Feature Selection based False Data Detection Scheme using Machine Learning for Power System

Deboleena Chakraborty
Dept. of Electrical Eng.
MNIT Jaipur
India
2019psm5441@mnit.ac.in

Ajay Kumar Verma
Dept. of Electrical Eng.
MNIT Jaipur
India
2021ree9532@mnit.ac.in

Satish Sharma
Dept. of Electrical Eng.
MNIT Jaipur
India
satish.ee@mnit.ac.in*

Rohit Bhakar
Dept. of Electrical Eng.
MNIT Jaipur
India
rbhakar.ee@mnit.ac.in*

***Abstract*—An electrical power grid is a conglomerate system that requires meticulous monitoring to ensure uninterrupted, secured and reliable grid operation by incorporating state estimation to ensure a better estimate of the power grid state through assessment of meter quantification. The state estimator operates on real-time inputs that are data and status information. Thereby, it becomes necessary to automatize and digitize the electric grid by enhancing the widespread installation of Remote Terminal (RTUs) and Phasor Measurement Units (PMUs) for improvising real-time wide-area system monitoring and control. However, the challenge of anomaly detection of the data obtained from the PMUs still exists as the PMUs data comprises different types of anomalies arising from both physical and cyber systems. This work proposes a machine learning-based scheme to detect the anomaly in the data. Principal Component Analysis algorithm is used as the feature selection algorithm to attain the important characteristics in the data and then a supervised classification algorithm is used to obtain the erroneous data in the PMU data streams.**

***Index Terms*—Phasor Measurement Units, Machine Learning, Principal Component Analysis, Logistic Regression**

I. Introduction

The rapid transformation of the traditional electricity grid into a "smarter grid" involves procuring the cyber as well as the physically present components of the power system. Integration of the various technologies on the communication side like Information Communication Technologies (ICTs) helps to sustain the backbone of the infrastructure of the smart grid, as it helps to enable rapid choice-making, demand-side management, location of the faults and mitigation, and precise monitoring of power system assets. The introduction of ICTs for power system utilities has increased awareness of the situation and introduced new setbacks to the smart grid. One prominent problem is false and inaccurate data obtained by various devices. Sensor data is gathered via the Supervisory Control and Data Acquisition (SCADA), PMUs and RTUs. Energy management systems (EMSs) use state estimation (SE) and measurement data from SCADA to visualize the power system's operating condition accurately. The quality of the measurement data obtained by the SCADA determines how precise the states of estimation are [1] [2].

The measurement data includes the real and reactive power and bus voltages positioned at significant positions in the power system. The control center ensures centralized monitoring and control capability using EMS and SE. The output of state estimation is further analyzed by the control center by performing contingency analysis and thereby reasoning the potential functional problems in the grid and the actions to be taken henceforth. Furthermore, innovative methodologies have been coined to detect and identify bad metered measurements, introduced by arbitrary and nonrandom causes. Given the current configuration of the power system including the conventional algorithms for bad data detection is known to the attacker, the system becomes vulnerable to the attacker as the safeguard protocols can be circumvented [3].

Most of the sensors and controllers on the grid were fabricated without cyber security considerations, thereby leaving the network at risk of loss of economy, damage to assets, or incessant blackouts. Modern-day research and findings in cyber-attack events and grid resiliency study these factors isolated. Hence, an inclusive and more idealized way is required to treat cyber threats against power system networks. These approaches aim to merge both the threat mechanisms in the ICT and the layers' physical impact on the power system [4].

A conclusive analysis and review of applying big data and machine learning in the smart grid is conducted in [5]. The idea and the need for reliable communication in this system bring forth a huge volume of information in the form of data that requires innovations over the traditional methodologies for rigorous analysis and decision-making. The Internet of things (IoT) added with the idea of smart grid provides cost-effectiveness to a large extent. Machine learning techniques are highly beneficial to use the advantages of the IoT. A cyber-constrained optimal power flow model for the urgent response from smart grids is presented in [6] considering the various impacts of cyber networks over the electrical power grids and vice-versa.

Prominent analysis of the data provided by highly performing devices like PMUs can help to cater to many power system anomalies [7]. This Data can be used in various protective applications, like fault detection and voltage stability analysis [8]. The challenge of malicious injection identification is presented in [9], and the distorted signals are recovered using a reliable PCA-based technique. Different harmful injection attack patterns are considered, and the impact of their al-

978-1-6654-9291-1/22/$31.00

teration and refabrication using the method is examined on an application for wide-area oscillation monitoring. A novel approach is proposed for various attack detection in the cyber domain, based on trees based algorithm and kernel principal component analysis for feature reduction [10]. It overcomes the problems of wrong control decisions compromising the security of the smart grid which leads to losses in finances, disturbances in the power network, or both. A detection algorithm is proposed in [11], to analyze the real-time False Data Identification (FDI) in the region, based on Principle Component Analysis (PCA). It furnishes an optimum solution for the FDI problem through its ability to get an idea about the correlation of the data. This provides an identical answer to the preceding FDI detection methods [12].

The data received from the power system through various devices should be checked with a higher importance in order to maintain a healthy electric grid system. Any anomaly in the data can have malicious consequences in the grid infrastructure. To check such anomalies, machine learning tools and algorithms can be used effectively [13]. This paper proposes a feature selection-based false data detection scheme using PCA for dimensionality reduction. Different machine-learning techniques are used as classification algorithms. The paper organizes as follows. Section II formulates the State Estimation problem and the different Machine Learning algorithms and terminologies. In Section III, the Logistic Regression model with PCA is presented. Section IV discusses the data handling with model selection and evaluation. The obtained results are presented in Section V followed by the conclusion in Section VI.

II. Problem Formulation

State estimation(SE) accurately represents an electrical power grid by estimating the states that include voltages and angles using the measurements available. It is regarded as the epicenter of SCADA & EMS systems, as the state evaluation plays a significant role in the whole power system operation. The obtained measurements contain sensor errors and noise. To approximate the state identically, state estimation often requires unessential measurements to minimize telemetered errors. Active and reactive power injection, transmission line flows, and generator bus voltages are used to estimate the electrical grid's states.

In the power system, where N is the total number of buses, $2N-1$ is the total number of the states to be estimated, which is because $N-1$ angles (One of them is ref. angle) and voltage magnitudes are N.
Thus, the state vectors are given as

$$x = [\Theta_1 + \Theta_2 + ... + \Theta_n] \tag{1}$$

The following is a measuring function for actual and reactive power injection measurements:

$$P_i = V_i \sum_{j=1}^{N} V_i(G_{ij}\cos\Theta_{ij} + B_{ij}\sin\Theta_{ij}) \quad \text{for} \quad i = 1, 2, ..., N \tag{2}$$

$$Q_i = V_i \sum_{j=1}^{N} V_i(G_{ij}\sin\Theta_{ij} - B_{ij}\cos\Theta_{ij}) \quad \text{for} \quad i = 1, 2, ..., N \tag{3}$$

The actual and reactive power flow measurements for the line connecting buses i and j are specified as

$$P_{ij} = V_i^2(g_{si} + g_{ij}) - V_iV_j(g_{ij}\cos\Theta_{ij} + b_{ij}\sin\Theta_{ij}) \tag{4}$$

$$Q_{ij} = -V_i^2(b_{si} + b_{ij}) - V_iV_j(g_{ij}\sin\Theta_{ij} - b_{ij}\cos\Theta_{ij}) \tag{5}$$

Measurements of active/reactive powers injection and power flow are both included in the measurement function $h(x)$ for ac SE. The power system values are obtained for sets of measurements "z" by minimizing the weighted least square error as.

$$J(x) = \sum_{i=1}^{m} (z_i - h_i(x))^2/\sigma_i^2 \tag{6}$$

State estimation operates on real-time inputs that is the data and status information. The On/Off status of the various switching devices determines the network configuration that changes whenever the devices operate. The RTUs and PMUs are installed on the system to administer the changes and monitor the following 1) the analog data in the form of MW and MVAr flows through the major lines, 2) the loading of active and reactive power of generators and transformers, and 3) magnitudes of the bus voltages of the system. The remote units and the entire set of measurements are scanned every few seconds and are received by the energy control center. It can be concluded that the state estimator extracts a larger database. To identify the false data obtained from SE, Machine learning (ML) techniques are very effective.

A. Machine Learning Approach

Machine Learning Classification algorithms are supervised learning techniques. In supervised learning, the output variables are categorical or like a boolean in nature that is 0 or 1, Yes or No. In such cases, classification algorithms are used to separate or demarcate the output categorical variables. They are used to classify the variables as 0 or 1 etc. These output variables are also called target variables or labels. Thus, the objective of the Machine Learning Classification algorithms is mainly, to classify the target variables and to predict them later. A discrete output function, y in a Machine Learning Classification algorithm can be written as:

$$y = f(x) \tag{7}$$

where the output function, y is categorical in nature.

Machine learning classification algorithms are comprised of Support Vector Machine(SVM), Logistic Regression, Naive Bayes algorithm and K-Nearest Neighbour etc.

1) Support Vector Machine: A supervised learning algorithm is used in machine learning to solve classification problems. This Classification algorithm aims to design the best line or boundary to divide an N-dimensional space into smaller subclasses to hold a new data point in the appropriate category for the future. This boundary is termed the hyperplane. It becomes important for SVM to identify the points to create the hyperplane. It is divided into two: Linear and Non-Linear SVM. The first one separates data linearly, that is, with the help of a single straight line. However, the second one is for the data that are non-linearly separable [14].

2) Logistic Regression: It is identified as the output categorical target variable by a set of independent variables [15]. The output of this algorithm is discrete, that is, 0 or 1, because it gives the probabilistic values of the corresponding input features. It uses an S-shaped sigmoid curve to classify the output based on a threshold value. Values more than the threshold value are flagged as 1 or Yes, and values less than the threshold value are flagged as 0 or No. This algorithm is highly significant because it produces the output as probabilities and classifies the input variables using discrete datasets.

$$log[y/1-y] = b_1x_1 + b_2x_2 + b_3x_3 + ... + b_nx_n \quad (8)$$

3) Naive Bayes algorithm: This algorithm is developed from the Bayes theorem. It is fundamentally useful when a high-dimensional training dataset is to be dealt with. It focuses on building fast machine-learning models that can make quick predictions. It is also a probabilistic classifier similar to the logistic regression ML classification algorithm. It evaluates the output variable based on probability. It is based on the Bayes theorem and evaluates the probability of a hypothesis with previous background, using the concept of conditional probability.

4) K-Nearest Neighbour: It assumes similarity between the new dataset and the previous ones, and the new case is put in the same category as the previous one. It then captures the available data, and a unique data point is classified similarly. It is a non-parametric algorithm that does not consider any pre-assumptions. Its working can be explained by first selecting the K number of the neighbours, followed by calculating the corresponding Euclidean distance. Now, the data points of the K nearest neighbours are based on the calculated Euclidean distance. Based on the neighbours, the data points are counted and then assigned to the new category.

B. Parameters of Machine Learning Algorithms

1) Confusion Matrix: When subjected to test data, the accuracy of the ML classification algorithm models is evaluated through the confusion matrix. The matrix has two dimensions, one being predicted values and the other absolute values and the total number of predictions as shown in Fig 1. The assessment of classification models is done based on the results obtained from test data. It enables calculating all the parameters for the model, such as accuracy, precision, recall, etc [16].

- **True Negative [TN]:** Prediction of the Model gives No, when the real or actual values are No.
- **True Positive [TP]:** Prediction of the Model gives Yes, when the real or actual values are Yes.
- **False Negative [FN]:** Prediction of the Model gives No, but the real or actual values are Yes.
- **False Positive [FP]:** Prediction of the Model gives Yes, but the real or actual values are No.

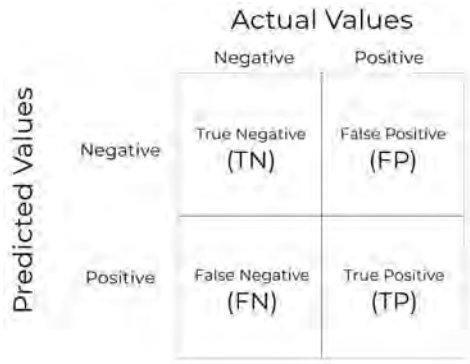

Fig. 1. Confusion Matrix

2) Classification Accuracy: The parameter evaluates the accuracy of the various classification problems. It measures how frequently the correct output is predicted by the model. The ratio of the total number of predictions to all the predictions obtained correctly by the classifier gives the Classification Accuracy. Accuracy can be measured by the following:

$$ClassificationAccuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

3) Precision: Precision measures the output counts correctly identified by the model or out of all classes. Precision can be calculated by the following formula:

$$Precision = \frac{TP}{TP+FP}$$

4) Recall: It is a measure of the total positive classes, predicted by the model correctly. The formula is given below:

$$Recall = \frac{TP}{TP+FN}$$

5) F-measure: When two models have a high recall or low precision or vice versa, the comparison of these models is not possible. Hence, F-score is used to measure recall and precision together. The value of the F-score gets maximized when the measure of recall is equal to the precision [16]. It can be calculated using the below formula:

$$F-measure = \frac{2*Precision*Recall}{Precision+Recall}$$

III. Proposed Algorithm

With the power system's growth, the size or number of features in the PMU dataset increases, and a reduction in the number of dimensions becomes integral in data mining to reduce the estimation costs, thereby improving the classifier's

efficiency, which can be affected by dispensable features. Principal Component Analysis (PCA) [17], is used to reduce the dimensionality of data streams and increase expressibility while concurrently reducing the information loss. PCA does it by creating new uncorrelated variables that eventually maximize variance. Obtaining the new variables, the principal components, helps obtain a solution to an eigenvalue or an eigenvector problem. The new variables given by the new data stream at hand, not a priori, make PCA an efficient data analysis tool.

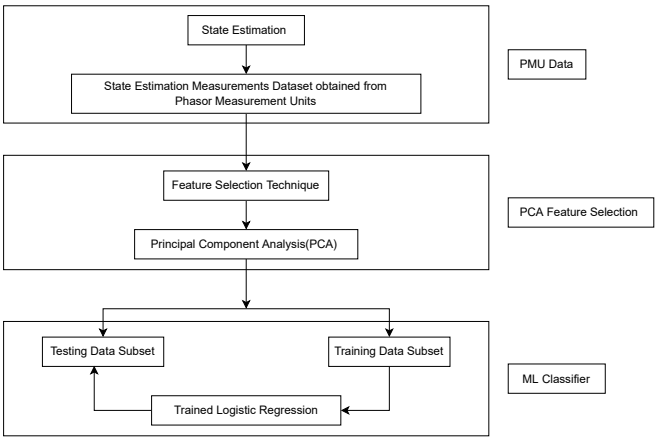

Fig. 2. Main Flowchart of the Proposed Machine Learning Approach

From Fig 2, the main flowchart of the proposed work is represented. As discussed in the previous sections, firstly, the state estimation measurement dataset is used for anomaly detection. The feature selection technique follows it, and the PCA is used to obtain the distinct features. The new dataset is then subjected to the Logistic Regression Model. A subset of the data is initially trained on the model and is then tested to identify the efficiency of the model. Exploratory Data Analysis is done that includes all the data pre-processing steps. The categorical column "status," which implies the status of the grid, is converted to binary values for binary classification. The data is separated into two data frames, X and Y, which contain voltage, current magnitudes, phase angles, and status values. The X data frame is bifurcated into train and test variables. The model is first evaluated with the train variables and then tested on the test variables. The correlation matrix is observed to check the variables with maximum correlation. It is followed by creating the Logistic Regression Model. Before building the model, the features of the data are checked using the Recursive Feature Elimination (RFE) technique. On observation, none of the features was to be eliminated. The logistic regression model is then fitted on the trained variables. The model evaluation is done by obtaining the predicted status values corresponding to the trained values. A data frame is created with the actual Status value, and the corresponding predicted probabilities.

IV. Data handling and Model Selection

The first step in the problem formulation is data analysis. As in the previous section, it was mentioned about the importance of state estimation. It is essential for the maintenance of the status of the grid. A similar data stream to quantify the above said, used in our case, is mentioned in the following section.

A. Dataset

The data set is obtained from a collection of power measurements and annotations available in the data repository at https://powerdata-explore.lbl.gov. Phasor measuring units (PMUs) [Powa, VMCMA14] and PQube3 power quality metres [Powb] made by the Lawrence Berkeley National Laboratory's Power Standards Laboratory in Alameda, California, are used to measure the different types of power. The dataset consists of current, voltage and power columns along with a column that includes the status of the grid, stating whether there is an attack or whether the data was malicious or not. The data are correctly identified as defective or not, which are in the column in the status as Y (Yes) and N (No). The status column Y/N is converted to 1/0 to make analyzing easier.

B. Training and Testing of Dataset

The Machine Learning Classification Algorithms from section 2 is applied to training data. The Algorithms are then evaluated on another part of the data called testing data. The purpose of supervised machine learning model creation is to evaluate and validate the model. To make the process unbiased, the performance for prediction of the model evaluation is essential. The data science library scikit-learn has a train-test-split feature to split the dataset into subsets. Bifurcating the dataset is essential for evaluating the model's prediction performance. The test set is needed for an evaluation in an unbiased manner of the final model. In this case, the test case is taken as 30% of the entire data. The model is trained on 70% of the data. Once the model is trained and required parameters like accuracy, F1-Score and precision-recall are obtained, it is then tested on the 30% of the test data. Thereafter, the same parameters are obtained for the test data. If the results are similar, it can be concluded that the model works well.

C. Model Selection and Evaluation

The division into two subsets of the data stream is obtained from the PMU: testing and training data. The training data is firstly subjected to all the models of Machine Learning algorithms as mentioned above. The parameters Classification Accuracy, Precision and Recall, are obtained. The same is observed in the testing data, and the values of parameters obtained are similar to that obtained from the training data. The value of the parameter obtained from the Logistic Regression Classifier [15] model is the best. The data stream is then subjected to the Principal component analysis (PCA). After the most suitable features are obtained, the updated data stream is again subjected to all the above models. It is observed that the Classification Accuracy value for the models either increased or remained the same. Here also, the most suitable

model is the Logistic Regression Classifier model. A similarity in the values of a few parameters obtained from all the models creates a dilemma to conclude.

Hence, the concept of cross-validation is used. It is a methodology for validating the model's efficiency by training it on the subset of input data and testing previous subsets of the data as input. This helps to validate the most suitable model finally. It can be said that it is a methodology to measure how a statistically enhanced model generalizes an independent dataset. Essentially in machine learning, the need to test the stability of the model is mandatory. It means that based only on the training dataset. The model cannot fit on the training dataset. Therefore, it becomes necessary for the model to fit on a new subset of the dataset that is not identical to the training dataset. The model is then tested on that subset before deployment, which sums up the concept of cross-validation. This is explained in the following chapter.

V. Results

In this paper, the proposed algorithm's results are applied to the data from Section IV-A and compared with other models with and without PCA.

A. Conclusions from F-score

The F-Score values without and with implementing PCA are presented in Tables I and II, respectively, respectively, for both the Train/Test samples. It is shown for both states, not defective(0)/defective(1). For the Logistic Regression model, there are minute differences in the values of the parameters without and with PCA. The values of Precision, F1-Score and Recall are for the defective states closer to 1, and there is no difference between the train and test values. This signifies that the model has performed well. Similar is the case for the Gaussian Naïve Bayes Model. For the K-Nearest Neighbour Model, the F-Score value without PCA for the defective state (1) does not have much difference, but it differs in the case of PCA. This signifies that the model has not performed well. For the Support Vector Machine Algorithm Model, the F1-Score value for the defective state (1) is different for both the cases without and with implementing PCA. This model hence has not performed well. Since there are similar results for the two models, therefore it becomes inconclusive. The method of Cross-Validation can help conclude this.

TABLE I
CONCLUDING TABLES FOR F-SCORE WITHOUT PCA

Model-Name	F-Score	Train/Test
Gaussian Naive Bayes Classifier	0.8	Train
Gaussian Naive Bayes Classifier	0.8	Test
Logistic Regression for Classification	0.84	Train
Logistic Regression for Classification	0.84	Test
K-Nearest Neighbors Classifier	0.83	Train
K-Nearest Neighbors Classifier	0.8	Test
SVM for Classification	1	Train
SVM for Classification	0.96	Test

TABLE II
CONCLUDING TABLES F-SCORE WITH PCA

Model-Name	F-Score	Train/Test
Gaussian Naive Bayes Classifier	0.84	Train
Gaussian Naive Bayes Classifier	0.84	Test
Logistic Regression for Classification	0.84	Train
Logistic Regression for Classification	0.84	Test
K-Nearest Neighbors Classifier	0.83	Train
K-Nearest Neighbors Classifier	0.78	Test
SVM for Classification	0.85	Train
SVM for Classification	0.83	Test

B. Conclusions for Accuracy

The Accuracy values are presented for all four models for both the train and test values, as shown in Table III. Since there is not much difference between the accuracy values for the models (Logistic Regression and Gaussian Naive Bayes Classifier), it becomes inconclusive. However, K-Nearest Neighbor and SVM cannot be considered as the accuracy for the test values are significantly lower than the train values due to overfitting. The inconclusiveness can be overcome by the Cross Validation method.

TABLE III
CONCLUDING TABLE FOR THE ACCURACY VALUES

Model-Name	Accuracy	Train/Test
Gaussian Naive Bayes Classifier	0.72	Train
Gaussian Naive Bayes Classifier	0.72	Test
Logistic Regression for Classification	0.72	Train
Logistic Regression for Classification	0.72	Test
K-Nearest Neighbors Classifier	0.73	Train
K-Nearest Neighbors Classifier	0.68	Test
SVM for Classification	0.79	Train
SVM for Classification	0.72	Test

C. Cross-validation of the models

The cross-validation of all four models is presented in Table IV. The difference between the train and test values is lowest in case of Logistic Regression. Thus we further proceed with the model of Logistic Regression. It signifies that the risk of overfitting is least in the case of Logistic Regression. An overfitted statistical model is one that when gets trained with data of a massive volume. When a model is trained with such a massive volume of data, the model learns by heart the noise that is integrated with a lot of inaccurate data entries in the data set. The model cannot, therefore, classify the data identically, as it learns all the random noise completely. Overfitting is mainly caused when the non-parametric and not linear methods are used since these machine learning algorithms have more freedom in fabricating the model on the dataset based and thus can really build not-so-reliable models. In order to obtain the threshold probability for LR model, a plot is drawn between the accuracy score, sensitivity (Recall) and specificity (Precision) for various probabilities, as shown in fig. 3. The threshold is obtained as 0.65. This signifies that the predicted probabilities beyond the threshold value that is greater than 0.65 will be flagged as erroneous data.

TABLE IV
CROSS-VALIDATION OF THE MODELS

Model-Name	Score	Train/Test
Gaussian Naive Bayes Classifier	0.7864	Train
Gaussian Naive Bayes Classifier	0.23052	Test
Logistic Regression for Classification	0.72114	Train
Logistic Regression for Classification	0.66689	Test
K-Nearest Neighbors Classifier	0.67648	Train
K-Nearest Neighbors Classifier	0.5897	Test
SVM for Classification	0.79211	Train
SVM for Classification	0.38195	Test

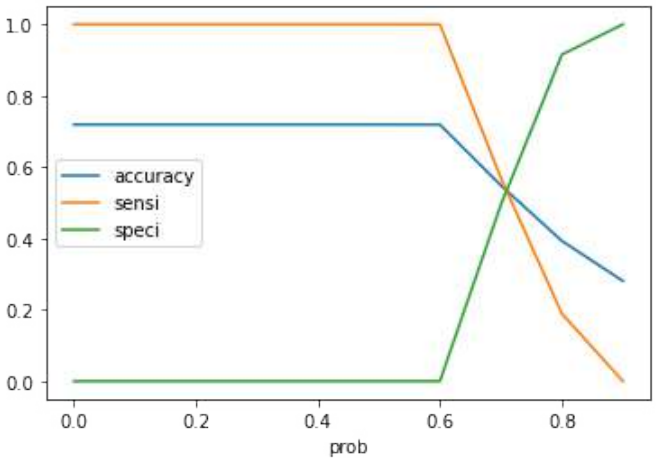

Fig. 3. Trade-Off between Accuracy, Sensitivity and Specitivity

VI. CONCLUSION

This paper proposes a PCA-based machine learning classification algorithm for identifying defective data that can be injected into the PMU data stream in smart grid communications networks. We have deployed PCA for the identification of particular and identical features. The optimally chosen features are then used as input to a logistic regression model. The focus was to show the utilization of machine learning classification algorithms for anomaly detection on the data injected by various malicious agents. This work also identifies when classification algorithms are applied to a particular dataset, and the parameters obtained from each algorithm yield similar results. It becomes inconclusive as to which classification algorithm is the most suitable one. This can happen due to many reasons and one of them being overfitting. In order to curb this, Cross-Validation is used. However, the optimal choice of the algorithm also depends on the type of data provided. The LR algorithm is most suited for the input data.

ACKNOWLEDGEMENT

This work is supported by Science & Engineering Research Board (SERB), DST (File no. SPG/2021/004202) and MHRD (India) grants for SPARC Project #1317.

REFERENCES

[1] SCADA, "Energy management systems," [online]. Available: https://inductiveautomation.com/resources/article/what-is-scada.

[2] EMS Gartner Glossar, "Energy management systems," [online]. Available: https://www.gartner.com/en/information-technology/glossary/energy-management-systems-ems.

[3] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security (TISSEC)*, vol. 14, no. 1, p. 13, 2011.

[4] C. Cameron, C. Patsios, P. C. Taylor, and Z. Pourmirza, "Using self-organizing architectures to mitigate the impacts of denial-of-service attacks on voltage control schemes," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3010–3019, May 2019.

[5] E. Hossain, I. Khan, F. Un-Noor, S. S. Sikander, and M. S. H. Sunny, "Application of big data and machine learning in smart grid, and associated security concerns: A review," *IEEE Access*, vol. 7, pp. 13 960–13 988, 2019.

[6] G. Huang, J. Wang, C. Chen, and C. Guo, "Cyber-constrained optimal power flow model for smart grid resilience enhancement," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5547–5555, Sep. 2019.

[7] V. Gayathry and M. Sujith, "Machine learning based synchrophasor data analysis for islanding detection," in *2020 International Conference for Emerging Technology (INCET)*, 2020, pp. 1–6.

[8] A. Karpilow, R. Cherkaoui, S. D'Arco, and T. D. Duong, "Detection of bad pmu data using machine learning techniques," in *2020 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2020, pp. 1–5.

[9] K. Mahapatra and N. R. Chaudhuri, "Malicious corruption-resilient wide-area oscillation monitoring using online robust pca," in *2018 IEEE Power Energy Society General Meeting (PESGM)*, 2018, pp. 1–5.

[10] M. R. Camana Acosta, S. Ahmed, C. E. Garcia, and I. Koo, "Extremely randomized trees-based scheme for stealthy cyber-attack detection in smart grid networks," *IEEE Access*, vol. 8, pp. 19 921–19 933, 2020.

[11] A. S. Musleh, M. Debouza, H. M. Khalid, and A. Al-Durra, "Detection of false data injection attacks in smart grids: A real-time principle component analysis," in *IECON 2019 - 45th Annual Conference of the IEEE Industrial Electronics Society*, vol. 1, 2019, pp. 2958–2963.

[12] S. Ahmed, Y. Lee, S. H. Hyun, and I. Koo, "Feature selection–based detection of covert cyber deception assaults in smart grid communications networks using machine learning," *IEEE Access*, vol. 6, pp. 27 518–27 529, 2018.

[13] M. Esmalifalak, L. Liu, N. Nguyen, R. Zheng, and Z. Han, "Detecting stealthy false data injection using machine learning in smart grid," *IEEE Systems Journal*, vol. 11, no. 3, pp. 1644–1652, 2017.

[14] M. Somvanshi, P. Chavan, S. Tambade, and S. V. Shinde, "A review of machine learning techniques using decision tree and support vector machine," in *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, 2016, pp. 1–7.

[15] S. S. Noureen, S. B. Bayne, E. Shaffer, D. Porschet, and M. Berman, "Anomaly detection in cyber-physical system using logistic regression analysis," in *2019 IEEE Texas Power and Energy Conference (TPEC)*, 2019, pp. 1–6.

[16] W. Zhe, C. Wei, and L. Chunlin, "Dos attack detection model of smart grid based on machine learning method," in *2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 2020, pp. 735–738.

[17] H. Abdi and L. J. Williams, "Principal component analysis," *WIREs Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.