Solar farm voltage anomaly detection using high-resolution μPMU data-driven unsupervised machine learning [☆]

Maitreyee Dey*[a,b], Soumya Prakash Rana[a], Clarke V. Simmons[b], Sandra Dudley[a]

[a]*School of Engineering, London South Bank University, London, United Kingdom*
[b]*Neuville Grid Data Limited, London, United Kingdom*

Abstract

The usual means of solar farm condition monitoring are limited by the typically poor quality and low-resolution data collected. A micro-synchrophasor measurement unit has been adapted and integrated with a power quality monitor to provide the high-resolution, high-precision, synchronized time-series data required by analysts to significantly improve solar farm performance and to better understand their impact on distribution grid behaviour. Improved renewable energy generation at large solar photovoltaic facilities can be realized by processing the enormous amounts of high-quality data using machine learning methods for automatic fault detection, situational awareness, event forecasting, operational tuning, and planning condition-based maintenance. The limited availability of existent data knowledge in this sector and legacy performance issues steered our exploration of machine learning based approaches to the unsupervised direction. A novel application of the Clustering Large Applications (CLARA) algorithm was employed to categorise events from the large datasets collected. CLARA has been adapted to recognize solar site specific behaviour patterns, abnormal voltage dip and spike events using the multiple data streams collected at two utility-scale solar power generation sites in England. Fourteen days of empirical field data (seven consecutive summer days plus seven consecutive winter days) enabled this analytical research and development approach. Altogether, ∼725 million voltage measurement data points were investigated, and automatic voltage anomaly detection demonstrated.

Keywords: Solar energy, condition monitoring, micro-synchrophasor measurement unit, electrical anomaly detection, unsupervised machine learning.

1. Introduction

Shifting to renewable energy sources, such as solar power with its smaller carbon footprint is essential to mitigate climate change. However, irregularities and abruptions in solar power generation bring challenges for power distribution system operators [1] and the energy market. In the UK, utility-scale (> 1 MWp) solar farm owners commonly suffer from burdensome plant failure rates, reduced equipment lifespan, unplanned outages, diminished energy output, and replacement overheads [2]. These problems can be countered through better condition monitoring and knowledge discovery that can automatically perceive abnormalities and trends that may indicate arising issues and even predict faults before they occur.

Today's industry standard Supervisory Control And Data Acquisition (SCADA) systems do not have the inbuilt capability to detect anomalous behaviour, predict faults, and diagnose failure modes

[☆] The short version of the paper was presented at virtual CUE2020, Oct 10-17, 2020. This paper is a substantial extension of the short version of the conference paper.

which lead to expensive complications [3]. Improved monitoring tools are beginning to provide more granular, higher accuracy time-stamped data [4]. Via the rapid development of digital technology and cloud computing, vast amounts of "big data" are produced through the plethora of installed sensors and advanced digitalisation infrastructures. Cost-effective exploration and rationalisation of the resulting plethora of structured big data can benefit facility operation and understanding of grid interactions [5].

To help observe and manage the distribution grid, the micro-synchrophasor measurement unit (μPMU or microPMU) was developed to provide ultra-high precision Root Mean Square (RMS) voltage and current phasor values at twice-per-cycle granularity (100/120Hz depending on grid frequency) with sub-100 nanosecond synchronised time stamps [6].

Transforming μPMU measurements into actionable information in real-life scenarios however remains a huge challenge [7]. Applying μPMU technology can improve operational monitoring of distribution networks and their performance via applications such as: stability assessment, state estimation, outage or disturbance detection, energy management, and so on [8]. These μPMUs could become an important part of smart-grid technology and rapid state variable solution, due to their lower cost, high accuracy, and short latency [9]. Establishing timely data transmission and effective large-scale data storage are implementation challenges [10]. The μPMU data volume grows massive with fast data reporting rates, thus pre-processing these big data to extract valuable information towards solar site maintenance and operation is challenging [11].

1.1. Background of Power Grid System μPMU Applications

Machine learning based data analytics can play a role in making large scale systems such as cities and their power grids more effective and user centred. Much research has been published on the challenges of optimising energy performance of low carbon cities using artificial intelligence [12, 13]. In future power systems, μPMUs and similar devices could provide enabling data for advanced distribution system operation [5]. For example, in a research study by Yigit, et.al. [2], linear state estimation has been performed to improve parallel computation for big power grid data screening, gathering, and processing, with the aim to provide intelligent grid monitoring via abnormal event detection. Several studies have been performed to investigate big μPMU data to identify anomalous events in the power distribution network [14]. In [4], a μPMU data-driven based method to classify power quality events to improve power distribution network performance has been presented. But this work involved rigorous data labelling by field experts and utility records, a luxury not always available if the execution of renewable infrastructure is to grow as it must to attain the worldwide targets published. The authors extended their previous work for determining the likely source of the power quality events on distribution systems through voltage and current phasor measurements [15]; however, results are limited to the examination of four event scenarios. Distributed generation can create problematic low-frequency oscillations (LFO). To deal with LFO, a two-stage method with Tunable Q-factor wavelet transformation and matrix pencil algorithm using the μPMU data has been proposed in [10].

Irregular fluctuations in power distribution by short-time local outlier probability and peak analysis have been analysed in [16]. The absolute deviation around the median combined with dynamic window size was employed in this paper for the purposes of event detection. A kernel Principal Component Analysis (kPCA) is obtained to build statistical models for anomaly detection in [17]. Here, expert knowledge has been used to train the event types (e.g. voltage sag) from the μPMU data using partially hidden structured support vector machine. In [18], a semi-supervised learning (SSL) approach has been applied to identify high impedance faulty events and their location, where the partial knowledge of the event is available a priori. Local outlier factor algorithm is used for the abnormal event detection purpose, where the PCA-based similarity search method was used to measure the differences of operation states between any two buses of the Western Electricity Coordinating Council (WECC) 179-bus power system, a case from the South China Power System (SCPS), and a case from the Guangdong Power System (GDPS) was proposed in [19]. A hidden structure semi-supervised machine learning method was proposed to detect events in power systems, where Minimum-redundancy-maximum-relevance (mRMR) is adopted to create a feature set (20 most informative features) and resolve the trade-off between relevancy and redundancy [20]. A three layered frequency events have been proposed in [21], where Granger causality was used to compare measurement relations across different locations, train sparse coding dictionary and the moving z-score mechanism was applied to detect the events from μPMU measurement installed by Lawrence Berkeley National Laboratory (LBNL) and Riverside Public Utility (RPU).

These studies reported employing data-driven conceptualisation to indicate inaccurate power event sensing is a significant task for researchers and site engineers. The μPMU provides high-resolution and substantial insight on the state of power system to which it is attached, is beneficial to prevent outages, reduce operational cost, and increase equipment lifespan. But, studies hitherto either utilise probabilistic approaches, data transformation into smaller domains to represent events, or labelling of a low number of data patterns to recognise abnormal events. Probability does not provide absolute decision, it expresses the chance of an event being normal or abnormal. Hence the event decisions are uncertain and cost-inefficient when it comes to big-data. The semi-labelling approaches are limited by knowledge and time complexity. The segregation of power data based on their quantifiable property has not yet been explored, which could be a fast and efficient automated site maintenance solution. These issues have been addressed through our study, where large-scale data has been collected by an operatively-paired μPMU plus power quality monitor (PQM) apparatus from existing UK solar sites. The resulting μPMU data has been analysed by employing fully unsupervised machine learning (UML) or clustering approach. This UML approach was performed by using the quantitative three-phase voltage magnitude. Voltage is often a key parameter when investigating or assessing the risk of operational problems or equipment failures on solar sites.

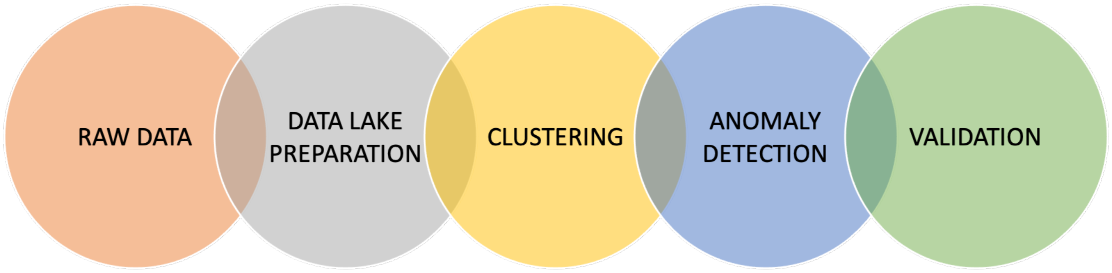

Figure 1: Process flow diagram of the contributed work.

The proposed framework comprised five stages: (a) grid data units were designed and installed on the solar farms to gather the time-series voltage measurement data, (b) masses of μPMU data were collected, pre-processed, and stored into a dedicated 'Data-Lake', (c) time-series feature vector approximation has been performed to find optimal feature vector length, (d) the Clustering LARge Applications (CLARA) algorithm was employed for the first time on this large-scale μPMU data to categorise and analyse the voltage dip events, (e) the proposed framework was tested on two real UK based solar farm sites over the period of two week's data for the summer and winter seasons, (f) the obtained results were validated through the standard power quality event monitoring data. The contributions are summarised by the flow diagram of Fig. 1.

2. Material and Methods

Engineers at Neuville Grid Data built a bespoke apparatus [22], the Grid Data Unit (GDU), to collect the high-resolution and accurate data from solar sites. The GDUs were installed at solar farm substations in Norfolk and Bedfordshire, England, UK. Both of the solar sites are connected to the UK Power Network. The collected data from these sites were investigated in this study and the overall process flow is shown in Fig. 2.

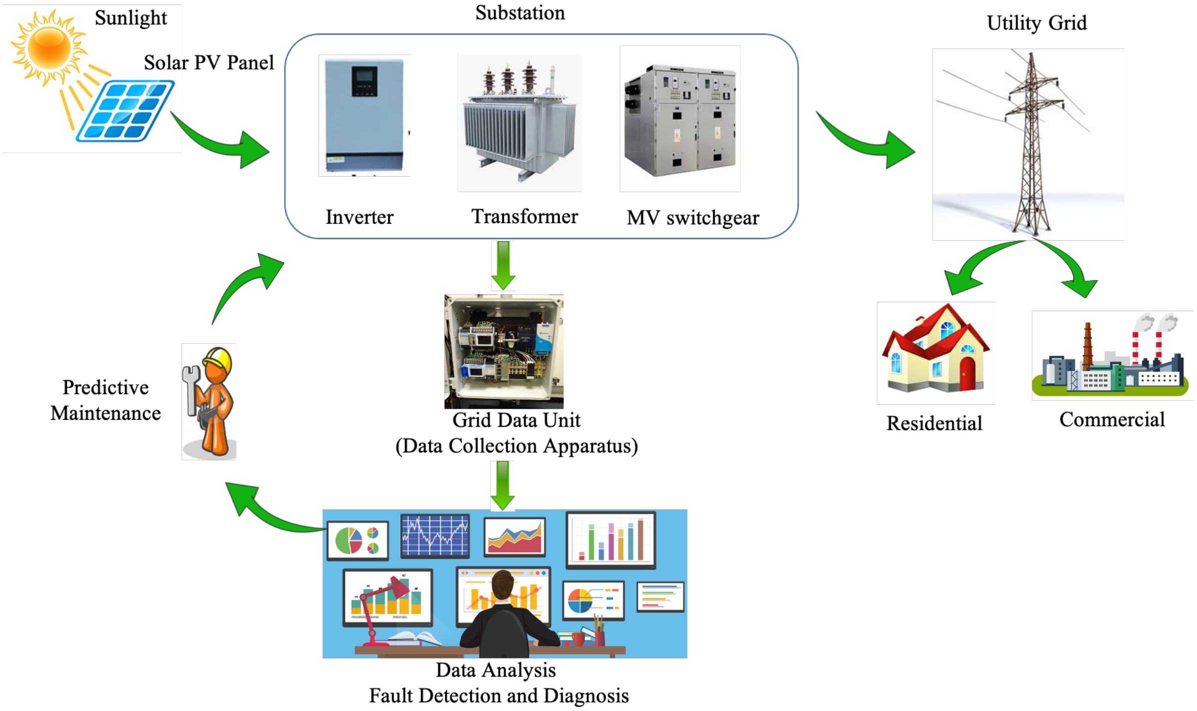

Figure 2: Proposed μ-PMU data-driven based solar farm anomaly event detection procedure.

2.1. Designed Apparatus

The GDU (shown in Fig. 3) incorporates a μPMU instrument, a PQM, simultaneous global positioning system (GPS) antenna for time-synchronization to sub-100ns, solid-state memory for data-buffering, and secure bidirectional 3G/4G cellular data telemetry equipment. This GDU provides mechanical protection, instrument power, and telecoms equipment for data backhaul.

The μPMU's underlying analogue-to-digital conversion (ADC) at 4MHz digital signal processing (DSP) results in RMS voltage and current phasor values for each phase at every half-cycle (i.e. on a 3-phase 50Hz installation collects 6 channels at a 100 Hz data-reporting rate). This can be subsequently down sampled to suit the post-processing techniques employed. The phasor amplitude and angle accuracies are ±0.05% and $\pm 0.01^o$, giving a total vector error (TVE) of ±0.01%. Fig. 4 shows the data collection apparatus, from signal acquisition through to data storage. The GDU apparatus is typically installed in the customer substation or inverter station. Calibrated split-core

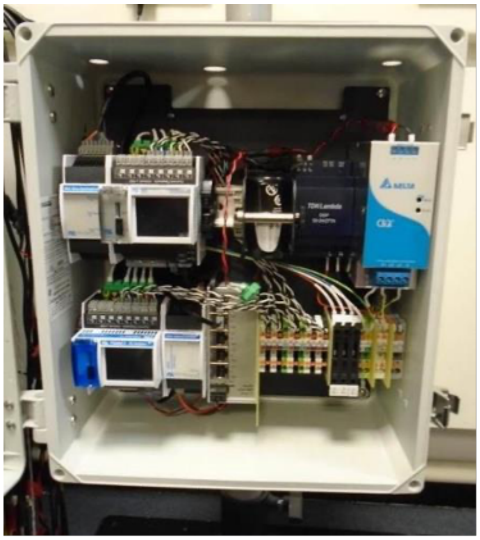

Figure 3: Grid Data Unit apparatus prototype.

current transformers monitor the 1A or 5A secondary outputs of the revenue-grade 11/33kV current transformers (CT), while voltage transformers (VT) on all three phases of the export cables provide proportional signals to the operatively paired μPMU and PQM.

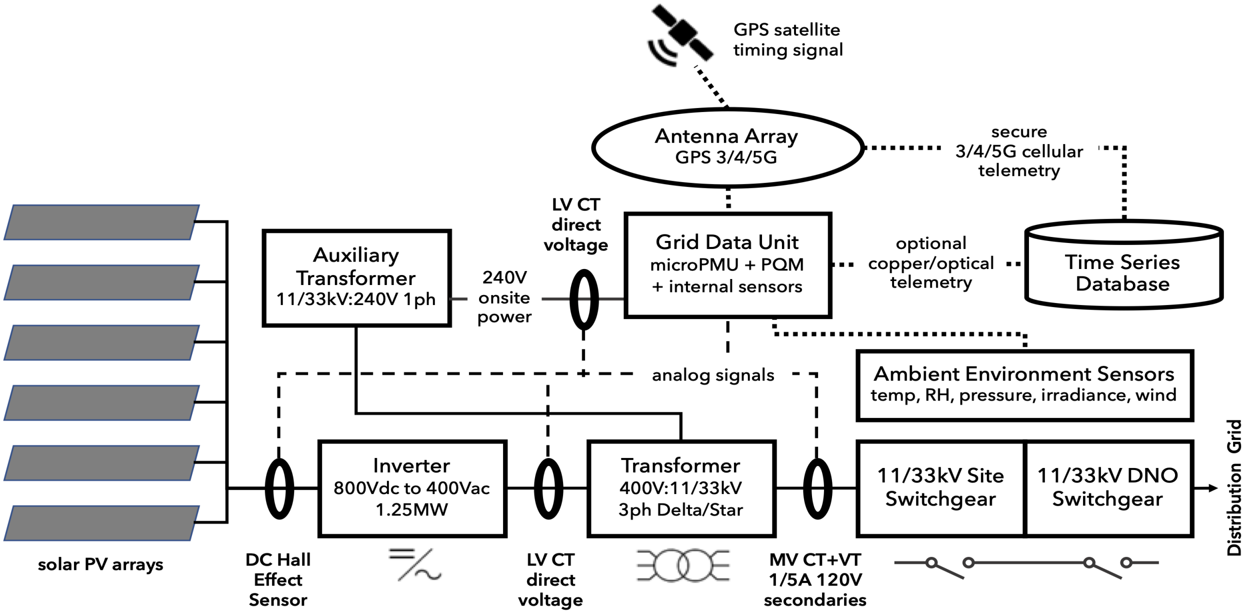

Figure 4: Utility-scale (> 1 MW) solar farm GDU installation.

2.2. Micro-synchrophasor Measurement Unit Data Characteristics

The μPMU operates in the frequency domain and collects measurements for each half-cycle at 10 milliseconds data reporting rates (i.e., 100Hz in Britain). Fig. 5 shows the actual power measurement generated by the solar farm on a single day during the summer of 2020. Fig. 5(a) depicts how the photovoltaic (PV) generated a three-phase current output, which varies with solar irradiance, subject to intermittent cloud cover. Fig. 5(b) depicts the corresponding three-phase (line-to-neutral) voltage measurements, where large voltage drops were found on several occasions.

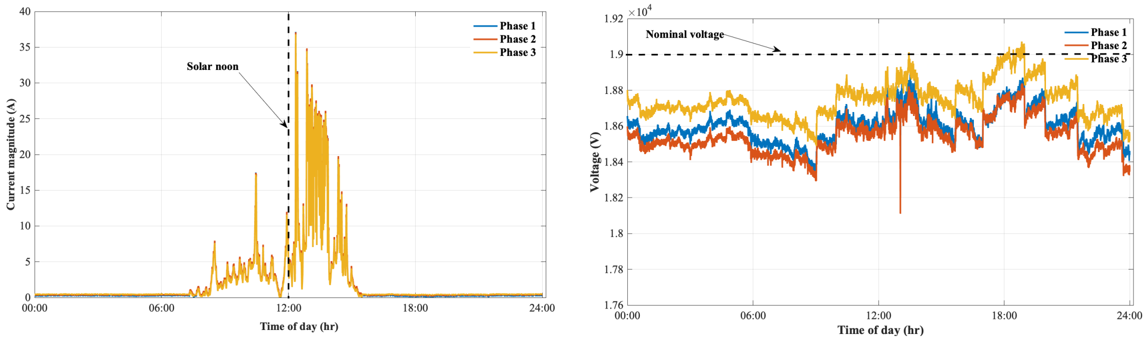

Figure 5: Collected μ-PMU solar farm data for a single day , (a) current measurements, (b) phase voltage measurements.

2.3. Proposed Approach

The anomaly detection method followed two consecutive and correlated ML steps: feature length approximation and clustering, using solar site μPMU big data. The data cleaning and preparation are key tasks on which the performance of tackling outliers depends. Thus, raw data was formatted from binary to columnar format and prepared to fit for the ML task. Different feature lengths of the data attribute were approximated with clustering to obtain optimally performing feature lengths for anomalous event detection and identification.

2.3.1. Feature Length Approximation

The three-phase, time-series voltage magnitude measurements are denoted by Eq. (1) and (2), where the voltage of each phase is considered as a $1-D$ feature vector. We used an approach that can transform the time series voltage feature vector in higher Euclidean space ($\mathbb{E}^s$), where s denotes the number of feature space and time domain voltage data (V_ϕ) of a time frame is considered as feature space to represent the voltage behaviour. The voltage feature engineering approach contemplates the time window (δ) required to collect n number of data points and the data reporting interval ($t = 10ms$) required to collect N number of bifurcate voltage magnitudes, aligned uniformly in ($N = \delta/t$) no. of rows (r).

$$\begin{gathered} V_\phi = \Big\{\alpha_x \in \mathbb{R} | \alpha_x, x = 1, 2, ..., n\Big\}, \\ \phi = 1, 2, 3, n = total\ no\ of\ datapoints\ in\ each\ phase. \end{gathered} \tag{1}$$

$$V_\phi^T = \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_N \\ \alpha_{1+N} & \alpha_{2+N} & \dots & \alpha_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_{(r-1)N+1} & \alpha_{(r-1)N+2} & \dots & \alpha_{rN} \end{bmatrix} \tag{2}$$

2.3.2. Clustering Large Applications

Once the optimal feature vector is obtained, the unsupervised learning or clustering method is subsequently applied to recognise distinct voltage behaviour of the data patterns without prior information. Clustering partitions the r number of patterns into k number of clusters with the aim to minimise intra-cluster distance and maximise inter-cluster distance. The unsupervised segregation follows partitioning based clustering method Clustering LARge Applications (CLARA) to perform this big voltage data clustering [23]. CLARA inherits and extends the methodology of partitioning around medoids (PAM) to overcome the limitations of k-medoids and k-means algorithms that operate on medoids for clustering the big data, reducing computational time. They are not sensitive to outliers like k-means. CLARA is fast and effective for large-scale data clustering as it does not check every neighbour of a node, only checking a sample of the neighbours of a node. CLARA takes a small percentage of the dataset to find medoids instead of the whole dataset. The PAM algorithm is applied to create an optimal set of medoids for the small sample. The performance of resulting medoids is calculated by averaging dissimilarity between each object within the full data set D and the medoid of its own cluster, defined as the below cost function:

$$C_{(\mu,D)} = \frac{\sum_{i=1}^{s} \Delta(w_i, w_j)}{s} \tag{3}$$

$$\Delta_{i,j} = \sqrt[2]{(w_i, w_j)^2} \tag{4}$$

Where, μ is the set of selected medoids, s is the selected sample, dissimilarity between objects (w_i, w_j) is measured by $\Delta_{i,j}$ and returns a medoid from μ which is closest to the object. CLARA iterates the sampling and clustering process and derives the final clustering result from the calculated medoids with minimal cost. The quality of clustering highly depends on the sample size as CLARA adopts a sampling approach.

2.3.3. Thresholding and Validation

Once the data are partitioned, a thresholding based on the cluster's mean (μ) and standard deviation (σ) is then applied on the Euclidean distance to detect the outlier events within the clusters. The three-sigma rule has been considered over the clustered data to highlight anomalous events. If each data point's distance exceeds the three-sigma rule ($3\sigma \pm \mu$) from its own centroid, then the data point is considered an outlier. The mathematical expression for thresholding is set experimentally and the outcomes are analysed later in this article. Once the outliers are identified, unusual voltage magnitudes are validated through power quality measurement data considering the event reporting time from both systems. The outcome and their validation are analysed in detail in the following section.

3. Result Analysis

The proposed method has been tested and implemented using data gathered from two pilot solar sites in England. The 13.2MWp Bedfordshire site comprises 8 inverter-stations that consist of transformers, Power Electronics brand inverters, and switchgear. The 8.0MWp site in Norfolk has 5 inverter-stations consisting of Gamesa central inverters, transformers, and switchgear. Both sites are physically located at approximately the same latitude. The sites were commissioned around the same time to similar designs. Having with very similar equipment, except for their inverters, and are therefore comparable matched pair. Our research investigated normal and abnormal voltage behaviour from the two-pilot sites for automatic anomalous event detection. This method experimented on the μPMU data collected by the GDU over a particular period between May to November 2020. The time span additionally covers both the summer and winter, which helps to account for seasonal aspects of both the solar farms and subsequent effects on the power grid. Physical details of both sites are summarized in Table 1.

Table 1: Site locations and details for the two pilots solar farm.

Site Name	Location	Commissioned	DNO Grid Connection	Inverter Type	MV Power Transformers (oil cooled)	Power generation capacity	Solar Altitude	Data Study Period
Langford	Bedfordshire, England	March 2015	UKPN 33kV	Central (Power Electronics)	4 x 400V:33kV1800kVA 2 x 400V:33kV1400kVA 2x 400V:33kV1250kVA	13,184 kWpdc	6.6	May to November 2020
Kenninghall	Norfolk, England	March 2015	UKPN 33kV	Central (Gamesa)	5 x 400V:33kV 1400kVA	7,999 kWpdc	5.8	May to November 2020

3.1. Data and Event Description

This is a first-ever study presented where real and high-resolution solar data is clustered using CLARA approach. With 100 samples/second data reporting rate, μPMU gathers $\sim$8.6 millions of voltage samples daily for each phase along with precise timestamp information. Primarily, three-phase voltage phasor measurements were studied here to detect anomalies as it has significant effect on the energy management functionality of the power distribution system. Seven consecutive summer days (from 1^{st} July to 7^{th} July 2020) and seven consecutive winter days (from 1^{st} November to 7^{th} November 2020) have been included in this study and analysed thoroughly. During these periods, several voltage issues have been noticed in the site data, some of them reported by the PQM data. The power quality voltage events found during this time span are summarized in Table 2. These anomalous events are known as voltage dip/ voltage sag events.

Table 2: Details of the occurred events during the experimented days.

Location	Event Type	Event Magnitude	Event Duration (Seconds)	Trigger Date	Trigger Day	Trigger Time (UTC)	Trigger Channel	Trigger Threshold
Langford	Voltage Dip	90.32% of nominal	0.718s	2020/07/02	Thursday	15:32:50.370	L1-L2	94.0% of nominal
Langford	Voltage Dip	91.95% of nominal	0.080s	2020/11/01	Sunday	10:59:36.347	L3-L1	94.0% of nominal
Kenninghall	Voltage Dip	92.43% of nominal	0.140s	2020/07/06	Monday	12:46:43.579	L1-L2	94.0% of nominal
Kenninghall	Voltage Dip	91.73% of nominal	0.589s	2020/07/06	Monday	12:47:15.139	L1-L2	94.0% of nominal
Kenninghall	Voltage Dip	92.84% of nominal	0.359s	2020/07/07	Tuesday	03:55:09.649	L1-L2	94.0% of nominal

3.2. Data Distribution Analysis

Initially, the probability distributions of the daily voltage trends were analysed. Fig. 6 and Fig. 7 show the three-phase voltage probability distribution for both solar farms under test: Langford and Kenninghall, respectively. The distributions presented in the first column represent the summer day voltage and the second column represent the winter days voltage samples. The grid connection phase-to-phase nominal AC voltage is $33kV$ with a phase-to-neutral voltage of $33kV/\sqrt(3) = 19.052kV$. The x-axis defines the voltage magnitude measurements in kV AC in the distribution plot and y-axis represents the probability density of each magnitude, where the red dotted line shows the nominal voltage for each phase. The distribution curves have been observed as left skewed depicting the voltage magnitudes as always lower than the nominal voltage. Also, the mean (μ) and standard deviation (σ) were measured for both sites and it was found that the seasonal voltage mean ranges between $\sim 18.6kV(18600V/1.8\times10^4V)$ to $\sim 18.8kV(18800V/1.8\times10^4V)$. However, for both solar farms σ is different for summer days compared to winter days. The σ of voltage for the Langford (site-1) is $\sim 14V$ whereas it is $\sim 66V$ for Kenninghall (site-2). The variation in σ from the two different sites indicate that the site-2 voltage measurements are more dispersed or spread out in relation to their corresponding μ. The higher σ indicates the all three-phase voltage measurements of Kenninghall during the winter days are not clustered around the mean but display a greater spread.

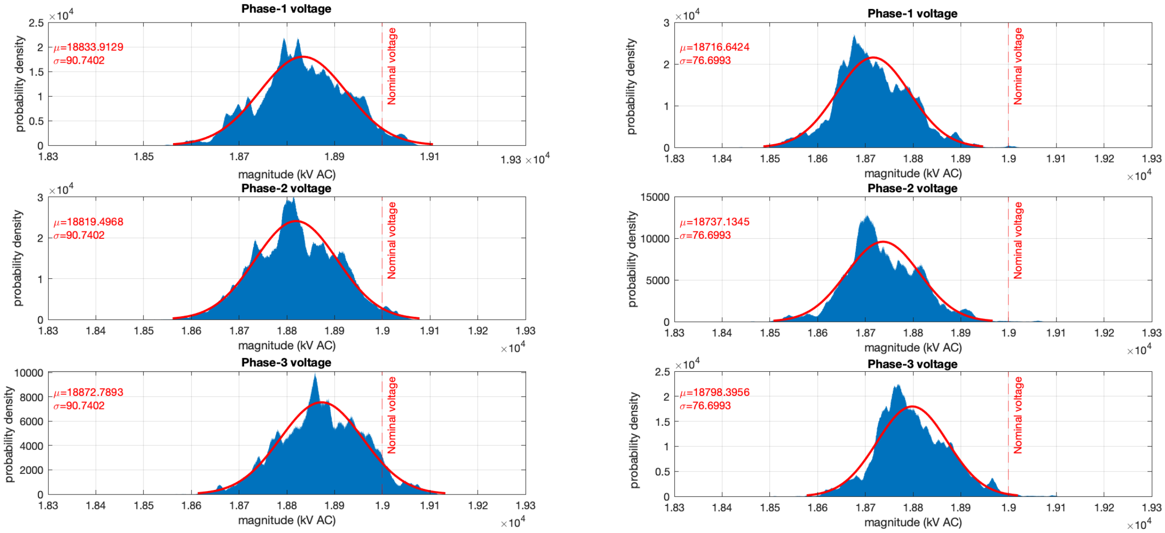

Figure 6: Site -1: Three phase voltage data distribution for a single day of a summer day (left) and winter day (right).

Conversely, the probability density of the voltage data is higher in Langford than at the Kenninghall site for both summer and winter days. The two sites clearly show contrasting probabilistic density, the voltage data of the Langford site (in Fig. 6) are unbiased with larger group of data population, whereas the data of Kenninghall site (in Fig. 7) contains smaller and sparser voltage groups. Analysis of voltage data distribution looks at the preparation specific to clustering for understanding the scale of magnitudes in different seasons and sites before execution. It provides the underlying groups of voltage magnitudes and intuition to make optimal default choice of initial parameters for initialising data transformation through clustering.

3.3. Feature Vector Approximation and Clustering with CLARA

Once the scale and distribution frequency of voltage magnitude had been pre-investigated, the voltage data was accumulated to make feature vectors representing different events through clustering. A feature vector was separately created for each voltage phase to study the variations, following the same data transformation mechanism. The voltage magnitudes were selected from pre-defined time frames and used as a feature vector for clustering (described in Section 2.3). The time frame was varied from 100ms to 10s with experimentally predefined intervals to obtain the best performing feature vector for clustering. Time windows of $100ms$, $1s$, $24s$, $3s$, $5s$, $8s$, and $10s$ were investigated for the optimal feature vector creation and to examine window size impact on the clustering outcomes. All the data from both sites was transformed in this way and the multi-featured data patterns are projected into the hyperspace to comprehend how each time window depended on the feature vector's potential for voltage anomaly detection. The feature vectors were employed to cluster the three phase voltages using CLARA. Each voltage phases and site were analysed separately using CLARA. The clustering outcomes obtained from these seven different feature lengths for the tested sites of Langford and Kenninghall are shown in Fig. 8 and Fig. 9 respectively. Fig. 8(a) to

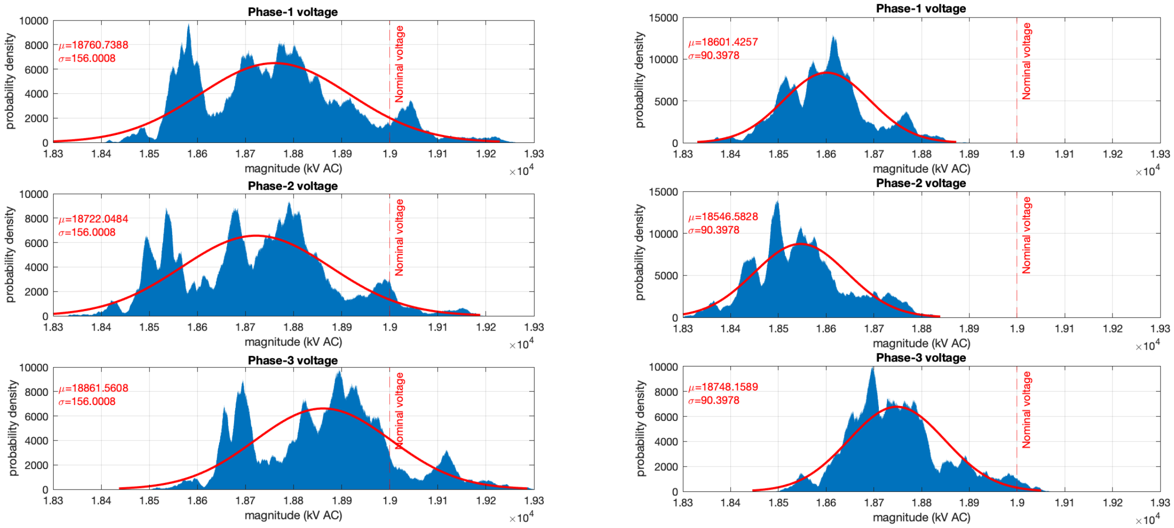

Figure 7: Site -2: Three phase voltage data distribution for a single day of a summer day (left) and winter day (right).

Fig. 8(g) demonstrate the clustering output using selected feature lengths from 100ms, 1s, 2s, 3s, 5s, 8s, and 10s respectively of the Langford site. The same numeric sequences have been used for clustering the Kenninghall data as shown in Fig. 9(a) to Fig. 9(g).

From this analysis, it was observed that the longer time windows represent lengthier feature vector/ feature dimensions, which decreasing the granularity of the data, losing the precise information needed for effective anomaly detection. Also, it has been found that the longer feature vector (i.e. magnitudes of longer time frame) produces more overlapping clusters, where the inter-cluster distances are shorter than the intra-cluster distances. Thus, segregating the normal voltages from any kind of abrupt changes that may have occurred becomes difficult. Conversely, the smaller feature dimension helps to capture distinct event behaviour and identify more granular information of an event that has occurred. The shorter feature length is preferable to detect finer changes therefore the "100ms" time window has been considered as the best performing feature vector for this study.

3.4. Anomaly Detection by CLARA

Once the feature vector's length has been approximated with 100ms time window, CLARA was performed on this optimal feature vector to detect voltage anomalies from the real data from both sites under test. Fig. 10 shows the clustering outcomes from a single day's μPMU data from the Langford solar farm (Site-1). The top row represents the μPMU data clusters from a summer day and the bottom row represents clustering outcomes of a winter's day. The three phases; phase-1, 2, and 3 are shown separately, and the red and yellow coloured data points indicate two different clusters i.e. two different voltage patterns. Outliers or anomalies were detected by measuring the mean and standard deviation of each cluster's Euclidean distance between the data points from their respective cluster centroids. The cluster mean and standard deviation were calculated, and it was experimentally found from repeated simulation that if any data point lies beyond the distance $d = (3\sigma \pm \mu)$ they should be considered as an outlier of that cluster (described in Section 2.3.3). Using this statistical approach, the data points that were found to be outliers are marked with blue colour for (cluster-1 patterns) and green colour for (cluster-2 patterns) in Fig. 10.

The same approach was performed on the second site's μPMU data and the resulting clusters are

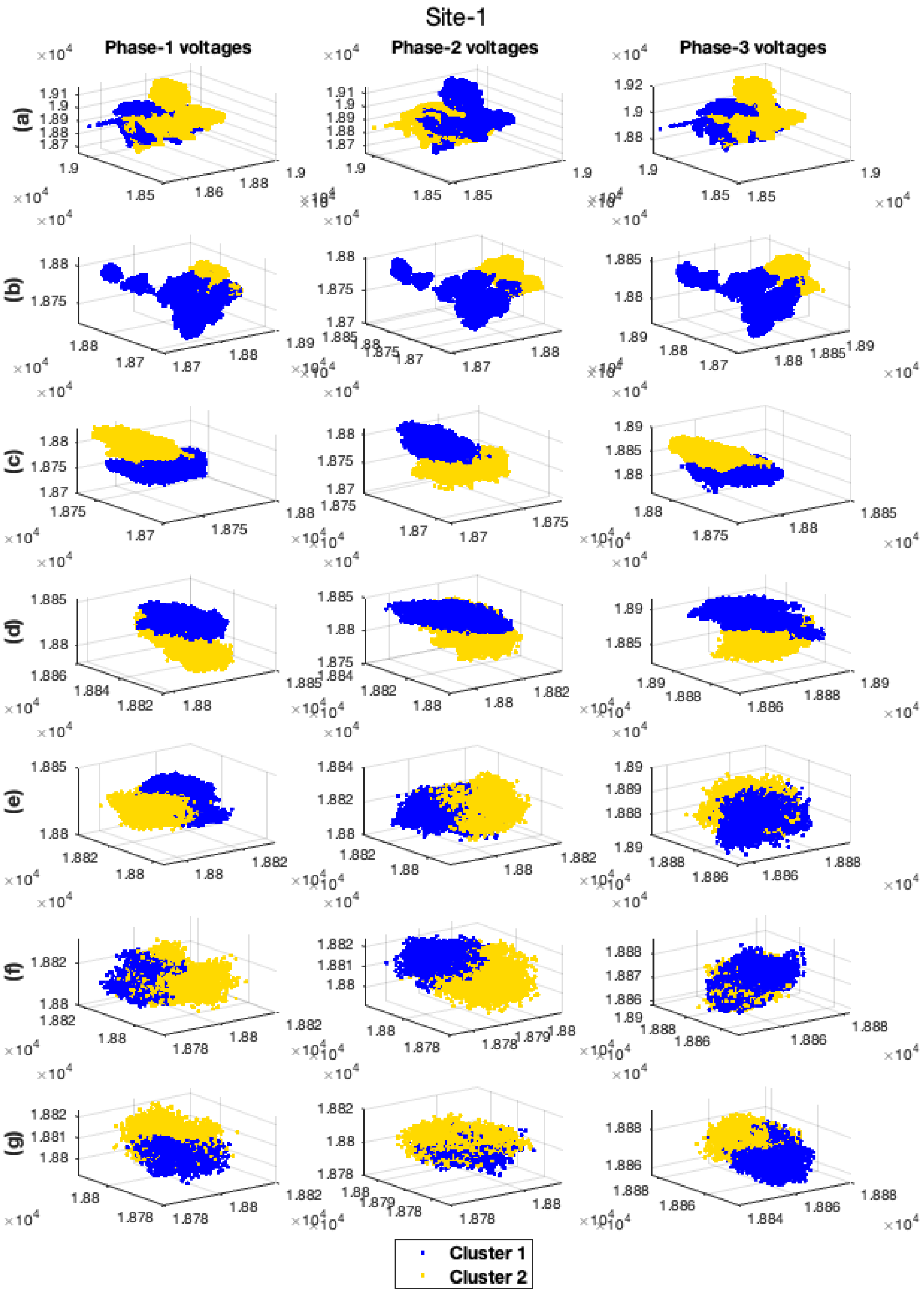

Figure 8: Site -1 feature length approximation.

depicted in Fig.11. The first row again shows the summer days' pattern, and the second row displays the winter days' pattern. The figures contain two clusters and their outliers. Outliers or anomalies, the unusual voltage behaviours have been detected through the stated approach and marked with blue (from cluster-1) and green (from cluster-2) colours as before. Though the data distribution of

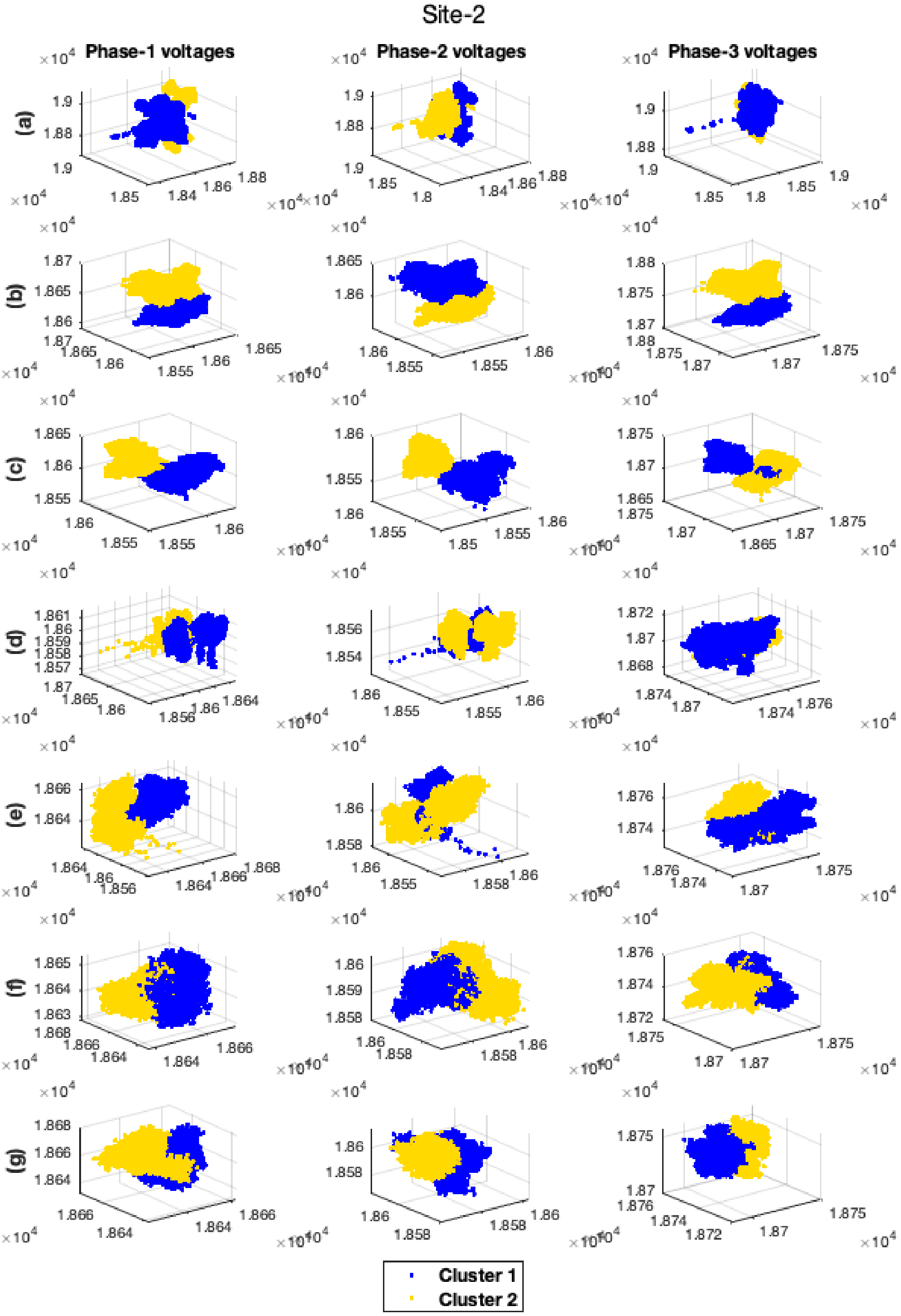

Figure 9: Site -2 feature length approximation.

both sites is different, the physical location parameters of both sites are similar enough to compare their energy generation performance. The big data clustering experiment was performed over all the μPMU data for the considered time span to capture any voltage fluctuations which had occurred due to anomalous PV effect. The segregated anomalies by CLARA were validated further using

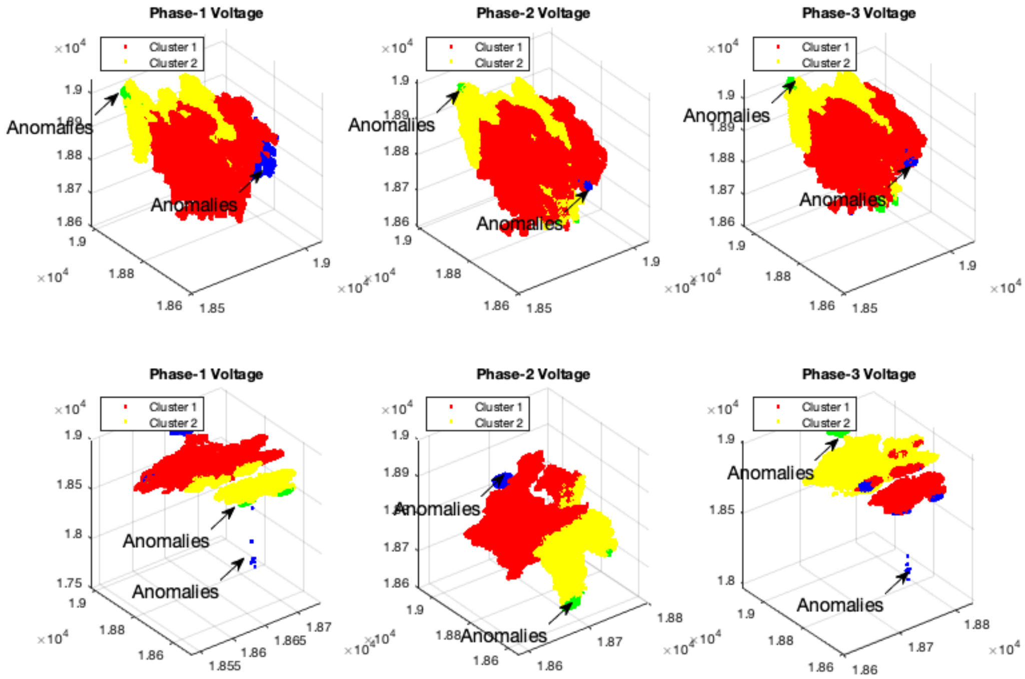

Figure 10: Site -1: three phase μPMU voltage data clustering for anomaly detection: (a) for a summer day, (b) for a winter day.

power quality measurements.

Performance of CLARA was compared with CLARANS, as both are useful for clustering big data [24]. However, CLARANS is known for its higher time complexity for time series data which reflected with collected time series voltage data too. Initial testing was performed on a single day's voltage measurement data (∼8.64 million of data points). CLARANS stuck to local minima when searching for local optimum and losses its effectiveness. Thus, CLARANS took a much longer time to converge, and it was challenging to evaluate the output quality from this method. CLARA takes a sample of voltage data points at the beginning of the search whereas CLARANS takes sample of voltage data point neighbours in each step of search. Therefore, CLARA converges quickly and efficiency in voltage data than CLARANS. Thus, CLARA was chosen for clustering the μPMU data.

3.5. Anomaly Validation via PQM

The detected outliers using CLARA along with the proposed feature approximation and thresholding were verified by expert engineers and the Power Standards Labs (PSL) standard Power Quality Monitoring (PQM) device. Precise time stamps and the reported event types from the PSL-PQM device were collated to compare with the identified anomalous events. According to the UK electricity supply the acceptable voltage to drift outside +10% to -6% is standard in public supply network. By following this, the PQM device reports an event at the point of time when any voltage measurement deviates outside of +10% to -6% from the nominal voltage. Although the duration and the characteristic of the event is not usefully captured due to the low-resolution data recording rate of the PSL device compared to the μPMU. During this research, all the events recorded by the PQM, and the anomalous events detected by this study were thoroughly inspected and cross-referenced by the authors and the engineers of the Neuville team. This proposed prototype can auto-detect anomalous behaviour far more precisely than the PQM device and therefore

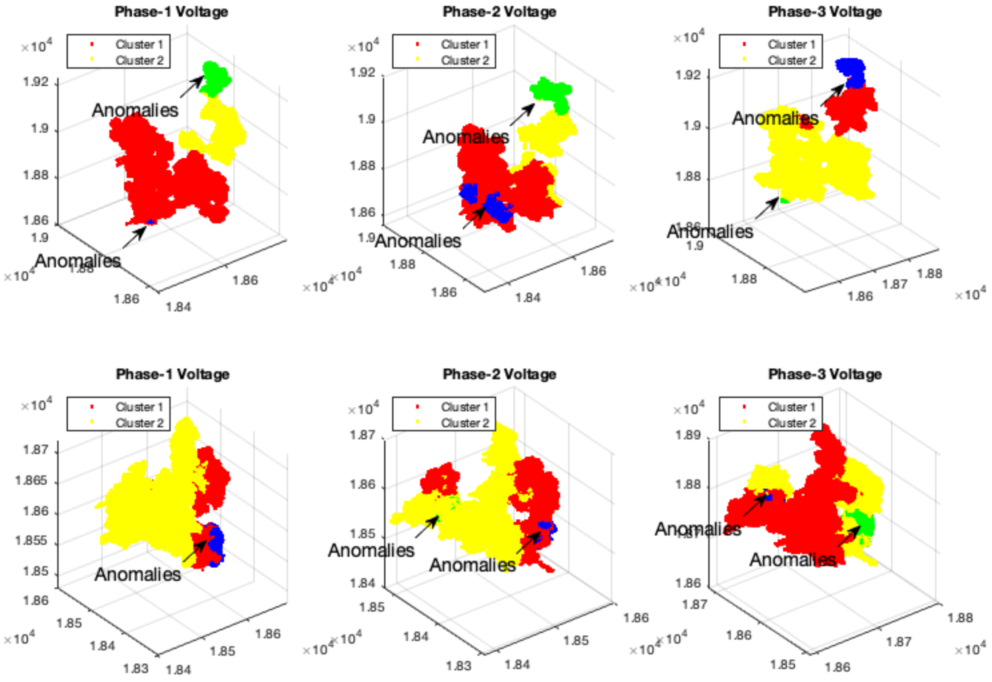

Figure 11: Site -2: three phase μ-PMU voltage data clustering for anomaly detection:(a) for a summer day, (b) for a winter day.

help in analysing their relationship with the distribution grid network. The work focusses on voltage performance monitoring and in particular "voltage dip/sag" event monitoring as this event occurred more often during the tested period. The algorithm was tested and validated on fourteen days consecutive historic data and a single day's results are displayed here from summer seasons of the two tested solar sites.

As expected, with its higher resolution μPMU reports a greater number of data points than the PQM and is able to capture important events. Those datapoints are clustered here for voltage anomaly identification. Also, μPMU can capture those events which may not drift outside the standard range, thus are not recorded by the PQM but do have an impact on the grid network. For example, there are two voltage dip events found from the Langford site's μPMU data on 2^{nd} July 2020 at 15:32 hr and 17.30 hr shown in Fig. 12 (top plot). The detected anomaly's time has been taken by the proposed method. It is to be noted that the first reported dip has been found from the PSL-PQM provided data (shown in Table 2), however the second voltage dip event at 17.30 hr of 2^{nd} July 2020, goes undetected by the PQM as it had not triggered the voltage dip event. Similarly, voltage dip events are also found in the Kenninghall site data on 6^{th} July 2020 at 12.47 hr shown in Fig. 12 (bottom plot). One event has been captured by the PQM, but the other event at 18.20 hr goes undetected. However, these events have been detected by the proposed method using the μPMU data. These voltage dips may have occurred due to either outages in a nearby generation unit, short-circuit, or overload that may have occurred on the grid. Thus, detecting a high proportion of unusual event is essential for comprehensive behaviour analysis and external grid effect on sites, demonstrating the usefulness of the method. The validation phase shows the effectiveness of the proposed method and its potential to capture more granular voltage trends beyond the power quality measurement mechanism. The method shows 100% accuracy for voltage dip event detection that are recorded by the PQM and an ability to detect those usual events that are sparse from the nominal voltage magnitude.

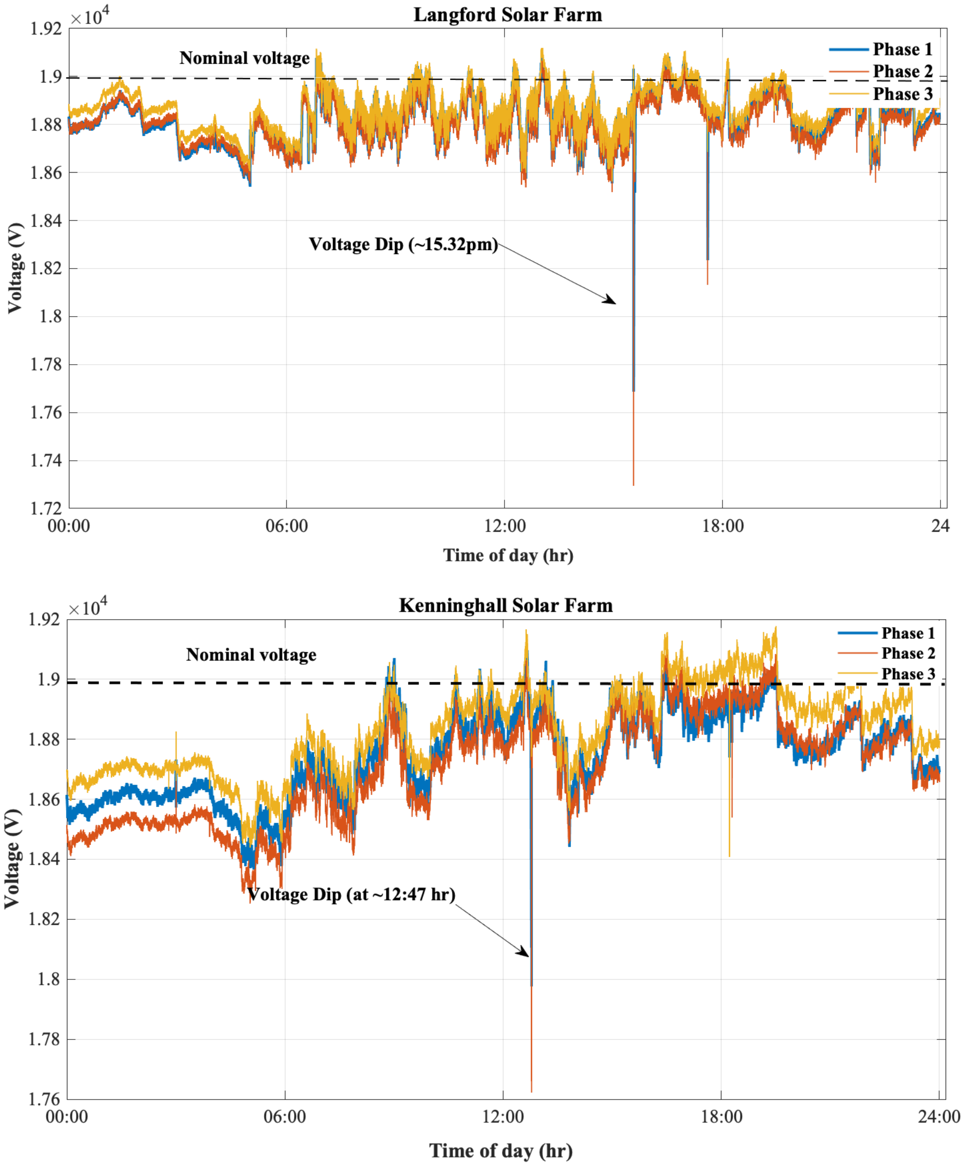

Figure 12: Voltage dips of the μPMU data on the tested day for Langford (top) and Kenninghall (bottom).

3.6. Discussion

Voltage dips and short voltage interruptions occur due to (a) abrupt changes in large loads/ high inrush and switching currents on the power network or on-site equipment (e.g. energising local transformers) and (b) various faults in the transmission and distribution networks [25]., which can adversely affect sensitive equipment and overall system operation. Thus, voltage dip can be an issue in solar farm operation and maintenance. Conversely, inspection of false issues, leaving anomalies undetected and undiagnosed can have a lasting and detrimental effects on a large solar farm's productivity. Hence, the proposed CLARA, followed by the statistically decided adaptive thresholding on high-resolution μPMU reported voltage data brings a cutting-edge solution to solar farm fault detection and diagnosis. Selected positive impacts from the work completed:

(a) A priori fault (anomaly) information from experts or technicians is not required to initiate the automatic voltage error detection, saving costly manual work, and time.

(b) Voltage spikes and dips can be detected quickly and automatically using unsupervised learning with near 100% certainty permitting a plug and pay future approach.

(c) Anomalies can be predicted to a useful extent; such cueing enabling avoidance of unexpected equipment failure through pre-emptive maintenance intervention leading to extended equipment service-life or planned equipment replacement.

(d) Operators can save costs and maximise revenues with tailored understanding of their site's unique behaviour patterns.

4. Conclusion and Future Work

The research is encouraged by the fact that there is over 8 GW of commonly underperforming utility-scale (> 1MWp) solar capacity across approximately 1,200 sites that were often hastily constructed in the UK between 2011 and 2017 to meet government subsidy programmes. The National Grid's Future Energy Scenarios 2020 report [26] projects 1.4GW of new UK solar installations every year through to 2050. Hence this study combined empirical big data analysis and machine learning to facilitate and improve solar farm operational and cost-efficiency through remote condition monitoring using CLARA for the first time on μPMU data for efficient cost, computational overhead management, and fast execution time.

The detected sparse voltage events will be further analysed. Both voltage anomaly detection and recognition of their cause will be studied for future work. Other anomalous events such as dip and flicker will be further studied to understand their relationship with weather conditions and electrically associated equipment. This will aid identifying the root of the outliers and motivate improvements to CLARA and the threshold decisions for a better performing solar data anomaly detection model. Anomaly recognition will be performed using historic data and the label information of events will come from the detection phase to predict future occurrences of voltage spike/dip enabling predictive solar site maintenance. Training ML models with voltage trends/profiles of specific faults (e.g. switchgear failure, transformer overheating, etc.) will be established for the betterment of diagnostic and control functionalities. The ongoing research predicts not only improved preventive and predictive site maintenance but also provides greatly improved data analytics for geographically dispersed assets and distribution grid purposes.

Acknowledgement

This work forms part of a Low Carbon London (LCLDN) project funded by the EU through its European Regional Development Fund (ERDF) with London South Bank University and Neuville Grid Data Ltd.

References

[1] R. C. Green, L. Wang, M. Alam, Applications and trends of high performance computing for electric power systems: Focusing on smart grid, IEEE Transactions on Smart Grid 4 (2) (2013) 922–931.

[2] M. Yigit, V. C. Gungor, S. Baktir, Cloud computing for smart grid applications, Computer Networks 70 (2014) 312–329.

[3] C. Liu, A. Akintayo, Z. Jiang, G. P. Henze, S. Sarkar, Multivariate exploration of non-intrusive load monitoring via spatiotemporal pattern network, Applied Energy 211 (2018) 1106–1122.

[4] A. Shahsavari, M. Farajollahi, E. M. Stewart, E. Cortez, H. Mohsenian-Rad, Situational awareness in distribution grid using micro-PMU data: A machine learning approach, IEEE Transactions on Smart Grid 10 (6) (2019) 6167–6177.

[5] Y. Zhang, T. Huang, E. F. Bompard, Big data analytics in smart grids: a review, Energy informatics 1 (1) (2018) 1–24.

[6] A. Von Meier, E. Stewart, A. McEachern, M. Andersen, L. Mehrmanesh, Precision micro-synchrophasors for distribution systems: A summary of applications, IEEE Transactions on Smart Grid 8 (6) (2017) 2926–2936.

[7] P. A. Pegoraro, K. Brady, P. Castello, C. Muscas, A. von Meier, Compensation of systematic measurement errors in a PMU-based monitoring system for electric distribution grids, IEEE Transactions on Instrumentation and Measurement 68 (10) (2019) 3871–3882.

[8] Z. Zhao, H. Yu, P. Li, X. Kong, J. Wu, C. Wang, Optimal placement of PMUs and communication links for distributed state estimation in distribution networks, Applied Energy 256 (2019) 113963.

[9] K. Jia, C. Gu, L. Li, Z. Xuan, T. Bi, D. Thomas, Sparse voltage amplitude measurement based fault location in large-scale photovoltaic power plants, Applied energy 211 (2018) 568–581.

[10] T. Jin, S. Liu, R. C. Flesch, W. Su, A method for the identification of low frequency oscillation modes in power systems subjected to noise, Applied Energy 206 (2017) 1379–1392.

[11] Z. Lv, H. Song, P. Basanta-Val, A. Steed, M. Jo, Next-generation big data analytics: State of the art, challenges, and future research topics, IEEE Transactions on Industrial Informatics 13 (4) (2017) 1891–1899.

[12] M. Dey, S. P. Rana, S. Dudley, Smart building creation in large scale hvac environments through automated fault detection and diagnosis, Future Generation Computer Systems 108 (2020) 950–966.

[13] M. Dey, S. P. Rana, S. Dudley, A case study based approach for remote fault detection using multi-level machine learning in a smart building, Smart Cities 3 (2) (2020) 401–419.

[14] N. Yu, S. Shah, R. Johnson, R. Sherick, M. Hong, K. Loparo, Big data analytics in power distribution systems, in: 2015 IEEE power & energy society innovative smart grid technologies conference (ISGT), IEEE, 2015, pp. 1–5.

[15] M. Farajollahi, A. Shahsavari, E. M. Stewart, H. Mohsenian-Rad, Locating the source of events in power distribution systems using micro-PMU data, IEEE Transactions on Power Systems 33 (6) (2018) 6343–6354.

[16] Y. Seyedi, H. Karimi, S. Grijalva, Irregularity detection in output power of distributed energy resources using PMU data analytics in smart grids, IEEE Transactions on Industrial Informatics 15 (4) (2018) 2222–2232.

[17] Y. Zhou, R. Arghandeh, I. Konstantakopoulos, S. Abdullah, A. von Meier, C. J. Spanos, Abnormal event detection with high resolution micro-PMU data, in: 2016 Power Systems Computation Conference (PSCC), IEEE, 2016, pp. 1–7.

[18] Q. Cui, Y. Weng, Enhance high impedance fault detection and location accuracy via μ-PMUs, IEEE Transactions on Smart Grid 11 (1) (2019) 797–809.

[19] S. Liu, Y. Zhao, Z. Lin, Y. Liu, Y. Ding, L. Yang, S. Yi, Data-driven event detection of power systems based on unequal-interval reduction of PMU data and local outlier factor, IEEE Transactions on Smart Grid 11 (2) (2019) 1630–1643.

[20] Y. Zhou, R. Arghandeh, C. J. Spanos, Partial knowledge data-driven event detection for power distribution networks, IEEE Transactions on Smart Grid 9 (5) (2017) 5152–5162.

[21] N. Duan, E. M. Stewart, Frequency event categorization in power distribution systems using micro PMU measurements, IEEE Transactions on Smart Grid (2020).

[22] C. V. Simmons, (Neuville Grid Data), Methods and apparatus for the sensing, collecting, transmission, storage, and dissemination of high-resolution power grid electrical measurement data, patent GB2579156 (granted 13 January 2021).

[23] L. Kaufman, P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis, Vol. 344, John Wiley & Sons, 2009.

[24] R. T. Ng, J. Han, Clarans: A method for clustering objects for spatial data mining, IEEE transactions on knowledge and data engineering 14 (5) (2002) 1003–1016.

[25] L. Weldemariam, V. Cuk, J. Cobben, Impact of voltage dips monitored in the mv networks on aggregated customers, Electric Power Systems Research 149 (2017) 146–155.

[26] Future energy scenarios, Tech. rep., National Grid ESO (July 2020).
URL https://www.nationalgrideso.com/future-energy/future-energy-scenarios/fes-2020-documents