An Efficient Cyber Attack Detection Using Machine Learning Techniques

S.S.Vinisha [1], G.Johncy [2]

[1] M.E. Student, Department of Computer Science & Engineering, St.Xavier's Catholic College of Engineering, Nagercoil, Tamil Nadu, India.

[2] Assistant Professor, Department of Computer Science & Engineering, St.Xavier's Catholic College of Engineering, Nagercoil, Tamil Nadu, India.

ABSTRACT

A new generation of technology has evolved which allows for the transmission of data between a utility company and its customers in real-time through new technologies such as Advanced Metering Infrastructure. The security and privacy of smart grid systems, which combine smart and legacy information and operational technologies, have grown in concern. We propose an information attack detection model for the smart grid based on XGBoost. It uses a modified k-means-smote oversampling method to obtain a balanced power data set, which solves the problem of data imbalance causing high false-positive rates in network attack detection. On the basis of oversampling data, feature selection is performed to reduce the dimension of the data. This will shorten the model training time, and accelerate the response speed of the network attack detection model. Finally, construct an XGBoost classifier model to identify several network attack modes in the data set. The paper studies machine learning models and proves that the network attack detection model improves the detection accuracy of smart grid information attacks significantly.

Keywords: *Smart grid, Machine Learning, Network Attack, XGBoost Algorithm.*

1.INTRODUCTION

Unlike other commodities like oil, food, etc. electricity cannot be stored for future supply. Electricity is one of the most important sources of energy that is being produced from electricity generating units and it is the fast-growing form of end-use energy consumption. Electricity networks are a pivotal element in erecting smart megacity infrastructures. Thus are a fundamental commodity that must be generated as per needed application which cannot be stored at large scales. The product cost heavily depends upon the source similar as hydroelectric power shops, petroleum products, nuclear and wind energy. In 2040 the global demand for electricity in the domestic sector will be projected to increase 1.4% [1] [2]. With the rapid growth of population and increased number of industries, it has become a difficult task for the grids to manage the demand of the electricity for household purposes and industrial purposes. The increased demand of electricity at particular hours of the day lead to several problems like short circuits, failure of transformers.

One of the crucial technologies of smart grid is the Advanced Metering Infrastructure (AMI) which depends on the implementation of smart home. In big data era, users have advanced prospects for network performance and the service quality [3]. In AMI, two-way communication between the utility and therefore the customers are achieved with the support of data and communication technologies. Electric utilities with reliable power services are widely located by the Smart meters (SMs) [4]. The Internet of things (IoT) relies on interconnected smart devices, thus allows the smart devices to gather sensitive information and carry out important functions, and these devices connect and communicate with each other at high speeds and make decisions according to indicator information. IoT can provide new criteria for sharing data and information into the whole grid, such as multi-hop communication, uses large numbers of sensors to extract significant information, and this information is analysed by artificial intelligence algorithms [5][6].

Smart grid is one of typical cyber-physical systems (CPSs), achieves the interactive integration in energy, power, communication and control. It consists of a set of computers, controllers, automation, and standard communication protocols, which are connected on the Internet. With these above advanced technologies, smart grid has significantly enhanced the intelligence and reliability in generation, transmission, distribution and consumption. Smart digital is one of the technologies that allows two-way communication between the utility and its customers, the detection of long distribution lines and transmission [7] [8].

The field of artificial intelligence is essentially when machines can do tasks that typically require human intelligence. It encompasses machine learning, where machines can learn by experience and acquire skills without human involvement [9]. Machine learning is an artificial neural network, algorithms inspired by the human brain, learn from large amounts of data.

Big data are generated inside the techniques of the safety troubles and in the operations of technology, transmission, transformation, distribution, consumption and dispatching of the smart grid. These massive data are very treasured assets for security situational awareness. Based at the long-term period tracking of the smart grid, security-related big data can be formed [10].

Electricity theft are one of the major problems that has brought enormous financial losses to electric utility companies worldwide. The smart grid could consolidate to the overall energy infrastructure of the smart city by creating additional energy storage using a community solar panel networks [11]. The main issue in the development of a smart grid is not located at the physical support but mainly in ensuring both security and privacy, which has come a major concern for the cyber security exploration community. An adversary can launch internal and external attacks (e.g., false data injection and denial of service attacks) in order to disrupt the operation of the mart grid. Examples include revision operations on the electricity data via wiretapping attack and distributed denial of service attacks on the network communication protocols (i.e., HTTP, TCP/IP, UDP) [12] [13]. Using power system datasets, this paper investigates a machine learning model to overcome this drawback.

2.RELATED WORK

Wang X et.al [13] put forward the recognition and partitioning of False Data Injection (FDI) strike in smart grid situated on the unknown input interval observer. A local logic judgment matrix-based detection and isolation algorithm is developed, in the presence of undetectable FDI attacks. Based at the mixtures of observable sensor cases, local control manage centres can further locate and isolate the assault set under structure vulnerability. The effectiveness of the evolved detection and isolation algorithms towards FDI assaults is demonstrated.

Ali S et.al [12] analyses an efficient DDoS attack detection technique based on multilevel auto-encoder based feature learning. By combining the efficient Multiple Kernel Learning algorithm and using multilevel features a final unified detection model is learned. The overall system is targeted towards more accurate and more efficient DDoS attack detection in the smart grid network. Experiments are performed on two benchmark DDoS attack detection databases and the results are compared with six state-of-the-art methods.

Shateri M et.al [11] provide privacy to the end users with minimal extra energy cost. Based on a model-free deep reinforcement learning algorithm a privacy-cost management unit is proposed and factual results evaluated on actual Smart Meters (SMs) data are confer to compare DDQL with the state-of-the-art. The extra cost because of the usage of RB was covered while privacy measured as the common relative distance of the grid load from a constant load.

Amin BR et.al [2] proposes a novel belief propagation-based algorithm to dig out both random and stealthy-type FDIAs in smart grids thus algorithm auspiciously ferrets random and stealthy attacks during corrupted single set data and multiple consecutive instance data. BP algorithm achieves 99.94% attack detection rate accuracy, which is crucially higher than the detection rate accuracy of state-of-the-art machine learning classifiers.

He X et.al [16] put forward the universal structure of the consortium block-chain of the smart grid accession and the chance scenarios of NTL problems inside the smart grid are analysed. A smart grid NTL detection model focused on the power grid association consortium block-chain is also initiated. They cleave the communication network domains comparatively HAN, LAN, and WAN in the smart grid. It brings forward a smart grid NTL problem detection scheme based totally on the energy gateway block-chain to work out the NTL complication in the smart grid system.

2.1 Existing Model

The existing system revealed a smart grid vulnerability that can be exploited by attacks on the network topology. Cyber-physical attacks present a complex risk that needs to be accounted for in traditional power system contingency analyses. The study uses Q-learning to identify critical attack sequences for transmission grids under sequential topology attacks. A realistic power flow cascading outage model is used to simulate the system behavior, where attackers can use the Q-learning to improve the damage of sequential topology attacks toward system failures with the least attack efforts. The reinforcement learning-based approach for vulnerability analysis of sequential attacks in power transmission grids is used. By considering overloading-related cascading outages and hidden line failures, the approach evaluates the blackout damage resulting from switching interdiction.

2.2 System Model

This novel idea of Smart grid information attack detection for the smart home is more efficient and reliable compared to previous methods. While smart grid systems are built with both smart and legacy information and operational technologies, security and privacy concerns have grown. Intelligent grid systems rely heavily on intrusion detection to detect attacks and alert the operators. The smart grid employs a model-based detection module and a machine learning-based prevention module that dynamically learns the best strategy against an attacker. IDS along with a dynamic machine learning-based prevention technique used to detect and prevent intrusions with a low false positive rate (FPR) and without prior knowledge of attacks. In the smart grid community, information attacks have become a serious issue that has caused massive losses.

3. PROPOSED SYSTEM

A network attack detection model of a smart grid based on XGBoost is proposed. It maps the information attack identification problem in the smart grid to a multi-class classification problem. In the attack dataset, first, fill in the missing values based on the average filling method in a pre-processing stage, then based on the k-means-smote algorithm, the data is oversampled to obtain a balanced data set, which reduces the false-positive rate of the model's identification of information attacks. In order to reduce the complexity of the model and shorten the training time of the model, the maximum correlation-minimum redundant feature selection method is used to reduce the original feature set. Finally, an XGBoost integrated classifier is constructed to identify the information attacks.

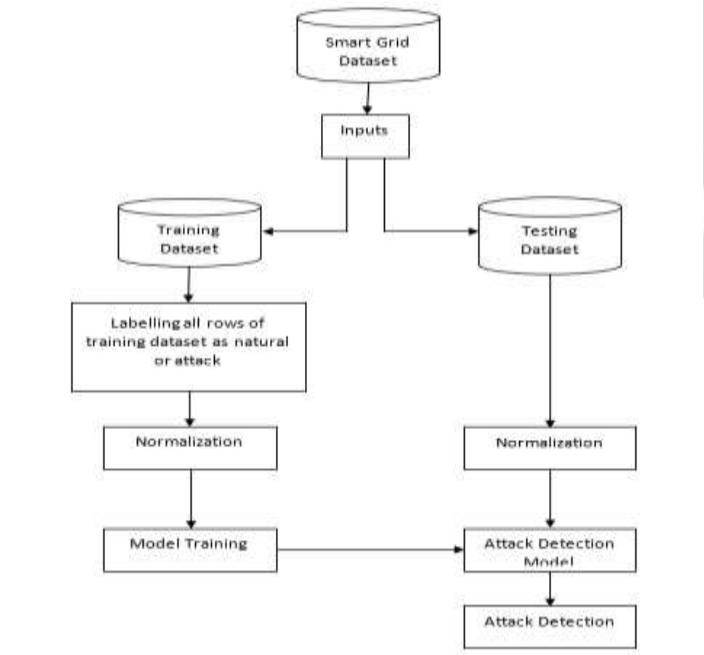


Fig 1.1 Proposed model for Attack Detection System

3.1 K-Means-Oversampling

K-Means-Oversampling oversamples the power data containing attack samples based on the improved k-means-oversampling method. Let x be the number of samples with the smallest number of samples in the data set, and y be the number of samples with the largest number of categories. The imbalance rate of the defined data set is (y−x)/x. At the point when the irregularity rate is higher than a set edge, an oversampling activity is performed.

Clustering, filtering, and oversampling are the three main steps of K-means-smote oversampling. Let the minority class sample be X, the label is Y, and k is the number of clusters generated by k-means clustering.

In order to determine the subset of features that are most conducive to characterizing cyberattacks, a two-stage feature selection method is proposed. In view of the common data hypothesis, in light of the greatest connection measure in the first component space, the most important element subset is chosen and marked. Suppose the target feature set is S, contains m features, N samples, and the data label is N. Define variables x, and y, the mutual information between them is shown in Eq. (1):

$$\mathrm{I}\ (\mathrm{x};\ \mathrm{y}) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)}\, dxdy \tag{1}$$

3.2 Probability Density Function

A probability density function is represented by p(i), p(j), and p(i, j), and the mutual information approximation between them in the definition of the maximum correlation criterion in the first stage is shown in Eq. (2):

$$\max D(S,c), D = \frac{1}{|s|} \sum_{xi \in S} I(x\mathrm{i}\ ; \mathrm{c}) \tag{2}$$

From Eq. (2), it can be known that the maximum correlation criterion only considers the mutual information between each information and the label, and does not consider the correlation between features. The classification accuracy will not be improved and the computational cost will increase when some features can be linearly represented by other features. A good subset of features. The minimum redundancy criterion is shown in Eq. (3):

$$\min\ R(S), R = \frac{1}{|s|2} \sum_{xi,xj \in S} I(x\mathrm{i}, \mathrm{xj}) \tag{3}$$

It is the remaining feature subset that is optimal for smart grid information attack identification after filtering out redundant features.

3.3 Prediction Accuracy

To improve prediction accuracy, XGBoost integrates multiple weak classifiers for integrated voting.

Assuming the original data set is D = {(xi, yi) : i = 1, 2,..., n, xi ∈ Rm, yi ∈ R}, The data contains a total of n samples and the data dimension is m. Each data sample xi has a label yi. Assuming that the XGBoost ensemble learning model integrates K regression trees, the XGBoost model is shown in Eq. (4):

$$\hat{y}_\mathrm{i} = \sum_{k=1}^{k} fk(xi), fk\ \in F \tag{4}$$

Let yˆi(t) is the predicted value of the i^{th} sample at the t^{th} iteration, and add an incremental function ft to optimize the objective function and improve the prediction accuracy during the t^{th} iteration. Then:

$$Obj^{(t)} = \sum_{i=1}^{n} l(yi, \hat{yi})^{(\mathrm{t-1})} + \mathrm{f_t(xi)} + \Omega(\mathrm{f_t}) \tag{5}$$

$$Obj^{(t)} = \sum_{i=1}^{n} \left[gift(xi)\ + \frac{1}{2} hift\,(xi)^2 \right] + \Omega(ft) \tag{6}$$

gi and hi represent the first and second derivatives of the loss function, respectively, and they are calculated as shown in Eqs. (7) and (8):

$$Obj^{(t)} = \sum_{i=1}^{n} \left[gift(xi)\ + \frac{1}{2} hift\,(xi)^2 + \gamma T + \|w\| 2 \right]$$

$$\sum_{j=1}^{T}\left[\left(\sum_{i\in Ij} gi\right)wj+\frac{1}{2}(\sum_{i\in Ij} h_i+\ \lambda w_j^2)\right]+\gamma^T$$

(7)

In Eq. (8), Ij = {i|q(xi) = j} represents the sample group on the leaf j, and the objective function is converted to the problem of solving the minimum value of the one-variable quadratic system of wj. The optimal leaf j weight remains unchanged when the tree structure is Eq. (8) shows:

$$w_j^* = \frac{G_j}{H_j + \lambda}$$

(8)

The optimal target value at this time is shown in Eq. (9):

$$Obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j\ +\lambda} + \lambda T$$

(9)

where Gj = Σi∈Ij gi, Hj = Σi∈Ij hi. A tree's structural score is represented by Obj. Precision increases with a smaller value. For the ongoing leaf hubs, the voracious calculation is utilized by XGBoost to partition the subtree. A split point is added to the current hub during every emphasis, and the hub that limits the goal capability and boosts the data gain is chosen to part. The data gain condition is displayed in Eq. (10)

$$\text{Gain } = \frac{1}{2}[\frac{G_L^2}{H_L\ +\lambda} + \frac{G_R^2}{H_R\ +\lambda} - \frac{(G_L+G_R)}{H_L\ +H_R+\lambda}] - \gamma$$

(10)

The hubs of each tree of the XGBoost model are doing a component split. The significance of an element can be assessed by the times a component is chosen as a split element. The higher the significance, the better the classifier can distinguish brilliant lattice network assaults.

3.4 XGBoost Algorithm

While utilizing XGBoost characterization, it is important to change the boundaries of the coach to work on its presentation. The selection of boundaries decides the precision of the XGBoost model. The normally utilized boundary change strategy is by and large the network search technique, yet the pursuit scope of the matrix search strategy is excessively limited, and finding the ideal parameters is difficult. This paper proposes an XGBoost arrangement calculation. The Python tool stash XGBoost is chosen to advance three significant boundaries in the XGBoost classifier: learn (learning_rate, ETA for short), the greatest profundity of the tree (max_depth), and test examining rate (subsample). Learning_rate: while refreshing leaf hubs, the weight will be increased by ETA. By decreasing the heaviness of the component, the advancement estimation process is more moderate. +e generally utilized esteem range is [0, 1], and the default esteem is 0.3. Max_depth: it controls the intricacy of the choice tree. The bigger the worth, the more perplexing the model, however, overfitting will happen. The default esteem is 6. Subsample: the subsample proportion of the preparation set implies that XGboost chooses the example proportion of the first traversing tree, which can actually forestall overfitting. The default esteem is 1.

3.5 Advantages

- Protects the smart grid from cyber attacks
- Achieves privacy preservation
- Detects network attacks and fraudulent transaction.

4. RESULT AND DISCUSSION

The first step for the smart grid information attack detection is to input the power dataset. It contains the field like Voltage Phase Angle and Magnitude, Current Phase Angle and Magnitude, Zero Voltage Phase Angle and Magnitude, Zero Current Phase Angle and Magnitude, Frequency for relays, Delta and Status Flag for relays. In the attack dataset, first fill in the missing values based on the average filling method in pre-processing stage, then based on the k-means-smote algorithm, the data is oversampled to obtain a balanced data set, which reduces the false-positive rate of the model's identification of information attacks. The Scatter Matrix is shown in Fig 1.2.

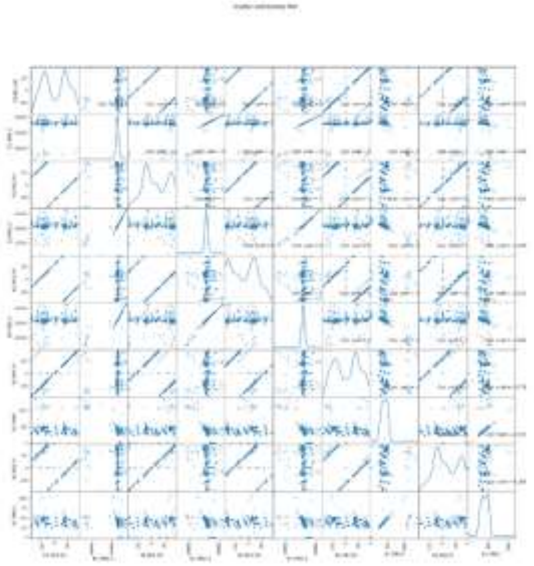
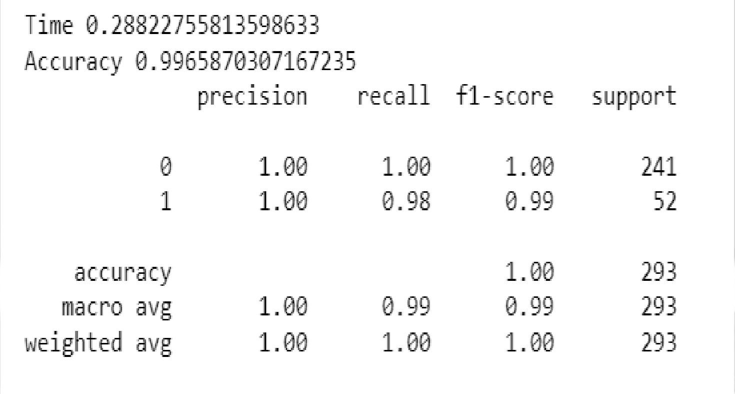


Fig 1.2 Scatter Matrix of PowerGrid dataset

In order to reduce the complexity of the model and shorten the training time of the model, the maximum correlation-minimum redundant feature selection method is used to reduce the original feature set. An XGBoost integrated classifier is constructed to identify the information attacks shown in Fig 1.3 and Fig 1.4.



```
Time 0.28822755813598633
Accuracy 0.9965870307167235
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       241
           1       1.00      0.98      0.99        52

    accuracy                           1.00       293
   macro avg       1.00      0.99      0.99       293
weighted avg       1.00      1.00      1.00       293
```

Fig 1.3 Prediction Result using XGBoost

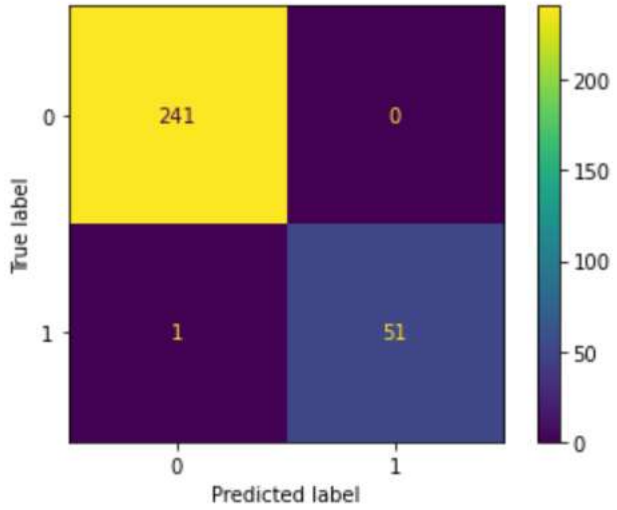


Fig 1.4 Label Prediction

The following table 1 displays the comparison results of proposed XGBoost with various methods such as the convolution neural network (CNN) approach, Recurrent Neural network using Long Short-Term Memory, Deep Neural Network and so on, and its accuracy comparison is represented in chart 1.

Table 1: Performance comparison of existing and proposed systems

Methods	**Accuracy**
DGNN	93.93%
DBN	92.50%
DNN	75.75%
STL	79.10%
RNN using LSTM	90.93%
RNN using FPBP	81.29%
NDAE	89.22%
RNN using GRU	93%
CNN	94.53%
MDPCA with DBN	82.08%
Proposed XGBoost	95.52%

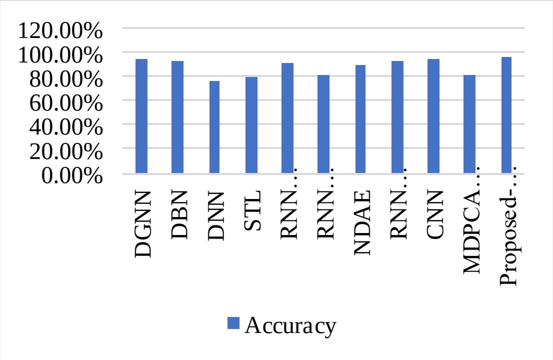


Chart 1: Graphical representation of the accuracy comparison

5.CONCLUSION

As the traditional electric grid system transitions to a smart grid system, the conventional power system methods present limitations in processing and analyzing the massive amounts of data that is now a norm with a smart grid. Thus, AI techniques are being developed and applied to many applications in smart grid systems with promising results. In this paper, a novel deep learning energy framework, for smart grids is proposed. This framework includes a novel deep learning-based scheme using XGBoost algorithm. Through performance evaluations using Smart Grid dataset, demonstrated the efficiency of the proposed framework. For future work, planned to study the performance of the framework with the integration of edge computing in the smart grid. Then, will consider an edge computing enabled network in the smart grid, where energy nodes can access and utilize computing services from an edge computing service provider. This integration may help the energy nodes achieve optimal energy management policy.

6.REFERENCE

[1] Otuoze, A.O., Mustafa, M.W. and Larik, R.M., 2018. Smart grids security challenges: Classification by sources of threats. *Journal of Electrical Systems and Information Technology*, *5*(3), pp.468-483.

[2] Amin, M., 2005. WOLL ENBERG B F. Toward a smart grid: power delivery for the 21st century. *IEEE Power and Energy Magazine*, *3*(5), pp.34-41.

[3] Corbett, Jacqueline, Katherine Wardle, and Chialin Chen. "Toward a sustainable modern electricity grid: The effects of smart metering and program investments on demand-side management performance in the US electricity sector 2009-2012." IEEE Transactions on EngineeringManagement 65.2 (2018): 252-263.

[4] Zhao, Shuai, et al. "Smart and practical privacy-preserving data aggregation for fog-based smart grids." IEEE Transactions on Information Forensics and Security 16 (2020): 521-536.

[5] Jawad, Muhammad, et al. "Machine learning based cost-effective electricity load forecasting model using correlated meteorological parameters." IEEE Access 8 (2020): 146847-146864.

[6] Alahakoon, D. and Yu, X., 2015. Smart electricity meter data intelligence for future energy systems: A survey. *IEEE Transactions on Industrial Informatics*, *12*(1), pp.425-436.

[7] Morello, Rosario, et al. "A smart power meter to monitor energy flow in smart grids: The role of advanced sensing and IoT in the electric grid of the future." IEEE Sensors Journal 17.23 (2017): 7828-7837.

[8] Alkahtani, Hasan, and Theyazn HH Aldhyani. "Intrusion detection system to advance internet of things infrastructure-based deep learning algorithms." Complexity 2021 (2021).

[9] Wang W, Lai Q, Fu H, Shen J, Ling H, Yang R. Salient object detection in the deep learning era: An in-depth survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021 Jan 12.

[10] Wang, Yong, Zhaoyan Xu, Jialong Zhang, Lei Xu, Haopei Wang, and Guofei Gu. "Srid: State relation-based intrusion detection for false data injection attacks in scada." In *European symposium on research in computer security*, pp. 401-418. Springer, Cham, 2014.

[11] Shateri M, Messina F, Piantanida P, Labeau F. Privacy-cost management in smart meters using deep reinforcement learning. In2020 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe) 2020 Oct 26 (pp. 929-933). IEEE.

[12] Ali S, Li Y. Learning multilevel auto-encoders for DDoS attack detection in smart grid network. IEEE Access. 2019 Aug 5;7:108647-59.

[13] Wang X, Luo X, Zhang M, Jiang Z, Guan X. Detection and isolation of false data injection attacks in smart grid via unknown input interval observer. IEEE Internet of Things Journal. 2020 Jan 13;7 (4):3214-29.

[14] Ismail, Muhammad, et al. "Deep learning detection of electricity theft cyber-attacks in renewable distributed generation." IEEETransactions on Smart Grid 11.4 (2020): 3428-3437.

[15] Srikantha P, Kundur D. A DER attack-mitigation differential game for smart grid security analysis. IEEE Transactions on Smart Grid. 2015 Aug 24;7(3):1476-85.

[16] He X, Wang J, Liu J, Yuan E, Wang K, Han Z. Smart Grid Nontechnical Loss Detection Based on Power Gateway Consortium Blockchain. Security and Communication Networks. 2021 Oct 14; 2021.

[17] Padin A, Kebede Y, Morgan M, Vorva D, Fozdar A, Kalvaitis R, Remley N, Tout S, Kallitsis MG. Diagnosing False Data Injection Attacks in the Smart Grid: a Practical Framework for Home-Area Networks. InProceedings of the SGIoT Conference 2017 Jul.

[18] T. A. Tang, L. Mhamdi, D. McLernon, S. A. R. Zaidi, and M. Ghogho, “Deep learning approach for network intrusion detection in software defined networking,” in Proc. Int. Conf. Wireless Netw. Mobile Commun., 2016, pp. 258–263.

[19] C. Yin, Y. Zhu, J. Fei, and X. He, “A deep learning approach for intrusion detection using recurrent neural networks,” IEEE Access, vol. 5, pp. 21 954–21 961, 2017.

[20] N. Shone, T.N.Ngoc,V.D. Phai, andQ. Shi, “A deep learning approach to network intrusion detection,” IEEE Trans. Emerg. Topics Comput. Intell., vol. 2, no. 1, pp. 41–50, Feb. 2018.

[21] Lakshminarayana S. Smart grid technology & applications. In 2014 POWER AND ENERGY SYSTEMS: TOWARDS SUSTAINABLE ENERGY 2014 Mar 13 (pp. 1-6). IEEE.