symmetry

MDPI

Article

Machine Learning-Based Intrusion Detection for Achieving Cybersecurity in Smart Grids Using IEC 61850 GOOSE Messages

Taha Selim Ustun [1,*], S. M. Suhail Hussain [2], Ahsen Ulutas [3], Ahmet Onen [4], Muhammad M. Roomi [5] and Daisuke Mashima [5]

[1] Fukushima Renewable Energy Institute, AIST (FREA), National Institute of Advanced Industrial Science and Technology (AIST), Koriyama 963-0298, Japan
[2] Department of Computer Science, School of Computing, National University of Singapore, Singapore 637551, Singapore; suhail@ieee.org
[3] Department of Electrical and Electronics Engineering, Necmettin Erbakan University, 42090 Konya, Turkey; a.ulutas@agu.edu.tr
[4] Department of Electrical and Electronics Engineering, Abdullah Gul University, 38170 Kayseri, Turkey; a.onen@agu.edu.tr
[5] Advanced Digital Sciences Center, Illinois at Singapore Pte Ltd., University of Illinois at Urbana-Champaign, Singapore 138602, Singapore; roomi.s@adsc-create.edu.sg (M.M.R.); daisuke.m@adsc-create.edu.sg (D.M.)
* Correspondence: ustun@ieee.org or selim.ustun@aist.go.jp

Citation: Ustun, T.S.; Hussain, S.M.S.; Ulutas, A.; Onen, A.; Roomi, M.M.; Mashima, D. Machine Learning-Based Intrusion Detection for Achieving Cybersecurity in Smart Grids Using IEC 61850 GOOSE Messages. *Symmetry* **2021**, *13*, 826. https://doi.org/10.3390/sym13050826

Academic Editors: Alfredo Alcayde, Raúl Baños Navarro and Kuo-Hui Yeh

Received: 2 April 2021
Accepted: 6 May 2021
Published: 8 May 2021

Abstract: Increased connectivity is required to implement novel coordination and control schemes. IEC 61850-based communication solutions have become popular due to many reasons—object-oriented modeling capability, interoperable connectivity and strong communication protocols, to name a few. However, communication infrastructure is not well-equipped with cybersecurity mechanisms for secure operation. Unlike online banking systems that have been running such security systems for decades, smart grid cybersecurity is an emerging field. To achieve security at all levels, operational technology-based security is also needed. To address this need, this paper develops an intrusion detection system for smart grids utilizing IEC 61850's Generic Object-Oriented Substation Event (GOOSE) messages. The system is developed with machine learning and is able to monitor the communication traffic of a given power system and distinguish normal events from abnormal ones, i.e., attacks. The designed system is implemented and tested with a realistic IEC 61850 GOOSE message dataset under symmetric and asymmetric fault conditions in the power system. The results show that the proposed system can successfully distinguish normal power system events from cyberattacks with high accuracy. This ensures that smart grids have intrusion detection in addition to cybersecurity features attached to exchanged messages.

Keywords: smart grid cybersecurity; GOOSE message security; IEC 62351; intrusion detection; artificial intelligence

1. Introduction

The integration of Information Technology (IT) with power systems gave birth to smart grids [1]. "In this fashion, more measurement can be done, and better operational decisions can be made. Power systems are operated more efficiently with smaller margins, in contrast to traditional procedures. Additionally, such connectivity enables novel applications that require coordination of more than one equipment in the system [2]. For instance, coordination of electric vehicles with renewable energy-based generators to mitigate their intermittency requires continuous message exchanges between these entities [3]. Alternatively, virtual power plant concept where different generation and storage devices act collectively to represent a much larger generation plant heavily relies on successful

power system from cyberattacks. It makes use of two key parameters of GOOSE messages, stNum and sqNum, as will be discussed below.

The major contributions of this work are as follows:

(a) A novel machine learning-based intrusion detection system is developed for IEC 61850 GOOSE messages.
(b) A realistic power system communication dataset is obtained. This dataset is used to train the proposed system. Then, the performance of the system is tested with test data where cyberattacks are included.
(c) Different machine learning algorithms are utilized, and their performances are contrasted. Results are reported to discuss which one of these algorithms is more suitable for intrusion detection in power system communication based on IEC 61850 messages. Evaluations are done in terms of training and attack detection times as well as attack detection accuracy."

The rest of the paper is organized as follows: Section 2 gives an overview of IEC 61850 GOOSE messages, their structure and operation style. Section 3 presents the proposed intrusion detection algorithm. Section 4 gives details about performance experiments, sample data and test data. Finally, conclusions are drawn in Section 5.

2. IEC 61850 GOOSE Messages and Cybersecurity Vulnerabilities

"IEC 61850 communication standard was initially developed to establish communication between substation devices [7]. However, it has received a lot of attention from researchers, engineers and companies alike. Its initial domain is extended several times so that it can be used for power system communication with a much larger pool of available devices. Researchers have worked towards developing models for novel devices such as energy routers [6], electric vehicles (EVs) [23], smart meters [24] or new smart grid applications such as virtual power plants [4], EV charging coordination schemes [3]. "The main reasons behind such a positive uptake are object-oriented modeling that allows for simple yet strong device modeling, interoperable communication systems that do not depend on certain company or a technology as well as robust message exchange services that are developed for power system applications [25]. As shown in Figure 1, there are three services utilized. Sample Value messages are used for periodic reporting of measurement values while Client-Server communication is used for ad-hoc message exchanges, notifications and reporting."

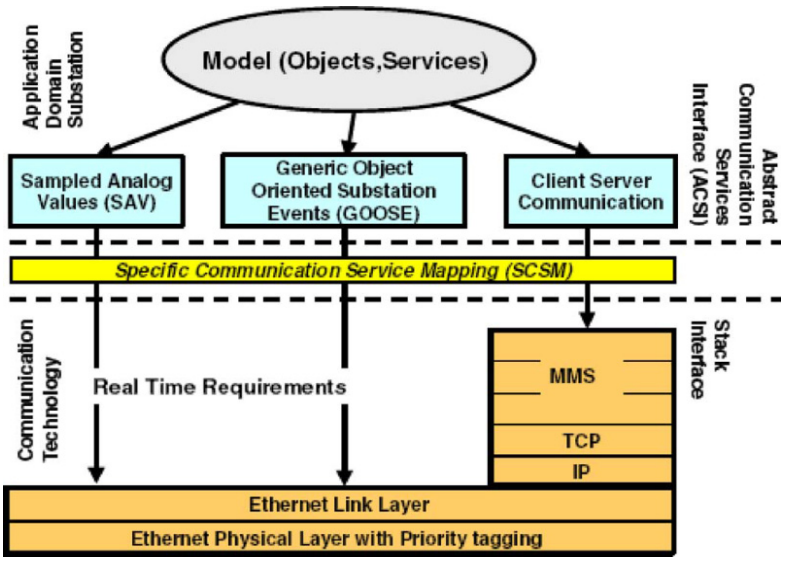

Figure 1. IEC 61850 communication stack.

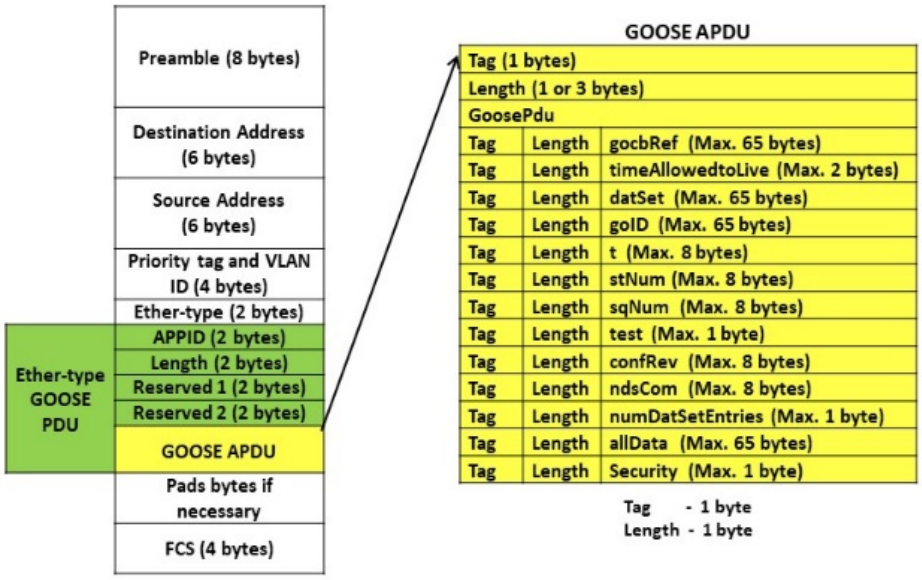

Figure 3. Message structure of GOOSE.

"For instance, as shown in Figure 1, GOOSE messages are directly mapped onto Ethernet layer, skipping TCP/IP and making transmission much faster. However, the downside is that there is no traditional sender and receiver addresses that can be used to protect messages and prevent cyberattacks. It is true that the GOOSE message structure as shown in Figure 3 includes destination and sources addresses, but these cannot be used for such purposes. The reason is that, firstly, these are Media Access Control (MAC) addresses and self-declared. Secondly, the destination address is not a real device's address." It is utilized to differentiate GOOSE streams from each other and can take any value within the range specified in IEC 61850 standard, as shown in Table 1.

Table 1. GOOSE and sampled value (SV) messages target address ranges.

Message Type	Address Range
GOOSE	01:0C:CD:01:00:00 to 01:0C:CD:01:01:FF
SV	01:0C:CD:04:00:00 to 01:0C:CD:04:01:FF

Secondly, as mentioned in [28], these messages do not include any cybersecurity mechanism whether it be message integrity, authentication or encryption. They are exchanged over the net with full visibility and can be read and viewed by any party [28]. The messages do not have any built-in mechanism to authenticate the sender which leaves the doors open for any imposter attack [29]. Similarly, there is virtually nothing stopping an entity from capturing a GOOSE message exchanged in the network, editing its contents and retransmitting it as a part of a replay or masquerade attack [30]. Some of these issues are identified and IEC 62351 Cybersecurity standard has been issued as a complementary to IEC 61850 communication standard. The proposed cybersecurity mechanisms are still in their infancy and require a lot of work to be widely implemented in power system communication infrastructure.

Nevertheless, IEC 62351 cybersecurity standard only recommends use of communication layer security mechanisms, such as implementing hash algorithms to check message integrity or using digital signatures to authenticate senders. There is no input on operational layer security. To ensure fully secure communication, a holistic cybersecurity approach is needed. For instance, if a hacker circumvents the security checks implemented at communication layer and gains access to the network, there is no system in place to detect this breach. Considering the sensitive nature of GOOSE message contents and that they are used to trigger actions in devices, this is a big problem." In order to fill this

and developing a behavior model as shown in Figure 5. Machine learning is utilized to develop a pattern for any given power system." Then, *stNum* and *sqNum* values are extracted from the incoming GOOSE message. Based on the event history, i.e., the event preceding this particular GOOSE message, and the regular behavior of the system, it is concluded whether there is any discrepancy. As explained earlier, if *stNum* is changing too often or *sqNum* values stay too low, this means way too many events are occurring in the system than usual. This indicates that a hacker has gained access to the system and is trying to trigger multiple events in a short period of time to effect usual operation.

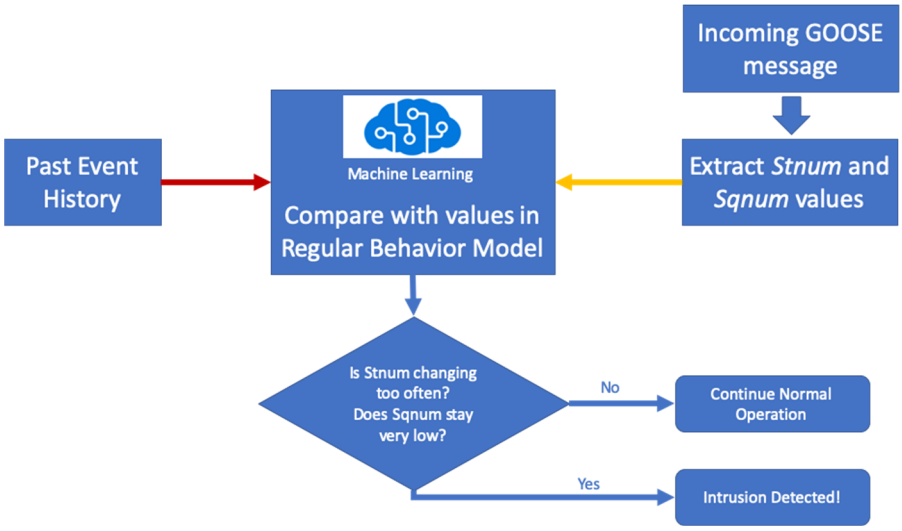

Figure 5. Machine learning-based event analysis.

Several machine learning algorithms are utilized, and their performances are compared in the next section. Before getting into test results and their analysis, an overview of these algorithms is given in the next sub-section.

Different Machine Learning Algorithms Utilized

As discussed in [31], "In order to measure success of prediction and contrast their performances, several algorithms are utilized in the proposed system. These are Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K Nearest Neighbor (k-NN) and Adaptive Boost (AdaBoost) algorithms.

Shown in Figure 6, DT algorithm utilizes decision tress with branches and leaves. In this fashion, it extracts conclusions from observations related to a particular item. In this approach, observations are represented as branches while the conclusions are the leaves. The algorithm is designed to progress towards the leaves. Since the goal of DT is to draw some conclusions and estimate the value of a target node, it is deemed suitable for the developed intrusion detection system where values for *stNum* and *sqNum* values are estimated in a broad sense.

A collection of DTs constitutes a RF. In other words, RFs utilize several DTs to make a decision and individual decisions from each DT are processed to reach a final conclusion in RF, as shown Figure 7. Decisions are made by following the most efficient path in each DT. RF is a bagging algorithm and it can be utilized to address over-fitting or accuracy issues encountered in DTs. The number of DTs is not limited and can be set as wished. In this particular study, 100 trees are used in RF."

k-NN is an instance-based learning technique where the input is classified based on some recent inputs. For this reason, it is also classified as a memory-based classification algorithm. An unclassified input is classified based on k number of neighbors, i.e., most recent k events. Since the intrusion detection system proposed in this paper requires keeping track of recent events and deciding whether the current event is an attack or not, k-NN is a very suitable algorithm. As shown in Figure 8, the number of k has a significant impact on the classification results. In the figure, in the classification of a new input, the green circle is the red triangle when k is selected as 3, while it becomes the blue square when k is 5. For this study, k is selected as 2. In other words, two previous events are utilized to make a decision on the incoming event.

Adaboost is a classifier based on a boosting method. It is utilized to lump together several weak classifiers, e.g., decision stumps, in order to build a much stronger classifier. Rather than being a distinct classifier on its own, Adaboost can be utilized on any classifier to identify its shortcomings and boost its performance. Its operation principle is given in Figure 9. Firstly, the input data are processed with the first classifier. Incorrectly classified training data are given a higher weight and the second classifier is run with these conditions. The output of the second classifier is treated the same before being fed to the third classifier as input. The unique feature about Adaboost is that the weight is updated in every single iteration. In this study, Adaboost is utilized with decision stumps. The motivation behind this selection is to see its impact on improving the performance of DT and compare with RF.

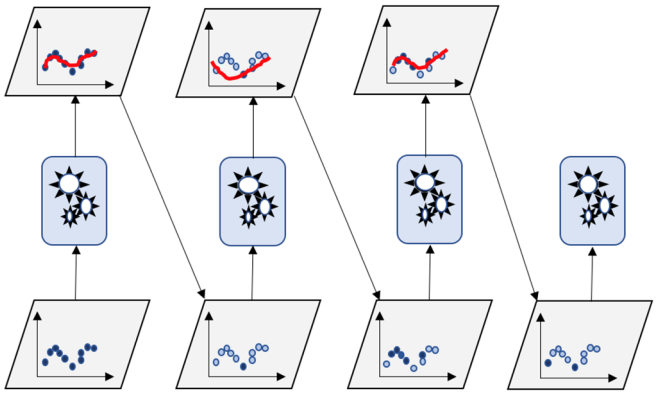

Figure 9. Instance weight updates of Adaboost classifier.

The next section presents the training data, test data and the test results for all the algorithms discussed above.

4. Intrusion Detection Performance Tests

In order to investigate the accuracy of the proposed intrusion detection system for GOOSE messages, several tests have been performed. Firstly, a realistic GOOSE dataset is developed for the generic power system given in Figure 10. In order to achieve this, firstly, emulators have been developed to generate and transmit GOOSE messages as per IEC 61850 rules [32,33]. Then, using these IEC 61850 emulators, desired GOOSE messages are created and sent as shown in Figure 11.

Once the clean dataset is acquired as in [33,34], random attack messages have been added to achieve the validation dataset. Figure 12 shows the set up used for creating the validation dataset. The attack GOOSE messages are published using the IEC 61850 emulators tools [32] and added to the clean dataset, creating the validation dataset. Figures 13 and 14 show *stNum* and *sqNum* values in this set, respectively. The attack data are shown in red and are added to the initial dataset. As it can be observed, regular behavior can be easily distinguished from the attack behavior.

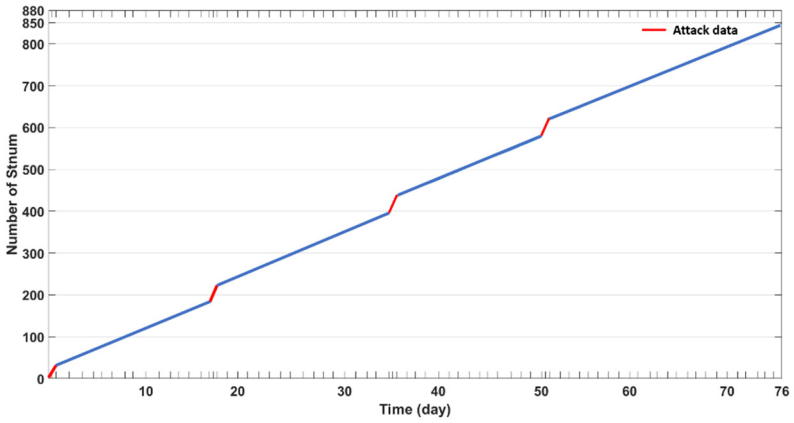

Figure 13. stNum values in dataset with attack data.

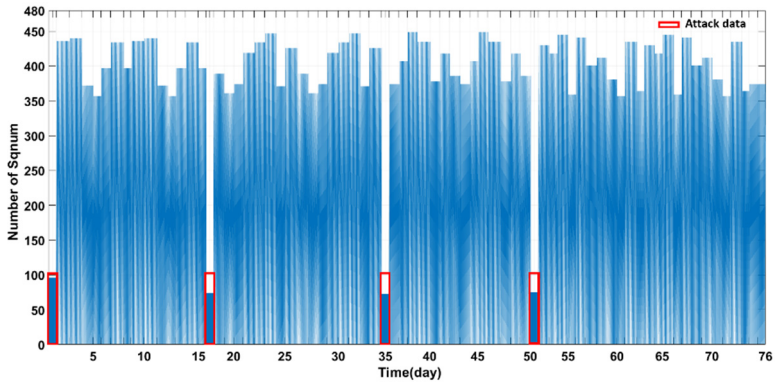

Figure 14. sqNum values in dataset with attack data.

Just to put into perspective, during normal operation the *stNum* value changes between 8 and 12 while the cyclic GOOSE message transmission continues until *sqNum* is somewhere between 350 and 450. On the contrary, attack data have a range of 15–21 for *stNum* and this is achieved in a much smaller time window, i.e., the rate of change is very high. Naturally, corresponding *sqNum* values stay low somewhere between 70 and 100. The time window was set as 76 days where the attacks are encountered only in 4 days. The distinct behavior of attack data is visualized by their steep slope, as shown in Figure 13. This corresponds to high rate of change of *stNum*. In contrast, normal operation data have a much more horizontal slope.

Cross validation studies are performed with a cross validation value of 7, performing seven distinct iterations on the dataset. As shown in Figure 15, the overall data are split into seven equal portions. In each iteration, a different portion is designated as the test fold, while the rest is used as training data. The benefit of this approach is that it mixes the training and test folds over the entire dataset. This eliminates the possibility of any lucky situations that may arise from a specific way of splitting the dataset. Every portion gets to be utilized as a test fold, thereby subjecting the proposed intrusion detection system to all possible combinations.

Performance tests have been performed in Python on two platforms for better comparison: (1) Intel Core i7-6700 @ 2.60 GHz with 16 GB RAM, and (2) Raspberry pi 3 (R-pi3). The results are reported in Tables 2 and 3. Firstly, it is safe to say that the proposed intrusion detection system is validated with these results. Regardless of the machine learning algorithm used, the system distinguishes regular operation from cyber-attacks in GOOSE messages. All the cases have reported accuracy higher than 94%. This confirms the intrusion detection approach design and shows that *stNum* and *sqNum* values can be used to detect cyberattacks on GOOSE messages

that only training time is affected by the volume of the dataset, while testing time stays constant. Considering that IEC 61850 standard stipulates that GOOSE messages need to be delivered within 3ms, this additional time introduced by the proposed system is very important.

Analyzing the performance test data for the platform with i7 processor in Table 2, it can be observed that DT can be safely used for intrusion detection in a system running GOOSE messages. The testing time required for this algorithm is less than 1 msec and is feasible for meeting IEC 61850 requirements. In contrast, Adaboost, RF and SVM algorithms require much longer processing times and this renders it impractical for GOOSE-based communication systems. These algorithms are deemed to be very robust and more accurate for complex systems. However, the data processing for the proposed system is very lightweight and timing has priority. The remaining algorithm, k-NN, can be utilized in a very fast system if GOOSE messages are guaranteed to arrive very rapidly, i.e., within 1 ms as the testing takes around 2 ms. The results with r-pi3 show that it is not practical to implement this IDS with slow systems. However, r-pi3 is a very old system and k-NN tests are performed in 3 msec. New generation r-pi or faster systems can be utilized to implement k-NN or in SVM. It is noteworthy to mention here that new IEDs are equipped with very strong processors such as i7, unlike their traditional and slow counterparts [35].

Finally, all of the algorithms except SVM have relatively short training times, considering that training is performed offline. This opens a path to the pseudo-online training approach where the system may collect data and retrain itself on a specific time window, e.g., 1 month or 3 months. The training times reported in Table 2 correspond to a 78-day dataset, i.e., 11 weeks. This will add value to the proposed system as it can learn the changing behavior of the power system and adjust its training. This will create a much more dynamic intrusion detection system that can respond to changing trends in the power system.

Test results show that DT achieves very high accuracy with much smaller training and detection times. Therefore, it can be deemed as the most suitable algorithm for the proposed intrusion detection system since it offers the best combination of higher accuracy and less time required.

5. Conclusions

Smart grid applications are getting more popular where different devices need to communicate and coordinate. For this to happen, a reliable infrastructure is needed. There have been efforts towards providing an interoperable communication platform for such purposes. However, the implementation of cybersecurity mechanisms to secure information exchange on such a large scale has lagged behind. There is imminent need for achieving cybersecurity in a power system, a cyber-physical system where message exchanges may have real, physical implications.

IEC 61850's GOOSE messages are widely used for instructing devices to perform certain operations. This makes them highly critical in cybersecurity assessments. This paper develops a machine learning-based intrusion detection system for GOOSE messages. Based on the frequency and nature of GOOSE messages, the system is able to differentiate *usual operation* from *attacks.* Performance tests have been performed with a realistic smart grid communication dataset. Furthermore, different machine learning algorithms were utilized to see their suitability for such use. The results show that the developed system can successfully detect cyber-attacks based on GOOSE message parameters with high accuracy. Although the performance of algorithms differs, all machine learning algorithms yield acceptable results and no over-fitting is observed.

Using algorithms other than the ones in this paper or using different parameter values can be a future extension of this work. Nevertheless, the current results show that the proposed intrusion detection system can successfully detect unauthorized access via GOOSE message analysis. Future work may focus on integrating this system with a honeypot.

21. Prisco, A.F.S.; Duitama, M.J.F. Intrusion detection system for SCADA platforms through machine learning algorithms. In Proceedings of the 2017 IEEE Colombian Conference on Communications and Computing (COLCOM), Cartagena, Colombia, 16–18 August 2017; pp. 1–6. [CrossRef]
22. Barbosa, R.R.R.; Pras, A. Intrusion Detection in SCADA Networks. In *Mechanisms for Autonomous Management of Networks and Services. AIMS 2010*; Stiller, B., De Turck, F., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6155. [CrossRef]
23. Nsonga, P.; Hussain, S.M.S.; Ali, I.; Ustun, T.S. Using IEC 61850 and IEEE WAVE standards in ad-hoc networks for electric vehicle charging management. In Proceedings of the 2016 IEEE Online Conference on Green Communications (OnlineGreenComm), Piscataway, NJ, USA, 14 November–17 December 2016; pp. 39–44. [CrossRef]
24. Liu, N.; Chen, J.; Luo, H.; Liu, W. A Preliminary Communication Model of Smart Meter Based on IEC 61850. In Proceedings of the 2011 Asia-Pacific Power and Energy Engineering Conference, Wuhan, China, 25–28 March 2011; pp. 1–4. [CrossRef]
25. Kim, H.J.; Jeong, C.M.; Sohn, J.-M.; Joo, J.-Y.; Donde, V.; Ko, Y.; Yoon, Y.T. A Comprehensive Review of Practical Issues for Interoperability Using the Common Information Model in Smart Grids. *Energies* **2020**, *13*, 1435. [CrossRef]
26. International Electrotechnical Commission. *IEC 62351-6: Power Systems Management and Associated Information Exchange—Data and Communications Security—Part 6: Security for IEC 61850*; International Standard: Geneva, Switzerland, 2020.
27. Boakye-Boateng, K.; Lashkari, A.H. Securing GOOSE: The Return of One-Time Pads. In Proceedings of the 2019 International Carnahan Conference on Security Technology (ICCST), Chennai, India, 1–3 October 2019; pp. 1–8. [CrossRef]
28. Cai, J.; Zheng, Y.; Zhou, Z. Review of cyber-security challenges and measures in smart substation. In Proceedings of the 2016 International Conference on Smart Grid and Clean Energy Technologies (ICSGCE), Chengdu, China, 19–22 October 2016; pp. 65–69. [CrossRef]
29. Hussain, S.M.S.; Ustun, T.S.; Kalam, A. A Review of IEC 62351 Security Mechanisms for IEC 61850 Message Exchanges. *IEEE Trans. Ind. Inform.* **2020**, *16*, 5643–5654. [CrossRef]
30. Ustun, T.S.; Farooq, S.M.; Hussain, S.M.S. A Novel Approach for Mitigation of Replay and Masquerade Attacks in Smartgrids Using IEC 61850 Standard. *IEEE Access* **2019**, *7*, 156044–156053. [CrossRef]
31. Ustun, T.S.; Hussain, S.M.S.; Yavuz, L.; Onen, A. Artificial Intelligence Based Intrusion Detection System for IEC 61850 Sampled Values Under Symmetric and Asymmetric Faults. *IEEE Access* **2021**, *9*, 56486–56495. [CrossRef]
32. Farooq, S.M.; Hussain, S.M.; Ustun, T.S. S-GoSV: Framework for Generating Secure IEC 61850 GOOSE and Sample Value Messages. *Energies* **2019**, *12*, 2536. [CrossRef]
33. Biswas, P.P.; Tan, H.C.; Zhu, Q.; Li, Y.; Mashima, D.; Chen, B. A Synthesized Dataset for Cybersecurity Study of IEC 61850 based Substation. In Proceedings of the 2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), Beijing, China, 21–23 October 2019; pp. 1–7.
34. Available online: https://github.com/smartgridadsc/IEC61850SecurityDataset (accessed on 18 December 2020).
35. Available online: https://selinc.com/products/3355/ (accessed on 2 December 2020).