Stealthy MTD Against Unsupervised Learning-based Blind FDI Attacks in Power Systems

Martin Higgins, Fei Teng, Member, IEEE, and Thomas Parisini, Fellow, IEEE

Abstract—This paper examines how moving target defences (MTD) implemented in power systems can be countered by unsupervised learning-based false data injection (FDI) attack and how MTD can be combined with physical watermarking to enhance the system resilience. A novel intelligent attack, which incorporates density-based spatial clustering and dimensionality reduction, is developed and shown to be effective in maintaining stealth in the presence of traditional MTD strategies. In resisting this new type of attack, a novel implementation of MTD combining with physical watermarking is proposed by adding Gaussian watermark into physical plant parameters to drive detection of traditional and intelligent FDI attacks, while remaining hidden to the attackers and limiting the impact on system operation and stability.

Index Terms—Cybersecurity, false data injection attacks, power systems state estimation, moving target defence, and physical watermarking.

I. INTRODUCTION

THE modern power system is increasingly dependent on communication integrated devices for efficiency, reliability and control. The higher levels of inter-connectivity in the infrastructure and a ubiquitous use of communications have resulted in new types of vulnerabilities which have not been fully covered by the existing defence frameworks. Occurrences such as the 2015 cyber-attack against distribution companies in Ukraine [1] have drawn attention to the field of defence against cyber-threats. The Ukraine attack took many months of infiltration and was successful in compromising the SCADA system and de-energizing a portion of the grid for a few hours. However, the attack itself was discovered almost instantly once implemented. If the attackers had opted for a stealthy attack type, such as FDI attacks, the attackers may have been able to continue attacking for months or years without being detected and the eventual consequences could have been much greater.

FDI attacks, first outlined in [2], involve altering system measurements to corrupt a network operator’s state estimation process and cause negative consequences such as line overloading or outage masking [3]. A comprehensive review of FDI attacks can be found in [4]. FDI attacks need to remain undetected by the network operator to be effective. To this end, FDI attacks compete with bad data detectors (BDD) within state estimation processes. In modern energy management systems (EMS), the BDD at the power system level relies on weighted-least squares (WLS) and chi-squared error testing [5], meaning an attacker needs to structure the attack based on the system model in order to remain undetected. Initial models for FDI attacks assumed full knowledge of the system and full access to meter measurements within the system [2]. An incomplete knowledge attack was introduced in [6], which showed a system could be attacked with only partial knowledge of the system topology and a subset of meter measurements. In [7] the blind FDI attack is introduced, which requires no system knowledge provided the attacker has access to all meters within the attacked grid system. The blind FDI attack uses independent component analyses (ICA) to map the inter-correlations of the visible meter measurements to create an approximation for the power flow model. A more effective version of the attack which utilizes partial susceptibility knowledge was developed in [8], allowing an islanded approach where the visible or ‘high knowledge’ parts of the system could be attacked by the standard-FDI attack while low information areas by the blind approach. Some recent studies enhanced FDI attacks by combining with other forms of attacks, such as denial of service (DoS) attack [9].

In addition, data-driven approaches have recently been applied to FDI attacks, although mostly from the defenders perspective [10][11]. For FDI attack, a data-driven form is suggested in [12], wherein the singular value decomposition is used to construct attack vectors without knowing the system measurement matrix. In [13], two strategies using subspace separation are suggested: one aims to use estimated system subspace to hide attack vectors and another aims to mislead BDD so that non-attacked measurements are removed. These methods allow for admittance values to be estimated but require a large number of historical measurements. In [14], sparse FDI attacks against wide area measurement systems and defense methods are explored. Using historical data to mount FDI by using multiple linear regression model was outlined in [15]. However, the above literature all focus on fixed network topology, while whether and how data-driven approaches can be applied to design FDI attacks under intentional or unintentional topology changes has not yet been investigated.

In fact, as FDI attacks are dependent on the characteristics of the physical system, a body of work has emerged to utilize the physical system to actively defend against the attacks. In particular, MTD is proposed through either transmission switching [16] or admittance perturbation via distributed flexible AC...
transmissions (D-FACTs) devices [17][18] to change physical system topology to proactively drive BDD. An analysis of MTD against FDI attacks is offered in [19] where they prove the susceptibility of isolated state measurements and design an algorithm for branch perturbation selection. Some limitations of MTD were explored in [20]. With the increasing capability of the attackers, there are growing interests in the research community to design new forms of MTD which can hide its existence to the attacker. One of the key state-of-the-art papers in this field is [21], which presents an enhanced hidden MTD model to make the topology change invisible to an attacker via identifying alternative topology and state combinations under the same power flow profile. Whilst this method is clearly effective, it relies on being able to find alternative topology and states to maintain constant power flows, which can be computationally expensive and even infeasible in a system with limited acceptable state ranges.

In this context, this paper examines the vulnerability of current MTD under unsupervised learning-based FDI attacks and develop a new form of stealth MTD to increase system resilience. Our main contributions are twofold:

- On the attacking front, this work introduces a novel new counter-MTD technique. Where previous FDI attacks have been designed against static systems, we seek to offer new attacking considerations in the presence of dynamic systems with MTD. The proposed intelligent attack under zero system knowledge assumption combines unsupervised learning and dimensionality reduction to identify underlying clusters associated with network topology and design the corresponding attack vector. The method is shown to be effective and stealthy against traditional MTD.

- From the defensive perspective, we introduce a new implementation of MTD to drive detection of traditional and intelligent FDI attacks in power systems. The proposed defence strategy combines MTD and physical watermarking concept [22], for the first time, to add a Gaussian watermark into physical plant parameters. As the added watermark mimics the underlying noise of the system, the physical changes driven by MTD stay hidden. The physical watermarking is combined with cumulative error monitoring to spot minor but sustained changes in the system to trigger alarm.

The rest of this paper is organized as follows. The problem formulation and underlying basis for FDI attacks and MTD through topology and parameter changes is outlined in section 2. Section 3 details the design of the proposed intelligent attack, justification for algorithm selection and demonstration of its effectiveness in circumventing MTD. Section 4 proposes the Gaussian style physical watermark in physical system parameters with cumulative error detection approach. Section 5 contains the results and analysis of the different types of MTD as applied to blind FDI attacks and Section 6 concludes the paper.

II. Problem Formulation

A. State Estimation

A static power system problem is considered, consisting of a set of $n$ state variables $\mathbf{x} \in \mathbb{R}^{n \times 1}$ estimated by analysing a set of $m$ meter measurements $\mathbf{z} \in \mathbb{R}^{m \times 1}$ and corresponding error vector $\mathbf{e} \in \mathbb{R}^{m \times 1}$. The non-linear vector function $\mathbf{h}(\cdot)$ relating meter measurements $\mathbf{z}$ to states $\mathbf{h}(\mathbf{x}) = (h_1(\mathbf{x}), h_2(\mathbf{x}), ..., h_m(\mathbf{x}))^T$ is shown by

$$\mathbf{z} = \mathbf{h}(\mathbf{x}) + \mathbf{e}. \quad (1)$$

With real power flow measurements under the non-linear expression defined by

$$P_{ij} = V_i^2 g_{ij} - V_i V_j g_{ij} \cos \Delta \theta_{ij} - V_i V_j h_{ij} \sin \Delta \theta_{ij}. \quad (2)$$

For simplicity and clarity, we first derive the initial formulation and condition based on the linear DC approximation of AC state estimation. A mathematical extension and simulations on original system are then preformed in later sections to demonstrate the applicability of the proposed methods in full AC state estimation.

As a result, the matrix formulation, represented by a linear regression model as a function of the Jacobian $\mathbf{H} \in \mathbb{R}^{m \times n}$ matrix and the state vector, can be expressed as:

$$\mathbf{z} = \mathbf{Hx} + \mathbf{e}. \quad (3)$$

The state estimation problem is to find the best fit estimate of $\hat{\mathbf{x}}$ corresponding to the measured power flow values of $\mathbf{z}$. Under the most widely used estimation approach, the state variables are determined by minimization of a WLS optimization problem as

$$\min J(\mathbf{x}) = (\mathbf{z} - \mathbf{Hx})^T \mathbf{W}(\mathbf{z} - \mathbf{Hx}). \quad (4)$$

$\mathbf{W}$ is a diagonal $m \times m$ matrix consisting of the measurement weights. A solution for a minimal $J(\mathbf{x})$ can be analytically obtained by taking the 1st derivative with respect to $\mathbf{x}$ and solving for 0, yielding $\hat{\mathbf{x}}$ defined by

$$\hat{\mathbf{x}} = (\mathbf{H}^T \mathbf{W})^{-1} \mathbf{H}^T \mathbf{Wz}. \quad (5)$$

B. Bad Data Detection

The current approach in power systems operation for bad data detection is to use the 2-norm of the measurement residual with a detection threshold $\eta$ [23]. The residual $\mathbf{r}$ is defined by the difference between the measured power flow values of $\mathbf{z}$ and the value calculated from the estimated state values $\hat{\mathbf{x}}$ and the known topology matrix $\mathbf{H}$

$$\mathbf{r} = ||\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}||_2. \quad (6)$$

Assuming the errors of state variable $\mathbf{x}$ are random, independent and follow a normal distribution with mean zero and unit $N(0, \sigma^2)$, a chi-squared distribution model $\chi^2_{m-\alpha}$ with
m - n degrees of freedom and confidence interval \( \alpha \) (typically 0.95 or 0.99) can be used to define the detection threshold as

\[
\eta = \sigma \sqrt{\frac{\chi_{m-n,\alpha}}{m-n}}. \tag{7}
\]

If \( r_t > \eta \) BDD alarms will trigger and the system operator will discard the result, removing the elements from the residual calculation with large values and replacing with an appropriate pseudo-measurement, based on historical data.

C. Constructing Attack Vectors

In the case of an infinitely resourced and knowledgeable attack, the attacker can gain full access to the metering infrastructure and change measured power flows in any desired manner. In this case, it is trivial to design the attack to maintain a residual at a given value. The attacker can choose any linear combination of \( \text{H}c \) where \( c \in \mathbb{R}^{m \times 1} \). The vector \( c \) can be selected to have the desired impact on the state vector \( \mathbf{x} \):

\[
\mathbf{z}_a = \mathbf{z} + \mathbf{a} = \mathbf{z} + \text{Hc}. \tag{8}
\]

The 2-norm residual remains unchanged as shown below:

\[
r_a = ||(\mathbf{z} + \mathbf{a}) - \text{H} (\hat{\mathbf{x}} + \mathbf{c})||_2 = ||\mathbf{z} - \text{H}\hat{\mathbf{x}}||_2. \tag{9}
\]

In a more realistic scenario, where the attacker has full access to the metering infrastructure but no understanding of how the network components interconnect or the branch admittance, the attacker has to commit a “blind” form of attack by estimating plausible attack vector models based on historical measurements. One way of achieving this is to utilize Blind Source Separation (BSS) techniques. This scenario has been outlined previously in [7]. The relationship between the state variables in a power system and latent independent variables \( \mathbf{y} \) under a fixed topology \( \text{H} \) can be described by

\[
\mathbf{x} = f(\text{H}, \mathbf{y}). \tag{10}
\]

In practice \( \mathbf{y} \) represents the loads of power system which vary independently while the topology is fixed but other underlying latent variables may exist for some systems. The state vector \( \mathbf{x} \) can be approximated as the first-order coefficient of the Taylor expansion \( \mathbf{A} \) around \( \mathbf{y} \).

\[
\mathbf{x} \approx \mathbf{A} \mathbf{y}. \tag{11}
\]

Returning to the state estimation problem, the system states can then be expressed in terms of load such that

\[
\mathbf{z} \approx \text{H} \mathbf{A} \mathbf{y} + \mathbf{e}. \tag{12}
\]

If the attacker can acquire \( \text{HA} \), an attack vector can be constructed with a value selected for a change in power flows \( \delta \mathbf{y} \) shown by

\[
\mathbf{z}_0 = \mathbf{z} + \text{H} \mathbf{A} \delta \mathbf{y}. \tag{13}
\]

A generalized form of blind source separation \( \mathbf{u} = \mathbf{Gv} \) can be used, where \( \mathbf{u} \) is the vector that can be directly observed, \( \mathbf{G} \) is the fixed vector known as the mixing matrix and \( \mathbf{v} \) the underlying vector of signals. The state estimation can be constructed in an equivalent manner such that:

\[
\mathbf{z} = \text{HAy} = \mathbf{Gy}. \tag{14}
\]

Provided the errors follow a Gaussian distribution and do not contain gross errors, \( \text{HA} \) can be extracted using independent component analysis as shown previously in [7] [24].

D. AC Extension of Blind Attack

Similar to the DC attack, AC FDI attacks must satisfy the system model to remain hidden such that

\[
\mathbf{z}_a = \mathbf{z} + \mathbf{a} = \mathbf{h} (\mathbf{x} + \mathbf{c}). \tag{15}
\]

This can be done without topology information either using the geometric approach [25] or a historical measurement based replay approach. Chin et al showed that where the vector angle between the normal power flows and attacking vector was defined by

\[
\mathbf{z}^T \mathbf{a} = \cos (\phi) \tag{16}
\]

the attack can bypass AC detection provided the vector space angle between the attacking vector and measurement vector was close to zero such that

\[
\mathbf{z}^T \mathbf{z}_a = 1. \tag{17}
\]

Under these considerations, a regression model can be extracted to attack the system. Alternatively, in the case of limited information, the attacker can implement a replay style attack which reuses a previous vector from historical measurements such that

\[
\mathbf{z}'_q = \mathbf{z}_{t-q} \tag{18}
\]

where \( q \) is used to denote a vector from a previous time period. Our AC simulations were built with this replay case in mind, but it should be noted both methods are susceptible to conventional MTD.

E. MTD through Topology Changes

Under AC state estimation, system measurements will consist of real power flows defined by Equation [2] and reactive power by

\[
Q_{ij} = -V_i^2(b_{ij} + b_{ij}^\theta) + V_i V_j g_{ij} \cos \Delta \theta_{ij} - V_i V_j b_{ij} \sin \Delta \theta_{ij}. \tag{19}
\]

For real power residual, error at the individual measurement level will be the difference between the measured flows and estimated value from the system model such that real power residual can be expressed as

\[
P_{ij} = -P_{ij}^o + V_i^2 g_{ij} - V_i V_j g_{ij} \cos \Delta \theta_{ij} - V_i V_j b_{ij} \sin \Delta \theta_{ij}. \tag{20}
\]

and reactive power flow residual can be expressed as
\[ r_{ij}^Q = -Q_i^{\text{ref}} - V_i^2(b_{ij} + p_{ij}^\text{th}) + V_iV_jg_{ij}\cos\Delta\theta_{ij} - V_iV_jb_{ij}\sin\Delta\theta_{ij}. \] (21)

In the AC state estimation, MTD can employ resistive as well as inductive components to introduce change. Alternatively, the SO can aim to force a state of non-convergence in the case of FDI which is done by violating the non-convergence criteria of the Newton-Raphson principle for power systems. Again, the alarm criteria will be the 2-norm value of the residual vector calculated by

\[ r_{nc} = \|z - h(\tilde{x})\|_2. \] (22)

We derive here the analytical expression of the impact on residual of topology change for a linear system under attack vector \( a = Hc \). Using the WLS formulation, \( r_n \) can be expressed as

\[ r_n = \|z + Hc - H(H^TWH)^{-1}H^TW(z + Hc)\|_2. \] (23)

The attacker is assumed to have static topology knowledge and construct the injected attack vector \( z_a \) as a function of the original topology \( H_o \). The new topology with MTD applied is \( H_n \), which is only known by the SO. As a result, the measurement vector under attack \( z_n \) will be

\[ z_n = z + H_n c. \] (24)

The SO estimates \( \tilde{x} \) via the WLS minimization using the visible \( z_n \) and \( H_n \). The min error estimate of \( \tilde{x}_n \) will utilize the new topology \( H_n \) while the attack vector is developed based on the old topology \( H_o \). Consequently, the new residual will be a product of the attack vector based on old topology \( H_o c \) and the WLS estimation based on the new topology as

\[ r_n = \|z + H_n c - H_n(H_n^TWH_n)^{-1}H_n^TW(z + H_n c)\|_2. \] (25)

Defining WLS minimization factor for the new topology as \( F_n \), which is fixed for a given topology as \( F_n = (H_n^TWH_n)^{-1}H_n^TW \), the residual 2-norm can be rewritten as

\[ r_n = \|z + H_n c - H_n F_n(z + H_n c)\|_2. \] (26)

Considering the old topology \( H_o \) as a function of the new and system change \( H_n + \Delta H \), the residual in terms of the new topology can hence be calculated as

\[ r_n = \|z + (H_n + \Delta H)c - H_n F_n(z + (H_n + \Delta H)c)\|_2. \] (27)

\( H_n F_n H_n \) is the idempotent matrix of \( H \) and therefore \( H_n F_n H_n c = H_n c \) the expression can be rearranged into

\[ r_n = \|(1 - H_n F_n)z + (1 - H_n F_n)\Delta Hc\|_2. \] (28)

As shown in Equation (28) any \( \Delta H \) will change the residual value \( r_n \). The aim of defender is to select a value for \( \Delta H \) such that under attack vector \( H_n c \), the new residual exceeds the alarm criteria (usually chi-squared criteria) \( r_n > \sigma \sqrt{\chi_{m-n,\alpha}} \).

### III. Clustering to Circumvent MTD

This section investigates how data-driven approach can be applied to explore the vulnerability of existing MTD. In particular, an efficient method is proposed to identify changes in the network caused by the implementation of DFACTS or switching through analysing the resultant power flow profiles. By doing so, the attacker can ensure only data points corresponding to the current configuration are used to create the blind attack. The proposed attack flow follows:

1. Observations of historical power flows are clustered into groups.
2. The clustering algorithm identifies the current power flow set to find corresponding measurements for the attack model.
3. The blind attack model is developed using only the data corresponding to the current power flow profile cluster.

This process is illustrated and compared with the normal blind attack in Figure 1. To achieve this, we propose a combination of density based spatial clustering of application with noise (DBSCAN) for the clustering and data preprocessing via T-distributed stochastic neighbour embedding (T-SNE) for dimensionality reduction.

#### A. Algorithm Design Considerations

1. **Curse of Dimensionality:** Due to the blind nature of the attack, no prior classification is possible and therefore an unsupervised learning method is required. Unsupervised learning methods often suffer from the curse of dimensionality. For example, hierarchical clustering has a time complexity of \( O(n^3) \) making it highly computationally intensive for large systems. To address this, we have opted to implement a layer of dimensionality reduction using T-SNE before applying the clustering algorithm. Native T-SNE itself has a time complexity of \( O(n^2) \) but can be reduced to \( O(n) \) by using optimization techniques as discussed in [26]. The brunt of the computational load is therefore taken by T-SNE which reduces the measurements of the network power flows into 2-dimensional space. Native K-means has a similar time complexity to native T-SNE but was unsuitable due to the requirement to pre-define the numbers of clusters.

2. **Choice of Unsupervised Learning Method:** DBSCAN is easy to implement and has been shown to have good benchmark performance when compared with a number of
different options [27]. Considerations on time complexity were also made. DBSCAN has a time complexity of $O(n^2)$ but this can be reduced to $O(n \log n)$ with parameter optimisation [28]. However as the data is dimension reduced before hand, this time complexity should not cause significant increases in processing time.

B. T-Distributed Stochastic Neighbours

High dimensionality in data sets can make effective analysis and visualization of data trends difficult. It can also significantly increase computational requirements [29]. T-SNE is a form of dimensionality reduction which works by constructing probability distributions over pairs of objects containing high dimensionality [30].

Consider a set of $N$ high dimension objects, $d_i$ and $d_j$ are two points within this set. $\sigma_i$ is the variance of the Gaussian centred on data point $d_i$. The closeness of these data points is defined by the conditional probability $p_{ji}$ that point $d_j$ would select $d_i$ as a neighbour given that the neighbours are picked proportionately to a Gaussian centred around $d_j$. This is given by

$$p_{ji} = \frac{\exp(-||d_i - d_j||^2 / 2\sigma_i^2)}{\sum_{k:j} \exp(-||d_i - d_j||^2 / 2\sigma_j^2)}.$$ (29)

The aim of T-SNE is to reduce these points into their low dimensional counterparts $g_j$ and $g_i$. These have an equivalent conditional probability $q_{ji}$ defined by

$$q_{ji} = \frac{\exp(-||g_i - g_j||^2)}{\sum_{k:i} \exp(-||g_i - g_j||^2)}.$$ (30)

If the map points $g_j$ and $g_i$ correctly model the similarity between the high dimensional sets, the conditional probabilities $p_{ji}$ and $q_{ji}$ will be equal. The positions of $g_i$ and $g_j$ are determined via gradient descent between the distributions $p$ and $q$, and is used to minimize the Kullback-Leiber divergence via cost function $C$ [31] shown by

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{ji} \log \frac{p_{ji}}{q_{ji}},$$ (31)

where $P_i$ is the conditional probability distribution over all data points given data point $d_i$ and $Q_i$ is the conditional probability distribution over every other map point, given map point $g_i$.

C. Intelligent Blind FDI Attack

This sub-section details the proposed intelligent blind FDI, as outlined in Algorithm 1. Once the attacker obtained adequate amount of measurement data, the attacker can initiate the attack algorithm. When the latest measurement data arrive, T-SNE is firstly applied for dimensionality reduction of the sets of power flow observations into a two dimensional space. The reduced form of the data set is then clustered via the DBSCAN algorithm into distinct subgroups of like measurements and the one corresponding to the current system topology is identified. The mixing matrix is subsequently derived based on this subgroup of data by using independent component analysis as per the normal blind attack. A vector of false data $z_f$ containing the desired attack bias will be then calculated based on the mixing matrix.

Algorithm 1 DBSCAN Blind-ICA attack

**Input:** A set of power flow observations $z_{obs}$
- $Y = \text{tsne} (z \text{ obs})$ Dimensionality reduction
- $idx = \text{dbscan}(Y, mpts, \epsilon)$ Cluster power flows

For $i = 1:|\text{length(unique(idx))}|$
- $j = [j, idx] \% \text{ Assign pf to cell}$
- $A[i] = j(j(:)=i,:) \% \text{ Assign obvs to cell}$
- $c = j(\text{end}) \% \text{ check what profile current z is}$
- $z = A(c) \% \text{ Select only corresponding z measurements}$
- $HA = \text{FastICA}(z) \% \text{ Run fastica for HA}$
- $z_f = z + HA \delta y \% \text{ Apply attack vector}$

**Output:** false data $z_f$

D. Performance Analysis

A case study is carried out on a system with 14 lines equipped with D-FACTS for MTD. As shown in Figure 2 the proposed algorithm successfully identify and cluster the potential topology sets, although only minor changes on topology (1% of base admittance) are applied. The computational performance of T-SNE and DBSCAN for different IEEE test cases systems is shown in Figure 3 and compared with hierarchical clustering with embedded cluster evaluation. The case studies are performed for 1000 sets of observations. We note that for small scale systems (such as the 5-bus case) the computational performance are similar but as the system becomes larger, the time to completion grows quickly for hierarchical clustering.

IV. PHYSICAL GAUSSIAN WATERMARKING WITH CUSUM

While physical watermarking has not been applied in the power system space, the concept has been proposed in control
systems such as in [22] where a watermark is added into a LQG-based control signals to drive detection. However, the papers in these areas aren’t true ‘physical’ watermarks as they only change signal parameter dependencies and not the underlying physical plant itself. At the same time, it should be noted that while MTD in the form of D-Facts control to change system topology has been explored, the use of watermarks in combination with MTD has not been investigated and there is an opportunity to incorporate a true physical watermark into the system plant to enhance the system security.

Previously, topology perturbation and transmission switching have been proposed as methods to drive detection of FDI attacks [17] [21]. These methods implement significant changes to line admittance as required by the change needed in residual (typically around 10-20% for D-Facts based changes), which may not only lead to interruption on system operation, but also provide opportunities for the data-driven attack to spot the existence of MTD and counter it. As the existence of clustering based intelligent attack proposed in Section III, it is crucial that the deployment of MTD can remain hidden to the attacker.

In this context, this section proposes a novel method to achieve this by combining MTD with physical watermarking, which makes the MTD itself indistinguishable from the noise profile of the system, and monitoring sequential errors for long-run trends by using cumulative summed monitoring (CUSUM). CUSUM is a sequential analysis technique which monitors for change detection over a number of measurements. Samples taken from the process are assigned a weighting and summed to monitor change detection. In this case, we will monitor the measured residual $r$ under MTD defined by

$$CEM_t = \sum_{j=1}^{t} r_j - T.$$  \hspace{1cm} (32)

where $CEM$ is the decision statistic, $T$ is the target value of residual dictated by monitoring the statistic under normal conditions and $t$ is the number of periods in a measurement set, with upper and lower control limits $CEM^+_{t}$ and $CEM^-_{t}$. As $r$ is an absolute value, the lower bound $CEM^-_{t}$ will be 0. $CEM^+_{t}$ can be selected based on engineering judgement from prior observations. Usually the upper bound can be defined in terms of the residual variance and mean value under no attack:

$$CEM^+_{t} = \bar{r} + B\sigma_r.$$  \hspace{1cm} (33)

where $B$ is defined by the user based on previous observations and minimising type 2 error.

The proposed defence strategy introduces these minor errors by using D-FACTS devices to alter the line admittance by a vector $w$. The size of admittance changes applied to each line is based on the output from a pseudo random number generator (PRNG), the seed value of which is only known by the network operator. This can be achieved with existing technology via a Unified Power Flow Controller (UPFC) in combination with a processing unit. The watermark may be applied selectively such that

$$w_m \in \{0, N(0, p)\}.$$  \hspace{1cm} (34)

where $p$ is the max change applied to the branch admittance.

The resulting power flow profile under physical watermarking will be equal to

$$z_w = (H + w)x + e.$$  \hspace{1cm} (35)

where $w$ represents the vector of admittance changes applied to branches and is known to the SO.

The impact of applying a Gaussian style watermark in physical system parameters is shown in Figure 4. Compared with direct binary perturbation, the proposed MTD show similar profile as underline noise and make it extremely hard for clustering algorithm to identify the existence of MTD or to counter it.

The key advantages of the proposed defence mechanism can be summarised as below, which will be validated in the next session:

1) As the proposed MTD is on magnitude with the noise levels, the change in power flow observations resulting from the MTD becomes difficult to be identified. Therefore, MTD stays stealthy to the attacker.
2) Due to the stealthiness of the proposed MTD, it significantly increases the chance of the detection of FDI attack and is specifically resilient to intelligent attack types such as the proposed DBSCAN blind-ICA attack.

3) The significantly-reduced magnitude of topology changes lead to less interruptions on the system stability and economic operation.

V. RESULTS AND ANALYSIS

This section assesses the performance of the proposed intelligent blind FDI in the presence of different forms of MTD on the standard IEEE 14-Bus and IEEE 118-bus test systems [32]. All simulations were implemented using the MATPOWER toolbox in MATLAB [33] and performed using Intel Core i7-7820X CPU with 64GB of ram running on a Windows 10 system.

A. Model Assumptions

The priority of this section is to capture the change in detection between the blind FDI technique and the proposed intelligent attack under different types of MTD. Some assumptions have been made across all simulations:

- Uncoloured Gaussian noise error of 1% noise-to-signal was added to meter values as error $e$ as seen previously in [34].
- A steady load assumption is made with load variation of around 0.1% for initial simulations in line with other state-of-the-art works in this field such as [21]. Additional case studies were performed with multiple load profiles.
- A minimum number of observations of 250 is assumed initially which rises to 1000 sequentially over the course of the simulation.

B. Transmission Switching

Figure 5 shows the impact of transmission line switching on the blind FDI attack and DBSCAN attack for the 14 bus and 118 bus cases. For the standard blind FDI attack, transmission switching is highly effective at introducing residual errors and driving alarms. With a single line switching the detection is 100% for the standard blind FDI attack. However, these large changes in the system flows make it easy for an attacker to identify the MTD. Compared with the standard attack, the DBSCAN attack out-performs the standard blind FDI whenever MTD is used. Detection remained low (less than 1%) with up to 15 lines being switched in/out across the network at different times. Even with 16 possible topologies in use the detection remained under 3%. Transmission switching is unlikely to be used for the sole purpose of attack detection due to the significant impact on the system operability.

C. Admittance Perturbation

Admittance perturbation is the most commonly proposed method of MTD for power systems in the current literature. This sub-section implements an admittance perturbation defence against the typical blind FDI attack and the proposed DBSCAN version. A known quantity of admittance is injected at given lines in the power system. The system operator is expecting to see the change in admittance reflected in the resulting power flows. If the attacker is unaware and does not reflect the new admittance in their attacking vector, the residual will increase significantly and BDD will be triggered. The results of admittance perturbation on detection of the standard and DBSCAN blind FDI attack are shown in Figure 6.

System models with branch admittance perturbations of 10% were implemented. The standard blind FDI attack performs poorly against this form of MTD. For a single line at 10% perturbation a detection level of over 95% is achieved. The detection rates for the DBSCAN informed attack were consistently low. This is due to the distinctive clusters of power flows emerging under the steady loads assumption.

D. Physical Gaussian Watermarking with Cumulative Errors

A novel implementation of MTD technique is trialled here. The admittance profile is varied using an PRNG with a profile equivalent to the underlying noise of the system. As we have used a 1% noise for our simulations the $p$ value is set equal to 1% to ensure that this profile is not visible to the attacker. This
is combined with cumulative error monitoring watching for sustained increased errors over 10 measurements. As shown in Figure 7 under the DBSCAN blind FDI attack, the CUSUM Gaussian watermark shows significant improvements. As additional lines are added these detection rates are close to 100% compared with under 10% for standard admittance perturbation. The is due to the difficulty DBSCAN algorithm has in identifying clusters for MTD on magnitude and identical to noise profile of the system. However, it should also be noted the increase in type-2 error resulting from this method with a trade off between lower alarm thresholds for cumulative errors and detection rates as shown in Figure 7. This cumulative approach requires multiple measurements which potentially could lead to the attacker having additional time to attack before being caught. Therefore, there exists a trade off between the speed to spot attacks and the magnitude of the added watermark. Figure 8 illustrates this for the 118-bus and 14-bus systems where for a lower level of added watermark, a larger number of measurement points are needed to break the 4 standard deviation threshold, while the gains between using a 2% and 10% watermark are marginal.

E. Load Variance Impact

In previous case studies, the simulations have been performed under steady load assumptions [21]. This session investigates the impact of large load variation on the performance of the DBSCAN attack. Case studies were performed on data sets containing ten load levels. As shown in Figure 9 large load variations reduce the effectiveness of the DBSCAN attack due to the increasing challenge to cluster the topology changes under high varying load. The rates of detection for both MTDs increase significantly and rise steadily with the number of lines perturbed. Once topology perturbation/switching is being applied to five or more lines detection rises above 40%. However, this impact can be mitigated by collecting more previous measurements. As shown in Figure 9, the high data point with 5k observations keeps the detection rate well below 40% even when 10 lines are perturbed.

F. Blind AC Replay Attack

We have implemented our clustering approach with a blind replay style attack against an AC state estimation. Under this attack the attacker attempts to inject a previously observed vector. The attacker is competing with MTD and wants to select the replay vector from a pool of values only containing those using the same topology configuration. In Figure 10 we can see that the distinctive cluster relationship exists within the AC model as shown previously for DC. Figure 11 demonstrates that the proposed pre-clustering algorithm preforms well in AC state estimation provided a large number of samples received. The non-linearity in the AC model significantly reduce the correction rate of clustering but increasing the number of observations allows good performance for the AC model.

VI. Conclusions & Further Work

This paper, for the first time, investigate how unsupervised learning and dimensionality reduction can be applied in blind FDI attacks to exploit the venerability of current forms of MTD. By incorporating a combination of T-SNE dimensionality reduction and the DBSCAN clustering algorithm, power flow observations can be clustered into their relative topology.
Further work on this topic entails enhancing the blind FDI model to model for other scenarios i.e. subset attacks, optimal design of physical watermarking scheme and analysing the effects of MTD on topological discovery techniques.

REFERENCES

[1] G. Liang, S. R. Weller, J. Zhao, F. Luo, and Z. Y. Dong, “The 2015 Ukraine Blackout: Implications for False Data Injection Attacks,” IEEE Transactions on Power Systems, 2017.

[2] Y. Liu, P. Ning, and M. K. Reiter, “False data injection attacks against state estimation in electric power grids,” ACM Transactions on Information and System Security, 2011.

[3] X. Liu, Z. Li, X. Liu, and Z. Li, “Masking Transmission Line Outages via False Data Injection Attacks,” IEEE Transactions on Information Forensics and Security, vol. 11, no. 7, pp. 1592–1602, 2016.

[4] R. Deng, G. Xiao, R. Lu, H. Liang, and A. V. Vasilakos, “False data injection on state estimation in power systems-attacks, impacts, and defense: A survey,” IEEE Transactions on Industrial Informatics, 2017.

[5] 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 Transactions on Smart Grid, 2017.

[6] M. A. Rahman and H. Mohsenian-Rad, “False data injection attacks with incomplete information against smart power grids,” in GLOBECOM - IEEE Global Telecommunications Conference, 2012.

[7] M. Esmailifalak, H. Nguyen, R. Zheng, and Z. Han, “Stealth false data injection using independent component analysis in smart grid,” in 2011 IEEE International Conference on Smart Grid Communications, SmartGridComm 2011, 2011.

[8] R. Deng and H. Liang, “False Data Injection Attacks with Limited Susceptance Information and New Countermeasures in Smart Grid,” IEEE Transactions on Industrial Informatics, vol. 15, no. 3, pp. 1619–1628, 2019.

[9] M. N. Kurt, Y. Yilmaz, and X. Wang, “Real-Time Detection of Hybrid and Stealthy Cyber-Attacks in Smart Grid,” IEEE Transactions on Information Forensics and Security, vol. 14, no. 2, pp. 498–513, 2018.

[10] Y. Wang, M. M. Amin, J. Fu, and H. B. Moussa, “A novel data analytical approach for false data injection cyber-physical attack mitigation in smart grids,” IEEE Access, vol. 5, pp. 26022–26033, 2017.

[11] S. Ahmed, Y. Lee, S. H. Hyun, and I. Koo, “Unsupervised Machine Learning-Based Detection of Covert Data Integrity Assault in Smart Grid Networks Utilizing Isolation Forest,” IEEE Transactions on Information Forensics and Security, vol. 14, no. 10, pp. 2765–2777, 2019.

[12] F. Wen and W. Liu, “An Efficient Data-Driven False Data Injection Attack in Smart Grids,” in International Conference on Digital Signal Processing, DSP, vol. 2018-November. Institute of Electrical and Electronics Engineers Inc., 2019.

[13] J. Kim, L. Tong, and R. J. Thomas, “Subspace methods for data attack on state estimation: A data driven approach,” IEEE Transactions on Signal Processing, vol. 63, no. 5, pp. 1102–1114, 2015.

[14] J. Hao, R. J. Piechocki, D. Kaleshi, W. H. Chin, and Z. Fan, “Sparse Malicious False Data Injection Attacks and Defense Mechanisms in Smart Grids,” IEEE Transactions on Industrial Informatics, vol. 11, no. 5, pp. 1198–1209, 2015.

[15] 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 Transactions on Power Systems, 2018.

[16] 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.

[17] K. L. Morrow, E. Heine, K. M. Rogers, R. B. Bobba, and T. J. Overbye, “Topology Perturbation for Detecting Malicious Data Injection,” in 45th Hawaii International Conference on System Sciences, 2012.

[18] C. Liu, M. Zhou, J. Wu, C. Long, A. Farraj, E. Hammad, and D. Kundur, “Reactive Perturbation for Enhancing Detection of FDI Attacks in Power System State Estimation,” in 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2017.

[19] Z. Zhang, R. Deng, D. Yao, P. Cheng, and J. Chen, “Analysis of Moving Target Defense Against False Data Injection Attacks on Power Grid,” IEEE Transactions on Information Forensics and Security, 2019.

[20] B. Li, G. Xiao, R. Lu, R. Deng, and H. Bao, “On Feasibility and Limitations of Detecting False Data Injection Attacks on Power Grid State Estimation Using D-FACTS Devices,” IEEE Transactions on Industrial Informatics, pp. 1–1, 2019.

[21] J. Tian, R. Tan, X. Guan, and T. Liu, “Enhanced hidden moving target defense in smart grids,” IEEE Transactions on Smart Grid, vol. 10, no. 2, pp. 2208–2223, 2019.

[22] Y. Mo, S. Weerakkody, and B. Sinopoli, “Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs,” IEEE Control Systems, vol. 35, no. 1, pp. 93–109, 2015.

[23] A. Monticelli, State Estimation in Electric Power Systems: A Generalized Approach, 1999.

[24] Z. H. Yu and W. L. Chin, “Blind False Data Injection Attack Using PCA Approximation Method in Smart Grid,” IEEE Transactions on Smart Grid, 2015.
[25] W. L. Chin, C. H. Lee, and T. Jiang, “Blind false data attacks against ac state estimation based on geometric approach in smart grid communications,” *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6298–6306, 11 2018.

[26] N. Pezzotti, B. P. Lelieveldt, L. Van Der Maaten, T. Höltt, E. Eisemann, and A. Vilanova, “Approximated and user steerable tSNE for progressive visual analytics,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 7, pp. 1739–1752, 7 2017.

[27] J. M. Dudik, A. Kuros, J. L. Coyle, and E. Sejdić, “A comparative analysis of DBSCAN, K-means, and quadratic variation algorithms for automatic identification of swallows from swallowing accelerometry signals,” *Computers in Biology and Medicine*, vol. 59, pp. 10–18, 4 2015.

[28] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” Tech. Rep., 1996. [Online]. Available: www.aaai.org

[29] H. Xie, J. Li, and H. Xue, “A survey of dimensionality reduction techniques based on random projection,” 6 2017. [Online]. Available: http://arxiv.org/abs/1706.04371

[30] L. Van Der Maaten and G. Hinton, “Visualizing Data using t-SNE,” Tech. Rep., 2008.

[31] D. G. E Silva, M. Jino, and B. T. De Abreu, “Machine learning methods and asymmetric cost function to estimate execution effort of software testing,” in *ICST 2010 - 3rd International Conference on Software Testing, Verification and Validation*, 2010, pp. 275–284.

[32] R. Christie, “Power Systems Test Case Archive.” [Online]. Available: http://labs.ece.uw.edu/pstca/

[33] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, “MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education,” *IEEE Transactions on Power Systems*, 2011.

[34] K. Khanna, B. Panigrahi, and A. Joshi, “AI based approach to identify compromised meters in data integrity attacks on smart grid,” *IET Generation, Transmission & Distribution*, vol. 12, no. 5, pp. 1052 – 1066, 2018.