Smart Grid Security Enhancement by using Belief Propagation

B M Ruhul Amin, *Member, IEEE,* Seyedfoad Taghizadeh, *Member, IEEE,* Sasa Maric, *Member, IEEE,* M J Hossain, *Senior Member, IEEE,* and Robert Abbas *Member, IEEE,*

***Abstract*—While the pace of emerging smart grids is increasing worldwide, the novel cutting edge technologies such as internet of things (IoT) and fifth generation (5G) communication networks, require more advanced cyber secure protections. False data injection attacks (FDIA) is a critical cyber-attack which can cause disrupt operations and subsequently black outs. Cleverly constructed false measurement vectors can circumvent the bad data detector (BDD) unit and mislead the state estimation process by creating stealthy type FDIAs. This paper proposes a novel belief propagation (BP) based algorithm to detect FDIA in smart grids. The proposed BP method in this paper operates via utilizing local sensor measurement data to calculate a local belief and send it as a message signal to the control center. Then the control center determines a final/global belief and compares the result with a predefined threshold value derived from the uncompromised measurement database. As a result, the BP based algorithm is able to detect the stealthy type FDIAs which bypass the BDD in state estimation process. Another novel feature of the proposed BP based algorithm is to detect FDIAs without using any historical cyber-attack data which are sketchy due to security constraints and infinitesimal in occurrence numbers. From the obtained results, it is explicit that proposed technique successfully detects random and stealthy FDIA attacks with relatively higher detection rate than the state-of-the-art machine learning classifiers such as Naive Bayes, Support-Vector Machines (SVM), RandomForest, OneR and AdaBoost. The proposed algorithm is tested on the IEEE 14 bus system by utilizing the load data from New York independent system operator.**

***Index Terms*—Smart Grid Security, False Data Injection Attack (FDIA), Belief Propagation (BP).**

NOMENCLATURE

τ	Bad data detection threshold
τ_b	Attack detection threshold
$\boldsymbol{\theta}$	System state vector
$\tilde{\boldsymbol{M}}$	Transformation matrix
$\boldsymbol{e}$	Gaussian measurement error vector
$\boldsymbol{h}(\theta)$	Function of state variables
$\boldsymbol{J}(\boldsymbol{\theta})$	Objective function
$\boldsymbol{z_t}$	Measurement vector at time t
$\boldsymbol{z}$	Measurement vector
$\boldsymbol{Z}_{d\times m}$	Time series measurement matrix
$\hat{\boldsymbol{\theta}}_{\boldsymbol{a}}$	Estimated system state vector during attack
$\hat{\boldsymbol{\theta}}$	Estimated system state vector
ψ	degree of freedom
$\mathbf{a}$	Attack vector
$\mathbf{H}$	System Jacobian matrix
$\mathbf{R}$	Known measurement co-variance
$\mathbf{r}$	Measurement residue
$\mathbf{W}$	Reciprocal of the measurement error variances
$\mathbf{z}$	Real power measurement matrix
$\mathbf{z_a}$	Compromised measurement vector
θ_k	Phase angle at bus k
θ_m	Phase angle at bus m
fn_i	False negative
fp_i	False positive
N_m	Number of buses connected to bus m
P_{km}	Active power flow from bus k to m
P_k	Active power injection at bus k
tp_i	True positive
v	Power flow measurement error
w	Power injection measurement error
x_{km}	Bus reactance
ϕ_i^G	Global belief
ϕ_i^L	Local belief
ψ_i^G	Global compatibility function
ψ_i^L	Local compatibility function
m_i^F	Final message
$z_{t,i}$	Measurement of i-th meter at time t
5G	Fifth generation
BDD	Bad data detector
BP	Belief propagation
CUMSUM	Cumulative sum
EMS	Energy management system
FDIA	False data injection attack
FNR	False negative rate
FPR	False positive rate
IoT	Internet of things
ISO	Independent system operator
MILP	Mixed integer linear programming
MITM	Man-in-the-middle
NYISO	New York independent system operator
OPF	Optimal power flow
PCA	Principal component analysis
PMU	Phasor measurement unit
ROC	Receiver operating characteristics
RTU	Remote terminal unit
SCADA	Supervisory control and data acquisition
SVM	Support vector decomposition
TNR	True negative rate
TPR	True positive rate
WLS	Weighted least square

B M R. Amin, S. Taghizadeh, S. Maric, and Robert Abbas are with the School of Engineering, Macquarie University, Australia. M. J. Hossain is with the School of Electrical and Data Engineering, University of Technology Sydney, Australia.

I. INTRODUCTION

CYBER-ATTACKS are significantly increasing in power industries around the globe in recent years [1]. Fortifying cyber defence is pivotal to the next generation power industry, well-known as a smart grid. As the electricity grid becomes increasingly distributed and reliant on emerging telecommunication and control technologies (e.g. 5G wireless and narrow band internet of technology (IoT) networks), vulnerabilities to cyber-attacks will increase, and more sophisticated approaches for detection and management will be required. Failing to detect cyber-attacks in such synergistic power systems can disrupt operations, hamper communication and stall the economy for hours e.g., Ukraine blackout 2015 [2], [3]. It is highly necessary to identify types, impacts, and solutions of cyber-attacks to ensure the secure operation of power systems.

A. Background and Related Works

Cyber-attack in power system could be initiated by insiders such as spiteful employees or outsiders such as professional hackers, organized criminals, etc. The attacker can exploit the flaws and vulnerabilities in software and communication protocols to electronically invade the power system operational networks. The proper cyber security measures need to be developed and implemented to substations, control centers, power generating stations, transmission and distribution infrastructures. Cyber-security applications cover all digital meters, protective devices, control center database, monitoring computers, software related to the analysis of the system status and measurement data, generation of the control actions, forecasting, predictions, real-time pricing of the energy management system and so forth. Cyber-attacks could be launched to mislead the state estimation process of the energy management system (EMS) in power systems. The state estimator uses network data, real-time measurements, load and production forecasts as inputs and, based on these information, predicts the most likely state of the network at a given moment. The monitoring and control of power systems fully rely on the successful estimation of systems' states, i.e., voltage magnitudes and angles. In practice, measurement residue based bad data detector (BDD) is employed in the state estimation process to detect anomalies in the measurement data set [4]. However authors in [5] have investigated how a false data injection attack (FDIA) can cause vulnerability in the state estimation process via utilizing the system topology and connectivity information which remains stealthy and bypasses the traditional BDD.

Besides, a significant number of research projects are being carried out to defend the FDIAs in power systems. The defense techniques are mainly categorized as detection based approach and protection based approach. The state estimation can be protected from FDIA by protecting all meter measurements or protecting a strategically selected set of sensor measurements. A greedy algorithm is proposed in [6] to facilitate the placement of secure measurement units e.g., phasor measurement units (PMUs) to defend against FDIAs. To determine the least number of protected measurements, a bi-level mixed integer linear programming (MILP) based model is proposed in [7]. The number of protected sensors will keep increasing to ensure the protection of larger systems. Therefore, a graphical method is proposed in [8] to find the optimized number of meter measurements in order to protect a selected number of state variables that have greater social/economic impacts. Another least budget defense strategy is discussed in [9]. However, all these methods are constraints by either protecting a set of measurements or by compromising the security index with limited budget.

Therefore, detection techniques are more realistic alongside the protection techniques to ensure maximum security of the system. In [10], a strategy is proposed, which rationally shuts down some pre-selected transmission lines in order to make transition in network's topology and subsequently identify the possible stealthy attacks. Although this strategy is claimed to be effective, it arises concerns about the possible transients, reducing life time of switching equipment and undesirable blackouts caused by shutting down some transmission lines. Authors in [11] and [12] considered the FDIA as sparse in nature due to the limited resources of the attacker for compromising a large number of meters and proposed novel detection methods based on the separation of nominal power grid states and anomalies to detect the attacks. In reality, the attacker can inject false data by invading communication channel or gaining access to the supervisory control and data acquisition (SCADA) database and for these attack cases, the attack vector is not bound to be sparse. An adaptive cumulative sum (CUMSUM) method is analyzed in [13], where authors considered a detection technique by directly evaluating the changes of measurement residue distribution during attack. The proposed method is unfeasible for larger networks as the computational complexity increases exponentially with the number of meters. In [14], authors proposed a robust sequential detector based on the generalized likelihood ratio for larger range of load variations and attack strategies. However, the stealthy type FDIAs which can bypass the measurement residue based BDD modules are not considered in the above mentioned method. A deep learning based scheme is adopted in [15] to detect FDIA for the purpose of power theft. Wavelet transform and deep neural networks could also be utilized to capture inconsistency in spatial and temporal correlations during normal and cyber-attack situations [16]. In a previous work [17], authors compared the performance of some state-of-the-art machine learning algorithms to detect and distinguish FDIAs and faults.

The Machine learning algorithm is a good choice to detect anomalies in any data set. However, a large amount of training data are required in machine learning or deep learning based methods to achieve successful detection rate. Though the power system operational data are available, but the historical cyber-attack data is very less. Among those infinitesimal data, most of the historical attack data are unavailable due to security constraints. In addition, attacks vary significantly and newer attacks will be able to bypass the machine learning and deep learning based detectors due to the lack of enough training data.

B. Contribution

A robust belief propagation (BP) based detection method is proposed in this paper to detect false FDIAs in the state estimation process. BP is a message passing algorithm which is often used in Bayesian networks and Markov random fields for performance inference [18]. It is highly effective and has low computational complexity which makes it an ideal solution for applications such as medical diagnosis, image processing [19], and intrusion detection [20], [21]. In [20] and [21], a BP based intrusion detection technique is utilized in cognitive radio network. To the best of authors' knowledge, BP algorithm has never been proposed to detect cyber-anomalies in power systems. In this paper, the standard BP algorithm is modified based on the bi-polarity of the power system measurement data. Appropriate selecting the belief threshold level is defined from the long historical uncompromised data to distinguish the attack measurements from the normal measurements. The below contributions are achieved while using the proposed BP method:

- BP method uses long historical normal operation data instead of using historical cyber-attack data. This provides simplicity in operation and lower computational burden.
- BP method successfully detects stealthy type cyber-attack data which can circumvent the conventional residue based BDDs.
- BP method exhibits higher detection rate than the existing state-of-art machine-learning algorithms.

To validate these contributions, the proposed BP algorithm is implemented on the IEEE 14 bus system by utilizing the real time load data from New York independent system operator. Employing the real-life data of NYISO in the test system will facilitate the engineers to realize the real-life scenarios intensively and prompt security measures before the critical situation occurs in power systems. The operation of the proposed method is evaluated through various case studies and its operation is compared with the state-of-art machine learning algorithms such as Naive Bayes, Support-Vector Machines (SVM), RandomForest, OneR and AdaBoost.

C. Paper Organization

The rest of the paper is organized as follows: The energy management system (EMS) model is presented in Section II. The stealthy FDIA construction is described in Section III. Section IV depicts the formulation of the proposed BP based detection method. Data generation, case studies and results are discussed in Section V. Finally, Section VI concludes the paper.

II. Energy Management System (EMS) Model

Efficient and coordinated control mechanism is a requisite for sustainable and reliable operation of power systems. Ambiguous and erroneous monitoring of the system status and measurements can prompt misleading control actions by the power system EMS. For proper monitoring and control, two types of data set are collected by the EMS control center. One type of data set is the topological and configuration data e.g., transformer settings, line impedance and status data regarding the circuit-breakers and switches. Other type of data is the measurement data at different nodes and branches. Both types of data are collected and analyzed for estimating present states of the system. The substantial control actions such as load forecasting, economic dispatch, contingency analysis, optimal power flow (OPF), etc. are determined based on these estimated states of the system. The smart remote terminal units (RTUs), digital communication channels and software based control center are subject to cyber-attacks. The cyber-vulnerable nodes for FDIA are depicted in Fig. 1.

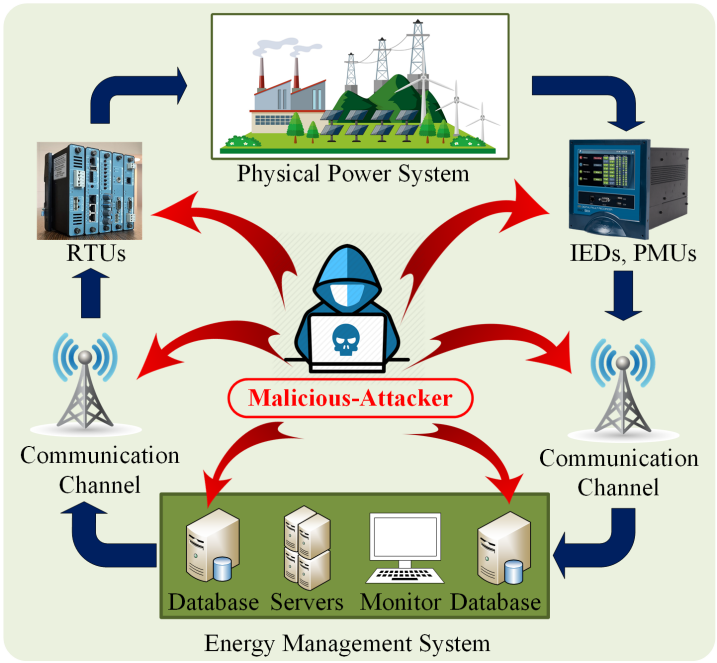


Fig. 1: Cyber-vulnerable nodes in SCADA based EMS subject to malicious data injection attacks

The state estimation process, and the bad data detector functionality are described below for further understanding of the EMS operations.

A. State Estimation

In the state estimation process, the network topology and parameters are considered as perfectly known and usually the voltage phasors (magnitudes and angles) are considered as system states. Full-non linear power flow equations are computationally intensive for an attacker to use and having full access to a significant amount of system information is required. Therefore, a simplified DC approximation model is adopted in this research for the explicit representation of the state estimation process and attack generation processes which could be much easier problem for an attacker. In the DC model, the bus voltages are normalized as equal to 1 and one of the voltage angles is considered as reference. So, the system state vector is represented as:

$$\boldsymbol{\theta} = [\theta_1, \theta_2, ...\theta_{N-1}]^T \tag{1}$$

The measurement vector z received from RTUs can be expressed as:

$$z = \boldsymbol{h}(\theta) + \boldsymbol{e} \tag{2}$$

where, $\boldsymbol{h}(\theta) = [h_1(\theta), h_2(\theta), ..., h_m(\theta)]^T$ is the function of state variables and $\boldsymbol{e} = [e_1, e_2, ..., e_m]$ is the Gaussian measurement error vector with known co-variance $\mathbf{R}$. The real power flow in DC state estimation process is obtained by neglecting all shunt elements and branch resistances.

The power flow measurements can be obtained by performing DC power flow analysis and expressed as:

$$P_{km} = \frac{\theta_k - \theta_m}{x_{km}} + v \tag{3}$$

where, θ_k and θ_m are phase angles at bus k and m, x_{km} is the branch reactance, and v is the measurement error. Similarly, power injection measurement at bus k can be expressed as:

$$P_k = \sum_{m \in N_m} P_{km} + w \tag{4}$$

where, N_m is the number of buses connected to bus m. Hence, the DC model for the real power measurement matrix can be expressed as:

$$\mathbf{z} = \mathbf{H}\boldsymbol{\theta} + \mathbf{e} \tag{5}$$

where, $\mathbf{z}$ includes only the real power flow and real power injection measurements. The system Jacobian matrix $\mathbf{H}$ is a function of branch reactances only.

As $\mathbf{H} \in \mathbb{R}^{m \times n}$ is a full rank matrix and further assuming $m \geq n$, the rank of $\mathbf{H} = n$. So, the solution for $\boldsymbol{\theta}$ can be formulated by using weighted least square (WLS) estimator and expressed as:

$$\underset{\boldsymbol{\theta}}{\mathbf{argmin}}\, J(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\left|\left|\mathbf{z} - \mathbf{H}\boldsymbol{\theta}\right|\right|^2 \tag{6}$$

and the solution for $\hat{\boldsymbol{\theta}}$ is given as:

$$\hat{\boldsymbol{\theta}} = (\mathbf{H^T W H})^{-1}\mathbf{H^T W z} \tag{7}$$

where, $\mathbf{W} = \mathbf{R}^{-1}$ is the reciprocal of the measurement error variances.

Now, the measurement residual can be calculated as;

$$\mathbf{r} = \mathbf{z} - \mathbf{H}\hat{\boldsymbol{\theta}} \tag{8}$$

This measurement residue is used as a key parameter to detect incorrect data in the BDD module.

B. Bad Data Detection

Data can be lost or modified through the communication channel. In the WLS based state estimation, the chi-square (χ^2) test is a common practice to detect bad data. As the estimated measurements are required for residue calculation, the chi-square test is only possible after estimating the system states. It is assumed that the noise samples in the communication channel are independent and follow a normal distribution with zero mean. Therefore, the objective function $\boldsymbol{J(\theta)}$ will follow the chi-square distribution with $\psi = (m-n)$ degree of freedom. By utilizing the Chi-square distribution table, a detection threshold $\chi^2_{(m-n),p}$ can be achieved considering a detection confidence with probability p

(e.g. 95%). If the objective function, $J(\boldsymbol{\theta}) \geq \chi^2_{(m-n),p}$, then bad data is suspected in the measurements, otherwise the measurements are free from bad data and the estimated states are considered for control actions.

III. Stealthy FDI Attack Construction

In most cases, the traditional BDD detects random data injections in the measurement data set and the state estimation process does not converge. However, if attackers are successful in collecting full or partial system data they can inject malicious data in such a way that the state estimation process converges and the malicious or compromised data circumvent the EMS's BDD [5], [22], [23], [24], [25]. The FDIA is also possible without the prior knowledge of the system topology and transmission line impedance by using measurement signals only [26], [27], [28]. The detail models of both attack types are discussed in following subsections.

A. Knowledge based FDI attack model

A man-in-the-middle (MITM) type FDIA is considered in this research work. The MITM attacker mindset is to construct such an attack vector $\mathbf{a}$ that the compromised measurement vector $\mathbf{z}_a = \mathbf{z} + \mathbf{a}$ circumvent the BDD module and mislead the state estimation process to generate false states of the system. For instance, a stealthy attack vector $\mathbf{a}$ is constructed based on the known system Jacobian Matrix $\mathbf{H}$ [5]. For the compromised measurement vector $\mathbf{z}_a$, the estimated states vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{a}}$ can be computed as:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\boldsymbol{a}} &= (\mathbf{H^T W H})^{-1}\mathbf{H^T W z}_{\boldsymbol{a}} \\ &= (\mathbf{H^T W H})^{-1}\mathbf{H^T W(z+a)} \\ &= \hat{\boldsymbol{\theta}} + (\mathbf{H^T W H})^{-1}\mathbf{H^T W a}\end{aligned} \tag{9}$$

If $\mathbf{a} = \mathbf{Hc}$, the 2-Norm of the measurement residual can be calculated as:

$$\begin{aligned}||\mathbf{z_a} - \mathbf{H}\hat{\boldsymbol{\theta}}_{\mathbf{a}}|| &= ||\mathbf{z} + \mathbf{a} - \mathbf{H}(\hat{\boldsymbol{\theta}} + (\mathbf{H^T W H})^{-1}\mathbf{H^T W a})|| \\ &= ||\mathbf{z} - \mathbf{H}\hat{\boldsymbol{\theta}} + (\mathbf{a} - \mathbf{H}(\mathbf{H^T W H})^{-1}\mathbf{H^T W a})|| \\ &= ||\mathbf{z} - \mathbf{H}\hat{\boldsymbol{\theta}} + (\mathbf{Hc} - \mathbf{Hc})|| \\ &= ||\mathbf{z} - \mathbf{H}\hat{\boldsymbol{\theta}}|| \leq \boldsymbol{\tau}\end{aligned} \tag{10}$$

where, $\boldsymbol{\tau} = \chi^2_{(m-n),p}$ is the BDD's threshold. As the 2-Norm measurement residue in both normal and attack cases are same and below the chi-square threshold, the BDD will be unable to detect the malicious injection in the measurement data.

B. Blind FDI attack model

From the discussion of the previous section it is explicit that the attack vector can be constructed by using the system information such as network topology and transmission line impedances. Usually these information of the system are confidential and periodically updated over time. However, a more practical approach is to use measurement signals to construct the stealthy attack vector.

A principal component analysis (PCA) technique can be adopted to construct stealthy attack vector utilizing the measurement signals only. PCA is a statistical process which utilizes an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. This technique can be utilized to approximate the Jacobian matrix $\mathbf{H}$ from the measurement data in order to construct successful attack vectors. Let us assume, $\boldsymbol{Z}_{d\times m}$ is a time series measurement matrix where each row represents a time instant and each column corresponds to measurement attributes. After a successful application of PCA to the measurement set $\boldsymbol{Z}_{d\times m}$, a transformation matrix $\tilde{\boldsymbol{M}}$ and a vector of principal components $\boldsymbol{x}$ are achieved. The PCA transformation can be expressed as:

$$\tilde{\mathbf{M}}^{\boldsymbol{T}}\mathbf{Z} = \mathbf{x} \tag{11}$$

As from the above system model discussion it is known that the Jacobian $\mathbf{H}$ is an $\boldsymbol{n}$ rank matrix, only $\boldsymbol{n}$ number of eigenvectors could be considered in Jacobian matrix construction. Therefore the measurement matrix can be expressed as:

$$Z \approx \begin{bmatrix} \tilde{M}_{1,1} & \tilde{M}_{1,2} & . & . & . & \tilde{M}_{1,n} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ \tilde{M}_{m,1} & \tilde{M}_{m,2} & . & . & . & \tilde{M}_{m,n} \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ . \\ . \\ . \\ x_n \end{bmatrix} \tag{12}$$

where, the approximated Jacobian matrix is a $m \times n$ amtrix.

Now, the attacker can easily construct attack vector $\boldsymbol{a}_{pca} = \boldsymbol{H}_{pca}\boldsymbol{c}$ from measurement signals, which will circumvent the BDD without detection. For the proof of stealthiness and further details of PCA approximation method readers are referred to the article [27].

IV. Proposed Belief Propagation based Detection Algorithm

Traditionally, BDDs are used to detect and identify malicious data in the power system. However, an attacker can construct attack vectors utilizing the topology of the system or measurement signals which bypass the traditional BDDs [5]. In this Section, a BP framework is described to detect FDIAs in the power system.

In the proposed framework, an iterative approach is considered to calculate the belief in every time instance. A belief can be defined as similarity function between the current measurement and the previous measurement. In order to diagnose and mitigate attacks effectively, beliefs are calculated in two levels, locally and globally. A local belief is calculated for each measurement sensor by utilizing the current measurement and historical measurement and a compatibility function. The purpose of the compatibility function is to generate local message for the control center. A global belief is calculated by considering all the local messages containing their beliefs and corresponds to the probability whether the measurement vector is under attack or not. A global compatibility function is also utilized to maintain the global belief between defined levels. The global belief is compared with a threshold value generated from the long historical uncompromised operational data. If the final belief is greater than the threshold, the state estimation process continues. Otherwise, the attack is detected and the compromised measurement vector is discarded from the estimation process. The detection procedure is repeated for the next measurement vector. If the attack vectors are continuously coming for longer time period, the system will go to safe mode operation by using default or forecast data. Continuing state estimation process after attack detection could be an area of future research.

The complete detection process is depicted in the Fig. 2 and the formulation of the local and the global belief are described in the following sections:

A. Local Belief

In the EMS, states of the system are obtained and observed by the independent system operator (ISO) at a regular time interval. For the DC state estimation technique used in this paper, the power injections in buses and the power flows through the branches are considered as measurement data. The measurement vector at time t is expressed as:

$$\boldsymbol{z}_t = [z_{t,1}, z_{t,2}, ..., z_{t,i}, ..., z_{t,N}]^T \tag{13}$$

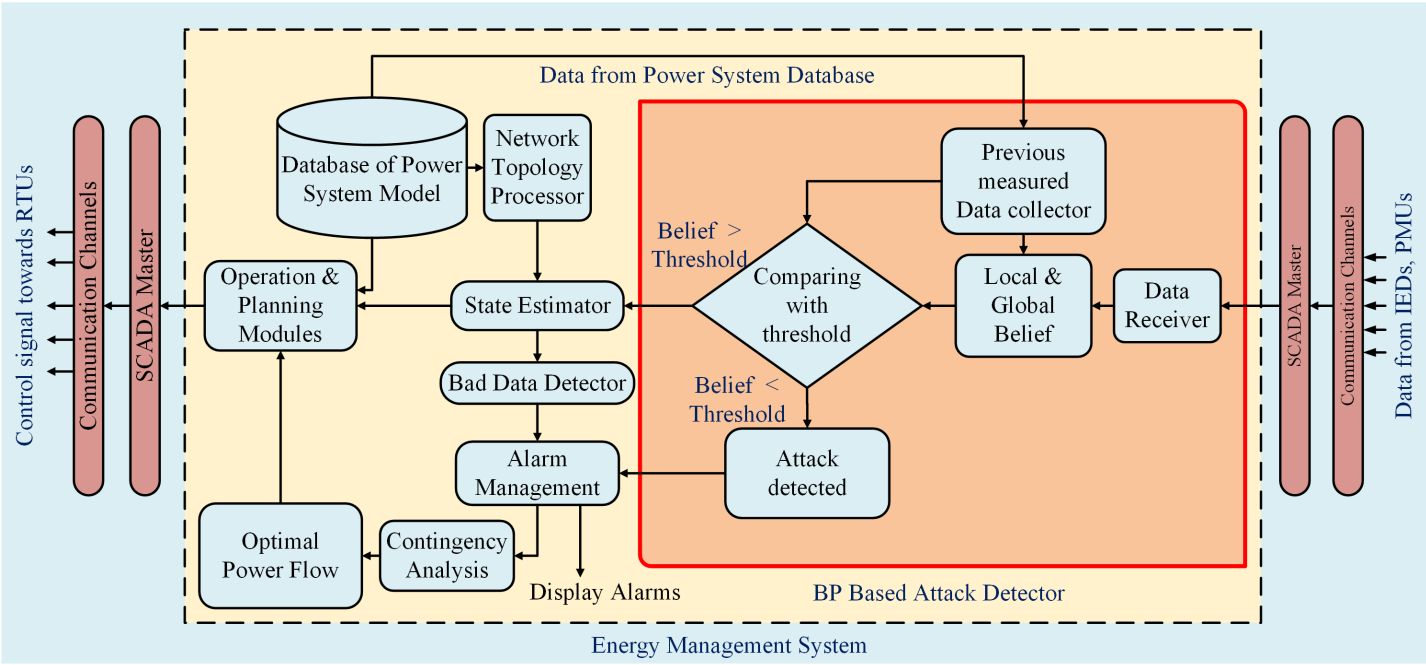


Fig. 2: Proposed EMS with BP detection algorithm

where, N is the total number of measurements. and $z_{t,i}$ is the measurement of i-th meter at time t.

Now, a local function ϕ_i is defined which corresponds to the local belief at meter i. In this proposed modified BP algorithm, an iterative approach is utilized to detect FDIAs in the measurement data set. For the first iteration $(l = 0)$, there is no historical information. Therefore, an uncompromised initial measurement, $z_{0,i}$ is considered to calculate the first local function. The first iteration starts from the measurement $z_{1,i}$, and the local function for i-th meter for the iteration $(l = 1)$ is formulated as:

$$\phi_i^L(l) = \left(1 - \frac{|a-b|}{|b-p|}\right), \tag{14}$$

where, $a, b, p \in \mathbb{R}$ and $p \ll a < b$. It is denoted that for $z_{t,i} < z_{(t+1),i}$, a = $z_{t,i}$ and b = $z_{(t+1),i}$; for $z_{(t+1),i} < z_{t,i}$, a = $z_{(t+1),i}$ and b = $z_{t,i}$. The constant p, is utilized to avoid false local belief calculation during polarity change of measurement data. This local function is an indicator of the difference between the current and previous measurement.

A compatibility function is derived to process the local belief for the local message generation purpose. The compatibility function is defined as:

$$\psi_i^L(l, l-1) = \begin{cases} \frac{1}{\phi_i^L(l)} & \text{if } a = b \\ 0 & \text{if } \phi_i^L(l) < 0 \\ 1 & \text{otherwise} \end{cases} \tag{15}$$

where, the parameters have the same meaning as described in the local belief calculation in (14).

B. Global Belief

The local belief does not represent the cyber situation of the overall system and the local function might give erroneous results for unexpected shutdown of large loads. A precise detection for the overall system can be achieved by using an upper-level detection method while using the local belief. Therefore, a global function called global belief is defined in this paper to detect FDIAs for the overall system. However, the overall peak demand of the system does not change abruptly between two consecutive time intervals and the demand variations reduced for shorter time intervals. Besides, the state estimation is performed only on the steady-state condition of the power system. Therefore, situations like faults are not considered in this study.

The global belief is calculated in a similar way to the local belief by using present and prior information and compatibility function. The information of the local function is treated as a message. Similar to the conventional BP method, all messages from all local functions are sent to the SCADA master to compile and calculate the final global message. The message containing local information of a sensor i is defined as:

$$m_i^L(l) = \phi_i^L(l)\psi_i^L(l, l-1) \tag{16}$$

where, $m_i^L(l)$ is a message from sensor i in the l-th iteration. After receiving all the messages from the sensors, a final message is calculated as:

$$m_i^F(l) = \frac{1}{N}\sum_{i=1}^{N} m_i^L(l) \tag{17}$$

The final message of iteration $l-1$ and l are utilized to generate the global belief. The formulation is similar to (14) and is expressed as:

$$\phi_i^G(l) = \left(1 - \frac{|a-b|}{|b-p|}\right), \tag{18}$$

where, $a, b, p \in \mathbb{R}$ and $p \ll a < b$ and when, $m_i^F(l-1) > m_i^F(l)$, a = $m_i^F(l)$ and b = $m_i^F(l-1)$; when, $m_i^F(l) > m_i^F(l-1)$, a = $m_i^F(l-1)$ and b = $m_i^F(l)$.

In similar to (15), the final compatibility function for iteration (l) is expressed as:

$$\psi_i^G(l, l-1) = \begin{cases} \frac{1}{\phi_i^G(l)} & \text{if } a = b \\ 0 & \text{if } \phi_i^G(l) < 0 \\ 1 & \text{otherwise} \end{cases} \tag{19}$$

where, the parameters have the same meaning as described in the global belief calculation in (18). Final belief is determined by multiplying the global belief with the compatibility function derived in (19). The final belief can be expressed as:

$$\phi_i^F(l) = \psi_i^G(l, l-1)\phi_i^G(l) \tag{20}$$

Where, $\psi_i^G(l, l-1)$ is the global belief and $\phi_i^F(l)$ is the global belief compatibility function.

An attack detection threshold τ_b can be obtained by observing final beliefs of a long historical data. For example, one year historical normal operational data represents all possible power flows and injections variations during days, months and seasons.

Now, if the global belief of the present measurement vector is below the threshold τ_b, attack is detected and the security alarm activated. Otherwise, the system is attack-free and the measurement is ready for the state estimation process. The complete process of the modified BP based detection technique is shown in Fig. 3 and summarized in Algorithm 1.

Algorithm 1 Defence Against FDI Attacks

1: Initialize $l = 0$
2: Initial power measurement obtained, z_0
3: At $l \geq 1$,
4: Obtain Measurement z_1
Calculate ϕ_i^L using Eq. 14
Calculate ψ_i^L using Eq. 15
5: Compile local messages using Eq. 16
Send local messages to the control center
6: Calculate final message $m_i^F(l)$ using Eq. 17
7: Calculate Global Belief ϕ_i^G using Eq. 18
8: Obtain Final Belief $\phi_i^F(l)$ by Eq. 19 and Eq. 20
9: if $\phi_i^F(l) < \tau_b$
Attack Detected
Buzz the alarm & Stop
else
Measurement data is attack free
Go for state estimation
$l = l + 1$
end

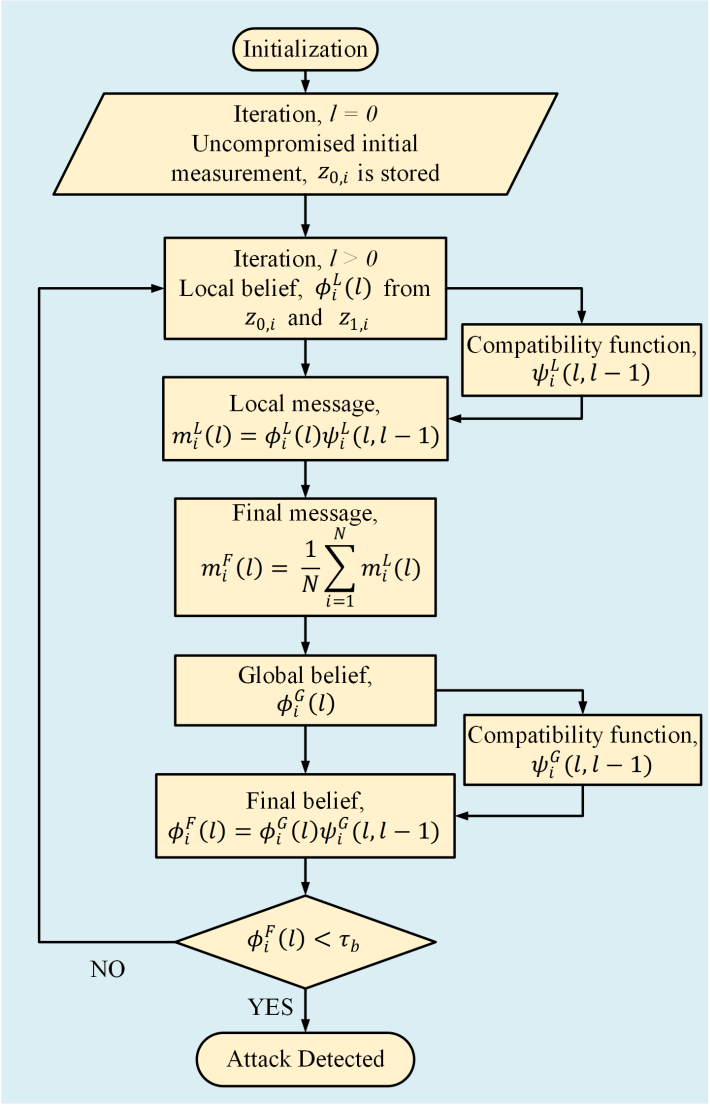


Fig. 3: Proposed BP detection algorithm flowchart

V. Result and Discussion

In this research work, the IEEE 14 bus test system is adopted to generate the measurement vector and evaluate the performance of the proposed BP based detection algorithm. There are 14 buses (nodes) and 20 branches (line sections) in the test model shown in Fig. 4. Total 54 measurements are obtained from 20 incoming and 20 outgoing power flow sensors and 14 power injection sensors respectively. The real time load data from the New York independent system operator (NYISO) are adopted to simulate the power system behavior in a more realistic manner [29]. The measurement samples are taken in every 5 minutes interval. A MATLAB tool MATPOWER is utilized to simulate the test model and analyze the performance of detection techniques [30].

A. False Data Injection Attack Construction

The BDD module in the power system EMS uses chi-square test to detect anomalies or bad data in the measurement data set. The IEEE 14 bus system has 54 measurements and 13 states, thus the degree of freedom is 41. Considering 95% confidence level and above mentioned degree of freedom, the threshold value of the BDD module is 56.94 [4]. Measurements having residual value less than the chi-square threshold are considered as normal data. Otherwise, measurements are considered as bad data and the state estimation process does not converge. For the proposed IEEE 14 bus test system, residual value of the original measurement data set is 4.45 and for measurement with randomly generated zero mean Gaussian noise is 5.85, which are quite below than

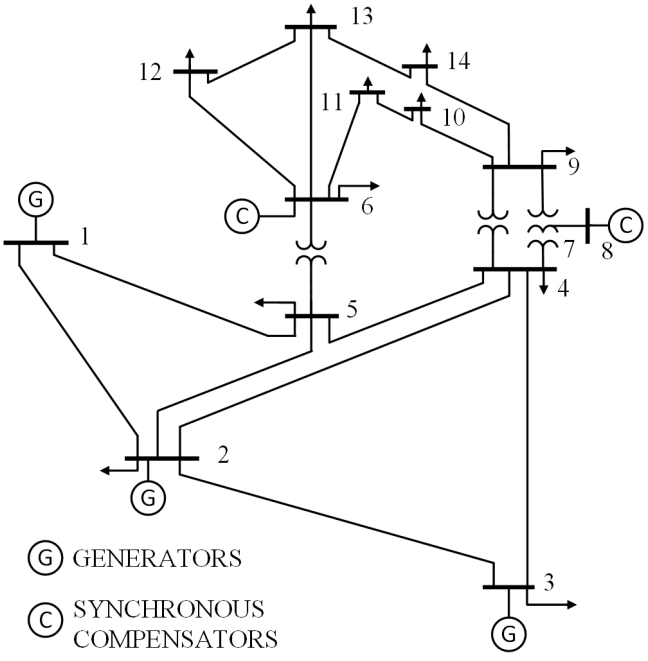


Fig. 4: The IEEE 14 Bus Test System

the threshold value. If random attack vectors are injected to the original measurements then the residual value is above the

threshold value, hence detected by the BDD module and the estimated states will not follow the original states. For evaluating the detection performance of the proposed algorithms, stealthy attack vectors are constructed by utilizing the known system topology matrix [5]. As shown in Fig. 5, the attack measurements are completely different than the original measurements (Fig. 5a) and the estimated states do not follow the original states (Fig. 5b). However, the estimated measurements are following the original measurements and the residue value during attack is similar as the residue value during normal operation. As a consequence, the attack measurement vector remain hidden in the traditional chi-square test based BDD module. In conclusion, it is explicit that stealthy attacks are possible to construct which successfully circumvent the traditional BDD module.

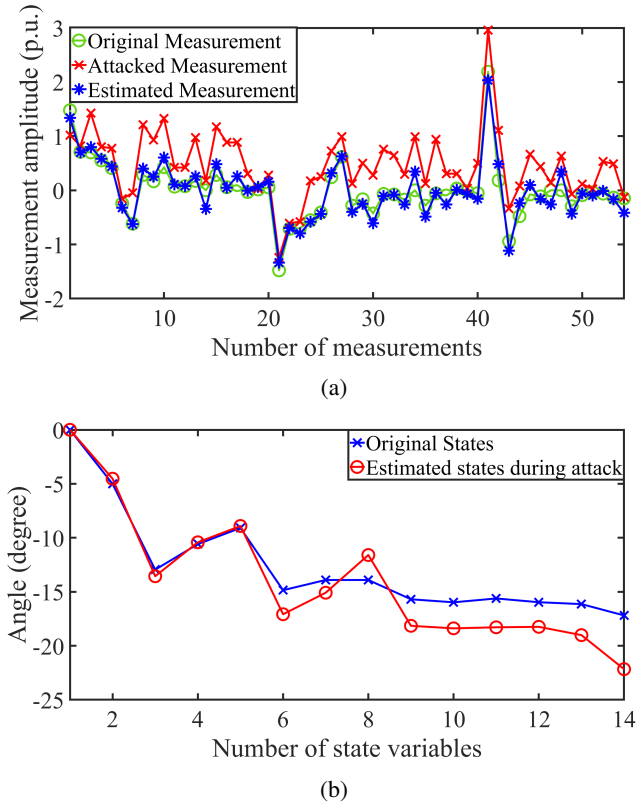


Fig. 5: (a) original measurements, received attacked measurements and estimated measurements during attack and, (b) original states and estimated states during attack.

B. Belief Calculation and Threshold Determination

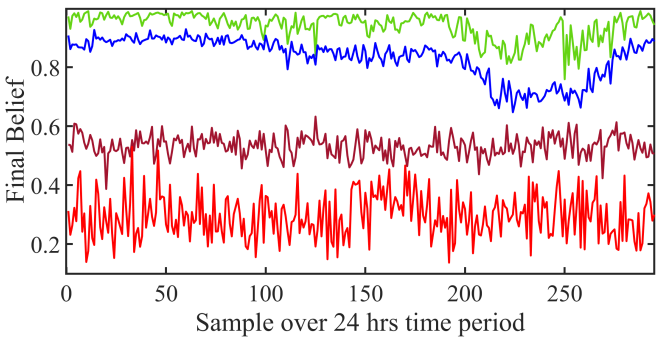


Fig. 6: BP over 24 hours time period for no noise (green), with noise (blue) and attack scenarios (stealthy attack - maroon, random attack - red) by using BP algorithm

Beliefs are calculated for no noise, with noise and attack scenarios by using the proposed BP algorithm. No noise scenario means the original measurements and with noise scenario means addition of Gaussian noise with original signals due to the use of communication channel for data transmission. The calculation is performed for 24 hours time period of a randomly selected regular day. Total 290 instances of 5 minutes interval are considered for the measurements signals. Local beliefs are calculated for each of the measurement nodes and global belief is determined from the calculated local beliefs. It is explicit from Fig. 6 that most of the final beliefs of original measurement (no noise) sets are closer to 1 and few beliefs are below 0.9 due to the heavy load fluctuations during peak hours. While noise is added to the original measurement signals, a decrease in the belief levels are noticed as shown in Fig. 6. However, most of the beliefs of original and noisy signals are above 0.8, whereas most of the beliefs of random and stealthy attack data are below 0.6. Therefore, it is clear from Fig. 6 that the beliefs of the original signal, signal with noise, random attack signal and stealthy attack signal fall in distinguishable distribution levels.

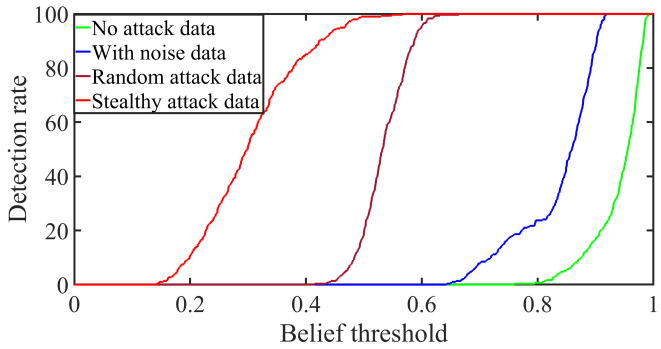
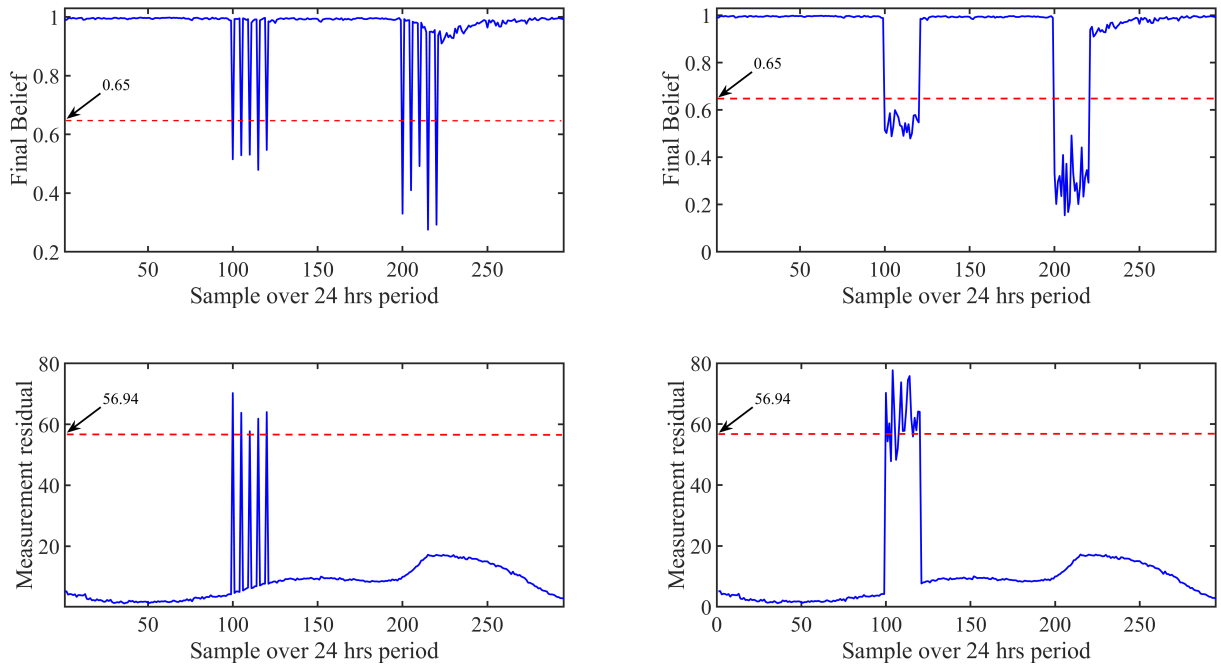


Fig. 7: Detection rate variation with the change of threshold value

The detection rate variation of original data, noisy data, random attack data and stealthy attack data are plotted in Fig. 7. As shown, the detection rate of the original signal is above 0.8. Therefore, all the normal data will be detected by the BP algorithm, if the threshold value is selected between 0.8 and 1. The detection rate of noisy data is above 0.65, hence the threshold value of beyond 0.65 results in detecting the noisy data. Considering the detection level of the normal and noisy data, the threshold value should be selected between 0.6 to 0.65 to make sure that the stealthy and random attack data, which are located below 0.6, are detected while normal and noisy data are ignored.

The accuracy (correctly and incorrectly detected normal and cyber attack data) of the BP algorithm versus changing the detection threshold value is analyzed and the results are summarised in Table I. In the table, the true positive rate (TPR) represents the rate of correctly detected attack which is also called the sensitivity of the classification test. Similarly, the true negative rate (TNR) means the rate of correctly detected normal data which is also called the specificity of the classification test. In addition, the false negative rate (FNR) means that the rate of falsely detected attack and the false positive rate (FPR) means the rate of falsely detected normal data. From Table I, the TPR increases as the threshold value increases and TNR decreases with the increase of the threshold levels. Similarly, FPR and FNR exhibit opposite trends while increasing the the threshold value. According to such a trade off between TPR and TNR and also FPR and FNR, the optimal value of 0.65 is determined via using numerical optimization algorithm, thus at the optimal belief threshold (0.65), the correctly attack detection rate is 99.94% and correctly original measurement detection rate is 99.98%, while FPR and FNR are



(a) Random and stealthy multiple single-attack

(b) Random and stealthy burst attack

Fig. 8: Cyber-attack detection for different case scenarios

located near to zero.

TABLE I: Detection rate of the original measurements and attack measurements for different belief threshold levels

Belief Threshold	True Positive Rate (%)	False Negative Rate (%)	True Negative Rate (%)	False Positive Rate (%)
0.50	58.18	41.82	100	00
0.55	83.08	16.92	100	0.00
0.60	98.05	01.95	100	0.00
0.65	99.94	0.06	99.98	0.00
0.70	100	0.00	99.85	0.15
0.75	100	0.00	99.28	0.72
0.80	100	0.00	97.33	2.67
0.90	100	0.00	76.71	23.29

C. Performance Evaluation of the Proposed Algorithm in detecting Random and Stealthy Attacks

The detection of falsely injected data are conducted for two different case studies as: single instance data set is corrupted for multiple time instances and multiple consecutive instance data are corrupted for consecutive time instances. For the first case, random false data are injected at sample no. 100, 105, 110, 115 and 120 and stealthy false data are injected at sample no. 200, 205, 210, 215 and 220. From the Fig. 8a, the final belief during both of the attacks (random and stealthy) is below the threshold 0.65, hence detected by the proposed BP algorithm. On the other hand, although the measurement residue during the random attack is above the chi-square threshold of 56.94 and is successfully detected by the BDD, during the stealthy attack from 200 to 220, the measurement residual is below the threshold and bypasses the BDD. This result can validate the ability of the proposed BP algorithm in detecting the random and stealthy attacks. Similar results are observed for multiple consecutive attacks and illustrated in the Fig. 8b.

D. Performance Evaluation of the Proposed Algorithm Compared with the Machine Learning Algorithms

In this section, performance of the proposed BP based detection algorithm is discussed in contrast to some of the state-of-the-art machine learning algorithm classifiers. Machine learning algorithms develop a mathematical model based on the supplied training data set and then make predictions on supplied data. The most efficient learner detects most of the supplied instance data correctly. The performance of the classifier depends on the learning method of the classifier and the quality of the training data. In this paper, five distinct types of popular classifiers are used to distinguish attack from the normal power system data. The classifiers and their categories are as follow:

- Probabilistic classification (Naive Bayes) [31]
- Non-probabilistic binary classification (SVM) [32]
- Decision tree learning (Random Forest)[33]
- Rule induction (OneR) [34]
- Boosting, a meta-algorithm for learning (AdaBoost) [35]

Total 424674 measurement instances are obtained from one year data flow where 212338 measurement instances are normal operational data (no noise and with noise) and 212336 are random and stealthy attack data. A machine learning tool WEKA is utilized to perform the case studies [36]. Three case scenarios are studied to clearly realize the effectiveness of the proposed algorithm.

As already mentioned, machine-learning based algorithms require to be trained by both normal operational data and historical attack data to detect attacks with the highest detection accuracy, whereas the proposed BP algorithm requires only normal operational data. To validate this, in case 1, only normal operational data (both original and noisy) are used on purpose to train the machine-learning classifiers and their detection accuracy rates are summarised in Table II. As shown, the machine-learning based algorithms exhibit low detection rate accuracy. Among them, SVM with (78.9923 %) shows the highest detection rate accuracy, whereas the proposed BP algorithm exhibits the maximum detection accuracy rate of 99.94 % which is because of its advanced algorithm that works based on normal operational data and does not require any historical attack data set.

A receiver operating characteristic (ROC) is plotted by using the true positive rate and false positive rate of the classifiers and the results are illustrated in Fig. 9. The results of all the above mentioned classifiers and proposed BP algorithm show that the proposed BP algorithm has the highest rate of correctly detected

TABLE II: Correctly classified instances (%) by using different machine learning algorithms

Cases/Machine Learning Algorithms	Naive Bayes	SVM	Random Forests	OneR	AdaBoost	Proposed BP
Case 1 (Learned from the historical normal operational data)	47.4139	78.9923	50.2768	69.9418	49.6201	99.94 (Using belief threshold)
Case 2 (Learned from normal operational data and historical attack data with stealthy attack)	25.0002	73.0324	75.0003	73.6906	72.8857	
Case 3 (Leaned from normal operational data and historical attack without stealthy attack)	97.4063	99.0082	99.9993	97.7734	49.8189	

attack among all the machine-learning based classifiers.

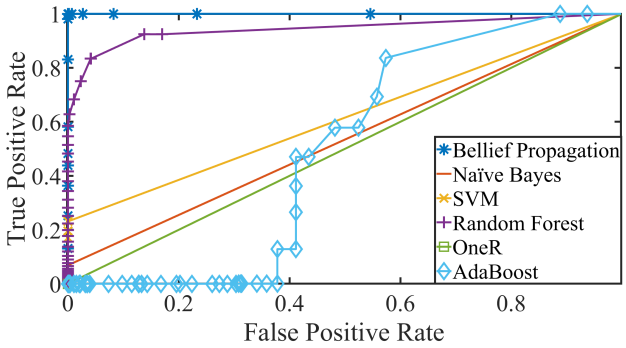


Fig. 9: ROC curve of proposed BP and different Machine Learning algorithms

In case 2, classifiers are trained with both normal operational data and historical cyber-attack data. In this test, it is considered that there are some stealthy attack data in the system, which are bypassed and not recognised by the BDD. As a result, the classifiers are not able to detect the stealthy data and even some of the normal operation data. The percentage of correctly classified instances (detection accuracy rates) are shown in Table II and as can be seen, none of the models achieve more than 75% detection rate accuracy. In the case 3, both cyber-attack data and normal operational data are used to train the classifiers, while there is no stealthy attack data in the system. At this case, which is not always realistic (no stealthy attack and availability of historical attack data), the machine-learning based classifiers can achieve above 95% detection rate which is similar to the proposed BP algorithm (Table I & II). As a consequence, the results of the preformed cases show that the proposed BP algorithm exhibits the highest rate accuracy with the least available data (only normal operational data) and with the existence of stealthy attacks in the system.

In order to provide a clearer picture of the classifiers' performance, three parameters: Precision, Recall and F-measure are calculated for the proposed BP algorithm and the state-of-art machine learning algorithms (Random Forest, Naive Bayes, OneR, Adaboost, and SVM). The Precision parameter measures the positive predictive value, and can be obtained from the confusion matrix of a classifier by using the following equation

$$Precision = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l} \tag{21}$$

where, l is the number of classes, $i = 1, 2, ..., n$, tp_i is the number of true positives and fp_i is the number of false positives.
The Recall parameter measures the true positive rate, and can be calculated by

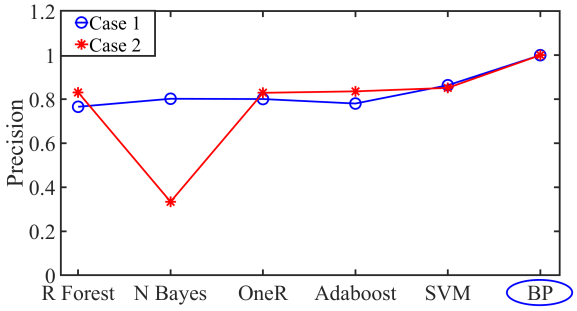
$$Recall = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fn_i}}{l} \tag{22}$$

where, fn_i is the false negative number and other parameters has the same meaning as Eq. (21).

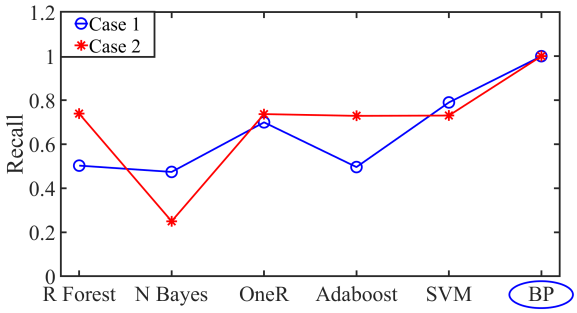
The F-Measure parameter is the harmonic mean of precision and recall. F-Measure is calculated by using the following equation

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{23}$$

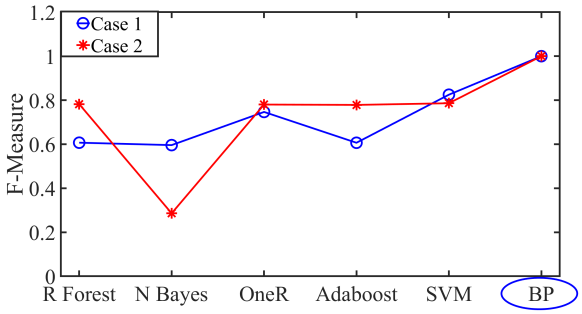
These three parameters (Precision, Recall and F-Measure) are variable according to the robustness of a classifier. Approaching 1.0 indicates the highest robustness of the classifier.



(a) Average precision over classification schemes



(b) Average recall over classification schemes



(c) Average F-Measure over classification schemes

Fig. 10: Precision, Recall and F-Measure over classification schemes

The performances of the classifiers are evaluated for the first two cases (case 1 and case 2) as mentioned in Table II and the results are depicted in Fig. 10. Case 3 is intentionally ignored due to the need for historical stealthy attack data which are realistically unavailable for classifier training purposes. Fig. 10a shows that the value of Precision of all the classifiers for case 1 and case 2 are below 0.9 showing the lack of robustness of the conventional methods. In contrast, the value of Precision is almost 1.0 for both

the cases while using the proposed BP detection algorithm. The same superior performance of the proposed BP is also clearly visible in Fig. 10b and Fig. 10c for the parameters Recall and F-measure.

VI. Conclusion

Stealthy attack data can cause negative consequences to the control decisions in a power system. In this paper, a BP based detection technique is adopted to detect stealthy type FDIAs which can bypass the traditional BDD module. Unlike other training based detection algorithm, the BP method works without any historical attack data. The proposed BP method is implemented on the IEEE 14 bus system by using the real time load data of NYISO and its performance is analyzed through different case scenarios. Key findings of the performed analysis are summarized as follows:

- The proposed BP algorithm successfully detects random and stealthy attacks during two case scenarios of corrupted single set data and multiple consecutive instance data.
- Using just normal operational data and during the existence of stealthy attacks in the system, the proposed BP algorithm achieves 99.94% attack detection rate accuracy which is significantly higher than the detection rate accuracy of state-of-the-art machine learning classifiers such as Naive Bayes, SVM, RandomForest, OneR and AdaBoost.
- The higher detection rate (robustness) of the proposed BP method than the existing state-of-art machine learning classifiers is validated via analyzing three parameters Precision, Recall and F-Measure.

Consequently, the simplicity of the proposed BP algorithm, its lower computational burden and subsequently higher detection rate accuracy, make it a robust and effective alternative for increasing the level of security and protection in smart grids. The future scope of this research is to extend the proposed method in distinguishing cyber-attacks from the common credible contingencies such as normal faults.

Acknowledgement

Author acknowledge the NYISO for sharing their data, and NSW Cyber Security Network for financial support.

References

[1] The industrial control systems cyber emergency response team: Incident reported by sector (fy 2016) and onsite assessments by sector (fy 2014-2016). [Online]. Available: https://ics-cert.us-cert.gov/Year-Review-2016

[2] G. Liang, S. R. Weller, J. Zhao, F. Luo, and Z. Y. Dong, "The 2015 ukraine blackout: Implications for false data injection attacks," *IEEE Trans. on Power Systems*, vol. 32, no. 4, pp. 3317–3318, July 2017.

[3] M. Zeraati, Z. Aref, and M. A. Latify, "Vulnerability analysis of power systems under physical deliberate attacks considering geographic-cyber interdependence of the power system and communication network," *IEEE Systems Journal*, vol. 12, no. 4, pp. 3181–3190, 2017.

[4] A. Abur and A. G. Expósito, *Power System State Estimation: Theory and Implementation*. Boca Raton, FL, USA:CRC, 2004.

[5] Y. Liu, P. Ning, and M. K. Reiter, "False Data Injection Attacks against State Estimation in Electric Power Grids," *ACM Trans. on Information and System Security*, vol. 14, no. 1, pp. 1–33, 2011.

[6] T. T. Kim and H. V. Poor, "Strategic protection against data injection attacks on power grids," *IEEE Trans. on Smart Grid*, vol. 2, no. 2, pp. 326–333, June 2011.

[7] X. Liu, Z. Li, and Z. Li, "Optimal protection strategy against false data injection attacks in power systems," *IEEE Trans. on Smart Grid*, vol. 8, no. 4, pp. 1802–1810, July 2017.

[8] S. Bi and Y. J. Zhang, "Graphical methods for defense against false-data injection attacks on power system state estimation," *IEEE Trans. on Smart Grid*, vol. 5, no. 3, pp. 1216–1227, May 2014.

[9] R. Deng, G. Xiao, and R. Lu, "Defending against false data injection attacks on power system state estimation," *IEEE Trans. on Industrial Informatics*, vol. 13, no. 1, pp. 198–207, Feb 2017.

[10] S. Wang, W. Ren, and U. M. Al-Saggaf, "Effects of switching network topologies on stealthy false data injection attacks against state estimation in power networks," *IEEE Systems Journal*, vol. 11, no. 4, pp. 2640–2651, 2015.

[11] L. Liu, M. Esmalifalak, Q. Ding, V. A. Emesih, and Z. Han, "Detecting false data injection attacks on power grid by sparse optimization," *IEEE Trans. on Smart Grid*, vol. 5, no. 2, pp. 612–621, March 2014.

[12] B. Li, T. Ding, C. Huang, J. Zhao, Y. Yang, and Y. Chen, "Detecting false data injection attacks against power system state estimation with fast go-decomposition approach," *IEEE Trans. on Industrial Informatics*, vol. 15, no. 5, pp. 2892–2904, May 2019.

[13] Y. Huang, J. Tang, Y. Cheng, H. Li, K. A. Campbell, and Z. Han, "Real-time detection of false data injection in smart grid networks: An adaptive cusum method and analysis," *IEEE Systems Journal*, vol. 10, no. 2, pp. 532–543, June 2016.

[14] S. Li, Y. Ylmaz, and X. Wang, "Quickest detection of false data injection attack in wide-area smart grids," *IEEE Trans. on Smart Grid*, vol. 6, no. 6, pp. 2725–2735, Nov 2015.

[15] Y. He, G. J. Mendis, and J. Wei, "Real-time detection of false data injection attacks in smart grid: A deep learning-based intelligent mechanism," *IEEE Trans. on Smart Grid*, vol. 8, no. 5, pp. 2505–2516, Sept 2017.

[16] 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. on Industrial Informatics*, vol. 14, no. 7, pp. 3271–3280, July 2018.

[17] B. M. R. Amin, A. Anwar, and M. J. Hossain, "Distinguishing between cyber injection and faults using machine learning algorithms," in *IEEE Region Ten Symposium (Tensymp)*, July 2018, pp. 19–24.

[18] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations." San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003, pp. 239–269.

[19] P. Ramachandra and M. Sartipi, "Compressive sensing based imaging via belief propagation," in *Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Nov 2011, pp. 254–256.

[20] T. Le and C. N. Hadjicostis, "Graphical inference for multiple intrusion detection," *IEEE Trans. on Information Forensics and Security*, vol. 3, no. 3, pp. 370–380, Sep. 2008.

[21] Z. Yuan, D. Niyato, H. Li, J. B. Song, and Z. Han, "Defeating primary user emulation attacks using belief propagation in cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 10, pp. 1850–1860, November 2012.

[22] X. Liu, Z. Bao, D. Lu, and Z. Li, "Modeling of local false data injection attacks with reduced requirement on network information," *IEEE Trans. on Smart Grid*, vol. 6, no. 4, pp. 1686–1696, 2015.

[23] J. Zhang, Z. Chu, L. Sankar, and O. Kosut, "Can attackers with limited information exploit historical data to mount successful false data injection attacks on power systems?" *IEEE Trans. on Power Systems*, vol. 33, no. 5, pp. 4775–4786, Sep. 2018.

[24] R. Deng and H. Liang, "False data injection attacks with limited susceptance information and new countermeasures in smart grid," *IEEE Trans. on Industrial Informatics*, vol. 15, no. 3, pp. 1619–1628, 2019.

[25] Y. Li and Y. Wang, "False data injection attacks with incomplete network topology information in smart grid," *IEEE Access*, vol. 7, pp. 3656–3664, 2019.

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

[27] Z. Yu and W. Chin, "Blind false data injection attack using pca approximation method in smart grid," *IEEE Trans. on Smart Grid*, vol. 6, no. 3, pp. 1219–1226, May 2015.

[28] A. Anwar, A. N. Mahmood, and M. Pickering, "Modeling and performance evaluation of stealthy false data injection attacks on smart grid in the presence of corrupted measurements," *Journal of Computer and System Sciences*, vol. 83, no. 1, pp. 58 – 72, 2017.

[29] New York Independent system operator (NYISO) real time actual load data -2018. [Online]. Available: http://mis.nyiso.com/public/P-58Blist.htm

[30] R. D. Zimmerman, C. E. Murillo-Sanchez, and R. J. Thomas, "Matpower: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Trans. on Power Systems*, vol. 26, no. 1, pp. 12–19, Feb 2011.

[31] P. Langley, W. Iba, and, and K. Thompson, "An analysis of bayesian classifiers." AAAI Press, 1992, pp. 223–228.

[32] J. C. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," *Microsoft Research*.

[33] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct 2001.

[34] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, no. 1, pp. 63–90, Apr 1993.

[35] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119 – 139, 1997.

[36] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann, Fourth Edition, 2016.