Improved Anomaly Detection and Localization Using Whitening-Enhanced Autoencoders

Wang, C.; Tindemans, Simon H.; Palensky, P.

DOI
10.1109/TII.2023.3268685

Publication date
2024

Document Version
Accepted author manuscript

Published in
IEEE Transactions on Industrial Informatics

Citation (APA)
Wang, C., Tindemans, S. H., & Palensky, P. (2024). Improved Anomaly Detection and Localization Using Whitening-Enhanced Autoencoders. *IEEE Transactions on Industrial Informatics*, *20*(1), 659-668. https://doi.org/10.1109/TII.2023.3268685

Improved Anomaly Detection and Localization Using Whitening-Enhanced Autoencoders

Chenguang Wang, *Graduate Student Member, IEEE*, Simon H. Tindemans, *Member, IEEE*, and Peter Palensky, *Senior Member, IEEE*

***Abstract*— Anomaly detection is of considerable significance in engineering applications, such as the monitoring and control of large-scale energy systems. This paper investigates the ability to accurately detect and localize the source of anomalies, using an autoencoder neural network-based detector. Correlations between residuals are identified as a source of misclassifications, and whitening transformations that decorrelate input features and/or residuals are analyzed as a potential solution. For two use cases, regarding spatially distributed wind power generation and temporal profiles of electricity consumption, the performance of various data processing combinations was quantified. Whitening of the input data was found to be most beneficial for accurate detection, with a slight benefit for the combined whitening of inputs and residuals. For localization of anomalies, whitening of residuals was preferred, and the best performance was obtained using standardization of the input data and whitening of the residuals using the *ZCA* or *ZCA-cor* whitening matrix with a small additional offset.**

***Index Terms*— Anomaly detection, Autoencoder, Renewable generation, Whitening transformation**

I. Introduction

MONITORING and control of large-scale engineering systems require accurate measurements and dependable communication infrastructure – and methods to process that data for operational awareness. An important example is the case of electrical power systems that are increasingly reliant on variable renewable generation [1]. In this context, it is important to detect anomalies in high-dimensional, highly variable observations from a multitude of sensors. For example, mild reductions in power generation caused by a wind turbine component malfunction or physical disturbance. Insufficient performance of anomaly detectors may threaten both the economic dispatch and secure control of power systems [2].

In recent years, with the development of deep neural network-related technologies, an unsupervised data-driven approach for anomaly detection has been proposed in the form of an autoencoder-based classifier [3]. It considers anomaly detection as a one-class classification task by learning patterns of normal operating states. This is well-suited to the inherent data imbalance in anomaly detection applications and the fast-evolving power grid [4]. On this basis, techniques have been designed to detect anomalies in renewable energy systems using autoencoder-based detectors. For example, in [5]–[7], the authors have proposed an autoencoder neural network to analyze anomalies of wind turbine components using power generation or other SCADA data. However, the basic autoencoder-based anomaly detector is based on thresholding of residuals (reconstruction errors) using a Euclidean distance metric. This does not account for significant dependencies between measurements, such as the spatial and temporal correlations of renewable resources [8]. The mismatch between the detector design and features of the data could have a negative impact on detection sensitivity and localization performance.

In view of this, some authors have proposed using the Mahalanobis distance has been utilized to measure the *residuals* and thus acquire more accurate classification boundaries for autoencoder-based anomaly detectors [9]. Authors of [10]–[12] reported autoencoder-based wind turbine fault detectors using the Mahalanobis distance. However, it has not yet been investigated how the modified detection boundaries impact the anomaly localization performance.

Apart from adjusting residuals, the correlated renewable generation data, which are the *input* of autoencoder network, can be decorrelated and standardized (i.e. whitened [13]) before fed into the autoencoder. This technique has been considered in the field of computer vision, with a focus on image and video data sets, for example, the image retrieval [14] and object recognition [15]. However, there has been little quantitative analysis of detection sensitivity and localization performance improvement in the context of utilizing an autoencoder-based detector with input data whitening [16]. Further, what is not yet clear is the impact of processing the inputs and residuals *together* on the capacity of a detector.

This paper bridges these identified gaps by investigating the impact of whitening input data and residuals and quantifying the improvement of detection sensitivity and localization performance using our proposed metrics. This is done in the context of two high-dimensional energy system use cases. The main contributions of this paper are listed below:

1) Comparative studies of different data processing methods, neural network configuration schemes, and whitening matrix selections are carried out, and their influences on anomaly detection sensitivity and localization accuracy are quantified using a variety of metrics.
2) We propose a combined whitening of the input features *and* of autoencoder residuals, which is shown to

This work was supported by the Chinese Scholarship Council. The authors are with the Department of Electrical Sustainable Engineering, Delft University of Technology, 2628 CD, Delft, The Netherlands (emails: {c.wang-8, s.h.tindemans, p.palensky}@tudelft.nl).

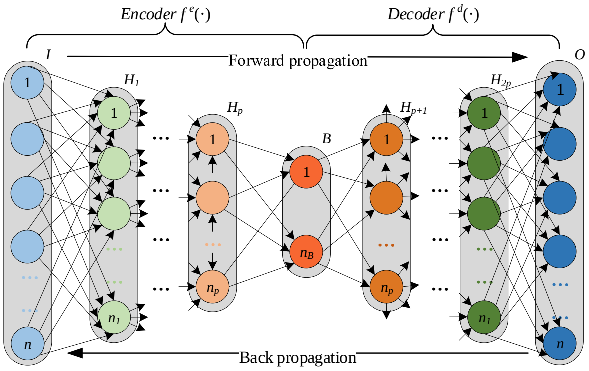


Fig. 1. The schematic of the autoencoder.

maximize detection sensitivity in two high-dimensional use cases: one for spatially correlated renewable wind power generation and one for electricity consumption time series.

3) A combination of input feature standardization and *ZCA-cor*- or *ZCA*-based residual whitening is shown to enhance the visibility of anomalies and thus achieve an outstanding localization performance of an anomaly detector. The performance is further enhanced by a tunable offset to the whitening transformation.

II. Anomaly Detection Mechanism

Anomaly detection is essentially a classification problem with the objective of distinguishing anomalous data from data that is considered 'normal' [17]. The most common approach is to treat anomaly detection as a supervised learning task, e.g. using SVMs (Support Vector Machines) [18] or deep neural network classifiers [19]. However, supervised learning requires a training data set with representative normal system operations and anomalies. Such data sets are in short supply because of the rarity of anomalies, unwillingness to share data, and evolving anomalies. Thus, it is difficult to learn a satisfactory discriminator of 'normal' and 'anomalous' scenarios on this basis [20].

Alternatively, anomaly detection can be approached as a one-class classification problem (e.g. using a one-class SVM approach [21]), where the detector is trained on examples of only 'normal' operation data using an autoencoder-based neural network. Observations with features that deviate substantially from those in the training data will be considered anomalies. There are two main advantages to this approach. First, the autoencoder-based mechanism avoids the need to gather or generate anomalous data to create balanced data sets for training the classifiers. Second, by focusing on what is normal only, the proposed mechanism is naturally prepared for unknown anomaly patterns.

A. Autoencoder Training

Autoencoders learn the most important features of the training data (i.e. normal power system measurements) by sending the measurements through an information bottleneck while attempting to reconstruct the training data with minimal error [3]. The structure of the autoencoder algorithm is depicted in Fig. 1. The dimension reduction process of mapping the n-dimensional input data to the code in the bottleneck layer B through hidden layers H_1 to H_p is named the *encoder*. Afterwards, the *decoder* decompresses the code to n-dimensional output data. Weight matrices K and bias vectors b are utilized in the encoding and decoding process as

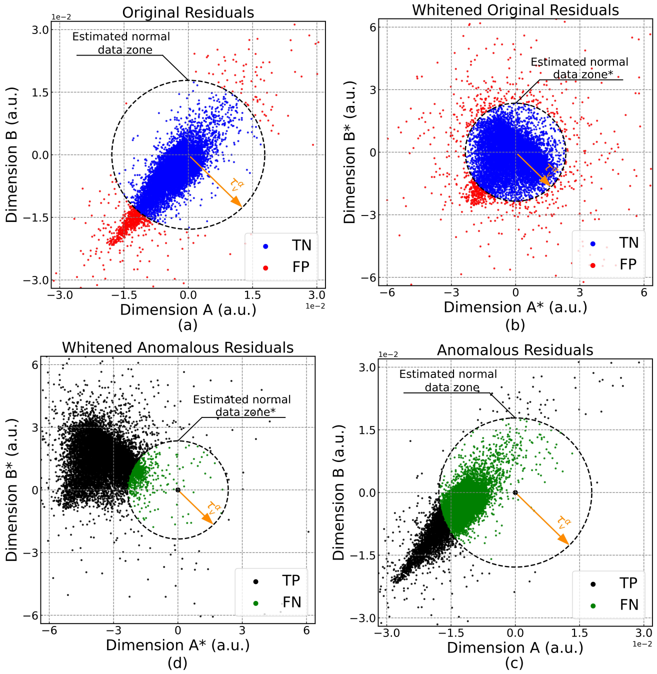


Fig. 2. Illustrative two-dimensional distributions of: (a) original residuals, (b) whitened original residuals, (c) synthetic anomalous residuals obtained by shifting, and (d) anomalous residuals whitened according to the normal data. The residuals are classified as TN (True Negatives), FP (False Positives), TP (True Positives), and FN (False Negatives) by comparing with a 95% threshold (α=95) calculated on a validation set.

$$x = g(K_p^e(\dots g(K_0^e z + b_0^e)\dots) + b_p^e)\,, \tag{1a}$$

$$\hat{z} = g(K_p^d(\dots g(K_0^d x + b_0^d)\dots) + b_p^d)\,, \tag{1b}$$

where K_p^e and K_p^d denote weight matrices for encoding and decoding process respectively, b_p^e and b_p^d are bias vectors, and $g(\cdot)$ represents a nonlinear element-wise activation function. $z \in \mathbb{R}^n$ refers to the input data vector, x is the data in the bottleneck layer B and vector $\hat{z} \in \mathbb{R}^n$ stands for the output.

The residual vector associated with a training observation z_i is given by $r_i = z_i - \hat{z}_i$. The corresponding reconstruction error R_i is commonly expressed as the 2-norm of r_i, and the objective of the training process is to minimize the mean value of all reconstruction errors $\mathcal{R}_i$ as

$$\min_{K,\,b} \quad \left\{ J := \tfrac{1}{k} \sum_{i=1}^{k} \|r_i\|_2 \right\}, \tag{2}$$

where k denotes the total number of the observations used for the autoencoder training process.

B. Anomaly Classification

In this work, an anomaly is defined as an observation that does not match the patterns inferred from data that are considered normal. Specifically, the observation is in a region of observation space that is estimated to contain less than a predefined fraction of data – thereby limiting the false positive rate of the assignment of observations to anomalies.

After the convergence of the objective J, the trained autoencoder is utilized to encode and decode the validation data, resulting in reconstruction errors $\mathcal{R}_v$, which are used to determine an anomaly threshold τ_v^α equal to the α^{th} percentile of $\mathcal{R}_v$; for example, at the value where an 'inflection point' occurs in the error distribution [17]. Finally, the reconstruction errors $\mathcal{R}_e$ of the test data are compared with τ_v^α to classify states into normal ($R_e \leq \tau_v^\alpha$) and anomalous ($R_e > \tau_v^\alpha$) data.

Geometrically, R_i denotes the Euclidean distance between the input z_i and reconstructed data $\hat{z}_i$. The τ_v^α stands for the maximum distance that a data point can be considered normal. Notably, the spatial set that is not further than τ_v^α is defined as an 'estimated normal data zone'. It is an n-ball in the space of residuals, with radius τ_v^α and centered on the origin.

C. Problem Formulation

Power system measurements exhibit spatial-temporal dependencies. For example, due to geographic factors, the scale and irradiance of renewable resources such as wind and solar are spatially dependent within a given region [8]. Autoencoder-based neural networks are trained to replicate these correlated inputs on the output side with minimal reconstruction errors.

Dependencies in inputs may also lead to correlated residuals. An example in Fig. 2 shows two dimensions of the residuals obtained in the case study of section IV. The residuals of normal test data shown in Fig. 2 (a) are classified into TN (True Negatives) and FP (False Positives) by the threshold τ_v^α. However, the assumption of a circular 'estimated normal data zone' is not appropriate for this ellipsoidal distribution. Fig. 2 (c) illustrates this with simulated anomalous data that is obtained by shifting the residuals. Many clearly anomalous points are within the normal circle and therefore not detected (FN, False Negatives). This reduces the probability that an actual anomaly is identified: the true positive rate (TPR), also known as detection sensitivity. The illustration in two dimensions also applies to residuals in higher dimensions.

III. Detector Enhancements

A. Data Whitening for Performance Improvement

In view of the elliptically distributed residuals and concomitant errors in anomaly detection, whitening (also known as sphering) the observations is a promising approach to improve detection performance. By removing the correlations between the residual components, the n-ball may better describe the normal data distribution, and anomalies may be detected more accurately. The potential effectiveness of this approach is depicted in Fig. 2 (b,d). Whitening can be applied in three different combinations of two approaches as:

- Whitening of the input data;
- Whitening of generated residuals;
- Combined whitening of the input data and residuals.

1) Whitening Transformations: We first summarize properties of the whitening transformation. Consider a random vector $Z = (z_1, \ldots, z_n)^T$, with the (non-singular) covariance matrix $\mathrm{Cov}(Z, Z) = \Sigma \in \mathbb{R}^{n\times n}$. We define the (also non-singular) *whitening transformation matrix* $W \in \mathbb{R}^{n\times n}$ such that

$$V = (V_1, \ldots, V_n)^T = WZ, \tag{3}$$

where the elements of the random vector v are uncorrelated and have unit variance: $\mathrm{Cov}(V, V) = \mathbf{1}$. We determine constraints on W by expanding

$$\begin{aligned}\mathrm{Cov}(V, V) &= \mathrm{E}[WZ(WZ)^T] - \mathrm{E}[WZ]\mathrm{E}[(WZ)^T] \\ &= W\big(\mathrm{E}[ZZ^T] - \mathrm{E}[Z]\mathrm{E}[Z^T]\big)W^T = W\Sigma W^T.\end{aligned} \tag{4}$$

This implies the constraint $W\Sigma W^T = \mathbf{1}$. As W is invertible, we multiply the W^{-1} and $(W^T)^{-1}$ from the left and right, respectively, and find after inversion:

$$W^T W = \Sigma^{-1}. \tag{5}$$

This does not determine the whitening matrix W uniquely. Among the infinite possible options of whitening matrices W, a few are commonly used [22]. In this paper, four approaches are studied: *PCA*, *ZCA*, *Cholesky*, and *ZCA-cor* [13].

The *PCA whitening* transformation is a widely used sphering approach due to its close relation to *principle component analysis* (*PCA*) [23]. It can be regarded as rescaling variances of all dimensions to one after a *PCA* procedure that omits the customary dimension reduction. The whitening matrix is

$$W_{PCA} = \Lambda^{-1/2} U^T, \tag{6}$$

where $\Lambda \in \mathbb{R}^{n\times n}$ is a diagonal matrix with the eigenvalues of covariance matrix Σ and the columns of $U \in \mathbb{R}^{n\times n}$ are the corresponding eigenvectors. It is closely related to the *ZCA* approach, which uses U to transfer the *PCA* whitened data back to the original coordinate system [13]:

$$W_{ZCA} = U\Lambda^{-1/2} U^T. \tag{7}$$

ZCA whitening has been used in various applications, such as a data pre-processing step for stochastic gradient decent [24] and image classification with convolutional neural networks [25]. The *Cholesky whitening* transformation is defined as

$$W_{Chol} = L^T, \tag{8}$$

where $L \in \mathbb{R}^{n\times n}$ is a lower triangular matrix with positive diagonal entries, obtained by *Cholesky* decomposition of the precision matrix (inverse covariance matrix): $\Sigma^{-1} = LL^T$. The final sphering approach considered in this paper is *ZCA-cor whitening* transformation [13]. It uses the whitening matrix

$$W_{ZCA-cor} = S^{-1/2} V^{-1/2}, \tag{9}$$

where $V \in \mathbb{R}^{n\times n}$ is the diagonal variance matrix and $S \in \mathbb{R}^{n\times n}$ denotes the correlation matrix (so that $\Sigma = V^{1/2} S V^{1/2}$). The *ZCA-cor whitening* approach maximizes the correlation of whitened and original components [13]. Unlike W_{ZCA}, $W_{ZCA-cor}$ is in general asymmetric.

In this paper, we consider both detection and localization performance of the anomaly detector. As the whitening procedure is transparent to the calculation of residual vectors (the

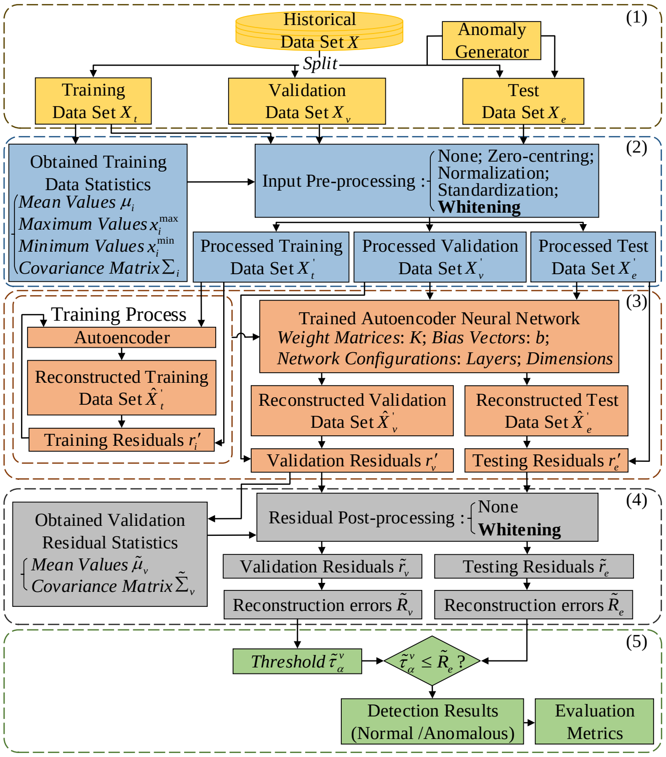

(1)
Historical Data Set X
Anomaly Generator
Split
Training Data Set X_t
Validation Data Set X_v
Test Data Set X_e
(2)
Obtained Training Data Statistics
Mean Values μ_i
Maximum Values x_i^{max}
Minimum Values x_i^{min}
Covariance Matrix Σ_i
Input Pre-processing :
None; Zero-centring; Normalization; Standardization; Whitening
Processed Training Data Set X_t'
Processed Validation Data Set X_v'
Processed Test Data Set X_e'
(3)
Training Process
Autoencoder
Reconstructed Training Data Set $\hat{X}_t'$
Training Residuals r_t'
Trained Autoencoder Neural Network
Weight Matrices: K; Bias Vectors: b; Network Configurations: Layers; Dimensions
Reconstructed Validation Data Set $\hat{X}_v'$
Reconstructed Test Data Set $\hat{X}_e'$
Validation Residuals r_v'
Testing Residuals r_e'
(4)
Obtained Validation Residual Statistics
Mean Values $\tilde{\mu}_v$
Covariance Matrix $\tilde{\Sigma}_v$
Residual Post-processing :
None
Whitening
Validation Residuals $\tilde{r}_v$
Testing Residuals $\tilde{r}_e$
Reconstruction errors $\tilde{R}_v$
Reconstruction errors $\tilde{R}_e$
(5)
Threshold $\tilde{\tau}_\alpha^v$
$\tilde{\tau}_\alpha^v \leq \tilde{R}_e$?
Detection Results (Normal /Anomalous)
Evaluation Metrics

Fig. 3. The proposed framework of data flow in the autoencoder neural network-based anomaly detector.

squared vector is used), the particular choice of W mostly affects the localization performance. It may also affect the detection performance, albeit indirectly, if whitening is applied to the input data, thus affecting the training of the autoencoder.

2) Input Whitening: The whitening transformation can be utilized to remove correlations from the data used for training and testing. First, this enables the autoencoder network to learn from less redundant inputs, which is generally desirable [15]. But more importantly, we hypothesize that the reduction in the input correlation may propagate to the residuals.

We consider a *whitened input data point*

$$z_w = W_z(z - \mu_z), \tag{10}$$

with W_z, the z-space whitening matrix, computed from the sample covariance of the training data z. The data is also (optionally) centered on the data mean μ_z. The residual $r_w \in \mathbb{R}^n$ is defined as $r_w = z_w - \hat{z}_w$, where $\hat{z}_w$ is the reconstructed data point. Inverting the whitening procedure gives $\hat{z}_a = W_z^{-1}\hat{z}_w + \mu_z$, which may be compared with the original z. The reconstruction error of the whitened data $\|r_w\|_2$ can be related to that of $r_a = z - \hat{z}_a$ as:

$$\begin{aligned}\|r_w\|_2 &= \|z_w - \hat{z}_w\|_2 = [(z-\hat{z}_a)^T W_z^T W_z (z-\hat{z}_a)]^{1/2} \\ &= [(z-\hat{z}_a)^T \Sigma_z^{-1}(z-\hat{z}_a)]^{1/2} \triangleq \|r_a\|_{\Sigma_z^{-1}}. \end{aligned} \tag{11}$$

Compared with (2), by taking correlations of original inputs into account, we are effectively measuring the Mahalanobis distance [26] between z and $\hat{z}_a$ instead of their standard Euclidean length. Whitening of the input data thus affects both the representation of the training data as well as the loss function used during training.

3) Residual Whitening: In contrast with applying whitening transformation before feeding data into the neural network, *residual whitening* reshapes the distribution of residuals *for a given trained autoencoder*. Concretely, the raw residual $r = [r_1, \ldots, r_n]^T$ is whitened as

$$r_s = W_r(r - \mu_r). \tag{12}$$

Here, the whitening matrix $W_r \in \mathbb{R}^{n\times n}$ is computed on the sample covariance of residuals from the *validation* data, because the training data set is used to train the autoencoder itself. μ_r represents the mean of raw residuals in the validation set, which should be approximately zero if the RMSE loss function was used during training. Accordingly, the reconstruction error is given by

$$\|r_s\|_2 = [(r-\mu_r)^T\Sigma_r^{-1}(r-\mu_r)]^{1/2} \triangleq \|r-\mu_r\|_{\Sigma_r^{-1}}. \tag{13}$$

Slightly different from (11), the reconstruction error in (13) denotes the Mahalanobis distance of a residual r from a set of residuals with mean μ_r and covariance matrix Σ_r.

B. Anomaly Localization Metrics

In many scenarios, when a likely anomaly has been detected, it is important to also identify which observation(s) triggered the anomaly detector. They may indicate a component malfunction or source of the physical disturbance. In the case study that follows, we will show that with a well-chosen whitening procedure, the values of the residual vector can be used to pinpoint the anomaly: the highest absolute residuals are the most likely locations of anomalies.

To quantify the dependability of the localization performance, we propose three metrics. The first of these is the *RMS Ratio*, which denotes the ratio of root mean square value of anomalous dimensions to that of normal dimensions in residual vectors. This is given by

$$\mathit{RMS\ Ratio} = \frac{1}{m}\sum_{l=1}^{m}\{[\frac{1}{|\mathcal{A}|}\sum_{j\in\mathcal{A}}(r_j^{(l)})^2]^{\frac{1}{2}}/[\frac{1}{|\mathcal{N}|}\sum_{j\in\mathcal{N}}(r_j^{(l)})^2]^{\frac{1}{2}}\}, \tag{14}$$

where $r_j^{(l)}$ denotes the j^{th} element of the residual of the l-th test data point. $\mathcal{A}$ is the set of anomalous dimensions (e.g., malfunctioning devices), and $\mathcal{N}$ represents the set of non-anomalous dimensions. Moreover, m refers to the total number of records in the test set. The *RMS Ratio* can be used to estimate if the anomalous stand out on average. A second, more stringent metric is introduced as well: *Gap Ratio*, which measures the average ratio of the smallest anomalous dimension to the largest normal dimension (absolute value). It is defined as

$$\mathit{Gap\ Ratio} = \frac{1}{m}\sum_{l=1}^{m}\{\min_{j\in\mathcal{A}}|r_j^{(l)}|/\max_{j\in\mathcal{N}}|r_j^{(l)}|\}. \tag{15}$$

The third metric, *OCR* (Ordinal Consistency Rate), calculates the proportion of samples for which the smallest absolute value of an *anomalous* dimension exceeds the largest absolute value of a *non-anomalous* dimension:

$$\mathit{OCR} = \frac{1}{m}\sum_{l=1}^{m}\mathbb{1}_{\min_{j\in\mathcal{A}}|r_j^{(l)}|>\max_{j\in\mathcal{N}}|r_j^{(l)}|}. \tag{16}$$

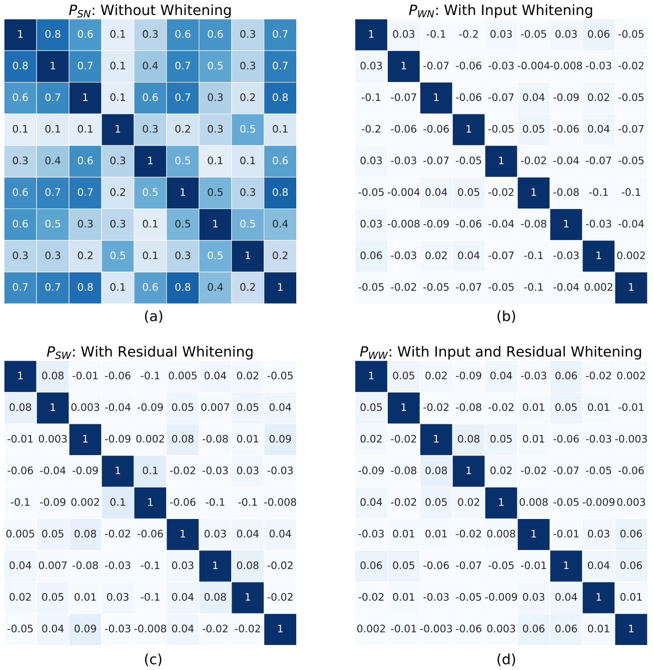


Fig. 4. The correlation coefficient matrices of normal testing residuals $\tilde{\boldsymbol{r}}_{\boldsymbol{e}}$ when utilizing four data processing combinations. The dimensions shown correspond to nine virtual wind farms ($\boldsymbol{B_{64}}$ is utilized).

C. Design of the Anomaly Detector

We integrate the anomaly detection mechanism and whitening schemes to give data processing options as well as explain which data is used in different stages. The proposed data flow and its transformation processes in the autoencoder neural network-based anomaly detector are depicted in Fig. 3 with the following five steps.

1) Data Partition: Given the historical data set X, the first step is to divide observations into training, validation, and test data set as X_t, X_v, and X_e with a specific ratio.

2) Input Pre-processing: In this step, statistics of the training data (mean, range, covariance) are computed, and these values are used for input processing of training, validation, and test data according to the selected method, e.g. input whitening shown in (10).

3) Training and Reconstruction: The weight matrices K and bias vectors b are updated iteratively to minimize the reconstruction loss in (2). Afterwards, the trained autoencoder neural network is utilized to reconstruct the validation and test set to $\hat{X}'_v$ and $\hat{X}'_e$. Then, the corresponding residual sets r'_v and r'_e are calculated, such as r_w in (11).

4) Residual Post-processing: If residual whitening is employed, the validation residuals are used to compute the whitening transformation. After executing residual whitening transformation shown in (12), the reconstruction errors of the validation set $\tilde{R}_v$ and test set $\tilde{R}_e$ are calculated as (13).

5) Detection Performance Evaluation: The anomaly threshold $\tilde{\tau}_v^\alpha$ is obtained as a quantile of the reconstruction errors $\tilde{R}_v$, corresponding to the desired true negative rate $\alpha\%$. The test data are classified by comparing reconstructions $\tilde{R}_e$ with the threshold $\tilde{\tau}_v^\alpha$. Consequently, the performance of the anomaly detector is assessed by calculating the evaluation metrics based on the predicted normal, anomalous states and the actual states.

IV. Case Study: Snapshot Data

In this section, the impacts of different data processing options, neural network configurations, and whitening matrix selections on anomaly detection and localization capacity of an autoencoder-based detector were investigated. To do so, we conducted a case study on the power generation from distributed wind farms. Although introduced anomalies are synthetic, the wind farm output is based on reanalysis data of historical wind speeds. This use of maximally realistic high-dimensional data ensures that the data won't be easily compressed into a low-dimensional manifold by an autoencoder.

To do so, we first described the process of modelling normal data patterns and then formulated anomalous scenarios. For these scenarios, we made anomaly detection performance comparisons by implementing various combinations of data transformations including whitening. Next, anomaly locating capacity evaluation was conducted by comparing four whitening transformations: *PCA*, *Cholesky*, *ZCA*, and *ZCA-cor*.

A. Experiment Scenario Formulation

Renewable power generation from spatially distributed wind farms is an increasingly relevant source of energy. The power output of each wind farm is highly variable due to variations in wind speed, but this may obscure other factors causing reduced performance. Given this, experiments were conducted to test if our proposed mechanism can detect and localize anomalies in the power output of wind farms with satisfactory capacity. Notably, without knowing any model-related information about wind farms and relying on the neural network-based data-driven methodology only, our proposed anomaly detection mechanism was trained on historical operation data and tested on both normal and anomalous scenarios. In this paper, we generated anomalous scenarios as reductions in power output of one or more wind farms. These could reflect unexpected malfunctions, disturbances, unscheduled outages, or unreported maintenance activities (from the perspective of system operators).

A realistic wind power data set was constructed as follows. We virtually placed a 100MW wind farm at each center of the 99 municipalities located in the North and South Holland provinces of the Netherlands. The wind power output was simulated on the basis of historical wind speeds at the 99 locations, obtained by MERRA-2 reanalysis and available from renewables.ninja [27]. After obtaining the historical wind data [28], the associated generated power outputs were calculated as described in [29]. For the purpose of this study, observations were snapshots of instantaneous power production. Ultimately, the whole generated data set $X \in \mathbb{R}^{87648\times 99}$, which includes 10 years' (2009-2018) hourly outputs of 99 wind farms, was divided into the training set X_t, validation set X_v and test set X_e with the proportion of 6, 2, and 2 years.

The autoencoder encoded and decoded the 99-dimensional data from the input to output layer. Both encoder and decoder were fully connected networks with 3 hidden layers with 200 neurons each, connected to a bottleneck layer with variable size. The bottleneck size is indicated as B_n, where n is the number of neurons in the bottleneck layer. The ReLU activation function was used, and the Adam Optimizer [30] was utilized to iteratively optimize the value of weight matrices K and bias vectors b. In this research, 5×10^3 training epochs were used, and the learning rate for training was 5×10^{-5}. Training and testing of the autoencoder were conducted using `tensorflow`.

For a comparative study of anomaly detection and localization performance, we made use of different combinations of input pre-processing methods and residual post-processing. The combinations, denoted as P_{xy}, are listed in Table. I.

TABLE I
DATA PROCESSING METHOD COMBINATIONS.

	Input Pre-processing	Residual Post-processing
P_{NN}	None	None
P_{SN}	Standardization	None
P_{WN}	Whitening	None
P_{NW}	None	Whitening
P_{SW}	Standardization	Whitening
P_{WW}	Whitening	Whitening

B. Impact of Whitening Transformation

Fig. 4 depicts the correlation coefficients of the testing residuals $\tilde{r}_e$ for 9 out of the 99 dimensions, for a variety of data processing methods. When the input data is standardized, but no whitening is performed (P_{SN}), high correlations among different dimensions are visible, which implies 'elliptically' distributed residuals according to the analysis in section II-C.

As expected, whitening effectively reduces these correlations. Whitening of the input data (P_{WN}) drastically reduces correlations between the residuals. Correlations are slightly lower still when whitening is applied directly to the residuals (P_{SW}). Note that the pairwise correlations are not zero due to differences between the validation set (used to determine the post-whitening matrix) and the test set. Finally, applying whitening on the inputs and the outputs (P_{WW}) produces similarly small correlations.

C. Anomaly Detection Performance Evaluation

Both the processing methods (P_{xy}) and the autoencoder configuration (B_x) influence the sensitivity of anomaly detection. This impact will be quantified in this section. For these tests, anomalies were generated by modifying the test data as follows. For each data point, we randomly selected one wind farm out of 99 and reduced its power output by a given amount (the *anomaly magnitude*, abbrev. *am*). This approach yields an anomalous test set consisting of 17,544 non-anomalous data points and an equal number of anomalous data points.

1) Receiver Operating Characteristic Curves: Receiver Operating Characteristics (ROCs) and the corresponding Area Under the Curve (AUC) were used to quantify the sensitivity and specificity of anomaly detection, as a function of the autoencoder structure, data processing method, and anomaly magnitude. ROC curves were constructed by varying the threshold τ_v^α. For all cases, *ZCA* was selected as the whitening matrix, and an anomaly magnitude of 10% was used.

For the first experiment, we used default processing P_{SN}: standardizing the input data and detecting anomalies from unprocessed residuals. Comparing the performance of all layer dimension configurations B_x, we can observe in Fig. 5 (a) that autoencoder networks configured as B_{32}, B_{64} and B_{96} have better detection performance. Specifically, at each false positive rate, the true positive rate (sensitivity) of these detectors are higher than others'. Accordingly, they also have larger AUC. This indicates that optimal detection is achieved with fairly wide autoencoders. The configuration B_{64} was used for all following experiments.

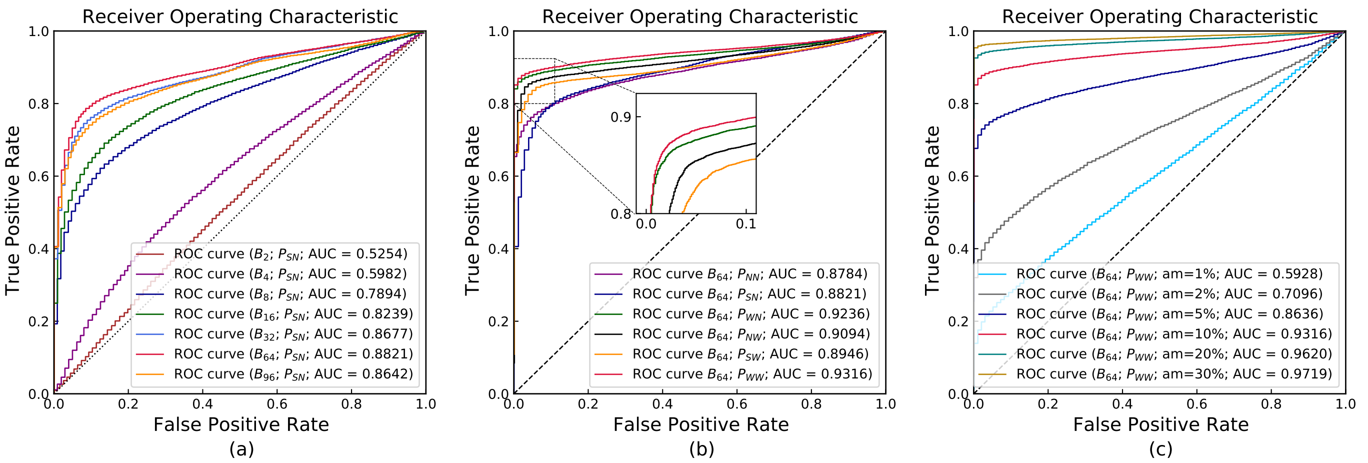
A comparison of the anomaly detection performance of all data processing approaches P_{xy} is shown in Fig. 5(b). We can observe that, i) detectors equipped with whitening transformation (P_{WN}, P_{NW}, P_{SW}, and P_{WW}) have higher anomaly detection sensitivity than the others; ii) performing whitening only on the inputs (P_{WN} and P_{WW}) renders higher detection sensitivity than whitening approaches performed to the residuals (P_{NW} and P_{SW}); iii) the detector using combined whitening (P_{WW}) slightly outperforms the detector just utilizing input whitening (P_{WN}). It can be concluded that the combined whitening approach P_{WW} is the best choice to improve overall anomaly detection sensitivity.

Moreover, we investigated the ability to detect anomalies of various magnitudes, using the selected processing strategy P_{WW}. Fig. 5 (c) shows that, as the anomaly rate increases from 1% to 30%, the ROC curves and AUC improve, reaching very high levels from 10%.

TABLE II
DETECTION PERFORMANCE COMPARISON BY MULTIPLE METRICS ($\boldsymbol{\alpha}$ = 99, ANOMALY MAGNITUDE = 10% AND $\boldsymbol{B_{64}}$ IS UTILIZED).

Processing method	TNR	TPR	PPV	F$_1$-score
P_{NN}	98.89%	56.26%	98.07%	71.50%
P_{SN}	98.77%	40.58%	97.06%	57.23%
P_{WN}	**98.98%**	84.06%	**98.80%**	90.83%
P_{NW}	98.86%	71.81%	98.44%	83.04%
P_{SW}	98.80%	66.52%	98.22%	79.32%
P_{WW}	98.88%	**85.14%**	98.70%	**91.42%**

2) Detection Performance Evaluation by Multiple Metrics : In addition to the detection sensitivity (true positive rate; TPR), we evaluated the performance of the anomaly detector by multiple metrics, namely specificity (TNR), precision (PPV), and F$_1$-score. Interested readers can refer to [31] for a detailed introduction to the metrics. For all cases, we used an anomaly magnitude of 10%, $\alpha = 99$, and layer configuration B_{64}. The experimental results are shown in Table. II. In all cases, the TNR is close to 99%, consistent with the choice of the threshold $\alpha = 99$. Similarly, the PPV scores are high across all processing methods, but a closer look at the TPR and F_1 metrics shows that - perhaps surprisingly - the P_{SN} processing scheme is least dependable, by a large margin. The schemes using whitening of input data outperform all others, with a



(a) (b) (c)

Fig. 5. Test set receiver operating characteristics of spatial data, using different configurations ($\boldsymbol{B_x}$), data processing methods ($\boldsymbol{P_{xy}}$), and anomaly magnitudes.

slight edge for the combined whitening procedure (P_{WW}) over the pre-whitening only procedure (P_{WN}).

To investigate the stability of the stochastic training process, model training was performed 15 times, and the variability of the sensitivity was monitored (results not shown). Whitening at the pre-processing stage resulted in higher sensitivity *and* a narrower range of results.

TABLE III
DETECTION PERFORMANCE COMPARISON BY DIFFERENT ANOMALY TYPES ($\boldsymbol{\alpha}$ = 99, $\boldsymbol{P_{WW}}$ AND $\boldsymbol{B_{64}}$ ARE UTILIZED).

Anomaly Types	TNR	TPR	PPV	F$_1$-score
$1 \times 10\%$	98.88%	85.14%	98.70%	91.42%
$2 \times 5\%$	98.88%	80.20%	98.62%	88.46%
$5 \times 2\%$	98.88%	67.84%	98.38%	80.30%
$10 \times 1\%$	98.88%	53.72%	97.96%	69.39%
$1 \times 30\%$	98.88%	95.35%	98.84%	97.06%
$2 \times 15\%$	98.88%	95.06%	98.84%	96.91%
$5 \times 6\%$	98.88%	91.95%	98.80%	95.25%
$10 \times 3\%$	98.88%	87.92%	97.74%	93.02%

3) Detection Performance of Different Anomaly Types: In the next experiment, the performance of detecting different anomaly types was investigated. The total anomaly magnitude (i.e. power reduction) was fixed at 10% and 30%, respectively, but distributed over 1, 2, 5 or 10 observations. The experimental results are shown in Table. III. Under each total anomaly magnitude, more concentrated anomalies (i.e., $1 \times 10\%$ and $1 \times 30\%$) render higher TPR and F$_1$-scores, but these decrease for more diffuse anomalies, as might be expected. TNR and PPV remain high for all anomaly patterns.

D. Anomaly Localization Performance Evaluation

In addition to quantifying the ability to detect anomalies (a binary classification), we next investigated the ability to localize the source of an anomaly, and how this depends on the configuration of the detector.

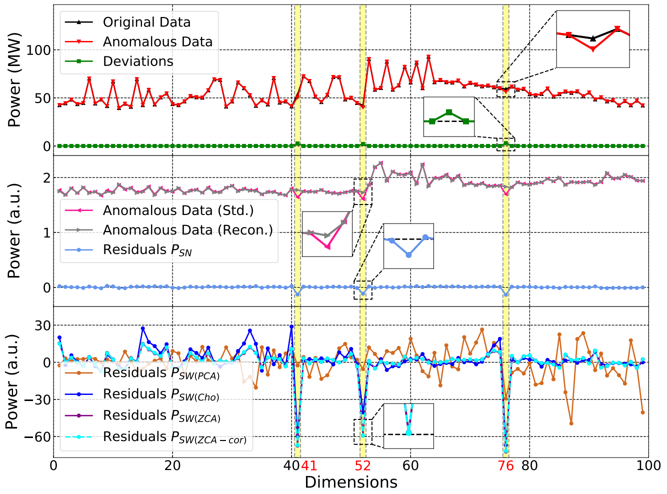


Fig. 6. Illustrative example that depicts the effect of different whitening approaches on anomaly locating capacity (anomaly magnitude = 5%; $\boldsymbol{P_{SN}}$, $\boldsymbol{P_{SW}}$, and $\boldsymbol{B_{64}}$ are utilized).

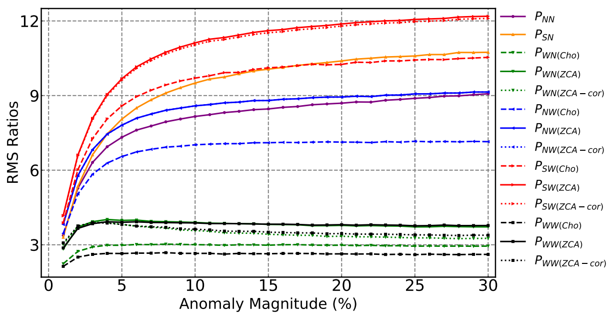


Fig. 7. Root mean square (RMS) ratios of the anomalous to normal dimensions for data processing/whitening combinations ($\boldsymbol{B_{64}}$ is utilized).

1) Visual Comparison of Localization Performance: We first considered an illustrative example of anomaly localization, in which the output of three wind farms (numbers 41, 52, and 76) was reduced by 5%. The input data is shown in Fig. 6 (top panel), where the anomalous locations are indicated in yellow. Looking at the residuals of standardized data (P_{SN}, middle

TABLE IV
LOCALIZATION CAPACITIES OF USING DIFFERENT DATA PROCESSING METHODS (ANOMALY MAGNITUDE = 5% AND $\boldsymbol{B_{64}}$ IS UTILIZED).

Processing method	Whitening matrix	RMS Ratio	Gap Ratio	OCR
$\boldsymbol{P}_{\textbf{NN}}$	/	7.32	1.60	73.18%
$\boldsymbol{P}_{\textbf{SN}}$	/	8.05	1.97	78.48%
$\boldsymbol{P}_{\textbf{WN}}$	*Cholesky*	2.98	0.37	7.17%
$\boldsymbol{P}_{\textbf{WN}}$	*ZCA*	3.98	0.68	23.40%
$\boldsymbol{P}_{\textbf{WN}}$	*ZCA-cor*	3.82	0.61	18.22%
$\boldsymbol{P}_{\textbf{NW}}$	*Cholesky*	6.55	1.39	68.04%
$\boldsymbol{P}_{\textbf{NW}}$	*ZCA*	7.81	1.65	75.98%
$\boldsymbol{P}_{\textbf{NW}}$	*ZCA-cor*	7.81	1.65	75.74%
$\boldsymbol{P}_{\textbf{SW}}$	*Cholesky*	8.58	1.59	71.78%
$\boldsymbol{P}_{\textbf{SW}}$	*ZCA*	**9.68**	**1.98**	82.23%
$\boldsymbol{P}_{\textbf{SW}}$	*ZCA-cor*	9.62	1.97	**82.60%**
$\boldsymbol{P}_{\textbf{WW}}$	*Cholesky*	2.65	0.30	2.44%
$\boldsymbol{P}_{\textbf{WW}}$	*ZCA*	3.90	0.65	20.52%
$\boldsymbol{P}_{\textbf{WW}}$	*ZCA-cor*	3.82	0.61	18.27%
$\boldsymbol{P}_{\textbf{SW}}$	*ZCA* - c	**9.72**	**2.07**	84.65%
$\boldsymbol{P}_{\textbf{SW}}$	*ZCA-cor* - c	9.67	2.06	**85.08%**

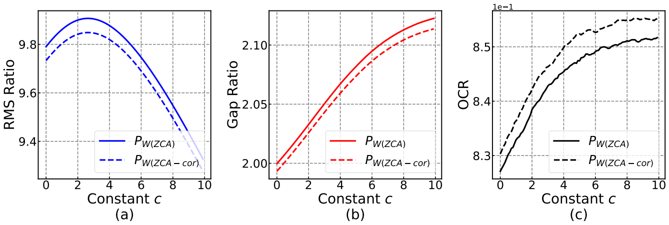


Fig. 8. Localization performance as a function of whitening offset c.

panel), it can be seen that the residuals of the three anomalous dimensions stand out. In the bottom panel, we can observe that the four whitening approaches affect the residual signals in different ways. *PCA* whitening mixes all coordinates and fully obscures the connection between the original perturbations and residuals. As a result, it will not be considered in the follow-up analysis. In contrast, the *Cholesky*, *ZCA*, and *ZCA-cor* whitened residuals all have peaks that are consistent with the actual anomalous dimensions, but the *Cholesky* method also produces residuals in non-anomalous locations (e.g., a peak at dimension number 40). This result is consistent with the fact that ZCA whitening maximizes the average cross-covariance between the original and whitened data and ZCA-cor maximizes their cross-correlation [13].

2) Statistical Localization Performance Comparison: For each point in the test set, we randomly selected 3 out of 99 wind farms and applied power reductions to generate anomalous test vectors, resulting in test sets of 17544 data points for each anomaly rate. Fig. 7 depicts the *RMS Ratio* for various data processing schemes, as a function of anomaly magnitude, and Table. IV shows numerical results for all three anomaly metrics for a fixed anomaly magnitude of 5%.

The most striking observation is that methods that perform whitening at the pre-processing stage (P_{Wx}) scored significantly worse on all localization metrics. Apparently, mixing of features before encoding helps to improve detection sensitivity (previous section), but is detrimental to localization performance. Moreover, for any data processing combination, both *ZCA* processing schemes scored higher than the *Cholesky* whitening scheme, and the best scores were obtained when the *ZCA* and *ZCA-cor* schemes are used for post-processing, in combination with standardization for pre-processing (P_{SW}). Here, *ZCA* scored very slightly higher on the *RMS ratio* and *gap ratio* metrics (typical separation), and *ZCA-cor* attained the highest scores on the *OCR* metric (ordering).

Finally, an additional enhancement of the method was introduced. The residual whitening transformation (12) mixes signals between dimensions. This may cause a peak (positive or negative) in one or more dimensions to affect the average value of other dimensions. In order to better separate this signal from the background, we applied a constant offset to the whitening matrix (12) as follows:

$$r_{s(c)} = (W_r - c\mathbb{1})r. \tag{17}$$

Here, $\mathbb{1}$ is a matrix of ones and c is a constant to be defined, so that a multiple of the sum-of-residual values ($\sum_i r_i$) is subtracted from the whitened feature vector. The change of localization performance as a function of c is shown in Fig. 8. An overall improvement is obtained for values larger than zero, although *RMS ratio* decreases after an initial increase. Results for the value $c = 5$ are included in Table. IV.

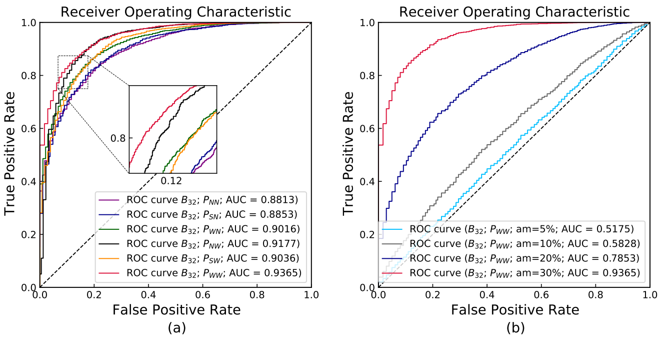
V. CASE STUDY: TIME SERIES DATA

A. Experiment Scenario Formulation

Apart from detecting anomalies in data 'snapshots' that correspond to spatially distributed locations, we further validate the anomaly detection performance of the whitening-enhanced autoencoder using a time-series data set. This data set contains half-hourly load profiles of 4,173 customers in London collected between 2011 and 2014 [32]. 4,000 customers were randomly selected, and data from 2013 (the most complete year) were used to create daily load profiles for groups of 100 customers. The resulting data set $X \in \mathbb{R}^{14600\times 48}$ was randomly divided into training, validation, and testing sets in blocks of one week with the proportion of 2:1:1. The dimensions of hidden and bottleneck layers were set as 96 and 32, respectively. The batch size was 16, the learning rate was 1×10^{-5}, and 350 training epochs were used. Other settings were kept the same as those in Section IV-A. Anomalous load profiles were created by modifying the test set. Specifically, 3 randomly selected data points out of each 48-dimensional load profile were reduced by a certain amount.

B. Anomaly Detection Performance

Fig. 9(a) compares the results obtained with different whitening methods, for a given anomaly magnitude of 30% (comparable with Fig. 5(b)). It is evident that also in this case, whitening-enhanced autoencoders (P_{NW}, P_{WN}, P_{SW}, and P_{WW}) outperform non-whitening-enhanced ones (P_{NN} and P_{SN}). A detector equipped with combined whitening (P_{WW}) has higher anomaly detection sensitivity than detectors



(a) (b)

Fig. 9. Test set receiver operating characteristics of time series data using different data processing methods ($\boldsymbol{P_{xy}}$) and anomaly magnitudes.

equipped with signal whitening transformation (P_{NW}, P_{SW}, P_{WN}).

In addition, the detection performance of the P_{WW} method was investigated for different anomaly magnitudes. As expected, and shown in Fig. 9(b), the ROC curves and AUC improve as the anomaly magnitudes increase. When comparing the results with Fig 5(c) for similar signal reductions, it is notable that detection performance is reduced, despite reductions occurring at three observed data points simultaneously instead of one. This is due to the larger stochasticity of electricity demand profiles compared to the highly correlated wind power snapshots, making it harder to distinguish normal and anomalous data, particularly at low anomaly magnitudes.

VI. Conclusion

Autoencoder neural networks are a powerful tool for the detection of unknown anomalies. A threshold for the (Euclidean) length of the residuals is used to identify anomalous states of a system. In this paper, we investigated how whitening-based decorrelation of the input features and residuals can improve the performance of the anomaly detector, for use cases of wind power generation at 99 different locations and daily electricity load profiles. Whitening of the *input* data was found to be most beneficial for detection performance, across multiple metrics, and a small further enhancement was obtained when both input data and the residuals were whitened (combined whitening). However, input whitening was found to reduce the ability to locate the source of anomalies. Three metrics were formulated to quantify this ability, and the best performance was obtained using standardization of the input data and whitening of the *residuals* using the *ZCA* or *ZCA-cor* whitening matrix with a small offset. In future work, we aim to extend the method to analyze spatial-temporal signatures and include contextual information. Further refinement of the neural network itself is also a promising research direction: applying batch normalization or modifying the objective function used during training may indirectly improve detection performance by smoothing the distribution of values of the non-anomalous data within the latent space or hidden network layers.

References

[1] Y. Wang, N. Zhang, C. Kang, M. Miao, R. Shi, and Q. Xia, "An efficient approach to power system uncertainty analysis with high-dimensional dependencies," *IEEE Transactions on Power Systems*, pp. 2984–2994, 2017.

[2] R. Liu, C. Vellaithurai, S. S. Biswas, T. T. Gamage, and A. K. Srivastava, "Analyzing the cyber-physical impact of cyber events on the power grid," *IEEE Transactions on Smart Grid*, vol. 6, no. 5, pp. 2444–2453, 2015.

[3] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*. ACM, 2014, p. 4.

[4] C. Wang, S. Tindemans, K. Pan, and P. Palensky, "Detection of false data injection attacks using the autoencoder approach," in *2020 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*. IEEE, 2020, pp. 1–6.

[5] H. Zhao, H. Liu, W. Hu, and X. Yan, "Anomaly detection and fault analysis of wind turbine components based on deep learning network," *Renewable energy*, vol. 127, pp. 825–834, 2018.

[6] L. Wang, Z. Zhang, J. Xu, and R. Liu, "Wind turbine blade breakage monitoring with deep autoencoders," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 2824–2833, 2016.

[7] Y. Li, W. Jiang, G. Zhang, and L. Shu, "Wind turbine fault diagnosis based on transfer learning and convolutional autoencoder with small-scale data," *Renewable Energy*, vol. 171, pp. 103–115, 2021.

[8] E. Vladislavleva, T. Friedrich, F. Neumann, and M. Wagner, "Predicting the energy output of wind farms based on weather data: Important variables and their correlation," *Renewable energy*, vol. 50, pp. 236–243, 2013.

[9] T. Denouden, R. Salay, K. Czarnecki, V. Abdelzad, B. Phan, and S. Vernekar, "Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance," *arXiv preprint arXiv:1812.02765*, 2018.

[10] G. Jiang, P. Xie, H. He, and J. Yan, "Wind turbine fault detection using a denoising autoencoder with temporal information," *IEEE/Asme transactions on mechatronics*, vol. 23, no. 1, pp. 89–100, 2017.

[11] J. Renman, "Deep autoencoder for condition monitoring of wind turbines-detecting and diagnosing anomalies," Master's thesis, Chalmers Tekniska Högskola, Gothenburg, Sweden, 2019.

[12] Y. Cui, P. Bangalore, and L. B. Tjernberg, "An anomaly detection approach based on machine learning and SCADA data for condition monitoring of wind turbines," in *2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*. IEEE, 2018, pp. 1–6.

[13] A. Kessy, A. Lewin, and K. Strimmer, "Optimal whitening and decorrelation," *The American Statistician*, vol. 72, no. 4, pp. 309–314, 2018.

[14] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[15] B. R. Kiran, D. M. Thomas, and R. Parakkal, "An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos," *Journal of Imaging*, vol. 4, no. 2, p. 36, 2018.

[16] N. Renström, P. Bangalore, and E. Highcock, "System-wide anomaly detection in wind turbines using deep autoencoders," *Renewable Energy*, vol. 157, pp. 647–659, 2020.

[17] C. Wang, K. Pan, S. Tindemans, and P. Palensky, "Training strategies for autoencoder-based detection of false data injection attacks," in *The 2020 IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe 2020)*, 2020, arXiv:2005.07158.

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

[19] J. James, Y. Hou, and V. O. Li, "Online false data injection attack detection with wavelet transform and deep neural networks," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3271–3280, 2018.

[20] L. Duan, M. Xie, T. Bai, and J. Wang, "A new support vector data description method for machinery fault diagnosis with unbalanced datasets," *Expert Systems with Applications*, vol. 64, pp. 239–246, 2016.

[21] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," *Advances in neural information processing systems*, vol. 12, 1999.

[22] G. Li and J. Zhang, "Sphering and its properties," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 119–133, 1998.

[23] J. H. Friedman, "Exploratory projection pursuit," *Journal of the American statistical association*, vol. 82, no. 397, pp. 249–266, 1987.

[24] S. Zhang, E. Nezhadarya, H. Fashandi, J. Liu, D. Graham, and M. Shah, "Stochastic whitening batch normalization," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10 978–10 987.

[25] K. K. Pal and K. Sudeep, "Preprocessing for image classification by convolutional neural networks," in *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. IEEE, 2016, pp. 1778–1781.

[26] P. C. Mahalanobis, "On the generalised distance in statistics," *Proceedings of the National Institute of Sciences of India*, vol. 2, pp. 49–55, 1936.

[27] S. Pfenninger and I. Staffell, "Renewables.ninja," *Renewables.ninja*, 2017. [Online]. Available: https://www.renewables.ninja/

[28] M. M. Rienecker, M. J. Suarez, R. Gelaro, R. Todling *et al.*, "MERRA: NASA's modern-era retrospective analysis for research and applications," *Journal of climate*, vol. 24, no. 14, pp. 3624–3648, 2011.

[29] I. Staffell and S. Pfenninger, "Using bias-corrected reanalysis to simulate current and future wind power output," *Energy*, vol. 114, pp. 1224–1239, 2016.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[32] S. Tindemans, G. Strbac, J. R. Schofield, M. Woolf, R. Carmichael, and M. Bilton, "Low Carbon London project: Data from the dynamic time-of-use electricity pricing trial, 2013. [data collection]," *UK Data Service, SN:7857*, 2016.