IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 July 2024, accepted 14 August 2024, date of publication 19 August 2024, date of current version 30 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3445811

RESEARCH ARTICLE

Improving FDI Detection for PMU State Estimation Using Adversarial Interventions and Deep Auto-Encoder

SALEH ALMASABI[1], **(Member, IEEE), ZOHAIB MUSHTAQ**[2], **NABEEL AHMED KHAN**[3], **AND MUHAMMAD IRFAN**[1]

[1]Electrical Engineering Department, Najran University, Najran 61441, Saudi Arabia
[2]Department of Electrical Electronics and Computer Systems, College of Engineering and Technology, University of Sargodha, Sargodha 40100, Pakistan
[3]Department of Electrical Engineering, Riphah International University, Islamabad 44000, Pakistan

Corresponding author: Saleh Almasabi (ssalmasabi@nu.edu.sa)

This work was supported by the Deanship of Scientific Research, Najran University, Saudi Arabia, through the Distinguished Funding Program under Grant NU/DRP/SERC/12/8.

ABSTRACT Concerns have been voiced about the growing significance of cyber-threats, especially in light of the potentially dire repercussions of false data injection (FDI) assaults. This work investigates FDI detection in phasor measurement units (PMU), focusing on instances where an attack can be launched simply by compromising one unit. Simulated post-processing adversarial interventions i.e., noise and non-linearity were introduced to train and fortify the system against possible attacks and to render it resilient to perturbations. By learning complex non-linear patterns from the data, a deep de-noising auto-encoder model is used to de-noise and learn genuine feature representations, improving overall reliability. The suggested framework performs better than conventional machine learning and 1-D CNN models when it comes to precisely estimating intrusion, as shown by a comparison study. By using an integrated strategy, power system monitoring and control become more accurate and resilient, successfully tackling the changing issues faced by contemporary electrical grids. The proposed adversarially robust framework is evaluated using Monte-Carlos simulations and on varying load conditions to better comprehend the impact of adversarial interventions on the FDI detection accuracy under different load characteristics and attack scenarios. The proposed framework yielded an average 98.3% in Monte Carlo simulations and an average of 96.5% accuracy under varying load conditions. Surpassing the conventional ML and 1-D CNN algorithms in successfully identifying FDI attacks under adversarial vulnerability.

INDEX TERMS False data injections, adversarial interventions, phasor measurement units, smart grids, intrusion.

I. INTRODUCTION

With the motorization of smart grids, the operation and control of the grid are becoming more reliant on real-time data, such as voltages and currents. This measurement data has enhanced the operation of the smart grid and made it more efficient and reliant. However, this moderation has also made the smart grade prone to cyber threats and cyber-attacks [1]. Accounting for 10.7% of all cyberattacks in 2022, the energy industry was the fourth most targeted. Energy is the most attacked industry in North America, with 20% of attacks occurring in this sector. 2015 saw a brief interruption in the provision of electricity to customers when hackers using the Black-Energy 3 malware remotely gained access to the information systems of three energy distribution businesses in Ukraine [2]. Manhattan, New York had widespread power disruptions on July 13, 2019, as a result of a cyber-attack on the city's electrical grid. Critical services were at risk and electrical power networks were affected as a result of the attack. Several assaults have recently threatened or affected

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

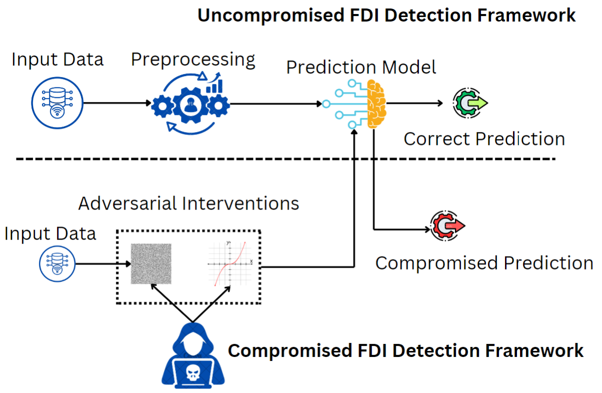

Uncompromised FDI Detection Framework
Input Data
Preprocessing
Prediction Model
Correct Prediction
Adversarial Interventions
Input Data
Compromised Prediction
Compromised FDI Detection Framework

FIGURE 1. Visual depiction of compromised and uncompromised FDI detection frameworks.

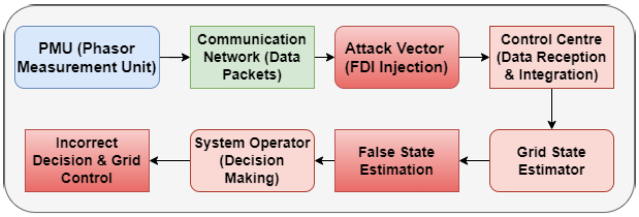

PMU (Phasor Measurement Unit)
Communication Network (Data Packets)
Attack Vector (FDI Injection)
Control Centre (Data Reception & Integration)
Grid State Estimator
False State Estimation
System Operator (Decision Making)
Incorrect Decision & Grid Control

FIGURE 2. A generic block illustration of False Data Injections in PMU based state-estimations.

electrical power systems, of which these destructive ones are only a few examples [2], [3].

One of these cyber threats is false data injection (FDI) attacks where the attacker injects false measurement data that resembles real data and can bypass bad data detection (BDD) [4]. For such attacks to be successful, the network's full information needs to be known. However, the partial knowledge of the network can be enough to launch successful attacks [5]. These attacks if successful, can affect the grid operation in many ways such as power flows [5], [6], frequency stability [7], and economic dispatch [8]. Several approaches have been used to mitigate the risk of FDIs. A phase lock Value-based approach was presented in [9], where FDIs time instances are detected with high accuracy, however, the capability to localize the intrusions and only detect their occurrences. Other unsupervised approaches were used such as Isolation Forset, Local Outlier Factor, and AutoEncoders [10], [11], [12]. These unsupervised approaches rely on anomaly detection and FDIs are identified based on their behavior in comparison with the normal measurement data. However, these approaches are effective for detecting FDIs as anomalies may not be as effective if the FDIs are behaving close to normal operational data.

On the other hand, supervised ML approaches, where the models are trained to detect FDIs and not rely solitary on FDIs' animality behavior. Support vector machine, convolutional neural network, and random forest are presented in [13], [14], and [15]. The ML-supervised-based approaches often focus on detecting FDIs without considering data noise or sensor noise and its effect on the detection accuracy. Also, the power grid requirement is often neglected such as having enough sensors (RTUs or PMUS) to have full observability for the state estimator. The power grid operation relies on monitoring the network through measurement data. These measurements are obtained from either legacy units (RTUs) or Phasor measurement Units (PMUs). The RTUs measure power flow through the grid, and FDI studies that focus on RTU-based attacks usually deal with DC-estimators which are linear [2]. Nonlinear estimators can also be used to estimate the status of AC power networks. Consequently, launching FDIs that avoid the BDD is challenging [11], [16], [17]. To recognize FDI assaults, [18] uses signal processing analysis in conjunction with the wavelet singular entropy (WSE) approach. Guan and Ge [19] investigate FDI and jamming attack detection using wireless sensor networks (WSNs).

PMUs, on the other hand, are more advanced and measure the bus voltages and currents in phasor form which enables more robust and efficient situational awareness [20]. While the PMU network facilitates better information sharing within a power system, it also creates significant cyber vulnerabilities. Numerous cyber-threats are known to target PMU networks [21], [22]. Although, PMU data is well time-stamped it is not totally immune to Direct Data Manipulation, meter tempering, data replay attacks of other data based interventions. In the context of False Data Injection (FDI) attacks the hackers intercept the data between in the communication network. There are a myriad of ways to intercept the PMU data communication, for instance Man-in-the-middle (MITM) attacks can position the attacker between the PMU and control center by means of compromised routing or by exploiting software vulnerabilities, weak wireless and credentials. A generic FDI injection route for PMU is shown in Fig. 2. Similarly, Distributed Denial of Service (DDoS) attacks, for example, have the ability to take PMUs offline and prevent them from transmitting measurements, so disrupting PMU operations. Furthermore, there is a chance that man-in-the-middle attacks will introduce harmful scripts into PMUs, changing their functionality without warning. Furthermore, PMUs' timestamps and measurement data can be altered by data spoofing attacks. Consequently, several studies have looked into how different mobile cyber-attacks are against PMUs. The global positioning system (GPS) technology has been the subject of investigations [23], [24]. Alexopoulos et al. [25] used a vulnerability analysis to start FDI attacks of PMUs on power networks. Chu et al. [26] examine the physical effects of FDI on the N-1 reliable power technology using real-time contingency analysis and a secured power dispatch. Distribution grids are protected against FDI assaults that cause over-voltage by employing a convex optimization method based on second-order cone programming [27]. Ding et al.'s [28] model of PMU positioning as a defensive mechanism against cyber threats used a bi-level technique. A load redistribution (LR) attack design, in which attackers and system operators

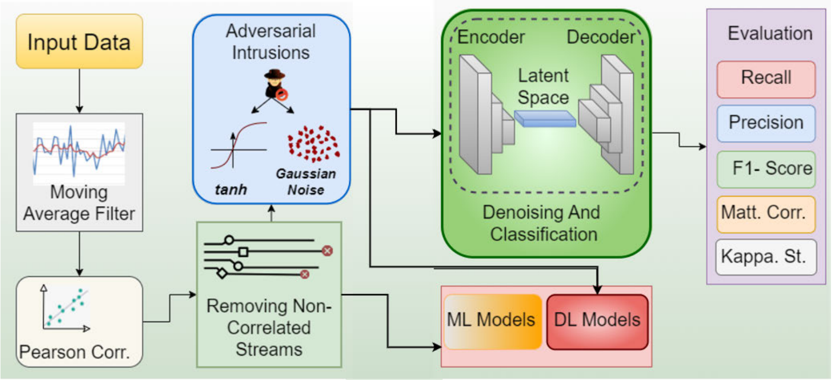


FIGURE 3. Block diagram of proposed methodology.

employ distinct resource allocations, makes use of insider threats to power networks. Liu and Wang [29]. Huang et al. [30] have improved a technique that was previously used to defend against coordinated cyber-physical attacks (CPAs) by reducing the number of PMUs.

As discussed in the literature for PMU based FDI detection, state-of-the art machine learning and deep learning based algorithms are being employed for efficient outcomes. However, with the raging interest in ML and DL frameworks for FDI detection, the susceptibility of these frameworks towards adversarial interventions have increased drastically as shown in Fig. 1. The adversarial interventions are attempts to modify the Machine Learning Model's input, or manipulate the learned parameters to compromise the predictions. Mostly, Machine learning models and deep neural networks are exploited for vulnerabilities which lead to incorrect outcomes. These adversarial interventions although are subtle, however, the effect they have on the model performance is sizeable. Most of these subtle adversarial attempts are made by introducing small perturbations in inference pipeline [31], rendering the decision boundary chaotic with no to little apparent visual cues of their presence. For sensitive applications of ML such as FDI detections in PMU, the ramifications of these adversarial attempts to compromise the model's integrity are dire [28], [32]. In one such study, Berghout et al. [33] enhanced PMU data resilience against dynamic disturbances and adversarial attacks through robust feature engineering, addressing data scarcity and imbalance with synthetic oversampling and adaptive learning. It emphasizes defense against false data injection by simulating adversarial attack scenarios, evaluating model performance under various levels of disturbances and synthetic perturbations, to ensure the reliability of FDI detection in power systems. Similar evaluations of machine learning classifiers is undertaken in a study by Kamal et al. [34] which investigates cyber-threats targeting event classification in micro-PMU measurements, focusing on poisoning attacks against SVM classifiers. It explores the impact of compromised event classification on utility operators and proposes a novel attack detection method to identify changes in SVM decision boundaries caused by poisoning attacks, aiding in the detection and evaluation of the number of poisoned data points in the training dataset. Adversarial attempts are discrete and potent, therefore it is imperative to expose the machine learning pipelines to simulated versions of these attacks or perturbations within the training data –in order to develop resilience and robustness. Following the similar approach, a study by Cheng et al. [35] addresses vulnerabilities of machine learning-based event classifiers to adversarial attacks on PMU data by proposing an adversarial purification method based on a diffusion model. It involves injecting noise into PMU data and utilizing a pre-trained neural network to remove both the added noise and perturbations from adversarial attacks, significantly improving classifier accuracy while maintaining real-time operability. The proposed method decreases the distance between original and compromised PMU data, reducing the impact of adversarial attacks, as validated by empirical results on a large-scale real-world dataset. Although machine learning based FDI detection frameworks for PMU have attained a high accuracy in classifying anomalous activity, these algorithms themselves are not immune to attacks [36]. Hence, it remains imperative to secure these frameworks at the application end also, to harness the true potential of these FDI detection algorithms to mitigate their failures. The works on False data injection for PMU's are getting popularity, especially the adversarial vulnerability of Machine Learning based FDI frameworks. The aforementioned works investigated and proposed novel methods be it feature engineering or synthetic perturbations to increase the resilience of ML models against the attacks. However, the existing works lack thorough analysis of the impact of the adversarial interventions within the feature space, for a refined understanding. Moreover, current methods often overlook the complexities of attacks in terms of non-linearity, which can significantly affect the decision boundary, making it challenging for ML models to accurately predict the FDI attacks. Hence, there is exists a need for further research to thoroughly analyze the impact of adversarial interventions to gain a refined understanding of their effects on the accuracy of FDI frameworks especially in PMU. As despite the transitioning of the smart grids from RTU's to PMU's, the PMU-based FDI's aren't given the same attention as RTU's [20]. FDI attack detection methods are mostly inclining towards ML or DL based frameworks, hence the classification model must itself be capable enough to withstand any interventions for a secure and reliable FDI framework.

In view of the gaps within the existing literature towards the adversarial vulnerability of ML models, this research proposed a strengthened and robust PMU based FDI pipeline using deep auto-encoders to increase the reliability of the False data detection. Not only were the performances of FDI detection models evaluated under non-linear perturbations and compared with our model, but the impact of adversarial interventions on the features were also analyzed via plot based correlations and dimensionality reduction. Moreover, FDI research usually does not consider the power grid operational conditions such as load change, sensor (PMU) locations, and the full observability of the grid. In this work, we investigate FDI attacks for PMUs and consider the operational conditions of the power grid; the main contributions can be summarized as follows:

- A novel adversarially robust deep-denoising auto-encoder based FDI detection framework for PMU is proposed, to identify FDI's location and instances with high-accuracy as given in Fig. 3
- The moving averages and correlations are utilized to identify FDI attacks streams from raw data. Synthetic noise and non-linear perturbations are introduced later to analyze the effects of adversarial intrusions on FDI detection framework and classification accuracy.
- A thorough visual analysis is made to understand the impact of the adversarial intrusions on the features, to better comprehend the behaviour of these disturbances in the feature space.
- The performance of the machine learning based classifiers was evaluated on these added adversarial perturbations and compared with the proposed auto-encoder based FDI classification paradigm under multiple scenarios.
- The proposed framework is tested on IEEE-14, IEEE-30, IEEE -39 and IEEE-118 bus system topologies using Monte - Carlos Simulations and varied load profiles.

The rest of the paper is organized as follows, Section II is based on the Measurement and State Estimation in Power Grids. Section III defines the Methodology and Section III-B is based on Simulating Adversarial Attempts. Results and Discussions are defined in Section IV and Finally the Conclusion is in Section V.

II. MEASUREMENT AND STATE ESTIMATION IN POWER GRIDS

State estimators use the data from either RTUs or PMUs, and then based on the acquired data the state estimation process becomes linear or nonlinear. The RTUs measure the voltage magnitudes, power flows, and power injections. The PMUs on the other hand, measure the voltages of the buses and current flows in phasor form. The measurement model can be described as follows

The state estimation process in power systems involves integrating complex representations of voltage (V) and current (I) measurements, crucial for accurately predicting the system's state [37], [38]. These complex measurements are often expressed as:

$$\begin{aligned} V_{cmplx} &= V_m \angle \theta_v \\ I_{cmplx} &= I_m \angle \theta_i \end{aligned} \tag{1}$$

Here, V_m and I_m represent magnitudes, while θ_v and θ_i denote phase angles.

The following are the basic state estimate equations that have been modified to incorporate these complex representations:

$$x_o = f\left(x_{pr}, t\right) + \omega \tag{2}$$

Here, x_o is the vector of state variables that need to be calculated while f is the state transition function that models the dynamics of a power system. x_{pr} depicts the former conditions, t is the control input vector and ω is noise in the process.

Taking into account the intricate nature of voltage and current data, the core state estimation equations, eq (2), capture the dynamics of the system. Also, the state variables (x) are limited within the bounds established by the previous estimate (x_{pr}). An objective function $J(x)$ is created to minimize the difference between the observed values (z) and the values anticipated by the state estimation $(h(x))$. Through optimization approaches, $E(x)$ is minimized to provide a full and accurate portrayal of the operational state of the power system, as well as a dependable estimate of its state variables.

$$\begin{aligned} E(x) = &\frac{1}{2}(z - h(x))^T L^{-1}(z - h(x)) \\ &+ \frac{1}{2}\left(x - x_{pr}\right)^T M^{-1}\left(x - x_{prev}\right) \end{aligned} \tag{3}$$

where

$E(x)$ is the objective function that is to be minimized.

L^{-1} = The covariance matrix encapsulating the measurement errors.

M^{-1} = The covariance matrix of the state estimation errors.

x_{pr} = The prior state estimate.

A. ATTACK MODEL

This subsection discusses FDI attacks against PMU-based smart grids. The adversaries potentially attempt to detectably fake the measurement and If this fraud is not discovered, the system's status may be incorrectly estimated, which may affect the operators' operational choices and could lead to overloading or tripping transmission lines. On the other hand, it is impossible to arbitrarily fabricate measurements. In literature, techniques like the Chi-square test and the

Largest Normalized Residual (LNR) are used as state estimators to spot manipulated or aberrant measurements. Therefore, to avoid BDD or LNR detection, the adversaries must use grid topology to mask these assaults for them to be successful [13], [39]. Monitoring the data streams (PMU measurements) or creating the attack vector (AV) with the help of the grid information are two strategies to evade detection.

The attack vector (Δa) corresponding to FDI, the original measurement equation is given by

$$a = q(h) + \epsilon \tag{4}$$

In the above equation, a is the output vector of measured instances and $q(h)$ gives a relation between the state variable h to the measurements taken and finally, ϵ is the measurement error. The attack vector (Δa) can be expressed in terms of perturbations added to the original measurements.

$$(\Delta a) = a_f - a \tag{5}$$

a_f = Erroneous measurements vector.
(Δa) = False Data Injection vector.

In order to generate a_f, a perturbation term taken as notation α, is scaled via a matrix (K) representing the grid information and inter-measurement relationship.

$$a_f = q\left(h_f\right) + \epsilon_f \tag{6}$$

In the above equation, $h_f = h + K.\alpha$ and $\epsilon_f = \epsilon + K.\alpha$. By expanding these above equations, we get:

$$a_f = q(h) + \nabla q \cdot (K \cdot \alpha) + \epsilon + K \cdot \alpha \tag{7}$$

where, ∇q is the Jacobian matrix for the measurement function q. Now, the FDI attack vector Δa can finally be expressed as

$$\Delta a = \nabla q \cdot (K \cdot \alpha) + K \cdot \alpha \tag{8}$$

Taking into account how sensitive the measurements are to variations in the state variables, this formulation captures the impact of the attack vector on the measurements. The grid information is captured by the matrix K, while the attacker's strategic choices about how to skew the measurements are represented by α. The attacker must maximize α to covertly modify measurements while taking into account the state estimators' detection measures and the grid layout.

III. METHODOLOGY

A. FILTERING AND CORRELATION

Phasor Measurement Unit (PMU) data must be carefully refined using a moving average $\overline{M}_{raw}$ and binary flags $\overline{A}_{raw}$in order to be ready for machine learning or deep learning-based classification. By doing this, the pipeline is stabilized and noise and fluctuations are reduced:

$$\overline{M}_{raw} = \frac{1}{w}\sum_{n=0}^{w} X_n^{raw} \tag{9}$$

$$\overline{A}_{raw} = \frac{1}{w}\sum_{n=0}^{w} Y_n^{raw} \tag{10}$$

The terms $\overline{M}_{raw}$ and $\overline{A}_{raw}$represent the moving average of the raw PMU measurement data X^{raw} and the binary flags that identify attacked samples Y^{raw}, respectively, in these equations. The temporal scope of influence is determined by the window size (w).

The most associated data streams impacted by False Data Injections (FDIs) are then determined using a Pearson correlation (C_r). The data is then segmented for modeling and validation through supervised learning, guided by the ground truth:

$$C_r = \frac{\Sigma_M \left(\left(\overline{M}_{raw} - \text{mean}\left(\overline{M}_{raw}\right)\right)\left(\overline{A}_{raw} - \text{mean}\left(\overline{A}_{raw}\right)\right)\right)}{\sqrt{\sum_{raw}\left(\overline{M}_{raw} - \text{mean}\left(\overline{M}_{raw}\right)\right)^2 \sum_M \left(\overline{A}_{raw} - \text{mean}\left(\overline{A}_M\right)\right)^2}} \tag{11}$$

In this correlation process, ($\overline{M}_{raw}$ assumes the lead as the moving average of raw PMU measurement data, while ($\overline{A}_{raw}$ functions as the moving average of binary flags identifying attacked samples. The correlation coefficient C_r facilitates the correlation between them, aiding in the validation of the accuracy of identified data streams influenced by FDIs.

The window size length that the **moving average filter uses** in [40] is also changed in line with the actual classification for the purpose of modifying the ground truths. The performance of the classifier was validated by achieving the true classifications. It is important to maintain an equal number of measurement samples in both the converted ground truth and the dataset acquired after the moving average step. The updated ground truth data will retain the majority class X_M and Y_M, owing to the majority voting criteria utilized in the ground truth transformation. For instance, $Y_{raw} = [1, 1, 1, 0]$ and the moving average window length is 4. Since "1" is the majority class in Y_{raw}, this scenario will therefore produce a ground truth value of $Y_M = [1]$.

B. ADVERSARIAL INTERVENTIONS AND DEEP AUTO-ENCODER

Important aspects of our proposed pipeline includes adversarial interventions, which are intended to expose the model to possible attack scenarios in the pre-processing phase. In order to simulate various complex assaults, the adversarial interventions involve adding noise and non-linear perturbations to the PMU data. Through utilizing these adversarial samples during training, the model has the ability to recognize and differentiate between authentic data and data that has been altered by adversaries. The resilience of the model and its capacity to identify FDIs in actual circumstances are much improved by this procedure.

At the heart of this adversarially robust framework is the deep auto-encoder, a specialized neural network. Denoising and feature learning are its two main tasks. The auto-encoder efficiently eliminates noise and distortions from the PMU data during the encoding and decoding process, guaranteeing that the system processes precise and clean data. The auto-encoder also picks up intricate, non-linear patterns in the data, capturing crucial feature representations necessary for precise FDI identification.

The proposed framework is intended to enhance the detection accuracy and reliability of FDIs by incorporating these adversarial interventions and utilizing the deep auto-encoder's capabilities. This dual strategy will preserve the integrity of the data required for state estimate and other crucial applications, strengthening the power system's protection against cyber-attacks while simultaneously guaranteeing the stability and security of the electrical infrastructure.

1) NOISE INJECTION

The representation of noisy measurements z_{noisy}embraces stochastic unpredictability reminiscent of adversarial intrusions. Each measurement z^i_{noisy} is modeled as a blend of the true measurement z^i and a Gaussian noise term ϵ^i.

$$z^i_{noisy} = z^i + \epsilon^i. \tag{12}$$

In vector form, this relationship is expressed as:

$$z^i_{noisy} = z^i + \epsilon^i. \tag{13}$$

z_{noisy} = vector of noisy measurements.
z = vector of true measurements.
ϵ = vector of Gaussian noise, introducing the stochastic nature inherent in adversarial intrusions.

2) NON-LINEARITY INJECTION USING THE HYPERBOLIC TANGENT

Introducing non-linearity involves applying the hyperbolic tangent (tanh) function element-wise to the true measurements, simulating complex manipulations similar to adversarial interventions. Each transformed measurement $z^i_{non-linear}$ is now a function of the corresponding true measurement z^i, providing a non-linear mapping that captures the intricate nature of adversarial manipulations.

$$z^i_{non-linear} = \tanh(\mathrm{z}^{\mathrm{i}}) \tag{14}$$

In vector form, this transformation is expressed as,

$$z_{non-linear} = \tanh(\mathrm{z}) \tag{15}$$

The combined effect of noise and tanh based non-linearity is then represented by,

$$z_{final} = \tanh(\mathrm{z} + \epsilon) \tag{16}$$

Leveraging tanh as the non-linear activation function enhances the model's ability to capture complex relationships present in power system dynamics and better prepares it for handling adversarial intrusions during the training process. This comprehensive representation with tanh based on non-linearity and injected noise provides a more realistic and challenging dataset for robust model training.

3) NOISE ROBUST- DEEP AUTO-ENCODER (NR-DAE)

Deep auto-encoders are known for learning inherent structures and relationship between the data. They've been extensively used in image reconstruction, noise reduction and anomaly detection due to their ability to learn the latent representations of the data and then reconstruct it. They've also found their applications in classification tasks, where an Auto-encoder network is first trained to learn the noise-free representations by using a clean version of the input data as ground truth. The network is trained on a Mean-squared-Error objective to increase the reconstruction power and amplify the ability to extract relevant features. Later the trained encoder part is detached and attached to a classifier for use in general classification tasks. This two step approach makes the feature extraction part robust to noise or any other perturbations, in constrast to a normal CNN network. This application is employed in this study to provide a robust False Data Detection framework and reduce any false predictions.

The deep de-noising auto-encoder named NR-DAE i.e., Noise Robust Deep Auto-Encoder given in Fig. 4 – involves key 1-D convolutional layers denoted by *Convx*. A convolutional layer *Conv*1 with Leaky Rectified Linear Unit (Leaky-ReLU) activation, succeeded by *MaxPool*1 for spatial reduction and *BatchNorm*1 for stabilization is used. Strictly followed by *Conv*2, which refines the encoded features, a *MaxPool*2 which further reduces spatial dimensions, and *BatchNorm*2 to ensures stability. Subsequently, *Conv*3 contributes to additional feature extraction, leading to the final spatial reduction through *MaxPool*3, resulting in the encoded representation denoted as **Encoded**. The decoding phase begins with *DeConv*1, representing a de-convolutional layer for spatial expansion, and *BatchNorm*3 for stabilization. *DeConv*2 and *DeConv*3 refine the decoding process for up-sampling with *BatchNorm*4. *DeConv*4 contributes to the final decoding, and *UpSample*3 increases spatial dimensions to generate the reconstructed output, denoted as **Decoded**. The complete auto-encoder model, integrating these layers, is compiled using the **Adam** optimizer and **Binary Cross Entropy** loss, facilitating the learning process for effective de-noising and reconstruction of input images.

After the successful training of the Auto-Encoder Network on the noisy data, the trained encoder is extracted and weights are frozen. Then trained encoder is then connected to a dense neural network and tuned for **50 Epochs** for training the parameters of the dense neural network in order to classify the input features as normal or FDI detected for each time stamp. Compilation parameters were similar to DAE with the metric replaced with the **Accuracy** and **Loss**, and the dense layers parameters were 64, 128 and 1 neuron for the last classification layer with a **Sigmoid** Activation.

IV. SIMULATION AND RESULTS

This section presents the ML approach for recognizing PMU-based FDIs. The methodology is assessed using the IEEE 14-bus,30-bus,39-bus and 118-bus systems. PMUs are

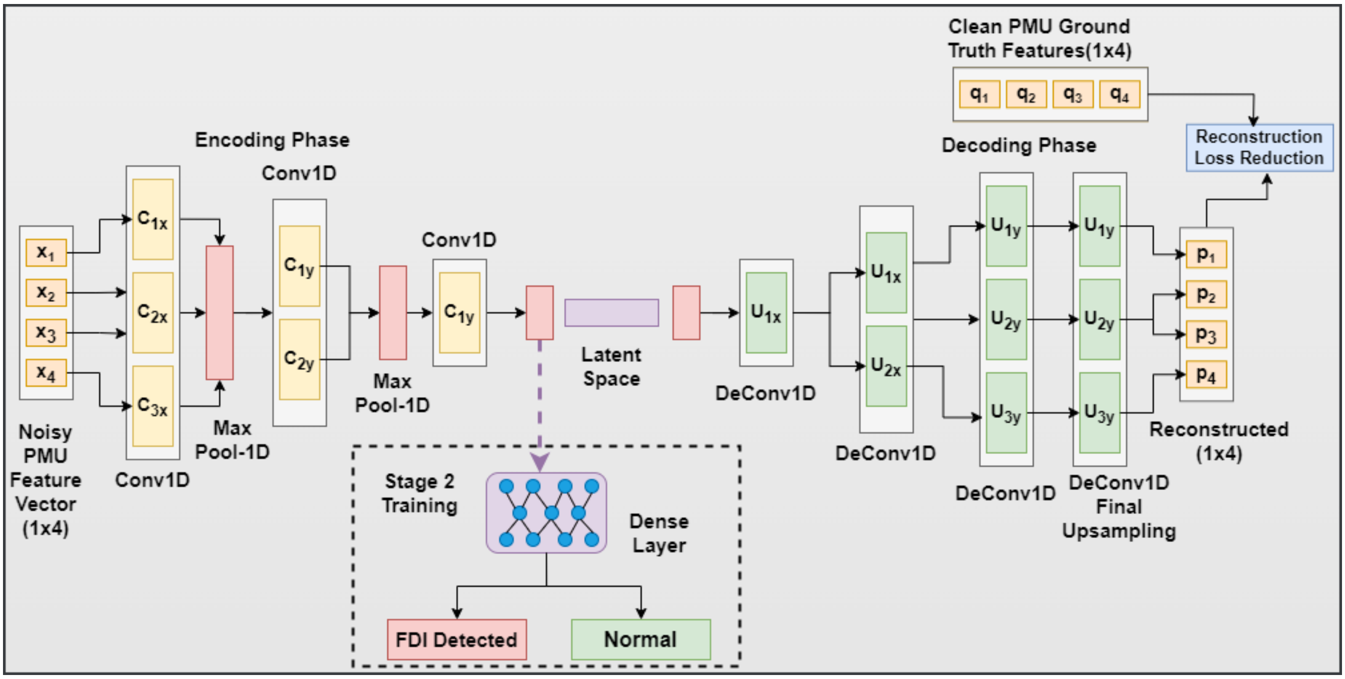


FIGURE 4. Noise robust - deep auto-encoder (NR-DAE) architecture.

deployed for both systems to ensure total observability, as shown in Fig. 5 and 5b [9].

The FDIs are tested on both systems using the procedure described in Section III. According to the proposed method, PMUs are assumed to measure signals at a rate of thirty samples per second. Every PMU is expected to measure the current flow of every bus that borders it.

A. CASE STUDIES

Each PMU is assumed to measure the voltage of the bus where the PMU is located, and the currents of all adjacent buses. Each PMU is sending the measurements at a speed of 30 samples per second. The meter errors of PMU measurements follow the normal distribution with zero mean and standard deviation of 10^{-3}. The tests are performed on the IEEE 14-bus,30-bus,39-bus and 118-bus test systems.

1) DATASET DETAIL

As discussed earlier two primary bus topologies IEEE-14 and IEEE-30 are taken into consideration for evaluation of our proposed work. To further validate the performance, IEEE-39 and IEEE-118 Bus topologies were included. A concise detail for each of the 4 bus topologies are given below:

- **IEEE-14 Bus** A very small scale system with 14-buses, 9 transmission lines and 5 generators. 4 PMU units are optimally placed with complete observability in reach due to the fewer number of components.
- **IEEE-30 Bus** A medium scale system with 30-buses, 6 generators and 41 transmission lines. The optimal PMU placement is crucial and challenging for complete observability. 10 PMU's are strategically located in the proposed study for full observability.
- **IEEE-39 Bus** Includes 39 Buses, 10 generators with 46 transmission lines. Optimal PMU placement is imperative because of increased number of components. Similar to IEEE-10 Bus, 10 PMU units are strategically placed for full observability.
- **IEEE-118 Bus** A very large network topology with 118 Buses, 19 generators and 177 transmission lines. A very delicate and complex PMU placement is required, usual 118 bus topologies employ 42 PMU units placed at strategic points for full observability.

2) EVALUATION SCENARIOS

- **Scenario I:** Every Monte Carlo simulation has a random attack duration and time. The attack's location and strength are maintained constant, though.
- **Scenario II:** Similarly, for the second scenario, every Monte Carlo simulation has a randomized attack duration and time. For every simulation, the attack's position is also different. The intensity of intrusions is constantly maintained, though.
- **Scenario III:** Every Monte Carlo simulation has a randomized attack duration and time. For every simulation, the attack position and intensity (av) are different.
- **Scenario IV:** This scenario involves the simultaneous attack of multiple random PMUs, with a random attack vector used in each Monte Carlo simulation. There is additional randomization in the attack's duration.

Table 1 gives a synopsis of the scenarios listed above. By varying the moving average window's size from 2 to 12,

TABLE 1. Specifications of scenarios used in FDI.

Test Scenarios	Number of Monte-Carlo Trials	Targeted PMU Location	Attack Strategy	Attack Duration
IEEE-14 Power System				
Scenario I	50	Consistent	Sustained	Short
Scenario II	50	A randomly selected PMU	Consistent	Long
Scenario III	50	A randomly chosen PMU	Dynamic	Variable
IEEE-30 Power System				
Scenario I	50	Consistent	Sustained	Short
Scenario II	50	A randomly selected PMU	Dynamic	Long
Scenario III	10	A randomly chosen PMU	Dynamic	Variable
Scenario IV	10	Multiple randomly selected PMUs	Dynamic	Variable
IEEE-39 Power System				
Scenario I	50	Consistent	Sustained	Short
Scenario II	50	A randomly selected PMU	Dynamic	Long
Scenario III	10	A randomly chosen PMU	Dynamic	Variable
IEEE-118 Power System				
Scenario IV	10	Multiple randomly selected PMUs	Dynamic	Variable

we have conducted numerous trials because these best match the scenario that was created. Post-processing incursions severely impair the suggested machine learning-based methodology in [41], making it difficult to detect fraudulent data streams. Changing the window size, however, has no effect on the ML model's performance and is applicable to both machine learning algorithms' identifications of the assaulting samples' data samples and the correlation's location of the impacted bus.

During the preprocessing stages, intentional perturbations, such as injecting noise and introducing non-linearities, were incorporated into the dataset. This perturbation aimed to simulate real-world challenges and evaluate the model's resilience to such variations. Interestingly, the performance of traditional machine learning models experienced a noticeable drop under these perturbations. The injected noise and non-linearities introduced complexities that conventional models struggled to adapt to, resulting in compromised performance. The effects of these perturbations on the features are illustrated and explained in different contexts thoroughly below.

The probability distribution plots in Fig. 6 depict the transformations in each feature's statistical distribution. Before perturbations, we observe certain patterns, such as normal or skewed distributions. Upon introducing noise and non-linearity, these distributions shift, widen, and take on new shapes. Deviations from the original patterns highlight the impact of perturbations on the feature's overall distribution.

Hexagon plots in Fig. 7a and 7b provide a visual representation of feature relationships. Before perturbations, clusters within these plots exhibit specific patterns. After injecting noise and non-linearity, clusters disperse or reshape, indicating changes in the inter-feature relationships. Shifts in hexagon patterns showcase how the perturbations alter the inherent structures of the feature pairs.

Similarly, the parallel coordinates plot in Fig. 10 visualizes the multivariate relationships among features through connecting lines. Before perturbations, these lines show certain smoothness and continuity. The introduction of noise and non-linearity can lead to irregularities, such as discontinuities, bends, or fluctuations in the lines. These alterations signify the perturbation-induced changes in the relationships between features.

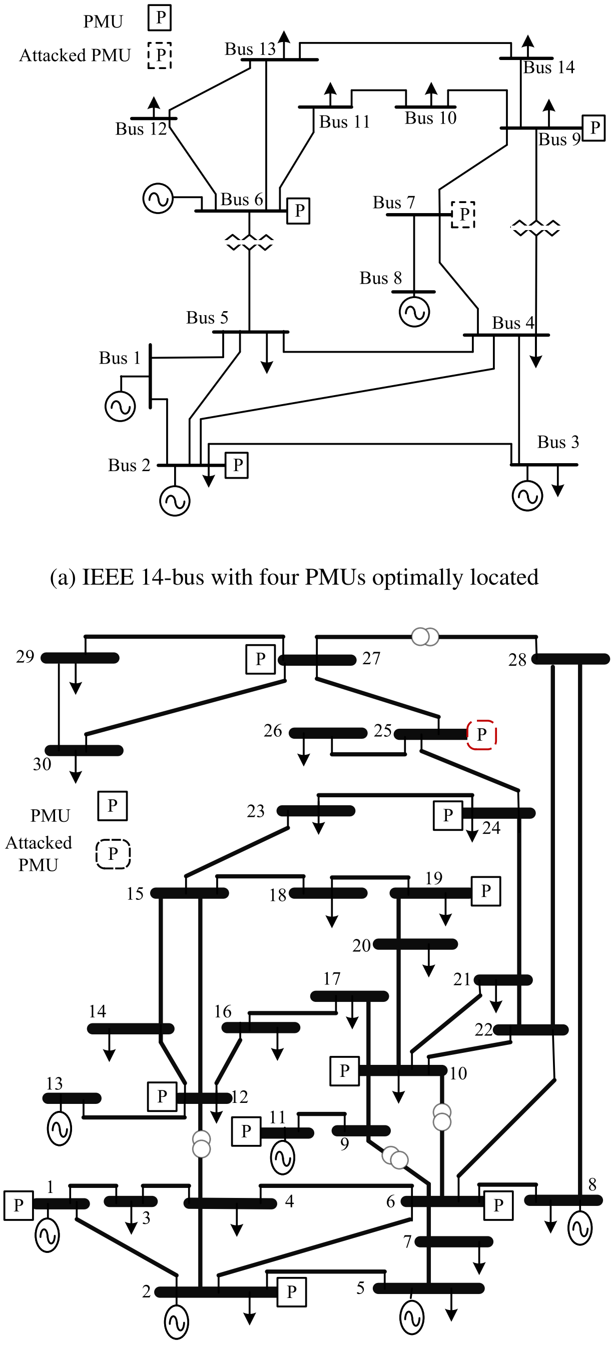
Pair-plot in Fig. 11 offers a detailed view of pairwise relationships between features. Before perturbations, the scatter plots, regression lines, and density distributions exhibit specific patterns. Injecting noise and non-linearity introduced changes in these visualizations, indicating shifts in the relationships between feature pairs. Distinct alterations in scatter plots or regression lines portray the effects of perturbations.

t-SNE visualizations given in Fig. 12 provide a two-dimensional representation of the high-dimensional feature space. Before perturbations, clusters have well-defined shapes and separations. Introducing noise and non-linearity can rearrange the points in the t-SNE plot, disrupting the original cluster patterns. Changes in the arrangement highlight the perturbation-induced alterations in the feature relationships and separability.

UMAP visualizations illustrated in Fig. 13 project high-dimensional data into a lower-dimensional space. Before perturbations, the distribution and proximity of points exhibit specific structures. Injecting noise and non-linearity lead to changes in these structures, with clusters shifting or merging. The alterations in the UMAP plot reveal how perturbations impact the relationships and organization of points in the feature space.

B. ATTACK BUS LOCATION IDENTIFICATION

Prior to subjecting the acquired features to adversarial perturbations, the data streams for each bus is processed via person correlation. This process is repeated for different bus topogies used within this study. The purpose of finding the pearson correlation between the corresponding data streams was to identify the highly correlated streams and separate them from others. As these correlated streams had a higher probability of being compromised, which is validated via ground truths. For our simulated data, different scenarios



(a) IEEE 14-bus with four PMUs optimally located

(b) IEEE 30-bus with 10 PMUs optimally located

FIGURE 5. IEEE test systems with full observability.

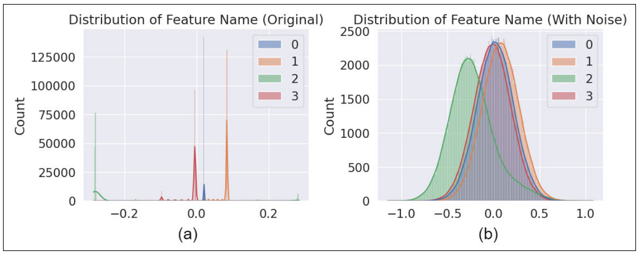


FIGURE 6. Distribution of features (a) Original features (b) Perturbed features.

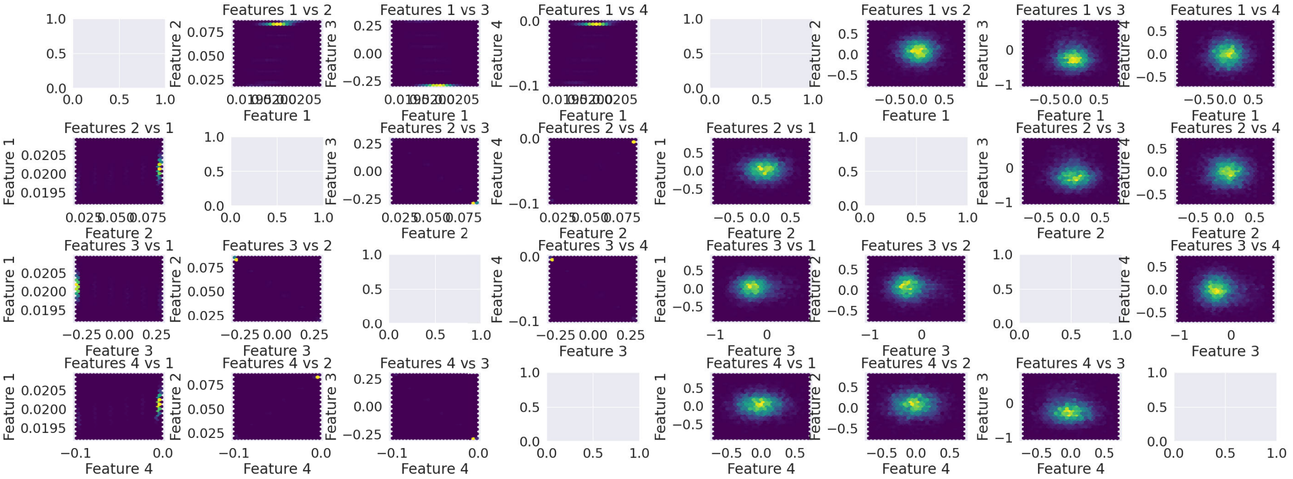
TABLE 2. Performance evaluation of proposed pipeline under different attack scenarios for IEEE-14,30,39 and 118 bus topologies.

Case	Attacked PMUs	Attack Vector	FDI Detection Accuracy
IEEE-14 Bus Test System	Scenario I: Case 1	7	0.9844
	Scenario I: Case 5	2	0.9811
	Scenario II: Case 9	6	0.9835
	Scenario II: Case 19	9	0.9866
	Scenario III: Case 2	7	0.9906
	Scenario III: Case 7	6	0.9868
IEEE-30 Bus Test System	Scenario II: Case 3	12	0.9867
	Scenario I: Case 7	8	0.9883
	Scenario III: Case 8	2	0.9910
	Scenario III: Case 5	24	0.9901
	Scenario IV: Case 2	11, 27	0.9905
	Scenario IV: Case 7	1, 12	0.9884
IEEE-39 Bus Test System	Scenario I: Case 5	19	0.9763
	Scenario II: Case 7	13	0.9728
	Scenario II: Case 20, 14	2	0.9802
	Scenario III: Case 2, 5	6	0.9891
IEEE-118 Bus Test System	Scenario IV: Case 7	21	0.9903
	Scenario IV: Case 5	7, 9	0.9896

were designed ranging from infecting a single random PMU unit to compromising multiple PMU units. The scenarios were iterated randomly for all four bus type topologies. With the primary topologies i.e., IEEE-14 Bus and IEEE-30 Bus types being subjected to all four scenarios and the remaining types to some of these for validation purpose as given in Table 2. For each scenario of the respective bus type, different streams of data were compromised which were detected proficiently with the correlation based method as shown in in the Fig. 8. Additionally, it can be seen in Fig. 9 – where 5 different data streams attained from IEEE-14 Bus topology for one of the scenario are shown in green and red. The Red streams are compromised by synthetic FDI attacks while the Green streams are normal. The attacked measurements are skewed to a very minimum measurements as in pragmatic circumstances, the attackers only compromise a few measurements that guarantee a successful attack due to the computational expense. The correlation between all streams are calculated and the two correlated streams are singled out, these are then compared with the ground truths to calculate the error and accuracy. As it can be seen, the predicted attack streams are detected with high accuracy.

C. EVALUATION ON PROPOSED NR-DAE METHOD

Our proposed auto-encoder named NR-DAE showcased a remarkable degree of robustness to these pre-processing perturbations. The inherent ability of auto-encoders to learn intricate patterns and abstract representations enabled them to effectively de-noise and reconstruct the input data, even in the presence of intentional disturbances. Where the conventional Machine Learning models flailed, our proposed method significantly showed adaptability and prowess. This resilience underscores the advantageous capacity of auto-encoders in handling real-world data variability, highlighting their potential for applications in scenarios where pre-processing challenges may arise. The evaluation of our proposed



(a) Hexagon Plot for Acquired Features

(b) Hexagon Plot for Perturbed Features

FIGURE 7. Hexagon plot for original and perturbed features.

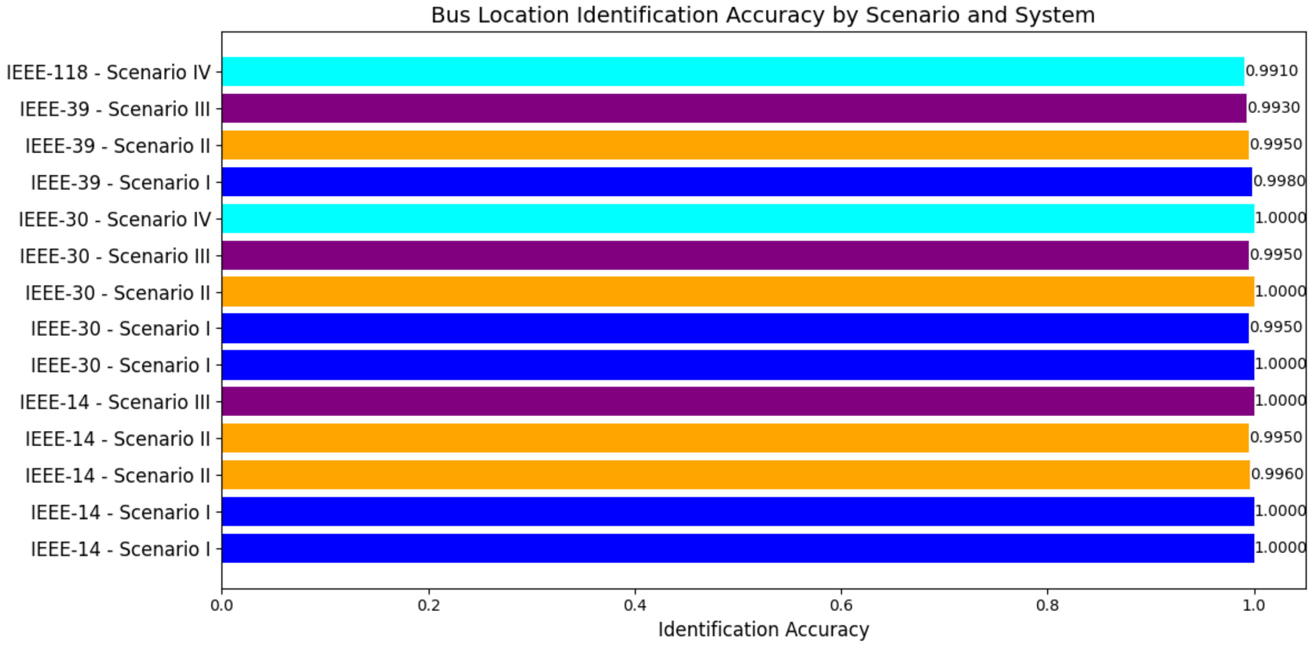


FIGURE 8. Pearson correlation based bus identification accuracy for false data injection (FDI) attacks.

adversarial intervention method on the same scenarios as [41] is extended, and compared in the below section.

1) EVALUATION ON SCENARIO: I

The comprehensive assessment of the presented scenarios in Table 3a elucidates distinct patterns in model performance. In Scenario I, a meticulous examination of confusion matrices, F1-scores, and ROC Curves reveals noteworthy outcomes. Auto-encoders, integral to the novel method, exhibit effectiveness in detecting false-injected data within the network, surpassing the performance of conventional approaches i.e., CNN and Machine Learning Models by achieving an accuracy of 98.44 and F1 score of 98.76, compared to the max F1-Score of approx. 90.25 achieved by XG-Boost and SVM. By observing the confusion matrix, it can be seen that the model's bias towards one class was more which might have been an implication of the simulating conditions. The overall efficacy of the novel method is evident in the training curves in Fig. 14. Table 1 and 2 provide additional insights into different variations of attacks with the targeted PMU location being the same for all three bus topologies, with the strategy

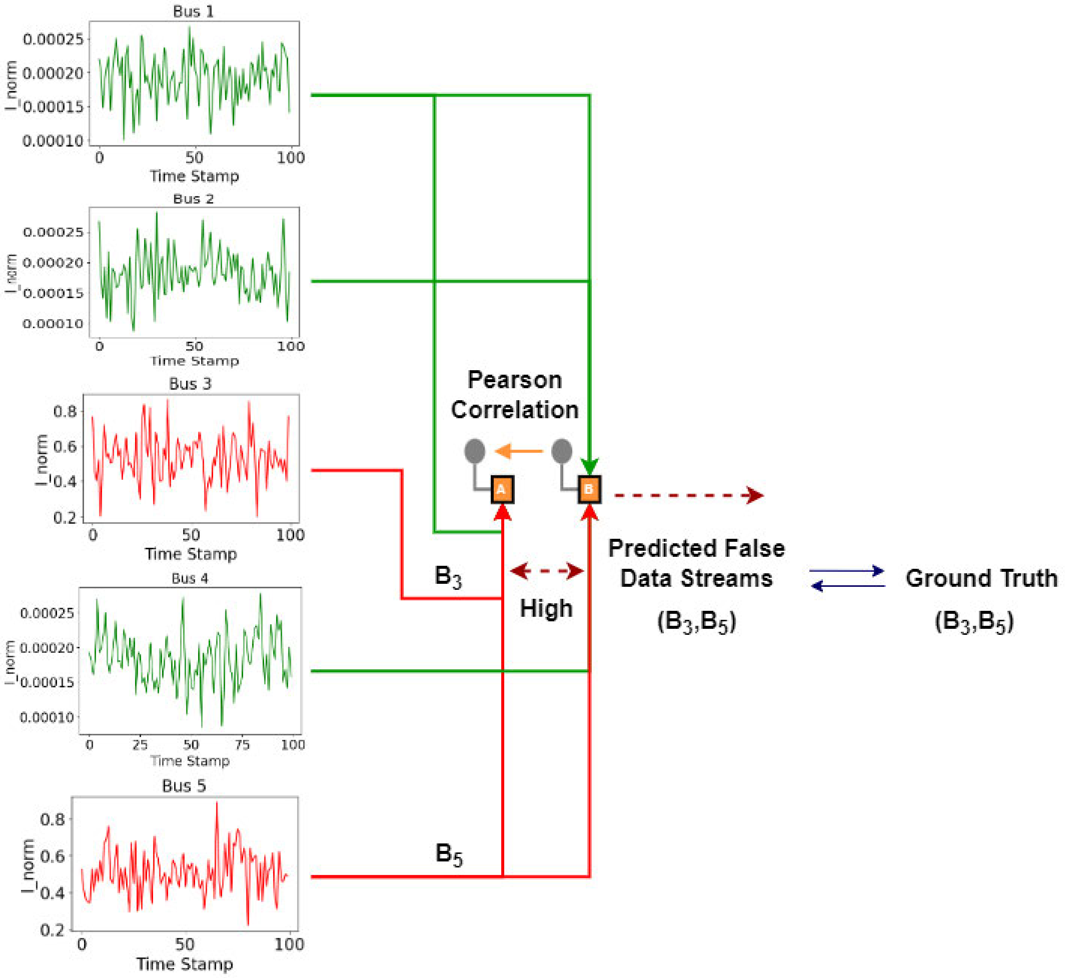


FIGURE 9. Location identification using person correlation.

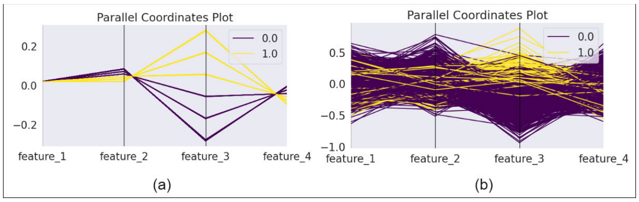


FIGURE 10. Parallel coordinate plot (a) Before perturbations (b) After perturbations.

and duration kept constant and short. This is reflected in the average F1-Score and Accuracy, as the short simulated duration of attack lacked the adequate sample diversity to learn the mappings properly. Yet the results from our proposed model were good enough even under these configurations. Moreover, the attack localization accuracy from the correlation technique showed lower variability in the accuracy values as for the scenario 1 the localization accuracy remained between 99.5% to 100 %, when just one random PMU was compromised. The performance of different bus system topologies was also good, with the highest accuracy value achieved by IEEE-30 by acquiring a value of 98.83%, followed by IEEE-14 Bus and IEEE-39 Bus systems.

2) EVALUATION ON SCENARIO: II

Shifting focus to Scenario II given in Table 1 and 2, the prominence of auto-encoders becomes apparent. Despite the intricacies introduced during later-stage perturbations, auto-encoders consistently outperform conventional machine learning and 1D-CNN model by a large margin, achieving an accuracy of 98.66%. This underscores the robustness of auto-encoders in capturing and reconstructing meaningful features, thereby enhancing overall detection outcomes. The improved accuracy in comparison to scenario I can be attributed to a longer attack duration and a larger sample size with the variability in attack vectors that were properly distinguished by the proposed techniques. However, the difference between the average results of the two scenarios is marginal and varies between different ML models. Furthermore, the confusion matrix for scenario II also showed a balanced classification between the fault samples and

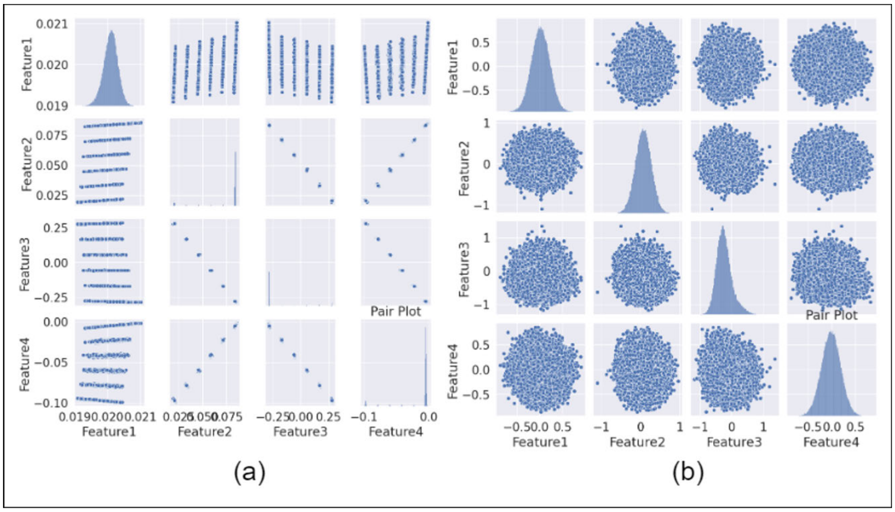


FIGURE 11. Pair-plot (a) Before perturbations (b) After perturbations.

TABLE 3. Performance evaluation of different ML and DL models for all scenarios.

Models	F1-Score	Accuracy
ML - Models		
XGB	90.25	90.04
QDA	80.37	80.12
SVM	90.25	90.11
DL – Models		
NR-DAE	98.76	98.44
CNN	96.02	95.36

(a) Performance Evaluation under Scenario I

Models	F1-Score	Accuracy
ML - Models		
XGB	89.36	90.01
QDA	80.55	80.05
SVM	90.02	90.54
DL – Models		
NR-DAE	99.01	98.66
CNN	94.26	94.02

(b) Performance Evaluation under Scenario II

Models	F1-Score	Accuracy
ML - Models		
XGB	90.07	90.46
QDA	81.53	82.31
SVM	91.65	91.37
DL – Models		
NR-DAE	99.12	99.10
CNN	97.98	97.64

(c) Performance Evaluation under Scenario III

Models	F1-Score	Accuracy
ML - Models		
XGB	89.98	89.71
QDA	81.02	81.68
SVM	90.74	90.15
DL – Models		
NR-DAE	99.01	99.05
CNN	96.54	96.08

(d) Performance Evaluation under Scenario IV

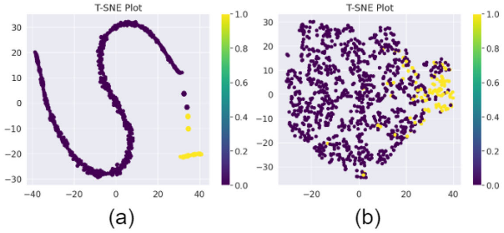


FIGURE 12. t-SNE (a) Before perturbations) (b) After perturbations.

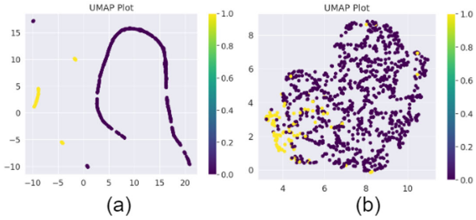


FIGURE 13. U-map (a) Before perturbations (b) After perturbations.

the normal samples contrary to scenario I because of the larger sample size for the attacked vector and dynamic conditions. Additionally, the high accuracy for FDI attack localization remained similar with little variations in contrast to Scenario I. The performance of all four IEEE bus system topologies was somewhat higher than that of scenario I, with the highest accuracy value of 98.67 % acquired by IEEE-30 Bus system, sharply followed by IEEE-14 with a value

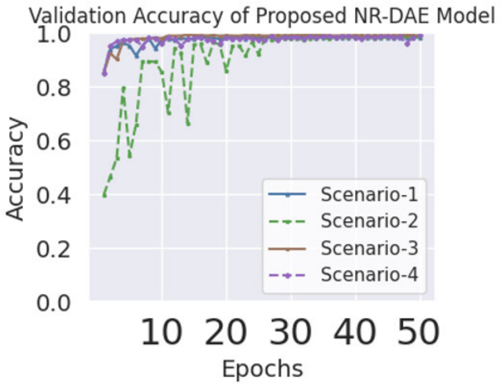


FIGURE 14. Val- accuracy of proposed RN-DAE model for all four scenarios.

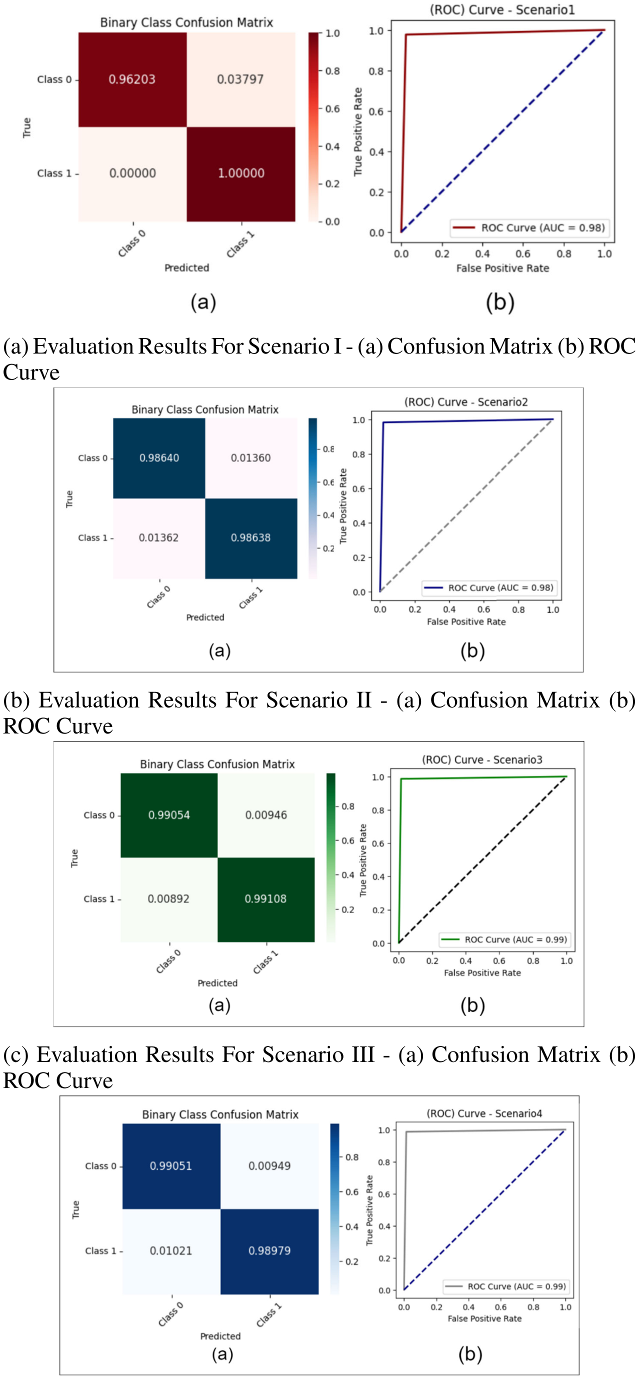
of 98.66%. However, IEEE-39 Bus system performed better under Scenario I then at Scenario II.

3) EVALUATION ON SCENARIO: III

The power system complexities of the IEEE-30 bus dataset highlight the difficulties caused by occlusions in Scenario III. Using their ability to learn complex representations, autoencoders perform exceptionally well in this case, achieving accuracy and F1-Score of 99.10 and 99.12, as seen by the confusion matrix and ROC curve in Fig. 15a and Table 3c. Because of their capacity to manage occlusions well, it is possible to distinguish between samples that have been attacked and those that have not, making them a reliable option for power system cyber-security applications. The key factor that distinguishes the operating conditions of scenario III from the other two is the variability in attack duration. The dynamic attack duration must've provided optimal data points which resonated well with the learning capabilities of the model, yielding high accuracy. It can also be observed that IEEE-30 Bus system topology performed exceptionally across all scenarios with the maximum accuracy value 99.10% achieved in scenario III, strictly followed by IEEE-14 Bus system yielding 99.06% and IEEE-118 Bus system moving along with 99.03% accuracy. It is evident that not only a single bus system topology performed well under these operating conditions but all four bus system topologies performed well under varied attack duration.

4) EVALUATION ON SCENARIO: IV

The resiliency of auto-encoders is seen even more in Scenario IV, which combines the difficulties of later-stage perturbations and occlusions. The complex interactions between noise and non-linearities presented a significant obstacle for conventional machine learning models, as demonstrated by the SVM, XGB, and QDA algorithms performance declines. Nevertheless, auto-encoders continuously beat other models, demonstrating their flexibility in handling intricate situations in the IEEE-30 bus dataset with accuracy and F1-Score of 99.05 and 99.01%, as illustrated in Fig. 15b and Table 3d. The simulating conditions for scenario IV were similar to scenario III, with the only difference being the number of compromised PMU units. Herein, multiple random PMU units underwent simulated attacks, and the scenario was implemented on only two bus type topologies i.e., IEEE-30 Bus and IEEE-118 Bus type. Both achieved a higher accuracy, similar to scenario III with a marginal decline in comparison. The identical performance metrics allude to a single common variable being responsible for the improved performance



(a) Evaluation Results For Scenario I - (a) Confusion Matrix (b) ROC Curve

(b) Evaluation Results For Scenario II - (a) Confusion Matrix (b) ROC Curve

(c) Evaluation Results For Scenario III - (a) Confusion Matrix (b) ROC Curve

(d) Evaluation Results For Scenario IV - (a) Confusion Matrix (b) ROC Curve

FIGURE 15. Evaluation results and ROC curves.

TABLE 4. Performance average measure of DL and ML models for all scenarios.

Models	Accuracy	Precision	Recall	Kappa Statistics	Mat. Correlations
Machine Learning – 1D (1x4) – Average of Four Scenarios					
QDA [41]	81.06	82.15	80.22	79.32	78.96
SVM [41]	90.32	91.45	89.55	88.73	88.71
XGB [41]	90.06	89.21	90.88	88.61	88.61
DT	80.84	81.56	79.95	79.53	79.22
LR	90.06	89.44	90.67	87.61	88.06
NR-DAE	98.76	99.03	98.85	97.69	97.82
CNN	95.35	95.02	95.52	95.01	95.11

which is the variable attack duration. The highest accuracy value was achieved by IEEE- 30 Bus system i.e., 99.05% and IEEE-118 Bus system also performed well with an accuracy of 99.03% at case 7.

D. EVALUATION ON VARIED LOAD PROFILES

To further evaluate the performance of our proposed adversarially robust framework, varied load profiles are used to analyze and assess the impact of different load characteristics on False Data Injection (FDI) detection. In this regard, four diffent scenarios were proposed with different load variations, to evaluate how variations in the electrical load affect the performance of FDI detection algorithms.

- **Load Scenario I:** Increasing Load by 10%
  Both the suggested auto-encoder model and the Machine Learning (ML) models show stable performance in Scenario I, where the electrical demand is increased by 10% as given in Table. The proposed auto-encoder model achieves an F1-Score of 93.27% and an Accuracy of 95.44%, whereas in ML models, SVM achieve an average F1-Score of 89.27% and an average Accuracy of 90.36% — followed sharply by XGB and QDA.
- **Load Scenario II:** Decreasing Load by 10% In comparison to Scenario I, there is a little drop in performance for some ML models in Scenario II due to a 10% reduction in electrical demand, while SVM retained its performance by achieving an accuracy of 92.13% whereas the XGB and QDA had their accuracy values at 87.27% and 80.05%. From Table 5b it can be observed, that the suggested auto-encoder model continues to outperform the ML models by attaining an average F1-Score of 97.21% and an average Accuracy of 96.67%. While the CNN based network also surpassed the ML models by achieving an accuracy of 94.10%.
- **Load Scenario III:** Increasing Load by 20%
  In comparison to Scenarios I and II, Scenario III was simulated with a 20% increase in load demand, which causes performance indicators to drop even further compared to the 10% increase in load. The suggested auto-encoder model continues to perform better than the ML models in spite of this drop, with an average F1-Score of 95.04% and an average Accuracy of 95.23%, compared to the highest ML models' F1-Score of 98.76% and Accuracy of 90.57%.
- **Load Scenario IV:** Decreasing Load by 20%
  Lastly, the performance of each model was seen to have been noticeably improved in Scenario IV, where the electrical load is reduced by 20%. The highest performing i.e., SVM obtained an F1-Score of 91.44% and an Accuracy of 92.15%, whereas the suggested auto-encoder model retained its stability with an average F1-Score of 97.99% and an average Accuracy of 98.63%.

1) RESULT ANALYSIS OF FDI DETECTION UNDER VARIED LOAD CONDITIONS

The dynamic nature of power grid networks and numerous model-inherent elements can be attributed to the observed difference in performance across various load scenarios. First, variations in the load cause the grid's distribution of electrical parameters to shift, which in turn causes changes in the patterns and characteristics of the signals. Performance may suffer if traditional machine learning models, such SVM and XGB, are unable to generalize to these changing patterns. On the other hand, the suggested auto-encoder model is more flexible to changing load conditions because of its capacity to identify and depict intricate relationships within the data. emphasizes how crucial it is to use cutting-edge machine learning methods that can adjust to the dynamic nature of power grid networks. Furthermore, the deep learning-based method extracts robust features by utilizing the hierarchical representations that it learnt during training, which improves its ability to distinguish between abnormal and normal grid activity under various load scenarios. Furthermore, typical machine learning models that rely on linear assumptions or predefined characteristics may face difficulties due to the inherent non-linearity and complexity of power grid dynamics. On the other hand, deep learning models—like CNN—are by nature better at identifying complex patterns and non-linear correlations, which enables them to continue operating steadily even under severe load variations. All things considered, the better performance of the suggested CNN and auto-encoder models under various load scenarios highlights how crucial it is to use cutting-edge machine learning methods that can adjust to the dynamic nature of power grid systems.

E. COMPARISON WITH THE EXISTING WORKS

The performance of the proposed approach is shown in Table 4, where it is compared with [41] and other popular methods. It can be seen that the measurement noise detrorieates the performance of QDA, SVM, and XGB of [41]. Comparing with the other ML and DL methods shows the superiority of our approach (NR-DAE).

A comparative evaluation is established in Table 6 to compare our proposed method to Phase Lock Value (PLV) from [9]. Our method suggests simulating constrained adversarial interventions within acquired features to assess their sensitivity under different scenarios. Although the

TABLE 5. Performance evaluation of different ML and DL models for varying load scenarios.

(a) Performance Evaluation under Load Scenario I

Models	F1-Score	Accuracy
ML - Models		
XGB	85.34	88.11
SVM	89.27	90.36
QDA	84.37	85.21
DL – Models		
NR-DAE	93.27	95.44
CNN	91.02	92.13

(b) Performance Evaluation under Load Scenario II

Models	F1-Score	Accuracy
ML - Models		
XGB	87.03	87.27
QDA	82.39	80.05
SVM	91.42	92.13
DL – Models		
NR-DAE	97.21	96.67
CNN	93.45	94.10

(c) Performance Evaluation under Load Scenario III

Models	F1-Score	Accuracy
ML - Models		
XGB	84.88	86.01
QDA	84.29	85.03
SVM	89.76	90.57
DL – Models		
NR-DAE	95.04	95.23
CNN	90.11	91.78

(d) Performance Evaluation under Load Scenario IV

Models	F1-Score	Accuracy
ML - Models		
XGB	90.39	91.68
QDA	87.63	88.72
SVM	91.44	92.15
DL – Models		
NR-DAE	97.99	98.63
CNN	94.28	94.02

TABLE 6. Performance comparison with existing literature.

Case		Proposed Method		[27]	
		(F1-Score)	Attack (location and instance)	(F1-Score)	Attack (location and instance)
IEEE 14-Bus	Scenario : 1	99.05	Identified	99.05	Time instance only
	Scenario : 2	99.01	Identified	99.99	Time instance only
	Scenario : 3	99.01	Identified	99.80	Time instance only
IEEE 30-Bus	Scenario : 1	99.10	Identified	-	Time instance only
	Scenario : 2	98.99	Identified	-	
	Scenario : 3	99.02	Identified	98.97	
	Scenario : 4	99.07	Identified	-	

performance metrics particularly F1 - Scores for PLV show a significant prowess in detecting the FDI, the reliability of the method under post-processing adversarial attempts can be put into question. The PLV methods usually rely on phase information, however a slight corruption in the phase information can compromise the integrity of the method. Moreover, the PLV method can be desensitized to true patterns by the added noise, resulting in misleading learning curves. Additionally, the PLV method is sensitive to window sizing as well. On the contrary, the underscoring contributions of our proposed method are locating compromised PMUs, providing invariance to window sizing, and the robustness to adversarial interventions at post-processing stages, resulting in a reliable detection.

V. CONCLUSION

This research tackles the escalating worries about cyber-threats, concentrating especially on the serious effects of false data injection (FDI) assaults on PMUs and the vulnerability of ML based FDI frameworks. To strengthen the FDI framework against adversarial attacks and improve the classification accuracy of False Data Injection attacks, a novel framework is proposed using auto-encoders. The proposed auto-encoder network performs well under a range of load circumstances and Monte Carlo simulations. The system's resistance to cyber-attacks is further increased by incorporating non-linear interventions and adversarial noise into the FDI framework during pre-processing. According to quantitative evaluation, the auto-encoder model achieves high FDI detection accuracy, averaging 98.3% identification accuracy in Monte Carlo simulations and 96.5% accuracy on average under different load scenarios. When adversarial noise is introduced, the model gains proficiency in distinguishing authentic data from fabricated inputs, which results in a notable enhancement in the accuracy of FDI detection when compared to traditional machine learning models.

Furthermore, the model's capacity to identify anomalous behavior is further enhanced by non-linear interventions, such as non-linear activation functions, which allow the model to capture the intricate correlations present in the data. By enhancing the power grid's entire security posture and protecting against cyber threats, the combined strategy not only increases the accuracy of FDI detection but also ensures the stability of electrical infrastructure. Subsequent studies can investigate sophisticated cybersecurity strategies and integrate cutting-edge technologies to adjust to changing risks in power system operations.

In the future, we will investigate the effects of different load characteristics on FDI detection. This will involve converting constant MVA loads into a mixture of different load profiles, including constant MVA, current, and admittance. This study will clarify how load variations affect the system's resilience and FDI detection accuracy, opening the door to the development of more flexible FDI detection frameworks. In addition, we will investigate cutting edge approaches like deep learning and machine learning with more refined adversarial interventions to improve the detection performance under various load conditions, supporting power grid security and reliability.

REFERENCES

[1] G. Liang, J. Zhao, F. Luo, S. R. Weller, and Z. Y. Dong, "A review of false data injection attacks against modern power systems," *IEEE Trans. Smart Grid*, vol. 8, no. 4, pp. 1630–1638, Jul. 2017.

[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. Power Syst.*, vol. 32, no. 4, pp. 3317–3318, Jul. 2017.

[3] P. Yuan, Q. Zhang, T. Zhang, C. Chi, X. Zhang, P. Li, and X. Gong, "Analysis and enlightenment of the blackouts in Argentina and new York," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2019, pp. 5879–5884.

[4] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 1, pp. 1–33, May 2011.

[5] L. Che, X. Liu, Z. Li, and Y. Wen, "False data injection attacks induced sequential outages in power systems," *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 1513–1523, Mar. 2019.

[6] L. Che, X. Liu, T. Ding, and Z. Li, "Revealing impacts of cyber attacks on power grids vulnerability to cascading failures," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 66, no. 6, pp. 1058–1062, Jun. 2019.

[7] X. Liu, Z. Li, Z. Shuai, and Y. Wen, "Cyber attacks against the economic operation of power systems: A fast solution," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 1023–1025, Mar. 2017.

[8] M. Jafari, M. A. Rahman, and S. Paudyal, "Optimal false data injection attacks against power system frequency stability," *IEEE Trans. Smart Grid*, vol. 14, no. 2, pp. 1276–1288, Mar. 2023.

[9] S. Almasabi, T. Alsuwian, E. Javed, M. Irfan, M. Jalalah, B. Aljafari, and F. A. Harraz, "A novel technique to detect false data injection attacks on phasor measurement units," *Sensors*, vol. 21, no. 17, p. 5791, Aug. 2021.

[10] Y. Song, Z. Yu, X. Liu, J. Tian, and M. Chen, "Isolation forest based detection for false data attacks in power systems," in *Proc. IEEE Innov. Smart Grid Technol. Asia (ISGT Asia)*, May 2019, pp. 4170–4174.

[11] A. Bhattacharjee, A. K. Mondal, A. Verma, S. Mishra, and T. K. Saha, "Deep latent space clustering for detection of stealthy false data injection attacks against AC state estimation in power systems," *IEEE Trans. Smart Grid*, vol. 14, no. 3, pp. 2338–2351, May 2023.

[12] Y. Zhang, J. Wang, and B. Chen, "Detecting false data injection attacks in smart grids: A semi-supervised deep learning approach," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 623–634, Jan. 2021.

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

[14] S. Wang, S. Bi, and Y. A. Zhang, "Locational detection of the false data injection attack in a smart grid: A multilabel classification approach," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8218–8227, Sep. 2020.

[15] D. Wang, X. Wang, Y. Zhang, and L. Jin, "Detection of power grid disturbances and cyber-attacks based on machine learning," *J. Inf. Secur. Appl.*, vol. 46, pp. 42–52, Jun. 2019.

[16] L. Mili, T. V. Cutsem, and M. Ribbens-Pavella, "Bad data identification methods in power system state estimation—A comparative study," *IEEE Trans. Power App. Syst.*, vol. PAS-104, no. 11, pp. 3037–3049, Nov. 1985.

[17] X. Liu and Z. Li, "False data attacks against AC state estimation with incomplete network information," *IEEE Trans. Smart Grid*, vol. 8, no. 5, pp. 2239–2248, Sep. 2017.

[18] M. Dehghani, A. Kavousi-Fard, T. Niknam, and O. Avatefipour, "A robust voltage and current controller of parallel inverters in smart island: A novel approach," *Energy*, vol. 214, Jan. 2021, Art. no. 118879.

[19] Y. Guan and X. Ge, "Distributed attack detection and secure estimation of networked cyber-physical systems against false data injection attacks and jamming attacks," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 4, no. 1, pp. 48–59, Mar. 2018.

[20] A. G. Phadke and J. S. Thorp, *Synchronized Phasor Measurements and Their Applications*, vol. 1. Cham, Switzerland: Springer, 2008.

[21] C. Beasley, G. K. Venayagamoorthy, and R. Brooks, "Cyber security evaluation of synchrophasors in a power system," in *Proc. Clemson Univ. Power Syst. Conf.*, Mar. 2014, pp. 1–5.

[22] W. Ding, M. Xu, Y. Huang, and P. Zhao, "Cyber risks of PMU networks with observation errors: Assessment and mitigation," *Rel. Eng. Syst. Saf.*, vol. 198, Jun. 2020, Art. no. 106873. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0951832019309494

[23] J. Xie and A. P. S. Meliopoulos, "Sensitive detection of GPS spoofing attack in phasor measurement units via quasi-dynamic state estimation," *Computer*, vol. 53, no. 5, pp. 63–72, May 2020.

[24] E. Schmidt, N. Gatsis, and D. Akopian, "A GPS spoofing detection and classification correlator-based technique using the LASSO," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 56, no. 6, pp. 4224–4237, Dec. 2020.

[25] T. A. Alexopoulos, G. N. Korres, and N. M. Manousakis, "Complementarity reformulations for false data injection attacks on PMU-only state estimation," *Electr. Power Syst. Res.*, vol. 189, Dec. 2020, Art. no. 106796.

[26] Z. Chu, J. Zhang, O. Kosut, and L. Sankar, "$N-1$ reliability makes it difficult for false data injection attacks to cause physical consequences," *IEEE Trans. Power Syst.*, vol. 36, no. 5, pp. 3897–3906, Sep. 2021.

[27] J. Khazaei and A. Asrari, "Second-order cone programming relaxation of stealthy cyberattacks resulting in overvoltages in cyber-physical power systems," *IEEE Syst. J.*, vol. 16, no. 3, pp. 4267–4278, Sep. 2022.

[28] W. Ding, M. Xu, Y. Huang, P. Zhao, and F. Song, "Cyber attacks on PMU placement in a smart grid: Characterization and optimization," *Rel. Eng. Syst. Saf.*, vol. 212, Aug. 2021, Art. no. 107586. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0951832021001344

[29] Z. Liu and L. Wang, "Defense strategy against load redistribution attacks on power systems considering insider threats," *IEEE Trans. Smart Grid*, vol. 12, no. 2, pp. 1529–1540, Mar. 2021.

[30] Y. Huang, T. He, N. R. Chaudhuri, and T. F. L. Porta, "Preventing outages under coordinated cyber–physical attack with secured PMUs," *IEEE Trans. Smart Grid*, vol. 13, no. 4, pp. 3160–3173, Jul. 2022.

[31] X. Niu, Y. Tong, and J. Sun, "Vulnerability assessment for PMU communication networks," in *Smart Computing and Communication*, M. Qiu, Ed., Cham, Switzerland: Springer, 2018, pp. 29–38.

[32] S. D'Antonio, R. Nardone, N. Russo, and F. Uccello, "A tamper-resistant storage framework for smart grid security," in *Proc. 31st Euromicro Int. Conf. Parallel, Distrib. Netw.-Based Process. (PDP)*, Mar. 2023, pp. 100–103.

[33] T. Berghout, M. Benbouzid, and Y. Amirat, "Towards resilient and secure smart grids against PMU adversarial attacks: A deep learning-based robust data engineering approach," *Electronics*, vol. 12, no. 12, p. 2554, Jun. 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/12/2554

[34] M. Kamal, A. Shahsavari, and H. Mohsenian-Rad, "Poisoning attack against event classification in distribution synchrophasor measurements," in *Proc. IEEE Int. Conf. Commun., Control, Comput. Technol. Smart Grids (SmartGridComm)*, Oct. 2021, pp. 327–332.

[35] Y. Cheng, K. Yamashita, J. Follum, and N. Yu, "Adversarial purification for data-driven power system event classifiers with diffusion models," 2023, *arXiv:2311.07110*.

[36] O. Bahwal, O. Kosut, and L. Sankar, "An adversarial approach to evaluating the robustness of event identification models," 2024, *arXiv:2402.12338*.

[37] M. Göl and A. Abur, "A fast decoupled state estimator for systems measured by PMUs," *IEEE Trans. Power Syst.*, vol. 30, no. 5, pp. 2766–2771, Sep. 2015.

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

[39] 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 Syst. J.*, vol. 11, no. 4, pp. 2640–2651, Dec. 2017.

[40] X. Liu, Z. Li, X. Liu, and Z. Li, "Masking transmission line outages via false data injection attacks," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 7, pp. 1592–1602, Jul. 2016.

[41] S. Almasabi, T. Alsuwian, M. Awais, M. Irfan, M. Jalalah, B. Aljafari, and F. A. Harraz, "False data injection detection for phasor measurement units," *Sensors*, vol. 22, no. 9, p. 3146, Apr. 2022.

SALEH ALMASABI (Member, IEEE) received the B.E. degree in electrical and electronics engineering from the King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia, in 2008, the M.S. degree from Wayne State University, Detroit, MI, USA, in 2014, and the Ph.D. degree from Michigan State University, East Lansing, MI, USA. He is currently an Assistant Professor with the Electrical Engineering Department, Najran University, Saudi Arabia. His research interests include power systems, smart grids, reliability, PMU applications, and cyber-physical security.

ZOHAIB MUSHTAQ received the B.Sc. degree from Islamia University, the M.S. degree from the Government College University, Lahore, and the Ph.D. degree in electrical engineering (artificial intelligence) from the National Taiwan University of Science and Technology, in 2020. He was an Assistant Professor with Riphah International University, Islamabad. He is currently an Assistant Professor of electrical engineering with the Department of Electrical Electronics and Computer Systems, University of Sargodha, Sargodha, Pakistan. He has published research articles in various IEEE and other reputable journals. His current research interests include neural networks, machine learning, deep learning, computer vision, and data science.

NABEEL AHMED KHAN received the degree in electrical engineering from Riphah International University, Islamabad, in 2022. Since then, he has been collaborating in research endeavors as a Research Associate in projects of diverse lineament. He has co-authored conferences and research papers in IEEE and other reputed journals. His research interests include deep learning, condition monitoring, signal processing, computer vision, machine learning, and data science.

MUHAMMAD IRFAN received the Ph.D. degree in electrical and electronic engineering from Universiti Teknologi PETRONAS, Malaysia, in 2016. He has two years of industry experience (Oct 2009–Oct 2011) and seven years of academic experience in teaching and research. Currently, he is an Associate Professor with the Electrical Engineering Department, Najran University, Saudi Arabia. He has authored several research papers in reputed journals, books, and conference proceedings (Google Scholar Citations of 3100 and H-index of 25). His research interests include automation and process control, condition monitoring, vibration analysis, artificial intelligence, the Internet of Things (IoT), big data analytics, smart cities, and smart healthcare.