A MODEL FOR THE SPAM DETECTION TECHNIQUE FOR IOT DEVICES

Dr.SHAIK MOHAMMAD RAFI[1] BABU BAKKA[2] P.RAJASEKHARAN[3] JAINA SAICHAND[4]
[1]PROFESSOR, DEPARTMENT OF ARTIFICIAL INTELLIGENCE,
[2] ASST. PROFESSOR, DEPARTMENT OF ARTIFICIAL INTELLIGENCE,
[3]ASST. PROFESSOR, DEPARTMENT OF ARTIFICIAL INTELLIGENCE,
[4] ASST. PROFESSOR, DEPARTMENT OF ARTIFICIAL INTELLIGENCE,
[1,2,3,4] SRI MITTAPALLI COLLEGE OF ENGINEERING

Abstract:

A network of millions of interconnected computing devices, housing sensors and actuators and communicating with one another over wireless or wired channels, is known as the Internet of Things (IoT). The Internet of Things (IoT) has expanded at a fast pace over the last decade, and by 2020, over 25 billion gadgets will be linked. Future years will see a meteoric rise in the amount of data produced by these gadgets. Not only does the volume of data generated by IoT devices rise, but the variety of modalities used to describe it also varies in terms of data quality, which is characterised by how quickly it is time- and location-dependent. When used to this kind of setting, machine learning algorithms have the potential to greatly enhance the safety and usefulness of IoT systems via biotechnology-based authorisation and abnormal detection. However, cybercriminals often study learning algorithms in order to find security flaws in intelligent systems that rely on the Internet of Things.

1. Introduction

Among the smart grid's most important parts is the advanced metering infrastructure (AMI), which includes both hardware (smart meters) and software (data management systems, communication networks, etc.). Utilities and end users are able to communicate with each other in both directions thanks to AMI. Because AMIs are structurally comparable to communication networks, power grids may benefit from methods developed for communication networks to prevent privacy breaches, harmful behaviours, and financial gain [1]. When compared to the danger of the individual parts, the system's infrastructure poses the greater threat. It gets increasingly difficult and complicated to track the system risk as the number of components vulnerable to attack grows [2]. Smart meter security risks include vulnerabilities in the network hub, distribution servers, links, management networks, firmware updates, hardware manipulation, and poor isolation between smart meters, power-line communication (PLC), and the smart meter's outlet. Aside from the displays, every action in the whole AMI is susceptible to changes in protocol architecture, network initialisation, and key management, which endanger the AMI infrastructure. It is difficult to exploit threats that interact with the hardware in order to change the memory. One of the most glaring dangers is meter manipulation, which manifests as changing the smart meter reading in order to provide the utility false information. Worst case scenario: this leads to overcharged bills and inaccurate data used for management and forecast. The smart meters generate massive volumes of time- and speed-varying data. Assuring the safety of smart home devices may be greatly aided by machine learning (ML) methods [3]. Integrity

attacks, which include the injection of fraudulent data, pose a cyber-physical hazard to contemporary smart grids. It was suggested that context extraction may be used to automatically identify malfunctioning Internet of Things (IoT) devices [4]. With the use of the collected data, a strong ensemble machine learning model was developed to identify any irregularities in the Internet of Things devices [5]. The goal of the approach suggested in [5] is to use ML model training to identify unusual occurrences in smart home datasets. Using a margin setting technique inside a data-centric paradigm, an analytical approach was suggested for detecting fake data injection (FDI) [6]. Analysis tools that make use of the massive volumes of data generated by the smart grid may detect integrity attacks like fake data injection (FDI). Anomaly identification in sensor data has been the subject of many approaches in the literature [5]. Problems with communication, massive data storage, security, and privacy are some of the network aspects that arise when Internet of Things (IoT) devices are integrated into smart home networks. Better energy consumption management is possible via analysis of data acquired from IoT devices, which allows for better monitoring of consumption trends [10]. When the devices linked to an AMI infrastructure are reliable, it becomes feasible to enhance energy efficiency with the aid of the infrastructure. Inaccurate measurements have negative consequences that may cause energy management to fail and cause a host of other problems. A smart home with Internet of Things (IoT) devices linked to the infrastructure has a number of problems, such as an absence of new goods made possible by edge computing, unreliable and scalable cloud infrastructure platforms, and guaranteed safe connections and data storage. To take advantage of security holes in the smart IoT system, a spam detection framework was suggested using machine learning models [3]. A spamicity score is applied to the Internet of Things devices in this study, which is an expansion of the previous work in [3]. This modification improves the algorithm's ability to run ML models simultaneously and handle time-series regression models rather than classification models. Determine the reliability of the Internet of Things (IoT) device in the smart home network by monitoring the readings of the appliances in the house every minute. Using autocorrelation analysis, the readings from the sensors were examined for any irregularities. For the purpose of securing smart devices via the detection of spam using various machine learning models, the algorithm assigns each Internet of Things device a spamicity score. Below is a summary of the key points made in this paper:
1. Improving prediction accuracy by delving into the structure of time-series data collected from smart home IoT devices.
2. Predicting the overall energy consumption of the IoT device and assigning it a feature significance score using machine learning modelling.
3. To improve the smart home environment's security, using feature priority ratings and energy prediction mistakes, calculate an IoT device spam score.

2. Related Work

One of the Internet of Things use cases that allows for the monitoring of energy being sent out or consumed is energy management, which may be used in the house as a specialised environment. Throughout the day, one may effortlessly switch between appliances that use less energy by monitoring all of the Internet of Things devices and their power use. When it comes to protecting IoT networks from cybercrime, IoT security is crucial. There are difficulties in controlling and managing the data flow caused by the massive amounts of data generated by IoT devices when connected to high-speed internet. One of the most promising approaches to controlling and managing data in the IoT is artificial intelligence (AI) [11]. Recent years have seen an influx of research into and implementation of deep learning algorithms for use with the vast array of IoT devices [12]. Machine learning, a branch of artificial intelligence, contributes to the Internet of Things (IoT) revolution in smart homes by analysing use patterns of internet-connected gadgets. To avoid using easily-spoofable IP addresses, machine learning is finding new uses in the Internet of Things security sector, such as making better use of network data collected from devices. Machine learning models have found many uses in cybersecurity, including but not limited to: analysing keystroke dynamics, detecting malicious URLs, fighting adversarial attacks, finding software vulnerabilities, automatically detecting intrusions, filtering spam emails, detecting phishing URLs, detecting credit card fraud, capturing network traffic, detecting botnet traffic, and Distributed Denial of Service (DDoS). In order to represent the temporal sequence using support vector regression for anomaly identification, Ma et al. created a framework [13]. A confidence value is assigned to each detection result that is created using the framework. To determine the best model weights, the authors developed a new version of the support vector regressor (SVR) that makes use of the Lagrange multipliers. A vote mechanism is used to choose the event duration n in the proposed online detection system, which is resilient. Both synthetic and actual data sets, including a Santa Fe Institute competition with a 1000-point time series, were used to validate the trials.

There are a number of supervised and unsupervised machine learning methods utilised in cybersecurity. Filtering out spam is one common use case for supervised learning. Two key ideas in machine learning are ensemble learning and time series analysis. By applying these ideas to past data, we may compare it to the present and spot any discrepancies that may arise in the future [14,15,16,17]. By using lengthy short-term memory approaches, Makkar et al. [3] created a system that can identify online spam. The study paper also used the method for spam categorisation in an IoT setting to provide a foundation for the Pagerank algorithm [18]. Malicious injections that generate false events are one kind of attack that wireless sensors are susceptible to. In order to guarantee the data integrity of the Internet of Things (IoT) network, Hau et al. [19] suggested a methodology for detecting fraudulent data injections in heterogeneous sensor data. Discussed in [20] is a review paper that examines the relative merits of several methods for protecting electrical power systems from hostile assaults. Some of the cyber-attacks offered by the Internet of Things include assaults that modify loads, attacks on data centres, and fraudulent command signals.

Using an algorithm to award an IoT device a spamicity score, the work highlighted in this study differentiates between good and poor network connections. In order to anticipate the time series load, the study discusses a regression issue and how several machine learning models assist with this task. The linked research highlights how ML analyses the time series data produced by IoT devices to help with spam identification in such devices.

3. Materials and Methods

Anomaly detection research utilising machine learning models in the IoT has shown encouraging results for detecting malicious internet traffic [21]. Additionally, in order to provide a secure network architecture, it is motivated to either identify abnormalities or apply a spamicity score to monitor the security of the network components. When it comes to electricity, a safe AMI infrastructure is crucial for smart grid security as a whole. The authors of the research addressed the possibility of connecting smart home gadgets to the cloud-based environment allowed by the

Internet of Things in their work [22]. When working with raw data, anomaly detection is a crucial stage for keeping an eye out for anything out of the ordinary. Bakar et al. compared the smart home environment to other security domains and emphasised the significance of anomaly detection in [23]. Given that the primary goal of this work is to provide an IoT spamicity score, this part delves into the data handling processes, statistical insights, anomaly detection in the data, and machine learning models used for prediction. Subsequently, it reveals the methodology that is utilised to compute the spamicity score.

3.1. Moving Average

One popular method for reducing the impact of data variations is time series decomposition, which is based on moving averages. You may find an uptrend or downtrend in the data using either a short-term or long-term moving average, depending on the period you choose. By capturing the trend from the previous day, moving averages make it easy to spot outliers in time-series energy data. Since the energy data does not anticipate abrupt changes in consumption numbers, every data point that differs from the moving average will be regarded as an anomaly. In order to get moving averages, one may use a number of different weighting schemes for the most recent data points, including:

- SMA, Or Simple Moving Average

The data points in the series were averaged over a certain time frame. When looking for broader patterns or cycles in data, moving averages may help level out the noise. Each data point is given equal weight in the simple moving average, independent of its occurrence x-1 days ago or the day before, making it one of the common averages. The SMA is calculated as the mean of the x data points. An SMA's simplicity and easy average price computation are two of its main benefits. Because it lends more weight to older data, the SMA isn't always the best choice, and that depends on the application type.

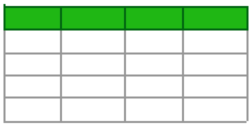
- The EMA, or its exponential moving average

Like SMA, an EMA averages data points over a certain time period, but unlike SMA, the weighting of each data point is not equal. The most recent data is given more weight, whereas data from farther back in time is given less weight. EMA gives greater weight to more current data than to previous data.

3.2. Machine Learning Models

The proposed algorithm is validated with the help of four ML models summarized below to identify the spamicity score. Regression methods are widely used in the short-term and medium-term power forecasting fields [**24**]. Several ML models are utilized for supervised machine learning; however, this paper uses ensemble methods, a set of ML techniques based on decision trees. The machine learning models utilized in the paper are described in **Table 1**.

Table 1. Various machine learning (ML) models for time series analysis.



(1)

Popular supervised machine learning models include Extreme Gradient Boosting (XGBoost) [25], which exhibits features such as efficiency, parallelisation, distributed and out-of-core processing, and so on. Only inside a single tree may many nodes be parallelised; this does not happen between trees. According to XGboost, complexity is defined as:
Ω(f)=γT+12λ∑j=1Tw2j

3.3. Spamicity Score
With the aid of the spamicity score, we may quantify the validity of the suggested technique [3]. An Internet of Things device's spamicity score indicates how trustworthy and dependable it is. The following formula calculates the spamicity score of all networked Internet of Things devices.
RMSE[i]=1n∑i=1n(yi−yP)2−−−−−−−−−−−−−√

S⇐RMSE[i]*Fi

where yi and yP denote the actual and predicted values, n denotes the number of samples, Fi denotes the feature importance vector, and S is the spamicity score of every IoT device. Multiplying the mistake rate by the feature significance score yields the spamicity score. Each round repeats the whole process of determining the spamicity score in

4. Results

4.1. Data Description
This study conducted its trials using a publicly available dataset of weather-enabled smart home IoT devices [33]. The dataset spans the time interval of one minute, beginning at 2016-01-01 08:00:00 and ending at 2016-12-31 23:00:00. What follows is an explanation of the power usage [kW] data statistic. The dataset that was taken into account for the study is summarised in Table 3. It sheds light on the dispersion of the data. A close approximation to 1 for the mean yields a value estimate for the whole dataset. In addition, the standard deviation, which displays the average distance of the data points from the mean, is near to ≈1.

4.2: Choosing Features
Step one in simplifying the model and understanding how characteristics affect energy prediction patterns as a whole is feature selection.
Feature selection is a technique for selecting a subset of features from a dataset by determining the relative importance of each feature [31]. Assigning a feature relevance score is another area where machine learning modelling comes in handy. The feature significance score is calculated for each machine learning model that was used to fit the data. All of the features that were evaluated using regression trees have their feature significance scores shown in Table 2.

4.3. Data Preprocessing
The data must be comprehensive and meet all criteria for data analysis, which is a crucial first step. Careful implementation of any preprocessing step is required to prevent dataset corruption. When a value was missing, the next valid observation was used to fill it in. Datasets with daily and monthly sampling were created from the original dataset according to the needs of the analyses.

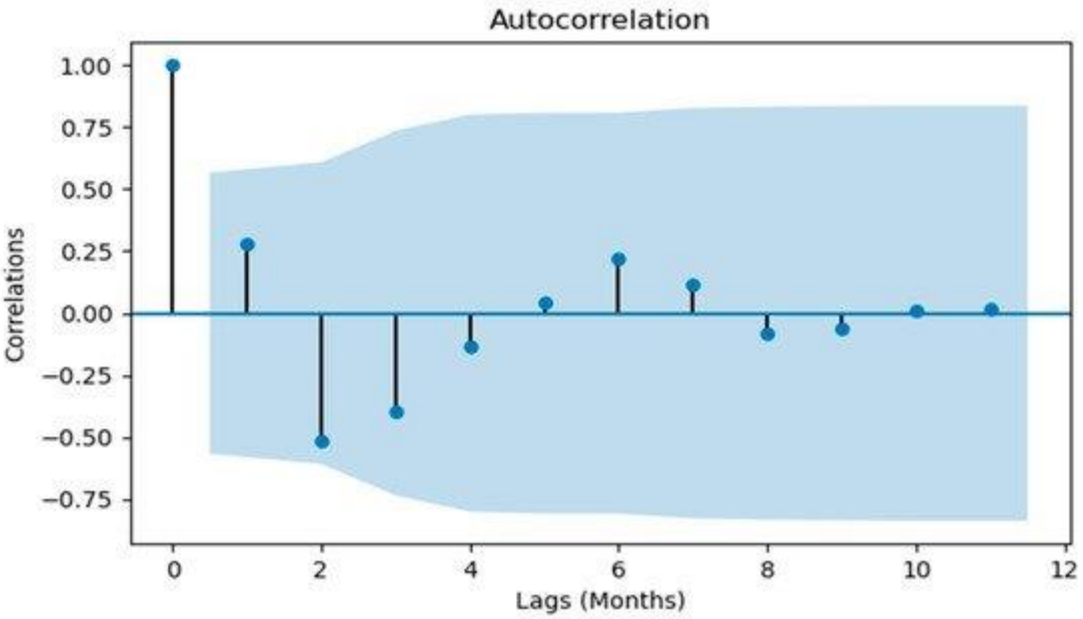


Figure 5. Autocorrelation of the power consumption [KW].

4.4. Data Statistics

The process of deserialising or detrending a time series might use a mix of methods, such as subtracting from the mean, differencing, log transformation, measuring percentage change, etc. Figure 6's data visualisation across a continuous span of time sheds light on the data structure via the density of the observations. Since there is a peak and values are declining exponentially, the right long line implies that the distribution is exponential rather than Gaussian.

.

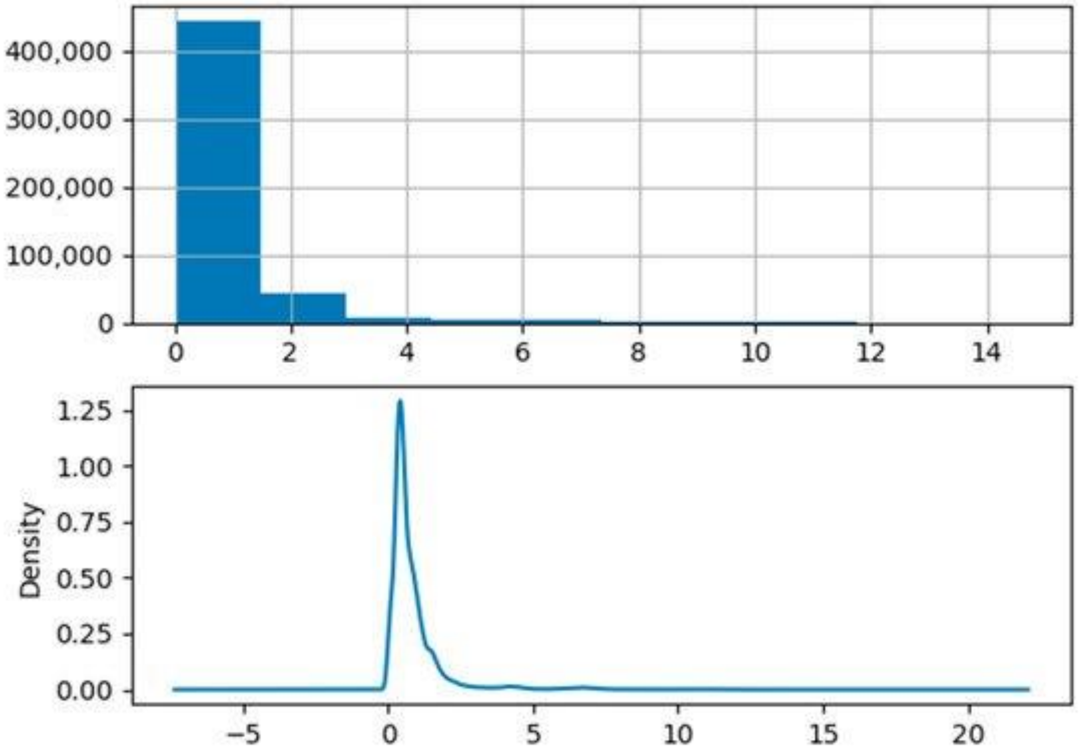


Figure 6. Histogram and density plot of the power consumption [KW].

By serially correlating a time-domain signal with itself as a function of delay, autocorrelation finds the degree to which the observations are similar across time. Finding out where the data has anomalies or noise is made easier by looking at the autocorrelation peaks. The consumption autocorrelation is seen in Figure 5.
Sticks that extend beyond the blue shaded zone are deemed statistically significant and were not caused by chance. The blue region represents the margin of uncertainty of the average readings.

5. Conclusions

Academics, industry developers, and researchers have recently focused on improving spam detection in IoT devices. This research explores the use of the spamicity score as a tool for comprehending the reliability of Internet of Things devices inside a smart home network. Every Internet of Things device gets a spam score based on the suggested methodology. To analyse the time-series data provided by smart meters, a number of ML models were subjected to extensive testing and experimentation. By assigning each smart home IoT device a spam score, we were able to use ensemble techniques of machine learning to ascertain the various IoT device contribution levels. According to the findings, the spamicity score of the devices plays a role in optimising the smart home environment for IoT device success.
You may determine the security of the IoT devices and get a spam score using the data supplied by the smart home in conjunction with meteorological characteristics.

References

1. Chapter 19: Admission Control-Based Load Protection in the Smart Grid—Security and Privacy in Cyber-Physical Systems. Available online: **https://learning.oreilly.com/library/view/security-and-privacy/9781119226048/c19.xhtml** (accessed on 30 April 2020).
2. Smart Meters—Threats and Attacks to PRIME Meters—Tarlogic Security—Cyber Security and Ethical Hacking. Available online: **https://www.tarlogic.com/en/blog/smart-meters-threats-and-attacks-to-prime-meters/** (accessed on 5 May 2020).
3. Makkar, A.; Garg, S.; Kumar, N.; Hossain, M.S.; Ghoneim, A.; Alrashoud, M. An Efficient Spam Detection Technique for IoT Devices using Machine Learning. IEEE Trans. Ind. Inform. **2020**. [**Google Scholar**] [**CrossRef**]
4. Choi, J.; Jeoung, H.; Kim, J.; Ko, Y.; Jung, W.; Kim, H.; Kim, J. Detecting and identifying faulty IoT devices in smart home with context extraction. In Proceedings of the 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, DSN 2018, Luxembourg, 25–28 June 2018; pp. 610–621. [**Google Scholar**] [**CrossRef**]
5. Tang, S.; Gu, Z.; Yang, Q.; Fu, S. Smart Home IoT Anomaly Detection based on Ensemble Model Learning from Heterogeneous Data. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019; pp. 4185–4190. [**Google Scholar**] [**CrossRef**]
6. Wang, Y.; Amin, M.M.; Fu, J.; Moussa, H.B. A novel data analytical approach for false data injection cyber-physical attack mitigation in smart grids. IEEE Access **2017**, 5, 26022–26033. [**Google Scholar**] [**CrossRef**]

7. Alagha, A.; Singh, S.; Mizouni, R.; Ouali, A.; Otrok, H. Data-Driven Dynamic Active Node Selection for Event Localization in IoT Applications—A Case Study of Radiation Localization. IEEE Access **2019**, 7, 16168–16183. [**Google Scholar**] [**CrossRef**]

8. Mishra, P.; Gudla, S.K.; ShanBhag, A.D.; Bose, J. Enhanced Alternate Action Recommender System Using Recurrent Patterns and Fault Detection System for Smart Home Users. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019; pp. 651–656. [**Google Scholar**] [**CrossRef**]

9. Gaddam, A.; Wilkin, T.; Angelova, M. Anomaly detection models for detecting sensor faults and outliers in the iot-a survey. In Proceedings of the 2019 13th International Conference on Sensing Technology (ICST), Sydney, Australia, 2–4 December 2019. [**Google Scholar**] [**CrossRef**]

10. Motlagh, N.H.; Khajavi, S.H.; Jaribion, A.; Holmstrom, J. An IoT-based automation system for older homes: A use case for lighting system. In Proceedings of the 2018 IEEE 11th Conference on Service-Oriented Computing and Applications (SOCA), Paris, France, 20–22 November 2018; pp. 247–252. [**Google Scholar**] [**CrossRef**]

11. Osuwa, A.A.; Ekhoragbon, E.B.; Fat, L.T. Application of artificial intelligence in Internet of Things. In Proceedings of the 9th International Conference on Computational Intelligence and Communication Networks, CICN 2017, Girne, Cyprus, 16–17 September 2017; pp. 169–173. [**Google Scholar**] [**CrossRef**]

12. Song, M.; Zhong, K.; Zhang, J.; Hu, Y.; Liu, D.; Zhang, W.; Wang, J.; Li, T. In-Situ AI: Towards Autonomous and Incremental Deep Learning for IoT Systems. In Proceedings of the 2018 IEEE International Symposium on High Performance Computer Architecture (HPCA), Vienna, Austria, 24–28 February 2018; pp. 92–103. [**Google Scholar**] [**CrossRef**]

13. Ma, J.; Perkins, S. Online novelty detection on temporal sequences. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 24–27 August 2003; pp. 613–618. [**Google Scholar**] [**CrossRef**]

14. Li, J.; Pedrycz, W.; Jamal, I. Multivariate time series anomaly detection: A framework of Hidden Markov Models. Appl. Soft Comput. J. **2017**, 60, 229–240. [**Google Scholar**] [**CrossRef**]

15. Flanagan, K.; Fallon, E.; Connolly, P.; Awad, A. Network anomaly detection in time series using distance based outlier detection with cluster density analysis. In Proceedings of the 2017 Internet Technologies and Applications (ITA), Wrexham, UK, 12–15 September 2017; pp. 116–121. [**Google Scholar**] [**CrossRef**]

16. Zhang, A.; Song, S.; Wang, J.; Yu, P.S. Time series data cleaning: From anomaly detection to anomaly repairing. Proc. VLDB Endow. **2017**, 10, 1046–1057. [**Google Scholar**] [**CrossRef**]

17. Wang, Y.; Zuo, W.; Wang, Y. Research on Opinion Spam Detection by Time Series Anomaly Detection. In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer Nature: Cham, Switzerland, 2019; Volume 11632, pp. 182–193. [**Google Scholar**] [**CrossRef**]

18. Makkar, A.; Kumar, N. Cognitive spammer: A Framework for PageRank analysis with Split by Over-sampling and Train by Under-fitting. Future Gener. Comput. Syst. **2019**, 90, 381–404. [**Google Scholar**] [**CrossRef**]

19. Hau, Z.; Lupu, E.C. Exploiting correlations to detect false data injections in low-density wireless sensor networks. In Proceedings of the CPSS 2019 5th on Cyber-Physical System

Security Workshop, Auckland, New Zealand, 8 July 2019; Volume 19, pp. 1–12. [**Google Scholar**] [**CrossRef**]

20. Mehrdad, S.; Mousavian, S.; Madraki, G.; Dvorkin, Y. Cyber-Physical Resilience of Electrical Power Systems Against Malicious Attacks: A Review. Curr. Sustain. Energy Rep. **2018**, 5, 14–22. [**Google Scholar**] [**CrossRef**]

21. Prasad, N.R.; Almanza-Garcia, S.; Lu, T.T. Anomaly detection. Comput. Mater. Contin. **2009**, 14, 1–22. [**Google Scholar**] [**CrossRef**]

22. Risteska Stojkoska, B.L.; Trivodaliev, K.V. A review of Internet of Things for smart home: Challenges and solutions. J. Clean. Prod. **2017**, 140, 1454–1464. [**Google Scholar**] [**CrossRef**]

23. Bakar, U.A.B.U.A.; Ghayvat, H.; Hasanm, S.F.; Mukhopadhyay, S.C. Activity and anomaly detection in smart home: A survey. In Smart Sensors, Measurement and Instrumentation; Springer International Publishing: Cham, Switzerland, 2016; Volume 16, pp. 191–220. [**Google Scholar**]

24. Massana, J.; Pous, C.; Burgas, L.; Melendez, J.; Colomer, J. Short-term load forecasting in a non-residential building contrasting models and attributes. Energy Build. **2015**, 92, 322–330. [**Google Scholar**] [**CrossRef**]

25. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [**Google Scholar**] [**CrossRef**]

26. Ruiz-Abellón MD, C.; Gabaldón, A.; Guillamón, A. Load forecasting for a campus university using ensemble methods based on regression trees. Energies **2018**, 11, 2038. [**Google Scholar**] [**CrossRef**]

27. Aharonu, Mattakoyya, et al. "Entity linking based graph models for Wikipedia relationships." *Int. J. Eng. Trends Technol* 18.8 (2014): 380-385.

28. Srinu, Nidamanuri, Sampathi Sivahari, and Mastan Rao Kale. "Leveraging Radial Basis Function Neural Networks for Rainfall Prediction in Andhra Pradesh." *2022 International Conference on Computer, Power and Communications (ICCPC)*. IEEE, 2022.

29. Muppavarapu, Rajasekhar, and Mastan Rao Kale. "An Effective Live Video Streaming System."