IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 May 2023, accepted 12 May 2023, date of publication 4 July 2023, date of current version 18 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3292059

RESEARCH ARTICLE

Realtime Detection of PMU Bad Data and Sequential Bad Data Classifications in Cyber-Physical Testbed

IMTIAJ KHAN, (Graduate Student Member, IEEE), AND VIRGILIO CENTENO, (Senior Member, IEEE)
Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060, USA
Corresponding author: Imtiaj Khan (imtiajkhan@vt.edu)

ABSTRACT Modern Smart Grids incorporate physical power grids and cyber systems, creating a cyber-physical system. phasor measurement units (PMUs) transmit time synchronized measurement data from physical grid to the cyber system. The System Operator (SO) in the cyber layer analyzes the data in both online and offline format and ensures the reliability and security of the grid by sending necessary command back to the PMUs. However, various physical events such as line to ground faults, frequency events, transformer events as well as cyberattacks can cause deviation in measurements received by the SO, which can be termed as ''bad data''. These bad data in turn can cause the SO to take a wrong restorative/mitigating strategy. Therefore accurate detection of bad data and identification of correct bad data type is necessary to ensure grid's safety and optimal performance. In this work we proposed a realtime sequential bad data detection and bad data classification strategy. At first, we have exploited the low rank property of Hankel-matrix to detect the occurrence of bad data in realtime. Secondly, we classify the bad data into two categories: physical events and cyberattacks. The algorithm utilizes the difference in low rank approximation error of multi-channel Hankel-matrix before and after random column permutations during physical events. If the cause of bad data is identified as cyberattack, our proposed algorithm proceeds to identify the cause of cyberattack. We have considered two possible cyberattack types: false data injection attack (FDIA) and GPS-spoofing attack (GSA). The proposed algorithm observes rank-1 approximation error of single-channel Hankel matrix containing unwrapped phase angle data to distinguish FDIA from GSA. Finally, the proposed algorithm is implemented in a realtime cyber-physical testbed containing PMU simulator and openECA. Results from the testbed using IEEE 13 node test feeder show that by choosing optimum parameters of Hankel-matrix, the bad data can be detected as well as the type of bad data can be correctly identified within less than 1 sec. of the occurrence of physical event or cyberattack. The bad data detection shows 100% accuracy for Hankel-matrix data-window greater than 140. Bad data can be classified as either cyberattack or physical event with perfect accuracy for data-window length greater than 73 for the threshold 0.1. A data-window length between 80 to 120 can distinguish GSA from FDIA, while GSA is implemented with varying phase angle shift of 0.1^o to 0.5^o. The realtime sequential model is also verified with IEEE 118 bus system simulated with SIEMENS PSS/E. Due to more complicated grid structure, IEEE 118 system requires more computational time to identify the bad data type, however that is still less than 2 sec, and can perform detection and classification with data-window length as small as 40.

INDEX TERMS PMU, GPS-spoofing, ad data, FDIA, Hankel.

I. INTRODUCTION

Microgrid refers to a cluster of interconnected electrical machines, local loads and distributed energy resources (DER)

The associate editor coordinating the review of this manuscript and approving it for publication was Giambattista Gruosso.

with the ability to function as a standalone system in island-mode as well as to work as grid-connected mode [1]. Penetration of single-phase and unbalanced loads can lead to the increase in power quality issues, that can affect the performance of microgrids (MGs) by causing abrupt changes in the power flow or by violating the operational limit [2]. Data centers in USA, which are considered as microgrid due to its ability to operate in islanding mode during power outage at the main grid, can suffer from voltage sags and harmonics [3]. Charging stations for electric vehicles (EV) can suffer from low order harmonics, causing total harmonic distortion (THD) greater than 1% [4], [5]. Three-phase unbalanced voltage, voltage fluctuation, harmonics etc. can hinder the operation of railway microgrid systems [6]. In addition, conductors breaking down and falling onto grounds/ physical objects can cause line to ground (LG) and phase to phase (PP) faults and subsequently increase the rate of rise of current and/or cause over-currents, over-voltage, under-voltage etc. These consequences can damage the grid performances and pose threats to human and wildlife safety [7]. Among the physical factors causing power quality issues in MG, LG faults are considered to be the most common type [8], [9], [10], [11].

Integration of microgrids with smart infrastructure including communication, monitoring and metering devices has pioneered the idea of smart grid (SG) that provides more reliable, resilient and robust operation [12]. Components of MGs are interconnected in a physical layer, and the smart communication systems and metering devices as well as monitoring equipments are interconnected within the cyber layer. The cyber layer is built on top of the physical layer, thereby making the whole SG a cyber-physical system (CPS) [13]. An important component of SG is phasor measurement units (PMU), which provides more reliable and relatively secured system-monitoring along with a faster reporting rate than that of conventional supervisory control and data acquisition (SCADA) system [14]. However, penetration of smart devices such as PMUs into the MGs increases the dependence on communication links between the different layers of CPS and requires secured data storage and analysis methods. Such dependencies on communication channel and data storage system raise the risk of cyberattack, particularly for critical infrastructures such as hospitals, military bases, data centers etc [15], [16].

Several researchers have assessed the vulnerabilities of PMU integrated MGs against cyber-attacks and proposed possible defense mechanism against such attacks. The most common vulnerabilities of PMUs in MGs is the third-party intrusion at the communication channel either between two portions of the cyber layer or between the physical and cyber layers. The third-party, also termed as the *attacker* can modify or take control over the data packets sent over the communication channel, thereby making it a man in the middle attack (MITM) [17], [18]. Of all types of MITM attacks, false data injection attack (FDIA) has been of particular interest among the researchers in recent year. FDIA can be best described as malicious data injection, or modifying the data packets by the intruders having the knowledge of system configuration [19], [20], [21]. The modified data can lead the system server taking unwanted actions and cause a misoperation in the grid. Recently, researchers proposed different generic attack models for PMU integrated smart grids, referred as GPS-spoofing Attack, that doesn't require the attacker to manipulate the highly secured communication channel or to have internal knowledge of the network parameters. PMUs use GPS 1 Pulse-per-Second (1 PPS) signal for time synchronization, and GPS-spoofing attack (GSA) target the GPS 1 PPS signal received by PMU [22], [23]. Attackers can spoof the actual GPS 1 PPS signal with a stronger electro-magnetic (EM) signal [24], consequently shifting the time-reference for the GPS-synchronized data in the PMU by manipulating the GPS 1 PPS signal [25]. Instead of modifying the measurements directly, GSA changes the timestamps of measurements and affects mainly Phase angle measurements due to the horizontal shift in timestamps. As a result, the measurements received by the system operator (SO) are taken with respect to a shifted time reference, making the SO perform faulty load-flow and stability analysis. These faulty analyses may force the SO to take unwanted action such as tripping a line serving critical loads and/or send wrong command back to the IEDs [26].

The proposed sequential scheme consists of two parts: detection and classification, which are executed sequentially in real time. In order to choose the most suitable algorithms for each part, the SO needs to address several challenges, such as accuracy, numerical complexity, realtime implementation, scalability etc.

For the detection part, there are several bad data detection schemes such as weighted least squares (WLS) [21], Kalman Filtering (KF) [27], software-defined networking model [28], active synchronous model [29], etc. Carefully designed FDIA can bypass a WLS based detection model [35]. KF is a more robust detection method and can be implemented in real time. However, since it relies on the measurements from the immediate previous state, a monotonous variation of time-series measurements during faults or cyberattacks may provide wrong detection flag to the SO. Whereas the models proposed in [29] and [30] demonstrate satisfactory performance in terms of anomaly detection, these models fail to distinguish cyberattack from physical events. The machine learning and big data approaches in [31], [32], [33], and [34] train the model with sensor measurements to compare the expected measurement with actual measurements. Nonetheless, these models suffer similar problem of failing to separate cyberattack from physical events. On top of that, it is difficult to implement these models in realtime as well.

The next challenge of bad data detection is realtime implementation. The Hankel-matrix based model utilized in this article uses time-series measurements over a moving data-window, therefore it is easily implementable in realtime.

Another challenge of algorithms is numerical efficiency. Implementing one algorithm for bad data detection in the first part and a different one for bad data classification in the second part is numerically inefficient. This is the case for the online detection model described in [35] and [36], both of them can be implemented in realtime. However, these models are numerically exhaustive and are not scalable to a different grid topology and operating set points. The models need to be re-trained for each change in topology/operating points.

The low rank approximation of Hankel-matrix is able to recover a large volume of missing data, and correct the bad data that exists in PMU measurements [37], [38], [39]. Even though the Hankel-matrix structure based models mentioned in [37], [38], and [39] can correct bad data accurately to provide reliable power system information, it is imperative to identify the cause of such bad data to increase the resilience of the system against future attack/ physical events. Identification of bad data types and locating the bad data source is critical for SO during HILF (high impact low frequency) incidents such as LG faults/ cyberattacks.

Therefore, with a goal of ensuring more resilient CPS, we have extended the bad data detection and correction models described in [37], [38], and [39] to a realtime sequential bad data classification algorithm. Hao et al. [37] utilizes low rank approximation of Hankel-matrix to distinguish measurement noises from physical events. We have exploited the similar concept to differentiate between physical events and cyberattacks.

The Hankel-matrix based model developed in [37] and [39] can be implemented in the detection (single channel Hankel-matrix) and classification (multiple channel Hankel-matrix) parts of the sequential implementation, ensuring numerical efficiency in realtime. Also, the Hankel-matrix based model analyzes each PMU channel individually to detect bad data, therefore it does not depend on grid topology. PMU channel from any additional node in the topology can be analyzed as a new Hankel-matrix with time-series measurements and provides similar satisfactory results as existing PMU channels.

The first step of classification, is differentiating cyberattacks from physical events. Only a very few works in existing literature focused on differentiating between physical events and cyberattacks. A no Table example of differentiating cyberattacks and physical events is the online machine learning (ML) based model [30]. Another ML based model that exclusively detects cyberattack and differentiates it from faults for differential relays is proposed in [40]. Both models have limitations of scalability and numerical complexity. ML based model requires training of large volume of measurements for a specific grid topology and operating conditions. The dataset is required to be trained for any change in grid topology and operating conditions, which is numerically exhaustive.

The proposed realtime classification model computes the low rank approximation error among the temporal measurements from neighboring PMUs. For each PMU node in the grid, the model uses multiple channel Hankel-matrix by taking measurements from the PMU nodes that are physically connected to it. As the model computes the low rank approximation error using measurements from only a few physically connected nodes, it is numerically less exhaustive. Additionally, the model does not depend on the grid topology or system operating condition since it analyzes the time-series measurements from only a cluster of physically connected nodes. For cyberattack, the temporal relation among physically connected nodes is different from the case with physical events, regardless of the grid topology. These attributes make the proposed model scalable to more complex topology and different operating conditions.

The second step of the proposed classification model is to identify the cyberattack type. For initial real-time detection model implemented in this paper, we consider only two cyberattack types: FDIA and GSA. Detection model of GSA can be formulated similarly as FDIA detection approach, with phase angle data used as measurement matrix [22], [41], [42]. Even though GSA can be detected in a similar way of FDIA using phase angle measurements only, these approaches fail to differentiate GSA from FDIA. Intruders attack different levels of the cyber layer for FDIA and GSA, such as the communication network for FDIA and the GPS signal for GSA. Therefore, differentiating FDIA from GSA is also important for the system operator to successfully restore the system after the attack. Distinguishing GSA from FDIA has been, however, an unexplored area until recently [43]. Performing low rank approximation of Hankel-matrix using unwrapped phase angle measurements can classify the cyberattack between FDIA or GSA.

The cyberattack classification model in [43] has the limitations of testing the algorithm with noiseless measurements, which don't reflect a real-world scenario. Another limitation of the model in [43] is that it requires multiple channel Hankel-matrix to compute the relative change in low rank approximation error of unwrapped phase angle measurements of affected PMUs. Nonetheless, the multiple channel Hankel-matrix fails to identify the affected PMU channel and the SO needs additional computational step to figure out which PMU is under GPS-spoofing attack. With an aim to improve the cyberattack classification performance in [43], we have proposed single channel Hankel-matrix based model using noisy PMU measurements. The proposed single-channel Hankel-channel of this article calculates low rank approximation error of each PMU channel individually, therefore the SO identifies affected PMU channel without further computational steps.

The Hankel-matrix based model used in this work exploits time-series voltage/ current phasor measurements in real-time and calculates the low rank approximation error for each PMU channel individually. For bad data classification, [30] and [40] uses measurements from sensors and relays to train the dataset, whereas the proposed realtime detection and classification model analyzes the spatial relation among

neighboring PMU channels with time-series phasor measurements of moving data-window. The PDC applies the multiple channel Hankel-matrix algorithm separately for voltage and current phasors, enabling SOs to identify and localize the measurements of the affected PMU channel in a complicated grid network. For classification of cyberattack, the Hankel-matrix uses unwrapped phase angle measurements, instead of voltage/current magnitudes or raw phase-angles, since only the unwrapped phase angle can demonstrate different behavior for GSA and FDIA, described in details in section IV.

The goal of this article is to propose a real time detection and subsequent classification of anomalies in the PMU measurements based on the bad data types such as: physical events, false data injection attack and GPS-spoofing attack. We have considered a sequential structure of the algorithm where the Bad Data is detected in the first step. The bad data is classified between physical event and cyberattack at the second step. If the cyberattack is identified, the algorithm uses low rank property of Hankel matrix on the unwrapped phase angle measurements to classify the attack between FDIA and GSA. The proposed sequential anomaly detection model is tested in a realtime testbed using a *PMU-simulator* and *openECA*. The major contributions of this paper are as follows:

- This article aims at accurately detecting and classifying bad data in realtime using measurements from existing PMU based infrastructure which provides fast (30 to 120 frame/second) data transmission;
- We have proposed a sequential realtime bad data detection and classification technique. The proposed model identifies the occurrence of bad data, as well as classifies the bad data among physical event, FDIA and GSA, in a realtime sequential manner;
- The proposed sequential classification model exploits the previous low rank approximation of Hankel-matrix based bad data detection and correction models by extending its application in differentiating bad data types, thereby providing the knowledge of bad data type to the SO to ensure proper system restoration and resiliency;
- The proposed technique utilizes Hankel-matrix based algorithm that is scalable to larger and more complex power grid, with combined detection and classification time being less than 1 sec., providing fast response opportunity for SO;
- With perfectly tuned data-window length, the proposed algorithm can achieve 100% accuracy in bad data detection;
- The proposed realtime model is tested and verified in a PMU-cybersecurity testbed;

The rest of the paper is organized as follows: section II discusses the bad data detection model using Hankel matrix. Section III describes the bad data classification with random column permutation of low rank properties of multi-channel Hankel-matrix. Section IV describes the classification of the cyberattack between FDIA and GSA, exploiting the unwrapped phase angle data. Section V briefly describes the testbed used to validate the proposed real-time approach. Section VI documents the results and analyzes the performance of proposed model in this paper. Section VII concludes the article.

II. BAD DATA DETECTION

As mentioned before, most common method to detect Bad Data injected by attacker is the state estimation model. The system operator estimates the state variable using AC and DC power flow equations, and flags bad data if the deviation between measurements and estimations exceeds threshold. The measurements are estimated with Weighted Least Squares (WLS) [21], [35] method. Bad Data is detection when the ℓ_2-norm of the residual between actual measurement and estimated measurement, also known as estimation residual, is larger than a predetermined threshold τ^h.

A well-designed FDIA can be stealthy enough to bypass WLS based BDD method [44]. However, more robust techniques such as Kalman Filtering (KF) [45] can be an effective tool to detect bad data caused by electrical events and FDIA. The KF estimator generates better estimations of state variables than conventional weighted least square based method [45]. At each time instance, KF estimates the state variable and measurements using measurements from previous timestamps. Furthermore, deviation based KF approach provides better result compared to the conventional KF estimator, particularly during FDIA since the estimation accuracy is impacted by malicious data injected to the measurement stream [46].

One advantage of the deviation based KF method (DKF) is that it incorporates the time-series variation of measurements. A similar advantage is available in the Hankel-matrix based detection method [37], which utilizes both temporal and spatial relation among the measurements from single or multiple PMU channels. The proposed bad data identification method in this work relies on both spatial and temporal relation of time-series PMU measurements from single and multiple channels. This approach has an edge over DKF method, as even a small deviation in measurements can be identified by comparing the spatial and temporal relations of measurements with neighboring PMU channels. Furthermore, applying Hankel-matrix based method in the first step of bad data detection reduces the time required to execute the proposed sequential algorithm of bad data classification in the second and third steps, providing SO with a faster restoration and mitigation opportunity.

A Hankel-matrix is created with the measurements for a specific time period T, starting from first timestamp to the timestamp t_s. Assuming there exists total W number of data points, the first row is creating by taking a small portion of W, such as $W - k + 1$, k is a positive integer. The second row is constructed by shifting the first row to one timestamp toward right. Consider a PMU measurement

matrix $Y = \begin{bmatrix} y_1 & y_2 & \dots & y_W \end{bmatrix}$ for a particular channel, and the measurements over time period T are $y_1, y_2, \dots, y_W$. The Hankel matrix *hank* for this dataset can be constructed as:

$$hank(Y) = \begin{bmatrix} y_1 & y_2 & \dots & y_{W-k+1} \\ y_2 & y_3 & \dots & y_{W-k+2} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ y_k & y_{k+1} & \dots & y_W \end{bmatrix} \tag{1}$$

The matrix $hank(Y)$ is a a $k \times W - k + 1$ matrix. W is referred as data-window length for the rest of the paper. The single channel Hankel matrix can be extended to a multi-channel analysis, with the modification in the measurement matrix Y. The new measurement matrix is a $M \times W$ matrix, considering M PMU channels, and can be expressed as:

$$Y_{mul} = \begin{bmatrix} y_{1,1} & y_{1,2} & \dots & y_{1,W} \\ y_{2,1} & y_{2,2} & \dots & y_{2,W} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ y_{M,1} & y_{M,2} & \dots & y_{M,W} \end{bmatrix} \tag{2}$$

The multi-channel Hankel-matrix $hank(Y_{mul})$ is as follows:

$$hank(Y_{mul}) = \begin{bmatrix} y_{1,1} & y_{1,2} & \dots & y_{1,W-k+1} \\ y_{2,1} & y_{2,2} & \dots & y_{2,W-k+1} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ y_{M,1} & y_{M,2} & \dots & y_{M,W-k+1} \\ y_{1,2} & y_{1,3} & \dots & y_{1,W-k+2} \\ y_{2,2} & y_{2,3} & \dots & y_{2,W-k+2} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ y_{M,2} & y_{M,3} & \dots & y_{M,W-k+2} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ y_{1,k} & y_{1,k+1} & \dots & y_{1,W} \\ y_{2,k} & y_{2,k+1} & \dots & y_{2,W} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ y_{M,k} & y_{M,k+1} & \dots & y_{M,W} \end{bmatrix} \tag{3}$$

The first step of the Hankel-matrix based bad data detection method is to exploit the low rank approximation (LRA) of the multi-channel Hankel-matrix $hank(Y_{mul})$, with M being total number of PMU channels. The key idea of obtaining LRA is taking the Singular Value Decomposition (SVD) of the $hank(Y)$ as $U\Sigma V^*$. The $hank(Y_{mul})$ can be approximated as rank-r ($r <$ rank-$hank(Y_{mul})$) by taking the largest r singular value such that the low rank approximation error, which is defined as eqn 4, remains less than a predefined threshold τ^r.

$$e^r(Y) = \frac{||U\Sigma^r V^* - Y_{mul}||_F}{||Y||_2} \times 100\%, \tag{4}$$

The low rank approximated equivalent of $hank(Y_{mul})$ can be defined as $hank(\hat{Y_{mul}}) = U\Sigma^r V^*$. Low rank approximated Hankel-matrix is also have same $k \times W - k + 1$ dimension as original Hankel-matrix $hank(Y_{mul})$

Each element $\hat{y_{i,j}}$, where $i = 1, 2, ..M, j = 1, 2, \dots, W$ comes from the low rank approximated $hank(\hat{Y_{mul}})$. At the second step, using the WLS method, the state variable d is estimated using the following relation [47]:

$$\hat{d} = (\hat{U^r}^T * \hat{U^r})^{-1} \hat{U^r} * hank(\hat{Y_{mul}}) \tag{5}$$

where $\hat{U^r}$ is the first r dominant left singular matrix from U. For each channel i, where $i = 1, 2, \dots, M$, the data at timestamp t_{s+1} is considered to be accep Table (not bad data) if the estimation residual $||y_{i,\hat{t_{s+1}}} - \Gamma_i||_2 \leq \tau^h$. The Γ is defined as $M \times 1$ matrix calculated from $\hat{U^r}\hat{d}$. Here τ^h refers to a predetermined threshold for bad data detection.

On the contrary, when the estimation residual exceeds the τ^h, it indicates the existence of estimation error at the timestamp t_{s+1}. However, a single occurrence such estimation error doesn't indicate physical event or cyberattack, since measurement noise or data transmission error might provide a discrete outlier in the measurement stream. To ensure the occurrence of bad data without any false positive case, we consider it as bad data only if there exists more than three consecutive estimation error over three consecutive moving time window with length T.

The main contribution of this section is the detection of bad data using low rank approximation of Hankel-matrix. The proposed method is scalable, since it only calculates estimation residuals separately for each individual PMU channel, therefore the size of the power grid does not impact the accuracy of detection. Moreover, the proposed model utilizes both spatial and temporal relationship among the PMU channels, thereby providing the SO with the ability to detect relatively smaller changes in measurements during a physical event or a cyberattack. This attribute gives the proposed model an edge over DKF.

III. CYBERATTACK VS PHYSICAL EVENT CLASSIFICATION

As the fault mitigation and system restoration techniques are different for cyberattack and physical events, differentiating these two types of faulty conditions is paramount for the system operator. The proposed Hankel matrix based model for BDD proposed in [37] can be extended to distinguish cyberattack and physical events. During physical event, there must exist a time-series correlation among the neighboring PMUs that are physically connected or topologically nearby. On the other hand, since the cyberattack is targeted to particular PMUs, there is no or little time-series correlation among the topologically neighbor PMUs. These differences can be exploited to identify the bad data type.

Generally under normal condition (without any physical event or cyberattack), random column permutation of Hankel-matrix would cause an increase in the low rank approximation error since the temporal relation among the elements of Hankel matrix is destroyed after column permutation. Since during physical event there exist temporal correlation among the data from neighboring PMU channels, a random column permutation will result in increased low

TABLE 1. Algorithm - identifying bad data type.

Initialization:	For a particular PMU channel I that is identified to contain bad data at timestamp t_s as in section II, O is the number of PMU channels that are physically connected to I. Receive data from all O PMU channels over data-window starting from timestamp t_{s+1} to t_{s+W+1}; W being a positive integer which is referred as data-window length. η is the threshold to identify bad data type; k is the Hankel-matrix parameter representing number of rows.
Step 1:	Create a $Ok \times W - k + 1$ Hankel-matrix H_A, similar to eqn 1, with voltage phasor measurements from O PMU channels over data-window length of W;
Step 2:	Calculate the low rank approximation error e^{rz} with varying rank r ($r \leq rank(H_A)$);
Step 3:	Do a random column permutation on the Hankel-matrix H_A and create a new matrix $\bar{H_A}$;
Step 4:	Calculate the low rank approximation error e^{rrz} with varying rank r ($r \leq rank(\bar{H_A})$);
Step 5:	If $e^{rrz} > \eta$ for r = 1, it is an physical event . Go to step 1 with the next data window , starting from timestamp t_{s+2} to t_{s+W+2} data sample;
Step 6:	If $e^{rrz} < \eta$ for r = 1, continue from step 2 with moving data-window until the data-window starting from timestamps $t_{s+W/3}$ to $t_{s+W/3+W}$ is covered. If $e^{rrz} > \eta e^{rz}$ is satisfied for any of the previous data-window, it is physical event;
Step 6:	Else, it is cyberattack.

rank approximation error. However, as for cyberattack, there is no or little temporal correlation among the data from neighboring PMU channels, therefore the low rank approximation error is already higher for cyberattack. A random column permutation will not change the low rank approximation error significantly. In short, observing the low rank approximation error before and after random column permutation will help the system server determine the cause of bad data in the measurements. Mathematically, the proposed cyberattack vs event identification model can be formulated as algorithm described in Table 1.

The proposed algorithm of differentiating physical event from cyberattack has a limitation against PMU calibration error, since despite being a physical event in nature, it lacks the proposed models' assumptions of the existence temporal relation among neighboring PMUs. However, there are realtime correction method of such events, that can provide PMU measurements with no or very little calibration error to the PDCs [48]. As a result of this pre-processing, the sudden change in measurements during cyberattack/ LG fault data will not be affected by such calibration error.

IV. CLASSIFICATION OF CYBERATTACKS

After the cause of bad data is identified as *cyberattack*, the SO tries to restore the breached communication network or use alternate communication medium. However, a GPS-spoofing attack doesn't require the attackers to breach the network communication between measurement devices to PDCs and/or to control center [25], [26]. This type of attack can be done by spoofing the GPS signal with a stronger electro-magnetic (EM) signal than that available from GPS satellites. The GPS receiver uses the spoofed GPS signal instead of the actual signal. As a result, GPS-spoofing attack needs to be mitigated in a different way than FDIA.

Previous works focused on the detection of GPS-spoofing attacks (GSA) treated this type of attack similar to FDIA, with a difference in measurement matrix z as in eqn 4. For FDIA, the measurement matrix can be either voltage or current magnitude or phase angles, however for GSA the measurement matrix is generally constructed with voltage/ current phase angle only as in [22], [41], and [42] since the impact of shift in time due to spoofing of GPS 1 PPS signal is reflected most at the phase angle measurements. Using phase angle measurements in algorithm in Table 1 will flag the bad data type as cyberattack for GSA. Similarly, if the FDIA targets phase angle measurements, executing algorithm in Table 1 with voltage/current phase angle measurements will also classify the bad data as cyberattack. It is difficult for the SO to determine whether the bad data is caused by spoofing GPS 1 PPS signal (GSA) or by corrupting the phase angle measurements directly (FDIA).

Generally, algorithm in Table 1 with the raw phase angle measurements cannot indicate whether it is GSA or FDIA, due to the similar behavior in raw phase angle data during GSA and FDIA. However, unwrapped angle measurements demonstrate a difference in behavior between GSA and FDIA. Angle unwrapping is particularly useful for missing data recovery due to the existence of wrapping-up of phase angler by 2π during its transition from $+/-\pi$ to $-/+\pi$. Venkatasubramanian [49] proposed an efficient angle unwrapping technique in real time, by adding the angle during each transition from $+\pi$ to $-\pi$ and subtracting the angle during each transition from $-\pi$ to $+\pi$, as in eqn 6.

$$\min_N |\theta_{i+1} - \theta_i + 360N| \tag{6}$$

$$ROC(i+1) = ROC(i) + N \tag{7}$$

The proposed model in [40] introduced the term $RollOverCounter(ROC)$, that is updated after each transition to store the number of such transition occurrence. ROC is added by $+1$ during $+\pi$ to $-\pi$ transition and is added by -1 during $-\pi$ to $+\pi$ transition, as described in eqn 7.

During false data injection attack (FDIA), the attacker is able to inject malicious data into the communication system and can modify the measurements received by SO from PMUs/ IEDs. When attackers target phase angle measurements, they corrupt the angular measurements with falsified values. For instance, at the timestamp t, a FDIA is initiated by targeting the phase angle measurements, and the phase angle measurements in eqn 6 changes from $\theta(t)$ to $\theta'(t)$. This can be represented as adding an attack value a(t) to the actual phase

angle $\theta(t)$. If the attacker initiates relentless attack as in [50], the next measurement $\theta(t)$ will also change to $\theta(t+1)$. At the transition of the amount of 2π, the actual angle value $\theta(t)$ is close to $+\pi$ and $\theta(t+1)$ is close to $-\pi$. The new phase angle values will be as follows:

$$\theta'(t) = \theta(t) + a(t)$$
$$\theta'(t+1) = \theta(t+1) + a(t+1) \tag{8}$$

Adding an attack vector to the phase angle data, as described in eqn 8, causes an increase or a decrease in the phase angle measurement. However, the time when there is a transition between $+/-\pi$ to $-/+\pi$ (Fig. 6) doesn't change with respect to neighboring PMUs. To make the attack stealthy, the attacker tries to make the $a(t)$ value as small as possible (preferably less than 1^o [32]). As a result, a small attack value $a(t)$, and will not change the transition point between $+/-\pi$ to $-/+\pi$. From eqn 6, it is clear that the N value remains same as the transition status between positive and negative remains same. Therefore FDIA doesn't change the ROC value in eqn 7, making the unwrapped angle graph under attack similar to the unwrapped angle graph during normal condition.

During GSA, the phase angle value shifts the time axis due to the spoofed 1 PPS. The transition point between $+/-\pi$ to $-/+\pi$ no longer occurs at the same timestamp as it would be under no-attack scenario, creating a times-shift in phase angle measurements. Assuming a time-shifted phase angle value of θ'' and an original phase angle value as θ,the relation between θ'' and θ can be expressed as:

$$\theta''(t) = \theta(t + \Delta T) \tag{9}$$

ΔT is the timestamp shift caused by GSA. The data the SO receives at the t^{th} timestamp is actually the data that the power grid generated ΔT timestamp ago. PMU adjust the time reference at each GPS 1 PPS signal, therefore the timestamp shift in eqn 9 will continue for next one second. Each transition point between $+/-\pi$ to $-/+\pi$ over the next 1 sec. period will also shifted, consequently changing the N value from eqn 6 at the time instance t. A change in N results in different ROC value from eqn 7 and causes distortion in unwrapped phase angle curve.

These two opposite behaviors of unwrapped phase angle curve during FDIA and GSA as discussed in this section can be exploited to distinguish these two types of attacks.

V. PROPOSED REAL-TIME TESTBED

The realtime testbed relies on the successful implementations of the algorithms described in section II to IV. At first the algorithm runs Hankel-matrix based model mentioned in section II to detect the bad data. If the occurrence of bad data is detected, the bad data type is identified using Hankel-matrix with multi-channels described in section III. Once the cause of bad data is identified as cyberattack, we apply the single channel Hankel-matrix algorithm in section IV on both the raw and unwrapped phase angle data.

TABLE 2. Algorithm - differentiating GSA from FDIA.

Initialization:	The number of PMU channels M, t_s is the current timestamp, T is the time period from 1^{st} timestamp to timestamp t_s , k is the Hankel matrix parameters, O is the number of PMU channels physically connected to the affected PMU channels, W is the data-window length over total T timestamps;
Step 1:	Create a $Mk \times W - k + 1$ matrix Hankel-matrix H , similar to eqn 1, with measurement from M PMU channels over data-window length of W, the data window starts from the 1^st timestamp and ends at the timestamp T;
Step 2:	For every channel i, calculate the estimation residual for the measurement at timestamp t_{s+1} using $\|\|y_{i,t_{s+1}} - \Gamma_i\|\|_2$;
Step 3:	If the estimation residual at timestamp t_{s+1} doesn't exceed threshold τ^h, measurement at t_{s+1} is not a bad data, proceed with the next time window starting from 2^{nd} timestamp and ending at timestamp t_{s+1} and perform BDD method from step 1;
Step 4:	If the residual at timestamp t_{s+1} exceeds threshold τ^h, measurement at t_{s+1} is a bad data, proceed to step 5;
Step 5:	Apply algorithm in Table I to identify the bad data type;
Step 6:	If identified as cyberattack, proceed to step 7;
Step 7:	Unwrap the angular data and create Hankel-matrix H_u similar to step 1, using measurements with timestamps starting from 2 to the timestamp t_{s+1} ;
Step 8:	Calculate the rank-1 approximation error e^{rc};
Step 9:	If the gradient of rank-1 approximation error is positive for more than three consecutive timestamp moving time window, then it is a GPS-spoofing attack;
Step 10:	Else, it is an FDIA;

For FDIA, only the raw phase angle data will flag anomaly, whereas for GSA both the raw and unwrapped phase angle data will flag divergences in the data. The detailed process is described in algorithm in Table 2.

The testbed used in this work, depicted in Fig. 1, contain three parts: the PMU channels, PDC, and the communication link. The synchrophaosor data transmission is performed with the *PMU simulator* model developed in [51]. The Phasor Data Concentrator (PDC) is modeled with *openECA* platform [52]. Both the *PMU simulator* and *openECA* are open source tools. The *PMU simulator* in the test system provides UTC timestamps starting at midnight (00:00:00) of Oct 10, 2022 with a sampling interval of 1/60 sec. The simulator transmit timestamped voltage and current phasors and frequency measurements received from physical power system to *openECA* through TCP or UDP communication protocols.

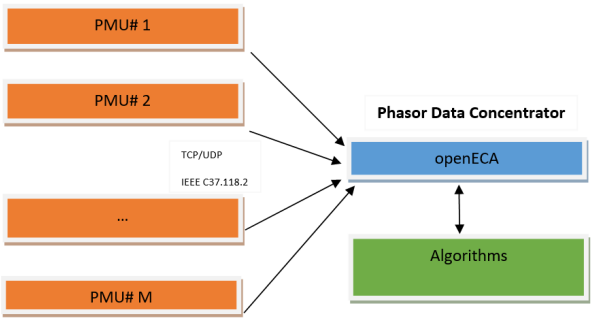

FIGURE 1. Proposed PMU-PDC testbed.

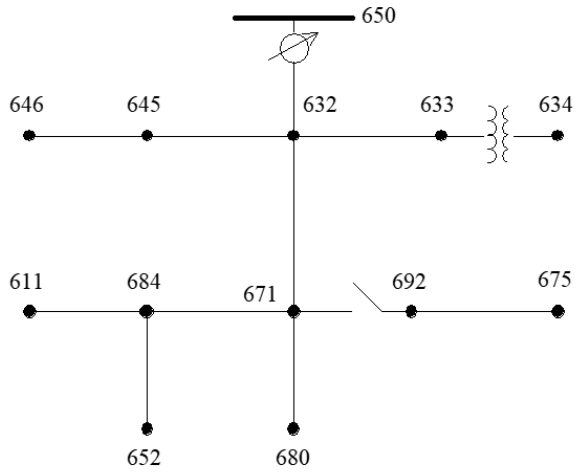

FIGURE 2. IEEE 13 node test feeder.

openECA aligns data from all the PMU channels with the timestamps and is able to communicate with the next layer of cyber system/ system server. The data format during all communication between PMU-PDC-PMU and PDC to SO is in IEEE C37.118.2 protocol [53]. The physical power system is modeled with IEEE 13 node test feeder system [54], simulated in *MATLABSIMULINK*, with PMUs added at buses 611, 632, 633, 634, 646, 671, 675, 680 and 692 (Fig. 2). The physical event is modeled with three separate line to ground faults at each of the lines connecting bus-671 to bus-692, bus-632 to bus-633 and bus-671 to bus-680. For each line, three types of line-to-ground faults are applied, i.e. single line to ground (SLG), double line to ground (DLG), and triple line to ground (TLG) faults. Furthermore, fault impedance during each type of fault is varied from 0.01 $p.u.$ to 0.1 $p.u.$, with a step size of 0.01 $p.u.$. This variations in fault impedance, location, and fault types generate a total $10 \times 3 \times 3 = 90$ number of physical events.

Two types of FDIA are modelled in this work to verify two different part of proposed sequential algorithm. The first type is done by adding an attack value a to the voltage magnitude measurements. The goal of the attacker is to avoid getting detected by SO, therefore a should follow a trade-off between small enough to avoid detection and large enough to cause significant impact to the system. To demonstrate the feasibility of our proposed realtime bad data detection and classification schemes, a variable attack value between 1% to 5% of the peak voltage magnitude measurements are added to the actual measurements. Second type of FDIA will be discussed later at this section.

The simulated PMU data from *PMU simulator* is transmitted at a rate of 60 frame per second to *openECA*. The *openECA* is enabled with *TestHarness* system that runs the user-defined algorithm to perform designated tasks. Algorithm in Table 2 can be implemented in the *TestHarness* system in real time. At each timestamp t_s, the algorithm receives N_s number of measurements from previous N_s times stamps, spanning from the measurement with timestamp $t_s - N_s + 1$ to the measurement with timestamp t_s. A major advantage of the proposed model is it is computationally efficient. The algorithms are executed in python, with an Intel Core i5, 1.8 GHz CPU and 8.00 GB RAM. The PMU-PDC testbed runs the *TestHarness* in realtime, and the *TestHarness* calls the python script at every 1 millisecond to 0.1 seconds, depending on the data-window length.

Considering the beginning of simulation as timestamp 00:00:00.000, the IEEE 13 bus system as in Fig. 2 runs over 50 second for each test case, generating $60 \times 50 = 3000$ measurements. The final timestamp of the measurements received by *openECA* from *PMU simulator* is 00:00:49.833. Initially, testbed is simulated for normal condition, i.e. no physical event or cyberattack. This normal condition provides bad data detection threshold τ^h. For each of 90 physical events represented by different line to ground faults described above, the testbed is simulated over 50 seconds, with the fault applied to the grid at the timestamp 00:00:14.117 and is removed at the timestamp 00:00:28.050. For the first type of FDIA, each of three different attack values are applied to the voltage measurement at the same timestamp 00:00:14.117 for respective simulation of the testbed, the three attack values being 1%, 3%, and 5% of rated peak voltage measurement. The FDIA targeting the voltage measurement is used for the second part of the proposed algorithm, that is identifying bad data type.

After the cause of bad data is identified as cyberattack, the proposed model uses phase angle data to distinguish GSA from FDIA as discussed in previous section. GPS-spoofing attack is modeled by shifting the timestamp at the PMU simulator by $\frac{\Delta\theta}{2\pi \times 60}$. The term $\Delta\theta$ represents the deviation in phase angle due to the GSA, as shown in eqn 9. As the most significant impact of GPS 1 PPS shift during GSA is reflected by a change in phase angle data, we have added a phase angle deviation $\Delta\theta$ to the actual measurements to demonstrate the effect of GSA. Phase angle deviation larger than 0.57 degrees results in a Total Vector Error (TVE) 1% [44], therefore the attacker must ensure the GPS 1 PPS shift to be small enough to produce an angle shift ($\Delta\theta$) less than 0.57^o. For a particular amount of GPS 1 PPS signal shift that is equivalent to $\Delta\theta$, the phase angle will be shifted by $\Delta\theta$ the next 1 second. In order to reflect two consecutive GPS 1 PPS shift, first phase angle

shift is applied at the timestamp 00:00:14.117, and the next 120 sample data is also shifted by the same degree, until the timestamp, 00:00:16.100.

Our goal is to differentiate GSA from FDIA, therefore a second type of FDIA should come into consideration with attack values applied to the phase angle data only. This second type of FDIA can be used for the third part of proposed algorithm, that is identifying cyberattack type. For the second type of FDIA, three different test cases are created with three different phase angle shifts with the values: 0.1^o, 0.3^o and 0.5^o. Each test cases are applied at the same 00:00:14.117 timestamp as previous faults and cyberattacks.

The combined sequential bad data detection and classification model is further tested with IEEE 118 bus system [55], simulated in Python-Siemens PSS/E to implement dynamics. We have created 20 separately physical events by applied line-ground faults with four different impedance (0.0005 p.u, 0.005 p.u., 0.05 p.u., and 0.5 p.u.) at five different branches. The five branches are: branch connecting bus 49 and bus 66, branch connecting bus 56 and bus 57, branch connecting bus 77 and bus 80, branch connecting bus 89 and bus 90, and branch connecting bus 100 and 103. The FDIA is simulated by applying an attack values of 1%, 3% and 5% of the peak voltage measurements. The physical events and cyberattacks for IEEE 118 bus systems are analyzed under same computational environment as it is for IEEE 13 node test feeder, and the LG faults and FDIA are applied at the timestamp 00:00:14.117 and are removed at the timestamp 00:00:28.05.

VI. SIMULATION RESULT

Each of the three parts of algorithm in Table 2 is implemented in the test harness sequentially, the first part, which is the BDD is being executed over times-series moving data-window in real time. If BDD indicates the existence of bad data, the second part, which is differentiating physical event and cyberattack, is executed for the next time-series moving data-window. When the type of bad data is identified as cyberattack, the third part is executed over the same data-window to determine the cyberattack type: FDIA or GSA.

A. PART I: BAD DATA DETECTION RESULTS

For each test case of physical events and cyberattacks, PMU simulator transmit data to openECA at GPS synchronized timestamp t_s, and openECA feeds the data into the test harness. The test harness runs the bad data detection algorithm using the measurement from timestamp t_s along with $T-1$ number of previous voltage and current measurements, constituting a time-window with size T. For a test case scenario under normal condition, i.e. no bad data, set V_m denotes the set of voltage magnitude measurements over 50 sec, period, and σ denotes the standard deviation. The bad data threshold τ^h can be determined using the relation as follows:

$$\tau^h = max(V_m) + 3\sigma(V_m) \tag{10}$$

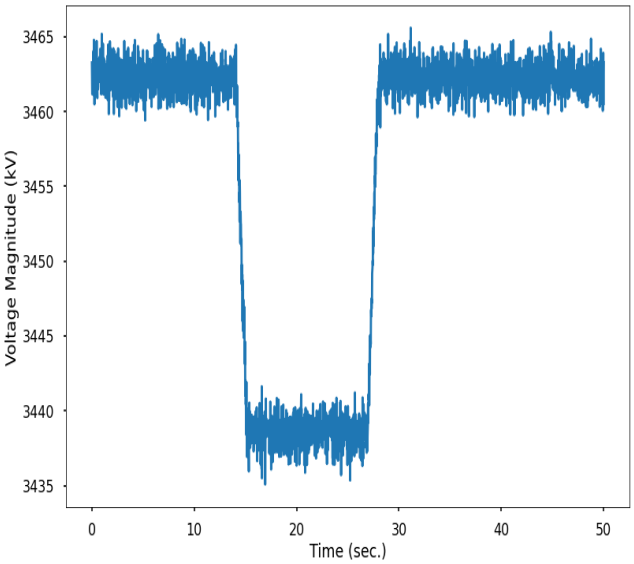

FIGURE 3. Noisy voltage magnitude measurement near bus 80.

At timestamp 00 : 00 : 14.117, a sample physical fault testcase with a TLG fault has been applied near bus 680 of IEEE 13 bus node test feeder. The TLG fault is removed at the timestamp 00 : 00 : 28.0167. The voltage magnitude measurements at bus 80 with random Gaussian noise of mean 0 and standard deviation 1 reflects the change over the fault duration (Fig. 3).

For the data-window length of 100, the estimation residual using Hankel-matrix as described in algorithm in Table 1 remains less than the threshold τ^h before the fault occurrence. However, just after the occurrence of the TLG fault, estimation residual exceeds the τ^h for more than 3 consecutive measurements, as in Fig. 4. Therefore, the system operator can detect the occurrence of bad data around timestamp 00 : 00 : 14.117.

The accuracy of Hankel-matrix based bad data detection method is expected to depend on the data-window length W. Smaller data-window length provides smaller dataset to estimate the next measurement, therefore it can be assumed that BDD accuracy increases with larger data-window length W. To check the dependence of accuracy on the data-window length W, total number of 99 testcases, including 90 physical events and 9 FDIAs are executed in the testbed. The proposed method provides 100% accuracy for data window length larger than 130. Smaller window length reduces the BDD accuracy, which confirms the expected relation between data-window length and accuracy.

B. PART II: BAD DATA CLASSIFICATION RESULTS

After the existence of bad data is detected at timestamp 00:00:14.117, the second part of algorithm in Table 2, which is identification of bad data type, is executed in the test harness. Theoretically, the low rank approximation error (eqn 4) for physical event is expected to change after random column

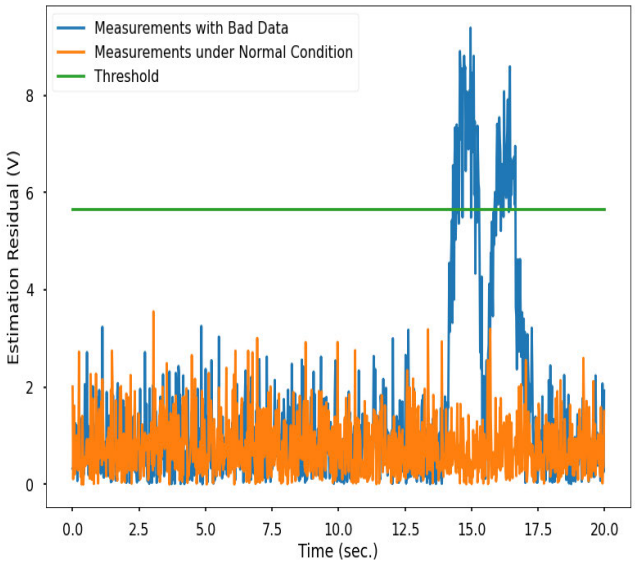

FIGURE 4. Bad data detection using hankel matrix.

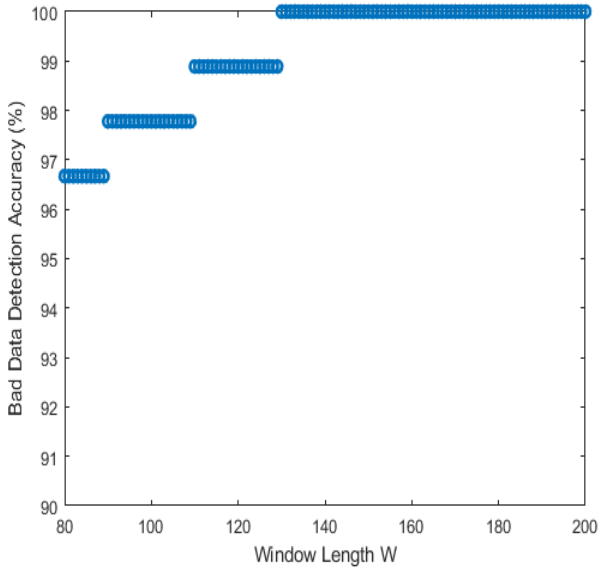

FIGURE 5. BDD accuracy variation with data-window length.

permutation, whereas for cyberattack the change is expected to be zero. Due to the existence of measurement noise and randomness of column permutation, the change in low rank approximation error before and after random column permutation may not be exactly zero. Therefore a threshold η as described in algorithm in Table 1 must be selected to indicate any significant change in the low rank approximation error. A heuristic value of $\eta = 0.1$ is selected as threshold to distinguish cyberattack and physical event.

We have executed two testcases in the testbed to verity the proposed algorithm: the first being a TLG fault at near bus 692 with fault impedance 0.01 p.u. and the second one being a FDIA with 1% deviation added to voltage magnitude measurements from bus 692, both initiated at the timestamp 00:00:14.117 separately. With a data-window length of 120, the bad data is detected at first for the timestamp 00:00:14.117, with similar results as shown in Fig. 4. When bad data is detected, the second part of algorithm in Table 2 is executed immediately, with the measurement from timestamp 00:00:14.117 and the previous 119 measurements, constituting total data-window length of 120. For each test case, algorithm uses multi-channel measurement matrix as depicted in eqn 2. Since bus 692 is physically connected with bus 671 and bus 675, the measurement matrix contains three rows, and the number O from algorithm in Table 1 is 3.

Measurements from the timestamp 00:00:14.117, which is the 847^{th} measurement, contain the first instance of bad data, the random column permutation based method sometimes fail to indicate the correct fault type. The bad data identification algorithm needs to applied for the next data set, that contains measurements from the timestamp next to 00:00:14.117, which is 00:00:14.133 or the 848^{th} measurements, and the previous 119 measurements. The process has to be repeated for moving data-window over time until the data-window contains enough measurement points to demonstrate significant change in low rank approximation error after random column permutation. For 120 data-window length, low rank approximation error after random column permutation shows a change larger than $\eta = 0.1$ for the first time for a data-window starting from 760^{th} measurement and ending at 880^{th} measurement. Fig. 6 indicates that the rank-1 approximation error is larger than 0.1 for TLG fault. The change in low rank approximation error before and after random column permutation over the same data-window for the testcase containing FDIA data remains approximately 0, implying a different behavior for cyberattack and physical event.

If the low rank approximation error indicates a change larger than threshold η after random column permutation for the data-window starting from measurements of timestamp $t_s - 2W/3$ to measurements of timestamp $t_s + W/3$, cause of bad data is identified as physical event, and the corresponding restorative actions need to be taken by the system operation. However, if the bad data type is identified as cyberattack, the third part of the proposed algorithm is executed using the same data-window that demonstrated the change in low rank approximation error as greater than η.

As mentioned before, data-window needs to contain a minimum number of measurements to indicate any change in low rank approximation error after random column permutation. algorithm in Table 1 has reported correct bad data type for $W \geq 73$. For instance, Fig. 7 depicts the failure of proposed model in identifying bad data type correctly when W = 70, since both of physical event and cyberattack display insignificant changes ($\ll \eta$) in rank-1 approximation errors.

C. PART III: DIFFERENTIATING FDIA FROM GSA RESULTS

In the next step, the algorithm computes low rank approximation error of unwrapped voltage phase angle measurements.

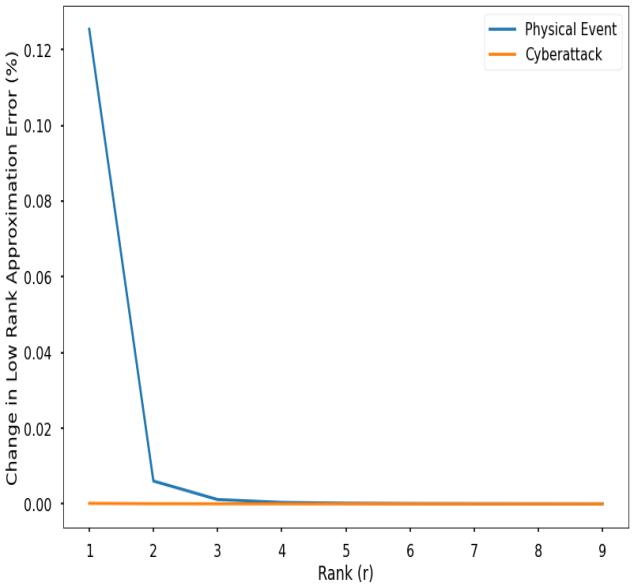

FIGURE 6. Bad data type identification: differentiating cyberattack from physical event W = 120.

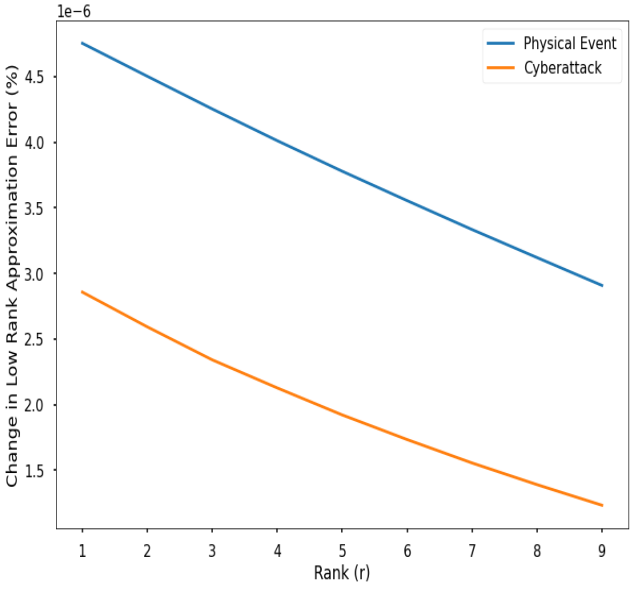

FIGURE 7. Bad data type identification: differentiating cyberattack from physical event with W = 70.

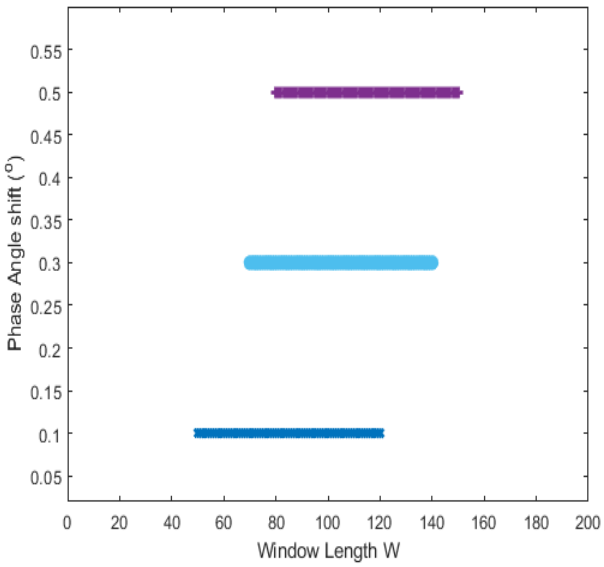

FIGURE 8. Data-window lengths' variations with different GSA for differentiating GSA from FDIA.

If the gradient of rank-1 approximation error is larger than 0 for three consecutive points, the type of cyberattack can be determined as GPS-spoofing attack, otherwise, it is FDIA. A testcase is executed with 0.3^o deviation added to voltage angle measurements of bus 692 over two second starting at the timestamp 00:00:14.117 to reflect shift in two consecutive GPS 1 PPS signal, thereby imitating GPS-spoofing attack. We have observed positive gradients of low rank approximation error for single channel measurements for more than 3 consecutive measurements. Another testcase of FDIA is executed with 0.3^o deviation added to voltage angle measurement of bus 692 over data-window length of 120. The result shows no positive gradient of low rank approximation error for the single channel phase angle measurements over same data-window length, confirming the FDIA as the cyberattack type in this testcase.

Smaller length of data-window W may provide insufficient dataset for the Hankel matrix to demonstrate any significant change in low rank approximation error. For GPS-spoofing attack, after GPS 1 PPS signal is shifted, phase angle measurements for the next 1 sec. are modified, along with the transition points from $+/-\pi$ to $-/+\pi$. As a result, the unwrapped phase angle data exhibit deviation for the next 1 sec., that is until the arrival of another GPS 1 PPS signal. Proposed Hankel-matrix based model relies on the sudden variation or break-point of unwrapped phase angle data at the moment of attack. For the case of a very large number of dataset is fed into Hankel-matrix, just one break-point in the unwrapped phase angle measurements is not sufficient to display any significant change in low rank approximation error. Therefore a there exists a trade-off between smaller and larger data-window length. Analyzing test-cases with different phase angle shift due to GSA over different data-window length confirms this assumption (Fig. 8). Larger phase angle shift (0.5 degree) demonstrates positive gradient in low rank approximation error over larger data-window length. From Fig. 8, it can be concluded that data-window length from 80 to 120 works best for distinguishing GSA from FDIA over wide range of phase angle variation.

Even though the proposed model is applied in realtime, there are computational and network constraints. Each part of algorithm in Table 2 require a fraction of second to be executed. Data-window length affects the computational time,

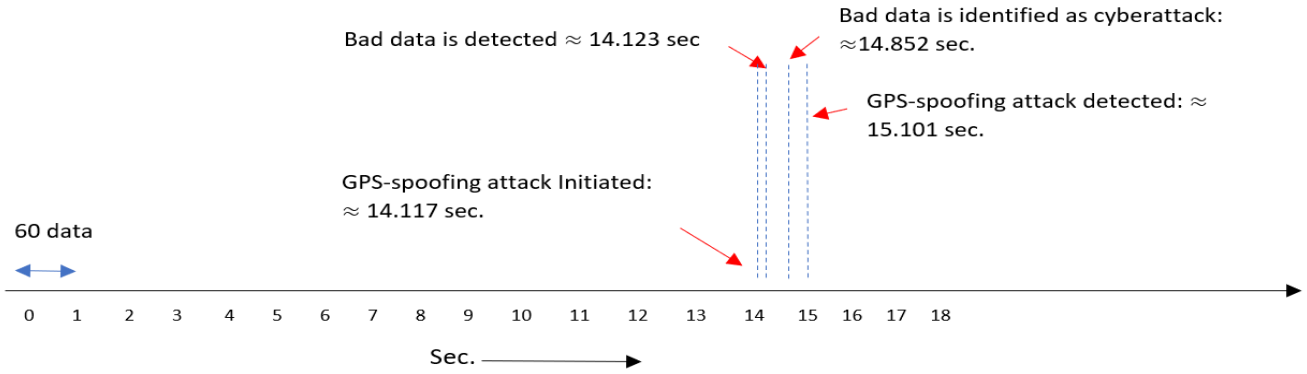

FIGURE 9. Time-series visualization of sequential realtime implementation of algorithm in Table 2 in testbed from Fig. 1 (IEEE 13 node test feeder).

TABLE 3. Computational time vs data-window length.

Data-window length W	BDD	bad data classification	Cyberattack type identification
80	0.006437 sec.	0.085991 sec.	0.228165 sec.
90	0.006327 sec.	0.095918 sec.	0.254029 sec.
120	0.011000 sec.	0.149227 sec.	0.405938 sec.
130	0.016136 sec.	0.176000 sec.	0.498000 sec.

the larger the data-window, longer it takes to execute the algorithm. Variations of average computational time of each portion of proposed realtime algorithm against data-window length W are tabulated in Table 3. Total computational time for data-window length 90 is $\approx$ 0.91 *sec*., including the time required to move the time-series data-window to include enough measurements for bad data type identification in subsection VI-B.

From the above analysis of the impact of data-window length on the effectiveness of each part of algorithm, data window length of 90-100 can be considered as optimum data-window length, since it can indicate the bad data occurrence, its type and the type of cyberattack in less than 1 sec. The effectiveness of the testbed used to implement the proposed realtime bad data detection model is visually depicted in Fig. 9. Bad data detection model is executed in realtime for moving time series data-window with length 90, using a test-case of GSA with 0.3^o deviation in phase angle measurement. Bad data is initialized at the timestamp 00:00:14.117 sec., and the proposed detection model indicates the occurrence of bad data after $\approx$ 0.006 *sec*. of computational time. The next two parts are executed sequentially. Bad data type identification requires 0.55 sec. to include enough measurements to the moving data-window so that it can demonstrate low rank approximation error change after random column permutation. It requires approximately 0.096 sec. to execute the low rank approximation error after random column permutation for data-window length of 90. Immediately after the detection of cyberattack, the third part requires approximately 0.25 sec. to identify GPS-spoofing attack. Considering the network delay of 0.033 sec. in *WIRESHARK* [56], we can correctly identify and locate the GPS-spoofing attack in the cyber-physical system in less than 1 sec.

D. RESULTS WITH IEEE 118 BUS SYSTEM

The dynamic simulation by SIEMENS PSS/E with LG faults applying at 5 different locations separately, each with 4 different fault impedances show similar result as of IEEE 13 node test feeder. For the LG fault of 0.005 p.u. impedance applied at the branch connecting bus 49 and bus 66, the data anomaly is detected within 0.02 second of timestamp 14.117 sec (Fig. 10). The low rank approximation error from Fig. 10 is calculated using the estimation of voltage measurement at bus 49, with data-window length of 100.

After detecting the existence of bad data, the voltage measurements are passed to the bad data classification step described in section III. The bad data is correctly classified as physical event (Fig. 11). The similar steps are applied for cyberattack, formulated by applying FDIA with attack value 0.1% of voltage measurements. Fig. 11 shows that the change in low rank approximation error is much larger for physical event than it is for cyberattack. This scenario is expected, since a larger system like IEEE 118 bus system, there are more interconnected buses to a particular node. As a result, a random column permutation will destroy the temporal relation among interconnected nodes more severely than it does for smaller system with small number of interconnected nodes. Also, a very short data-window, for example the window length of 40, is enough to detect this change in low rank approximation error.

The computational time for sequential detection and classification with varying data-window length is illustrated in Table 4. The additional computational time in IEEE 118 bus system comes from the multiple channel Hankel-matrix from eqn 3, the larger system with more interconnected nodes leads to a larger multiple channel Hankel-matrix. Including the communication delay between PMUs and PDCs, the total bad data detection and classification time with data-window length of 90 is approximately 2.9519 sec. However, with

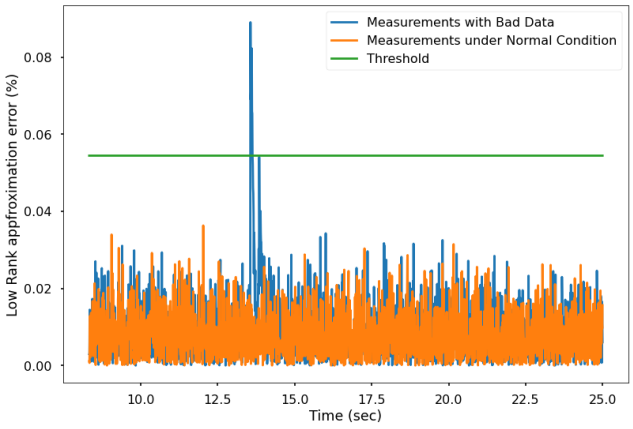

FIGURE 10. BDD using Hankel-matrix for IEEE 118 bus system.

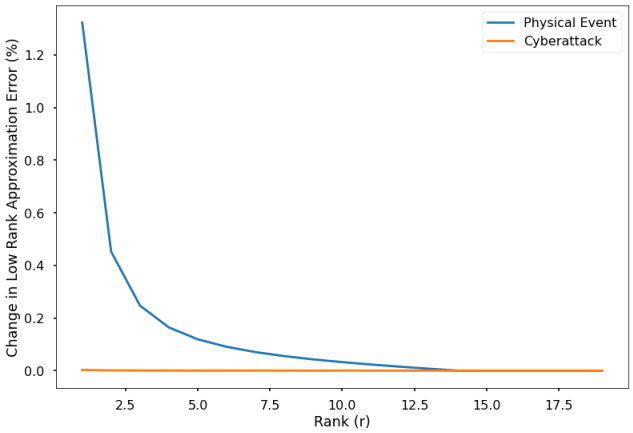

FIGURE 11. Bad data type identification: differentiating cyberattack from physical event with W = 40.

TABLE 4. Computational time vs data-window length for IEEE 118 bus system.

Data-window length W	BDD	bad data classification	Cyberattack type identification
40	0.001999 sec.	0.271844 sec.	0.238516 sec.
80	0.006996 sec.	0.468730 sec.	0.224139 sec.
100	0.013992 sec.	0.623647 sec.	0.415231 sec.
120	0.020988 sec.	0.845527 sec.	0.504000 sec.

smaller data window of length 40, total bad data detection and classification time is approximately 1.8323 sec, less than 2 sec.

VII. CONCLUSION

In this work we proposed a realtime bad data detection and bad data type identification strategy. At first, we have exploited the low rank property of Hankel-matrix to detect the occurrence of bad data in realtime. Secondly, we classify the bad data and differentiate in two categories: physical events and cyberattacks. The algorithm utilizes the difference in low rank approximation error of multi-channel Hankel-matrix before and after random column performance during physical events. If the type of bad data is determined to be cyberattack, our proposed algorithm proceeds to identify the cause of cyberattack. We have considered two possible cyberattack types: false data injection attack (FDIA) and GPS-spoofing Attack (GSA). The proposed algorithm observes rank-1 approximation error of single-channel Hankel-matrix containing unwrapped phase angle measurements to distinguish FDIA from GSA. Finally, the proposed algorithm is implemented in a realtime cyber-physical testbed containing PMU simulator and openECA.

Bad data can be detected with 100% accuracy if the Hankel-matrix data-window length is larger than 130. The second part of the algorithm can correctly identify bad data type for data-window length of more than 73. Results from the testbed show that the optimum size of Hankel-matrix data-window lies within the range of 90-120. In this range, the bad data can be detected as well as the type of bad data is correctly identified within less than 1 sec. of the occurrence of physical event or cyberattacks, considering the network latency. For larger system such as IEEE 118 bus, the proposed sequential model can correctly detect and classify bad data with smaller data-window length due to temporal relations among higher number of interconnected nodes. Smaller data-window leads to smaller computational time for bad data detection, however, the increase in the size of multiple channel Hankel-matrix for larger grid system causes longer computational time for bad data classification step. Therefore, there exists a trade-off between data-window length and computational time for larger and mroe complex grid system.

The proposed realtime Hankel-matrix based sequential bad data detection and classification provides accurate results within a very short period. Nevertheless, application of proposed model for highly sophisticated coordinated attacks, where the attackers have much detailed knowledge of the grid architecture and can mimic a temporal relation among neighboring PMUs like line faults, is yet to be explored. Future works can focus on improving the Hankel-matrix based algorithm for coordinated cyberattacks.

REFERENCES

[1] D. Ton and M. Smith, "The U.S. department of energys microgrid initiative," *Electr. J.*, vol. 25, no. 8, pp. 84–94, 2012, online: [Online]. Available: https://www.energy.gov/sites/prod/files/2016/06/f32/The

[2] F. Nejabatkhah, Y. W. Li, and H. Tian, "Power quality control of smart hybrid AC/DC microgrids: An overview," *IEEE Access*, vol. 7, pp. 52295–52318, 2019, doi: 10.1109/ACCESS.2019.2912376.

[3] M. Glinkowski, L. Simmons, D. Loucks, D. Becker, I. Bitterlin, B. Campbell, H. Handlin, P. Lembke, T. Earp, D. LeRoy, and A. Lynch, "Data center power system harmonics: An overview of effects on data center efficiency and reliability," Tech. Rep., 2014.

[4] C. Su, J. Yu, H. Chin, and C. Kuo, "Evaluation of power-quality field measurements of an electric bus charging station using remote monitoring systems," in *Proc. 10th Int. Conf. Compat., Power Electron. Power Eng. (CPE-POWERENG)*, Jun. 2016, pp. 58–63.

[5] Q. Li, S. Tao, X. Xiao, and J. Wen, "Monitoring and analysis of power quality in electric vehicle charging stations," in *Proc. 1st Int. Future Energy Electron. Conf. (IFEEC)*, Nov. 2013, pp. 277–282.

[6] S. M. M. Gazafrudi, A. Tabakhpour Langerudy, E. F. Fuchs, and K. Al-Haddad, "Power quality issues in railway electrification: A comprehensive perspective," *IEEE Trans. Ind. Electron.*, vol. 62, no. 5, pp. 3081–3090, May 2015.

[7] K. D. Pham, R. S. Thomas, and W. E. Stinger, "Operational and safety considerations in designing a light rail DC traction electrification system," in *Proc. IEEE/ASME Joint Railroad Conf.*, Apr. 2003, pp. 171–189.

[8] Y. Pan, P. M. Silveira, M. Steurer, T. L. Baldwin, and P. F. Ribeiro, "A fault location approach for high-impedance grounded DC shipboard power distribution systems," in *Proc. IEEE Power Energy Soc. Gen. Meeting-Convers. Del. Electr. Energy 21st Century*, Jul. 2008, p. 16.

[9] X. K. Wang, X. Yang, J. H. He, B. Kirby, C. Y. Dong, and D. Writer, "Grounding fault location in DC railway system," in *Proc. 22nd Int. Conf. Exhib. Electr. Distribution (CIRED)*, Jun. 2013, pp. 1–4.

[10] C. Y. Dong, J. H. He, X. K. Wang, J. F. Xu, L. Yu, and Z. Q. Bo, "High-resistance grounding fault detection and location in DC railway system," in *Proc. 11th IET Int. Conf. Develop. Power Syst. Protection (DPSP)*, Apr. 2012, pp. 1–5.

[11] J. Park, "Ground fault detection and location for ungrounded DC traction power systems," *IEEE Trans. Veh. Technol.*, vol. 64, no. 12, pp. 5667–5676, Dec. 2015, doi: 10.1109/TVT.2015.2388785.

[12] S. Sahoo, S. Mishra, S. Jha, and B. Singh, "A cooperative adaptive droop based energy management and optimal voltage regulation scheme for DC microgrids," *IEEE Trans. Ind. Electron.*, vol. 67, no. 4, pp. 2894–2904, Apr. 2020.

[13] A. Vasilakis, I. Zafeiratou, D. T. Lagos, and N. D. Hatziargyriou, "The evolution of research in microgrids control," *IEEE Open Access J. Power Energy*, vol. 7, pp. 331–343, 2020.

[14] S. Mousavian, J. Valenzuela, and J. Wang, "A probabilistic risk mitigation model for cyber-attacks to PMU networks," *IEEE Trans. Power Syst.*, vol. 30, no. 1, pp. 156–165, Jan. 2015, doi: 10.1109/TPWRS.2014.2320230.

[15] C. K. Veitch, J. M. Henry, B. T. Richardson, and D. H. Hart, "Microgrid cyber security reference architecture," Sandia Nat. Lab., Albuquerque, NM, USA, Tech. Rep., SAND2013-5472, 2013.

[16] S. Sahoo, J. C. H. Peng, A. Devakumar, S. Mishra, and T. Dragicevic, "On detection of false data in cooperative DC microgrids—A discordant element approach," *IEEE Trans. Ind. Electron.*, vol. 67, no. 8, pp. 6562–6571, Aug. 2020.

[17] M. Conti, N. Dragoni, and V. Lesyk, "A survey of man in the middle attacks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 2027–2051, 3rd Quart., 2016.

[18] S. Sahoo, T. Dragicevic, and F. Blaabjerg, "Multilayer resilience paradigm against cyber attacks in DC microgrids," *IEEE Trans. Power Electron.*, vol. 36, no. 3, pp. 2522–2532, Mar. 2021, doi: 10.1109/TPEL.2020.3014258.

[19] J. Zhao, L. Mili, and M. Wang, "A generalized false data injection attacks against power system nonlinear state estimator and countermeasures," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 4868–4877, Sep. 2018.

[20] J. Kim, L. Tong, and R. J. Thomas, "Subspace methods for data attack on state estimation: A data driven approach," *IEEE Trans. Signal Process.*, vol. 63, no. 5, pp. 1102–1114, Mar. 2015.

[21] J. Hao, R. J. Piechocki, D. Kaleshi, W. H. Chin, and Z. Fan, "Sparse malicious false data injection attacks and defense mechanisms in smart grids," *IEEE Trans. Ind. Informat.*, vol. 11, no. 5, pp. 1–12, Oct. 2015, doi: 10.1109/TII.2015.2475695.

[22] Y. Fan, Z. Zhang, M. Trinkle, A. D. Dimitrovski, J. B. Song, and H. Li, "A cross-layer defense mechanism against GPS spoofing attacks on PMUs in smart grids," *IEEE Trans. Smart Grid*, vol. 6, no. 6, pp. 2659–2668, Nov. 2015.

[23] Z. Zhang, S. Gong, A. D. Dimitrovski, and H. Li, "Time synchronization attack in smart grid: Impact and analysis," *IEEE Trans. Smart Grid*, vol. 4, no. 1, pp. 87–98, Mar. 2013.

[24] B. Pardhasaradhi, P. Srihari, and P. Aparna, "Navigation in GPS spoofed environment using M-best positioning algorithm and data association," *IEEE Access*, vol. 9, pp. 51536–51549, 2021.

[25] D. P. Shepard, T. E. Humphreys, and A. A. Fansler, "Evaluation of the vulnerability of phasor measurement units to GPS spoofing attacks," *Int. J. Crit. Infrastruct. Protection*, vol. 5, nos. 3–4, pp. 146–153, Dec. 2012.

[26] C. Konstantinou, M. Sazos, A. S. Musleh, A. Keliris, A. Al-Durra, and M. Maniatakos, "GPS spoofing effect on phase angle monitoring and control in a real-time digital simulator-based hardware-in-the-loop environment," *IET Cyber-Phys. Syst., Theory Appl.*, vol. 2, no. 4, pp. 180–187, 2017, doi: 10.1049/iet-cps.2017.0033.

[27] M. M. Rana, L. Li, and S. W. Su, "Cyber attack protection and control of microgrids," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 2, pp. 602–609, Mar. 2018.

[28] D. Jin, Z. Li, C. Hannon, C. Chen, J. Wang, M. Shahidehpour, and C. W. Lee, "Toward a cyber resilient and secure microgrid using software-defined networking," *IEEE Trans. Smart Grid*, vol. 8, no. 5, pp. 2494–2504, Sep. 2017.

[29] Y. Li, P. Zhang, L. Zhang, and B. Wang, "Active synchronous detection of deception attacks in microgrid control systems," *IEEE Trans. Smart Grid*, vol. 8, no. 1, pp. 373–375, Jan. 2017.

[30] G. Intriago and Y. Zhang, "Online dictionary learning based fault and cyber attack detection for power systems," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Jul. 2021, pp. 1–5, doi: 10.1109/PESGM46819.2021.9637891.

[31] Y. Zhao, X. Jia, D. An, and Q. Yang, "LSTM-based false data injection attack detection in smart grids," in *Proc. 35th Youth Academic Annu. Conf. Chin. Assoc. Autom. (YAC)*, Oct. 2020, pp. 638–644, doi: 10.1109/YAC51587.2020.9337674.

[32] B. D. Barros, N. K. D. Venkategowda, and S. Werner, "Quickest detection of stochastic false data injection attacks with unknown parameters," in *Proc. IEEE Stat. Signal Process. Workshop (SSP)*, Jul. 2021, pp. 426–430, doi: 10.1109/SSP49050.2021.9513837.

[33] F. Ünal, A. Almalaq, S. Ekici, and P. Glauner, "Big data-driven detection of false data injection attacks in smart meters," *IEEE Access*, vol. 9, pp. 144313–144326, 2021, doi: 10.1109/ACCESS.2021.3122009.

[34] S. Mohammadi, F. Eliassen, Y. Zhang, and H. Jacobsen, "Detecting false data injection attacks in peer to peer energy trading using machine learning," *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 5, pp. 3417–3431, Sep. 2022, doi: 10.1109/TDSC.2021.3096213.

[35] D. Huang, X. Shi, and W. Zhang, "False data injection attack detection for industrial control systems based on both time- and frequency-domain analysis of sensor data," *IEEE Internet Things J.*, vol. 8, no. 1, pp. 585–595, Jan. 2021, doi: 10.1109/JIOT.2020.3007155.

[36] J. J. Q. Yu, Y. Hou, and V. O. K. Li, "Online false data injection attack detection with wavelet transform and deep neural networks," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3271–3280, Jul. 2018, doi: 10.1109/TII.2018.2825243.

[37] Y. Hao, M. Wang, J. H. Chow, E. Farantatos, and M. Patel, "Modelless data quality improvement of streaming synchrophasor measurements by exploiting the low-rank Hankel structure," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 6966–6977, Nov. 2018, doi: 10.1109/TPWRS.2018.2850708.

[38] S. Zhang and M. Wang, "Correction of corrupted columns through fast robust Hankel matrix completion," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2580–2594, May 2019, doi: 10.1109/TSP.2019.2904021.

[39] M. Yi, M. Wang, T. Hong, and D. Zhao, "Bayesian high-rank Hankel matrix completion for nonlinear synchrophasor data recovery," *IEEE Trans. Power Syst.*, early access, Mar. 10, 2023, doi: 10.1109/TPWRS.2023.3254909.

[40] A. Mohammad Saber, A. Youssef, D. Svetinovic, H. H. Zeineldin, and E. F. El-Saadany, "Anomaly-based detection of cyberattacks on line current differential relays," *IEEE Trans. Smart Grid*, vol. 13, no. 6, pp. 4787–4800, Nov. 2022, doi: 10.1109/TSG.2022.3185764.

[41] S. Siamak, M. Dehghani, and M. Mohammadi, "Dynamic GPS spoofing attack detection, localization, and measurement correction exploiting PMU and SCADA," *IEEE Syst. J.*, vol. 15, no. 2, pp. 2531–2540, Jun. 2021, doi: 10.1109/JSYST.2020.3001016.

[42] S. Barreto, M. Pignati, G. Dán, J. Le Boudec, and M. Paolone, "Undetectable timing-attack on linear state-estimation by using rank-1 approximation," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3530–3542, Jul. 2018, doi: 10.1109/TSG.2016.2634124.

[43] I. Khan and V. Centeno, "Detecting GPS-spoofing attack on PMU data with phase angle unwrapping technique and low-rank approximation of Hankel matrix," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Jul. 2022, pp. 1–5, doi: 10.1109/PESGM48719.2022.9916859.

[44] R. Deng, G. Xiao, R. Lu, H. Liang, and A. V. Vasilakos, "False data injection on state estimation in power systems—Attacks, impacts, and defense: A survey," *IEEE Trans. Ind. Informat.*, vol. 13, no. 2, pp. 411–423, Apr. 2017, doi: 10.1109/TII.2016.2614396.

[45] K. Manandhar, X. Cao, F. Hu, and Y. Liu, "Detection of faults and attacks including false data injection attack in smart grid using Kalman filter," *IEEE Trans. Control Netw. Syst.*, vol. 1, no. 4, pp. 370–379, Dec. 2014, doi: 10.1109/TCNS.2014.2357531.
[46] C. Pei, Y. Xiao, W. Liang, and X. Han, "A deviation-based detection method against false data injection attacks in smart grid," *IEEE Access*, vol. 9, pp. 15499–15509, 2021, doi: 10.1109/ACCESS.2021.3051155.
[47] G. M. De Mijolla, S. Konstantinopoulos, P. Gao, J. H. Chow, and M. Wang, "An evaluation of algorithms for synchrophasor missing data recovery," in *Proc. Power Syst. Comput. Conf. (PSCC)*, Jun. 2018, pp. 1–6, doi: 10.23919/PSCC.2018.8442776.
[48] A. Xue, H. Kong, F. Xu, J. Zhao, N. Chang, J. H. Chow, and H. Hong, "Correction of time-varying PMU phase angle deviation with unknown transmission line parameters," *CSEE J. Power Energy Syst.*, vol. 9, no. 1, pp. 315–325, Jan. 2023, doi: 10.17775/CSEEJPES.2021.07280.
[49] V. Venkatasubramanian, "Real-time strategies for unwrapping of synchrophasor phase angles," *IEEE Trans. Power Syst.*, vol. 31, no. 6, pp. 5033–5041, Nov. 2016, doi: 10.1109/TPWRS.2016.2538209.
[50] Y. Liu and L. Cheng, "Relentless false data injection attacks against Kalman-filter-based detection in smart grid," *IEEE Trans. Control Netw. Syst.*, vol. 9, no. 3, pp. 1238–1250, Sep. 2022, doi: 10.1109/TCNS.2022.3141026.
[51] [Online]. Available: https://iitbpdc.sourceforge.net/
[52] [Online]. Available: https://github.com/GridProtectionAlliance/openECA
[53] *IEEE Standard for Synchrophasor Data Transfer for Power Systems*, IEEE Standard C37.118.2-2011 (Revision of IEEE Std C37.118-2005), Dec. 2011, doi: 10.1109/IEEESTD.2011.6111222.
[54] [Online]. Available: https://cmte.ieee.org/pes-testfeeders/resources/
[55] [Online]. Available: https://electricgrids.engr.tamu.edu/electric-grid-test-cases/ieee-118-bus-system
[56] [Online]. Available: https://www.wireshark.org

IMTIAJ KHAN (Graduate Student Member, IEEE) received the B.Sc. degree from the Bangladesh University of Engineering and Technology. He is currently pursuing the Ph.D. degree in electrical engineering with Virginia Tech. He was a Summer Intern with ISO New England, in 2021, and a Research Intern with the Mitsubishi Electric Research Laboratory, in 2022. He is currently working on the cybersecurity of PMU data. His primary research interests include the security of smart grids, synchrophasor measurements, and inverter-controlled microgrids.

VIRGILIO CENTENO (Senior Member, IEEE) received the M.S. and Ph.D. degrees in electrical engineering from the Virginia Polytechnic Institute and State University (Virginia Tech), Blacksburg, VA, USA, in 1988 and 1995, respectively. From 1991 to 1997, he was a Project Engineer in the development of phasor measurement units with Macrodyne, Inc., Clifton Park, NY, USA. In fall 1997, he joined the faculty of Virginia Tech as a Visiting Professor, where he became an Associate Professor, in 2007. His research interests include wide-area measurement and its applications to power system protection and control.