Downloaded from orbit.dtu.dk on: Apr 25, 2024

DTU Library

SALVAGE D2.2 Description of the developed algorithms for intrusion detection in smart grid components

Kosek, Anna Magdalena; Korman, Matus; Heussen, Kai; Tyge, Emil; Jonasdottir, Anna Hildigunnur

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Kosek, A. M., Korman, M., Heussen, K., Tyge, E., & Jonasdottir, A. H. (2016). *SALVAGE D2.2 Description of the developed algorithms for intrusion detection in smart grid components*.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

SALVAGE D2.2
Description of the developed algorithms for intrusion detection in smart grid components

Anna Magdalena Kosek (DTU), Matus Korman (KTH), Kai Heussen (DTU), Emil Tyge, Anna Hildigunnur Jonasdottir

October 11, 2016

SmartGrids ERA-Net
Cyber-phySicAl security for Low-VoltAGE grids (SALVAGE)

Project partners:
KTH - Royal Institute of Technology
DTU - Technical University of Denmark
PWR - Wroclaw Institute of Technology

Revision history

Issue	Date	Changed page(s)	Cause of Change	Implemented by
0.1	03-06-2015	All	First draft	A.M. Kosek
0.2	22-08-2016	All	Second draft	A.M. Kosek
0.3	13-09-2016	1,4-11,13-20	Third draft	A.M. Kosek, M. Korman, K. Heussen
1.0	11-10-2016	19-20	Final Report	K. Heussen, M. Korman, A.M. Kosek

Contents

SALVAGE project

The purpose of the SALVAGE project is to develop better support for managing and designing a secure future smart grid. This approach includes cyber security technologies dedicated to power grid operation as well as support for the migration to the future smart grid solutions, including the legacy of ICT that necessarily will be part of it. The objective is further to develop cyber security technology and methodology optimized with the particular needs and context of the power industry, something that is to a large extent lacking in general cyber security best practices and technologies today. In particular the focus of the project will be on smart grid with many small distributed energy resources, in particular LV substation automation systems and LV distribution system.

Scope of the report: This report presents developed model-based anomaly detection techniques used for intrusion detection in smart grid.

1 Intrusion detection system (IDS)

Intrusion detection systems (IDS) gather and analyze the information from a computer network or a system in order to discover malicious activities or violations of policy [1]. Two general types of detection techniques are used in IDS: anomaly-based or signature-based. Current IDS focus on the analysis of software and network traffic, but do not usually take the physical component of a cyber-physical system into consideration.An IDS that is well suited for the application in smart grids needs to address both on-line and post-mortem analysis of the state of the observed cyber-physical system and detect anomalies in operation of both cyber and physical components [2].

1.1 SALVAGE approach to IDS

The cyber-physical IDS architecture proposed in the SALVAGE project consists of two main parts: an analysis of the behavior of the observed cyber-physical system and components, and a joint analysis of the cyber-physical system 1. The behavior analysis and characterization of the physical power system is performed with two components: DER and power system analysis, the evaluation of the cyber vulnerabilities is performed in the cyber security analysis component. The joint cyber-physical analysis combines the information from both physical and cyber security components and presents the outcomes to the power system operator.

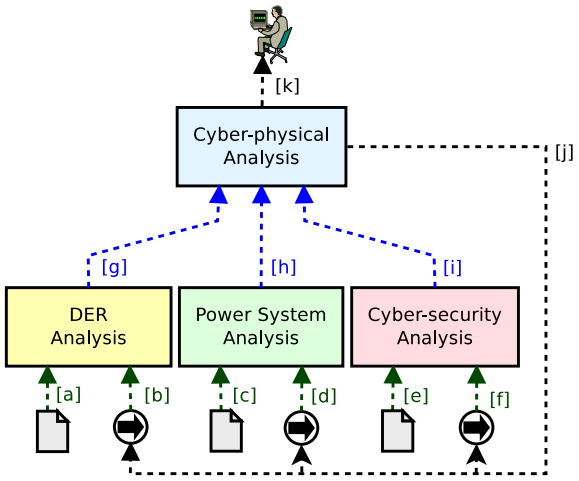

Figure 1: *Architecture of the cyber-physical intrusion detection system*

The main building blocks of the architecture: DER, power system and cyber-security analysis use models of the part of the observed system and live or historical measurements to each perform their partial analysis. The information exchange between the IDS components is as follows:

(a) DER behavior models (producers, consumers and prosumers);

(b) measurements from DERs (includes smart meter measurement);

(c) power grid model or state estimation;

(d) power grid measurements;

(e) model of the ICT infrastructure (e.g., field devices, communication networks, instances of data communication, control systems, and information systems of different kind);

(f) ICT attack indicators – real (e.g., real-time warnings and alerts from a security information and event management (SIEM) system), and/or assumptions about the presence of the attacker;

(g) vector of probability of the DER being compromised;

(h) vector of power system state or risk assessment;

(i) vector of time-to-compromise for top 5%, 50% and 90% of the generally assumed population of attackers, whose skills correspond to the skills of professional penetration testers;

(j) attack assumptions for automated evaluation of multiple scenarios, enabling the validation of multiple hypotheses;

(k) assessment of the cyber-physical system state, including a list of investigated problems and details supporting each hypotheses.

The initial SALVAGE IDS system had a unidirectional structure, where DER, power system and cyber-security analysis was fed forward to the cyber physical analysis component. With time the architecture was extended with the return loop from the cyber-physical analysis component back to the DER, power system and cyber-security analysis components (see arrow (j) in figure 1). The extended architecture supports the idea of a hypothesis testing approach to cyber-physical security. A similar approach is currently being instigated in the cyber-security research. The University of Illinois investigates Network Hypothesis Testing Methodology (NetHTM) [1], a set of techniques for performing and integrating security analyses applied at different network layers, in different ways, to pose and rigorously answer quantitative hypotheses about the end-to-end security of a network. In SALVAGE we investigate a set of hypotheses for a defined purpose of the cyber-physical power system attack. For example in the paper presented in section 4.1 we test the hypothesis that DERs maliciously influence the voltage in the low voltage distribution feeder.

The different parts of the entire cyber-physical IDS as depicted in figure 1 perform each a different type of analysis. The function of the DER analysis module is described below in this document.

The module for power system analysis performs load flow based calculations from electrical measurements (e.g., from substations and/or smart meters). This provides information about how safely the different parts of the electrical grid operate (i.e., how close to their margins), which further enables the consideration of (1) how sensitive each part of the network is to a particular attack, (2) what attack might be currently ongoing given the state of the power grid, and (3) which attacks might be particularly harmful to the power grid given its state. All of these three considerations, which happen on the higher level of the module for cyber-physical analysis, allow the module to evaluate and further formulate hypotheses, adjust the assumptions (see arrow (j) in figure 1), and manage the entire intrusion detection process.

The module for cyber-security analysis performs a probabilistic simulation of cyber attacks within the IT infrastructure that exists alongside the physical infrastructure of the power grid,

[1] (online April 2016) http://publish.illinois.edu/science-of-security-lablet/ hypothesis-testing-for-network-security/

and in the cyber-neighborhood of it. In order for the module to simulate cyber-attacks, a comprehensive model of the IT infrastructure is used together with security information and assumptions (in likeness with the model of a power grid used together with power measurements in order to calculate load flow). A brief overview of the IT model is depicted in figure 2. The

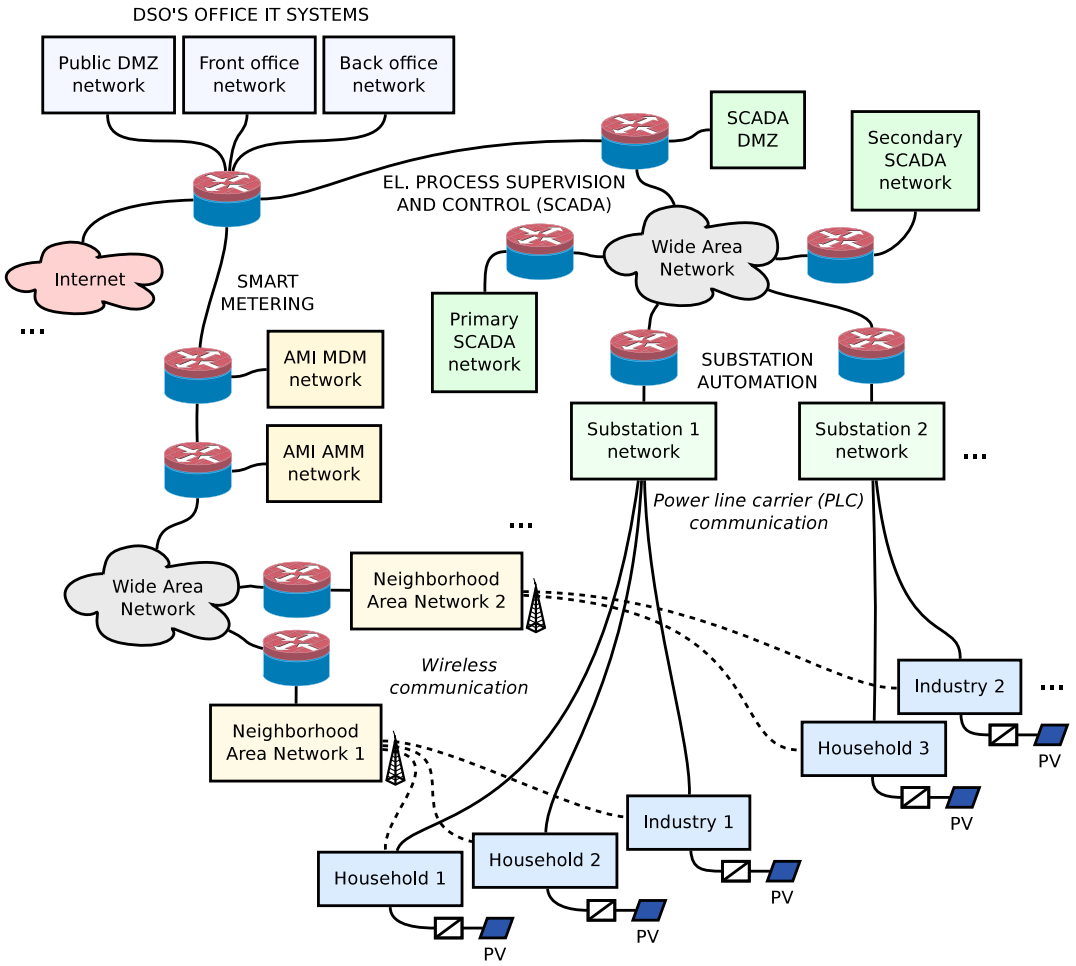

Figure 2: *A brief conceptual overview of the IT model*

security information can be indications from a SIEM system, which are then transformed into security assumptions and embedded into the IT model prior to a simulation. The term security assumptions covers the following: (1) the placement of the attacker in the IT infrastructure (i.e., what does the attacker have access to, what components has the attacker compromised and in what way); (2) any modifications [to the baseline model] of the set of cyber-components modeled and their interconnections; (3) any modifications [to the baseline model] of the properties (i.e., parameters) of the different modeled cyber-components (e.g., operating systems, routers, software applications, security management processes) that are found to have significance to the ease with which an attacker can compromise them given different prerequisites; and hence the overall cyber-security. Examples of these properties are the use of data execution prevention on an operating system using a processor with Intel x86/x64 architecture; the use of rigorous security patching process or system hardening process for cyber-components in a given network segment; or the fact that a software has been subject to significant security scrutiny during

its development (i.e., thorough security reviews and testing). The information the module for cyber-security analysis provides, is a reachability map across the entire IT-infrastructure, evaluating the distribution of time it would take the attacker to compromise each reachable cyber-component, in each reachable mode of compromise defined for it. This further enables the consideration of (1) what components might be compromised and so cause physical disturbances to the power grid; (2) what cyber-components are highly vulnerable and so potentially next-in-line to be compromised, given a set of assumptions. Similarly as with the module for power system analysis, these two considerations happen on the level of the module for cyber-physical analysis, which combines the information from the cyber-security analysis together with that of power system analysis and DER analysis, in order to evaluate and further formulate hypotheses, adjust assumptions, and manage the entire intrusion detection process.

2 Model based anomaly detection

Chandola [3] defines anomalies as *"patterns in data that do not conform to a well defined notion of normal behavior"*. Anomaly detection is a method of discovering anomalies and can be divided into three types: supervised, semi-supervised, unsupervised. Supervised methods use a fully labelled training set to train a classification method which distinguishes normal behaviour from different types of anomalies. Semi-supervised methods (so called model-based anomaly detection) use partially labelled data to create a model of normal behaviour and compare the model output to the observed network or system behaviour. Unsupervised methods assume that the total number of anomalies is small in comparison to the normal data points in the training set. Based on this assumption, statistical anomaly based techniques analyze operational data in order to distinguish between normal and anomalous operation through statistical inference tests. The results or anomaly detection are either labels (classes) or scores (numerical).

Three categories of anomaly detection can be distinguished: point, contextual and collective. The point anomaly detection takes the global view of the data [3]. The contextual or conditional anomalies were introduced in [4] and are defined as data points that are anomalous in a specific context and acceptable in another context. The collective anomaly appears when a related data instances are anomalous with respect to the entire data set. In this case a sequence or a simultaneous occurrence of events or data points is unusual, but the appearance of a single event or data point is not anomalous.

In SALVAGE we have investigated contextual model-based anomaly detection for DER analysis, this research was based on classical and ensemble model-based anomaly detection.

2.1 Classical model-based anomaly detection

In the classical model-based anomaly detection method, normal DER behaviour is modelled in the *DER model* component (figure 3). The output of the model is compared to sensor measurements (or target data) in the *Anomaly Detection* component. Differences between normal and observed DER behaviour can originate form several sources: sensor error, model error, DER fault, or malicious or verified DER control. The output of the model-based anomaly detection is either a label (class) or an anomaly score for every data input.

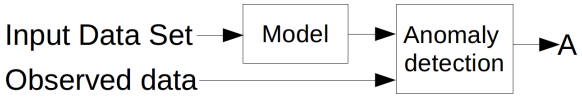

Figure 3: *Flow diagram of the model-based anomaly detection*

2.2 Ensemble model-based anomaly detection

Ensemble learning combines several models to produce a prediction to solve classification and regression problems [5]. The increased robustness and accuracy of ensemble methods over single model methods was reported in [6]. Ensemble learning consists of three steps: generation, pruning and integration. First several redundant models are generated, then the set of models is pruned by removing some of the generated models, finally the base model results are combined to create the ensemble prediction [5]. An overview of ensemble regression approaches for generation, pruning and integration are presented in [5]. The ensemble is evaluated by the degree of agreement between predictions represented by their overall spread. The ensemble prediction is usually evaluated in terms of an average of the individual predictions (mostly using equal weight averaging).

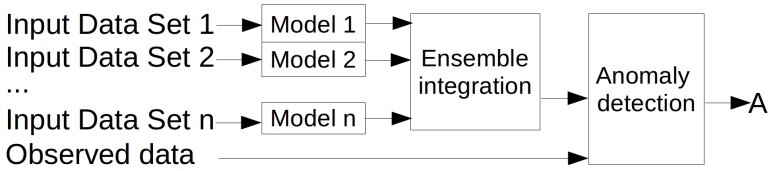

Figure 4: *Ensemble model-based anomaly detection architecture.*

The ensemble model-based anomaly detection (EM-AD) uses two or more DER normal behaviour models which produce the same output variables based on disjoint sets of inputs. The additional *Model merging* component calculates the final model output that is next compared to the observed output in the *Anomaly detection* component.

3 Model-based anomaly detection in SALVAGE IDS

3.1 On-line

In the on-line model-based anomaly detection a hypothesis or a test hypothesis is tested on the current state of the observed system. The current state can be partially measured and estimated with use of a model. The measurement and estimation of the same value or a state is being compared to find anomalies, Additionally comparison of historical and current states can be performed in order to determine the state transition. For example in work presented in section 4.1 we compare historical and current voltage state in order to determine if the influence of the control signal is malicious or aiding.

3.2 Off-line (post-mortem)

The investigated post-mortem model-based anomaly detection consist of two steps. First the semi-supervised normal model is trained with use of historical data. In this step a normal model of the system should be inferred form the data. There are several ways to obtain the

training data: semi-supervised and unsupervised approach. In the semi-supervised approach the historical data can be labeled in order to extract the normal model, this can be done with use of known characteristic of the normal behaviour for example a correlation between model inputs and outputs. In this case the data is filtered and a subset is used to train the normal model. In the unsupervised approach, it is assumed that most of samples are representing the normal behaviour and a small set of outliers represent the anomaly. Statistical models can be used to model normal behaviour. The second step of the post-mortem model-based anomaly detection is the analysis of the historical data with use of the normal model calculated in the previous step. In this step model based anomaly detection is performed on the historical data set.

3.3 Tools used for method validation

3.3.1 Simulation

In order to simulate the measurements form the physical system we have used a co-simulation setup with load flow simulation (PyPower[2]), load and production simulators (mosaik-csv[3]) with real data form Pecan Street project [4]. Simulations are managed by the co-simulation orchestrator mosaik [5]. This co-simulation setup was presented in paper form section 4.1.

3.3.2 Demonstration

The demonstration presented in section 5.1 was done in SYSLAB DTU laboratory. The software was written in Java and trained models are executed with R scripts executed form Java code. For cyber analysis, the software uses alerts form OPC UA server implementation in SYSLAB.

3.3.3 Analysis

Data analysis was performed with R, all ANN models were trained with use of *nnet* package [7]. The analysis was used in papers from sections 4.1 and 4.2, and work in progress presented in section 5.1.

4 Published papers

4.1 Contextual anomaly detection for cyber-physical security in Smart Grids based on an artificial neural network model

Authors: Anna Magdalena Kosek, Energy System Operation and Management, Department of Electrical Engineering, Technical University of Denmark

Publication: 2016 Joint Workshop on Cyber-Physical Security and Resilience in Smart Grids (CPSR-SG2016) part of Cyber Physical Systems week 2016 (CPSweek 2016) April 2016

[2]https://github.com/rwl/PYPOWER
[3]https://bitbucket.org/mosaik/mosaik-csv
[4]http://www.pecanstreet.org/
[5]http://mosaik.offis.de/

Abstract: This paper presents a contextual anomaly detection method and its use in the discovery of malicious voltage control actions in the low voltage distribution grid. The model-based anomaly detection uses an artificial neural network model to identify a distributed energy resource's behavior under control. An intrusion detection system observes distributed energy resource's behavior, control actions and the power system impact, and is tested together with an ongoing voltage control attack in a co-simulation set-up. The simulation results obtained with a real photo-voltaic rooftop power plant data show that the contextual anomaly detection performs on average 55% better in the control detection and over 56% better in the malicious control detection over the point anomaly detection.

Full text: Appendix B

4.2 Ensemble Regression Model-Based Anomaly Detection for Cyber-Physical Intrusion Detection in Smart Grids

Authors: Anna Magdalena Kosek and Oliver Gehrke, Energy System Operation and Management, Department of Electrical Engineering, Technical University of Denmark

Publication: 2016 IEEE Electrical Power and Energy Conference

Abstract: The shift from centralized large production to distributed energy production has several consequences for current power system operation. The replacement of large power plants by growing numbers of distributed energy resources (DERs) increases the dependency of the power system on small scale, distributed production. Many of these DERs can be accessed and controlled remotely, posing a cybersecurity risk. This paper investigates an intrusion detection system which evaluates the DER operation in order to discover unauthorized control actions. The proposed anomaly detection method is based on an ensemble of non-linear artificial neural network DER models which detect and evaluate anomalies in DER operation. The proposed method is validated against measurement data which yields a precision of 0.947 and an accuracy of 0.976. This improves the precision and accuracy of a classic model-based anomaly detection by 75.7% and 9.2%, respectively.

Full text: Appendix B

5 Work in progress

Two anomaly detection methods are currently being developed in WP2 of the SALVAGE project. The first laboratory demonstration uses regression and classification models in an on-line method that detects data integrity anomalies form a PV inverter (section 5.1). The second uses data form EcoGrid project to determine anomalies in the price response of family houses in the demand side management experiment (section5.2).

5.1 OPC UA data integrity monitor for DERs

In this work a first version of the on-line data integrity monitor was created for a PV plant with an OPCUA interface. The cyber-attack target is integrity of the data produced by the PV

plant. The attack investigated in this work was modification of PV set-point value or active power production. We considered data in transit integrity. The developed monitor uses the data produced by the OPC UA server and additionally meteorological data and measurement from an independent active power mater. The preliminary architecture of the monitor is presented in figure 5.

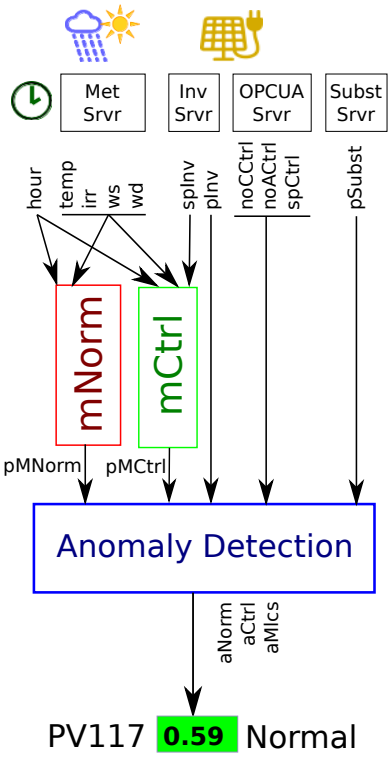

Figure 5: *OPCUA PV monitor architecture.*

The components of the architecture are as follows:

- **Met Srvr** is a meteorological server near by the investigated PV plant providing temperature (*temp*), solar irradiance (*irr*), wind speed (*ws*) and wind direction (*wd*).
- **Inv Srvr** is a OPCUA server providing power measurement of the PV inverter, in this work we only consider active power measurement (*pInv*) and active power set-point that the PV is following (*spInv*).
- **OPCUA Srvr** is a OPCUA server providing OPCUA specific security events, for example new client connection, authentication status, requested data and modified data. These alerts and events are being aggregated and the OPCUA PV monitor calculates the total current number of connected clients (*noCCtrl*), the number of connected authenticated controllers (*noACtrl*) and the last set-point set by any controller (*spCtrl*).
- **Subst Srvr** is a substation server provided the independent power measurement from the PV inverter at the point of common coupling. The OPCUA PV monitor reads only active power measurement (*pSubst*).

- **nNorm** is a PV normal model component that takes hour-of-the day (*hour*) calculated form the computer host local time (*hour*), temperature (*temp*), solar irradiance (*irr*), wind speed (*ws*) and wind direction (*wd*) and outputs predicted active power production of the PV (*pMNorm*).

- **nCtrl** is a PV normal behaviour with control model component that takes hour-of-the day (*hour*) calculated form the computer host local time (*hour*), temperature (*temp*), solar irradiance (*irr*), wind speed (*ws*), wind direction (*wd*) and inverter setpoint (*spInv*) and outputs predicted active power production of the PV in the context of the control signal (*pMCtrl*).

- **AnomalyDetection** is a classification model that inputs normal power prediction (*pMNorm*), normal power prediction in the context of control (*pMCtrl*), PV power measured by the inverter (*pInv*), the total current number of connected clients (*noCCtrl*), the number of connected authenticated controllers (*noACtrl*), the last set-point set by any controller (*spCtrl*) and active power measurement at the common point of coupling (*pSubst*). The model outputs the probability for each class of anomaly: Normal, Controlled and Malicious (*aNorm, aCtrl, aMlcs*).

5.1.1 PV normal model

The PV plant normal model inputs are: solar irradiance $[kW/m_2]$ (*irr*), wind speed $[m/s]$ (*ws*), wind direction $[deg]$ (*wd*),and hour of the day (*hour*). The model output is active power production $[kW]$ (*power*). The input significance analysis of the linear model based on the same inputs and outputs as the presented model, using test statistics under the null hypothesis, shows that all inputs are significant. The training data set consisted of 1 second measurement from a PV at SYSLAB, DTU Risø Campus from June 2016. An overview of the training data is presented in figure 6. The set consists of several days of data, the missing days: 3rd, 7th and 8th of June does not influence the model quality.

Artificial neural network (ANN) was used to model the normal behaviour of the PV plant. Package *nnet* for R was used to calculate the feed-forward neural networks with a single hidden layer [7]. In order to determine the number of the hidden neurons in the ANN, ten models of 1 to 20 hidden neurons were trained, comparison of their normalised RMSE (root mean squared error) is presented in figure 7.

The used ANN consists of 4 inputs, 10 hidden neurons, bias unit and one output. The formula to train the neural network, as described below, uses regularization parameter $decay = 0.0006$. The data set used to train the ANN was normalised in order to improve the model accuracy.

```
nnet.formula(formula = pInv ~ irr + hour + ws + wd, data = dataN,
    size = 10, decay = 6e-04, maxit = .Machine$integer.max)
```

The residuals obtained form the model are presented in fig 8. The obtained RMSE is 0.9589178. The graphical representation of the mapping between model inputs and outputs is presented in figure 9.

5.1.2 Controllable PV normal model

A complex controllable PV model was replaced with a simple model that inputs the output of the normal model active power production $[kW]$ (*power*) and inverter set-point (*spInv*)

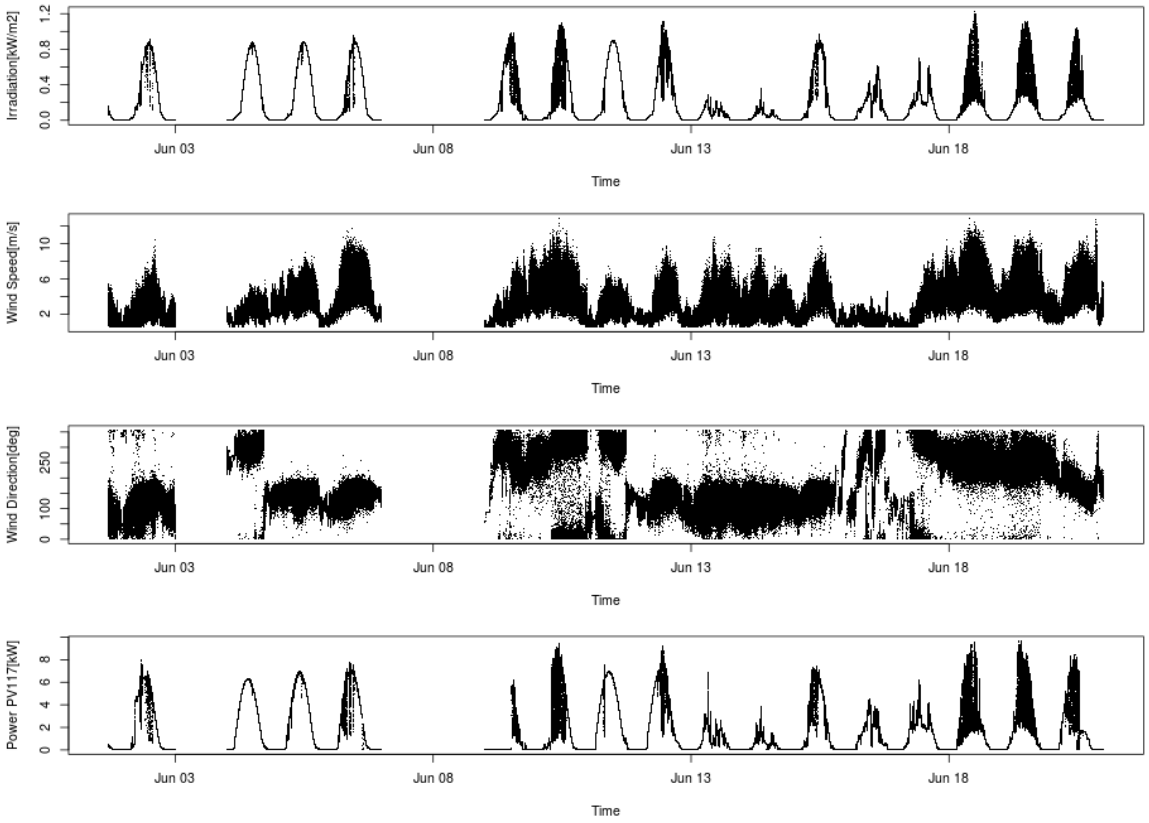

Figure 6: *Normal model training data.*

and outputs predicted active power production of the PV in the context of the control signal (*pMCtrl*). In the controllable PV normal model if the estimated power is higher that set-point, then the model outputs the set-point, otherwise it outputs the estimated power. The difference between measured power, power estimated form normal and controllable PV normal models are presented in figure 10

5.1.3 Investigated attacks

The objective of the considered attacks is data integrity while performing control on a PV plant. The general idea is that the monitor would discover if the PV is being controlled and at the same time the data reported form the PV inverter is being actively modified in order to hide the effect or presence of the malicious control actions. Seven attacks were considered in this scenario:

- the set-point data is being modified in order to hide the control action:
  - **Attack A1** set-points (*spInv* and *spCtrl*) are modified to 110%
  - **Attack A2** set-points (*spInv* and *spCtrl*) are modified to 90%
  - **Attack A3** set-points (*spInv* and *spCtrl*) removed
- OPCUA client authentication and registration in OPC UA server can be removed in order to hide that the external client is controlling the PV plant.

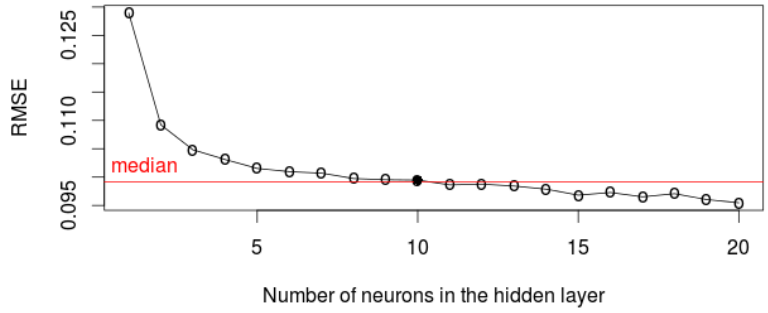

Figure 7: *RMSE of the hidden layer neurons ANN training sessions (from 1 to 20 neurons).*

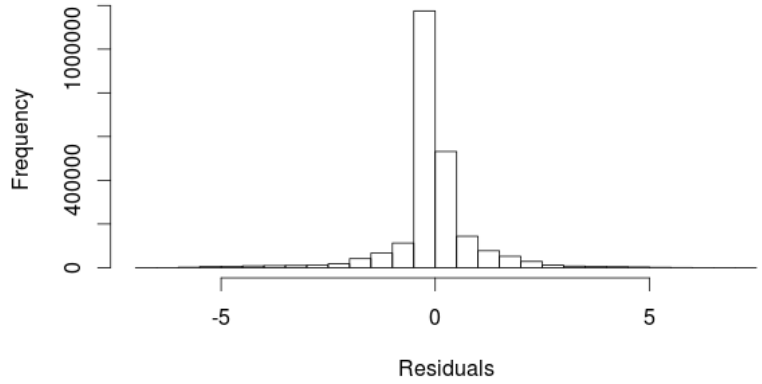

Figure 8: *Histogram of the normal model's residuals.*

  - **Attack A4** authentication removed
  - **Attack A5** authentication and client registration removed

- modification of the power production data
  - **Attack A6** active power production *pInv* modified 110%
  - **Attack A7** active power production *pInv* modified 90%

In practice attacks A4 and A are difficult to execute on the OPCUA server due to its security features, hoverer the delivery of the alert could be delayed or the message could be lost in transit. Additionally the integrity of the event message could be compromised in transit.

5.1.4 Anomaly-Detection model trained with synthetic attack data

The training data set for the *AnomalyDetection* model consisted of 1 second measurement from a PV at SYSLAB, DTU Risø Campus from 22-24 June 2016. The PV data was modified to introduce different attacks 5.1.3. The classification model can output either a score- the probability for each class of anomaly: Normal, Controlled and Malicious, or a label - the anomaly class (labels: Normal, Controlled and Malicious).

In order to determine best type of the classification model from *nnet* package [7] for the investigated problem, several models were trained and their confusion matrix, accuracy, precision and sensitivity are used to choose the model type, as presented in table 1.

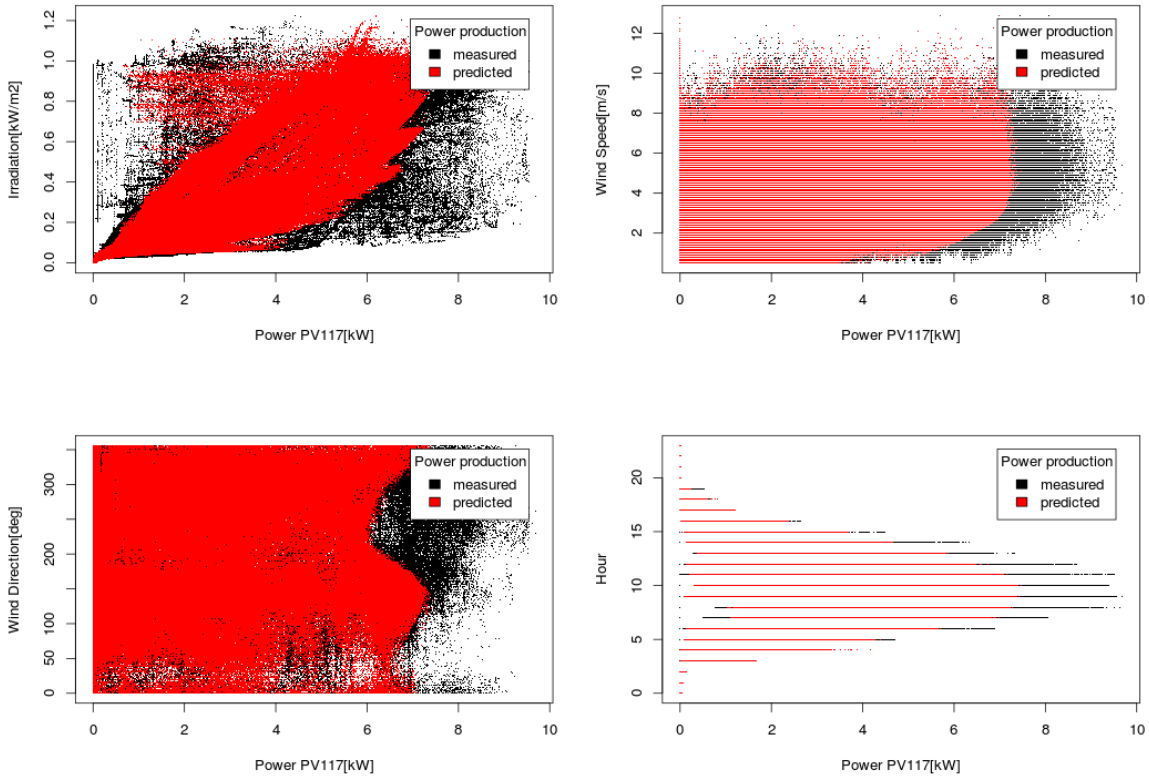

Figure 9: *Model's prediction compared to the expected output for each input.*

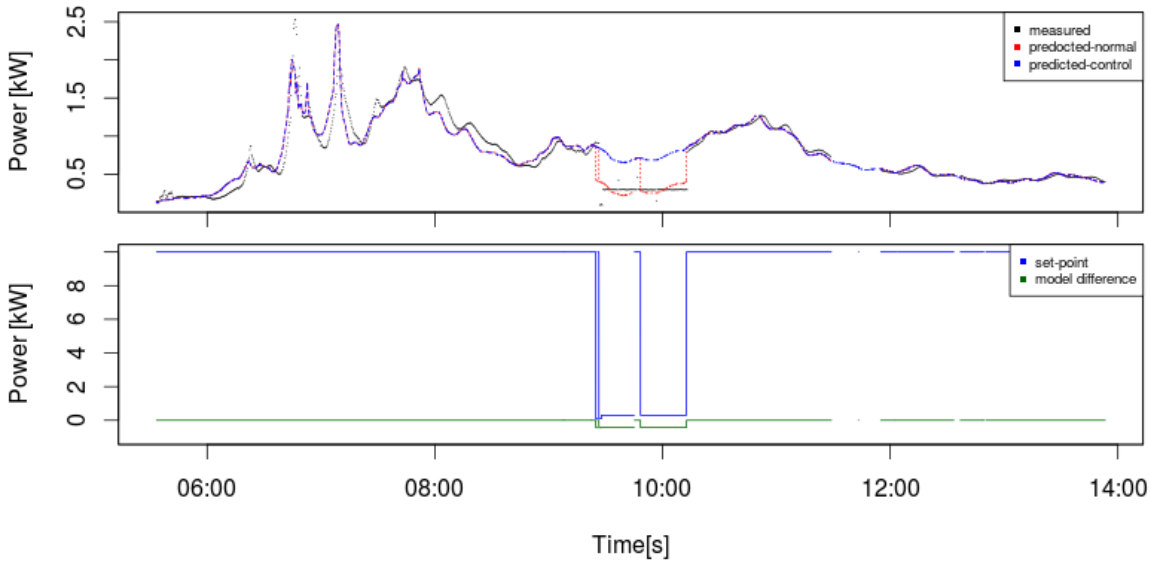

Figure 10: *Comparison of the PV normal and Controllable PV normal models. Set point is the control signal that the PV is following, and the model difference is the the diference between normal and controller model.*

The confusion matrix, as described by Fawcett in [8], is a compilation of instances of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), evaluated from

Table 1: *Confusion matrix of the AnomalyDetection model trained with real data.*

Algorithm	Class	TP	TN	FN	FP	SEN	ACC	PREC
log-linear	Normal	23497	211859	0	0	1.000	1.000	1.000
	Controlled	6	208683	26592	75	0.000	0.887	0.074
	Malicious	185186	23503	75	26592	1.000	0.887	0.874
softmax	Normal	23497	211859	0	0	1	1	1
	Controlled	15282	208654	11316	104	0.575	0.951	0.993
	Malicious	185157	38779	104	11316	0.999	0.951	0.942
softmaxSkip	Normal	23497	211859	0	0	1	1	1
	Controlled	15140	208314	11458	444	0.569	0.949	0.972
	Malicious	184817	38637	444	11458	0.998	0.949	0.942
entropy	Normal	0	211859	23497	211859	0	0.474	0
	Controlled	26598	208758	26598	185261	0.5	0.526	0.126
	Malicious	185261	50095	185261	26598	0.5	0.526	0.874
entropySkip	Normal	0	197761	23497	211850	0	0.457	0
	Controlled	12782	208758	26598	184970	0.325	0.512	0.065
	Malicious	184970	50095	185261	12782	0.500	0.543	0.935
lout	Normal	0	18181	127071	1667	0	0.124	0
	Controlled	0	144567	2352	0	0	0.984	-
	Malicious	0	129423	17496	0	0	0.881	-
logit	Controlled	0	48189	784	0	0	0.984	-
	Malicious	4184	42357	1648	784	0.717	0.950	0.842

a population of results. To measure the correctness of the anomaly detection we calculate three significance measures: accuracy (ACC), precision (PREC) and sensitivity (SEN). Accuracy is a description of random errors, precision is the fraction of predicted instances that are relevant and sensitivity (also called recall) is the fraction of predicted instances that are retrieved. Accuracy, precision and sensitivity are calculated as follows:

$$ACC = \frac{TP + TN}{TP + FP + FN + TN}$$

$$PREC = \frac{TP}{TP + FP}$$

$$SEN = \frac{TP}{TP + FN}$$

Several supervised classification methods were considered in order to train the *AnomalyDetection* model. The types of investigated classification models:

- **log-linear:** fitting multinomial log-linear models with use of artificial neural networks, using *multinorm* function from *nnet* R package
- **logit:** logistic regression is calculated with *glm* R function. Logistic regression can be calculated only for two classes, therefore class Normal and Controlled were combined.

- **softmax:** feed-forward neural network using softmax function as activation function in the output layer
- **softmaxSkip:** recurrent neural network using softmax function as activation function in the output layer
- **entropy:** feed-forward neural network using maximum conditional likelihood (east-squares) used for training
- **entropySkip:** recurrent neural network using maximum conditional likelihood (east-squares) used for training
- **lout:** feed-forward neural network using linear function as activation function in the output layer.

The confusion matrix, sensitivity, precision ad accuracy scores for each model are presented in table 1. Neural network model has a single hidden layer with 10 neurons, regularization parameter is set to $decay = 0.0004$. In the case when precision cannot be calculated the sum of TP and FP is equal to zero.

As seen in table 1, simple *logit* and *lout* models perform badly with identifying true positives in the data. Both *entropy* and *entropySkip* models recognize *Controlled* and *Malicious* classes with 0.512 to 0.543 accuracy but score low on sensitivity and precision. The *log-linear* model recognises well the *Normal* behaviour, but number of false-negatives for *Controlled* and false-positive for *Malicious* behaviours is alarmingly high. The *softmax* model performs best with only low score for sensitivity in the *Controlled* class of 0.575. Changing the ANN to a recurrent does not improve sensitivity precision and accuracy, therefore in this work we have decided to use the *softmax* model for the anomaly detection.

5.1.5 Anomaly Detection model with real attack data

The same ANN network as presented in section 5.1.4: ANN model with a single hidden layer with 10 neurons, regularization parameter set to $decay = 0.0004$ was used in this section. In this case the model was trained with data from the SYSLAB laboratory, with attacks as described in section 5.1.3.

Table 2: *Test and Validation set class samples*

Set	Normal	Controlled	Malicious	Total
Training	2856	350	662	3868
Validation	1318	459	456	2233

The properties of the training and the validation data sets are presented in table 2. The training set consists of recorded labelled data from 2016-08-24, 2016-08-19, 2016-07-08, and 2016-08-25 (date format *yyyy-mm-dd*) and the validation data from 2016-08-26. The number of attack cycles of a length of around 1 minute for each attack type are presented in table 3.

The ANN model developed on a random sample of size of 80% of the training set and tested with a random sample of size of 20% of the training set, the results of testing different ANN

Table 3: *Test and Validation set attack samples*

Set	No attack	A1	A2	A3	A4	A5	A6	A7	Total
Training	3207	64	129	160	90	69	73	76	3868
Validation	1777	94	43	82	49	1	43	144	2233

types are presented in table 4. It is clearly visible that softmax model achieves the highest sensitivity, precision and accuracy.

The selected soft-max ANN model was used in the intrusion detection system as presented in figure 5. The overall results for the soft-max ANN model are sensitivity of 0.627, accuracy of 0.915 and precision of 0.935 (as seen in table 4). When the prediction for every attack is considered (as shown in table 5), the lowest sensitivity is observed for attack A1 and A7, while the accuracy and precision is between 0.912 and 1. The sensitivity measures how many relevant samples are selected, in other words how complete the results of the prediction are. The sensitivity can also be treated as probability that a randomly selected relevant sample is retrieved in a search. The different sensitivity measures for each attack recognition points to an issue that the model represents some attacks better than other and therefore recognises relevant samples better.

5.1.6 Conclusions

The proposed intrusion detection system uses information form the PV power production, meteorological conditions and cyber-security events to discover cyber attacks on the PV operation. The proposed method was verified with experimental data. Further improvements of the method mainly focus on the model improvements to decrease the method sensitivity in the attack recognition. ¡in order to achieve this improvement, more varied attack data is required for the model training. Additional long therm monitoring tests are needed to asses the repeatability of results presented in this report and applicability of the developed model to fall, winter and spring seasons.

Table 4: *Confusion matrix of the AnomalyDetection model with real attack data*

Algorithm	Class	TP	TN	FN	FP	SEN	ACC	PREC
softmax	Normal	1318	901	0	14	1.000	0.994	0.989
	Controlled	439	1618	20	156	0.956	0.921	0.738
	Malicious	286	1757	170	20	0.627	0.915	0.935
softmaxSkip	Normal	982	276	633	342	0.608	0.563	0.742
	Controlled	46	1589	119	479	0.279	0.732	0.088
	Malicious	79	1475	374	305	0.174	0.696	0.206
entropy	Normal	633	468	1270	275	0.333	0.416	0.697
	Controlled	39	2068	165	374	0.191	0.796	0.094
	Malicious	86	1780	453	327	0.160	0.705	0.208
entropySkip	Normal	623	508	1356	270	0.315	0.410	0.698
	Controlled	37	2068	165	487	0.183	0.764	0.071
	Malicious	123	1780	453	401	0.214	0.690	0.235
lout	Controlled and Normal	0	6131	568	0	0	0.915	-
	Malicious	0	5431	1268	0	0	0.811	-
logit	Normal	1194	174	421	444	0.739	0.613	0.729
	Controlled	0	2068	165	0	0	0.93	-
	Malicious	132	1317	321	463	0.291	0.649	0.222
log-linear	Normal	984	274	631	344	0.609	0.563	0.741
	Controlled	36	1702	129	366	0.218	0.778	0.090
	Malicious	115	1392	338	388	0.254	0.675	0.229

Table 5: *Confusion matrix of the Anomaly Detection model*

Attack	TP	TN	FN	FP	SEN	ACC	PREC
None	1757	286	20	170	0.989	0.915	0.912
A1	36	2139	58	0	0.383	0.974	1
A2	33	2190	10	0	0.767	0.996	1
A3	69	2151	13	0	0.841	0.994	1
A4	49	2184	0	0	1	1	1
A5	0	2232	1	0	0	1	-
A6	43	2190	0	0	1	1	1
A7	56	2089	88	0	0.389	0.961	1

5.2 Residential demand response: behaviour model and anomaly detection

Residential demand response is maturing from a concept to real-world applications, and it is considered a significant resource of localized flexibility. In particular in cases where the heat and cooling needs of buildings are satisfied by electric heating or heat pumps. As demand response is maturing from a vision to real-world applications, it is also becoming a potential target for cyber attacks. A real-time demand response system can be viewed as a cyber-physical

system: a physical structure, yet with a behaviour that is strongly ICT-dependent. Therefore, there should be physical (non-ICT based) indicators of anomalous behaviour. In this work we investigate the observable characteristics of individual households consumption behaviour with respect to real-time demand response.

The demand response behaviour is partly governed by physical properties of the process, partly by autonomous behaviour of residents, and in part by the local control systems, which may be parametrized by local users. Combined this leads to new challenges for reliability and security of operation, as the required open control systems also offer more entry points for cyber-attacks.

The data analysis reported a wide variety of typical behaviours, which indicated that anomalies need to be identified in the characteristic time-domain response behaviour of an individual (or group of) loads. The load response behaviour as modelled by the finite impulse response (FIR) characterizes the behaviour. Consequentially, a change of this behaviour has been formulated as criterion for anomalies. It has been shown that some information on characterizing the system response could be extracted, and an intuitive interpretation of extracted parameters may be given. The identification of behavioural abnormalities in demand response data was expected to be challenging, and so it is not surprising that the accuracy of anomaly detection has not been convincing.

The main innovations of the presented approach are:

- method and metrics to define similarity of response characteristics based on FIR,
- the demonstration of a continuously identified the price responsiveness of individual households.

This *online* system requires about 12-24 hours to fully converge from one behaviour characteristic to another, behavioural anomalies may be discovered more quickly. The results form this work are to be published in in scientific conference.

Title: Behaviour Signatures of Residential Demand Response applied to Cyber-physical Anomaly Detection

Authors: Kai Heussen, Emil Tyge and Anna Magdalena Kosek, Energy System Operation and Management, Department of Electrical Engineering, Technical University of Denmark

Publication: This work has been submitted for publication at a relevant scientific conference. A major part of this investigations has been carried out in context of a student project at DTU and on the basis of a data set obtained by the EcoGrid.eu project [9].

References

[1] P. E. Proctor, *Practical intrusion detection handbook*. Prentice Hall PTR, 2000.

[2] A. M. Kosek, "Contextual anomaly detection for cyber-physical security in smart grids based on an artificial neural network model," in *2016 Joint Workshop on Cyber-physical Security and Resilience in Smart Grids*. IEEE, 2016.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[4] X. Song, M. Wu, C. Jermaine, and S. Ranka, "Conditional anomaly detection," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, no. 5, pp. 631–645, 2007.

[5] J. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. D. Sousa, "Ensemble approaches for regression: A survey," *ACM Comput. Surv.*, vol. 45, no. 1, pp. 10:1–10:40, Dec. 2012.

[6] N. García-Pedrajas, C. Hervás-Martínez, and D. Ortiz-Boyer, "Cooperative coevolution of artificial neural network ensembles for pattern classification," *Evolutionary Computation, IEEE Transactions on*, vol. 9, no. 3, pp. 271–302, 2005.

[7] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. New York: Springer, 2002, iSBN 0-387-95457-0. [Online]. Available: http://www.stats.ox.ac.uk/pub/MASS4

[8] T. Fawcett, "An introduction to {ROC} analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861 – 874, 2006, {ROC} Analysis in Pattern Recognition. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016786550500303X

[9] N. P. Lund, P., R. D. Grandal, S. H. Sørensen, M. F. Bendtsen, G. L. Ray, E. M. Larsen, J. Mastop, F. Judex, F. Leingruber, K. J. Kok, and P. A. MacDougall, "EcoGrid EU - A Prototype for European Smart Grids, Overall evaluation and conclusion," 2015.

Ensemble Regression Model-Based Anomaly Detection for Cyber-Physical Intrusion Detection in Smart Grids

Anna Magdalena Kosek, Oliver Gehrke
Technical University of Denmark
Department of Electrical Engineering
Energy System Operation and Management
Email:{amko,olge}@elektro.dtu.dk

***Abstract*—The shift from centralised large production to distributed energy production has several consequences for current power system operation. The replacement of large power plants by growing numbers of distributed energy resources (DERs) increases the dependency of the power system on small scale, distributed production. Many of these DERs can be accessed and controlled remotely, posing a cybersecurity risk. This paper investigates an intrusion detection system which evaluates the DER operation in order to discover unauthorized control actions. The proposed anomaly detection method is based on an ensemble of non-linear artificial neural network DER models which detect and evaluate anomalies in DER operation. The proposed method is validated against measurement data which yields a precision of 0.947 and an accuracy of 0.976. This improves the precision and accuracy of a classic model-based anomaly detection by 75.7% and 9.2%, respectively.**

***Keywords*—Data-driven modelling, machine learning, cyber-physical security, model-based anomaly detection, ensemble regression, power system.**

I. Introduction

Power systems are critical infrastructures for industry, transportation, health care, water and food supply, telecommunication and financial systems. Cybersecurity in power grids is a topic of increasing concern [1], and a considerable effort is required to secure the infrastructure from cyber-attacks. This includes securing legacy systems and designing new systems with security in mind [2]. The discipline of cybersecurity analyzes threats, vulnerabilities and risks for computing systems and proposes defense mechanisms [3]. Initially, cybersecurity in power systems has focused on communication standards [4] including Advanced Metering Infrastructure [5], and SCADA security [6]. More recently, a new type of approach has been used which takes the cyber-physical nature of power systems into account, i.e. the interaction between the physical power system and the ICT infrastructure used in its operation (e.g. [7]).

Cybersecurity measures can be categorized along the time domain as preventive, real-time or post-mortem. Intrusion detection systems (IDS) gather and analyze the information from a computer network or system in order to discover malicious activities or violations of policy. Two general types of detection techniques are used in IDS: anomaly-based or signature-based. Current IDS focus on the analysis of software and network traffic, but do not usually take the physical component of a cyber-physical system into consideration. In this paper we investigate an intrusion detection method based on physical component models. Using the example of a photovoltaic (PV) generator as the potential target of a cyber-attack [2], we analyze operational data to detect anomalies in its operation which may be further classified as resulting from unauthorized or malicious control inputs.

In the data mining context, anomaly detection is concerned with identifying rare data instances or events that do not match an expected pattern. Applications include the detection of financial fraud, identification of manufacturing faults and monitoring of computers in data centers [8]. Three types of anomaly detection techniques can be distinguished: supervised, semi-supervised and unsupervised. Supervised methods use a fully labelled training set to train a classification method which distinguishes normal behaviour from different types of anomalies. Semi-supervised methods (so called model-based anomaly detection) use partially labelled data to create a model of normal behaviour and compare the model output to the observed network or system behaviour. Unsupervised methods assume that the total number of anomalies is small in comparison to the normal data points in the training set. Based on this assumption, statistical anomaly based techniques analyze operational data in order to distinguish between normal and anomalous operation through statistical inference tests.

In this work we investigate model-based anomaly detection for DERs. Anomaly detection with a single regression model has been used for PV fault diagnostics [9], wind turbine fault detection [10], and discovering cyber-attacks on SCADA [6]. In [9], the authors use redundant linear and non-linear models to detect different faults in PV operation. Anomaly detection in control systems with use of a linear model of the normal behaviour is investigated in [11]. Contextual anomaly detection was used in a cyber-physical IDS (Intrusion Detection System) to detect malicious voltage control actions [12]. Here, the proposed on-line method utilizes a model which is trained on data known to have no malicious control actions or sensor faults, therefore the trained normal model is accurate.

This paper continues the work presented in [12]. The proposed method recognises anomalies in pre-recorded data of DER operation; it therefore focuses on post-mortem analysis, detecting past occurrences of control events. The contribution of this paper is as follows: a) a novel model-based anomaly detection method using ensemble regression, in section II-A; b) a new method for selecting model training set to improve the anomaly detection performance, in section II-B; We further verify of the proposed method against the DER operation data, in section IV and perform quantitative comparison of the proposed method against single model anomaly detection, in section VI.

II. Cyber-physical intrusion detection system

The concept of a cyber-physical intrusion detection system (CP-IDS) was proposed in the SALVAGE project [13] as presented in figure 1 [12]. The CP-IDS uses data from the observed cyber-physical system and analyses it under three aspects: DER operation, power system vulnerability and cyber-security threat. The outcome of this analysis is passed to a cyber-physical analysis component. Here, all three aspects are combined into a joint cyber-physical security assessment.

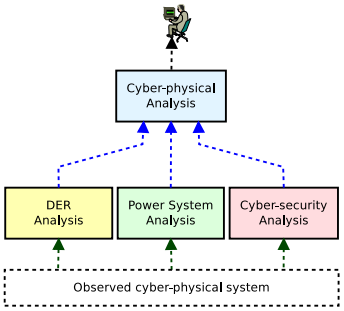

Fig. 1. Cyber-physical IDS architecture [12].

Work presented in [12] is an on-line power system and component analysis for the purpose of discovering malicious voltage control events. In this paper we focus on the DER analysis component of the CP-IDS and propose a method for off-line (post mortem) DER control detection. In this context, four operational states of a DER can be distinguished:

- Normal operation: a DER behaves as expected and its operation is not influenced by external set points. An internal DER controller may or may not govern the operation of the unit.
- Faulty operation: the operation of a DER deviates from normal due to a fault at the unit or in its electrical network environment.
- Verified control: a DER behaves as expected under a verified control scheme, or according to authorized external set points.
- Malicious control: a DER is operated under an unverified control scheme or according to unauthorized external set points.

In this paper we define a DER behaviour anomaly as either verified or malicious control, and consider faulty operation as part of normal DER operation to exclude it from the detection algorithm. After an introduction to model based-anomaly detection in section II-A we describe data cleaning and selection in section II-B and model training in section III. We present a model based anomaly detection method with ensemble regression models in section V and apply it to a PV plant data set in section IV. In section VI we compare this approach to several single model approaches and evaluate the method for selecting model training set.

A. Model-based anomaly detection

In the proposed model-based anomaly detection method, normal DER behaviour is modelled in the *DER model* component (figure 2). The output of the model is compared to sensor measurements (or target data) in the *Anomaly Detection* component. Differences between normal and observed DER behaviour can originate form several sources: sensor error, model error, DER fault, or malicious or verified DER control. The output of the model-based anomaly detection is either a label (class) or an anomaly score for every data input.

Fig. 2. Flow diagram of the model-based anomaly detection

Ensemble learning combines several models to produce a prediction to solve classification and regression problems [14]. The increased robustness and accuracy of ensemble methods over single model methods was reported in [15]. Ensemble learning consists of three steps: generation, pruning and integration. First several redundant models are generated, then the set of models is pruned by removing some of the generated models, finally the base model results are combined to create the ensemble prediction [14]. An overview of ensemble regression approaches for generation, pruning and integration are presented in [14]. The ensemble is evaluated by the degree of agreement between predictions represented by their overall spread. The ensemble prediction is usually evaluated in terms of an average of the individual predictions (mostly using equal weight averaging).

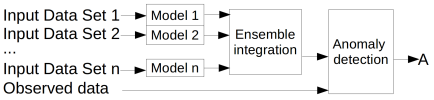

Fig. 3. Ensemble model-based anomaly detection architecture.

The proposed ensemble model-based anomaly detection (EM-AD) uses two or more DER normal behaviour models which produce the same output variables based on disjoint sets of inputs. The additional *Model merging* component calculates the final model output that is next compared to the observed output in the *Anomaly detection* component.

In this paper we apply the EM-AD method to a PV component and implement it as a proof of concept, using

historical time series of power and meteorological measurements obtained from a PV plant. The model building method is presented in section II-B.

B. Model building

The semi-supervised anomaly detection uses partially labelled data to train the normal model. Since the historical data has not been labelled, we use correlation analysis as a method for selecting a training set to improve the normal model and consequently enhance the anomaly detection performance. The chosen model building stages are as follows: data cleaning, aggregation, data scoring with correlation analysis, model data labelling and selection, removal of missing values, normalization, ANN model creation with supervised model training.

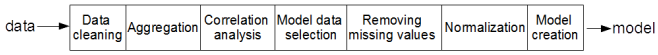

Fig. 4. The proposed model building method.

Figure 4 shows the data preparation and model building processes. The following section describes these processes in detail.

1) Data cleaning: Data cleaning detects and removes errors and inconsistencies in data in order to improve its quality [16]. In this paper, the observations of modelled phenomena are produced by sensors. Many errors can be hidden in raw sensor data; therefore the data needs to be cleaned before it can be used for modeling and analysis. The classification of data quality problems in data sources was proposed in [16]. According to this classification, the data can originate from a single measurement unit (single-source) or several measurement units (multi-source). Single source problems include schema and instance level issues. The schema level issues can be addressed with the mechanism of data storing and its integrity constrains. Instance-specific problems cannot be prevented at the schema level and include word misspellings and sensor errors. Multi-source data quality issues are consequences of integrating multiple sources of data. Resulting from conflicts due to different data models and representations, overlaps and contradictions can appear in the integrated data. In this work we consider both singe and multi-source data quality issues. In section IV-B of this paper we focus on single-source instance problems and will not cover sensor errors and multi-source instance level problems considering inconsistent timing.

2) Aggregation: The aggregation process targets two issues: model training time and missing values. Firstly the aggregation decreases the amount of the data that need to be processed to train the model, reducing the computation time for the model training. Secondly the column-wise aggregation based on mean or average on a matrix with some missing values (represented as NA) uses all available information from partially missing data samples. This process allows integrating the partially missing samples into the data set without removing them.

3) Correlation analysis: The analysis of data correlation serves two purposes: filtering data for the normal behaviour model and discovery of sensor faults. This data selection step is based on the assumption that the output of the model is correlated to one or more of its features. In this work we use the standard Pearson product-moment correlation coefficient of two variables. The proposed correlation analysis takes a defined subset of features and the model output and calculates its total correlation. Let $x_i^{(j,k)} = \{x_i^{(j)}, x_i^{(j+1)}, ..., x_i^{(j+k)}\}$ be a subset starting from sample $j \in N_0$ of size $k \in N_1$, where $k <= n$, of the i^{th} feature, and $y^{(j,k)} = \{y^{(j)}, y^{(j+1)}, ..., y^{(j+k)}\}$ is a variable that is the matching subset starting from sample j of size k of the output. The correlation Pearson product-moment correlation coefficient $corr(x_i^{(j,k)}, y^{(j,k)})$ is calculated as follows:

$$corr(x_i^{(j,k)}, y^{(j,k)}) = cov(x_i^{(j,k)}, y^{(j,k)})/(\delta x_i^{(j,k)} \delta y^{(j,k)}) \quad (1)$$

Where $cov(x_i^{(j,k)}, y^{(j,k)})$ is the covariance of variables x_i^k and y^k, and $\delta x_i^{(j,k)}$ and $\delta y^{(j,k)}$ are their respected standard deviations. The correlation calculated in equation 1 serves as a normality score for model data selection.

4) Model data selection: Samples of all features from the training set are evaluated based on the calculated correlation score. The proposed method allocates a sample $(x_i^{(j,k)} y^{(j,k)})$ into one of two groups: normal behaviour and suspicious behaviour. For a chosen $\alpha \in [0, 1]$, samples with $corr(x_i^{(j,k)}, y^{(j,k)}) > \alpha$ are allocated to normal behaviour group. If $corr(x_i^{(j,k)}, y^{(j,k)}) <= \alpha$ or if $corr(x_i^{(j,k)}, y^{(j,k)})$ does not exist, the samples are allocated to the suspicious behaviour group and are removed from the training set. Note that the correlation cannot be calculated if the standard deviation of $x_i^{(j,k)}$ or $y^{(j,k)}$ is zero. In this case the correlation is assigned a NA value. In sensor data this kind of feature can be observed for periods with long sensor failures.

5) Normalization: Vector normalization or scaling is usually performed before ANN model fitting. In this work normalisation was used. This is done to adjust values used for training by scaling them into the set $[0, 1]$. Large differences between values in the training set have an influence on the model weights which affects the model's ability to learn and aids generalization [17].

III. ANN MODEL CREATION

An artificial neural network (ANN) is a machine learning algorithm used to estimate unknown functions depending on several parameters. An ANN consists of interconnected neuronal nodes which perform simple calculations on outputs from neighbouring nodes in the previous layer. The result is passed to the next layer of the network.

We consider an ANN with $n \in N_1$ input variables $x = [x_0, x_1, x_2, ..., x_n]^\top$, where $x_n \in R$, and $x_0 = 1$ is a bias unit. The output variable of the considered ANN is $y \in R$. Let $a_i^{(j)}$ be the activation of neuron i in layer j, where $j \in 1, 2, .., l$, and l is the number of layers. $\Theta^{(j)}$ is a matrix of weights controlling the function mapping from layer j to layer $j + 1$. The considered hypothesis function approximated by the ANN is $h_\Theta(x) \in R$. Any layer L_j of the ANN consists of s_j neurons $a^{(j)} = [a_0^{(j)}, a_1^{(j)}, a_2^{(j)}, ..., a_{s_j}^{(j)}]^\top$. The size of the layer j can be different for every hidden layer. The input layer L_1 is of

size n, corresponding to the features vector. The output layer L_3 is of size 1 since the considered hypothesis function is $h_\Theta : R^n \rightarrow R$. The neural network architecture, including the number of inputs, outputs, layers and neurons in each layer, as well as the selection of the transfer function, describes an artificial neural network. Supervised learning methods for training ANN use the training examples $x_0, x_1, x_2, ..., x_n, y$ to calculate weight matrices $\Theta^{(1)}, \Theta^{(2)}, ..., \Theta^{(l-1)}$. The neural network architecture and the calculated weight matrices are jointly used for the approximation of an unknown function representing the relationship between input features and output variables. This way an artificial neural network can be trained to approximate transfer functions, especially unknown non-linear relationships.

A. ANN model training

The next step of data processing is the creation of an ANN model from the data by supervised training. The ANN training method chosen for training is called feed-forward training method (or forward propagation). Let's consider an ANN with $l \in N_1$ layers, $n \in N_1$ inputs, one output, a single set of model features $x = [x_0, x_1, x_2, ..., x_n]^\top$ and the output variable $y \in R$. Each layer L consists of S_l neurons. The neuron activation function is the sigmoid function, as defined in equation 2.

$$g(z) = 1/(1 + e^{-z}) \tag{2}$$

The Cyberenko theorem proves that the sigmoid function fulfills the universal approximation theorem which states that a single layer feed-forward artificial neural network can approximate continuous functions. The sigmoid activation function is therefore used to add non-linearity to the artificial neural network.

The forward propagation algorithm takes the vector x as an input and assigns it to the first layer $a^{(1)}$, therefore $a^{(1)} = x$. Neurons $a^{(2)}, a^{(3)}, ..., a^{(l)}$ can be constructed with the following vectorised equations: $\forall j \in [2, l] \quad a^{(j)} = g(\Theta^{(j-1)} a^{(j-1)})$.

The matrices $\Theta^{(1)}, \Theta^{(2)}, ..., \Theta^{(l-1)}$ are model weights. $\Theta^{(j)}$ is a matrix of weights controlling the function mapping from layer j to layer $j + 1$, for any $j \in (2, l)$, additionally $\Theta^{(j)} \in R^{s_{j+1} \times s_j + 1}$. Because the layer L_l is an output layer, $h_\Theta(x) = a^{(l)}$, therefore the hypothesis in the forward propagation algorithm is as follows:

$$h_\Theta(x) = g(\Theta^{(l-1)} g(\Theta^{(l-2)} ... g(\Theta^{(1)} x))) \tag{3}$$

Forward propagation takes the features $x_1, x_2, ..., x_n$ and modifies them with matrices $\Theta^{(1)}, \Theta^{(2)}, ..., \Theta^{(l-1)}$ and the sigmoid function g to create better suited features $a^{(1)}, a^{(2)}, ..., a^{(n)}$. In order to calculate the matrices $\Theta^{(1)}, \Theta^{(2)}, ..., \Theta^{(n-1)}$, the cost function J with least-squares fitting is described as follows:

$$J(\Theta) = \frac{1}{2n} \sum_{i=1}^{n} (h_\Theta(x^{(i)}) - y^{(i)})^2 \tag{4}$$

In feed-forward ANN, the problem of over-fitting can be solved with regularization [18] which is used to minimise the $\Theta^{(1)}, \Theta^{(2)}, ..., \Theta^{(l-1)}$ weights of the model. The cost function J with regularization is as follows:

$$J_R(\Theta) = \frac{1}{2n} \sum_{i=1}^{n} (h_\Theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2n} \sum_{k=1}^{l-1} \sum_{i=1}^{S_l} \sum_{j=1}^{S_{l+1}} (\theta_{i,j}^{(k)})^2 \tag{5}$$

where S_l is a number of units without a bias unit in the layer, λ is a regularization parameter called weight decay, and $\theta_{i,j}^{(k)}$ is an element of the matrix $\Theta^{(k)}$. Ripley [18] suggests to use $\lambda = 10^{-4} - 10^{-2}$ as a regularization parameter for least-squares fitting.

By minimising the cost function $J_R(\Theta)$, the ANN model weights $\Theta^{(1)}, \Theta^{(2)}, ..., \Theta^{(l-1)}$ can be computed. The Broyden–Fletcher–Goldfarb–Shannon (BFGS) algorithm [19] is used for solving the unconstrained nonlinear optimization problem of minimising the cost function $J_R(\Theta)$. While the BFGS algorithm is not guaranteed to converge, the Hessian matrix can be inspected in order to check if a secure local minimum has been found. There are many solutions to the optimisation problem and the weights are initialised at random at the start of the process, therefore the results might differ.

IV. PV MODEL ENSEMBLE GENERATION

The DER modeling process presented in section II-B is applied to a data set from a single PV inverter at the SYSLAB laboratory at the Technical University of Denmark. Two different models, a meteorological and a neighbourhood model, are presented in sections IV-C and IV-D and used in the EM-AD (section V).

A. Data sources

The data used for this study has been recorded from a 10kWp PV array located in Risø, Denmark in October 2014. The active power consumption data is recorded from the inverter in 1 second intervals. Meteorological data at the same time resolution - irradiation, temperature, wind speed and direction - is obtained from a meteorology mast about 600 m away from the PV site.

The data cleaning and preparation procedure presented in section II-B is used to pre-process the data before the ANN model can be trained. Since all presented models use the same data for training, the process of data cleaning is identical.

B. Data preparation

In this paper we focus on single-source instance problems removing discovered sensor errors and multi-source instance level problems considering inconsistent timing. Easily observed sensor failures result in missing data or false measurements. In the considered data set the observed false measurements were either constant values or inconsistent values, for example negative solar irradiation. Inconsistent value errors were present in the data set due to a sensor logging error. The threshold between consistent and inconsistent values is determined manually and inconsistent values are removed with the first filtering step.

TABLE I. DATA PROPERTIES OF SOLAR IRRADIATION (12.10.2014).

Data	Min.	1st Q	Median	3rd Q	Max.	NA's
Recorded	-31.450	-0.001	-0.001	0.014	31.450	0
Filtered	-0.001	-0.001	-0.001	0.015	0.087	4543

In the second filtering step, a 3rd order Butterworth low pass filter has been used to remove high frequency components appearing in the data due to sensor errors. The properties of the recorded and filtered data is presented in table I. Both the first and the third quadrant did not change significantly after the filtering process. The median remained the same, therefore it can be concluded that mostly outliers have been removed from the data set. This filtering process removed around 5% of the irradiation data set. Once the data is clean and uniform it can be aggregated to 1-minute values. The resulting time-series is then randomly divided into 3 sets: a training set D_t of size 14841, a validation set D_v of size 14901, and a cross-validation set D_{cv} of size 14898 samples.

C. Meteorological model

The meteorological model is based on the assumption that meteorological data can be used to model the yield of a PV panel. The available meteorological data (solar irradiation, wind speed, wind direction, ambient temperature) was used to construct the PV production model. The input significance analysis of the linear model based on the same inputs and outputs as the presented model, using test statistics under the null hypothesis, shows that all inputs are significant.

1) Model data selection with correlation analysis: The correlation between irradiation and yield data from the training set D_t was calculated using equation 1 with $k = 44640$ corresponding to a single day (see figure 5). The data set was extended with the normality score which is equal to the daily correlation. All data points with a score larger or equal to 0.2 were included in the normal behaviour model.

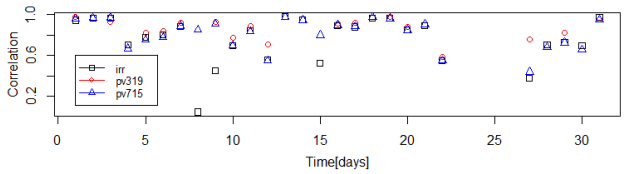

Fig. 5. Correlation analysis for irradiation, active power production from PV319 and PV715 with active power from PV117.

All data points for which the correlation could not be calculated due to zero standard deviation, were excluded from the training set. Based on correlation analysis, training data was selected for the ANN meteorological model, excluding data recorded on the 8th and 23-26th of October (figure6).

From the initial size of 14841, correlation analysis decreases the size of the training set D_t to 12480 rows. After excluding all rows where any of the data points is NA (removed by data cleaning), the size of the actual training set becomes 11739 rows. In the next step of the data preparation process, the training set is normalized.

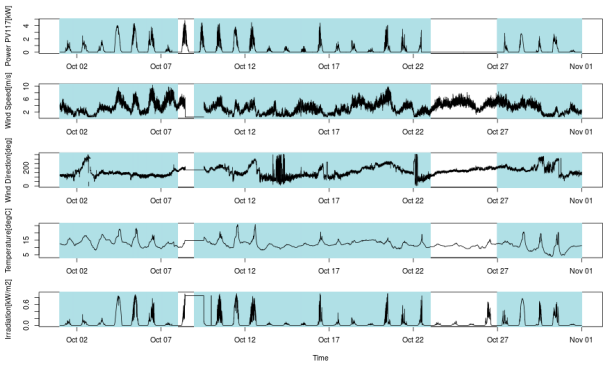

Fig. 6. Meteorological model normal training data set.

2) Model training: The proposed ANN network consists of 5 input neurons (representing solar irradiation, wind speed, wind direction, ambient temperature and time of day), one hidden layer with 10 neurons and a bias unit. The network has a single output neuron (PV117 power production) and 71 weights. The used transfer function g is sigmoid (as in equation 2) and the regularization parameter λ is set to 0.0006. The package *nnet* (Feed-Forward Neural Networks and Multinomial Log-Linear Models) [18] for the R scripting language has been used for the creation of the supervised learning ANN model. The *nnet* package calculates the Θ parameters of a single-hidden layer neural network, as described in section III-A. The model output for each input is presented in figure

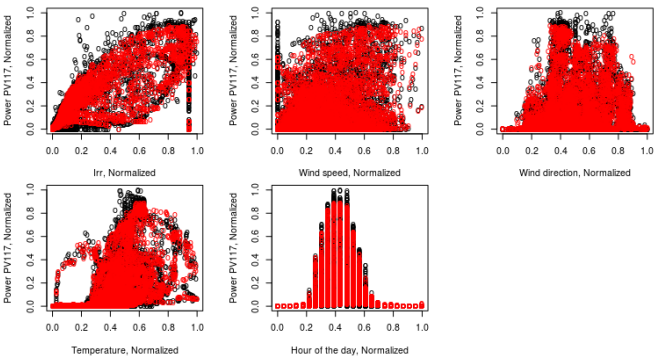

Fig. 7. Actual and predicted PV117 power consumption mapped to the inputs of the ANN meteorological model.

7. The root mean square error (RMSE) is calculated in order to evaluate the model. The RMSE for the training set is 1.13 versus 1.116 for the validation set and 1.118 for the cross-validation set. The small difference in RMSE between the validation and cross-validation sets indicates that the model generalises well.

D. Neighbourhood model

The proposed model uses data from two neighboring PV systems to take advantage of correlations between the three systems. In the SYSLAB laboratory the distance between

PV117 and PV715 is $630m$, compared to $340m$ between PV117 and PV319. In this investigation PV117 is being modelled. The correlation between the active power productions of PV117, PV319 and PV715 is calculated by using data from the training set D_t and equation 1 with $k = 44640$, corresponding to a single day (see figure 5). All days with a correlation larger

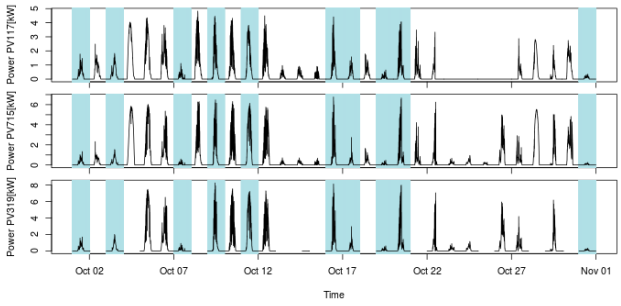

Fig. 8. Neighbourhood model training data with days selected by the correlation analysis in blue.

or equal to 0.2 have been included in the model. In this case, the 2nd, 4th, 8th, 13-15th, 21st, 23-26th and 30th of October have been removed from the training set. After the correlation analysis, the size of the training set D_t decreased from 14841 to 8201 samples.

The proposed ANN network consists of 3 input neurons (PV715 and PV319 power production and time of day), one hidden layer with 8 neurons and a bias unit. The network has a single output neuron (PV117 power production). The used transfer function g is sigmoid (as in equation 2) and the regularization parameter λ is 0.0006. Similarly to the meteorological ANN model, the R package *nnet* is used to find the Θ parameters of the model, as described in section III-A. The output of the model compared to its input with use of data from the validation set is presented in figure 9. The RMSE for the training set is 0.96, versus 0.570 for the

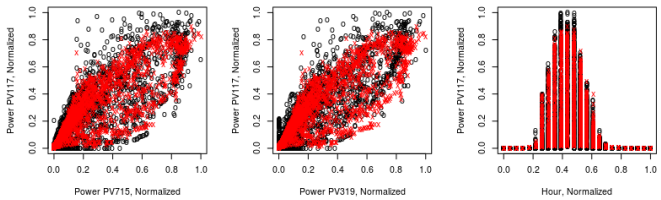

Fig. 9. Actual and predicted PV117 power consumption mapped to the model inputs for ANN neighbourhood model.

validation set and 0.572 for the cross-validation set. The small difference in RMSE values between the validation and cross-validation sets indicates that the model generalises well.

V. EM-AD for a PV plant

The architecture of the EM-AD is presented in figure 10. Sensor data of solar irradiation, wind speed, wind direction, ambient temperature, hour of day and power consumption of two neighbouring PVs (PV319 and PV715) are used as input. The proposed ensemble regression is composed of two regression models. The models were generated from disjoint parameter sets and a contextual parameter (hour of day), creating redundant heterogeneous ANN regression models of active power production as presented in sections IV-C and IV-D. The ensemble model set was not pruned because the set contains only two models. The ensemble integration is usually calculated as a linear combination of the predictions [14]. Here the ensemble power prediction P' is calculated from predictions for each model P^N and P^M as follows: $P' = \alpha P^N + \alpha P^M$, where $\alpha = 1/2$ corresponds to equal weight averaging.

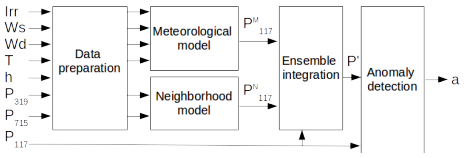

Fig. 10. Architecture of the proposed PV ensemble regression model anomaly detection (EM-AD)

The ensemble prediction P' is weighted with the anomaly score in the anomaly evaluation component. The anomaly score is based on the correlation analysis for both ANN models as presented in figure 5. Partial anomaly scores a_M and a_N are calculated for both models as in equation 6.

$$a = \begin{cases} 1 & corr \geq 0.2 \\ 10 & corr < 0.2 \end{cases} \quad (6)$$

The anomaly score a_s combines the partial scores for the models a_M and a_N and is calculated as $a_s = 1/(a_M \cdot a_N)$. The anomaly score a_s is multiplied by the difference between the ensemble prediction P' and measured power P to calculate the anomaly $a = a_s \cdot (P' - P)$. The chosen anomaly threshold is $\epsilon = 0.1$, therefore only observations with $a > \epsilon$ are considered.

VI. Results

In the considered scenario one month of the historical active power production of a single PV plant is analysed. In the analysed period of time the PV should have not been controlled, the considered anomalous cyber event is curtailment of the PV active power production to zero. The cyber event is being discovered by the proposed EM-AD, in this section we present the anomaly detection results for the EM-AD and compare it to other anomaly detection techniques.

The degree of agreement between the ensemble predictions is given by their overall spread ($s = P^N - P^M$) with the first and third quadrant at -0.011 and 0.006, respectively, and a standard deviation of 0.511. This indicates that the models generally agree in their predictions. Table II presents nine approaches for model-based anomaly detection which were performed using the October 2014 PV data set. The evaluated models are M (meteorological), N (neighbourhood), MN (joint model with inputs from M and N), EMN (ensemble of M and N). The used training sets are: cor (correlated days for the data set), full (entire data set). Two anomaly detection

TABLE II. RESULTS CONFUSION MATRIX AND STATISTICAL MEASURES

Model	Train	AD	TP	TN	FP	FN	ACC	PPV	NPV	FNR	TPR	TNR	FPR	FDR
M	full	M-AD	1198	37445	4531	1466	0.866	0.209	0.962	0.550	0.450	0.892	0.108	0.791
M	cor	M-AD	1633	41305	671	1031	0.962	0.709	0.976	0.387	0.613	0.984	0.016	0.291
N	full	M-AD	1543	39242	2734	1121	0.914	0.361	0.972	0.421	0.579	0.935	0.065	0.639
N	cor	M-AD	1736	41660	316	928	0.972	0.846	0.978	0.348	0.652	0.992	0.008	0.154
MN	full	M-AD	764	38723	3253	1900	0.885	0.190	0.953	0.713	0.287	0.923	0.077	0.810
MN	cor	M-AD	764	41756	220	1900	0.953	0.776	0.956	0.713	0.287	0.995	0.005	0.224
EMN	full	M-AD	1447	38730	3246	1217	0.900	0.308	0.970	0.457	0.543	0.923	0.077	0.692
EMN	cor	M-AD	1709	38614	3362	955	0.903	0.337	0.976	0.358	0.642	0.920	0.080	0.663
EMN	cor	EM-AD	1709	41880	96	955	0.976	0.947	0.978	0.358	0.642	0.998	0.002	0.053

methods are used: M-AD (model based anomaly detection) and EM-AD (ensemble regression model anomaly detection). The confusion matrix is a compilation of instances of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), evaluated from a population of results. To measure the correctness of the anomaly detection we calculate eight significant measures: accuracy (ACC), precision (PPV), negative predictive value (NPV), false negative rate (FNR), sensitivity (TPR), specificity (TNR), false positive rate (FPR) and false discovery rate (FDR).

The proposed EM-AD with the correlation training selection approach achieves an accuracy of 0.976, which improves the accuracy by 0.4-11.1% for single model AD, 2.3-9.2% for joint model AD, and 7.3-7.6% over the method without correlation ensemble integration. The precision of the proposed method is 0.947, which improves the precision by 23.8-73.8% for single model AD, 10.1-58,6% for joint model AD, and 61-63.8% over the method without correlation ensemble integration. EM-AD with correlation training data selection additionally keeps low values for FNR of 0.358, FPR of 0.002 and FDR of 0.053. While the specificity has improved only by 10.6% at best, totalling to 0.998, the sensitivity is 0.642 which presents an improvement of up to 35.5% over other presented methods.

VII. CONCLUSION

This paper proposes a novel ensemble model anomaly detection method with non-linear regression models and anomaly scores based on correlation analysis (RM-AD) used for cyber-physical intrusion detection in smart grids. The models are presented and evaluated and the ensemble integration and anomaly detection methods are described in detail. A proof-of-concept RM-AD analysing a data set from a PV plant is presented and compared to other M-AD approaches. Future work will include automatic ensemble model set generation, an investigation into whether a larger ensemble can improve the prediction accuracy, and alternative interpolation methods for missing data.

This research was conducted as part of the SALVAGE project (Cyber-physical security for low-voltage grids) funded by ERA-Net Smart Grids.

REFERENCES

[1] "Security Guideline for the Electricity Sector: Physical Security," North American Electric Reliability Corporation(NERC):Critical Infrastructure Protection Committee, Tech. Rep., June 2012.

[2] A. McIntyre, "Renewable systems interconnection study: Cyber security analysis," Sandia Natl. Lab., Tech. Rep., February 2008.

[3] J. Moteff, "Risk management and critical infrastructure protection: Assessing, integrating, and managing threats, vulnerabilities and consequences," in *Library of Congress Washington DC Congressional Research Service*. DTIC Document, 2005.

[4] G. N. Ericsson, "Cyber security and power system communication — essential parts of a smart grid infrastructure," *Power Delivery, IEEE Trans.*, vol. 25, no. 3, pp. 1501–1507, 2010.

[5] F. M. Cleveland, "Cyber security issues for advanced metering infrastructure (AMI)," in *PES GM - Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE*. IEEE, 2008, pp. 1–5.

[6] D. Yang, A. Usynin, and J. W. Hines, "Anomaly-based intrusion detection for SCADA systems," in *5th Intl. Topical Meeting on Nuclear Plant Instrumentation, Control and Human Machine Interface Technologies*, 2006, pp. 12–16.

[7] S. Sridhar, A. Hahn, and M. Govindarasu, "Cyber–physical system security for the electric power grid," *Proc. of the IEEE*, vol. 100, no. 1, pp. 210–224, 2012.

[8] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[9] M. Sanz-Bobi, A. M. San Roque, A. de Marcos, and M. Bada, "Intelligent system for a remote diagnosis of a photovoltaic solar power plant," in *Journal of Physics: Conference Series*, vol. 364, no. 1. IOP Publishing, 2012, p. 012119.

[10] A. Zaher, S. McArthur, D. Infield, and Y. Patel, "Online wind turbine fault detection through automated SCADA data analysis," *Wind Energy*, vol. 12, no. 6, p. 574, 2009.

[11] A. A. Cárdenas, S. Amin, Z.-S. Lin, Y.-L. Huang, C.-Y. Huang, and S. Sastry, "Attacks against process control systems: risk assessment, detection, and response," in *Proceedings of the 6th ACM symposium on information, computer and communications security*. ACM, 2011, pp. 355–366.

[12] A. M. Kosek, "Contextual anomaly detection for cyber-physical security in Smart Grids based on an artificial neural network model," in *Joint Workshop on Cyber-Physical Security and Resilience in Smart Grids (CPSR-SG2016), CPSweek 2016*, 2016.

[13] SALVAGE project - Cyber-physical security for low-voltage grids. http://salvage-project.com. Accessed: 2016-01-03.

[14] J. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. D. Sousa, "Ensemble approaches for regression: A survey," *ACM Comput. Surv.*, vol. 45, no. 1, pp. 10:1–10:40, Dec. 2012.

[15] N. García-Pedrajas, C. Hervás-Martínez, and D. Ortiz-Boyer, "Cooperative coevolution of artificial neural network ensembles for pattern classification," *Evolutionary Computation, IEEE Transactions on*, vol. 9, no. 3, pp. 271–302, 2005.

[16] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, 2000.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[18] B. D. Ripley, *Modern applied statistics with S*. Springer, 2002.

[19] J. D. Head and M. C. Zerner, "A Broyden— Fletcher —Goldfarb —Shanno optimization procedure for molecular geometries," *Chemical physics letters*, vol. 122, no. 3, pp. 264–270, 1985.

Contextual anomaly detection for cyber-physical security in Smart Grids based on an artificial neural network model

Anna Magdalena Kosek
Energy System Operation and Management
Department of Electrical Engineering
Technical University of Denmark
amko@elektro.dtu.dk

***Abstract*—This paper presents a contextual anomaly detection method and its use in the discovery of malicious voltage control actions in the low voltage distribution grid. The model-based anomaly detection uses an artificial neural network model to identify a distributed energy resource's behaviour under control. An intrusion detection system observes distributed energy resource's behaviour, control actions and the power system impact, and is tested together with an ongoing voltage control attack in a co-simulation set-up. The simulation results obtained with a real photovoltaic rooftop power plant data show that the contextual anomaly detection performs on average 55% better in the control detection and over 56% better in the malicious control detection over the point anomaly detection.**

***Keywords*—anomaly detection, intrusion detection system, smart grid, data analysis, cyber-physical security**

I. Introduction

Cyber security is an increasing interest and worry in power systems. The main concerns consider new control paradigms, on-line access to a range of power system components and DERs (Distributed Energy Resources), and enormous data exchange and collection introduced by the so called Smart Grid. The power system security is mostly concerned with cyber security of AMI (Advanced Metering Infrastructure), [1], SCADA (Supervisory Control and Data Acquisition) security [2] and communication standards [3]. A new cyber-physical approach to the smart grid security was introduced in [4] and addresses a tight coupling between the physical power system and the ICT (Information and Communication Technology). Nine major research topics emerged from the combination of these two fields: vulnerability research, impact analysis, mitigation research, cyber-physical metrics, data and model developments, security validation, interoperability, cyber forensics and operator training [5]. A smart grid compatible IDS (Intrusion Detection System) needs to address both on-line and post-mortem analysis of the state of the observed cyber-physical system and detect anomalies in operation of both cyber and physical components. An anomaly detection method identifies rare data instances or events that do not match an expected pattern [6]. The development of models used for anomaly detection requires cyber-security and power system expertise, and additionally, if data driven models are required, data analysis knowledge. Once both cyber and physical anomaly detection analysis is performed, cyber-physical metrics need to be developed to combine the information from both domains to address the tight relations between the power system and the ICT domains. Anomaly detection with regression models has been used for discovering cyber-attacks on a SCADA system [7], a wind turbine fault detection [8], a PV (photovoltaic power plant) fault diagnostics [9]. A special case of a PV attack against voltage control in distribution power grids has been described in [10].

Two types of anomaly detection can be distinguished: point and contextual. The point anomaly detection takes the global view of the data [6]. The contextual or conditional anomalies were introduced in [11] and are defined as data points that are anomalous in a specific context and acceptable in another context. For example for spatial data, the location of a measurement is its context. For time series, time is the context for each measurement [6]. The advantage of the contextual over point anomaly detection is the detection accuracy. The disadvantage is that this method requires context data, which is not always available. Two methods for contextual anomaly detection exist: reduction to a point anomaly detection problem and utilizing the structure in data [6]. The reduction to point anomaly detection problem technique divides the data into contextual groups and analyses behaviour attributes for each context separately, reducing the problem to several point anomaly detections. Thos method produces a model for each context, as a consequence several models are used to represent a single system. In case of the time contextual data, models for every year, month, day of the month, minute and so on would have to be created. Contextual anomaly models utilising the structure of the data modify the structure the training data to include the date adding separately: year, month, day and so on as input variables, the modified input data is then used for training of a single contextual model.

In the energy domain the contextual anomaly detection have been previously used for recognising user behavior in a residential dwelling based on non-parametric belief propagation for energy efficiency [12]. In [12] a user behaviour is categorised as unusual equipment usage or bursty occupancy and is used to adjust the energy management schedule. Authors

978-1-5090-1164-3/16/$31.00 ©2016 IEEE

in [13] propose use of on-line contextual anomaly detection for fault diagnostics of power transformers. In this paper we propose to use contextual anomaly detection utilizing the structure in data for cyber-physical IDS. According to the author's knowledge, contextual anomaly detection have not been used to identify control actions.

II. Anomaly based IDS

The proposed cyber-physical IDS architecture consists of two main parts: an analysis of the behaviour of the observed cyber-physical system and components, and a joint analysis of the cyber-physical system (figure 1). The behaviour analysis and characterisation of the physical power system is performed with two components: DER and power system analysis, the evaluation of the cyber vulnerabilities is performed in the cyber security analysis component. The joint cyber-physical analysis combines the information from both physical and cyber security components and presents the outcomes to the power system operator. In this work we consider the physical

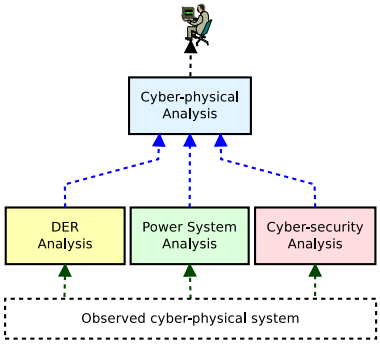

Fig. 1. Cyber-physical IDS architecture.

part of the proposed IDS and focus on DER and power system analysis. Additionally this paper introduces an on-line method to combine information produced by these components in the cyber physical analysis component. The proposed on-line anomaly based IDS architecture is presented in figure 2. The IDS consists of three parts: DER, power system and cyber-physical analysis.

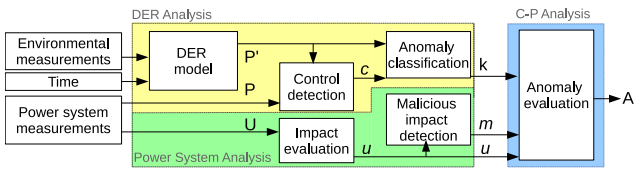

Fig. 2. IDS with anomaly detection and power system stability evaluation.

A. DER analysis

The objective of the DER analysis is to identify a suspicious behaviour of a DER unit and associate it with a DER control action. The DER analysis consists of: a DER model, a control detection mechanism and an anomaly classification. In order to detect if a DER is being controlled, an anomaly in its behaviour need to be discovered. An anomaly based detection uses a normal behaviour model predicting a DER power production or consumption (P') and compares it to the power measurement from the DER (P). The difference between these values is identified as anomaly $\alpha = P' - P$. Since the model introduces errors to anomaly detection, a threshold $\tau = 0.1$ was chosen to eliminate some of the model errors. Additionally this paper considers controllable power production (when P is negative), where curtailment is the only possible control action, therefore all positive $\alpha > 0$ are treated as an error. The final anomaly c associated with a curtailment action is as follows:

$$c = \begin{cases} 1 & \alpha > \tau \\ 0 & \alpha \leq \tau \end{cases} \tag{1}$$

The anomaly classification checks if the discovered anomaly is within the possible DER operation time β, therefore:

$$\beta = \begin{cases} 1 & P' < 0 \\ 0 & P' \geq 0 \end{cases} \tag{2}$$

The anomaly classification produces output $k = c\beta$ identifying all significant and possible curtailments of a DER, here classified as control anomalies k.

B. Power system analysis

Power system analysis consists of two components: impact evaluation and malicious impact detection. The impact evaluation depends on the attack hypothesis, in this paper the considered attack influences the power stability by causing under- or over-voltage. The impact analysis takes under consideration the voltage limits and creates a piece-wise function u evaluating measured voltages U. Let's consider n as the nominal voltage value and $0.9n$ is considered under-voltage and $1.1n$ is over-voltage, the proposed function is as follows:

$$u(U) = \begin{cases} 1 & U \geq 0.9n \\ -20n(x - 0.9n) & 0.9n < U < 0.95n \\ 0 & 0.95n \leq U \leq 1.05n \\ 20n(x - 0.05n) & 1.05n < U < 1.1n \\ 1 & U \leq 1.1n \end{cases} \tag{3}$$

The impact evaluation function u is presented in figure 3. The

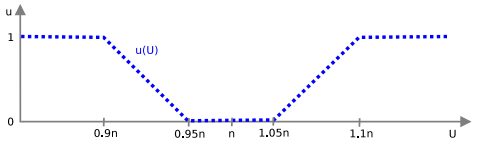

Fig. 3. Impact evacuation function.

output of the impact evaluation function is used in malicious impact detection component that evaluates the voltage difference u_{diff} for each time t: $u_{diff} = u_{t-1} - u_t$. The output of the malicious impact detection component u_{diff} is a measure of the state of the voltage from one point in time to another. If the u_{diff} is positive, when $u_{t-1} > u_t$, the voltage have

improved between time $t-1$ and t. If the u_{diff} is negative, when $u_{t-1} < u_t$, the voltage have changed and is closer to under- or over-voltage between time $t-1$ and t. If the u_{diff} is equal to zero, the voltage have not significantly changed, it might have changed inside of the $\pm 0.05n$ range or have not changed at all. The impact classification m is as follows:

$$m = \begin{cases} 1 & u_{diff} < 0 \cup (u_{diff} = 0 \cap u = 1) \\ 0 & otherwise \end{cases} \quad (4)$$

The impact is classified as malicious if the voltage has changed towards under or over-voltage. Additionally, we assume that the malicious impact occurs in case the under or over-voltage is present and that this state didn't change between time $t-1$ and t.

C. Cyber-physical analysis

Cyber-physical analysis evaluates the recognised anomalies. The DER analysis provides the control evaluation k, the power system analysis component brings the malicious impact evaluation m and impact estimation u. In this work, we focus on the following three control anomaly cases: **normal control**, when $m = 0$, $u = 0$ and $k = 1$; **suspicious control**, when $m = 1, u > 0$ and $k = 1$; and **malicious control**, when $m = 1, u = 1$ and $k = 1$. The proposed anomaly based IDS and its three main analysis components have been tested in simulation, the implementation and results are presented in sections III, IV and V.

III. DER MODEL

This paper proposes use of contextual anomaly detection for DER analysis and evaluates its use on an example DER: a residential rooftop PV panel. The time is used as a contextual attribute for PV production prediction as shown in figure 4. The anomalous behaviour is defined as PV's response to a control signal resulting in curtailment of its power production. The model input data is described in section III. Several modeling

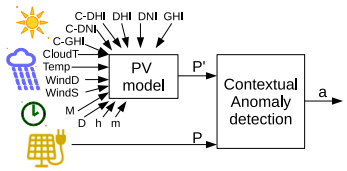

Fig. 4. Contextual anomaly detection with a PV mode.

approaches can be used to model a normal behaviour of a PV or any other DER. Three types of popular models are white, gray and black box models [14]. White box models use the known physical properties of the system. Gray box model and black box models both use the available system input and outputs to determine the system model. Gray box model combines a partial theoretical structure or partial physical system model and use model training methods to estimate parameters. In the black box model, only inputs and outputs of the system are know and the operation of the system is unknown. Machine learning methods are used to identify properties of the system by supervised model training. The black box model of a DER might be less accurate than a white box model, but can be adjusted to any unknown DER, that need to be observed and modelled. In this paper we investigate the use of machine learning technique, specifically artificial neural network (ANN), to model a DER. PV arrays have been previously modelled with use of ANN [15]. In the context of power system ANN was used together with anomaly detection in only in few cases: distribution feeder fault detection [16], detecting anomalies at substation level of abnormal measurements [17]. The design of the PV model presented in this paper is based on availability of input data for its training. Similar PV models have been proposed in numerous publications, for example a PV model using irradiation, ambient temperature, voltage, active power and current training set was proposed in [18]. According to author's knowledge no other contextual ANN PV models have been developed, where time is considered as a context.

A. Model training data

1) PV power production data: The real PV production data was recorded by the Pecan Street Smart Grid Demonstration Program project that started in 2010. The objective was to implement an open platform Energy Internet Demonstration [19] with real residential consumers. The primary sight of the demonstration was at Austins Mueller community in Austin, Texas. One of the project outcome is a Dataport[1] database containing anonymized data of home electricity use, PV power, EV charging, and demand response data recorded while participating in the utility programs. The PV active power production was recorded by an energy monitoring system from eGauge. The considered solar power production used in this research is a rooftop PV produced by SunEdison, from a single-family home (referred in Dataport as house 774) in Austin, Texas. The data used in this research is 1 minute active power production in kW from 1st January 2013 to 31st January 2014.

2) Meteorological data: The Meteorological data was acquired from National Solar Radiation Data Base (NSRDB)[2] developed by NREL (National Renewable Energy Laboratory). The used data comes from a meteorological station in Texas, Austin (latitude 30.29, longitude -97.7) from 1st January 2013 to 1st February 2014. The data is recorded every 30 minutes, the chosen data points, defined in the Glossary of Solar Radiation Resource Terms [3], are as follows. Diffuse Horizontal Irradiance (**DHI**) $[w/m^2]$ (diffuse sky radiation) - the radiation component that strikes a point from the sky, excluding circumsolar radiation. Direct Normal Irradiance (**DNI**) $[w/m^2]$ (beam radiation) - the amount of solar radiation from the direction of the sun. Global Horizontal Radiation (**GHI**) $[w/m^2]$ (global horizontal irradiance) total solar radiation. Three measures of clear sky irradiance: clear sky diffuse horizontal irradiance, direct normal radiance and global horizontal radiation (**C-DHI**, **C-DNI** and **C-GHI**) $[w/m^2]$ - measurement of DHI, DNI and GHI excluding the influence of clouds. Cloud type (**CloudT**)

[1] https://dataport.pecanstreet.org/
[2] https://nsrdb.nrel.gov/
[3] http://rredc.nrel.gov/

is another available meteorological data, NSRDB records 13 cloud types: clear, probably clear, fog, water, super-cooled water, mixed, opaque ice, cirrus, overlapping, overshooting, unknown, dust, smoke. Additional meteorological information are ambient temperature (**Temp**) [c], wind direction (**WindD**) [Degrees], and wind speed (**WindS**) [m/s].

3) Contextual attributes: The time stamp from each measurement was transformed into a vector M, D, h, m, where $M \in [1, 12]$ is a month, $D \in [1, 31]$ is a day, $h \in [0, 23]$ is an hour and $m \in [0, 59]$ is a minute. The relationship between contextual attributes (hour and month) and power production is presented in figure 5. For the purpose of this study, the training set was combined from the weather and PV production data together with the time information from 1st January 2013 to 31st December 2013. A total of 525540 data rows were divided into 80% training set (420660 data rows) and 20% validation set (104880 data rows). Before the 30 minute meteorological data was combined with 1 minute power production and time data, linear interpolation was performed on the weather data. This training set was used for the PV model training. The data from 1st January 2014 to 31st January 2014 was used in the simulation as on-line data. The PV production data used for simulation was modified in order to simulate PV control. In the simulation, an instantaneous and constant curtailment is assumed.

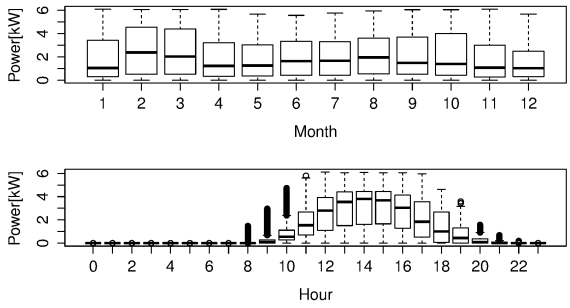

Fig. 5. Box plot of time of the day and month, and PV power production.

B. ANN models

Let's consider a single layer feed-forward ANN with $n \in N_1$ inputs, one output, and a single set of model features $x = [x_0, x_1, x_2, ..., x_n]^\top$ and output variable $y \in R$. The hidden layer consists of $h \in N_1$ neurons. In order to train the ANN, the forward propagation algorithm is used. The ANN model hypothesis is as follows:

$$H_w(x) = w_0 + \sum_h w_{1h}\phi(\alpha_h + \sum_k w_{kh}x_i) \qquad (5)$$

Here, w is model weights, ϕ_0 is the output function and ϕ_1 is the activation function. In this work the neural network is build to model a non-linear continuous function. According to Cyberenko theorem, sigmoid activation function of a single layer feed-forward ANN fulfills the universal approximation theorem, therefore a ANN with sigmoid activation function, as in equation (6), can approximate continuous functions.

$$\phi(z) = 1/(1 + e^{-z}) \qquad (6)$$

The weights are chosen to minimise the cost function with least squares. In forward-feed ANN problem of over-fitting can be minimised with regularization [20], that is used to minimise the weights of the model, the cost function J with regularization is as follows:

$$J(w) = \sum_i ||H_w(x^{(i)}) - y^{(i)}|| + \lambda \sum_h \sum_k w_{kh}^2 \qquad (7)$$

Ripley [20] suggests to use $\lambda = 10^{-4} - 10^{-2}$ as a regularization parameter for least-squares fitting. Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [21] was used for solving unconstrained nonlinear optimization problem of minimising the the cost function $J(w)$.

The point ANN model (ANN-P) consists of 10 input neurons 15 hidden neurons and one output neuron. The regularization parameter $\lambda = 0.0006$. The contextual ANN model (ANN-P) consists of 14 input neurons 20 hidden neurons and one output neuron. The regularization parameter is $\lambda = 0.0006$. Both numbers of the hidden neurons and the regularization parameter for each model were chosen to minimise the root mean square error (RMSE) of the model prediction. As presented in table I the ANN-C model is more accurate than the ANN-P model based on RMSE.

TABLE I. COMPARISON OF THE DEVELOPED MODELS

Model				Residuals		
Name	Prm	Ctxt prm	RMSE	Min	Median	Max
ANN-P	10	-	0.88	-3.82	-0.14	4.25
ANN-C	14	m, D, H, M	0.43	-3.11	0	3.81

IV. SIMULATION

A co-simulation set-up was used to obtain results in this paper. The co-simulation [4] combines the PV, house and meteorological station emulators, power load flow solver PYPOWER part of the MATPOWER package for Python, implementation of monitors and attacker. Open source co-simulation orchestrator mosaik[5] synchronises the operation of all simulations and programs and exchanges data between them. Two scenarios were chosen to demonstrate the IDS system presented in section II. Both scenarios consider operation a LV distribution grid and consists of two feeders with houses and rooftop PVs. For each scenario two use cases are presented: normal operation and under attack. Use cases test hypothesis that an attacker controls PV operation in order to influence voltage on the line, leading to reduction of power quality. An autonomous monitor with IDS proposed in this paper observes each PV plant and tests the scenario hypothesis. The objective of the monitor in each use case is to determine if the PV control leads to over- or under-voltage on the line. According to EN50160 European standard the nominal value of voltage in LV grid is

[4]https://pypi.python.org/pypi/PYPOWER
[5]http://mosaik.offis.de/

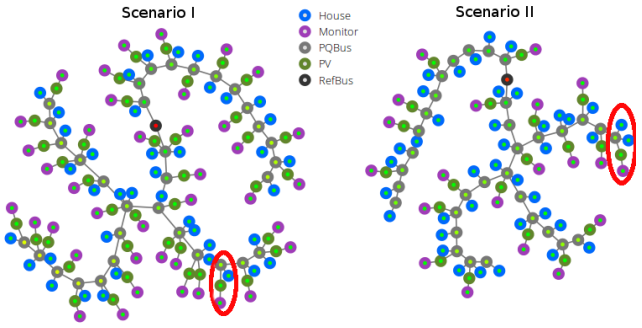

Fig. 6. System configuration for Scenario I and II.

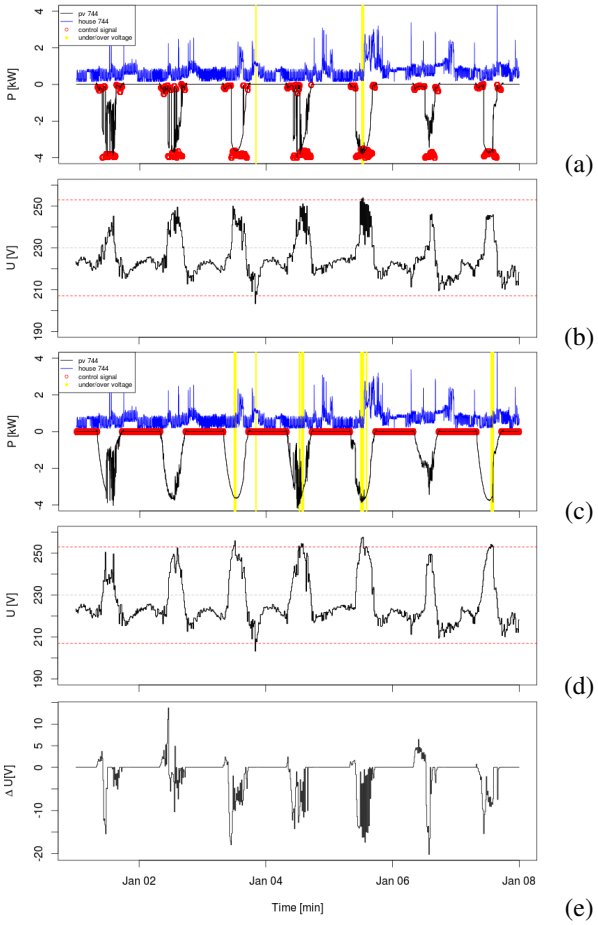

Fig. 7. Scenario I: house and PV load pattern (a) and voltage (b)- normal behaviour; house and PV load pattern (c) and voltage (d)- behaviour under attack; (e) voltage difference between the normal behaviour and the attack.

$230v$, over-voltage is defined as 10% increase of the nominal voltage ($253v$), under-voltage is 10% decrease of the nominal voltage ($207v$). In this section we present monitoring results from a single PV (refereed to PV number 744 in [19]). Two types of monitors: contextual and point anomaly detection have been implemented in the co-simulation set up.

Scenario I considers 100% residential PV penetration. The system configuration used for this scenario consists of 40 houses and PVs, divided into two feeders 12 sets of houses and PVs on feeder A and 28 sets of houses and PVs on feeder B (see figure 6). Ten houses and corresponding PVs have been created from real house data III-A1 and replicated to create 40 prosumers. The actors in the normal operation use case are: houses, PVs, monitors and an aggregator. The aggregator reads the voltage from each PQbus (connection point to the grid from both house and the PV) and curtails the PV in case the voltage is reaching over-voltage. The outcome of the aggregator operation is presented in figure 7(a). In total 45 minutes of the operation voltage problems are visible (30 minutes over-voltage and 15 minutes under-voltage). In the use case under attack, the actors are as follows: houses, PVs, monitors and an attacker. The attacker gathers information about the active power production of each PV and voltage on each PQbus. The attacker sends control signals to each PV in order to reach either under or over-voltage. It is visible in figure 7(c) that the attacker's decision was not to curtail the PV operation and increase over-voltage, as presented in 7(d). The voltage problems increased to 240 minutes (where 225 minutes of over-voltage and 15 minutes of under-voltage). The difference between voltages for the normal operation and the attack use case is presented in figure 7(e), it is visible that the voltage is mostly decreased in this scenario.

In Scenario II 50% of the houses are equipped with rooftop PVs. The system configuration for this scenario is as follows: it consists of 40 houses and 20 PVs are divided into two feeders 12 houses and 5 PVs on feeder A and 28 houses and 15 PVs on feeder B (see figure 6). Similarly to the normal use cases from Scenario I, the aggregator is controlling the PV in order to meet the voltage limits, as presented in figures 8(a,b). There are several voltage problems: 15 minutes of over-voltage and 135 minutes of under-voltage. In the attack use case, the attacker is aiming at increasing the over- and under-voltage minutes by controlling the PV. It is visible in figure 8(c) that attacker decides to curtail the PV744 to $0kW$, which leads to a decrease in voltage. The total number of voltage problems is increased to 420 which all minutes are under-voltage. The voltage difference between use cases in the Scenario II is presented in figure 8(e). It is visible that voltage have been significantly decreased in this scenario.

V. Results

Two presented models are tested for each use case in two scenarios. The results are divided into accuracy of the control detection and overall results of the malicious control detection. A confusion matrix and accuracy calculations are used to evaluate the control and attack results. The confusion matrix is a collection of occurrences of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN) evaluated from a population of results. The accuracy is calculated as follows:

$$Acc = (TP + TN)/(TP + FP + FN + TN) \quad (8)$$

As seen in table II, the accuracy of the control action detection for point detection ranges between 0.39 and 0.58, where

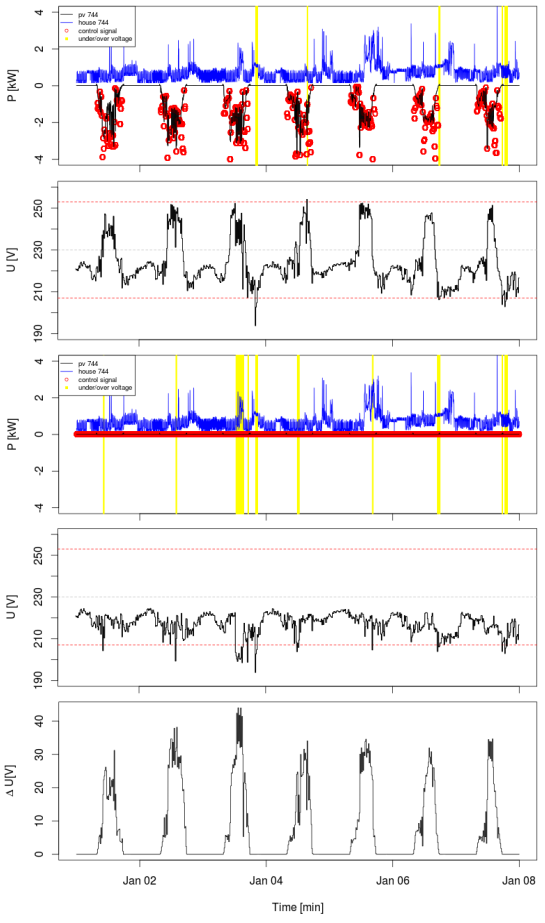

Fig. 8. Scenario II: house and PV load pattern (a) and voltage (b)- normal behaviour; house and PV load pattern (c) and voltage (d)- behaviour under attack; (e) voltage difference between the normal behaviour and the attack.

contextual anomaly accuracy is between 0.79 and as much as 0.94 for attack use case in Scenario II. Both methods recognised less control actions during attack in Scenario I than in Scenario II. On average the accuracy of detection for the contextual method increases by 0.37 over the point method that accounts to 55% in the presented scenarios. As presented in table III, the discovery of malicious control is performed well by both point and contextual detection, scoring 0.99 or 1 accuracy. For the attack in Scenario II both methods have 0.93 accuracy. However the attack case of the Scenario I is more problematic or both methods however, contextual anomaly recognised 4 times more true positives than point anomaly detection, increasing the accuracy by 56%.

VI. Conclusion

An on-line IDS detecting malicious DER control two attack on voltage scenarios in the LV grid is described and tested in this paper. The IDS consists of a DER analysis with a contextual anomaly detection and a power system analysis with an impact analysis. The simulation results obtained from the

TABLE II. Confusion matrix of control detection

Use case	TP	TN	FP	FN	Acc
contextual anomaly					
Scenario I: normal	2033	5956	218	1873	0.79
Scenario I: attack	6	8863	1208	3	0.88
Scenario II: normal	2034	5956	218	1872	0.79
Scenario II: attack	3498	5953	213	416	0.94
point anomaly					
Scenario I: normal	1770	2194	3980	2136	0.39
Scenario I: attack	9	5498	4573	0	0.55
Scenario II: normal	1466	2194	3980	2440	0.36
Scenario II: attack	3693	2194	3972	221	0.58

TABLE III. Confusion matrix of malicious control detection

Use case	TP	TN	FP	FN	Acc
contextual anomaly					
Scenario I: normal	0	45	0	0	1
Scenario I: attack	44	15	0	181	0.25
Scenario II: normal	0	150	0	0	1
Scenario II: attack	249	141	0	30	0.93
point anomaly					
Scenario I: normal	0	45	0	0	1
Scenario I: attack	11	15	0	214	0.11
Scenario II: normal	0	149	1	0	0.99
Scenario II: attack	249	140	1	30	0.93

chosen scenarios confirm that a contextual anomaly detection is more accurate than point anomaly detection.

In the present implementation the IDS analysis in limited to a simple voltage use case. A more broad analysis modules need to be added for other power system malicious control. The presented DER model is calculated from the near past historical data, in the next implementation the model needs to be recalculated periodically or be based on a large set of data. The presented IDS is designed to a local produce the IDS only associated with a control of a single DER. If the underlying model is recalculated periodically the ANN training execution complexity should be considered. Additionally the presented co-simulation set-up allows implementation of different attack profiles, future work can include implementation of different attack profiles.

Acknowledgements

This research has been conducted as part of the SALVAGE project (Cyber-physical security for low-voltage grids) funded via ERA-Net SmartGrids programme.

References

[1] F. M. Cleveland, "Cyber security issues for advanced metering infrastructure (ami)," in *PES GM - Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE*. IEEE, 2008, pp. 1–5.

[2] A. Creery and E. Byres, "Industrial cybersecurity for power system and scada networks," in *Petroleum and Chemical Ind. Conf., 2005. Industry Appl. Soc. 52nd Annual*. IEEE, 2005, pp. 303–309.

[3] G. N. Ericsson, "Cyber security and power system communication essential parts of a smart grid infrastructure," *Power Delivery, IEEE Trans.*, vol. 25, no. 3, pp. 1501–1507, 2010.

[4] Y. Mo, T. H.-J. Kim, K. Brancik, D. Dickinson, H. Lee, A. Perrig, and B. Sinopoli, "Cyber–physical security of a smart grid infrastructure," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 195–209, 2012.

[5] A. Hahn, A. Ashok, S. Sridhar, and M. Govindarasu, "Cyber-physical security testbeds: Architecture, application, and evaluation for smart grid," *Smart Grid, IEEE Transactions on*, vol. 4, no. 2, pp. 847–855, 2013.

[6] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[7] D. Yang, A. Usynin, and J. W. Hines, "Anomaly-based intrusion detection for SCADA systems," in *5th intl. topical meeting on nuclear plant instrumentation, control and human machine interface technologies (npic&hmit 05)*. Citeseer, 2006, pp. 12–16.

[8] A. Zaher, S. McArthur, D. Infield, and Y. Patel, "Online wind turbine fault detection through automated SCADA data analysis," *Wind Energy*, vol. 12, no. 6, p. 574, 2009.

[9] M. Sanz-Bobi, A. M. San Roque, A. de Marcos, and M. Bada, "Intelligent system for a remote diagnosis of a photovoltaic solar power plant," in *Journal of Physics: Conference Series*, vol. 364, no. 1. IOP Publishing, 2012, p. 012119.

[10] Y. Isozaki, S. Yoshizawa, Y. Fujimoto, H. Ishii, I. Ono, T. Onoda, and Y. Hayashi, "On detection of cyber attacks against voltage control in distribution power grids," in *Smart Grid Communications (SmartGridComm), 2014 IEEE International Conference on*. IEEE, 2014, pp. 842–847.

[11] X. Song, M. Wu, C. Jermaine, and S. Ranka, "Conditional anomaly detection," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, no. 5, pp. 631–645, 2007.

[12] Z. Zhao, W. Xu, and D. Cheng, "User behavior detection framework based on nbp for energy efficiency," *Automation in Construction*, vol. 26, pp. 69–76, 2012.

[13] V. M. Catterson, S. D. McArthur, and G. Moss, "Online conditional anomaly detection in multivariate data for transformer monitoring," *Power Delivery, IEEE Transactions on*, vol. 25, no. 4, pp. 2556–2564, 2010.

[14] M. E. Khan and F. Khan, "A comparative study of white box, black box and grey box testing techniques," *Int. J. Adv. Comput. Sci. Appl*, vol. 3, no. 6, 2012.

[15] F. Almonacid, C. Rus, L. Hontoria, M. Fuentes, and G. Nofuentes, "Characterisation of si-crystalline pv modules by artificial neural networks," *Renewable Energy*, vol. 34, no. 4, pp. 941–949, 2009.

[16] S. Ebron, D. L. Lubkeman, and M. White, "A neural network approach to the detection of incipient faults on power distribution feeders," *Power Delivery, IEEE Transactions on*, vol. 5, no. 2, pp. 905–914, 1990.

[17] M. Martinelli, E. Tronci, G. Dipoppa, and C. Balducelli, "Electric power system anomaly detection using neural networks," in *Knowledge-Based Intelligent Information and Engineering Systems*. Springer, 2004, pp. 1242–1248.

[18] A. El Shahat, "Pv cell module modeling & ann simulation for smart grid applications," *Journal of Theoretical and Applied Information Technology*, vol. 16, no. 1, pp. 9–20, 2010.

[19] "Pecan Street Smart Grid Demonstration Program - Final Technology Performance Report," Pecan Street Inc., Tech. Rep., February 2015.

[20] B. D. Ripley, *Modern applied statistics with S*. Springer, 2002.

[21] J. D. Head and M. C. Zerner, "A BroydenFletcherGoldfarbShanno optimization procedure for molecular geometries," *Chemical physics letters*, vol. 122, no. 3, pp. 264–270, 1985.