**International Journal of Engineering Sciences Paradigms and Researches (IJESPR)**
**(Vol. 36, Issue 01) and (Publishing Month: November 2016)**
**(An Indexed, Referred and Impact Factor Journal)**
**ISSN: 2319-6564**
**www.ijesonline.com**

# CAN ACTIVE LEARNING BENEFIT THE SMART GRID? A PERSPECTIVE ON CONTROL THE DATA SHORTAGE

Archana Mandal
Aryan Institute of Engineering & Technology, Bhubaneswar

***Abstract*— In the earlier decade, a lot of tries were given to the field of working with an unrivaled insightful organization system by using the power of man-made intellectual ability. Beyond question, AI is as of now expecting an unavoidably huge part in essentially all aspects of power systems. Regardless, in certified practice, there is much greater proportion of unlabelled data than the one named by human subject matter experts. In this work, we try see focus on beating the data lack in insightful system. The unique learning system will be proposed to offer a possible response for keeping an eye on the data deficiency challenge. Furthermore, we will give a discussion on present status of-the-craftsmanship and the cutoff points in past work. We trust this work can be a fair helper for researchers to extra the material audit in the near future.*Index Terms*—Smart Grid, Power Systems, Active learning, Machine Learning, Weakly Supervised Learning.**

## I. INTRODUCTION

Machine learning (ML) has increasingly attracted efforts from both of the academia and industrial community of smart grid (SG) in the past decade [1]–[3]. Using these ML-based decision-making methods will enable automatic and autonomous control with involving minimal manual programming of the operation logic for power system [4]. Besides, the reduction of human participation in the system operation and monitoring meanwhile benefits the overall decision-making process with fast response, which facilitates the in-time tracking of system status and quick operation actions.

In line with this philosophy of automatic and autonomous decision-making, many subcategories of ML methods, such as supervised learning, unsupervised learning and reinforcement learning, have been widely used in theoretical study and practical deployment for smart grid applications [5]–[7]. However, the electric grid is a very complex system and one of the largest infrastructure in modern society, which is not possible to work reliably and independently without using any expert knowledge summarized in a systematical way. Combining the ML-enabled automatically generated information with the ancillary expert programmable knowledge is so far still the most reasonable solution to guarantee the system operation accuracy and effectiveness at the same time. Nonetheless, in real practice of SG (see Fig. 1) with generation side, dispatching side and consumption side, especially considering information flow with privacy concerns, accurately labelled data by human experts are extremely limited. Thereby, in this paper, we suggest a conceptual framework of active learning (AL) [8] for overcoming the various conventional difficulties in the smart grid problems with some but limited amount of human involvement. The AL methods had been successfully applied in many aspects in data mining projects, e. g., image retrieval [9], text classification [10], bioinformatics data classification [11], speech emotion recognition [12], and bird sound classification [13]. Motivated by the aforementioned successes, we want to introduce the AL concept and talk about the potentials and challenges in the applications of SG.

The main contributions of this work are: First, we discuss the opportunities and challenges in the area of ML for SG applications. Second, we propose the perspective on the AL for overcoming the data scarcity by introducing and describing the main methodologies. Third, we discuss the current limitations and point out future research directions for attracting more efforts to the relevant research community. The rest of this paper will be organised as follows: Firstly, Section II lists the challenges existing in typical applications of the AI-based SG domain, i. e., *smart building energy management*, *fault detection and diagnosis*, *cyber-security surveillance*, *distribution network operation* and *non-intrusive load monitoring*. Secondly, the AL method will be given and explained in Section III. Thirdly, we will have a discussion on current state-of-the-art and the limitations in Section IV before we make a conclusion in Section V.

## II. APPLICATIONS AND CHALLENGES

### A. Smart Building Energy Management

Most of the existing building energy management systems (BEMs) generally implement the entire building control as a whole and based on limited amount of human behavior

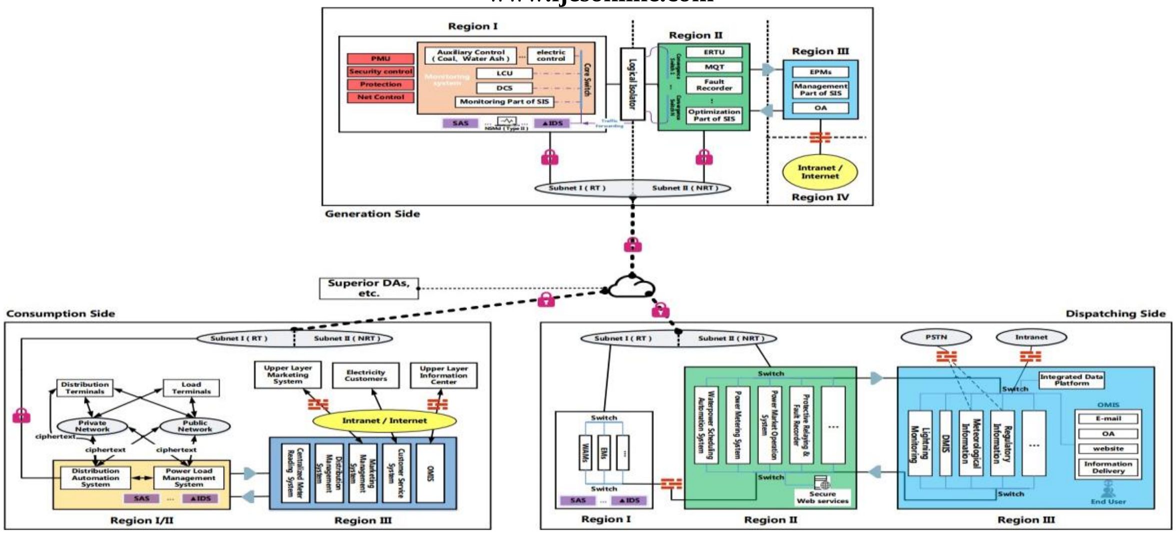


Fig. 1. Potential application scenario for obtaining optimal intrusion detection system with limited labelled data.

data [14]. The BEMs operators are usually unable to consider micro-zone comfort level in different areas (e. g., floors, rooms, corridors, etc.) according to occupancy levels and zone-specific dynamic environmental conditions. These as-a-whole building control strategies are also usually designed to only target the pure purpose of energy saving, without efficiently leveraging the increasing number of installed flexible building equipment (e. g., HVAC, lighting, electric heater, etc.) for grid-interactive service in the modernizing power grid.

In order to treat building as a living entity, which exchanges (transacts) energy with external environment, and make smart buildings more intelligent, occupancy comfortable, grid-interactive friendly and energy efficient, the AL strategies using limited amount of human interactive data will enable customizing the zone-specific operation strategy for building equipment by taking into account the zonal temperature, occupant satisfaction, amount of air flow, $CO_2$ levels, lighting levels and humidity in different areas. For example, the local HVAC thermostats, zone-specific lighting and plug loads can be adjusted according to the AL-enabled automatically labeled knowledge of occupancy level in a particular zone in a building for energy saving purpose [15]. Its noteworthy that by leveraging the AL technology for collection of human interactive data, "energy saving" in the context of smart buildings with zone-specific intelligent control strategy does not simply mean to use less energy or peak load shedding, but aim to avoid unnecessary energy consumption, in other words helping in improving the comfort level, and use energy more productively and efficiently.

### B. Fault Detection and Diagnosis

In Smart Grid environment, various types of faults are often existing there with harmful impact on the reliability and resilience of system operation. For decades research onthe fault detection and diagnosis for power system, many methods are proposed with more or less involvements of expert knowledge, including manual design of the crafted fault features for machine learning methods [16] [17] (see Fig. 2). In either monitoring data pre-processing stage or later on analysis stage, purely automatic and autonomous methods are usually hardly believed dependable without applying any human-defined rules. Therefore, by using a hybrid framework of semi-supervised learning technology, especially AL-based methods, the labeling and analysis process of fault detection and diagnosis for power system can be significantly simplified. For some rare event faults, we can even use a specific sparse- instance-based AL to overcome the data scarcity issue. More discussion can be found in Section III.

*C. Cyber-Security Surveillance*

Modern power system in smart grid and energy internet environment is keep continuing extending its components as cyber-physical-system (CPS), with many vulnerability and cyber-security issues [18]. Using ML technologies to deal with cyber-security surveillance in smart grid is a new trend and research paradigm for improving the system resilience [19], [20]. However, most works were using supervised learning and assuming some certain types of attacks predefined, such as Denial-of-service (DoS) attack, direct-access attacks, eaves- dropping attacks, and phishing attacks [21]. In practice, most attacks have heterogeneous sources and hardly known for their previous labeling information. Through application of AL- based methods, which provides *confidence score* information, the cyber-security surveillance problems can be transferred to an identification and probabilistic attack estimation problem using a mix of historical human labeled cyber-attack dataand real-time automatic monitoring data. In this way, we can make the best use of the experience collected from historical attacking samples and provide confidence score distribution

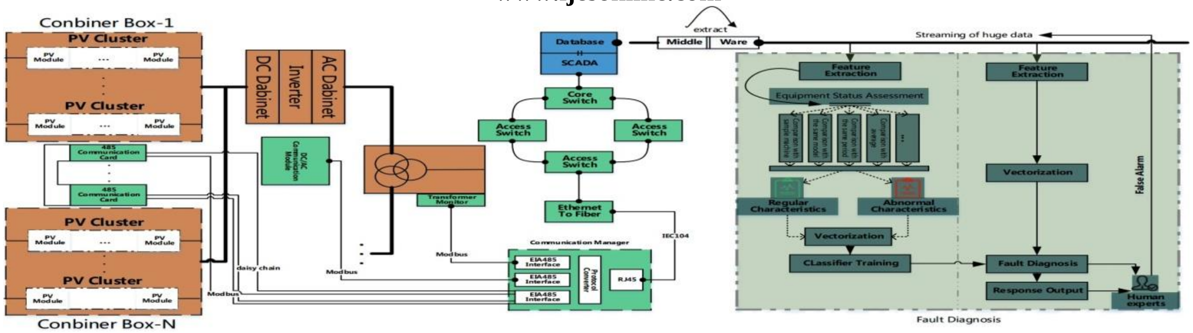


Fig. 2. Machine learning framework for fault detection diagnosis.

(similar to confidence interval) instead of only estimation of the attacking results.

*D. Distribution Network Operation*

Distribution network operation is a traditional problem for power system with involving various decision-making process, such as topology identification, microgrid reconfiguration, demand-side management and so on [22]. Most utilities or distribution system operators that are responsible for these problems have abundant operation data with the collection of historical information. Many operation problems are also solved via rule-based methods with expert judgement and manual analysis [23]. In most scenarios, ML-enabled methods hardly cover all the comprehensive operation conditions, even keeping updating and modifying the original operation rules. However, through AL-based approaches and semi-supervised learning framework, the summarized operational rules could be embedded in the automatic labeling loop in the AL algorithms (Section III). Immediate feedback in operating the distribution network can be obtained via membership query synthesis with human expert, and the delay-tolerated feedback could be processed at the end of pool-based sampling and training process [8]. It is believed the AL-based methods will be a promising solution candidate for shifting the conventional rule-based distribution network operation to a highly automatic and autonomous decision-making process.

*E. Non-intrusive Load Monitoring*

Non-intrusive load monitoring (NILM) is a way for homeowners and building managers to monitor energy consumption on an appliance-by-appliance basis without having to install dedicated sensors across an entire house or office building. NILM is not a new concept - the first non-intrusive devices date back almost twenty years to processes developed at MIT [24]. But those early devices were far from user-friendly: setting them up required installation by a trained electrician and a manual, one-by-one synchronization with every appliance in the household. Nowadays, with many advance in data analytic and machine learning technologies, especially by leveraging the new paradigm of AL strategy, the NILM implementation could be more accurate and efficient [25].

The proposed AL-based framework (see Fig. 3) mainly

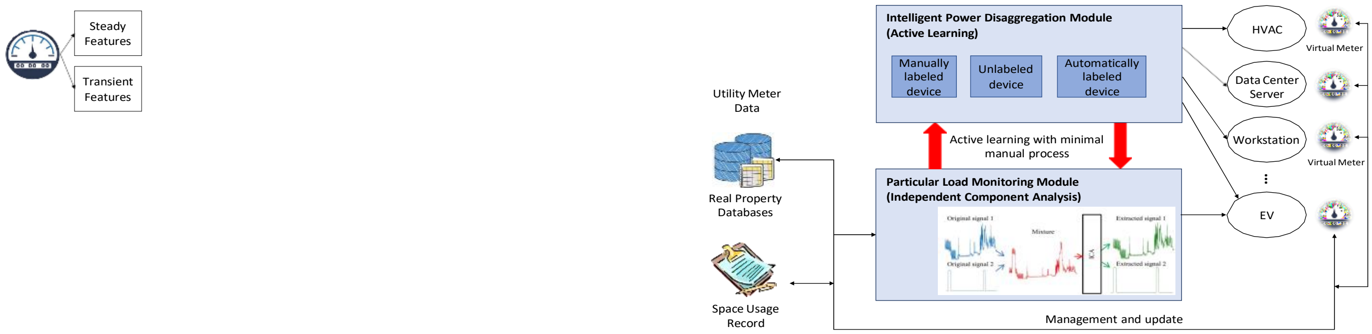


Fig. 3. Diagram of AL-based NILM System.

consists of two interdependent modules: intelligent power disaggregation (IPD) module and particular load monitoring module (PLM), which make best use of the collected utility meter data, real property data, space usage record and automat-ically labeled human behavior data to update the most recent findings or detection. These two modules can work either independently or coordinately for different tasks in different scenarios. Their functionality and operation are introduced respectively as follows:

a) IPD module uses AL to label the unknown appliance types using automatically generated features. In contrast to the conventional power disaggregation methods, like Hidden Markov Model and dictionary model, the AL methods can be waived from the handcrafted features and additional pro- gramming for feature extraction. Different utility meter data fidelity can determine how to choose the steady features (e.g. real power, reactive power, phase current) or transient features(e. g. voltage noise, start-up current) as input for IPD module.

b) PLM module mainly uses independent component anal- ysis (ICA) to extract load profile of certain particular electric appliance, like EV, from aggregated mixture energy consump-tion measurement. Meanwhile, the PLM module will also provide feedback to IPD AL module to readjust the power disaggregation result and help updating key findings, as well as significant parameters, in other disparage data sources (e. g. real property database) for management purpose.

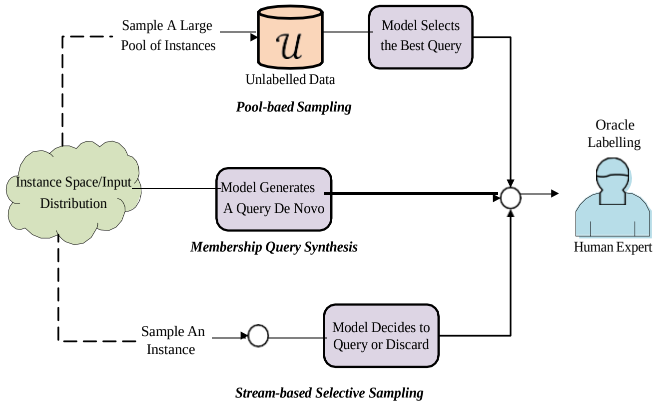


Fig. 4. Diagram of Three Main AL Scenarios [8].

---

**Algorithm 1:** PL

---

**repeat**

- Randomly select $N$ samples $D_N$ from the pool of unlabelled data $U$.
- Ask human expert to annotate $D_N$ .
- Remove $D_N$ from $U$: $U \leftarrow U \setminus D_N$.
- Add $D_N$ to the labelled data $L$: $L \leftarrow L \cup D_N$ .

**until** *iteration reaches a pre-defined number, or the trained model achieves an acceptable performance.*

---

## III. Active Learning

AL [8] was originally proposed for handling the issues raised from the fact that, large amount of data are unlabelled in nature. On one hand, it is always 'hungry for data' (specifically, the accurately labelled data) from an insight of machine learning community to build strong models. On the other hand, letting experienced human experts annotate the data is expensive, time-consuming, and even difficult in some areas which needs much domain knowledge and training, e. g., medicine, education, power system. In this line, one approach, i. e., semi-supervised learning (SSL) [26] was proposed to exploit the unlabelled data without involving any human expert's efforts. Nevertheless, annotation in power system needs sufficient human knowledge in real practice. Therefore, we introduce AL in this work.

The main mechanism of AL is to find the 'most informative' data in the unlabelled data. The initially trained weak learner can make a query to select those 'most informative' data for human experts to annotate (oracle labelling), which can efficiently improve the learner's performance on the corresponding task. There are several different AL scenarios for the learner to ask queries (see Fig. 4). In this work, we mainly discuss on the scenario of pool-based sampling, which maybe close to the real circumstance in the SG applications.

---

**Algorithm 2:** Sparse-Instance-based AL (SI-AL)

---

**repeat**

- Initially, train a model $C$ based on $L$.
- Randomly select $N$ samples $D_N$ from $U$ that are predicted by $C$ as belonging to the *sparse class*.
- Ask human expert to annotate $D_N$.
- Remove $D_N$ from $U$: $U \leftarrow U \setminus D_N$.
- Add $D_N$ to $L$: $L \leftarrow L \cup D_N$ .

**until** *iteration reaches a pre-defined number, or the trained model achieves an acceptable performance.*

---

**Algorithm 3:** Least-Confidence-Score-based AL (LCS-AL)

---

**repeat**

- Initially, train a classifier $C$ based on $L$.
- Predict $U$ by $C$ and rank the data by its prediction *confidence score*.
- Randomly select $N$ samples $D_N$ from the last $\epsilon N_U$ of the ranked data in $U$, $N_U$ is the number of instances in $U$, $\epsilon$ is a pre-defined parameter.
- Ask human expert to annotate $D_N$ .
- Remove $D_N$ from $U$: $U \leftarrow U \setminus D_N$.
- Add $D_N$ to $L$: $L \leftarrow L \cup D_N$ .

**until** *iteration reaches a pre-defined number, or the trained model achieves an acceptable performance.*

---

The counterpart of AL is passive learning (PL), which randomly select the unlabelled data to ask for human experts annotation (see Algorithm 1). In the paradigm of PL, thetrained learner (classifier) is not involved. Different with PL, AL fully considers involving the initially trained learner in finding the 'most informative' data. AL ask for human expert annotation by making an query which is based on a variety of strategies [27]. There are two typical strategies, i. e., sparse- instance-based AL (SI-AL), and least-confidence-score-based AL (LCS-AL) were demonstrated to be efficient in [12], [13], [28]. Among them, SI-AL thinks the data predicted tobe *sparse class* can be significant to improve the learner's performance (see Algorithm 2) while LCS-AL selects the data which achieves the lowest *confidence score* to be annotatedby human experts (see Algorithms 3). The detailed methods and definitions can found in Qian's doctoral thesis [29]. The capacity of the AL strategies are usually dependent on tasks. One possible future direction is to investigate and compare different AL strategies for their performances in specific SG applications.

## IV. DISCUSSION

### *A. Current Limitations*

Firstly, even though ML-based SG applications are increasingly attracting attentions from both of the academia and the research community, the available publicly accessible database is still extremely limited. In particular, limitation of the database restrains the development of advanced ML methods (e. g., deep learning [30]) in relevant studies. Besides, insufficient database cannot train an initial learner well, which will dramatically effect the following performance of AL in its paradigm. It is also difficult to make a fair and standard comparison between algorithms.

Secondly, the specific tasks that needing the AL approach are worth investigating deeply. In this paper, we firstly introduce our perspectives on overcoming the data scarcity challenge in SG applications. We need to take more real practices into account.

Thirdly, developing real-time products which involve the AL paradigm needs more sophisticated design. It can be seen that, only efficient and robust AL paradigm can be widely used in real SG applications.

### *B. Future Outlook*

We will make more comprehensive review of the literature and discuss in deeply about the AL in SG. Moreover, we will conduct some simulation experiments with data collected from different scenarios for showcasing the AL-enabled results. We will develop more advanced and applicable AL algorithms to improve the models' performances.

## V. Conclusion

In this paper, we introduced a perspective on AL for enabling the automatic and autonomous decision-making with overcoming the data scarcity in various SG applications. We discussed the opportunities and challenges existing in ML and AL for specific application problems, providing suggestion for the novel decision-making framework design. It is believed that AL-based methods combined with semi-supervised philosophy will enable one more solution candidate to deal with conventional SG problems, especially by leverage of the limited amount of expert knowledge. In addition, the main mechanism and typical strategies of AL were described for better understanding of its algorithm features. This work can facilitate relevant studies and bring more attention to the power energy research community.

## ACKNOWLEDGMENT

This work was partially supported by the JSPS Postdoctoral Fellowship for Research in Japan (ID No. P19081) from the Japan Society for the Promotion of Science (JSPS), Japan. X. Zha, K. Qian and T. Chen are the *Corresponding Authors*.

## References

[1] H. Jiang, J. J. Zhang, W. Gao, and Z. Wu, "Fault detection, identification, and location in smart grid based on data-driven computational methods," *IEEE Transactions on Smart Grid*, vol. 5, no. 6, pp. 2947–2956, 2014.

[2] M. Esmalifalak, L. Liu, N. Nguyen, R. Zheng, and Z. Han, "Detecting stealthy false data injection using machine learning in smart grid," *IEEE Systems Journal*, vol. 11, no. 3, pp. 1644–1652, 2014.

[3] M. Ozay, I. Esnaola, F. T. Y. Vural, S. R. Kulkarni, and H. V. Poor, "Machine learning methods for attack detection in the smart grid," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 8, pp. 1773–1786, 2015.

[4] M. Kandil, S. M. El-Debeiky, and N. Hasanien, "Long-term load forecasting for fast developing utility using a knowledge-based expert system," *IEEE Transactions on Power Systems*, vol. 17, no. 2, pp. 491–496, 2002.

[6] T. Zufferey, A. Ulbig, S. Koch, and G. Hug, "Unsupervised learning methods for power system data analysis," in *Big Data Application in Power Systems*. Elsevier, 2018, pp. 107–124.

[5] A. Asheibi, D. Stirling, and D. Sutanto, "Analyzing harmonic monitoring data using supervised and unsupervised learning," *IEEE Transactions on Power Delivery*, vol. 24, no. 1, pp. 293–301, 2008.

[7] D. Ernst, M. Glavic, and L. Wehenkel, "Power systems stability con- trol: reinforcement learning framework," *IEEE Transactions on Power Systems*, vol. 19, no. 1, pp. 427–435, 2004.

[8] B. Settles, "Active learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, 2012.

[9] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proc. ACM MM*. Ottawa, Canada: ACM, 2001, pp.107–118.

[10] S. Tong and D. Koller, "Support vector machine active learning with ap- plications to text classification," *Journal of Machine Learning Research*,vol. 2, no. 11, pp. 45–66, 2001.

[11] Y. Liu, "Active learning with support vector machine applied to gene ex- pression data for cancer classification," *Journal of Chemical Informationand Computer Sciences*, vol. 44, no. 6, pp. 1936–1941, 2004.

[12] Z. Zhang and B. Schuller, "Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition," in *Proc.INTERSPEECH*, Portland, OR, USA, 2012, pp. 362–365.

[13] K. Qian, Z. Zhang, A. Baird, and B. Schuller, "Active learning forbird sounds classification," *Acta Acustica united with Acustica*, vol. 103,no. 3, pp. 361–364, 2017.

[14] M. A. Hannan, M. Faisal, P. J. Ker, L. H. Mun, K. Parvin, T. M. I. Mahlia, and F. Blaabjerg, "A review of internet of energy based building energy management systems: Issues and recommendations," *IEEE Access*, vol. 6, pp. 38 997–39 014, 2018.

[15] Y. Peng, A. Rysanek, Z. Nagy, and A. Schlüter, "Using machine learning techniques for occupancy-prediction-based cooling control in office buildings," *Applied Energy*, vol. 211, pp. 1343–1358, 2018.

[16] J. Cordova, C. Soto, M. Gilanifar, Y. Zhou, A. Srivastava, and R. Arghan- deh, "Shape preserving incremental learning for power systems fault detection," *IEEE Control Systems Letters*, vol. 3, no. 1, pp. 85–90, 2018.

[17] T. Chen, A. Mutanen, P. Järventausta, and H. Koivisto, "Change detec- tion of electric customer behavior based on amr measurements," in *Proc.Eindhoven PowerTech*. IEEE, 2015, pp. 1–6.

[18] C.-C. Sun, A. Hahn, and C.-C. Liu, "Cyber security of a power grid: State-of-the-art," *International Journal of Electrical Power & Energy Systems*, vol. 99, pp. 45–56, 2018.

[19] H. Karimipour, A. Dehghantanha, R. Parizi, K. Choo, and H. Leung, "A deep and scalable unsupervised machine learning system for cyber-attack detection in large-scale smart grids," *IEEE Access*, 2019.

[20] Y. Cui, F. Bai, Y. Liu, and Y. Liu, "A measurement source authentication methodology for power system cyber security enhancement," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 3914–3916, 2018.

[21] K. Pan, A. Teixeira, M. Cvetkovic, and P. Palensky, "Cyber risk analysis of combined data attacks against power system state estimation," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3044–3056, 2018.

[22] L. Ma, T. Zhu, H. Gao, Z. Lu, and J. Qi, "Evaluation and sensitivity analysis for economical operation of active distribution network," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 13, no. 5, pp. 715–724, 2018.

[23] F. Bellizio, S. Karagiannopoulos, P. Aristidou, and G. Hug, "Optimized local control for active distribution grids using machine learning tech-niques," in *Proc. PES GM*. Portland, OR, USA: IEEE, 2018, pp. 1–5.

[24] G. W. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[25] A. Zoha, A. Gluhak, M. Imran, and S. Rajasegarar, "Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey," *Sensors*, vol. 12, no. 12, pp. 16 838–16 866, 2012.

[26] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 3, no. 1, pp. 1–130, 2009.

[27] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Madison, WI, USA, Computer Sciences Technical Report 1648, 2009.

[28] K. Qian, Z. Zhang, A. Baird, and B. Schuller, "Active learning for bird sound classification via a kernel-based extreme learning machine," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 1796–1804, 2017.

[29] K. Qian, *Automatic General Audio Signal Classification*. Munich, Germany: Technical University of Munich, 2018, Doctoral Thesis.

[30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.