Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

TRUSTWORTHY ARTIFICIAL INTELLIGENCE FRAMEWORK FOR PROACTIVE DETECTION AND RISK EXPLANATION OF CYBER ATTACKS IN SMART GRID

Md. Shirajum Munir
Sachin Shetty

Virginia Modeling, Analysis, and Simulation Center
Old Dominion University
1030 University Blvd,
Suffolk, VA 23435, USA.

Danda B. Rawat

Department of Electrical and Computer Science
Howard University
2400 Sixth Street NW,
Washington, D.C. 20059, USA.

ABSTRACT

The rapid growth of distributed energy resources (DERs), such as renewable energy sources, generators, consumers, and prosumers in the smart grid infrastructure, poses significant cybersecurity and trust challenges to the grid controller. Consequently, it is crucial to identify adversarial tactics and measure the strength of the attacker's DER. To enable a trustworthy smart grid controller, this work investigates a trustworthy artificial intelligence (AI) mechanism for proactive identification and explanation of the cyber risk caused by the control/status message of DERs. Thus, proposing and developing a trustworthy AI framework to facilitate the deployment of any AI algorithms for detecting potential cyber threats and analyzing root causes based on Shapley value interpretation while dynamically quantifying the risk of an attack based on Ward's minimum variance formula. The experiment with a state-of-the-art dataset establishes the proposed framework as a trustworthy AI by fulfilling the capabilities of reliability, fairness, explainability, transparency, reproducibility, and accountability.

1 INTRODUCTION

The smart grid infrastructure is expected to deploy huge amounts of distributed energy resources (DERs) such as renewable energy sources, consumers, prosumers, and so on to meet the goal of around 40% (U.S. Energy Information Administration: Annual Energy Outlook 2023 with Projections to 2050) cost reduction by 2050. Thus, the operation of such deployed DERs in a smart grid significantly increases the risk of cyber-attacks through their control/status messages (Nafees et al. 2023; Bitirgen and Filik 2023; Karimipour et al. 2019; Kurt et al. 2019). As a result, it is critical to recognize adversarial strategies and evaluate the impact of the attacker's DER. Therefore, a trustworthy smart grid controller must monitor and manage each DERs operation for the entire smart-grid cyber-physical systems.

To boost confidence in smart grid controllers and enable secure power transactions between DERs, it is crucial to develop a trustworthy artificial intelligence (AI) (Dwivedi et al. 2023; Wing 2021; Floridi 2019; Chatila et al. 2021; Liang et al. 2022; Munir et al. 2023) mechanism for detecting potential cyber threats and measuring the severity of the risk. However, in order to establish trustworthy AI mechanisms for smart grid controllers, several technical metrics such as reliability, fairness, explainability, transparency, reproducibility, and accountability on threat detection must be needed to be fulfilled.

In this work, we address the following research challenges:

- How to proactively detect the potential cyber threat in a smart grid environment, where millions of DERs are connected and controlled by the smart grid controller?

979-8-3503-6966-3/23/$31.00 ©2023 IEEE

- In order to guarantee a secure power transaction, how can the root cause of a possible cyber attack be identified in a smart grid controller?
- When there are no apparent indicators to differentiate among types of attacks, how can the grid controller distinguish between the various attack types and measure the risk?

To address the above research challenges, in this work, a trustworthy artificial intelligence (AI) technique is studied for the purpose of proactively identifying and explaining the cyber risk brought on by the control/status message of DERs. We summarize our key contributions as follows:

- First, we design a system model of a trustworthy smart grid controller for assuring a secure power transmission among the distributed energy resources. Then, we formulate a decision problem that can detect the potential cyber threat in an adaptive smart grid environment while analyzing the root cause of such threat by determining the contribution among the features of a status message.
- Second, we propose a trustworthy artificial intelligence framework to solve the formulated decision problem in smart grid controllers. In particular, we design an AI pipeline that is capable of facilitating the deployment of any AI algorithms for potential cyber threat detection, can analyze root causes based on Shapley (Shapley et al. 1953; Munir et al. 2022; Dubey 1975; Lundberg and Lee 2017) value interpretation, and dynamically quantifying the risk of an attack based on Ward's minimum variance (Ward Jr 1963) formula.
- Third, we implement the proposed framework in a simulation environment and tested it with the state-of-the-art cyber-physical system SCADA WUSTL-IIOT-2018 (Teixeira, M., Zolanvari M., and Jain R. 2018).
- Finally, we do the analysis of the implemented framework in terms of reliability, transparency, fairness, accountability, and explainability of the proposed trustworthy artificial intelligence framework. We have found that the proposed framework can detect potential attacks with at least 99% reliability while ensuring transparency, fairness, and explainability by analyzing the contribution of the feature on decisions and quantifying the risk of each potential attack.

The rest of the paper is organized as follows. Section 2 discusses some of the interesting related work. The system model and problem formulation of the considered trustworthy smart grid controller are described in Section 3. The proposed trustworthy AI framework is presented in Section 4 and experimental analysis is discussed in Section 5. Finally, we conclude our discussion in Section 6. A summary of notations is presented in Table 1.

2 RELATED WORK

The problem of cyber attack detection for smart grid infrastructure has been studied in (Karimipour et al. 2019; Kurt et al. 2019; Sakhnini et al. 2019; Munir et al. 2021; Bitirgen and Filik 2023). In (Karimipour et al. 2019), the authors proposed an unsupervised learning-based anomaly detection by measuring the statistical correlation among features. The work in (Kurt et al. 2019) studied a partially observable Markov decision process for cyber-attack detection and proposed a model free reinforcement learning (RL) to solve it. The authors in (Sakhnini et al. 2019) investigated the problem of false data injection (FDI) attack detection for the smart grid by deploying a support vector machine (SVM), K nearest neighbor (KNN), and artificial neural network (ANN)-based supervised learning mechanism. Further, the authors in (Munir et al. 2021) proposed a data-informed policy-based model-free RL scheme for detecting cyber threats and controlling DERs for connected or disconnected from the main grid based on the attack situation. Recently, the authors (Bitirgen and Filik 2023) proposed a hybrid deep learning model by combining particle swarm optimization (PSO) and convolutional neural networks-long short-term memory (CNN-LSTM) for FDI detection in the smart grid environment.

However, these works (Karimipour et al. 2019; Kurt et al. 2019; Sakhnini et al. 2019; Munir et al. 2021; Bitirgen and Filik 2023) do not investigate the problem of root cause analysis of a particular cyber-attack nor do they account the quantify the potential risk, when there are no apparent indicators to differentiate among types of attacks. Dealing with explainability, reliability, and fairness in the detection of cyber attacks and measuring cyber risk with fairness, transparency, and accountability is challenging due to the intrinsic nature of millions of distinct DERs in smart grid cyber-physical systems. Therefore, in this work, we propose a trustworthy AI framework that can proactively recognize and explain the cyber risk posed by the control/status message of DERs.

3 SYSTEM MODEL AND PROBLEM FORMULATION OF TRUSTWORTHY SMART GRID OPERATION

3.1 System Model of Trustworthy Smart Grid Controller

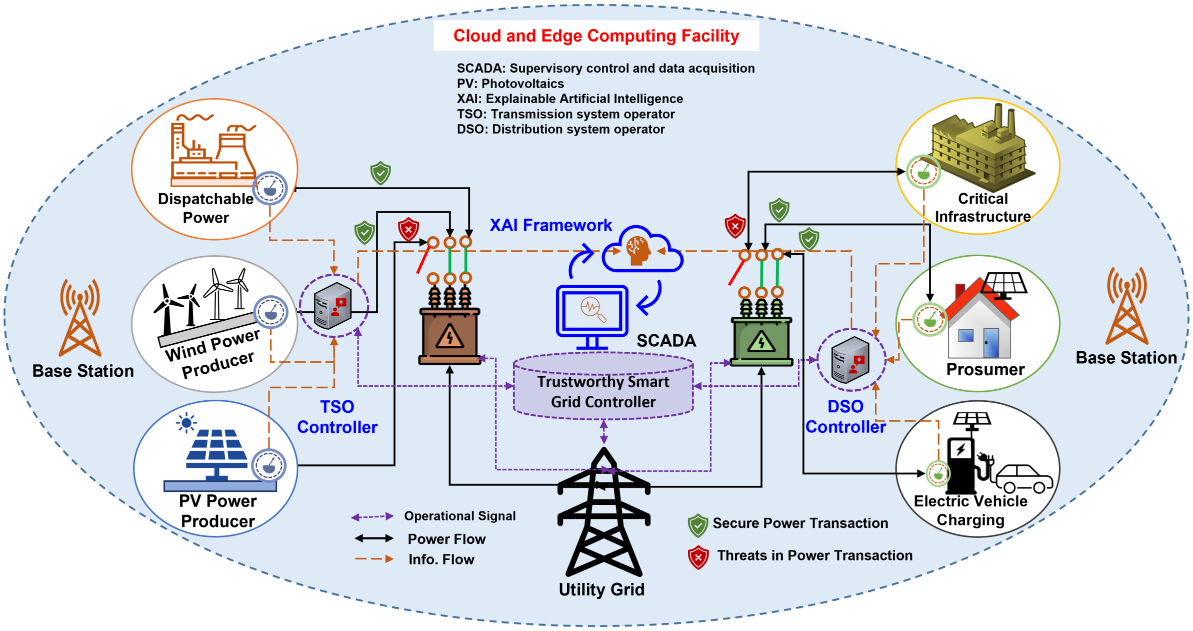

Cloud and Edge Computing Facility
SCADA: Supervisory control and data acquisition
PV: Photovoltaics
XAI: Explainable Artificial Intelligence
TSO: Transmission system operator
DSO: Distribution system operator
Dispatchable Power
Wind Power Producer
PV Power Producer
Base Station
TSO Controller
XAI Framework
SCADA
Trustworthy Smart Grid Controller
DSO Controller
Critical Infrastructure
Prosumer
Base Station
Electric Vehicle Charging
Utility Grid
Operational Signal
Power Flow
Info. Flow
Secure Power Transaction
Threats in Power Transaction

Figure 1: A system model of trustworthy smart grid controller.

We consider a smart grid environment, where a set $\mathscr{D} = \{1,2,\ldots,d,\ldots,D\}$ of distinct distributed energy resources (DERs) (Nafees et al. 2023; Munir et al. 2021; Munir et al. 2022) such as dispatchable power sources, wind sources, photovoltaics sources, prosumer, critical infrastructure consumers, electric vehicle (EV) charging station, and so on are deployed (as seen in Figure 1). The transmission system operator (TSO) controller transfers energy from source DERs to the distribution system operator (DSO). The DSO controller is responsible to distribute energy among the consumer DERs such as critical infrastructure, EV charging, prosumers, and others. Considering a *trustworthy smart grid controller (TSGC)* that can monitor and control the activities of each DER for the entire smart grid cyber-physical systems. We assume that, physically, the TSGC is an entity of the supervisory control and data acquisition (SCADA) system of the considered smart grid infrastructure.

In the considered system model, three kinds of communication and energy flows occur power transmission, information flow, and operational message flow. In Figure 1, the yellow dashed line indicates information flow, the black solid line represents energy flow, and the purple dashed line presents operational signals among the DERs.

In general (Teixeira, M., Zolanvari M., and Jain R. 2018; Munir et al. 2021), the at first control/status messages are exchanged between the generator DERs and TSO controller before transacting energy to the DSO. Similarly, control/status messages are exchanged between DSO controllers and consumer DERs

Table 1: Summary of Notations.

Notation	Description
$\mathscr{D}$	Set of distributed energy resources (DERs)
$d \in \mathscr{D}$	Each control/status messages
$\mathscr{X} = \{0,1,2,\ldots,x,\ldots,X\}$	Set of attack types including trusted message
M	No. of features
$y_d \in \mathbf{Y}$	Threat detection decisions variable
$\Upsilon_m \in \boldsymbol{\Upsilon}$	Shapley coefficient (root cause)
$g(.)$	Risk of severity
z_d	Tuple of features
ω_M	Weight of the model for M features
Z^t	Total number of control/status messages at time slot $t \in T$

before distributing the energy to the consumer DERs and vice versa. However, it is essential to assure trust in DERs before establishing energy flow or transaction while control messages may carry threats in the considered smart grid. In this system model, we consider five types of potential cyber threats: 1) port scanner, 2) address scan, 3) device identification, 4) aggressive mode, and 5) exploit (Teixeira, M., Zolanvari M., and Jain R. 2018; Munir et al. 2021) in a DER control/status message $d \in \mathscr{D}$. Therefore, we define a set $\mathscr{X} = \{0,1,2,\ldots,x,\ldots,X\}$ that includes X cyber threats and $x = 0$ represents trusted control/status message. Thus, X cyber threats can occur in smart grid cyber-physical systems and are initiated by the DERs $\forall d \in \mathscr{D}$. Thus, mathematically, a binary indicator can represent the observed cyber threat of each DER control/status message $d \in \mathscr{D}$,

$$y_d = \begin{cases} 1, & \text{if } x \in \mathscr{X}, x \neq 0, \\ 0, & \text{trustworthy}, \end{cases} \tag{1}$$

where $y_d = 1$ denotes the control message of DER $d \in \mathscr{D}$ is an attack, and 0 trustworthy.

The considered TSGC can observe the characteristics and behavior of each DER $d \in \mathscr{D}$ by analyzing each of the control/status messages. In which, each control/status message consists of the following features: source port a, the total number of packets b, the total number of bytes c, the number of packets at source DER α, the number of packets sent to destination β, and the number of source byte size γ (Teixeira, M., Zolanvari M., and Jain R. 2018; Munir et al. 2021). Therefore, we represent these features as a tuple $z_d : (a, b, c, \alpha, \beta, \gamma)$, $d \in \mathscr{D}$. However, in a dynamic case, we consider M features, then each tuple represents as $z_d : (z_{d1}, \ldots, z_{dm}, z_{dM}))$. Then, for each time slot $t \in T$, the total number of control/status messages are presented as Z^t, where $\forall z_{dm} \in M \times Z^t$ and T in a finite time domain.

In this system model, our goal is to proactively detect the potential cyber threat y by analyzing each control/status message before executing the power transaction command. However, only detecting the potential cyber threat does not buy the system as a trustworthy smart grid control. Therefore, the system model must contain such a mechanism that can explain the root cause of the detected threat along with a confidence score. We consider a linear model that can predict the potential cyber threat y of DER control message $d \in \mathscr{D}$. For each feature m, the model is defined as follows (Munir et al. 2022):

$$y_d = \hat{h}(z_{dm}) = \omega_0 + \omega_1 z_{d1} + \cdots + \omega_M z_{dM}, \tag{2}$$

where ω_M denotes a weight of the model (2). Therefore, z_{dm} presents a feature value, where $m = 1, \ldots, M$ and ω_m represents a weight of feature m.

In this work, our aim is to find the reasoning behind a potential cyber attack from the control/status message of each DER. Therefore, we can define a contribution function of feature m on $\hat{h}(z_{dm})$ (Shapley et al. 1953; Munir et al. 2022; Dubey 1975; Lundberg and Lee 2017). Then, thus, we can formulate a score function based on the contribution of the attack decision and define it as follows (Shapley et al. 1953; Dubey 1975):

$$\Upsilon_m(\hat{h}(z_{dm})) = \omega_m z_{dm} - \mathbb{E}[\omega_m Z^t], \tag{3}$$

where $\mathbb{E}[\omega_m Z^t]$ represents an expectation of effect for feature m. Here, our goal is to detect potential cyber threat $\mathbf{Y} \in \forall y$ that is initiated by DER $d \in \mathscr{D}$ and find the root cause behind that decision $\forall \Upsilon_m(\hat{h}(z_{dm})) \in \boldsymbol{\Upsilon}, m \in M$. Therefore, we need to formulate a problem for a trustworthy smart grid controller that can not only detect the potential attack but also can find the root cause to build trust in such a decision.

3.2 Problem Formulation of Trustworthy Smart Grid Controller

In this section, we design a decision problem for the proposed trustworthy smart grid controller that can coordinate between TSO and DSO. In this formulation, we consider two decision variables: 1) threat detection decisions $y_d \in \mathbf{Y}$, and 2) quantifying root cause $\Upsilon_m \in \boldsymbol{\Upsilon}$ of such decisions that are affected by the M features of the control/status message of each DER $d \in \mathscr{D}$. The objective is to minimize the square error between actual occurrence $\hat{y_d}$ and predicted occurrence y_d of each control/status message $d \in \mathscr{D}$ in the smart grid controller. Thus, we formulate the decision problem of the smart grid controller as follows:

$$\min_{y \in \mathbf{Y}, \Upsilon_m \in \boldsymbol{\Upsilon}} \frac{1}{|\mathscr{D}||T|} \sum_{t=1}^{T} \sum_{d=1}^{|\mathscr{D}|} \left(\hat{y_d} - y_d\right)^2, \tag{4}$$

$$\text{s.t.} \quad \omega_m z_{dm} \leq \mathbb{E}[\omega_m Z^t], \forall m \in (1,\ldots,M), \tag{4a}$$

$$\hat{y_d} \geq \omega_0 + \omega_1 z_{d1} + \cdots + \omega_M z_{dM}, m \in M, d \in \mathscr{D}, \tag{4b}$$

$$\Upsilon_m \leq \sum_{m=1}^{M} (\omega_m z_{dm} - \mathbb{E}[\omega_m Z^t]), \forall \Upsilon_m \in \boldsymbol{\Upsilon}, \tag{4c}$$

$$y_d |\mathscr{D}| \leq |\mathbf{Y}|, \tag{4d}$$

$$y_d \in \{0,1\}, \forall d \in \mathscr{D}. \tag{4e}$$

Constraint (4a) ensures that the contribution of feature $m \in \mathscr{M}$ must be smaller or equal to the expectation of effect m for a control/status message of DER $d \in \mathscr{D}$. Constraint (4b) ensures the linear predicted model (2) never exceeds the actual value during estimation. We establish a coupling between prediction and contribution (i.e., effect) in constraint (4c). Finally, constraints (4d) and (4e) assure that the total number of threats prediction does not bigger than the number of status messages and each status message $d \in \mathscr{D}$ only has one decision, respectively.

The formulated problem (4) leads to a combinatorial optimization problem in both the time and space domain due to potential threats are depended on time and are characterized by the nature of each DER status message. The problem (4) is hard to solve in polynomial time; however, it can be solved through heuristic approximation. In this work, we propose a Shapley-based (Shapley et al. 1953; Munir et al. 2022; Dubey 1975; Lundberg and Lee 2017) trustworthy artificial intelligence framework to solve the formulated problem. The proposed framework can proactively detect cyber threats that are generated by DER, find the root cause by analyzing the features of a status message, and elaborate on the severity of the cyber risk.

4 PROPOSED TRUSTWORTHY ARTIFICIAL INTELLIGENCE FRAMEWORK

We illustrate the high-level system design of the proposed trustworthy AI framework in Figure 2. In particulate, Figure 2 demonstrated the overall AI pipeline for enabling a trustworthy smart grid controller. The developed framework incorporates a regression-based predictive mechanism for cyber threat detection, a

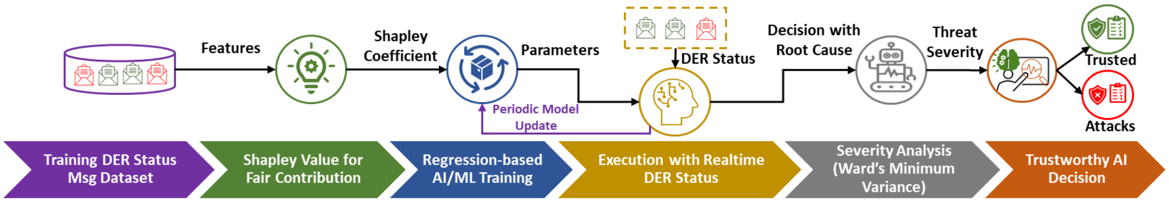

Features
Shapley Coefficient
Parameters
DER Status
Periodic Model Update
Decision with Root Cause
Threat Severity
Trusted
Attacks
Training DER Status Msg Dataset
Shapley Value for Fair Contribution
Regression-based AI/ML Training
Execution with Realtime DER Status
Severity Analysis (Ward's Minimum Variance)
Trustworthy AI Decision

Figure 2: The proposed trustworthy AI framework for the smart grid controller.

Shapley-based explainer for root cause analysis, and Ward's minimum variance-based hierarchical clustering for characterizing threat severity.

The root cause of a status message d in feature m can be estimated as follows:

$$\begin{aligned}\sum_{m=1}^{M} \Upsilon_m(\hat{h}(z_{dm})) = \sum_{m=1}^{M} (\omega_m z_{dm} - \mathbb{E}[\omega_m Z^t]) \\ = \hat{h}(z_{dm}) - \mathbb{E}[\hat{h}(Z^t)],\end{aligned} \tag{5}$$

where ω_m is the weight of contributing feature m and z_{dm} denote the each feature m in a feature tuple z. $\mathbb{E}[\omega_m Z^t]$ represents expected effect by the feature m in DER status message d. Therefore, the root cause of a particular threat detection decision can be calculated as follows:

$$\begin{aligned}\Upsilon_m = \frac{1}{|\mathscr{M}|} \sum_{\mathscr{M} \subseteq \mathscr{D} \setminus \{m\}} \left[|\mathscr{M}| \times \begin{matrix}|\mathscr{D}| \\ |\mathscr{M}|\end{matrix}\right]^{-1} [\hat{h}(z_{dm})_{\mathscr{M}} - \hat{h}(z_{dm})_{\mathscr{M} \setminus m}] \\ = \frac{marginal\ contribution\ of\ m}{number\ of\ coalitions\ \mathscr{M} \subseteq \mathscr{D} \setminus \{m\}},\end{aligned} \tag{6}$$

where $y_d = \hat{h}(z_{dm})$ in (2). Thus, Υ_m provides the root cause of the potential cyber attack $y_d = \hat{h}(z_{dm})$ for each feature $m \in \mathscr{M}$ of DER status message $d \in \mathscr{D}$. Now, we need to characterize the severity of the potential threat risk. However, the challenge here is to adapt to the unknown behavior of a potential cyber threat in smart grid controllers. We consider each X_i can represent a disjoint type of cyber threat or trusted energy transaction control/status message, where $\forall X_i \in \mathscr{X}$. Therefore, for two instance (i.e., i and $i+1$), we can define Ward's minimum variance (Ward Jr 1963) formula as follows:

$$g(\boldsymbol{\Upsilon}, \mathscr{X}) = \frac{|X_i||X_{i+1}|}{|X_i| \cup |X_{i+1}|} ||\mu_{X_i} - \mu_{X_{i+1}}||^2_{\sim \forall \Upsilon_m \in \boldsymbol{\Upsilon}}, \tag{7}$$

where μ_{X_i} and $\mu_{X_{i+1}}$ denote centroid of $|X_i|$ and X_{i+1}. We can rewrite (7) as follows:

$$g(\boldsymbol{\Upsilon}, \mathscr{X}) = \sum_{x \in X_i \cup X_{i+1}} ||x - \mu_{X_i \cup X_{i+1}}||^2_{\sim \forall \Upsilon_m \in \boldsymbol{\Upsilon}} - \sum_{x \in X_i} ||x - \mu_{X_i}||^2_{\sim \forall \Upsilon_m \in \boldsymbol{\Upsilon}} - \sum_{x \in X_{i+1}} ||x - \mu_{X_{i+1}}||^2_{\sim \forall \Upsilon_m \in \boldsymbol{\Upsilon}}. \tag{8}$$

An algorithmic procedure of the proposed mechanism for the trustworthy smart grid controller is shown in Algorithm 1. Physically, Algorithm 1 will be executed by the smart grid controller. The input of the Algorithm 1 will be control/status messages that are sent by DERs $\forall d \in \mathscr{D}$ and each message contains $|\mathscr{M}|$ features. The Algorithm 1 can be run in a finite time domain and is also capable of working as a watchdog in an infinite time domain. All of the received control/status messages will be filtered as a potential cyber threat or trustworthy y_d and quantifying the root cause of that decision $\boldsymbol{\Upsilon}$. Based on that findings, the Algorithm 1 characterizes the risk of severity $g(.)$ among the decisions. Lines from 4 to 11 estimate the predicted decision based on the regression and calculate each feature's contribution $\Upsilon_m \in \boldsymbol{\Upsilon}$ based on the Shapley value interpretation. In particular, line 8 executes the regression model and line 9 calculates the

Algorithm 1 An algorithm for trustworthy smart grid controller

```
Require: $\forall m \in \mathscr{M}, \forall d \in \mathscr{D}, Z^t$, $z_d$: $(a, b, c, \alpha, \beta, \gamma)$, $d \in \mathscr{D}$, $\forall \hat{y} \in \hat{Y}$
Ensure: $y_d \in \mathbf{Y}$, $\Upsilon_m \in \boldsymbol{\Upsilon}$, $g(.)$
 1: while $T \neq T_{max}$ do
 2:     for $\forall d \in \mathscr{D}$ do
 3:         Select a feature $z_{dm} \in Z^t$
 4:         for $\forall m \in \mathscr{M}$ do
 5:             Calculate: $\omega_m z_{dm}$ and $\mathbb{E}[\omega_m Z^t]$
 6:             if $\omega_m z_{dm} \leq \mathbb{E}[\omega_m Z^t]$ then                ▷ Constraint (4a)
 7:                 Calculate: $(\hat{h}(z_{dm}) - \mathbb{E}[\hat{h}(Z^t)])$
 8:                 Estimate: $\omega_0 + \omega_1 z_{d1} + \cdots + \omega_M z_{dM}$
 9:                 Calculate: $\Upsilon_m \leftarrow \frac{1}{|\mathscr{M}|} \sum_{\mathscr{M} \subseteq \mathscr{D} \setminus \{m\}} \left[ |\mathscr{M}| \times \genfrac{}{}{0pt}{}{|\mathscr{D}|}{|\mathscr{M}|} \right]^{-1} [\hat{h}(z_{dm})_{\mathscr{M}} - \hat{h}(z_{dm})_{\mathscr{M} \setminus m}]$ ▷ Using (6)
10:             end if
11:         end for
12:         if (4b) and (4c) then
13:             Evaluate $\frac{1}{|\mathscr{D}||T|} \sum_{t=1}^{T} \sum_{d=1}^{|\mathscr{D}|} \left( \hat{y}_d - y_d \right)^2$        ▷ Generic loss function evaluation for AI models
14:             Get Threat Decision: $y_d$
15:             Get Explanation: $\Upsilon_m$
16:         end if
17:     end for
18:     $\mathbf{Y} \leftarrow y$, $\boldsymbol{\Upsilon} \leftarrow \Upsilon_m$
19:     Severity Analysis: $g(\boldsymbol{\Upsilon}, \mathscr{X})$                ▷ Attack risk analysis using (8)
20: end while
21: Return: $\mathbf{Y}, \boldsymbol{\Upsilon}$, $g(\boldsymbol{\Upsilon}, \mathscr{X})$
```

contribution of each feature in Algorithm 1. Line 13 evaluates a loss function for regression-based AI model training such as random forest, extra tree, adaboots, and so on. Finally, cyber attack severity is analyzed in line 19 of Algorithm 1. The computational complexity of the proposed Algorithm 1 relies on the number of features $|\mathscr{M}|$ of a control/status message in a smart grid controller. The average case Computational complexity of the proposed Algorithm 1 leads to $\mathscr{O}(2^{|\mathscr{D}| \times |\mathscr{M}|} + |\mathscr{X}|^2 \log |\mathscr{X}|)$, where $|\mathscr{X}|$ is the number of disjoint types of cyber threats in smart grid controller.

5 EXPERIMENTAL ANALYSIS

5.1 Experiment Setup

The proposed trustworthy AI framework for a smart grid controller is implemented in Python platform (Scikit-learn: Ensemble Methods ; SHAP: Welcome to the SHAP Documentation ; Scikit-learn: sklearn.cluster.AgglomerativeClustering). In particular, we have developed several ensemble machine learning methods (Scikit-learn: Ensemble Methods) such as Extra Tree, Random Forest, Gradient Boosting, AdaBoost, and Linear Regression for testing the proposed trustworthy smart grid controller. We have utilized the SHAP (SHAP: Welcome to the SHAP Documentation) Python library for calculating Shapley value-based root cause analysis of the potential cyber attack. Then, we used Scikit-learn library to implement the Agglomerative (Scikit-learn: sklearn.cluster.AgglomerativeClustering) clustering for severity characterization of the potential cyber risk. A summary of the experimental setup is given in Table 2.

The implemented framework is tested using the state-of-the-art cyber-physical system SCADA WUSTL-IIOT-2018 (Teixeira, M., Zolanvari M., and Jain R. 2018). In this dataset, each control/status message consists of 6 features along with the true label. The features include the source port, the total number of

Table 2: Summary of Experimental Setup.

Description	Value
No. of DERs status/control message sessions	2000 (Teixeira, M., Zolanvari M., and Jain R. 2018)
No. of training sessions	1400 (Teixeira, M., Zolanvari M., and Jain R. 2018)
No. of testing sessions	600 (Teixeira, M., Zolanvari M., and Jain R. 2018)
Types of attacks	5 (Teixeira, M., Zolanvari M., and Jain R. 2018)
No. of control/status message features	6 (Teixeira, M., Zolanvari M., and Jain R. 2018)
Cluster affinity	Euclidean
Cluster linkage	Complete
Cluster method	Ward

Table 3: Reliability Analysis on AI Decisions (Score between 0 to 1).

AI Methods	TN	FP	FN	TP	R^2 Training Score	R^2 Test Score
Random Forest	0.996	0.003	0.0	1.0	0.997	0.993
Extra Trees	0.996	0.003	0.0	1.0	0.998	0.993
Gradient Boosting	0.996	0.003	0.0	1.0	0.997	0.993
AdaBoost	0.996	0.003	0.0	1.0	0.997	0.993
Linear Regression	0.996	0.003	0.0	1.0	0.946	0.935

packets, the total number of bytes, the number of packets at source DER, the number of packets sent to the destination, and the number of source bytes. We have considered 2000 DER control/status messages and divided these into 70% for training and the rest of them are used for testing. In this dataset, five types of potential cyber threats are considered: 1) port scanner, 2) address scan, 3) device identification, 4) aggressive mode, and 5) exploit; these are labeled into one category. Therefore, there are no given clues that can differentiate between the types of attacks. To overcome such challenges, we have utilized the concept of unsupervised learning for measuring the severity of cyber risk.

5.2 Results and Discussion

In this experiment, we focus on establishing a trustworthy smart grid controller by validating the proposed trustworthy AI framework for SCADA. Thus, in order to establish a trustworthy AI model, we need to fulfill a few of the metrics (Wing 2021; Floridi 2019; Chatila et al. 2021; Liang et al. 2022; Munir et al. 2023) such as *reliability*, *transparency*, *fairness*, *accountability*, *reproducibility*, and *explainability*. Therefore, we justify the proposed trustworthy AI framework in the following subsections.

5.2.1 Reliability

In order to measure the reliability of the proposed trustworthy AI framework, we consider a set of metrics (Scikit-learn: Metrics and scoring: quantifying the quality of predictions) such as true negative (TN), false positive (FP), false negative (FN), true positive (TP), R^2 regression score for training and testing, mean squared error (MSE) and mean absolute error (MAE). In Table 3, we have presented TN, FP, FN, and TP rates for 600 DER control/status messages, where we have found 0.003 and 1.0 as the FP and TP rates, respectively for all of the AI/ML methods. Table 3 also evidenced a higher R^2 score (i.e., coefficient of determination) for all cases. Further, Figure 3 demonstrates the reliability of the defined objective function 4 (i.e., mean squared error) during testing, where we achieved a minimum error rate 0.0015 by the Extra Tree model in the trustworthy AI framework.

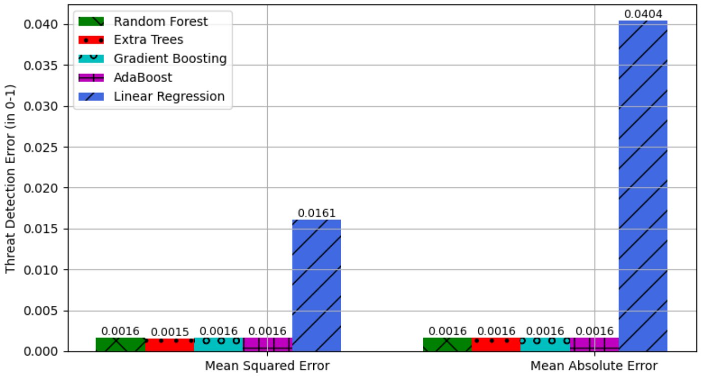
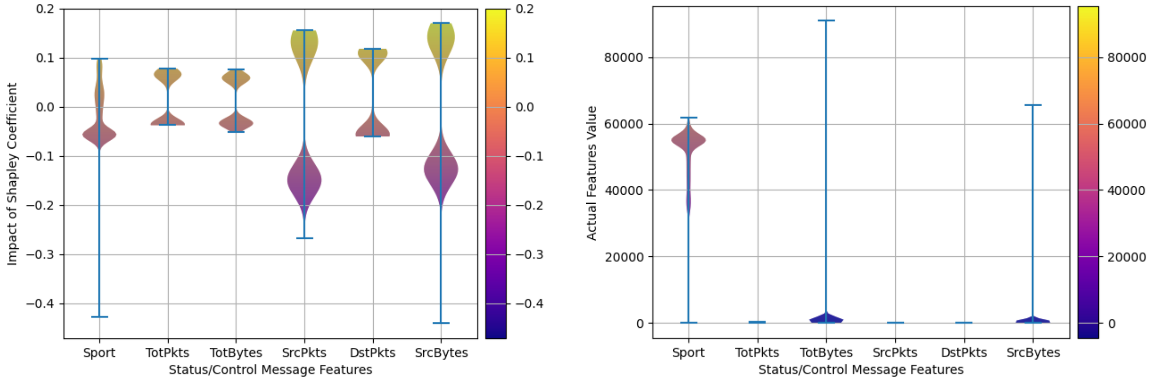


Figure 3: Trend analysis of mean squared error and mean absolute error among the ensemble-based regression AI/ML models during execution (i.e., testing).



(a) Contribution of each feature on attack detection from DER control/status messages.

(b) Features correlation of DER messages in WUSTL-IIOT-2018 dataset (Teixeira, M., Zolanvari M., and Jain R. 2018).

Figure 4: Explanation of the impact of Shapley value coefficient in cyber attack detection.

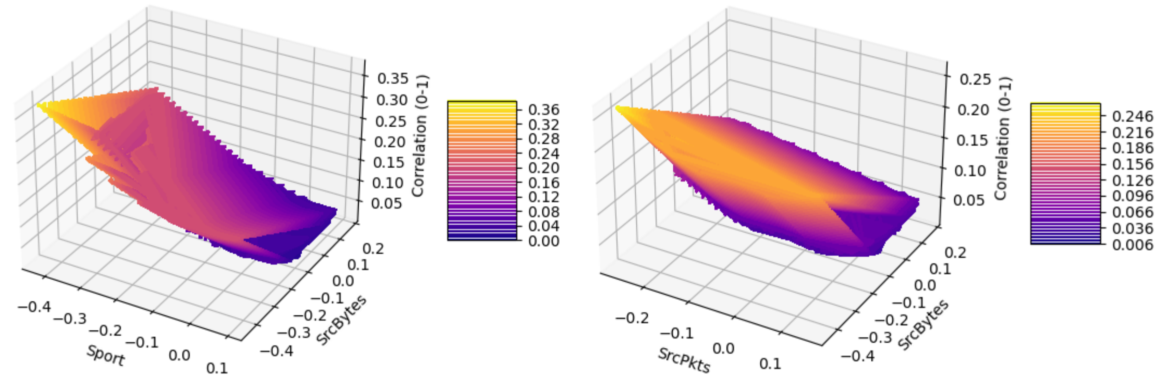
5.2.2 Transparency, Fairness, and Accountability

To assure transparency, fairness, and accountability of the proposed trustworthy AI framework, we have demonstrated the impact of the Shapley value coefficient for DER cyber attack detection in Figures 4a. In particular, it is clearly understandable in Figure 4a that the source port (Sport), the size of the source packet (SrcPkts), and the number of source bytes (SrcBytes) have more contribution for detecting trusted and threat DER messages. However, feature correlation (in 4b) of the raw data does not convey the right insight. As a result, the proposed Shapley value-based feature contribution in attack detection assures fairness and transparency in the AI model's decision.

We have illustrated around 35% effect of generated bytes (SrcBytes) in source port (Sport) in Figure 5a while Figure 5b depicts that generated bytes (SrcBytes) and source packet (SrcPkts) have around 25% effects on DERs' attack detection. This evidence assures the accountability of the proposed trustworthy AI framework.

5.2.3 Reproducibility and Explainability

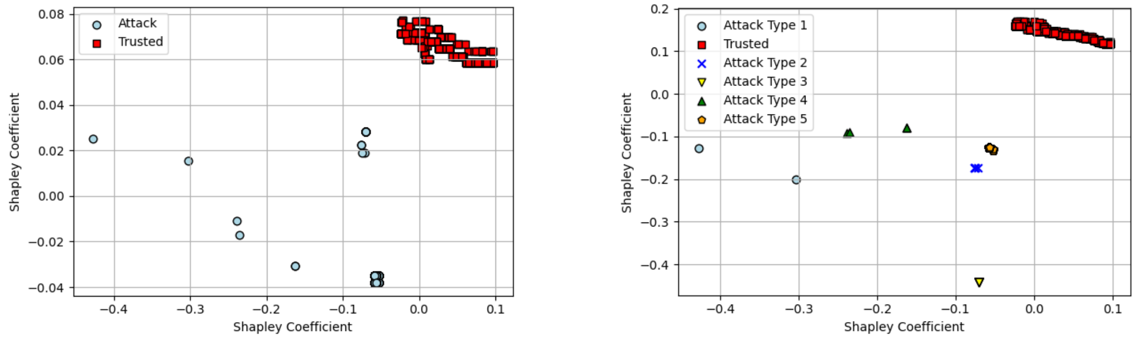
We assure the reproducibility and explainability of the proposed trustworthy AI framework by incorporating Ward's minimum variance-based hierarchical clustering scheme. In particular, the trustworthy AI framework



(a) Effect of generated bytes in source port of DERs on attack detection.

(b) Effect of generated bytes and source packet of DERs on attack detection.

Figure 5: Characterizing evidence on most prominent contributed features in DER control/status messages.



(a) Shapley value-based explanation of threat and trusted DER messages.

(b) Characterization of the severity of five types of cyber attacks in unknown labels.

Figure 6: Outcomes of agglomerative (hierarchical) clustering for the severity explanation.

explains the Shapley coefficient-based root cause and severity of the potential cyber risk that are shown in Figures 6a, 6b, and 7. The explanation between trusted DER messages and potential attacks is shown in Figures 6a, where we can clearly observe that the potential threats do not follow any particular characteristics. Therefore, we differentiate among the unknown types of cyber threats in Figure 6b that can interpret the root cause of such threats based on the Shapley coefficient.

The trustworthy smart grid controller can observe the severity of potential cyber risk from the DER status/control messages. In Figure 7, we quantify the severity of cyber risk during the execution. In Figure 7, the x-axis represents the Shapley coefficient of features and the y-axis is control messages. Figure 7 indicates that the variation of Sport and SrcByte introduced a high risk of potential attack. To this end, the experimental analysis establishes the proposed AI framework as a trustworthy AI mechanism that can satisfy reliability, transparency, fairness, accountability, and explainability.

6 CONCLUSION

In this paper, we have introduced a new trustworthy artificial intelligence framework for assuring trusted power transmission among the DERs. The proposed trustworthy AI framework can proactively predict and explain the root cause of potential cyber-attacks in smart grid controllers while measuring the severity

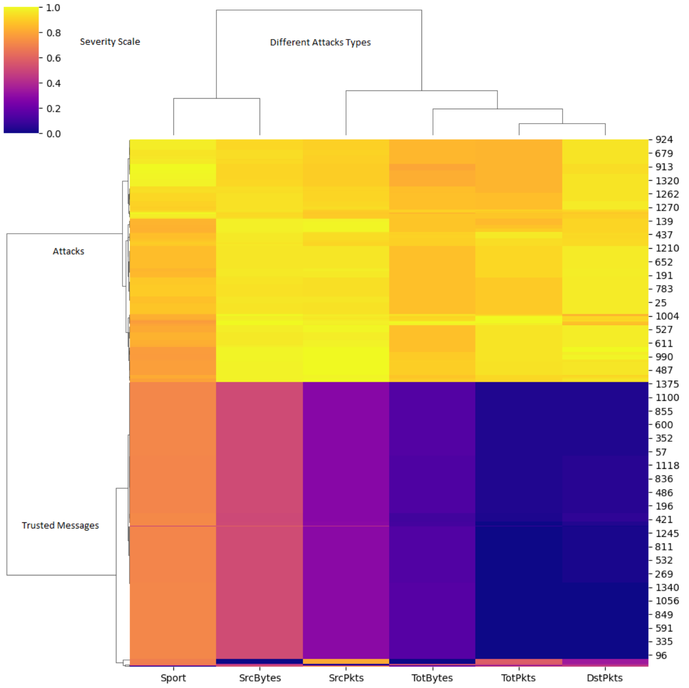


Figure 7: Risk quantification and explanation among the cyber threats in smart grid controller by Ward's minimum variance-based hierarchical clustering.

of such attacks. The developed trustworthy AI framework can select prominent features by deploying Shapley value interpretation and train numerous regression-based ML/AI techniques by executing the mean squared error loss function. Further, it can interpret the root cause and severity of the potential cyber-attacks by employing Ward's minimum variance-based hierarchical clustering. Experimental analysis shows the efficacy of the proposed framework in terms of reliability, fairness, and explainability. That ensures the trustworthiness of the proposed AI framework. In the future, we will investigate a neuro-symbolic AI scheme for autonomous recovering system faults in a smart grid framework.

ACKNOWLEDGMENTS

This work is supported in part by DoD Center of Excellence in AI and Machine Learning (CoE-AIML) under Contract Number W911NF-20-2-0277 with the U.S. Army Research Laboratory.

REFERENCES

Bitirgen, K., and U. B. Filik. 2023. "A Hybrid Deep Learning Model for Discrimination of Physical Disturbance and Cyber-Attack Detection in Smart Grid". *International Journal of Critical Infrastructure Protection* 40:100582.

Chatila, R., V. Dignum, M. Fisher, F. Giannotti, K. Morik, S. Russell, and K. Yeung. 2021. "Trustworthy AI". *Reflections on Artificial Intelligence for Humanity* 12600:13–39.

Dubey, P. 1975. "On the Uniqueness of the Shapley Value". *International Journal of Game Theory* 4(3):131–139.

Dwivedi, R., D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, and R. Ranjan. 2023. "Explainable AI (XAI): Core Ideas, Techniques, and Solutions". *ACM Computing Surveys* 55(9):1–33.

Floridi, L. 2019. "Establishing the Rules for Building Trustworthy AI". *Nature Machine Intelligence* 1(6):261–262.

Karimipour, H., A. Dehghantanha, R. M. Parizi, K.-K. R. Choo, and H. Leung. 2019. "A Deep and Scalable Unsupervised Machine Learning System for Cyber-Attack Detection in Large-Scale Smart Grids". *IEEE Access* 7:80778–80788.

Kurt, M. N., O. Ogundijo, C. Li, and X. Wang. 2019. "Online Cyber-Attack Detection in Smart Grid: A Reinforcement Learning Approach". *IEEE Transactions on Smart Grid* 10(5):5174–5185.

Liang, W., G. A. Tadesse, D. Ho, L. Fei-Fei, M. Zaharia, C. Zhang, and J. Zou. 2022. "Advances, Challenges and Opportunities in Creating Data for Trustworthy AI". *Nature Machine Intelligence* 4(8):669–677.

Lundberg, S. M., and S.-I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions". *Advances in Neural Information Processing Systems* 30:4768–4777.

Munir, M. S., D. H. Kim, S. W. Kang, L. Zou, and C. S. Hong. 2021. "Intelligent Grid Shepherd: Towards a Resilient Distributed Energy Resources Control System". In *2021 22nd Asia-Pacific Network Operations and Management Symposium (APNOMS)*, 398–401.

Munir, M. S., K. T. Kim, A. Adhikary, W. Saad, S. Shetty, S.-B. Park, and C. S. Hong. 2023. "Neuro-Symbolic Explainable Artificial Intelligence Twin for Zero-Touch IoE in Wireless Network". *IEEE Internet of Things Journal (Early Access)*:1–1.

Munir, M. S., K. T. Kim, K. Thar, D. Niyato, and C. S. Hong. 2022. "Risk Adversarial Learning System for Connected and Autonomous Vehicle Charging". *IEEE Internet of Things Journal* 9(16):15184–15203.

Munir, M. S., S.-B. Park, and C. S. Hong. 2022. "An Explainable Artificial Intelligence Framework for Quality-Aware IoE Service Delivery". In *ICC 2022 - IEEE International Conference on Communications*, 4787–4793.

Munir, M. S., N. H. Tran, W. Saad, and C. S. Hong. 2021. "Multi-Agent Meta-Reinforcement Learning for Self-Powered and Sustainable Edge Computing Systems". *IEEE Transactions on Network and Service Management* 18(3):3353–3374.

Nafees, M. N., N. Saxena, A. Cardenas, S. Grijalva, and P. Burnap. 2023. "Smart Grid Cyber-Physical Situational Awareness of Complex Operational Technology Attacks: A Review". *ACM Computing Surveys* 55(10):1–36.

Sakhnini, J., H. Karimipour, and A. Dehghantanha. 2019. "Smart Grid Cyber Attacks Detection Using Supervised Learning and Heuristic Feature Selection". In *2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE)*, 108–112.

Scikit-learn: Ensemble Methods. https://scikit-learn.org/stable/modules/ensemble.html, Accessed 26th April 2023.

Scikit-learn: Metrics and scoring: quantifying the quality of predictions. Metricsandscoring:quantifyingthequalityofpredictions, Accessed 26th April 2023.

Scikit-learn: sklearn.cluster.AgglomerativeClustering. https://scikit-learn.org/stable/modules/ensemble.html, Accessed 26th April 2023.

SHAP: Welcome to the SHAP Documentation. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html, Accessed 26th April 2023.

Shapley, L. S. et al. 1953. *A Value for N-person Games*. Princeton University Press Princeton.

Teixeira, M., Zolanvari M., and Jain R. 2018. "WUSTL-IIOT-2018 Dataset for ICS (SCADA) Cybersecurity Research". Last modified June 28, 2020. https://www.cse.wustl.edu/~jain/iiot/index.html.

U.S. Energy Information Administration: Annual Energy Outlook 2023 with Projections to 2050. https://www.eia.gov/outlooks/aeo/pdf/AEO2023_Release_Presentation.pdf, Accessed 26th April 2023.

Ward Jr, J. H. 1963. "Hierarchical Grouping to Optimize an Objective Function". *Journal of the American Statistical Association* 58(301):236–244.

Wing, J. M. 2021. "Trustworthy AI". *Communications of the ACM* 64(10):64–71.

AUTHOR BIOGRAPHIES

MD. SHIRAJUM MUNIR received the Ph.D. degree in Computer Engineering from Kyung Hee University (KHU), Republic of Korea, in 2021. He is currently working as a Research Assistant Professor at School of Cyber Security, Old Dominion University, USA. His research interests include machine learning, data science, trustworthy artificial intelligence and stochastic models, wireless network, sustainable edge computing, healthcare, Internet of Things network management, future internet, and resilient smart grid. His email address is mmunir@odu.edu.

SACHIN SHETTY received the Ph.D. degree in modeling and simulation from Old Dominion University, in 2007. He is currently the Associate Director of the Virginia Modeling, Analysis, and Simulation Center, Old Dominion University. He holds a joint appointment as an Professor with the Department of Computational, Modeling, and Simulation Engineering. His research interests include the intersection of computer networking, network security, and machine learning. His email address is sshetty@odu.edu.

DANDA B. RAWAT is the Executive Director, Research Institute for Tactical Autonomy (RITA) - a University Affiliated Research Center (UARC) of the US Department of Defense, Associate Dean for Research and Graduate Studies, a Full Professor in the Department of Electrical Engineering Computer Science (EECS), Founding Director of the Howard University Data Science Cybersecurity Center, Director of DoD Center of Excellence in Artificial Intelligence Machine Learning (CoE-AIML) at Howard University, Washington, DC, USA. His email address is danda.rawat@howard.edu.