Smart Grid Energy Fraud Detection Using Artificial Neural Networks

Vitaly Ford [1], Ambareen Siraj[2], William Eberle [3]
Computer Science Department
Tennessee Tech University
Cookeville, TN, USA
[1] vford42@students.tntech.edu
[2] asiraj@tntech.edu
[3] weberle@tntech.edu

***Abstract*—Energy fraud detection is a critical aspect of smart grid security and privacy preservation. Machine learning and data mining have been widely used by researchers for extensive intelligent analysis of data to recognize normal patterns of behavior such that deviations can be detected as anomalies. This paper discusses a novel application of a machine learning technique for examining the energy consumption data to report energy fraud using artificial neural networks and smart meter fine-grained data. Our approach achieves a higher energy fraud detection rate than similar works in this field. The proposed technique successfully identifies diverse forms of fraudulent activities resulting from unauthorized energy usage.**

Keywords—fraud detection; neural networks; smart meter data

I. INTRODUCTION

The Smart Grid incorporates many technological innovations introducing a plethora of opportunities for making the traditional power grid more efficient, reliable, and flexible with the integration of multiple types of sensors, alternative power sources, and smart home appliances [1]. While the Smart Grid continues to provide improvements primarily in energy efficiency, at the same time it is opening up many opportunities for security violations in different domains [14]. In this paper, we discuss the fraud detection aspect of smart meters in the customer domain.

The modernized power grid provides controllable access to consumers' fine-grained energy consumption data collected by smart meters via two-way wired and wireless connections.

We consider two different types of energy fraud in this work:

- Fraud type 1: the consumer reports less energy consumed than actually used; and
- Fraud type 2: more energy used by rogue connections than actually used by consumer.

Various mechanisms that can be used for carrying out these energy fraud are:

- Unauthorized tapping to electricity line.
- Bypassing the smart meter.
- Implanting different kinds of chips inside a smart meter to slow down its readings.

By applying machine learning techniques, we are able to model consumer's energy consumption behaviour (ECB) under normal conditions. With knowledge of the ECB, individual consumer's energy consumption can then be closely monitored to detect anomalies indicating fraudulent activities. We demonstrate the effectiveness of our approach by experimenting with the energy consumption data archived by the Irish Social Science Data Archive Center [2]. This anonymized energy data was gathered from actual consumers over almost a two-year period.

The paper is structured as follows. Section 2 describes related work in energy fraud detection. Section 3 describes data extraction and pre-processing methods. In section 4, our fraud detection approach is reported. Section 5 summarizes experimental results and analysis. Section 6 concludes with future work.

II. RELATED WORK

Cabral et al. in [17] applied self-organizing maps for detecting energy fraud by learning historical consumer energy consumption behavior and comparing it with present measurements. The researchers specifically focused on high voltage electricity consumer data supplied by a Brazilian electrical energy distribution company's database. The database contained weekly (Monday to Friday, 15 minute intervals) aggregated consumer energy consumption entries. For simulating fraud, several consumers were selected for an intentional drop of 30% on their energy consumption. The researchers reported that in 85% of cases, their system could identify the 30% drop in consumer energy consumption behaviour and raised an alarm.

Monedero et al. [18] utilized neural networks and statistical techniques for detecting non-technical losses (fraudulent activities) in electrical consumption. The researchers used actual energy consumption data generated by a small town in Spain and a hostelry business. It was noted that the overall number of consumers in the database were 72,927 and among them, only 22 were related to fraudulent activities. The proposed neural network system employed

978-1-4799-4546-7/14/$31.00 ©2014 IEEE

several consumer characteristics, such as 1) minimum, maximum, and average values of the bills during the previous year, 2) (maximum – minimum)/amount of contract power, 3) difference between the energy consumption in month *M* under investigation and the average consumption in month *M* over the previous year, 4) difference between the consumption in month *M* and the maximum consumption for month *M* over the previous year. The statistical analysis was conducted in a few stages. First, the researchers normalized the data. Assuming that the given data were normally distributed, the sample variance was calculated and a statistical outlier detection technique was employed. The resulting success rate of energy fraud detection was 50% after applying both neural networks and statistical analysis.

Nagi et al. [19] applied support vector machines for detecting fraudulent activities in a power grid. The proposed fraud detection framework consists of several steps including data pre-processing, feature extraction, classification, data post-processing, and identifying suspected customers. The researchers used two main features for analysing data: energy consumption on an hourly basis and the credit worthiness rating of consumers that was automatically generated by the billing system. For building the support vector machine classifier, 330 profiles without any abrupt changes or fraudulent activities were used and 53 samples were marked as abnormalities. The experiments were conducted with data containing energy consumption samples from three cities in the state of Kelantan, Malaysia. They noted that the percentage of energy fraud detected by the utility company in the past eight years was less than 1% of the total number of customers in each city. However, the estimated rate of fraudulent activities in those cities was about 35%. As a result of their experiment, they reported a 26% detection hit rate corresponding to the rate of identifying fraudulent activities. For improving the hit rate index, the researchers used a decision making system including human knowledge and expertise in energy fraud cases.

Costa et al. [20] proposed an approach for detecting energy fraud applying neural networks and data mining techniques. Their technique comprised of data cleaning and integration, data selection and transformation, data mining, and pattern recognition. In the data mining process, several key attributes were utilized for developing the model. Some of those attributes are: location, business class, activity type (e.g. residence, drugstore), mean consumption, query debts, and voltage. A neural network multi-layer perceptron was used for pattern recognition as the last stage of the data mining process. The experiment was conducted with real data from a Brazilian electric power distribution company. They used 21,583 records of consumer energy consumption measurements. Their classifier recognized 945 fraudulent activities among 1,453 fraudsters (65% fraud detection rate), and 2,261 energy fraud cases among 20,130 non-fraudsters.

As we can see, although energy fraud seems to be a very common and real problem, existing research in fraud detection are still not able to perform very well, demonstrating an unsatisfactory fraud detection rate. Also, most of these work use outside knowledge about the data to help investigation. Outside knowledge may not be easily attainable and may not apply uniformly to all consumers/utility providers. Also, in most of the simulated experiments, fraudulent activities were introduced in a consistent manner, which does not conform to real world scenario where things happen in random nature.

III. Data Extraction and Preparation

Before the neural network approach is described, it is imperative to introduce the data to be analysed for deriving computational intelligence. The dataset under investigation consists of smart meter consumption data from approximately 5,000 residential households and 600 businesses. These representative samples were anonymously collected during almost two years (2009 – 2011) by the Irish Social Science Data Archive Center. The data values were captured every 30 minutes during that time. The raw data are stored in six different files, each having around 24 million entries corresponding to diverse power meter readings. Table I represents a small sample of one of these data files where the data is represented in three columns. The first column represents the smart meter ID which is linked to a particular household. The second column shows the date and time associated with the meter reading and the third column is the energy consumption measurement in kilowatts-per-hours (kW/h).

TABLE I
Raw Data File Structure

Meter ID	Encoded date/time	Energy consumption value kW/h
1392	19503	0.140
1392	19504	0.138
…	…	…
1187	22028	1.367
1187	22029	1.425
1392	19940	0.234

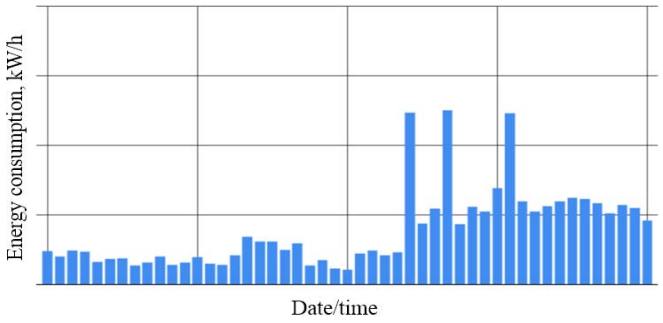

Fig. 1. A sample one-day period of energy consumption.

Fig. 1 shows an overview of energy consumption measurements in a single day for a particular consumer.

It is well known that the effectiveness of data processing depends not only on the data analysis algorithms used but also on the quality of the data. Therefore, for achieving better results in energy fraud detection, a suite of data pre-processing techniques is employed before any further analysis. In this section, we describe the two main procedures utilized for pre-processing the raw energy consumption data before they are analysed: *data cleaning and feature selection*. In addition, to

facilitate faster access to data in the files, we applied some indexing and compression algorithms widely used in information retrieval systems.

A. Data Cleaning

Data cleansing/cleaning is a data mining process which focuses on identifying and correcting inaccurateness, incompleteness, and inconsistency in raw data. It generally involves a few tasks such as identifying and addressing missing values, reducing noise and inconsistencies, and detecting outliers [3].

At a first stage, the data are explored for missing values. Since energy consumption data are time series data, it is quite straightforward to find all missing points as they correspond to missing time stamps. When such rare missing data points are found for a particular day and time, it was replaced with average energy consumption value for that particular day and time [16].

The second cleaning step involves identifying and eliminating outliers. In this regard, the mean and standard deviation (σ) for each time stamp comprising of weekday/time pair for each month is computed. Considering 7-day weeks and 24-hour days with 2 energy consumption measurements per hour, there is $7 \times 24 \times 2 = 336$ different time stamps. After calculating σ, all measurements that do not lie within three σ of the mean (we assume that around 1 % of the data are outliers or noisy/unimportant data) are removed from set. The assumption is that these outliers should not be considered as fraud and may correspond to peak energy consumption activities during holidays or special occasions such as birthdays, celebrations.

Lastly, we check the data with respect to real world inconsistencies in date/time stamps. For example, if we have a time stamp for the 13th month, we either attempt to correct the inconsistency or remove the entry.

B. Feature Selection

The experimental smart meter data contain several important parameters such as month, week, time, and consumer energy consumption values in kW/h. We decided to concentrate on all of the above-mentioned except the "month". The main reason behind this is that although seasonal weather changes can significantly influence the energy consumption measurements, conducting experiments with "month" feature included did not show any significant improvements in results. It is thus omitted for simplicity.

C. Indexing and compressing

The data files contain millions of various meter ID entries. In order to efficiently process the reading of consumer energy, we created an index which maps each meter ID to its own measurement data. Such an index provides quicker data retrieval from all files for a particular ID. In addition, the index is compressed by delta encoding widely used in information retrieval systems to store or transmit data in the form of differences between sequential data rather than complete files [13].

IV. Fraud Detection Approach

Machine learning is widely used in many areas of research for deriving computational intelligence. It allows understanding of the underlying behavior of complex systems. A neural network is one of the most common machine learning approaches used in fraud detection due to its noise-robustness and fast response qualities [7, 8, 9, 10, 11, 12].

A neural network consists of three different types of layers [3]. The first layer includes various features of the dataset under investigation as input. The second layer usually comprises of a few hidden layers having a varied number of nodes. And the last layer is an output layer representing the classification result of the neural network. The main goal of applying a neural network technique in this scenario is to learn consumption behavior per consumer and predict future energy consumption measurements. We believe that it is possible to detect energy fraud by comparing neural network predicted values with real measurements and applying statistics to analyse any deviations.

Our energy fraud detection approach consists of the following.

1) Selection of Input/Output Parameters for Neural Network Structure

Careful articulation of input/output parameters in a neural network structure is very important for performance and precision.

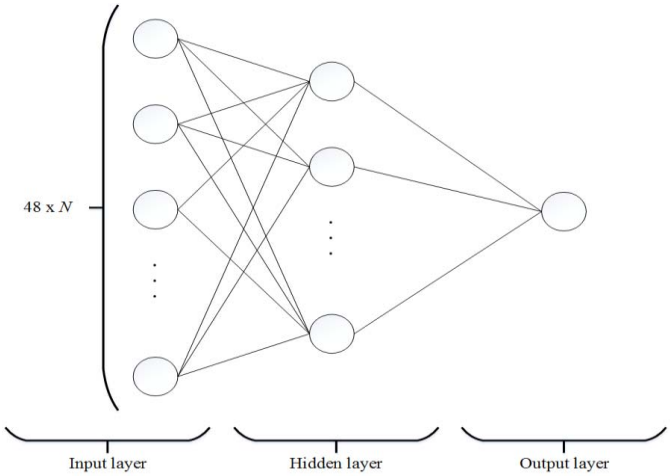

Fig. 2. ECB Neural network structure.

The number of consecutive days (N) serves as the input layer attribute for the neural network. Each day contains 48 measurements due to the fact that meter readings in the dataset are split into 30-minute intervals. Therefore, the input layer contains a total number of $48 \times N$ nodes.

The hidden layer is adjustable and in our experiment, it consists of only 1 layer of nodes for simplicity. The output layer consists of only one attribute, which represents the expected value in the smart meter read*ing data series following the* consecutive data points. Fig. 2 shows the ECB neural network structure employed in this research.

2) Generation of Training and Validation Datasets

Fig. 3 demonstrates generation of training and validation datasets. The training set contains energy consumption values of a certain consumer gathered during a particular period of time, for instance, 3 weeks. Multiple unique "training instances" are generated using the training set by selecting data points one by one and marking them as output nodes for the training instance. Afterwards for known output node values, the input layer nodes are generated by selecting K-consecutive energy consumption measurements (in this case, K is $48 \times N$). The process repeats until end of the training set.

The same procedure is repeated for generating the validation dataset with a different (month/year/season) set of smart meter measurements.

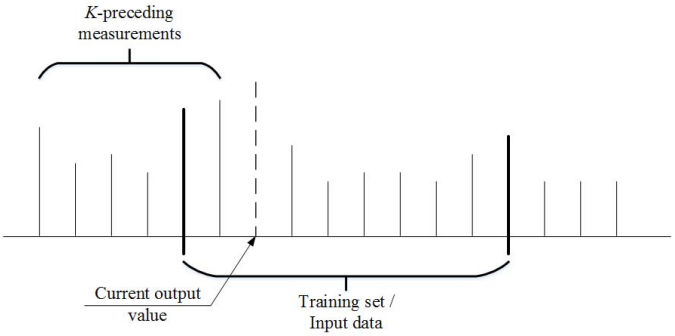

Fig. 3. Training and validation dataset generation.

3) Training the Neural Network

The neural network is trained using the training dataset which results in learning of the consumer energy consumption behavior.

4) Prediction

This task is accomplished by running the neutral network using the validation dataset. The output of the neural network predicts the value of the expected data point in the smart meter reading data series following the consecutive data points fed as input.

5) Detection of Deviation

The Root Mean Squared Error (RMSE [15]) serves as the deviation indicator between the predicted value and the actual value in the validation dataset and is calculated as in (1):

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} \left(y_t' - y_t\right)^2}{n}}, \qquad (1)$$

where y_t' is a predicted energy reading by the neural network, y_t is an actual energy consumption reading in the validation set, and n is the total number of instances. If the RMSE is above a certain threshold, then it is inferred that the training set depicts a different behavior than the validation set and therefore the data in the validation set possibly correspond to an energy fraud.

V. Experiments and Analysis

A. Experimental setup

Several diverse experiments were conducted over different periods of time and varied consumer energy consumption data. *WEKA* [4], which provides many machine learning solutions for data mining, was used as the primary experimental tool.

The experimentation consisted of two major steps. At first, the ECB neural network was trained by feeding the energy consumption measurements from the training set. One of the major challenges in this stage was to determine the number of data records adequate for training. The dataset is a dynamic time series that varies depending on many different factors. For example, the consumer ECB profile may deviate based on daytime period, holidays, special occasions, weather seasons, and consumer habits. For this reason, to achieve quality results and high performance, the neural network structure should contain a minimal set of specific parameters of interest related to the energy measurements under investigation. In addition, while the neural network training dataset should be minimal to save on training time, at the same time, we need to make certain that there is adequacy of data to train on - otherwise the neural network will yield weak performance in terms of detecting energy fraud. After trial and error analysis, we determined that three weeks of consecutive data (in other words, $K = 48 \times 3$ as in Fig. 3) were adequate enough for training the neural network to build a profile for a particular consumer's energy consumption behavior over the period of time. Secondly, we simulated the scenario when a malicious actor implants a chip inside a smart meter, slowing down the meter readings for a random period of time and thus, resulting in deviations from actual energy consumption.

Two types of energy fraud discussed in the Introduction section were simulated by introducing random noise deviating from 0 to 0.5 kW/h (imitating the energy fraud type 1 described in section 1) and from -0.5 to 0 kW/h (imitating the energy fraud type 2). The ECB neural network was validated with both "pure" data (without noise) and "fraudulent" data (with noise).

The RMSE measurement was used to define the normal/fraudulent label for the ECB under investigation. If the RMSE value is within a certain threshold (in this scenario, we assigned 0.5 kW/h as the threshold), then the data were considered normal. Otherwise, the data were identified as a potential fraud.

B. Experimental results

We conducted three different types of experiments to provide a detailed insight into energy fraud detection.

1) Experiment 1

First, we assumed that consumer ECB can be predicted a year ahead. We trained the neural network on August, 2009 data and validated the fitted model on August, 2010. Then we repeated the experiment using data from other months (October and December). Experiment 1 results are presented in Fig. 4, which demonstrates that both types of energy fraud are above the threshold (the straight bold line across) and,

consequently, are denoted as fraudulent activities by this approach.

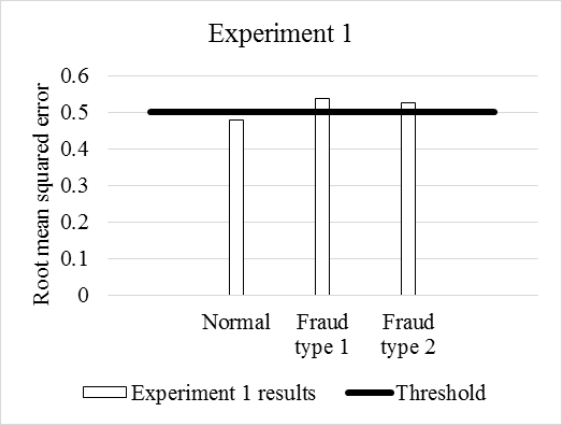

Fig. 4. Experiment on monthly models.

2) Experiment 2

The objective of the second experiment was to demonstrate if a neural network trained on three consecutive weeks is able to predict and analyse the following week to identify simulated fraudulent activities.

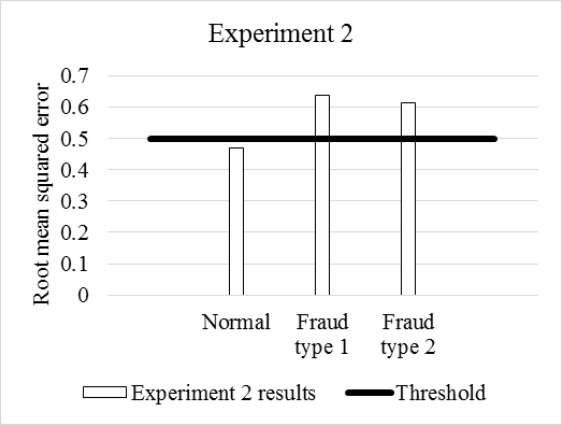

Fig. 5. Experiment on weekly models.

For instance, we utilized September 6 – 26, 2009 data to train the neural network and validated the produced model on the subsequent week (September 27 – October 3, 2009) energy measurements. Experiment 2 results are presented in Fig. 5. Both types of fraud were identified as their RMSEs are above the threshold.

3) Experiment 3

The third experiment set was aimed to examine if energy fraud can be detected in the same weather season. For evaluation, we trained the neural network with data from one of the three months of a particular weather season and, then validated the model with the two months of the same season. For example, we selected June, 2010 as the training data and validated the neural network with July and August, 2010. Experiment 3 results are presented in Fig. 6. Both types of fraud were identified as their RMSEs are above the threshold.

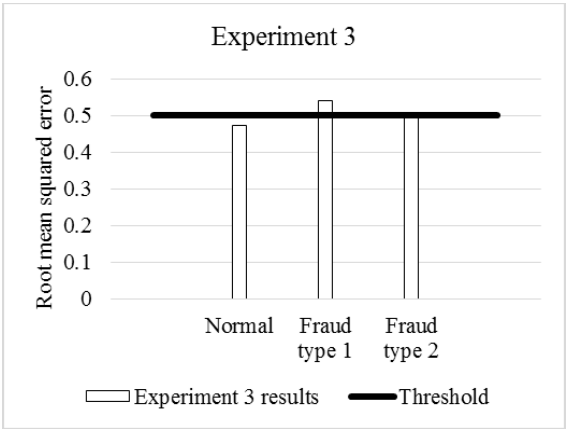

Fig. 6. Experiment on seasonal models.

C. Analysis of results

Table II shows the confusion matrix [5] created to evaluate the performance of the ECB neural network. The confusion matrix reports *true positives*, *true negatives*, *false positives*, and *false negatives*. It provides a mechanism for a detailed analysis, with more insight into the performance of the classifier than traditional accuracy metric. *True negatives* (TN) depict classifying normal data as normal activities. *False negatives* (FN) correspond to the rate of fraudulent activities classified as normal behavior. *True positives* (TP) shows the ratio of fraudulent activities correctly classified as fraud in the data. Lastly, *false positives* (FP) define the ratio of normal behavior classified as fraud.

TABLE II
CONFUSION MATRIX

TN	75.00%
TP	93.75%
FN	6.25%
FP	25.00%

Overall, the evaluation results demonstrate that energy fraud can be successfully detected for the same weather season, for the same month in a year, or even for subsequent weeks. In 93.75% cases (denoted by TP) the ECB neural network was able to detect simulated fraudulent activities or abnormal behavior in the energy consumption data. However, the *false positives* (25.00%) measurement reveals a relatively high rate of normal activities labelled by the neural network as fraud. This is caused by frequent changes in consumer energy consumption behavior due to weather seasons, holidays, new home appliances, or family vacations. A more comprehensive model of consumer energy consumption behavior would help to improve detection.

VI. CONCLUSION

This research demonstrates that a neural network can be used as a computational intelligent tool for detecting different types of energy fraud in customer energy consumption behavior. Real world historical energy consumption data from varied periods of time were used for analyzing and building

the energy consumption behavior profile and then the profile was used to identify abnormal or fraudulent activities.

Several improvements can be made as future work. First, we plan to tune the neural network parameters as well as its input layer structure to reduce the number of false negatives. Second, we plan to conduct more experiments by automating the neural network training and validation processes for analyzing different data sets concurrently. Third, comparison to other machine learning approaches can be done to develop a more robust fraud detection system. Fourth, the simulation process of fraudulent activities can be made more realistic.

ACKNOWLEDGMENT

This work is supported by the Center for Energy Systems Research of Tennessee Tech University. We are thankful to the Irish Social Science Data Archive Center for allowing access to the Smart Meter Data that was instrumental to this research.

REFERENCES

[1] National Institute of Standards and Technology, "Smart Grid Cyber Security Strategy and Requirements," *NISTIR 7628*, Vol. 1, August 2010.

[2] Commission for Energy Regulation, Irish Social Science Data Archive, *ucd.ie*. [Online]. [Accessed: March 5, 2013]. Available: http://www.ucd.ie/issda/data/commissionforenergyregulation/

[3] J. Han, M. Kamber, and J. Pei, *Data mining: Concepts and techniques*, 3rd ed., M. Kaufmann, Ed. MA, USA: Elsevier, 2012.

[4] The University of Waikato, "WEKA 3: Data Mining Software in Java," *cs.waikato.ac.nz*. [Online]. [Accessed: February 3, 2014].

[5] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote Sensing of Environment*, Vol. 62 (1), pp. 77–89, 1997.

[6] M. Jawurek, F. Kerschbaum, and G. Danezis, "SoK: Privacy Technologies for Smart Grids – A Survey of Options," *research.microsoft.com*. [Online]. [Accessed: April 22, 2014]. Available: http://research.microsoft.com/pubs/178055/paper.pdf

[7] P. Keung, J. Karel, and C. Bright, "Neural Networks for Insurance Fraud Detection," 2009, *uwaterloo.ca*. [Online]. [Accessed: May 12, 2014]. Available: https://cs.uwaterloo.ca/~cbright/reports/STAT840-project.pdf

[8] A. M. R. Serrano, J. P. C. L. da Costa, C. H. Cardonha, A. A. Fernandes, and R. T. de Sousa Junior, "Neural Network Predictor for Fraud Detection: A Study Case for the Federal Patrimony Department," 2012, *icofcs.org*. [Online]. [Accessed: May 12, 2014]. Available: http://www.icofcs.org/2012/ICoFCS2012_10.pdf

[9] S. Ghosh and D. L. Reilly, "Credit card fraud detection with a neural-network," *System Sciences,* in *Proceedings of the Twenty-Seventh Hawaii International Conference*, vol.3, pp. 621-630, Jan. 4-7, 1994.

[10] H. Verrelst, E. Lerouge, Y. Moreau, J. Vandewalle, C. Stormann, and P. Burge, "A rule based and neural network system for fraud detection in mobile communications," *ftp.cordis.europa.eu*. [Online]. [Accessed: May 12, 2014]. Available: ftp://ftp.cordis.europa.eu/pub/ist/docs/ka4/10396.pdf

[11] R. Patidar and L. Sharma, "Credit Card Fraud Detection Using Neural Network," *International Journal of Soft Computing and Engineering (IJSCE)*, Volume 1, Issue NCAI2011, pp. 32-38, June 2011.

[12] L.-J. Park, "Learning of Neural Networks for Fraud Detection Based on a Partial Area Under Curve," in *Proceedings of the Second International Symposium on Neural Networks*, Chongqing, China, pp. 922-927, May 30 - June 1, 2005.

[13] Microsoft, "Binary Delta Compression Technology," *download.microsoft.com*. [Online]. [Accessed: May 12, 2014]. Available: http://download.microsoft.com/download/e/1/c/e1c654c0-39a6-446d-bdfe-6d7a153c67ec/BDC_v2.doc

[14] S. Olmstead and A. Siraj, "Smart Grid Insecurity: A New Generation of Threats," in *Proceedings of the International Conference on Security and Management (SAM'11)*, Las Vegas, NV, July 18-21, 2011.

[15] M. P. Anderson and W. W. Woessner, *Applied Groundwater Modeling: Simulation of Flow and Advective Transport* (2nd ed.), Academic Press, 1992.

[16] J. W. Grzymala-Busse, L. K. Goodwin, W. J. Grzymala-Busse, and X. Zheng, "Handling Missing Attribute Values in preterm birth data sets," in *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, Springer-Verlag Berlin Heidelberg, pp. 342-351, 2005.

[17] J. E. Cabral, J. O. P. Pinto, E. M. Martins, and A. M. A. C. Pinto, "Fraud Detection in High Voltage Electricity Consumers Using Data Mining," in *Proceedings of the Transmission and Distribution Conference and Exposition*, pp. 1-5, April 2008.

[18] I. Monedero, F. Biscarri, C. Leon, J. Biscarri, and R. Millan, "MIDAS: Detection of Non-Technical Losses in Electrical Consumption Using Neural Networks and Statistical Techniques," in *Proceedings of the Computational Science and Its Applications Conference – ICCSA*, Vol. 3984, pp. 725-734, May 8-11, 2006.

[19] J. Nagi, K. S. Yap, S. K. Tiong, and S. K. Ahmed, "Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines," in *Proceedings of IEEE Transactions on Power Delivery*, Vol. 25, No. 2, pp. 1162-1171, April 2010.

[20] B. C. Costa, B. L. A. Alberto, A. M. Portela, W. Maduro, and E. O. Eler, "Fraud Detection in Electric Power Distribution Networks Using an ANN-Based Knowledge-Discovery Process," *International Journal of Artificial Intelligence & Applications*, Vol. 4, No. 6, pp. 17-23, November 2013.