Data Driven Machine Learning Model for Condition Monitoring and Anomaly Detection in Power Grids

Komal Saleem*, Bugra Alkan* and Sandra Dudley-McEvoy*
*School of Engineering
London South Bank University, 103 Borough Rd, SE1 0AA, London, United Kingdom
Email(s): saleemk2@lsbu.ac.uk, alkanb@lsbu.ac.uk, dudleyms@lsbu.ac.uk

***Abstract*—The power system complexity and associated stability problems are greatly linked to the increasing penetration of unconventional energy sources and loads, such as renewable energies. The application of renewable for climate change, sustainability, and Net Zero come at the cost of deteriorated power quality, faults, instability, and disturbances in the power system. It gives rise to various problems such as equipment malfunctioning, power factor problems, transformer heating, inertia, voltage sags/swells, transmission lines overloading, etc. This requires and adjudicates the need for efficient monitoring and identification of faults and anomalies happening in the power system so as to accordingly mitigate these in a timely manner. The fault data however is not readily available and requires on-site inspection and accumulation. This paper thus aims at developing a synthetic database for various abnormal power system conditions captured from a well-known Kundr's two-area system. These include symmetrical and asymmetrical faults, frequency, and phase variations, as well as voltage amplitude disturbances (sag/swell). The synthetic database is then combined with artificial intelligence techniques to enable fault detection and identification featuring low linear complexity and small memory requirements. The paper includes a benchmark study for three unsupervised anomaly detection algorithms, evaluating their performance in terms of both Area under the ROC Curve (AUC) and the execution time. The results show that iForest and iNNE provide competitive results in detecting anomalies of all fault types, with iNNE providing significantly better execution time performance.**

***Index Terms*—Renewable penetration, power system faults, grid disturbances, anomaly detection, data analytics, unsupervised learning.**

I. Introduction

The emerging integration of smart grid technologies such as renewables, electric vehicle, etc. into the power system over the recent years has grab researchers' attention towards data acquisition, system monitoring, control, and protection [1]. Despite the positive impact of these technologies towards net zero and grid support, they may also contribute to system disturbances and faults, which deteriorate power quality and results in system instability. Phasor measurement units (PMUs) are thus used to record synchronised measurements of power system dynamics and provide great potential for real time monitoring of an electrical grid [2]. The data is clock synchronized using Global Position System (GPS). The data recorded by PMUs is huge and is significantly challenging to monitor, record, and analyse system disturbances, affecting the operation and stability of power system enormously. Therefore, to cope with these challenges, machine learning and artificial intelligence techniques are being used for data processing and advance anomaly detection that helps in preventing power system blackouts.

In the literature, methods for detecting system faults and anomalies are classified into three categories: supervised, semi-supervised, and unsupervised. Supervised methods use labelled data, such as 1 for a fault and 0 for no fault, whereas semi-supervised algorithms use both labelled and unlabeled data. Unsupervised methods, on the other hand, do not require labelled data during the process. Labelling is often considered as a time-consuming process; therefore, unsupervised approaches have received a significant attention from the scientific community as a viable alternative to labelled data analysis [3]. Unsupervised anomaly detection techniques are roughly classified into four primary domains [4]: i) clustering-based approaches, ii) density-based approaches, iii) relative density-based approaches, and iv) ensemble-based approaches. Clustering techniques group data points with similar feature values into the same cluster using a distance function, such as the Euclidean, Hamming, or Minkowski distance. The distance to the cluster's centroid is then used to calculate the scores for each data point that is contained within the cluster. As a result, data points that are significantly offset from the cluster centres are labelled as anomalies. Data points are classified as anomalies in density-based approaches based on whether or not they are in low-density zones. These methods frequently compute anomaly scores by substituting density for nearest neighbour (NN) distance. According to [5], these methods fail to account for local anomalies that exist in high-density areas but have a low density score in comparison to their neighbours. This gap is filled by relative-density techniques, which define anomalies as data points with a low relative density in comparison to their neighbours. When it comes to detecting anomalies, these methods frequently use the ratio of an instance's density to the density of its surroundings as a measure of relative density. Both the density and relative density approaches have a number of limitations, the most significant of which are: i) an increase in computational complexity, especially when working with high-dimensional datasets; and ii) a sensitivity to the size of neighbouring clusters [6], [7]. An ensemble-based approach first applies a

This work is supported by Innovate UK (IUK) smart grant under project reference 10004690.

set of methods (such as decision-trees) to a number of distinct subsets of a dataset, and then employs a voting mechanism that takes all of these subsets into account to arrive at an overall anomaly score. The main disadvantage of these approaches is that they are computationally expensive. Isolation-based approaches begin with the presumption that anomalies are more prone to being isolated. As a result, these approaches make an effort to designate datapoints within a given dataset as anomalies based on a metric that evaluates the datapoints' propensity to be isolated. According to [8], these methods do not necessitate time-consuming and expensive NN queries. This is the primary advantage that makes their linear time complexity possible.

The motivation of this research is early anomaly detection and identification to minimise the instability issues in power systems. The main contribution of this paper are (i) is developing Kundur's two area systems real time model in speedgoat, creating synthetic database for different symmetric and asymmetric faults, (ii) using developed database for unsupervised anomaly detection algorithms such as, Rapid Distance-Based Outlier Detection, Isolation Random Forest, and Isolation-based Nearest Neighbor Ensemble, (iii) benchmarking of anomaly detection algorithms through analysis and comparative assessment of results.

II. System Model

A Kundur's two area system is used to generate various type of faults so as to develop a synthetic database. The system is well known for its use in testing and analysing the power system dynamics and their impact on the overall operation of network [9]. It has two areas each having two generating stations and are connected to each other through a transmission line, as can be seen from Fig. 1. The system ratings and parameter values have been provided in Table I and Table II. The p.u. base values for generators and transmission lines are $S^G_{base} = 900MVA$, $V^G_{base} = 20kV$ and $S^T_{base} = 100MVA$, $V^T_{base} = 230kV$, respectively.

Different type of symmetrical and asymmetrical faults are added to the system at bus 7 and their resulting effect on system voltages and currents are recorded by using Phasor Measurement Unit (PMU) and stored in the synthetic database.

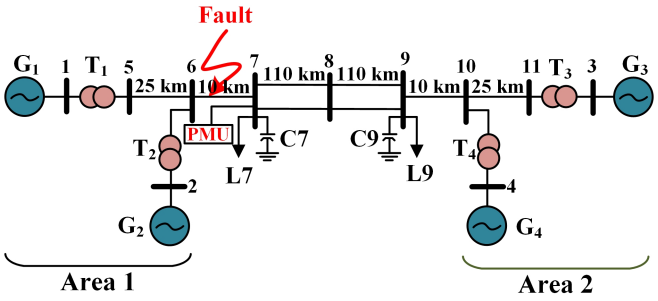


Fig. 1: The Kundr's two area system.

TABLE I: System ratings for Kunder's two area system.

Power system components						
Generators				Transformers		
	P(MW)	Q(MVAr)	V(p.u.)		S(MVA)	V(kV)
G_1	700	185	$1.03\langle 20.2°$	T_1	900	20/230
G_2	700	235	$1.01\langle 10.5°$	T_2	900	20/230
G_3	719	176	$1.03\langle -6.8°$	T_3	900	20/230
G_4	700	202	$1.01\langle -17°$	T_4	900	20/230

TABLE II: System parameters for Kunder's two area system.

System Parameters (p.u.)	Values	System Parameters (p.u.)	Values
X_d	1.8	H_1	6.5
X_d	1.8	H_2	6.175
X_q	1.7	X_T	0.15
X'_d	0.3	r_{line}	0.0001
X''_d	0.25	x_{line}	0.001
R_a	0.0025	b_{line}	0.00175

III. Fault Data Models and Definitions

The data model is developed by creating a synthetic database consisting of various power system faults by introducing these faults to a power system bus in Kundr's two-area system. PMUs are used because they provide great potential for monitoring an electrical grid by recording synchronized current voltage and frequency measurements.

Worth mentioning, in the normal operating state, all the conductors in power systems exhibit coherent properties. Any dissimilarity and deviation indicate the existence of anomalies in the system. These anomalies are referred to as faults and have a negative impact on the overall operation and stability of the system. A number of faults occur and exist in the power system, but in this work symmetrical, asymmetrical short circuit faults, magnitude variation faults, frequency as well as phase variations are considered, briefly explained below.

A. Asymmetrical faults

Asymmetrical faults make the power system unbalanced and result in oscillating active/reactive powers - strike usually 1 or 2 phase(s) of a three-phase system. These faults also result in the flow of abnormal high currents through the equipment or transmission lines. If these faults are allowed to persist even for a short period, it leads to extensive damage to the equipment. Different types of asymmetrical faults are briefly explained below:

1) Single phase to ground fault: The most commonly occurring fault (70% to 80%) in power systems is the single phase to ground fault in which one out of the three-phase conductors is grounded by a temporary or permanent cause (such types of failures may occur due to lightning, falling off a tree, high-speed wind, etc.). An example is shown in Fig. 2a, in which phase$-a$ suffers from a phase-to-ground fault at 0.06 s lasting for 0.08 s where the voltage is reduced to 20% and the phase current is increased enormously (a peak of ≈400 A).

2) Two phase to ground fault: When two phases came in contact with the ground, the fault is named as two-phase to-ground fault as shown in Fig. 2b. It results in voltage reduction

and significant increases in the current magnitude for the faulty phases and also, the flow of ground current becomes considerable. The probability of occurrence of this fault is 10%. As seen in Fig. 2b, with normal initial conditions, the two phases (a and b) are subjected to a fault at 0.06 s which results in a drop of 30% voltage and a rise of current (300 A peak).

3) Phase to phase fault: When two phases came in contact with each other, phase-to-phase fault occurs, shown in Fig. 2c. Heavy winds are the major cause of this fault during which the swinging of overhead conductors may touch together. However, it is less severe in nature and the percentage of occurrence of this type of fault is between 15% to 20%. In the examples provided in Fig. 2c, a voltage reduction of 20% is observed between the fault period of 0.006 s to 0.0014 s.

B. Symmetrical faults

Symmetrical faults also referred to as balanced faults do not affect the symmetry of the power system and occur when all three phases are simultaneously short-circuited.

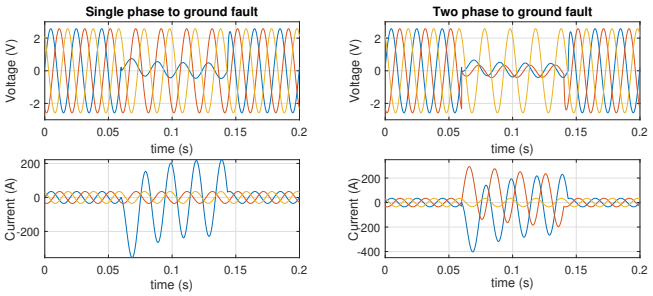
1) Three phase to ground fault: This type of fault occurs when all three phases of the power system are grounded simultaneously, as shown in Fig 2d. The voltage approaches almost zero and a significant rise in currents is observed. This type of fault is the least occurring fault nearly 2% to 3%. However, even though the system remains in a balanced condition, these faults may result in severe damage to the equipment. As in the testing phase, these faults help in identifying the optimal size of protective devices (such as circuit breakers, etc.).

2) Frequency variation faults: It is important to maintain a stable power system frequency because the equipment and appliances are designed to operate at a certain frequency (e.g. 50 Hz). Any variation in the frequency is referred to as the frequency fault, an example is shown in Fig. 3a. These variations occur as a result of short or open circuit faults and sudden load variations. For instance, increased demand for electricity would result in a drop in frequency and likewise, a sudden drop in load would rise the system frequency. Thus, it is important to monitor and take appropriate actions to restore system frequency.

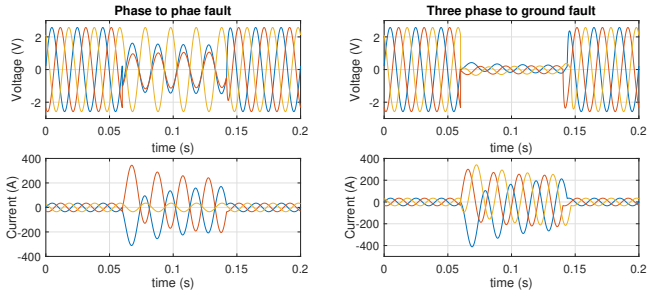
3) Voltage amplitude variation faults:: The short-duration increase or decrease in the voltage amplitude is referred to as voltage faults [10]. An increase in amplitude beyond nominal value (caused by the disconnection of large load) is referred to as the voltage swell and likewise, amplitude decrease is signified as voltage sag (usually caused by inrush currents). These surges are dangerous for sensitive equipment (e.g, computers, controllers, etc.) and result in malfunctioning. An example of voltage swell is shown in Fig. 3b.

IV. Selected Anomaly Detection Algorithms

Within the scope of this investigation, three anomaly detection algorithms featuring a low linear time complexity and small memory requirement have been taken into consideration. These are Rapid Distance-Based Outlier Detection (S_p), Isolation Random Forest (iForest) and Isolation Using Nearest Neighbor Ensemble (iNNE).

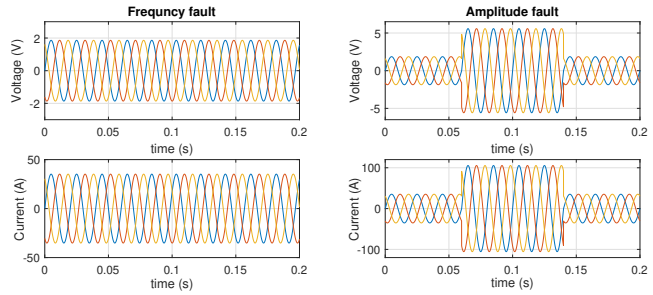


(a) Single phase to ground fault (b) Two phase to ground fault



(c) Phase to phase fault (d) Three phase to ground fault

Fig. 2: Examples of asymmetrical and symmetrical power system faults.



(a) Frequency fault (b) Amplitude fault

Fig. 3: Frequency and amplitude variation faults.

A. Rapid Distance-Based Outlier Detection (S_p)

The S_p algorithm, which was initially proposed by [11], performs a single sampling and computes the anomaly score of an instance $x \in \Re^d$ by employing the NN distance in the following way:

$$S_p(x) = \min_{y \in S} \parallel x - y \parallel \tag{1}$$

Despite the use of a relatively modest sample size, S_p has been found to outperform k-NN-based algorithms in terms of both time complexity and accuracy [8].

B. Isolation Random Forest (iForest)

The iForest, initially proposed by [12], is an isolation-based methodology that makes use of ensemble trees (also known as isolation trees or iTree) that are constructed through the use of randomly selected subset of size ψ. Isolation trees are made up of nodes, each of which carries out a random split on an attribute of a feature space. The split point is a

randomly selected real value that is positioned between the lowest possible value and the highest possible value of the selected attribute in the sample. The iForest method iteratively applies this procedure to all subsets until one of them becomes a singleton, yielding a binary tree with a depth of $2\psi - 1$. Here, anomaly score of a data point (x) is measured using path length h(x) on the tree, and the score is normalized and averages on the number of iTrees as follows:

$$Q_{tree}(x) = 2^{-\overline{h(x)}/c(\psi)} \tag{2}$$

where $\overline{h(x)}$ is the average of $h(x)$ on t iTrees and $c(\psi)$ is defined as $c(\psi) := 2H(\psi 1)2(\psi 1)/t$, where H denotes the harmonic number. iForest has been found to be very efficient in detecting anomalies while maintaining a linear time complexity [12].

C. Isolation-based Nearest Neighbor Ensemble (iNNE)

iNNE, which is an isolation-based approach that was primarily proposed by [8], is quite similar to iForest in that it isolates instances in a sub-sample and then constructs an ensemble from several sub-samples. iNNE includes two main stages: *i)* training stage and *ii)* evaluation stage. During the training stage, the algorithm creates t number of randomly selected hyperspheres from subsamples of size of ψ. In evaluation stage, each test instance is evaluated against t sets of hyperspheres, and the isolation scores are averaged to produce the anomaly score. The isolation score $I(x)$ for instance (x) is calculated using the below formula:

$$I(x) = \begin{cases} 1 - \frac{\tau(\eta_{cnn(x)})}{\tau(cnn(x))}, & \text{if } x \in \bigcup\limits_{c \in S} B(c) \\ 1, & \text{otherwise,} \end{cases} \tag{3}$$

where, $cnn(x) = \underset{c \in S}{argmin}\{\tau(c) : x \in B(c)\}$, and $B(c)$ is a hypersphere centred at c and isolates instance x from the rest of the instances in S. This hypersphere's radius $\tau(c)$ is a measure of the degree of isolation of c. The larger the radius, the more isolated c is, and vice versa. Finally, the anomaly score is calculated as follows:

$$\overline{I} = \frac{1}{t} \sum_{i=1}^{t} I_i(x) \tag{4}$$

According to [8], iNNE algorithm is more memory efficient and runs much faster, especially on large data sets with thousands of dimensions or millions of instances, than NN-based methods like Local outlier factor (LOF) (see [5] for LOF).

V. Empirical Study

A. Experimental Setup

All of the experiments are carried out on a computer that has an Intel(R) Core(TM) i7-10850H CPU operating at 2.70 GHz, 2712 MHz, and having a total of 32 GB of installed physical memory. The overall data consists of a total of 180 datasets (30 randomly generated datasets for each fault type), each of which includes 6 channel data with 600k data points and corresponding labels for the purposes of validation. The datasets are normalised utilising a method known as min-max normalisation because this is necessary for the distance and density based methods, which call for normalised data input. The area under the ROC (Receiver Operating Characteristics) curve (or simply AUC) (see [13]) is the metric that is utilised as the standard for measuring the accuracy of anomaly detection, and the execution time is the metric that is utilised for comparing the effectiveness of each approach. It is important to keep in mind that the iNNE, iForest, and S_p techniques are all randomised. As a result, the results of their AUC are reported as an average based on 20 separate tests utilising a variety of random seeds.

Noteworthy, the Kundur's two area system model is developed in MATLAB Simulink with PMU connected to bus 7 and validated in real-time using Speedgoat real-time simulator for various faults and grid conditions. Furthermore, the data-driven algorithms are also benchmarked in MATLAB R2022b environment using the created synthetic datasets. In order to tune the algorithm parameters, a parameter selection method based on a grid search has been implemented. Throughout the course of the evaluations, the value of ψ was searched for within the range of 2, 4, 8, 16, and 32 for the iForest and iNNE algorithms, whereas the total number of clusters for these algorithms was changed from 50 to 300 (with an increment of 10). The efficiency of S_p method was searched for a wide variety of sample sizes, with the possible range falling anywhere from 10 to 100 (with an increment of 10).

B. Results

The box-plot depiction shown in Figure 4 displays the AUC results of six different fault types obtained by the S_p, iForest, and iNNE algorithms, respectively. Table III shows the best results, standard deviation in terms of AUC, optimal parameter settings (t or ψ) and computation times, respectively.

TABLE III: Benchmark results for S_p, iForest, and iNNE algorithms on datasets of various fault types (only best models).

Fault Type	Method	Best AUC	Std.	Time (sec)	Best ψ	Best t
3 Phase to Grd.	Sp	0.79066	0.05565	0.18	10	-
	iForest	0.99978	0.00002	7.12	8	150
	iNNE	0.99993	0.00002	3.25	4	100
2 Phase to Grd.	Sp	0.74502	0.11431	0.16	50	-
	iForest	0.99991	0.00001	21.26	32	300
	iNNE	0.99993	0.00002	7.52	8	250
1 Phase to Grd.	Sp	0.81953	0.08709	0.16	10	-
	iForest	0.98631	0.00410	14.87	16	250
	iNNE	0.99982	0.00001	9.28	16	250
Phase to phase	Sp	0.76735	0.05403	0.42	50	-
	iForest	0.94693	0.01019	8.75	32	100
	iNNE	0.99989	0.00007	8.08	8	300
Amplitude Var.	Sp	0.63118	0.10645	0.16	20	-
	iForest	0.94991	0.00633	25.44	32	300
	iNNE	0.93152	0.00432	13.42	32	250
Frequency	Sp	0.96519	0.00280	0.87	100	-
	iForest	0.94501	0.00329	23.50	32	300
	iNNE	0.98059	0.00152	10.58	16	300

It has been observed that the iForest and iNNE models, with the exception of the ψ=2 variants, produced competitive AUC results for all different types of fault datasets. According

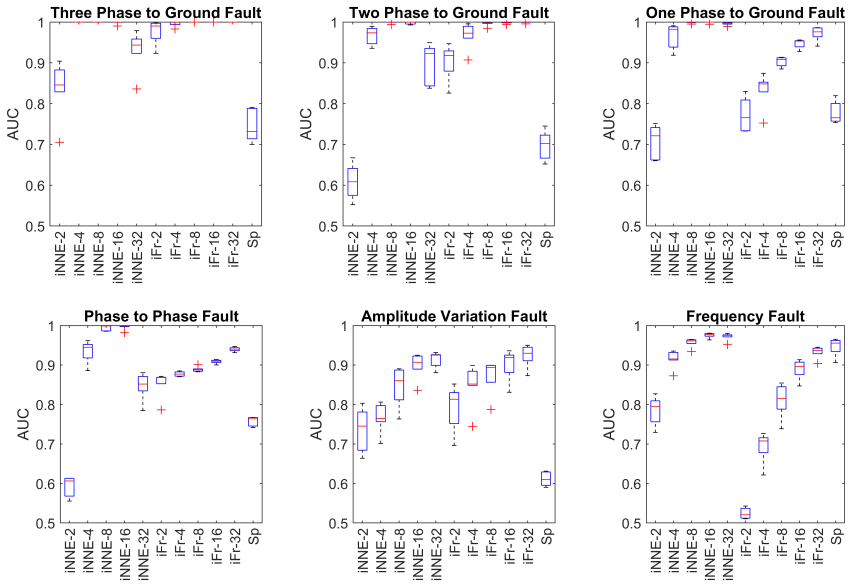


Fig. 4: A box-plot illustration of the results of the AUC computation using the S_p, iForest, and iNNE algorithms. The findings provide an averaged performance evaluation across a range of cluster numbers (50,..,300) (for iForest and iNNE) and sample sizes (10,..,100) (for S_p) for each of the 6 fault categories.

to the findings, these models offer a significantly low standard deviation across repetitions, and as a result, they have the potential to be regarded as reliable solutions. On the other hand, when compared to the other algorithms, the S_p algorithm provided the best computation time while receiving relatively lower AUC scores (excluding frequency fault datasets). We think that this is because the S_p is utilising a global anomaly detection. It is interesting to note that as the value of ψ increases, the estimates of AUC produced by both the iForest and iNNE models become increasingly accurate. It should also be mentioned that, in low ψ values, the iNNE algorithm runs noticeably more quickly than the iForest algorithm. However, this needs to be looked into more with experiments that use high ψ values and data sets with a lot of dimensions.

VI. Conclusion

This study proposes a benchmark study of unsupervised anomaly detection algorithms, such as Sp, iForest, and iNNE, using synthetically generated symmetric and asymmetric power system fault data. The dataset was generated in speedgoat using Kundur's two area systems real time model. The results show that iForest and iNNE provide competitive results in detecting anomalies of all fault types, with iNNE providing significantly better execution time performance. It is discovered that S_p provides an extremely efficient execution time but is unable to provide competitive AUC results. Future work includes faults added at multiple locations in the power system and the design of algorithms that can detect the anomaly as well as the position of fault.

References

[1] S. R. Sinsel, R. L. Riemke, and V. H. Hoffmann, "Challenges and solution technologies for the integration of variable renewable energy sources—a review," *renewable energy*, vol. 145, pp. 2271–2285, 2020.

[2] Z. Ali, K. Saleem, R. Brown, N. Christofides, and S. Dudley, "Performance analysis and benchmarking of pll-driven phasor measurement units for renewable energy systems," *Energies*, vol. 15, no. 5, p. 1867, 2022.

[3] H. Fanaee-T and J. Gama, "Event labeling combining ensemble detectors and background knowledge," *Progress in Artificial Intelligence*, vol. 2, no. 2, pp. 113–127, 2014.

[4] T. R. Bandaragoda, K. M. Ting, D. Albrecht, F. T. Liu, and J. R. Wells, "Efficient anomaly detection by isolation using nearest neighbour ensemble," in *2014 IEEE International conference on data mining workshop*. IEEE, 2014, pp. 698–705.

[5] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 93–104.

[6] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger, "The r*-tree: An efficient and robust access method for points and rectangles," in *Proceedings of the 1990 ACM SIGMOD international conference on Management of data*, 1990, pp. 322–331.

[7] G. O. Campos, A. Zimek, J. Sander, R. J. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle, "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study," *Data mining and knowledge discovery*, vol. 30, no. 4, pp. 891–927, 2016.

[8] T. R. Bandaragoda, K. M. Ting, D. Albrecht, F. T. Liu, Y. Zhu, and J. R. Wells, "Isolation-based anomaly detection using nearest-neighbor ensembles," *Computational Intelligence*, vol. 34, no. 4, pp. 968–998, 2018.

[9] P. S. Kundur and O. P. Malik, *Power system stability and control*. McGraw-Hill Education, 2022.

[10] A. Khoshkbar Sadigh and K. Smedley, "Fast and precise voltage sag detection method for dynamic voltage restorer (dvr) application," *Electric Power Systems Research*, vol. 130, pp. 192–207, 2016.

[11] M. Sugiyama and K. Borgwardt, "Rapid distance-based outlier detection via sampling," *Advances in neural information processing systems*, vol. 26, 2013.

[12] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, no. 1, pp. 1–39, 2012.

[13] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.