data [17]. Other works directly use the readings with periods of one hour or 30 minutes [5], [12], [20], [24]. In this work, we focus on the impact of using multi-resolution consumer data for NTL detection. 15-minute resolution data from smart meters and monthly readings prior to change of measurement technology are used.

III. Proposed approach

A. Multi resolution

In a stage of technological update of data systems, it is as important to maintain historical records as it is to generate added value with the new information available. In general, there are several years of monthly energy consumption data prior to the installation of smart meters (SM). With the new AMI, it is possible to have energy consumption data every fifteen minutes for the same clients since the installation of the SM. The objective is to identify fraudulent behavior by analyzing the clients' consumption curves. For this, we propose to use a convolutional neural network architecture with two inputs. The time series of monthly data passes through a one-dimensional convolutional network from which relevant features are extracted [25]. In a second input, the energy consumption every 15 minutes of the last three months is used to form a 90x96 image (see Fig. 1). We tested experimentally including different time intervals, and we empirically observed that there is a performance saturation when 90 or more days of data are considered. For the network input image, each of the 96 columns corresponds to the energy consumption at one time of the day, and each row corresponds to a different day. This image is fed into a two-dimensional convolutional network.

Convolutional 2D networks have proven to be very powerful for identifying patterns in images. In these networks, pattern detection is translation invariant, which in our approach is equivalent to time

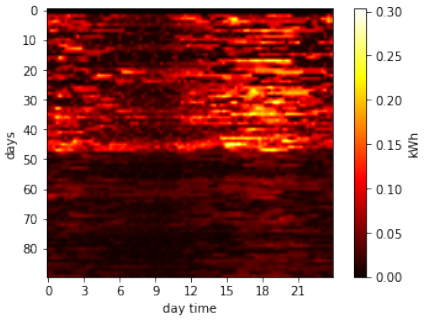


Fig. 1: Example of fraudulent energy consumption in 2D. The y-axis corresponds to the days and the x-axis to the time of day. Time advances from top to bottom and from left to right.

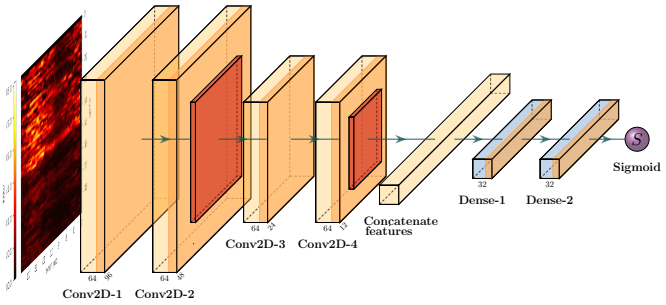


Fig. 2: CNN2D architecture for fraud detection with fifteen minute active energy consumption data.

invariance, which is desired in the context of NTL. In this way, the algorithm learns to identify suspicious energy consumption and can then find a similar pattern in another client in another region of the image (another date).

In addition to translation invariance, convolutional networks drastically reduce trainable parameters compared to fully connected networks. This is based on the fact that images have local structures (lines, shapes, colors, etc.) that can be captured by a set of filters (kernels). The detection of small parts allows coding more complex structures in a hierarchical way. Each filter is equivalent to a neuron that has connections to only one region of the image. The convolution operation allows each filter to operate with all areas of the image, sharing the values of the trainable parameters (kernel weights). Even though one could use for both (the high-resolution and the low-resolution) inputs 1D convolutional layers, it has been proved that re-arranging the high-resolution data into 2D images allows more efficient feature extraction with kernels of more compact and narrow support [24] Intuitively, since rows correspond to the consumption of each day, by reshaping this 1D vector into the selected 2D matrix (image), the consumption at the same time of the day between two consecutive days becomes a neighbor pixel in the 2D domain, allowing compact kernels to fit patterns associated with activities that occur a specific times during the day.

For the case of a single channel image I the convolution operation with a kernel K is presented in the following equation,

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n),$$

where m and n are the dimensions of the kernel, (i,j) the coordinates of a pixel, and S the output layer of the convolution. Each convolution layer consists of a filter bank characterized by its length and height dimensions, in this work of size 3x3. The depth is given by the dimension of the input data volume. In the case of deep learning, several concatenated convolutions are usually used. In this way, with the first layer of 64 3x3 filters, the data volume at the output will be depth 64. A 3x3 filter in the second layer has depth 64, this is equivalent to 576 trainable parameters in each filter. The activation of each neuron is given by a non-linear function, in this case, we use rectified linear unit (ReLU).

Convolution layers are the main building block of a CNN classifier architecture. Other widely used layers in a CNN classifier are pooling, batch normalization, dense, and softmax (or sigmoid for two classes). Pooling layers reduce the size of the data while creating multi-scale feature representations. Max pooling is used in the present work. Once features are extracted from the data, an multi layer perceptron (MLP) classifier can be trained by concatenating dense layers. To do this, the data volume of the last convolution layer is flattened and used as input of the first fully connected layer. The proposed NTL detector architecture with CNN2D is presented in the Fig. 2, this is used to process the high-resolution (2D) components of the input. We used convolutional layers as they are one of the most standard and efficient components in the state of the art deep network architectures, however, the main ideas of the paper are agnostic to this choice and could be implemented with other architectural choices such as recurrent neural networks. In section IV we compare the performance of different architectures classifying high resolution energy consumption data.

Detecting fraud using only monthly energy consumption data is a difficult task. In the past, we have tried to extract features from time series. However, for NTL the use of raw data in a one-dimensional convolution network, or machine learning algorithms such as xgboost or random forest, have shown to have better results [19]. Inspired by these previous findings, in this work, we use one-dimensional convolutional network for feature extraction. As with the 2D convolution network, adding dense layers at the end, an output layer with a neuron and a sigmoid function conclude the network architecture.

Finally, to merge the information from each branch (the

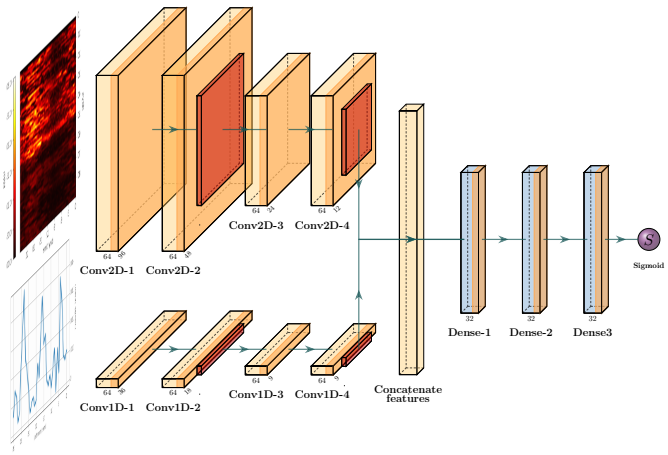


Fig. 3: multi-resolution architecture for NTL detection. The upper part represents the input of high resolution data in 90x96 image format and its processing with four layers of 2D convolution. The lower entry corresponds to the series of low frequency consumption (monthly). This data passes through a 1D convolution network. The values at the output of each branch are concatenated to enter a series of fully connected neural network layers. The output is a score given by a Sigmoid function.

low-resolution and the high-resolution energy measurments) we combine the last convolution layer of each network described before, feeding them into a MLP and simultaneously trains all the parameters using a single loss function (Fig. 3).

B. Metrics

We use AUC_PR as a metric for model comparison and hyperparameter selection. We also use $P@10\%$ as the primary performance metric in test runs, which is the precision obtained when performing inspections by labeling 10 % of the samples as positive. The set of solutions for which a fixed percentage of the data is labeled as positive (i.e., a fixed number of inspections) corresponds to a line in the precision-recall space as we show next. Defining TP, TN, FP, and FN the number of true positive, true negative, false positive, and false negatives, respectively, the number samples labeled as positive can be computed as $N_i = TP + FP$. Let N_P denote the number of positives $(TP + FN)$ and N_i the number of inspections to be carried out, we can express $Recall = TP/(TP + FN) = TP/N_P$ and $Precision = TP/(TP + FP) = TP/N_i$. Combining both equalities, we obtain

$$Precision = \frac{N_P}{N_i} Recall,$$

which is a line that passes through the origin with slope N_P/N_i. We add the operating line at $10\%(P@10\%)$ in the model comparison graphs with PR curves.

IV. Experiments

A. Data

1) UTE_SM, data set with real fraud: Starting in 2019, the Uruguayan power generation and distribution company (UTE) began a campaign to update energy meters, shifting from electronic or electromechanical meters to smart meters. The data set used in this work has data corresponding to 10,596 clients with smart meters inspected by UTE technicians between January 2019 and December 2020. For each client, the energy consumed every 15 minutes is available from the installation date to the date of inspection. In addition, there is a monthly consumption history for the last three years prior to the inspection date. Data is labeled as a member of the positive class (label equal to 1) if an irregularity was found and to the negative class otherwise (label equal to 0). The database has 772 positive cases representing 7.3 % of the total. In the experiments, we used 90 days of smart meters data per client. 90% of the records in the database have at least 90 days of valid data. Zero padding was performed for the 10% of the customers for which less than 90 days of smart meter readings were available. Some examples of fraudulent consumption profiles are shown in Fig. 4.

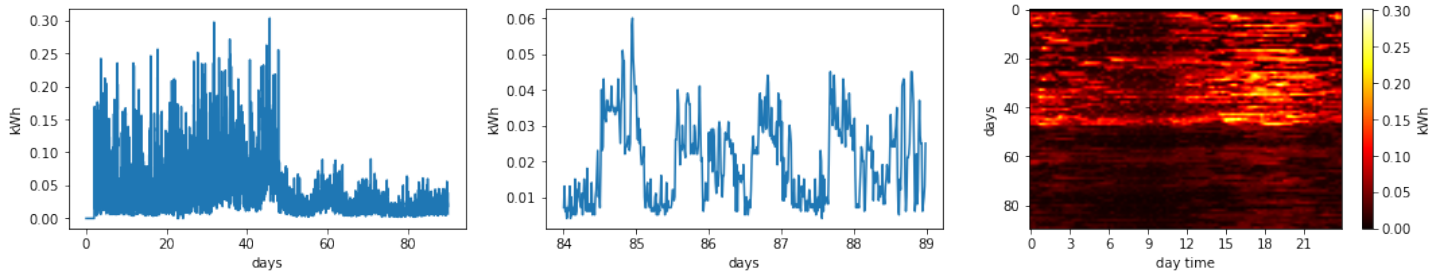
2) CER_NTL data set): CER_NTL is a synthetic fraud electrical energy consumption data set. Fraudulent patterns are created to simulate typical frauds on real consumption data from a public access database. The CER dataset was created by the Energy Regulatory Commission of Ireland to analyze customers' behavior on energy consumption [26]. The database contains, among other information, the report of active energy consumed every half hour in 6,435 homes for a period of 17 months.

UTE technicians' expertise and other academic works were taken into account for the fraud simulation [11]–[13]. Measurement readings can be partially or totally affected. Cyber attacks or meter modifications can affect the AMI permanently or for periods of time. Using byPass on the meter terminal block or in installations with a second connection without a meter may have an activation switch that allows fraud by time windows. The percentage of energy stolen depends on how the fraud is carried out. Similar models have been presented in other works [11], [13], [27]. In this work, we use a set of random variables to model human behavior when fraud takes place.

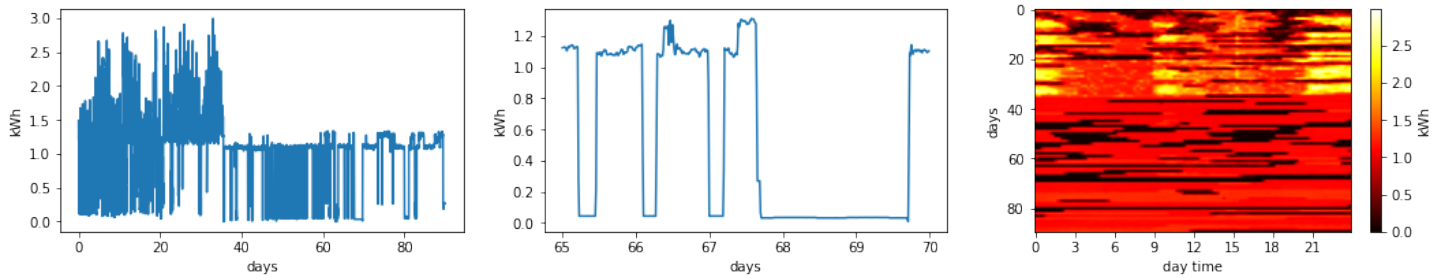
- Fraud 1: Constant proportional decrease over time

$$\hat{p}_{t_i,n} = \nu p_{t_i,n}; \;\; \nu \in [min, max]$$

where $p_{t_i,n}$ is the energy consumption value of the customer n at the instant t_i and $\hat{p}$ the consumption modified with fraud. For each customer n a fixed value of ν is assigned



(a) Illustration of a profile of a real fraud associated with the implementation of an electrical bridge in one of the phases of the meter's terminal.



(b) Illustration of a profile of fraud associated with the implementation of an intermittent electrical bypass in the meter.

Fig. 4: Examples of actual fraudulent profiles from the UTE SM data set. From left to right: the complete 90-day time series, a five-day series fragment of the period in which the fraud is active and representation of images of the smart meters' data of 90 days.

with uniform distribution $\nu \sim U[0.3, 0.7]$ and a random start date of fraud $t_{f,n}$ within the 17 months of data.

- Fraud 2: Proportional decrease in daily time windows.

$$\hat{p}_{t_i,n} = \delta_{t_i} p_{t_i,n} \text{ where } \delta_{t_i} = \begin{cases} \alpha \text{ if } t_{start} \leq t_i \leq t_{start} + l \\ 1 \text{ otherwise.} \end{cases}$$

α models the forgetfulness in a few days of the fraud triggering. $[\nu, 1]$ values with Bernoulli distribution. Let $F \sim$ Bernoulli(p)

$$\alpha = (F-1)\nu + F.$$

Fraud is committed daily in a window of time t_{start} and l model the noise at the start time and duration of the window. $t_{start} \sim N(\mu_s, \sigma_{ini})$, the mean start time value $\mu_s[n]$ is set for each customer with uniform probability within an hourly range $\mu_s \sim U(t_{min}, t_{max})$. The actuation duration is also a random variable with Gaussian distribution $l \sim N(\mu_d[i], \sigma_d)$. The mean value $\mu_d[i]$ is fixed over the days and is assigned to each customer using a uniform probability distribution $\mu_d \sim U(l_{min}, l_{max})$

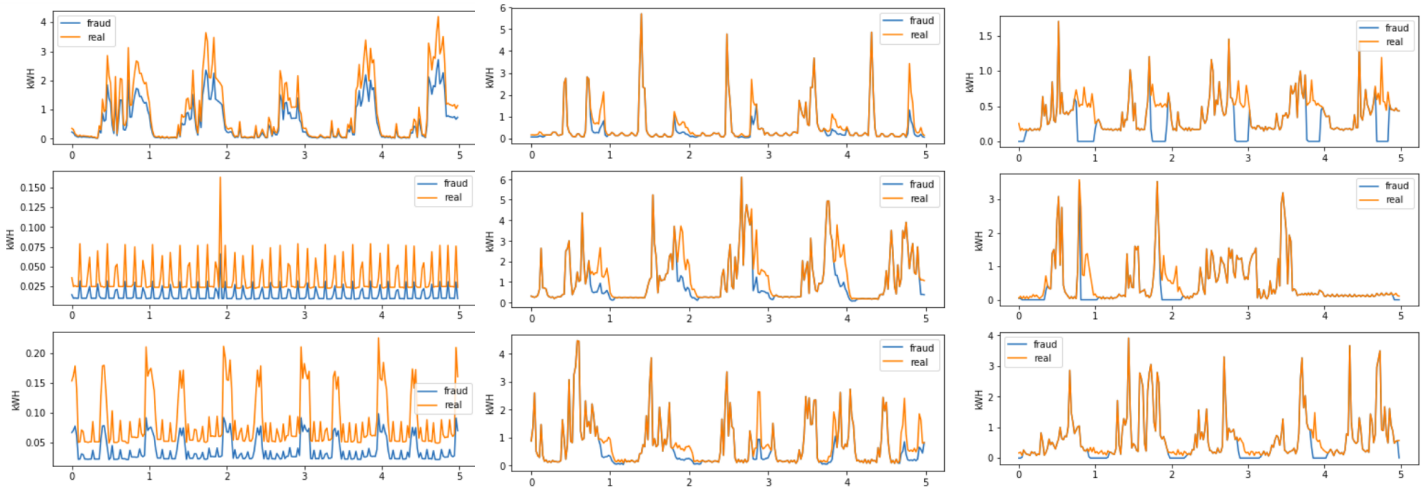
- Fraud 3: Total decrease in time windows during peak demand hours. This fraud models a zero consumption reading during a time windows and is a particular case of fraud 2 when $\alpha = 0$

$$\hat{p}_{t_i,n} = \delta_{t_i} p_{t_i,n} \;\; / \;\; \delta_{t_i} = \begin{cases} 0 \text{ if } t_{start} \leq t_i \leq t_{start} + l \\ 1 \text{ otherwise} \end{cases}$$

In this model, $t_{min} = 5pm$ and $t_{max} = 11pm$, this time interval is set as, usually, fraud occurs within the period of maximum consumption. For most residences this time is in the afternoon between 5 and 11 pm.

- Total decrease in energy reading in time windows without time bands. This fraud is similar to the third type with the difference that the fraud windows can be in at any time of the day. In other words, μ_s has no time slot restrictions. Likewise, each client has its own pattern of fraud.

Using the models of fraud described above, we generated the dataset CER_NTL, which contains 8% of positive (generated) samples with: 4% fraud type 1 with a theft range between 30% and 70 % ; 2% type 2 fraud with the same percentage decrease in consumption in windows between 5 pm and 11 pm. The windows last on average between 2hs and 6hs ($\mu_s \sim U[2, 6]$) with a variance of 1 h for the starting time ($\sigma_s = 1$) and a variance in the fraud duration of 1 h ($\sigma_d = 1$); 1% of the fraud corresponds to the type 3 parameters as fraud 2; and finally, 1% of type 4 fraud with windows of longer duration and without taking into account peak times. In this case, the daily duration of the fraud can range from 4 hours to 12 hours with a variance of 1 hour both in average and in duration for each client. The fraud generation



(a) Type 1 (b) Type 2 (c) Type 3

Fig. 5: Examples of synthetic frauds generated from real consumption data. Five consecutive days of active energy consumption are plotted for different consumption curves from CER data set. In orange the original consumption and in blue the modified (fraudulent) sample.

model with the parameters in the previously described values is made available in [1] to facilitate the experiment's reproducibility.

B. Implementation details

1) Test data set : To assess the prediction capacity of new fraud from the past data, we split our dataset into train and test sets by sliting customers into two sets, those that were inspected before a given date (train) and those inspected after it (test). This model the conditions at deployment, where inspections (potential new fraud) are planned based on prior data. The splitting date is set to obtain 15% of the samples for testing and 85% for training. The test data has a proportion of fraud of 8%. From the training set, 20 % of the data is taken as a validation set.

2) Models: We experimentally evaluated five architectures described next.

CNN1D: Time series classifier applied to monthly electricity consumption data. We use the following layer structure:

- $[Conv1D + BatchNormalization + ReLU + Conv1D + BatchNormalization + ReLU + MaxPooloing + Dropout] * N/2$
- $Flatten + [Dense + BatchNormalization + ReLU] * M + Dropout$
- $Dense(1) + Sigmoid$

where N is the number of convolution layers and M the number of fully connected layers.

Wide&Deep: NTL detection architecture proposed in [3]. For the sake of a fair comparison, our implementation takes as input the high-resolution measurements of the smart meter instead of the daily value of the original proposal. We use 90 days of data prior to the inspection (90x96 for UTE_SM and 90x48 for CER_NTL).

LSTM: Recurrent neural network for time series classification. The architecture includes an LSTM layer and a fully connected layer. The number of hidden units in the LSTM layer is included in the model hyperparameters for fine-tuning. The input for each time step is an array of full-day consumption measures. Like the other models tested, we consider 90 days of data as the model input.

CNN2D: Image classification algorithm applied to smart meter data. The layer structure is exactly the same as CNN1D with the difference that it uses convolutions in two dimensions and therefore max-pooling in 2D.

CNN_MR: This two-input architecture makes it possible to identify non-technical losses in energy consumption by combining historical monthly consumption data with smart meter readings. The convolution layers for both data inputs are the same as those used in the CNN1D and CNN2D models. Features extracted from the SM images and from the monthly consumption data are concatenated to train an MLP (see Fig. 3). The initialization weights of the convolution layers are those obtained by training CNN1D and CNN2D. The hyperparameter search for this architecture is only done in the MLP fully connected layers as shown in Table I.

Normalization layers are added to avoid gradient vanishing during training. Normalization is performed by computing the mean and standard deviation of each mini batch of data after

[1] https://github.com/pmassaferro/NTL_SmartMeters

Hyperparameters	CNN1D	CNN2D	CNN_MR
layers CNN	[2,4]	[2,4]	-
filters	[32,64]	[32,64]	-
kernel size	[3,5]	[(3,3)]	-
learning rate	$[10^{-2} : 10^{-4}]$	$[10^{-2} : 10^{-5}]$	$[10^{-3} : 10^{-5}]$
drop out	[0, 0.3]	[0,0.3,0.5]	[0, 0.3]
layers FC	[2,3]	[2]	[2,3,4]
neurons FC	[32,64]	[32,64,128]	[32,64]

TABLE I: Range of values for hyperparameters´search according to the selected architecture.

Data Set	Precision	Recall	F1	AUC_PR	AUC_ROC
Validation	0.15	0.23	0.18	0.16	0.70
Train	0.20	0.35	0.25	0.20	0.77

TABLE II: Results of the CNN_MR model on the training and validation data sets of UTE_SM.

each layer. This normalization implies the inclusion of two more training parameters for each activation layer. The implementation used (Keras) also computes the statistics with sliding windows, to then be used to normalize the test set, adding two more parameters for each activation layer.

3) Fine tuning: Within the search for hyper parameters we have included, in addition to the learning rate, variables that determine the complexity of the network, such as the number of hidden layers and the number of filters or neurons per layer. On the other hand, we also included dropout layers as a regularization agent to reduce overfitting. In order to cover a wide range of options, a random search is used combined with early stopping. During each training the AUC_PR is monitored in the validation data set. The unbalance of classes is taken into account during the training assigning different weights to each class when computing the loss function. The ranges of values for each hyper parameter are presented in Table I.

C. Models training

The first stage of multi-resolution model training consists of training the CNN2D and CNN1D models. The weights obtained in the convolution layers of these models are then used to initialize the CNN_MR model. Then, the convolution layer weights are fixed and the fully connected layers tuned. Figure 6 illustrates the training and validation loss as a function of the number of training epochs. Hyper-parameters (such as the number of filters) are optimized considering the validation loss. The highest performance model consists of four convolution layers and 64 filters, with a dropout rate of 30%. Table II presents the results obtained for the multi-resolution model on the training and validation datasets of UTE_SM.

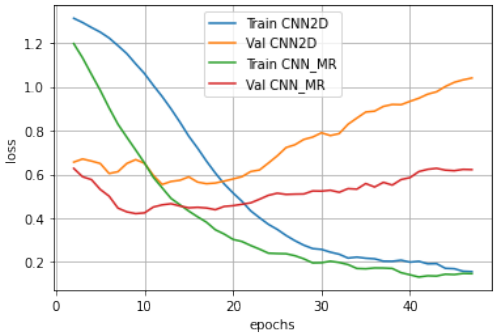


Fig. 6: Validation and training loss by epochs for CNN2D and CNN_MR models.

Model	Precision	Recall	F1	AUC_PR	AUC_ROC
CNN1D	0.10	0.12	0.11	0.11	0.56
Wide& Deep	0.15	0.17	0.16	0.11	0.55
LSTM	0.18	0.21	0.19	0.13	0.61
CNN2D	0.19	0.21	0.19	0.15	0.66
CNN_MR	**0.20**	**0.23**	**0.22**	**0.18**	**0.69**

TABLE III: Results on the test data set of UTE_SM. We considered 36 monthly consumption data as the low-resolution input (in the CNN1D and CNN_MR models) and 90 days of high-resolution data (Wide&Deep, LSTM, CNN2D and CNN_MR).

D. Results

The results obtained by the fraud detection models in both data sets (CER_NTL and UTE_SM) are presented next. The results tables include the overall evaluation metrics for the AUC_PR and AUC_ROC algorithms. P@10% (Precision when a fixed 10% of the customers are inspected), Recall, and F1 are also reported, setting the decision threshold to predict 10% of the test data set as positive.

1) Results on UTE_SM database : The results obtained from fraud detection in the UTE_NTL data set show that the CNN2D algorithm trained with three months of smart meter data achieves better results than those obtained by CNN1D using 36 months with low-frequency values. The CNN_MR multi-resolution model outperforms those models that leverage a single resolution despite of their architectural components (CNNs or LSTMs). Table III presents the results obtained on the UTE_SM data set for models studied. Observing the PR curves in Fig. 7 it can be seen that CNN_MR is not only optimal at the selected operation point (black dotted line), it is better for all decision thresholds. In addition, we observed that the type of layers selected to build the model have a reduced impact on the final model performance, suggesting that the idea of leveraging multi-resolution information can be implemented in practice with a variety of architectural choices.

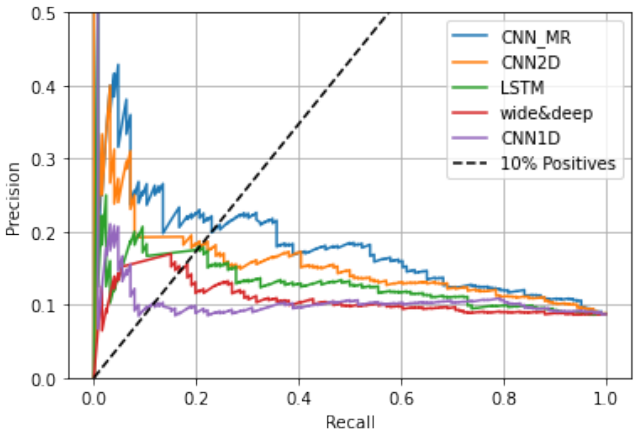


Fig. 7: Precision-Recall curves tested on UTE_SM test dataset.

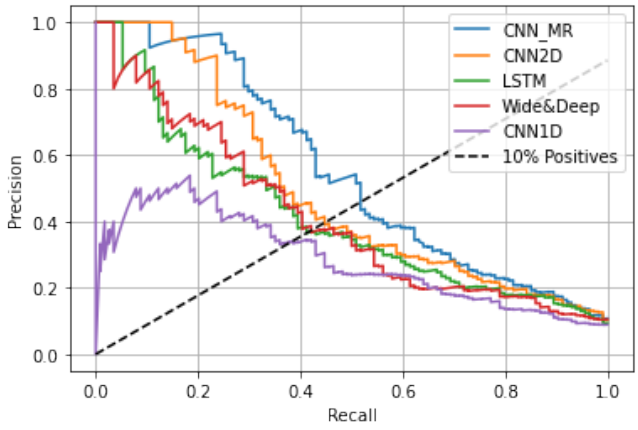


Fig. 8: Precision-Recall curves tested on CER_NTL test dataset.

Model	Precision	Recall	F1	AUC_PR	AUC_ROC
CNN1D	0.34	0.38	0.36	0.28	0.76
Wide&Deep	0.36	0.40	0.38	0.36	0.79
LSTM	0.37	0.41	0.39	0.41	0.82
CNN2D	0.40	0.45	0.42	0.49	0.84
CNN_MR	**0.46**	**0.52**	**0.49**	**0.55**	**0.86**

TABLE IV: Results on the CER_NTL test data set. We considered 17 monthly consumption data as the low-resolution input (in the CNN1D and CNN_MR models) and 90 days of high-resolution data (Wide&Deep, LSTM, CNN2D and CNN_MR).

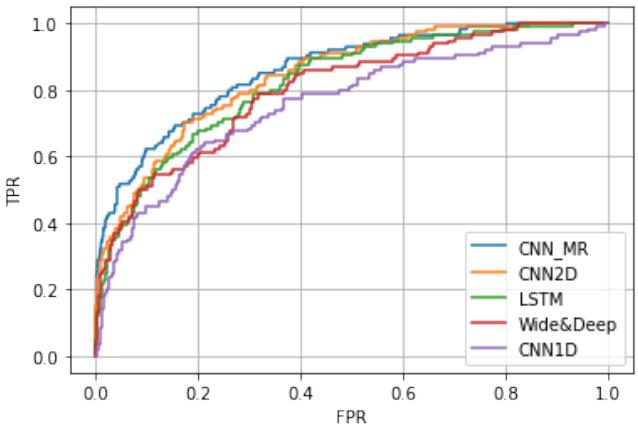


Fig. 9: ROC curves based on CER_NTL test dataset.

2) Results on CER_NTL database: As in the previous experiments, the multi-resolution algorithm achieves the best results outperforming models that consider independently low resolution (CNN1D) or high-resolution (Wide&Deep, LSTM and CNN2D) data. Again, multi-resolution models outperform their single resolution counterparts for any decision threshold (see Figs. 8 and 9). The CNN2D model achieves a P@10% of 40% while CNN1D model obtained 34%. Considering that the data has a proportion of 8% of fraudulent examples, the results more than quadruple a random classification. Other architectures such as LSTM and Wide&Deep taking three months of high-resolution data as input achieve superior performances than those obtained with three years of monthly data by CNN1D. The most relevant of the results is that the proposed multi-resolution architecture (CNN_MR) exceeds the performance of all tested algorithms, reaching a precision of 46% (see Table IV).

To analyze what each data source provides and understand why the multi-resolution model achieves better performance than the individual low-resolution (17months) and high-resolution (3months) models, we present in the table V the performance of P@10% for each of the four types of fraud included in the data set. It can be seen that the CNN1D algorithm is superior at detecting time-constant and proportional-decrease fraud (type 1), while CNN2D is much better at detecting fraud in time windows within each day. The multi-resolution model has the best overall performance and performs well in all four types of fraud. Type 2 fraud is the most difficult, coinciding with being the one that introduces the fewest modifications on the original measurement curve (see Fig. 5.) Type 3 fraud with a total decrease in the reported energy consumption within peak hours is detected at 100%, and for type 4, it obtains 94% with the multi-resolution model. CNN1D also achieves significant results detecting type 4 fraud since the windows are of long duration, significantly affecting the monthly consumption.

V. Discussion and Conclusions

We present a multi-resolution convolutional neural networks architecture for fraud detection on smart grids inspired by the recent shift of the grid infrastructure, where the new generation of smart meters is replacing older versions of digital and electrome-

Model	Fraud 1	Fraud 2	Fraud 3	Fraud 4	all
CNN1D	**0.52**	0.07	0.08	0.63	0.34
Wide&Deep	0.328	0.714	0.269	0.63	0.36
LSTM	0.34	0.86	0.19	0.63	0.37
CNN2D	0.22	**1.00**	**0.35**	0.94	0.40
CNN_MR	0.36	**1.00**	0.31	**1.00**	**0.46**

TABLE V: Precision results (P@10%) by fraud type for the CER_NTL sample.

chanical meters. Our multi-resolution approach shows superior performance to other NTL detection algorithms trained exclusively on smart meter data (CNN2D, LSTM and Wide&Deep). Experiments support our main hypothesis that multi-resolution information can be leveraged during technological infrastructure transitions to minimize NTL. Results are consistent both for simulated experiments on a publicly available dataset and on real data collected on the field by UTE company. Our results also suggest that the architectural choices (i.e., the particular layers that compose the network solution) are not critical and comparable results can be obtained leveraging CNN or LSTM components. One of the advantages of using convolutional architectures relies on their ease of interpretation, as we illustrate in in Fig. 10 where the activation layer associated with feature kernels is presented.

The relative performance of the models tested matches for the experiments performed on real (UTE_SM) and simulated (CER_NTL) fraud. Interestingly, despite of the relative coherence among models, a significant performance gap can be observed when detecting real fraudulent profiles compared with simulated cases. Although the different types of simulated fraud have a physical basis for their effect on consumption curves, and in that regard, they accurately simulate different electrical configurations associated with fraud, the lack of a model associated with the human behavior and energy usage patterns of a fraudulent customer could explain the performance mismatch.

We also observed that when compared independently, 3 months of fine-resolution data has more predictive power than 36 months of low-resolution data. This is great news for utility companies since it provides quantitative proof that novel smart meters improve the ability to detect and prevent NTL. Moreover, it shows that even a few months of high-resolution data can outperform more than a year of monthly readings, suggesting that companies do not have to wait for too long after new meters are installed to start detecting abnormal activities. Finally, our modular model can leverage both datasets from previous and future infrastructures. Our model's high-resolution and low-resolution components can potentially be fine-tuned and trained independently (freezing portions of the network) to leverage heterogeneous datasets that combine customers with a variety of meters.

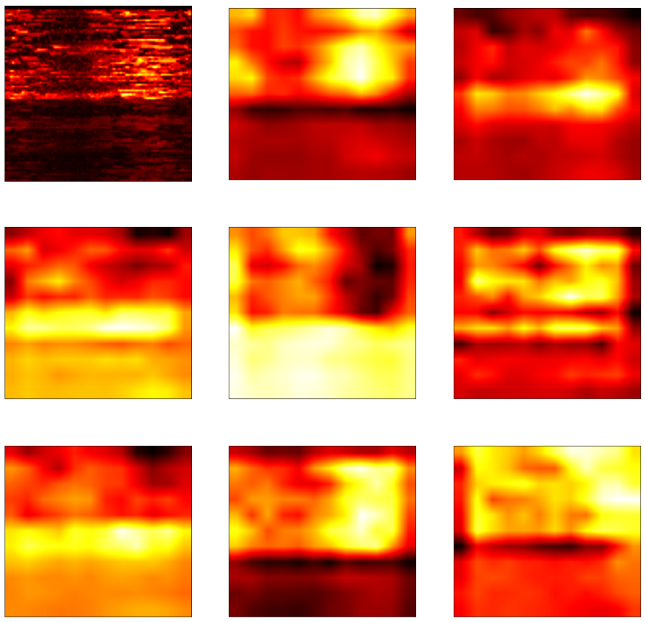


Fig. 10: Top left: an example of the 2D image associated with the high-resolution consumption profile of a fraudulent customer; the remaining images illustrate examples of activation layers for this test input for a random subset of kernels of the CNN2D model. As we can observe (see e.g., the top right activation) some learned features are associated with abrupt consumption shifts, suggesting some kernels are associated with directional edge detection.

ACKNOWLEDGMENT

The authors would like to thank UTE for providing the datasets and for sharing their expertise in particular, authors greatly acknowledge engineers Marcelo Álvarez, Gonzalo Caudullo, Ibero Fomichov, Alexander Martins and Andrés Jorysz.

REFERENCES

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

[2] M. S. Saeed, M. W. Mustafa, N. N. Hamadneh, N. A. Alshammari, U. U. Sheikh, T. A. Jumani, S. B. A. Khalid, and I. Khan, "Detection of non-technical losses in power utilities—a comprehensive systematic review," *Energies*, vol. 13, no. 18, p. 4727, 2020.

[3] 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.

[4] M. M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gómez-Expósito, "Detection of non-technical losses using smart meter data and supervised learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2661–2670, 2018.

[5] T. Hu, Q. Guo, H. Sun, T.-E. Huang, and J. Lan, "Nontechnical losses detection through coordinated biwgan and svdd," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 1866–1880, 2020.

[6] A. Aldegheishem, M. Anwar, N. Javaid, N. Alrajeh, M. Shafiq, and H. Ahmed, "Towards sustainable energy efficiency with intelligent electricity theft detection in smart grids emphasising enhanced neural networks," *IEEE Access*, vol. 9, pp. 25 036–25 061, 2021.

[7] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohamad, "Nontechnical loss detection for metered customers in power utility using support vector machines," *IEEE transactions on Power Delivery*, vol. 25, no. 2, pp. 1162–1171, 2010.

[8] P. Massaferro, H. Marichal, M. Di Martino, F. Santomauro, J. P. Kosut, and A. Fernández, "Improving electricity non technical losses detection including neighborhood information," in *2018 IEEE PES General Meeting (GM) - IEEE Power and Energy Society, Portland, Oregon, USA, 5-9 aug*. IEEE, 2018, pp. 1–5.

[9] M. Di Martino, F. Decia, J. Molinelli, and A. Fernández, "Improving electric fraud detection using class imbalance strategies." in *International Conference on Pattern Recognition and Methods, 1st. ICPRAM.*, 2012, pp. 135–141.

[10] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3125–3148, 2018.

[11] P. Jokar, N. Arianpoo, and V. C. Leung, "Electricity theft detection in ami using customers' consumption patterns," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216–226, 2015.

[12] R. Punmiya and S. Choe, "Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 2326–2329, 2019.

[13] S.-C. Yip, W.-N. Tan, C. Tan, M.-T. Gan, and K. Wong, "An anomaly detection framework for identifying energy theft and defective meters in smart grids," *International Journal of Electrical Power & Energy Systems*, vol. 101, pp. 189–203, 2018.

[14] K. Zheng, Q. Chen, Y. Wang, C. Kang, and Q. Xia, "A novel combined data-driven approach for electricity theft detection," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, pp. 1809–1819, 2018.

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

[16] W. Hu, Y. Yang, J. Wang, X. Huang, and Z. Cheng, "Understanding electricity-theft behavior via multi-source data," in *Proceedings of The Web Conference 2020*, 2020, pp. 2264–2274.

[17] M.-M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gómez-Expósito, "Hybrid deep neural networks for detection of non-technical losses in electricity smart meters," *IEEE Transactions on Power Systems*, vol. 35, no. 2, pp. 1254–1263, 2019.

[18] P. Massaferro, J. M. Di Martino, and A. Fernández, "Fraud detection in electric power distribution: An approach that maximizes the economic return," *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 703–710, 2019.

[19] ——, "NTL detection: Overview of classic and dnn-based approaches on a labeled dataset of 311k customers," in *2021 IEEE NA Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2021.

[20] Y. Wang, Q. Chen, D. Gan, J. Yang, D. S. Kirschen, and C. Kang, "Deep learning-based socio-demographic information identification from smart meter data," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2593–2602, 2018.

[21] G. Fenza, M. Gallo, and V. Loia, "Drift-aware methodology for anomaly detection in smart grid," *IEEE Access*, vol. 7, pp. 9645–9657, 2019.

[22] P. Glauner, J. Meira, P. Valtchev, R. State, and F. Bettinger, "The challenge of non-technical loss detection using artificial intelligence: A survey," *International Journal of Computational Intelligence Systems 10.1 (2017): 760-775.*, 2017.

[23] T. Hu, Q. Guo, X. Shen, H. Sun, R. Wu, and H. Xi, "Utilizing unlabeled data to detect electricity fraud in ami: A semisupervised deep learning approach," *IEEE transactions on neural networks and learning systems*, 2019.

[24] S. Li, Y. Han, X. Yao, S. Yingchen, J. Wang, and Q. Zhao, "Electricity theft detection in power grids with deep learning and random forests," *Journal of Electrical and Computer Engineering*, vol. 2019, 2019.

[25] 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)*. IEEE, 2016, pp. 272–279.

[26] Commission for Energy Regulation (CER), "CER Smart Metering Project - Electricity Customer Behaviour Trial, 2009-2010 [dataset]." 2012, 1st Edition. Irish Social Science Data Archive. SN:0012-00. https://www.ucd.ie/issda/data/commissionforenergyregulationcer/.

[27] F. G. Viera Nuñez, "Modelado y detección de fraudes en redes inteligentes de distribución de energía eléctrica." Master's thesis, Udelar.FI, IIE, 2020.

Pablo Massaferro received the B.Sc. and M.Sc. from Universidad de la República, Uruguay, in 2008 and 2018 respectively. Currently is a Ph.D. student and holds a position as Assistant Professor in the Signal Process Department of Universidad de la República, Uruguay. His main research areas are machine learning, source separation and data science applied to electrical power systems .

J. Matias Di Martino received the B.Sc. and Ph.D. degrees in Electrical Engineering from "Universidad de la Republica", Uruguay, in 2011 and 2015 respectively. During 2016-2017 he was a Research Associate at ENS (Paris) and currently he is Research Assistant Professor at the Department of Electrical and Computer Engineering, Duke University, US. His main areas of interest are applied optics, machine learning, facial and behavioral analysis and image processing.

Alicia Fernandez Full Professor of Signal Processing at the Electrical Engineering Institute(IIE), Universidad de la República. Since 1989, she works at the IIE, in telecommunication and signal processing areas. Her main research interests are signal processing and pattern recognition with focus in biomedical image analysis, biometric identification, anomaly detection and big data analysis.