D. Pliatsios, P. Sarigiannidis, K. Psannis, S. K. Goudos, V. Vitsas, and I. Moscholios, "Big Data against Security Threats: The SPEAR Intrusion Detection System," presented at the 2020 3rd World Symposium on Communication Engineering (WSCE), Oct. 2020

This is the author's preprint version, the published version is available at the IEEE Xplore Library: https://ieeexplore.ieee.org/document/9275580

# Big Data Against Security Threats: The SPEAR Intrusion Detection System

Dimitrios Pliatsios[†], Panagiotis Sarigiannidis[†]✉, Konstantinos Psannis[‡], Sotirios K. Goudos[§], Vasileios Vitsas[¶] and Ioannis Moscholios[‖]

[†] Department of Electrical & Computer Engineering, University of Western Macedonia
50100, Kozani, Greece, {dpliatsios, psarigiannidis}@uowm.gr

[‡]Department of Applied Informatics, University of Macedonia
54636, Thessaloniki, Greece, kpsannis@uom.edu.gr

[§]School of Physics, Aristotle University of Thessaloniki
54124, Thessaloniki, Greece, sgoudo@physics.auth.gr

[¶]Department of Information and Electronic Systems Engineering, International Hellenic University
57400, Thessaloniki, Greece, vitsas@it.teithe.gr

[‖]Department of Informatics & Telecommunications, University of Peloponnese
22100, Tripolis, Greece, idm@uop.gr

***Abstract*—The environmental concerns, the limited availability of conventional energy sources, the integration of alternative energy sources and the increasing number of power-demanding appliances change the way electricity is generated and distributed. Smart Grid (SG) is an appealing concept, which was developed in response to the emerging issues of electricity generation and distribution. By leveraging the latest advancements of Information and Communication Technologies (ICT), it offers significant benefits to energy providers, retailers and consumers. Nevertheless, SG is vulnerable to cyber attacks, that could cause critical economic and ecological consequences. Traditional Intrusion Detection Systems (IDSs) are becoming less efficient in detecting and mitigating cyberattacks, due to their limited capabilities of analyzing the exponentially increasing volume of network traffic. In this paper, we present the Secure and PrivatE smArt gRid (SPEAR) platform, which features a Big Data enabled IDS that timely detects and identifies cyber attacks against SG components. In order to validate the efficiency of the SPEAR platform regarding the protection of critical infrastructure, we installed the platform in a small wind power plant.**

***Index Terms*—Big data, Cyber attack, Intrusion detection system, Smart grid.**

## I. Introduction

Electricity is one of the greatest discoveries of the 19th century that led to a revolutionary progression of economy and society. Nowadays, it is considered as a fundamental commodity and the most widely used form of energy as it can be transferred in very long distances. The electric grid is a massive interconnected network that is used to deliver electricity to consumers.

Due to environmental concerns and limited availability of conventional energy sources (i.e., coal, gas and oil), the demand for cleaner energy and more efficient use of the current one increases. Power-demanding appliances, such as entertainment systems, heating, ventilation and air conditioning systems and electric vehicles heavily affect the rate of power generation. Moreover, the increasing demand at peak times requires more power sources in order to avoid declines in power quality and blackouts. These challenges highlight the need to reassess how electricity is generated and distributed. Smart Grid (SG) is an appealing novel paradigm in response to the aforementioned issues [1]. SG aims to intelligently coordinate the behaviors and actions of all stakeholders involved in energy generation and supply, in order to efficiently deliver economic and environmental-friendly energy.

SG faces security challenges similar to typical computer networks since all devices involved in SG are susceptible to cyber-attacks [2], [3]. Threats against SG are versatile and sophisticated, such as unauthorized intrusion, Denial of Service (DoS), physical damage to grid and devices, confidential data theft and market fraud. Conventional threat and attack mitigation methods such as traffic and flow analysis, network forensics and intrusion detection techniques have limitations and are not always adequate to support large-scale networks and state-of-the-art attacks. The incorporation of machine learning techniques is an established approach to enhance the aforementioned methods and remove their limitations [4].

In this work, we present the Secure and PrivatE smArt gRid (SPEAR) platform, which features an Intrusion Detection System (IDS) tailored to the security requirements of the SG. In order to efficiently manage the huge volume of data generated from the SG devices, Big Data analytics are integrated into the IDS. Furthermore, a Visual-based Intrusion Detection System (V-IDS) is leveraged in order to alert the SG administrators, provide information about the cyberattack, and facilitate the decision-making process. As a result, the SPEAR solution is capable of timely detecting and identifying cyberattacks in a SG infrastructure, composed of large numbers of SG devices.

The paper is structured as follows. Section II provides the background for this work, by providing an overview of the SG concept and the Big Data analytics. In Section III, we present the SPEAR platform that aims to provide a Big Data-enabled threat detection solution, tailored to the security requirements

of the Smart Grid. In Section IV, we discuss the integration of the proposed solution into a realistic scenario that features a small wind power plant. Finally, we conclude this work in Section V.

## II. Background

### A. Related Work

Due to their advantages, Big Data analytics have been widely used for the detection and identification of cyberattacks in various environments. The authors in [5] presented a novel malware behavioral classifier based on the Big Data methodology. The dataset was built by collecting network traffic from an ISP. The authors selected specific dataset features to base their classification. Various classification algorithms were used such as the J48 and Random Forest.

Rathore et al. [6] developed a real-time IDS that works in an ultra-high speed Big Data environment. They extracted features from DARPA, KDD99 and NSL-KDD datasets by using Forward Selection Ranking and Backward Elimination Ranking algorithms. It utilized Naive Bayes, Support Vector Machine, Random Forest, Conjunctive Rule, REPTree and J48 algorithms from the Spark machine learning library.

The authors in [7] proposed a fast and scalable Hadoop-based IDS for large volumes of data. KDD99 and 10%-KDD99 were used to build the dataset. The system utilized a parallel version of Binary Bat algorithm for feature selection and Naive Bayes algorithm for classification. Additionally, the authors evaluated the system in terms of attack type accuracy.

In [8], the authors designed an Apache Spark based solution for detecting intrusions in SG. The dataset consists of measurements taken by Phasor Measurement Units in a 2-line, 3-phase power system. They experimented with both Correlation based Feature Selection and Principal Component Analysis (PCA) in order to reduce the number of dataset features. Deep Neural Networks, Support Vector Machine, Naive Bayes, Decision Trees and Random Forest algorithms were utilized for classification. The performance of these algorithms was compared considering the raw dataset as well as the dataset with reduced features.

Gupta and Rani [9] proposed a scalable framework for zero-day malware detection. Their dataset includes 200.000 samples (50.000 clean files and 150.000 malware files) built from VxHeaven, Nothing and Virus Share. A customized Logistic Regression was used to extract the most important features. Towards classification, they used Naive Bayes, Random Forest and Support Vector Machine algorithms from Apache Spark.

The authors in [10] developed an anomaly-based IDS for the SG that utilizes operational data from a real power plant. For the anomaly detection, several machine learning methods utilized, such as One Class-Support Vector Machine, Isolation Forests, Angle-based and Stochastic-based Outlier Detection, Principal Component Analysis, and deep fully connected Autoencoders. The evaluation results show the efficacy of the proposed approach and the improvement in the detection accuracy, due to the complex data representation that is utilized in the proposed IDS.

### B. The Smart Grid Concept

The SG concept emerged in order to address the shortcomings of the existing power grid. Compared to traditional power grid, smart grid has improved energy efficiency, reliability and superior integration with alternative energy sources. These features are realized by leveraging advanced sensor and measurement technologies, modern communication and information techniques and novel computing and control algorithms [11], [12].

SG provides significant advantages to all the involved stakeholders, such as energy providers, retailers and consumers. Energy providers are involved in the generation and distribution of the power from the power plants to the customers. SG provides tools and techniques to the providers for accurately detecting outages and equipment failures, thus reducing operation and maintenance costs. Power grid planning [13] leads to improved power generation and distribution efficiency. The continuous monitor of the power generation and consumption, effectively reduces energy waste. Finally, SG seamlessly integrates alternative energy sources to the grid [14]. Furthermore, SG enables retailers to maximize their revenue, through providing better knowledge of the energy market. SG collects detailed information about the customers, which allow retailers to compose customer profiles and better plan their economic strategy [15]. Finally, customers can monitor their power consumption in real time and better plan their power usage and take advantage of customized billing plans to reduce their energy charges.

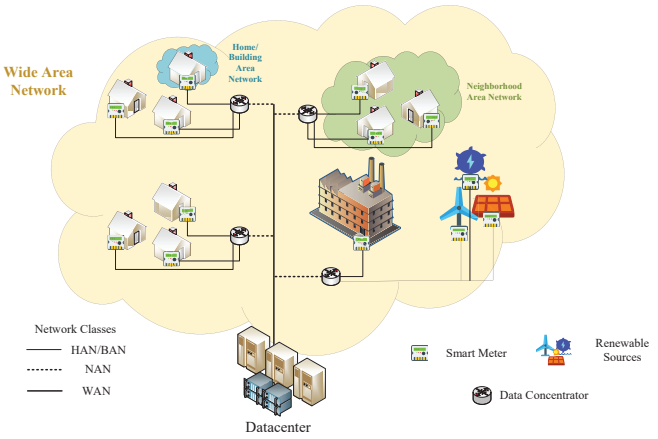


Fig. 1. AMI Hierarchical Architecture

### C. Smart Grid Architecture

The Advanced Metering Infrastructure (AMI) paradigm is a combination of technologies, which is responsible for data collection, transfer and management [16]. In the AMI, a large number of smart meters is deployed in every facet of SG. Specifically, the smart meters are deployed in power plants, throughout the distribution network and in the consumer premises. A smart meter is an electronic device capable of measurement and data transmission in short time intervals.

The hierarchical architecture of the AMI is shown in Fig.1, where a WAN consisting of smaller NANs and HANs is illustrated. In each building, a smart meter is installed sending consumption information to a data concentrator. Smart meters are also installed throughout the SG infrastructure such as power plants, distribution network and renewable energy sources. Then, the data from the concentrators are sent to a datacenter for further processing.

The Home/Building Area Network (HAN/BAN) covers the area closest to the user premises and involves sending smart meter data and receiving control commands. Short range and low data-rate communications, such as ZigBee, Wireless Fidelity (WiFi), Bluetooth and Power Line Communications (PLCOM) are common technologies used in HAN/BAN.

The Neighborhood Area Network (NAN) is responsible for transferring data from the user premises to either data concentrators or substations. Thus, it requires communication technologies with larger range and higher data rates, such as cellular communication technologies, Worldwide Interoperability for Microwave Access (WIMAX) and Digital Subscriber Line (DSL).

Finally, the Wide Area Network (WAN) is the top class in the hierarchical architecture of AMI. It transfers vast amounts of data as data from HANs and NANs are aggregated in this class. Therefore, it requires long range coverage and very high data rates. Optical technologies and cellular communication technologies can be used effectively for this class. Also, satellite communications can be used for critical sections of the network.

### D. Big Data Analytics

Big Data is an abstract concept that refers to data sets which are too large for conventional data technologies to capture, manage and process [17]. The 4Vs model, namely Volume, Velocity, Variety and Value is the most common model to define the Big Data. Volume refers to the collection of massive data from multiple sources. Velocity indicates the speed at which the data are generated and processed. Variety describes the various data types (i.e., structured and unstructured data, text and video). Value is the insights extracted from these data (e.g, discovering a trend, spotting cost reduction from a different logistical perspective).

Data analysis is a process that aims to extract useful information from datasets. The most widely used data analysis methods include statistical analysis, regression and correlation analyses, and cluster analysis. Statistical theory is the basis of statistical analysis. Descriptive statistics can summarize datasets using various indexes (e.g., mean and deviation). Inferential statistics draw conclusions from data subjected to random variations. Regression and correlation analyses are often used mutually. Regression analysis involves a number of processes for discovering relationships between variables. Specifically, it reveals how a specific variable changes, while other variables are modified. On the other hand, correlation analysis attempts to quantify the association between variables, using correlation coefficients (e.g., Pearson product-moment, Spearman's rank and Kendall's rank). Finally, cluster analysis aims to classify a set of objects based on some features. A cluster is a group of classified objects that have high homogeneity. Common clustering algorithms include k-means clustering, hierarchical clustering, fuzzy c-means clustering and Gaussian clustering.

## III. The Secure and PrivatE smArt gRid (SPEAR) Project

### A. The SPEAR Concept

The ever-increasing integration of critical power infrastructure to the SG [18], along with the proliferation of sophisticated cyberattacks [19], call for the development of even more advanced IDS. To this end, we introduce the Secure and PrivatE smArt gRid (SPEAR) research project, which is funded by the Horizon 2020 framework programme of the European Union. One of its main objectives is to develop methods, processes, and tools that leverage the aforementioned concepts in order to accurately and timely detect novel smart grid threats, such as Advanced Persistent Threats, DoS and Distributed Denial of Servce (DDoS) attacks.

Within SPEAR, cybersecurity is be considered in all domains, components, and subsystems of the smart grid. The security platform provides an integrated security solution for satisfying the SG security requirements by timely detecting cyber threats against the AMI. SPEAR brings an added value by introducing a second level of defense in the Security Information and Event Management (SIEM) tools, where Big Data analytics in conjunction with visual-aided IDS will timely detect a stealth cyberattack, if it is missed by the basic SIEM tool due to lack of processing time. In this respect, SPEAR supports an enriched anomaly detection system, minimizing the time needed for detecting sophisticated cyber-attacks.

### B. SPEAR Platform Architecture

The architecture of the SPEAR platform is shown in Fig. 2. It consists of five main components, namely the Data Acquisition Parsing and Storage (DAPS), the Big Data Analytics (BDA), the Message Bus, and the Visual-based Intrusion Detection System (VIDS). Table I lists the main information elements utilized in the SPEAR platform, namely the Raw Smart Grid Data, the Processed Smart Grid Data, and the Security Events. Additionally, Table II summarizes the interactions between the platform's components, along with a brief description of each interaction.

The DAPS component is responsible for acquiring and storing all data from the Network Capturer and Parser (NCP) instances. The NCP instances are deployed on the SG elements, and they monitor and capture the network traffic. The pre-processed data are forwarded to the BDA and VIDS components. The DAPS component utilizes the following tools: T-shark network analyzer, CICFlowMeter flow analyzer, Elastic Filebeat, Apache Kafka, Elastic Logstash, and Elasticsearch.

The BDA component carries out machine learning and deep learning processing in order to detect potential cyber-attacks and anomalies. Specifically, the BDA component con-

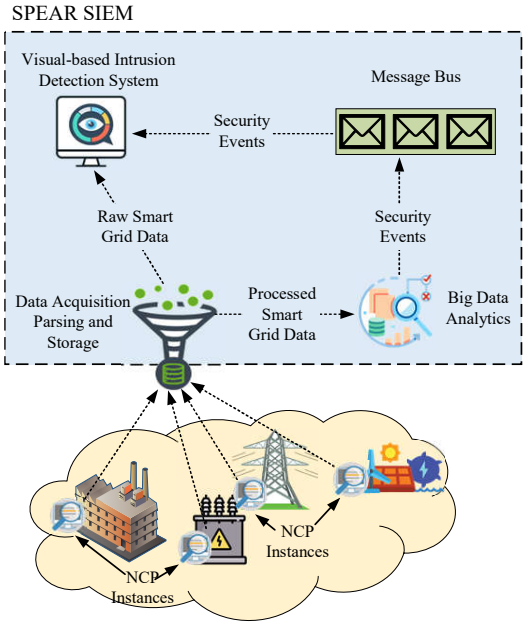


Fig. 2. SPEAR Platform Architecture

TABLE I
DESCRIPTION OF THE SPEAR INFORMATION ITEMS

| Item | Description |
|---|---|
| Raw Data | Raw data refer to representations that can be recognized, produced, or processed by a software asset of the SG. |
| Processed Data | DAPS collects SG data and structures them appropriately for Big Data analytics |
| Security Events | Security events are asynchronous security incidents that involve a security violation of the SG, and they detected using advanced data analytics techniques |

stitutes an anomaly-based IDS that detects cyber-attacks and anomalies by analyzing network flows, application layer data, and operational data. In particular, BDA is able to detect potential anomalies and cyberattacks based on the network traffic, transport and application layers of the Open Systems Interconnection (OSI) model, as well as specific SG values such as current, voltage and phase. Additionally, it can handle huge volumes of data that originate from multiple sources. After the analysis, the corresponding security events are generated and forwarded to the Message Bus. The BDA component utilizes Python-based machine learning libraries, such as Numpy, Pandas, Scikit-learn, Tensorflow, Keras, and PyTorch in order to construct efficient anomaly detection models.

The Message Bus component is based on the Apache Kafka, and it enables the communication among all the SPEAR SIEM components that exchange security events. This component handles the asynchronous exchange of events, based on the publish-subscribe paradigm.

Finally, the VIDS offers an intuitive visual approach of potential threats to the administrator, facilitating the detection of anomalies. The VIDS receives the security events from the Message Bus and utilizes visualization methods to generate graphs from these events. In addition, it offers intuitive graphical user interfaces, which display the results in both web-based and mobile-based applications. Finally, all these interfaces generate notifications, in order to enable the fast reaction from the administrators.

## IV. RENEWABLE ENERGY PLANT CASE STUDY

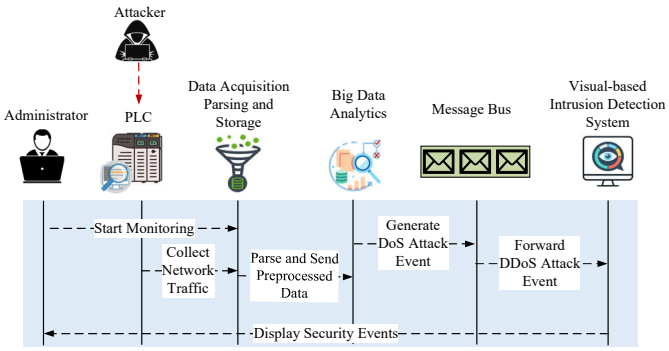


Fig. 3. Renewable Energy Plant Case Study

This subsection includes a realistic applied scenario, where a small wind power plant is considered, following the European Program for Critical infrastructure Protection challenges [20]. Within the context of the SPEAR research project, a small wind power plant was selected. The wind power plant features some unique characteristics such as a) it constitutes a high-technology power plant, where the cyberattacks can result in devastating consequences, b) it represents an example of a renewable energy utility, c) it constitutes a roadmap in order to evaluate the SPEAR architecture towards securing renewable energy smart grid utilities.

The wind power plant infrastructure includes turbines, generators, electrical systems, and transformers. The infrastructure's monitoring and control equipment is controlled by Programmable Logic Controllers (PLCs) through network interfaces. The SPEAR SIEM tool is installed in the plant's control center, while a number of NCP instances are deployed to the infrastructure.

The SPEAR SIEM tool monitors the traffic to/from the wind power plant to detect any suspicious activity. Specifically, the system will monitor the Human Machine Interface (HMI) and the smart devices that are used for monitoring and controlling the power generation process. The BDA and VIDS components, will respectively detect anomalies in the data packets and report them to the administrator.

Fig. 3 shows an attack detection scenario, where a hacker launches a DoS attack against a PLC . The SG administrator uses the DAPS to collect the data generated from the SG and captured by the NCP instances. The collected SG data include mainly network traffic and log data, produced by the SG elements. The DAPS also pre-processes the collected data, by parsing the communication protocols traffic, as well as the operational information included in the payload.

TABLE II
SUMMARY OF COMPONENT INTERACTIONS

| Component 1 | Component 2 | Interaction Description |
|---|---|---|
| Data Acquisition, Parsing, and Storage | Big Data Analytics | DAPS provides the parsed SG data to the BDA, in order for the BDA to perform data analytics to detect incidents and anomalies in the SG |
| Data Acquisition, Parsing, and Storage | Visual-based Intrusion Detection System | DAPS provides the parsed SG data to the VIDS, in order to visualize the SG status |
| Big Data Analytics | Message Bus | BDAC generates the security events and forwards them to the Message Bus in order for them to be distributed to the VIDS |
| Message Bus | Visual-based Intrusion Detection System | VIDS receives the security events from the Message Bus and provides a visual representation to the SG administrators |

The BDA retrieves the pre-processed SG data and applies near real-time data analytics in order to detect anomalies in the SG. In this demonstration scenario, the DDoS attack is detected and the corresponding security event is generated. The security event is forwarded to the Message Bus, in order to be received by the rest of the components.

Lastly, the VIDS retrieves the security event from the Message Bus and visualizes it in order to support the decision making process.

## V. CONCLUSION

Smart Grid utilizes advanced measuring techniques along with communication technologies in order to improve its efficiency and reliability, reduces carbon emissions by integrating green energy resources and provides minimized operating and maintenance costs. Nevertheless, Smart Grid is facing security challenges and is vulnerable to cyberattacks. The leverage of Big Data analytics in intrusion detection is a promising method in defending and mitigating new and sophisticated cyberattacks.

In this work, we presented the concept and the architecture of the SPEAR platform, which leverages Big Data analytics to deliver a novel threat detection and identification platform, tailored to the SG security requirements. Finally, a wind power plant was selected in order to validate the efficiency of the SPEAR platform in the protection of critical infrastructure.

## ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 787011 (SPEAR).

## REFERENCES

[1] G. Dileep, "A survey on smart grid technologies and applications," *Renewable Energy*, vol. 146, pp. 2589–2625, 2020.

[2] P. I. Radoglou-Grammatikis and P. G. Sarigiannidis, "Securing the smart grid: A comprehensive compilation of intrusion detection and prevention systems," *IEEE Access*, vol. 7, pp. 46 595–46 620, 2019.

[3] D. Pliatsios, P. Sarigiannidis, T. Lagkas, and A. G. Sarigiannidis, "A survey on scada systems: Secure protocols, incidents, threats and tactics," *IEEE Communications Surveys & Tutorials*, 2020.

[4] 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, 2016.

[5] E. Bocchi, L. Grimaudo, M. Mellia, E. Baralis, S. Saha, S. Miskovic, G. Modelo-Howard, and S.-J. Lee, "Magma network behavior classifier for malware traffic," *Computer Networks*, vol. 109, pp. 142–156, 2016.

[6] M. M. Rathore, A. Ahmad, and A. Paul, "Real time intrusion detection system for ultra-high-speed big data environments," *The Journal of Supercomputing*, vol. 72, no. 9, pp. 3489–3510, 2016.

[7] P. Natesan, R. Rajalaxmi, G. Gowrison, and P. Balasubramanie, "Hadoop based parallel binary bat algorithm for network intrusion detection," *International Journal of Parallel Programming*, vol. 45, no. 5, pp. 1194–1213, 2017.

[8] K. Vimalkumar and N. Radhika, "A big data framework for intrusion detection in smart grids using apache spark," in *Advances in Computing, Communications and Informatics (ICACCI), 2017 International Conference on*. IEEE, 2017, pp. 198–204.

[9] D. Gupta and R. Rani, "Big Data Framework for Zero-Day Malware Detection," *Cybernetics and Systems*, pp. 1–19, 2018.

[10] G. Efstathopoulos, P. R. Grammatikis, P. Sarigiannidis, V. Argyriou, A. Sarigiannidis, K. Stamatakis, M. K. Angelopoulos, and S. K. Athanasopoulos, "Operational data based intrusion detection system for smart grid," in *2019 IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*. IEEE, 2019, pp. 1–6.

[11] N. S. Nafi, K. Ahmed, M. A. Gregory, and M. Datta, "A survey of smart grid architectures, applications, benefits and standardization," *Journal of Network and Computer Applications*, vol. 76, pp. 23–36, 2016.

[12] S. K. Goudos, P. Sarigiannidis, P. I. Dallas, and S. Kyriazakos, "Communication protocols for the iot-based smart grid," in *IoT for Smart Grids*. Springer, 2019, pp. 55–83.

[13] B. Nasiri, C. Wagner, U. Häger, and C. Rehtanz, "Distribution grid planning considering smart grid technologies," *CIRED-Open Access Proceedings Journal*, vol. 2017, no. 1, pp. 2228–2232, 2017.

[14] N. Javaid, G. Hafeez, S. Iqbal, N. Alrajeh, M. S. Alabed, and M. Guizani, "Energy efficient integration of renewable energy sources in the smart grid for demand side management," *IEEE Access*, vol. 6, pp. 77 077–77 096, 2018.

[15] C. M. Cheung, S. R. Kuppannagari, R. Kannan, and V. K. Prasanna, "Load demand user profiling in smart grids with distributed solar generation," in *2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*. IEEE, 2020, pp. 1–5.

[16] A. Hansen, J. Staggs, and S. Shenoi, "Security analysis of an advanced metering infrastructure," *International Journal of Critical Infrastructure Protection*, vol. 18, pp. 3–19, 2017.

[17] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, Jan 2014.

[18] H. Farhangi, "A road map to integration: Perspectives on smart grid development," *IEEE Power and Energy Magazine*, vol. 12, no. 3, pp. 52–66, 2014.

[19] Y. Tang, Q. Chen, M. Li, Q. Wang, M. Ni, and X. Fu, "Challenge and evolution of cyber attacks in cyber physical power system," in *Power and Energy Engineering Conference (APPEEC), 2016 IEEE PES Asia-Pacific*. IEEE, 2016, pp. 857–862.

[20] European Program for Critical Infrastructure Protection (EPCIP). Protection of critical infrastructure. [Online]. Available: https://ec.europa.eu/energy/en/topics/infrastructure/protection-critical-infrastructure