CHALMERS
UNIVERSITY OF TECHNOLOGY

Silent Signals: Applying AI to Reveal Hidden Power Quality Trends

Long-Term Power Quality Pattern Discovery and Anomaly Detection Using Feature-Based Unsupervised Learning

Master's thesis in Complex Adaptive Systems

Lukas Wallén, Jacob Walter

DEPARTMENT OF ELECTRICAL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's thesis 2026

Silent Signals: Applying AI to Reveal Hidden Power Quality Trends

Long-Term Power Quality Pattern Discovery and Anomaly Detection Using Feature-Based Unsupervised Learning

Lukas Wallén, Jacob Walter

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Division of Electric Power Engineering
Chalmers University of Technology
Gothenburg, Sweden 2026

Silent Signals: Applying AI to Reveal Hidden Power Quality Trends
Long-Term Power Quality Pattern Discovery and Anomaly Detection Using Feature-Based Unsupervised Learning
Lukas Wallén & Jacob Walter

Supervisor: Fredrik Stigmarker, Unipower AB
Examiner: Jimmy Ehnberg, Department of Electrical Engineering, Chalmers University of Technology

Master's Thesis 2026
Department of Electrical Engineering
Division of Electric Power Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Generated image using M365 Copilot from the prompt "I'm doing a master thesis project that is called: Silent Signals: Applying AI to Reveal Hidden Power Quality Trends. I want to generate an image for this project. Emphasis on AI and power quality in the image. Without a title or subtitle.".

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

Silent Signals: Applying AI to Reveal Hidden Power Quality Trends
Long-Term Power Quality Pattern Discovery and Anomaly Detection Using Feature Based Unsupervised Learning
Lukas Wallén, Jacob Walter
Department of Electrical Engineering
Chalmers University of Technology

Abstract

Power quality (PQ) monitoring is essential for ensuring the efficient operation of modern electrical grids, where increasing electrification and the integration of non-linear loads introduce complex disturbances. While traditional methods focus on short-term events, long-term PQ trends remain underexplored due to the high dimensionality and volume of the data. This thesis investigates the application of unsupervised machine learning techniques to discover long-term anomalous patterns in PQ data without relying on labelled datasets.

Daily statistical features were extracted from root mean square (RMS) voltage, total harmonic distortion (THD) voltage, and voltage unbalance factor (VUF), collected from two locations in Sweden. Three models were implemented and evaluated: A Self-Organising Map (SOM), Autoencoder (AE), and Variational Autoencoder (VAE). Anomaly detection was carried out using model-specific error metrics: quantisation error (QE) for SOM and reconstruction error (RE) for AE-based models alongside a density-based outlier score (OS).

The results demonstrate that all models are capable of identifying anomalous days in long-term PQ behaviour. Distance-based methods (QE and RE) showed higher consistency and better alignment with meaningful deviations compared to density-based OS. The VAE approach exhibited the strongest performance in distinguishing anomalous patterns, likely due to its regularised latent space, which enhances sensitivity to deviations from normal behaviour. The models showed limited agreement on specific anomalies, this highlights that anomaly detection is strongly model-dependent, as each model captures different aspects of the data structure.

Overall, the study confirms that unsupervised machine learning is a promising tool for long-term PQ analysis, enabling automated detection of subtle anomaly disturbances. The findings support the integration of such methods into PQ monitoring systems to improve data-driven decision-making and facilitate a transition toward more proactive grid management.

Keywords: Power quality, AI, machine learning, anomaly detection, unsupervised learning, autoencoder, self-organising map.

Acknowledgements

We would like to express our deepest gratitude to Unipower for giving us the opportunity to carry out this thesis project within the company. Their support, expertise, and encouragement have been invaluable throughout the entire process. We would also like to express an extra thanks to our supervisor Fredrik Stigmarker. We especially appreciate the time, insights, and guidance provided by him and the engineers and staff who helped shape our understanding and contributed to the quality of this work. Collaborating with Unipower has been a rewarding experience that has strengthened both our professional skills and our enthusiasm for the field.

We would also like to extend our heartfelt thanks to our families. Their patience, understanding, and constant encouragement have supported us through long workdays, late nights, and the inevitable challenges of academic research. Without their unwavering belief in us, this thesis would not have been possible.

Finally, we are grateful to friends, colleagues, and teachers, who has contributed directly or indirectly to this journey. Your support has meant more than words can express.

Lukas Wallén, Jacob Walter, Gothenburg, June, 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

AC	Alternating Current
AI	Artificial Intelligence
AE	Autoencoder
BMU	Best Matching Unit
DC	Direct Current
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
KDE	Kerndel Density Estimate
KLD	Kullback-Leibler Divergence
ML	Machine Learning
MSE	Mean Squared Error
OS	Outlier Score
PQ	Power Quality
PQDs	Power Quality Disturbances
QE	Quantisation Error
RE	Reconstruction Error
RMS	Root Mean Square
SOM	Self-Organising Map
THD	Total Harmonic Distortion
VAE	Variational Autoencoder
VUF	Voltage Unbalance Factor

Contents

1
Introduction

In this chapter, an overview of the thesis will be given in addition to the purpose, the research questions and limitations of the thesis.

1.1 Background

An ideal AC electricity signal consists of a pure sinusoidal waveform with constant amplitude and frequency. In practice, this ideal condition is rarely achieved. Loads and other system components cause deviations in amplitude, phase, waveform shape, and may introduce additional frequency components. PQ describes the impact that these deviations have on the signal [1], and such deviations from ideal conditions are referred to as PQDs. Some examples of causes of PQDs are an industry plant starting after downtime or heavy machinery being switched on [2].

As the world undergoes increasing electrification, maintaining sufficient PQ and reducing PQDs has become increasingly important. PQ monitoring is necessary to ensure reliable operation and compliance with international, national, and application specific standards across all levels of the electrical grid [3, 4, 5]. Insufficient PQ handling can lead to several problems, such as equipment malfunction (e.g. computers restarting, controllers losing synchronisation, or sensors providing incorrect data) and reduced equipment lifetime [6]. Electrical insulation, capacitors, and semiconductor devices may all degrade faster due to repeated exposure to PQDs. In rare but severe cases, PQDs can also contribute to power outages and consequently financial losses.

There are several methods to reduce the impact of PQDs, such as surge protection divides, uninterruptible power supplies, and filters [7]. Although these protective measures exist, PQ remains relevant because PQDs cannot be completely eliminated. Electrical networks are continuously exposed to changing load conditions, switching events, faults, nonlinear devices and inverter-based resources. In addition, no protective measure is fail-safe. Devices can break and degrade. As a result, PQ must be continuously monitored and mitigated rather than assumed to be permanently solved.

Some examples of PQDs are voltage magnitude variations, voltage unbalance, and harmonic distortion [1]. Voltage magnitude variations are primarily caused by changes in load demand and power flow [8] , where increased current through network

impedances leads to voltage drops, while distributed generation can cause voltage rise. Voltage unbalance occurs due to asymmetry between phases [9], typically originating from uneven distribution of single-phase loads or unequal line impedances. Harmonic distortion is introduced by non-linear devices such as inverters, rectifiers, and other power electronic equipment [10], which draw non-sinusoidal currents and consequently distort the voltage waveform. These disturbances can have significant consequences for electrical systems and connected equipment [7]. Voltage variations may lead to equipment shutdown or malfunction, while harmonic distortion increases thermal stress in transformers and cables due to additional losses. Voltage unbalance is particularly critical for rotating machines, where it causes uneven currents that result in excessive heating, torque oscillations, and reduced lifespan.

While extensive research exists on short-term voltage magnitude, unbalance and harmonic events such as voltage sags, swells, and transients [11, 12], long-term patterns in the mentioned PQDs remain comparatively underexplored [13]. Studying these long-duration variations offers substantial value. It enables a shift from reactive maintenance to preventative strategies, improves energy efficiency, and helps organisations reduce operational costs. Long-term trends can uncover slow-developing forms of system degradation, reveal the impact of new loads, and expose structural weaknesses in the electrical network [14]. However, PQ data collected over days, weeks, or months is vast, complex, and highly multidimensional. Detecting subtle anomalies or emerging trends in such datasets is challenging for humans and deterministic algorithms, which are typically optimised for clearly defined events.

In this project the selected PQ parameters chosen for analysis are RMS voltage, VUF and voltage THD. These parameters quantify voltage magnitude, voltage unbalance, and harmonic distortion, and were chosen because they represent part of key PQ voltage indicators defined in international, European, and Swedish regulations [3, 4, 5]. They also reflect relevant emerging challenges in modern electrical networks, including rising harmonic distortion from non-linear loads and increasing voltage unbalance due to distributed generation [14, 15]. For these reasons, RMS, THD, and VUF provide a foundation for monitoring long-term PQ behaviour that is aligned with standards and relevant to the industry.

ML offers a promising approach for analysing the selected PQ parameters from a long-term perspective [16, 17]. Unlike rule-based methods, ML models excel at learning underlying patterns directly from data, discovering non-linear relationships, and identifying anomalies that do not fit predefined categories. This makes ML particularly suited for uncovering slow-evolving or previously unknown patterns in large-scale RMS, VUF and THD measurements [18, 19], using unsupervised learning. Beyond the research interest, the use of ML driven analysis presents a practical opportunity for companies collecting PQ data. Integrating ML analysis tools into the product portfolio would enable users to process complex PQ data more efficiently, automate analysis tasks, and reduce manual work, while gaining insights into the health and stability of their electrical systems.

This master's thesis project was conducted at Unipower AB. Unipower is a company that specialises in PQ measurement and analysis [20]. Unipower manufactures PQ meters that collect data from electrical networks. These meters are used by companies in various sectors, including industrial, transmission and energy production. Therefore, the interest in investigating whether ML models can distinguish anomalies from the selected PQ parameters is of interest both for industrial parties, i.e Unipower, and societal parties.

1.2 Purpose

The purpose of this thesis is to investigate whether ML models are capable of distinguishing long-term anomalous behaviour in the electric grid using unsupervised learning, and evaluate their performance based on the following PQ quantities:

- RMS for all three voltage phases U_1, U_2, U_3
- Voltage unbalance factor VUF
- Total harmonic distortion THD for all phase voltages.

In addition, the thesis investigates whether such models can identify outliers or anomalous behaviour that deviates from typical operational patterns, potentially indicating faults or disturbances.

1.3 Related work

As mentioned previously, PQ assessment requires monitoring quantities that reflect changes in voltage magnitude, waveform distortion, and phase-related imbalance. In addition to the inclusion of these phenomena in the norms, they are also included in the IEEE 1159-2019 standard for monitoring PQ [21].

A similar set of PQ parameters used in this thesis has been investigated in other research papers, both with and without ML. For example, in [22], Liu et al. analyse both harmonic distortion and voltage unbalance when assessing PQ performance in shipboard microgrids, the result being that the voltage unbalance and THD can degrade system performance and reliability, emphasising the need for improved monitoring and control in systems. In [17], D.K. Nishad et al. proposed using ML optimisation to reduce voltage THD, voltage unbalance and voltage variations, the result being that ML was able to increase stability of the metrics used due to fast response from the ML algorithms. The article suggests that further research could be necessary for other complex configurations.

Even though the authors could not find as much research focusing on long term PQ measurements, there exists some research with this perspective. In [23] A. Ostrowska et al. studied the impact of a microgrid in a distribution network over a five week period. The study used statistical evaluation of voltage RMS, voltage THD and VUF, among other parameters, with the result suggesting that a statistical rep-

resentation of RMS, THD, and VUF works for representational purposes. In [16], Furlano Bastos and Santoso proposed ML models designed to analyse power systems using continuous (long-term) voltage RMS data, and extract rare events, with the result being ML was able to identify rare events and aiding with rebalancing PQ monitor datasets.

This related research shows that using the suggested PQ parameters to analyse long-term trends using ML models is relevant.

1.4 Research questions

- Can unsupervised machine learning models such as SOM and AEs be used to consistently discover long term anomalies given the defined set of PQ parameters mentioned in section 1.2?
- Which of these models is most suitable for discovering anomalies in PQ data?

1.5 Limitations

This project is a master thesis, it is therefore limited in terms of time and resources. The following limitations have been identified:

- The data is collected from several PQ meters from two geographical locations (Landskrona, Sweden and Ulricehamn, Sweden).
- The time window that will be used during the project will be limited to 1 day. A time window of a day is considered detailed enough to give a proper approximation of the PQ parameters considered.
- The lack of ground truth labels makes it difficult to validate whether anomalies correspond to real PQ issues.
- ML anomaly detection identifies *that* something deviates, not *why*. The root cause and/or explanation of found events/types of days requires further analysis.

2
Theory

The theory and explanations regarding all relevant models, algorithms and parameters will be explained in this chapter.

2.1 Selected power quality parameters

This section provides a detailed explanation about the selected PQ parameters.

2.1.1 Voltage unbalance factor

In an ideal three-phase system, all three phase voltages (U_1, U_2, U_3) have equal magnitudes and are separated by 120°. Unbalance can occur due to uneven loads or other disturbances, causing the phase voltages to differ in magnitude and/or phase, breaking the ideal 120° symmetry [24, 25].

One way to quantify unbalance is by defining positive sequence voltage U_{pos}, negative sequence voltage U_{neg}, and zero sequence voltage U_{zero}. U_{pos} represents the balanced part of the system, rotating in the normal direction. U_{neg} represents the unbalanced part, rotating in the opposite direction. U_{zero} represents the part of the system where all three phase voltages are equal, stationary and in phase. One can think of these definitions as a transformation from the original phase voltages. Mathematically, these are defined accoring to Equation 2.1:

$$\begin{aligned} U_{zero} &= \frac{1}{3}(U_1 + U_2 + U_3) \\ U_{pos} &= \frac{1}{3}(U_1 + \alpha U_2 + \alpha^2 U_3) \\ U_{neg} &= \frac{1}{3}(U_1 + \alpha^2 U_2 + \alpha U_3). \end{aligned} \tag{2.1}$$

Where U_1, U_2, U_3 are phasors (complex numbers with magnitude and phase), $\alpha = e^{j120^\circ}$, and j is the imaginary unit. In a perfectly balanced system, $U_{zero} = U_{neg} = 0$ and $|U_{pos}| = U_{phase}$, where U_{phase} is the RMS voltage of each phase.

In summary, the negative sequence voltage indicates the part of the system rotating opposite to the normal power flow, and is a direct measure of unbalance. However, unbalance can also be expressed as VUF according to Equation 2.2:

$$\text{VUF} = 100 \cdot \frac{|U_{neg}|}{|U_{pos}|} \tag{2.2}$$

as shown in [25]. By definition, VUF is expressed as a percentage and is bounded between 0% (perfectly balanced) and 100% (extremely unbalanced).

2.1.2 Root mean square voltage

The voltage U of an AC sinusoidal source at time t and angular frequency ω, with values ranging between U_0 and $-U_0$ can be defined according to equation 2.3:

$$U = U_0 \sin(\omega t). \tag{2.3}$$

RMS voltage U_{RMS} is the DC equivalent of an AC source which can be described by equation 2.4:

$$U_{RMS} = \frac{U_0}{\sqrt{2}}. \tag{2.4}$$

as shown in [26]. In Europe the standard RMS value is $U_{RMS} = 230\,\text{V}$ but the AC voltage value varies between $U_0 = \pm 325\,\text{V}$ with a certain frequency.

2.1.3 Total harmonic distortion

In an AC voltage signal, there is one main frequency component f_1, referred to as the fundamental frequency [8]. In the ideal case, the signal consists only of this component. In practice, however, non-linear loads, disturbances, and other system effects cause the signal to become distorted. As a result, the voltage waveform contains additional frequency components besides the fundamental.

One class of these additional frequency components is called harmonics. Harmonics are defined as integer multiples of the fundamental frequency, as described by Equation 2.5:

$$f_n = n f_1. \tag{2.5}$$

Where f_n is the frequency of the nth harmonic component and n is a positive integer. Consequently, the voltage waveform may be represented as a sum of sinusoidal components at integer multiples of the fundamental frequency. Each harmonic component f_n has a corresponding RMS voltage value V_n.

THD is a measure of the amount of harmonic distortion present in a signal. For a voltage waveform, THD is defined as the ratio between the RMS value of all harmonic components and the RMS value of the fundamental component [8]. This is expressed in Equation 2.6.

$$\text{THD} = \frac{\sqrt{V_2^2 + V_3^2 + \cdots + V_N^2}}{V_1}. \tag{2.6}$$

Where V_1 is the RMS value of the fundamental component, and $V_2, V_3, \ldots, V_N$ are the RMS values of the harmonic components. In practice, THD is calculated up to a specified maximum harmonic order N. In many cases, $N = 50$ [27]. THD only accounts for harmonic components, i.e. frequency components at integer multiples of the fundamental frequency, and does not include other types of frequency components.

2.2 Clustering and outlier score

HDBSCAN is a density-based clustering algorithm that produces a representation of the data by organising clusters according to their persistence across varying densities [28]. Density in this case is referred to how packed the data points are. In other words, HDBSCAN observes which clusters remain stable across density levels. Data points that fail to consistently belong to clusters are naturally identified as noise. This hierarchy enables identification of clusters with arbitrary shape and varying density, while being less sensitive to cluster size and shape than other clustering methods.

In addition to producing a clustering structure, HDBSCAN provides an associated anomaly score, referred to as OS [29]. The OS shows how anomalous a data point is by checking whether it lies in a dense cluster or in sparser, less common areas. Data points located near the core of dense clusters are assigned low outlier scores, whereas data points in sparse or peripheral regions of the clusters receive higher scores. The outlier score is bounded between zero and one, where values close to zero indicate typical observations and values close to one indicate strong outliers .

The OS can be used as an unsupervised anomaly detection measure to identify anomalous days and to partition days into anomalous and normal categories based on their degree of deviation from typical patterns.

2.3 Machine learning models

In this section, the ML models used in this thesis will be explained.

2.3.1 Self-organising map

A SOM is an unsupervised artificial neural network used for dimensionality reduction and clustering [30]. It produces a low-dimensional representation of high-dimensional data while preserving the structure of the original dataset. During training, the neurons in the SOM are updated together with their respective neighbours. After training, the SOM provides a grid of neurons representing the distribution of the data, and new observations can be mapped to the grid by finding the neuron whose weight vector best matches the input, called the BMU.

As stated above, two fundamental calculations are performed during SOM train-

ing. Firstly, the similarity between a data point and neurons are calculated, usually by Euclidean distance [30]. Secondly, the neighbourhood of the BMU is determined and the neurons within the neighbourhood is updated. This neighbourhood function is often a function dependent on a predefined hyperparameters, such as a radius from the BMU, a learning rate, and the specific position of neurons within said radius. A standard function that is often used is the gaussian neighbourhood function [31]. An illustrative figure of a SOM is shown in Figure 2.1.

The original SOM paper introduces the learning and neighbourhood mechanisms [30], but does not emphasise QE as a performance metric. In later literature, QE is often used to assess map quality, similarities between data points, and anomalous data points [32, 33]. QE is defined according to Equation 2.7 below:

$$QE(x) = ||x - w_{BMU}||. \tag{2.7}$$

Where $QE(x)$ is the quantisation error for an arbitrary data point x, w_{BMU} is the weight vector of the BMU corresponding to x. High quantisation error means that the data point is not represented by its BMU well, which can indicate outlier days or rare operating regimes [34].

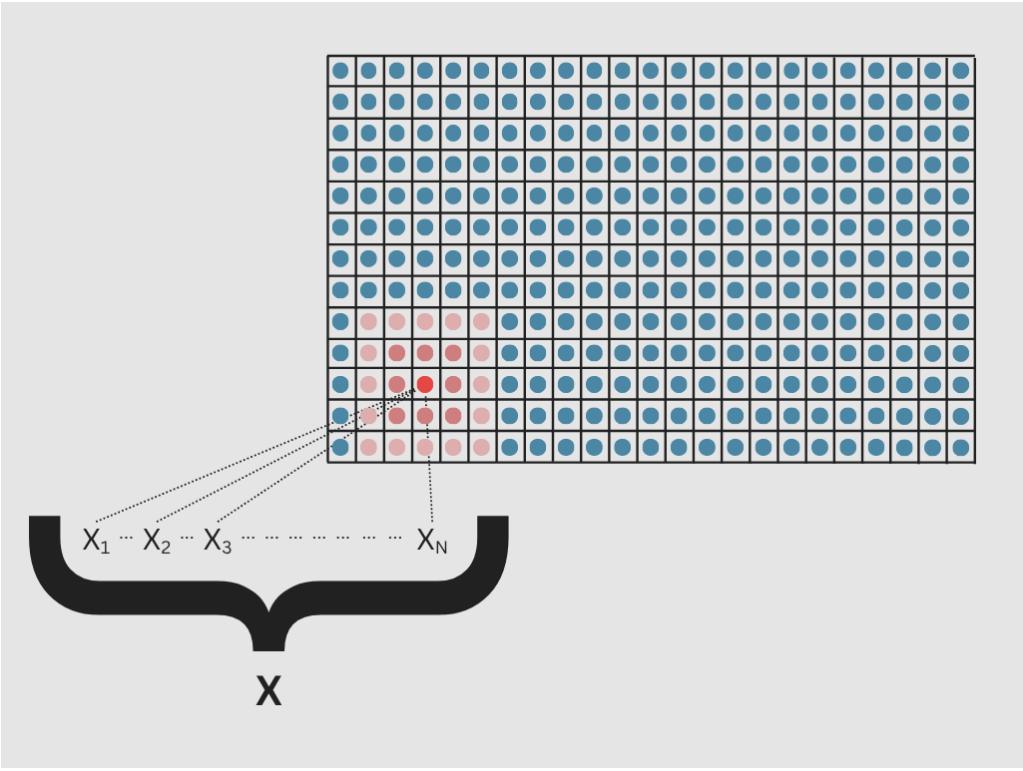


Figure 2.1: Illustrative figure of the self-organising map workflow. The rectangle represents the grid of the SOM. A data point x with dimensionality N is mapped to its BMU in the SOM, shown in red. The weights of the BMU will be updated, along with the weights of the neighbours around it, shown in light red/pink. Neurons will be less and less influenced by the weight update caused by x as the distance to the BMU grows. Neurons far away from the BMU, shown in blue, are not affected.

The ability of a SOM to reduce dimensionality while preserving the structure of data makes it well suited for clustering, as it groups similar inputs together and separates distinct operational regimes. Each neuron in the SOM represents a prototype input, allowing for interpretable visualisation of typical behaviour, and identification of atypical data points through large quantisation errors [32, 33]. Additionally, the SOM provides a geometrical representation of the data through its grid structure, enabling direct visualisation of how anomalies are distributed within the learned topology.

2.3.2 Autoencoder

An AE is an unsupervised artificial neural network that learns latent representations of input data [35]. An AE consists of an encoder, that reduces the dimensionality of the input to a latent space representation, and a decoder, that transforms the latent representation back to its original size. These two parts of the AE cooperate during training and learns how to efficiently transform data to a latent dimension, and reproduce it again.

An AE is able to effectively learn and represent normal frequent inputs, while struggling to reconstruct inputs that are rare or non-existent in the data set, such as outliers and anomalies. This ability can be quantified by using the reconstruction error, shown in Equation 2.8 below:

$$RE(x) = ||x - \hat{x}||. \tag{2.8}$$

Where $RE(x)$ is the reconstruction error for input x, $\hat{x}$ is the reconstructed input. Inputs with low RE are well represented by the model. Inputs with high RE indicate that the AE cannot reconstruct it well, potentially being an anomaly or unusual [35, 36]. Moreover, RE is often used as the loss function during training, in that case conventionally referred to as MSE. An illustrative figure of the structure of an autoencoder is visualised in Figure 2.2.

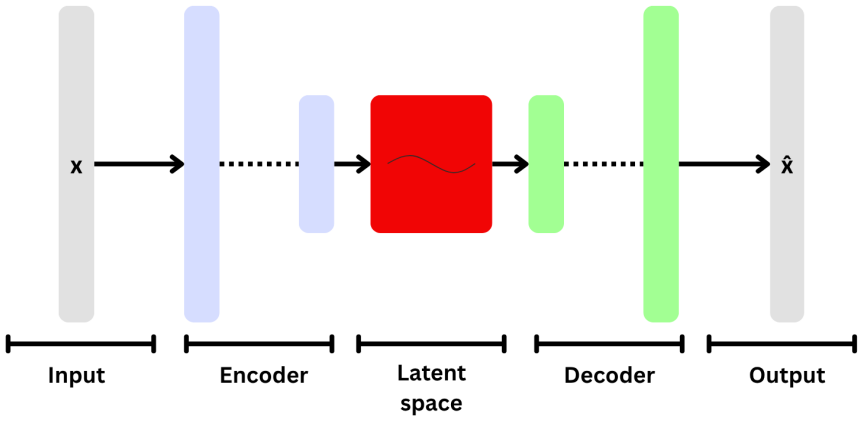


Figure 2.2: Basic structural layout of autoencoders. The difference lies in the latent space where the AE and VAE operates a bit differently.

An autoencoder learns a compact, non-linear representation of the data in an unsupervised manner [36]. This makes it suitable for anomaly detection in complex datasets. By reconstructing inputs from a latent space, it captures correlations between features and provides a reconstruction error that highlights anomalies.

2.3.3 Variational autoencoder

The VAE is a probabilistic extension of the AE. While a conventional AE learns a deterministic mapping from the input space to a fixed latent representation, a VAE models the latent space as a continuous probability distribution [37]. This probabilistic nature encourages the latent representations of normal data to form a smooth distribution, rather than arbitrary point embeddings. As a result, nearby points in the latent space correspond to similar reconstructions, and the model learns a compact representation of the underlying data distribution.

In a VAE, the encoder does not output a single latent vector for each input. Instead, it predicts the parameters of a distribution, typically a mean μ and standard deviation σ. This is called the posterior distribution over the latent variables for each input [37]. During training, this distribution is constrained to remain close to a chosen prior distribution, commonly a standard Gaussian with mean 0 and standard deviation 1. This is done by minimising the KLD, which is a metric that measures the similarity between two distributions [38]. This term penalises unlikely latent representations and limits the model's capacity to arbitrarily encode inputs. In practice, the encoder predicts the logarithm of the variance, $\log(\sigma^2)$, rather than the standard deviation σ, in order to improve numerical stability [37].

The overall training objective therefore consists of two terms: a reconstruction error, which measures how accurately the model reconstructs the input data (see Equation 2.8), and a KLD term, which restricts the latent space. This balance encourages the model to represent normal data efficiently while avoiding overfitting.

For anomaly detection, this formulation is appealing because the VAE is trained to model the distribution of normal data explicitly. Inputs that deviate from this learned distribution tend to require unlikely latent representations, which results in increased reconstruction error [39].

2.4 Jaccard index

The Jaccard index is a measurement for determining the similarity between two sets of arbitrary size [40, 41]. The Jaccard index $J(A, B)$ of the two sets A and B is defined as the size of the intersection divided by the size of the union of the two sets, see Equation 2.9:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{2.9}$$

By definition, $0 \leq J(A, B) \leq 1$ for all sets A and B. It can therefore be interpreted as a percentage of how similar two sets are, and is often used in research as a comparison measurement.

2.5 Kernel density estimate

KDE provides a complementary view of the data distribution alongside a histogram representation [42]. Unlike histograms, the KDE represents the data as a continuous probability density function, offering a smoother and more detailed depiction of the underlying distribution. This allows for a more nuanced comparison between different datasets when assessing model generalisation.

3
Method

The method follows a workflow consisting of three main stages. First, raw data is processed by extracting daily time windows of the selected PQ parameters. Statistical measures are computed over these time windows and then normalised. Second, ML models are trained to learn a representation of the dataset. Finally, anomaly detection is performed by extracting RE/QE from the trained model and comparing them with OS obtained from the HDBSCAN algorithm, enabling the identification of anomalies from two perspectives.

3.1 Data processing

The data was extracted using Unipower's system, PQSecure. The data consists of time series of the selected PQ parameters, where each value is the 10-minute aggregate value. The raw data was extracted by 11 PQ meters in the Swedish town of Landskrona and 9 PQ meters in the Swedish town of Ulricehamn, giving a total of 20 PQ meters with data collected from January 2019 until the middle of March 2026, resulting in a total of 28113 data points. For simplicity, these data points will be referred to as days in this thesis. Next, the data was split into a training set (26339 days) and test set (1774 days), with roughly 7% of the data being used for the test set.

Next, the daily time windows of the PQ paremeters were extracted. Since 144 10 minute values are the equivalent of 1 day (24 hours), each day and parameter is represented by 144 values. For these daily time windows, and for each of the seven PQ parameters, the following five quantities were calculated to get a statistical representation of the data:

- Mean μ
- Standard deviation σ
- Maximum value max
- Minimum value min
- Percentile range $p = p_{95} - p_5$ (also referred to as robust range)

The variables p_{95} and p_5 are the 95th percentile and the 5th percentile, respectively. These statistical quantities were chosen to extract important characteristics of the signals while maintaining a reasonable feature space. The mean represents the central tendency of the PQ signal over the day, while the standard deviation quantifies

its variability. The minimum and maximum values provide information about extreme operating conditions. The robust range offers a measure of signal spread that is less sensitive to outliers than the full range, thereby improving stability under noisy conditions commonly present in PQ data. Together, these statistics represent relevant characteristics of each daily window, enabling the ML models to distinguish between different operating regimes and event patterns without relying on the raw time series [43, 44].

The data was then normalised per meter to have mean zero and unit variance. The structure of the final input vector x is shown in Equation 3.1 below:

$$x = \begin{bmatrix} \mu_{U_1} \\ \sigma_{U_1} \\ \vdots \\ \max_{U_{neg}} \\ \min_{U_{neg}} \\ \vdots \\ \max_{THD,U_1} \\ \min_{THD,U_1} \\ \vdots \end{bmatrix}. \tag{3.1}$$

Where μ_{U_1} is the mean value of the phase one voltage U_1 for this specific time window. Similarly, σ_{U_1} is the standard deviation of U_1, $\max_{THD,U_1}$ is the maximum value of the THD for U_1 for this specific time window, and so on. Additionally, the date and the meter ID were extracted as metadata for later analysis. In other words, this data was extracted but was not used in the training of the models. Instead, it was used after training to find the date and meter ID corresponding to a specific input.

3.1.1 Programming language and libraries

The programming language used was *Python* [45]. It is one of the programming languages suitable for ML tasks. The libraries used are presented in Table 3.1 below:

Table 3.1: *Python* libraries used.

Library	**Purpose**
matplotlib [46]	Visualisation of data
pytorch [47]	Implementation of AE/VAE
minisom [48]	Implementation of the SOM
hdbscan [49]	Clustering/calculation of OS
pandas [50]	Data management/numerical calculations
numpy [51]	Numerical calculations

3.2 Selected machine learning models

The models selected for this thesis are the SOM, AE and the VAE. The SOM and AE were chosen for their strong ability to model high-dimensional, non-linear, and unlabelled data, which is characteristic of PQ data [30, 36]. Both methods perform unsupervised learning and enable meaningful low-dimensional representations of complex data structures. The VAE is an extension of the conventional AE that introduce probabilistic and distribution-based constraints on the latent space [39]. These properties improve robustness, and interpretability of learned representations. Consequently, the VAE provides a natural and well-motivated progression from the standard AE and are therefore well suited for the objectives of this study. The AE and VAE will henceforth collectively be referred to as the AE variants.

3.2.1 Self-organising map structure

The hyperparameters of the SOM are shown in Table 3.2 below:

Table 3.2: Hyperparameters for the SOM

Model	SOM
Grid size	30x30
Neighbourhood function	Gaussian
Learning rate	0.5
Neighbourhood radius	15
Number of iterations	300 000

3.2.2 Autoencoder structure

The hyperparameters of the AE are listed in Table 3.3 below. In addition to the regular hyperparameters, weight decay and dropout layers were included to reduce overfitting.

Table 3.3: Hyperparemeters for the AE

Model	AE
Loss function	MSE
Optimisation method	AdamW
Learning rate	$1 \cdot 10^{-4}$
Weight decay parameter	$1 \cdot 10^{-5}$
Number of epochs	1000
Activation function	ReLU
Dropout rate	0.1

Moreover, the structure of the AE is shown in Table 3.4

Table 3.4: Structure of the AE.

Encoder Layers	Size	Type
1	In: 35, Out: 32	Linear (input)
2	-	ReLU
3	-	Dropout
4	In: 32, Out: 16	Linear
5	-	ReLU
6	-	Dropout
7	In: 16, Out: 8	Linear
8	-	ReLU

Latent dimension	Size	Type
–	8	Linear

Decoder layers	Size	Type
1	-	ReLU
2	In: 8, Out: 16	Linear
3	-	ReLU
4	In: 16, Out: 32	Linear
5	-	ReLU
6	In: 32, Out: 35	Linear (output)

3.2.3 Variational autoencoder structure

The hyperparameters of the VAE are listed in Table 3.5 below:

Table 3.5: Hyperparameters for the VAE

Model	VAE
Loss Function	MSE + KLD
Optimisation method	AdamW
Learning rate	$1 \cdot 10^{-4}$
Number of epochs	1000

Next, the structure of the VAE is shown in Table 3.6:

Table 3.6: Structure of the VAE.

Encoder Layers	Size	Type
1	In: 35, Out: 32	Linear (input)
2	-	ReLU
3	In: 32, Out: 16	Linear
4	-	ReLU
Latent Parameter Heads	**Size**	**Type**
μ	In 16, Out: 8	Linear
$\log(\sigma^2)$	In: 16, Out: 8	Linear
Decoder layers	**Size**	**Type**
1	-	ReLU
2	In: 8, Out: 16	Linear
3	-	ReLU
4	In: 16, Out: 32	Linear
5	-	ReLU
6	In: 32, Out: 35	Linear (output)

3.3 Evaluation methodology

Two methods are proposed for clustering, the first is clustering using the respective error metrics of each mode (RE/QE). The second is clustering using OS.

3.3.1 Reconstruction error/quantisation error

Each input day will have its RE (for the AE variants) or QE (for the SOM) calculated, depending on which model is used. These are methods of quantifying how anomalous the day is according to the models, and is often used in research to determine rare anomalies in data, both for SOM and the AE variants [52, 53]. Moreover, a threshold will be set to determine where the line for anomalous days will be set. In this paper, the 99th percentile of the RE/QE is set, to balance minimising false positives and capturing meaningful anomalies [54]. This is a common approach for anomaly detection in research.

3.3.2 Outlier score

Similar to the RE/QE pipeline, each input day will have its OS calculated using the HDBSCAN algorithm [29]. This is another approach for outlier detection in research. Next the 99th percentile threshold will be set for OS as well.

These anomaly detection methods capture different aspects of the data. The RE and QE pipelines focus on how well data points conform to the learned structure of the dataset, either through reconstruction or proximity to representative prototypes [33, 52, 53]. In contrast, the OS is based on local data density, identifying points

that lie in sparse regions in feature space [29]. Combining these approaches therefore provides a more comprehensive basis for anomaly detection in PQ data.

3.3.3 Self-organising map grid visualisation

To further analyse the behaviour of the SOM, detected anomalies are visualised on the trained SOM grid. This enables a geometric interpretation of how anomalous observations are distributed within the learned topology. By leveraging the inherent structure of the SOM, this visualisation provides insight into the relationships between normal and anomalous data points.

3.3.4 Histogram distribution plots with kernel density estimate

To evaluate how well the models and anomaly scores generalise to unseen data, the distributions of training and test data will be visualised using histogram density plots. These plots illustrate how error values are distributed by grouping similar values into bins, allowing for a comparison between training and test behaviour. In addition, days identified as anomalous will be extracted and analysed in separate histogram plots. Comparing these distributions provides an indication of whether the model behaves consistently on unseen data. This gives an assessment of generalisation of the models in the absence of ground truth. KDE will also be used to provide a continuous representation of the histogram distribution. This will allow a better understanding between the training and test data.

3.3.5 Agreement of the models using Jaccard index overlap

To examine the agreement between the models, the Jaccard index is calculated pairwise between the different anomaly detection methods. This metric quantifies the overlap in identified anomalous days, providing insight into the consistency between models and highlighting whether any approach deviates from the others.

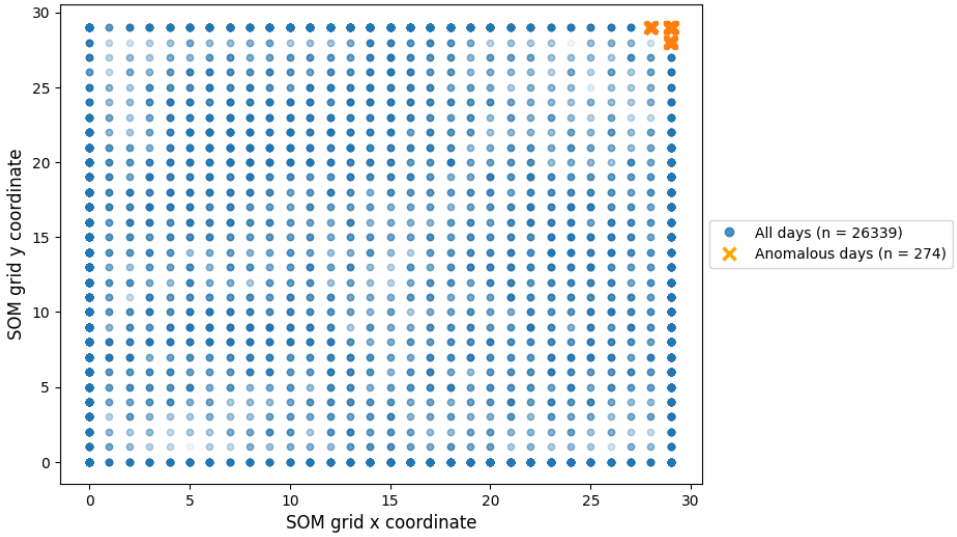
4
Results

This chapter presents the results obtained using the methodology described in chapter 3. The findings address the research questions by comparing the models and the two anomaly detection methods.

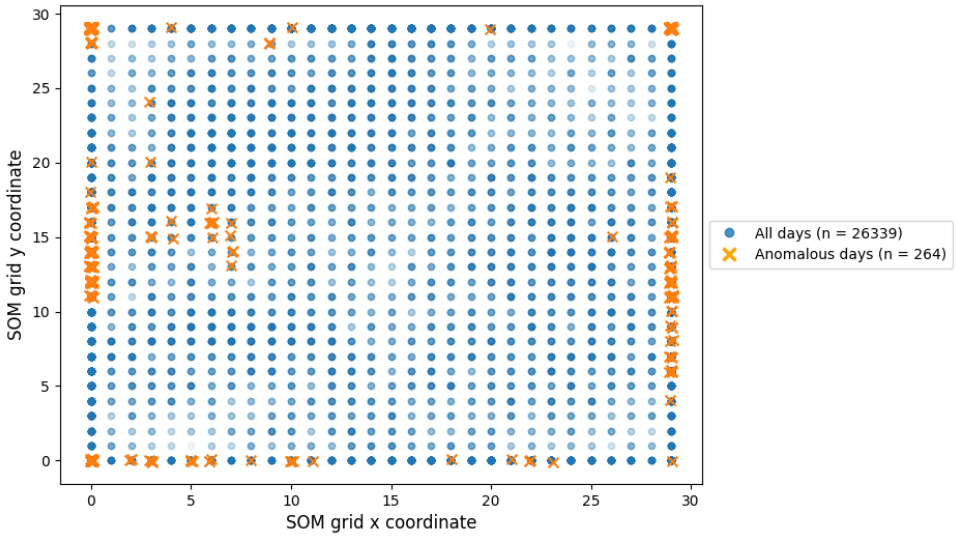
In the figures of this chapter, n denotes the number of days. When multiple subsets of days are considered within a figure, n is used separately for each subset and is specified when necessary.

4.1 Self-organising map results

Figure 4.1 illustrates how the data is mapped onto the SOM grid. This representation provides an overview of how the model organises the data and highlights possible anomalous regions in the map. It is worth reiterating that the anomalous days are not the same in the two anomaly detection methods, but rather independently found and classified accordingly.



(a)



(b)

Figure 4.1: The grid of the SOM and the anomalies according to OS and QE, shown in subfigure a) and b), respectively. Each day, shown in blue, is mapped to a neuron. The neurons are located at each integer coordinate in the figure. The transparency of the blue points represent the number of days mapped to that specific neuron. The more transparent, the less days mapped to a neuron. The orange crosses are days considered anomalies. Some statistical noise is added to the crosses to better visualise the occurrence of several anomalies in one location. Figure 4.1a has more anomalies (274) compared to Figure 4.1b (264). This is because there were several data points with identical OS. All of these were selected to be considered anomalous rather than excluding some of them to enforce exactly 1% anomalies.

Figure 4.1a shows the outliers concetrated in the top right corner. This indicates that the SOM and OS pipeline creates a region in the top right corner of the grid where all anomalies are mapped to. In other words, this pipeline manages to separate

anomalies from normal days, but it does not distinguish between anomaly types. On the other hand, Figure 4.1b displays several anomaly clusters spread out over the grid. This means that the SOM and QE pipeline identifies different types of anomalies, as indicated by the mapping of anomalies to different regions.

4.1.1 Comparison of the anomaly detection methods

It is of interest to compare the two anomaly detection methods by analysing which days they consider anomalous and to what extent. This is shown in Figure 4.2.

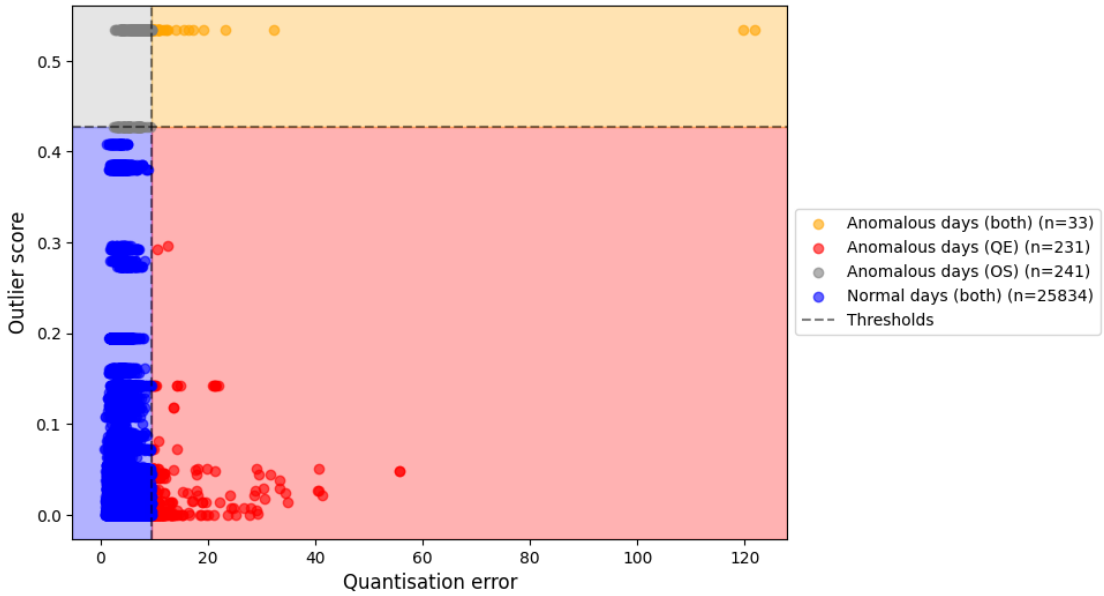


Figure 4.2: QE of the SOM plotted against the OS of the SOM for the training days. Each point represent one day. The figure is divided into four colour-coded zones by the two thresholds. The zones represent if a day is considered anomalous by both, none, or one of the scores. There is a discrete pattern on the OS axis, with many days having identical OS. This is because each data point is mapped to a neuron, and the OS is calculated for each neuron and then assigned to each data point of that neuron. This gives data points with the same BMU identical OS.

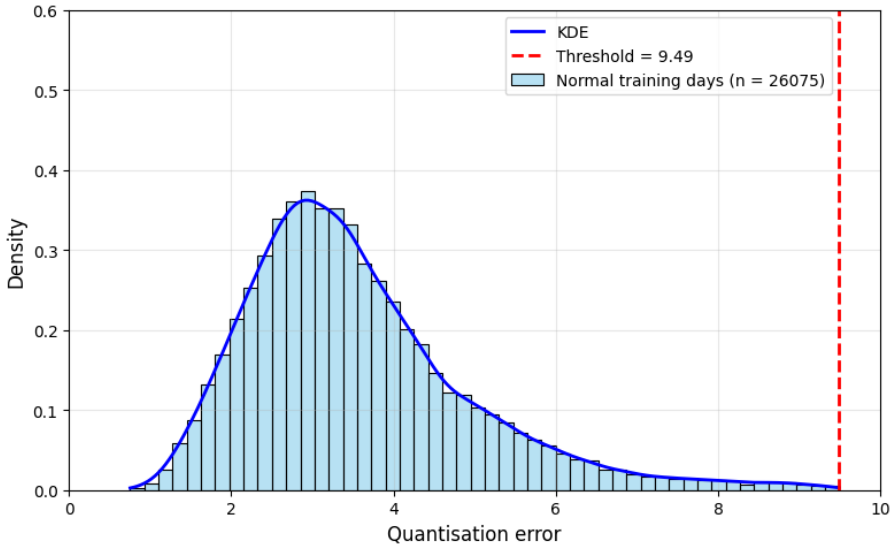
Figure 4.2 shows that the two anomaly detection methods behave differently. The majority of days have a relatively low QE, while the OS method has more spread. Two days can be seen in the top right corner. These two days are considered extremely anomalous by both methods. There are several days in the grey and the red area, respectively. This means that the models disagree strongly on which days are anomalous. On the other hand, both models consider the vast majority of days to be normal (the blue region). There are also some days in the red region with quite high QE (in the range 40-60) while having quite low OS. This indicates strong disagreement for a few days between these two methods.

While OS is defined to be between 0 and 1, the maximum OS assigned was around

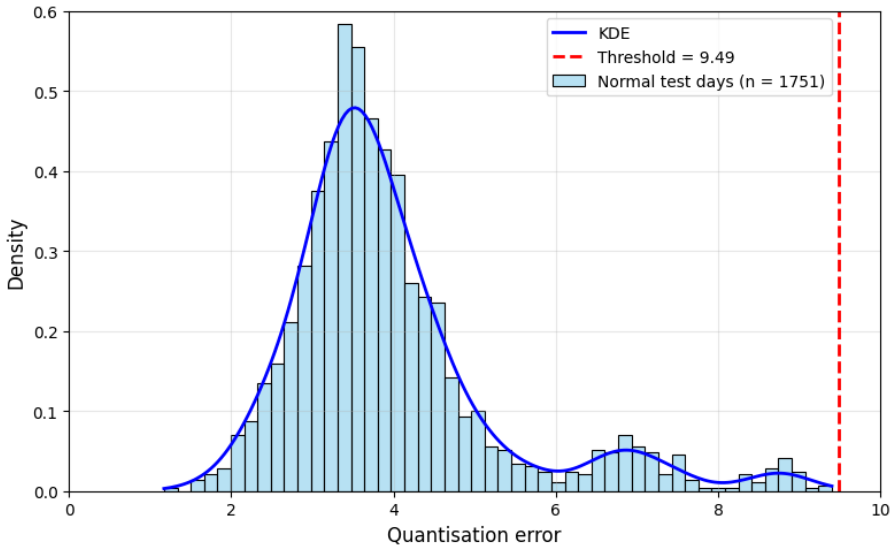
0.55. In other words, the OS pipeline does not consider any of the days to be extremely anomalous.

4.1.2 Quantisation error histogram distribution

In order to further evaluate the model, it is of interest to see how it performs on unseen (test) data and compare this with seen (training) data. This is visualised in Figure 4.3 below.



(a)



(b)

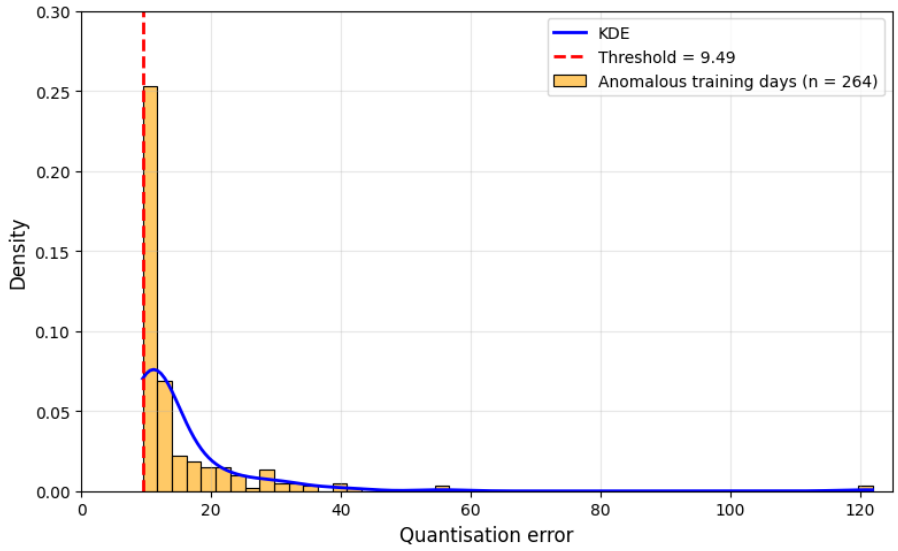
Figure 4.3: Histogram distribution plots of the normal training days and test days for the SOM. The dashed red line represents the threshold calculated from the training set. The solid blue line is the KDE.

Figure 4.3a shows that the dataset is skewed to the right with a mean QE close to

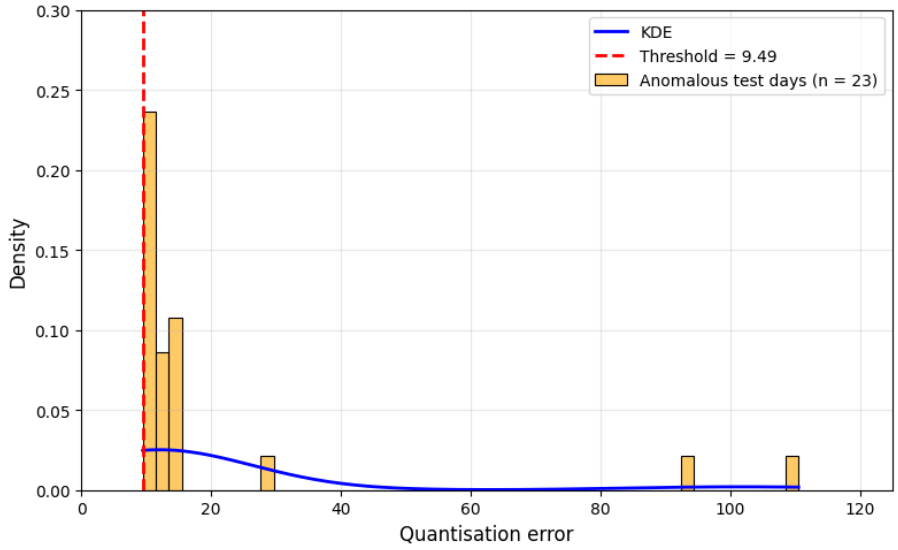
3. The data has a tail to the right. This indicates that as the QE increases, the density of days is reduced. In other words, there are few days with high QE.

Figure 4.3b has a mean QE close to 3.8 and is also skewed to the right with a right tail. There are noticable larger densities of days when the QE is around 7 and 9, compared to Figure 4.3a. This indicates that the model had a harder time mapping the test set to the SOM. In other words, the test set is in general more erroneous compared to the training set according to the SOM. The two distribution plots show some similarity in having a mean value skewed to the right, followed by a right tail.

To further analyse the distribution of the data, the anomalous days of the training and test set are extracted and visualised in Figure 4.4 below.



(a)



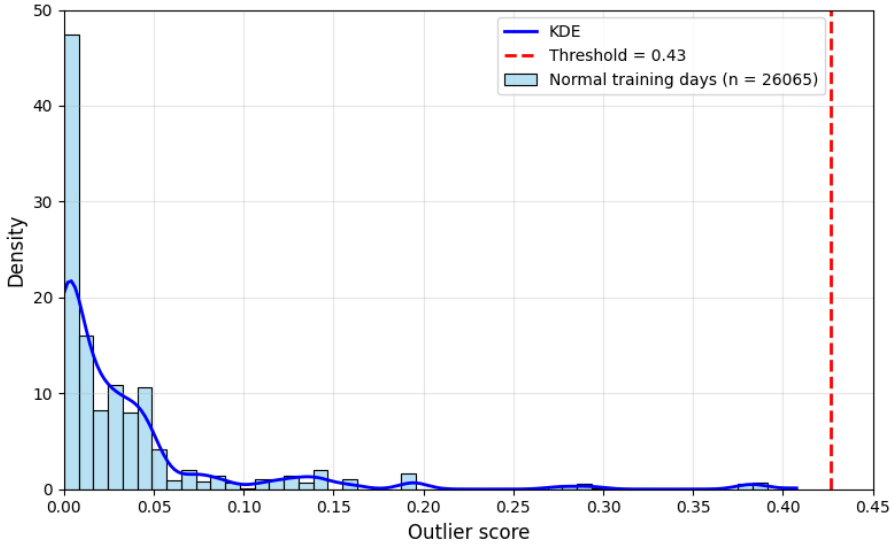
(b)

Figure 4.4: Histogram distribution plots of the anomalous training days and test days for the SOM. The dashed red line represents the threshold calculated from the training set. The solid blue line is the KDE.

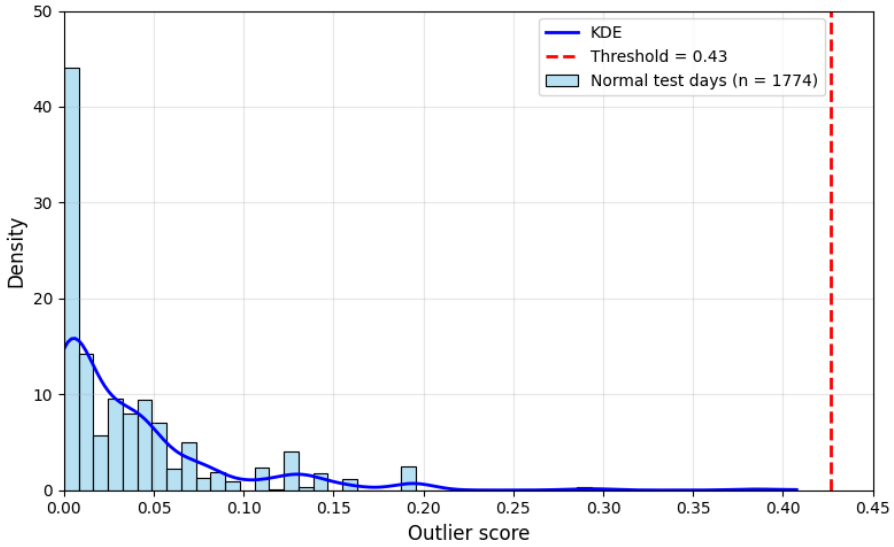
Figure 4.4a shows a tall histogram close to the threshold (when QE is 9.49) followed by a right tail. There are extreme outliers with a QE of around 120. Similarly, Figure 4.4b shows a large density close to the threshold, with a right tail, and some days with large QE close to 100. The KDE in Figure 4.4a has a more significant peak compared to Figure 4.4b. This means that the anomalous days of the training set are more concentrated close to the threshold compared to the test set. These figures show some similarity, indicating that the model generalises well to unseen data.

4.1.3 Outlier score histogram distribution

Next, similar figures displaying the distribution of OS over the dataset is shown, starting with Figure 4.5:



(a)



(b)

Figure 4.5: Histogram distribution plots of the normal training and test days for the SOM. The dashed red line represents the threshold calculated from the training set. The solid blue line is the KDE.

Figure 4.5a and Figure 4.5b show similar shape and size. Indicating that the model generalises well for this OS pipeline. However, the peak of the KDE differs. There is small density of days when the OS is around 0.4 in both sub figures.

There were no identified anomalous test days for the OS pipeline. In other words,

all days in the test set had an OS lower than threshold. Due to this, no figure is shown for this case. No found anomalous test days can indicate that the threshold is too high for this case, or that the test set is too small.

4.2 Autoencoder results

Here, the results from the AE will be presented. For comparison, similar figures have been used as for the previous model (see section 4.1).

4.2.1 Comparison of the anomaly detection methods

Figure 4.6 shows the training days and their classification according to the two anomaly detection methods.

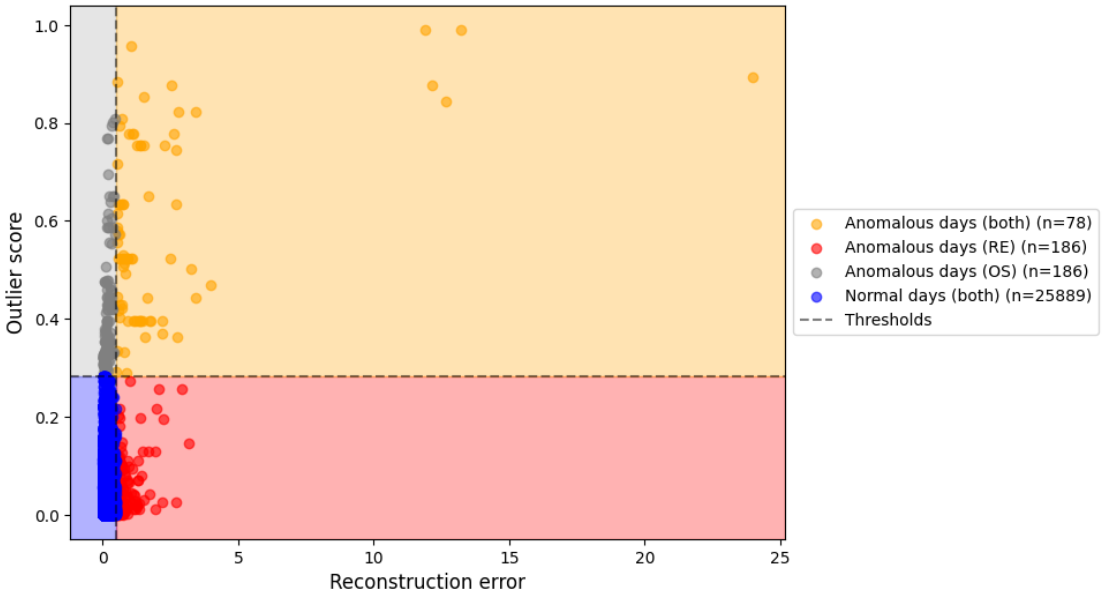
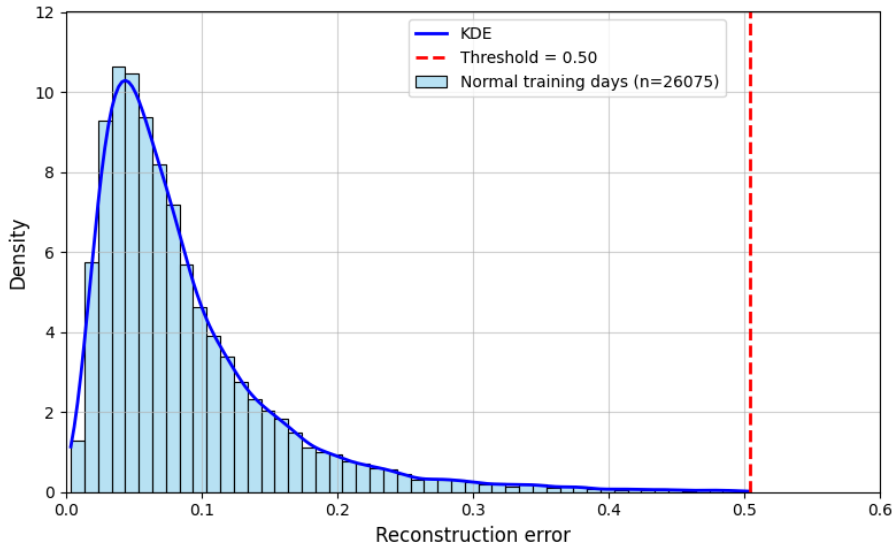


Figure 4.6: RE of the AE plotted against the OS of the AE for the training days. Each point represent one day. The figure is divided into four colour coded zones, by the two thresholds. The zones represent if a day is considered anomalous by both, none, or one of the scores.

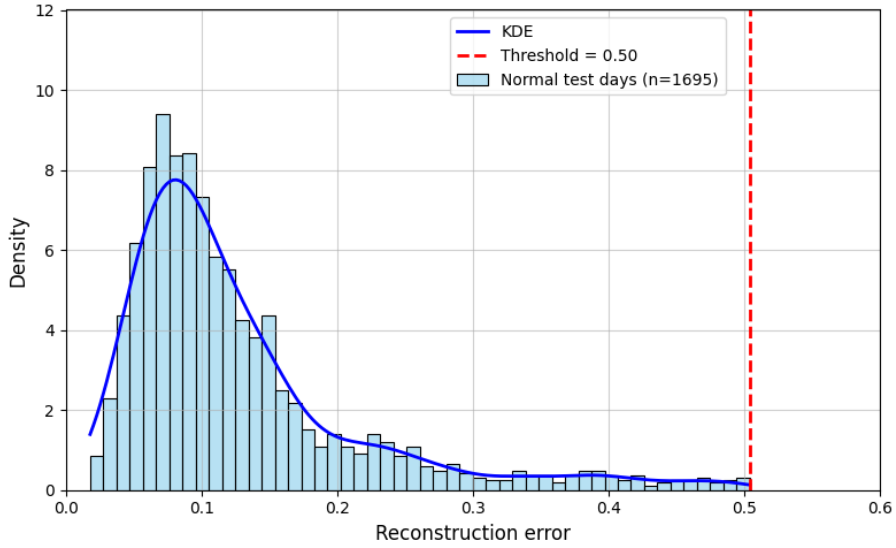
Figure 4.6 shows that the majority of days are considered normal for both methods. There are several days where the two methods do not agree on anomalous behaviour. The RE is low for most days, while OS is more spread out. Some correlation can be identified for days with high RE ($12 \leq \text{RE} \leq 25$) and high OS ($0.8 \leq \text{OS} \leq 1.0$). This indicates that the pipelines agree on extreme anomalous events.

4.2.2 Reconstruction error histogram distribution

The distribution of the RE is presented below, starting with Figure 4.7.



(a)

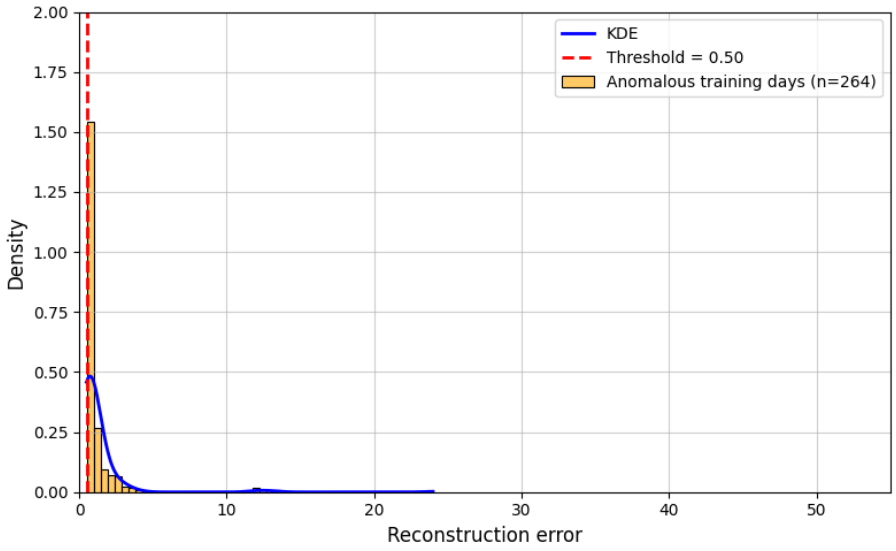


(b)

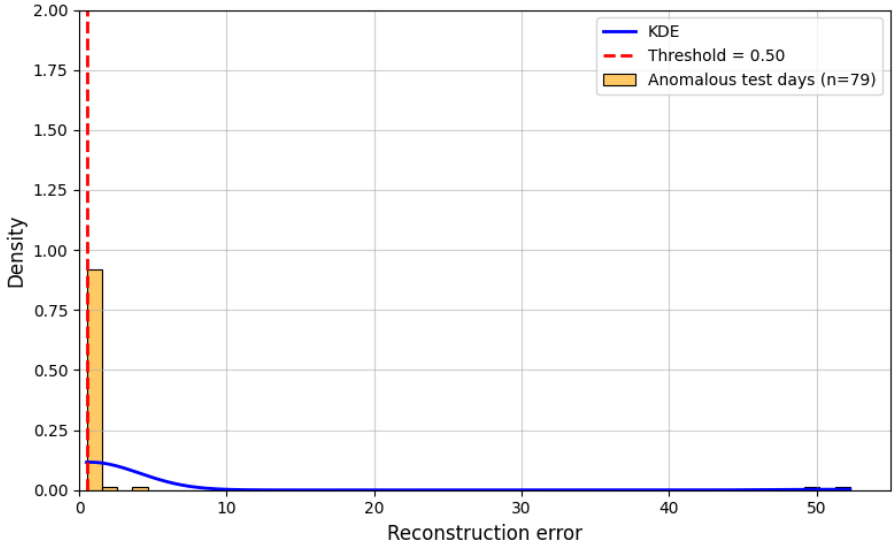
Figure 4.7: Histogram distribution plots of the normal training days and test days for the AE. The dashed red line represents the threshold calculated from the training set. The solid blue line is the KDE.

Figure 4.7 shows a right-skewed distribution of the normal days in the training and test days. The distribution of the test set (Figure 4.7b) show a larger mean, a larger tail and a lower maximum KDE, compared to the training set in Figure 4.7a. This shows that the model considers the test set more difficult to reconstruct. Nevertheless, the distribution shape of the training and test are fairly consistent, which is an indication that the model can interpret unseen data well.

Next, the distribution of the anomalous days of the training and test set are investigated in Figure 4.8.



(a)



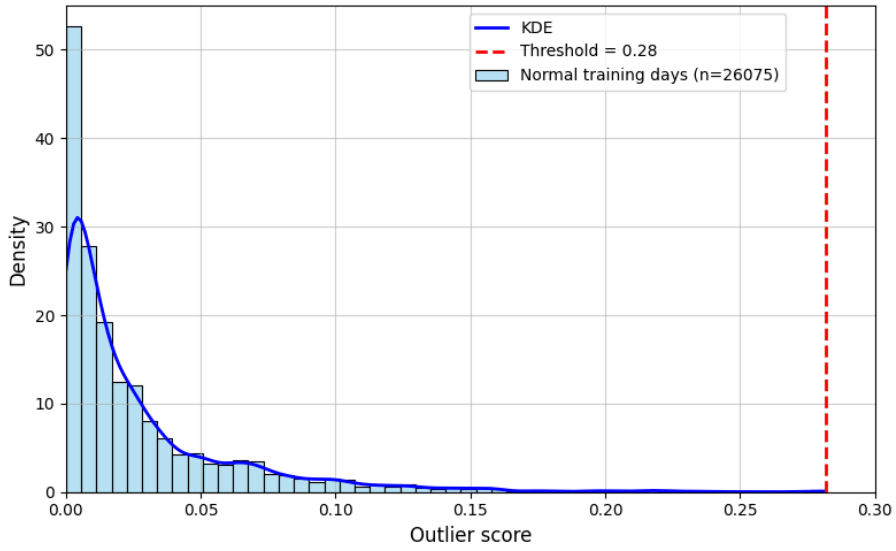
(b)

Figure 4.8: Histogram distribution plots of the anomalous training days and test days for the AE. The dashed red line represents the threshold calculated from the training set. The solid blue line is the KDE.

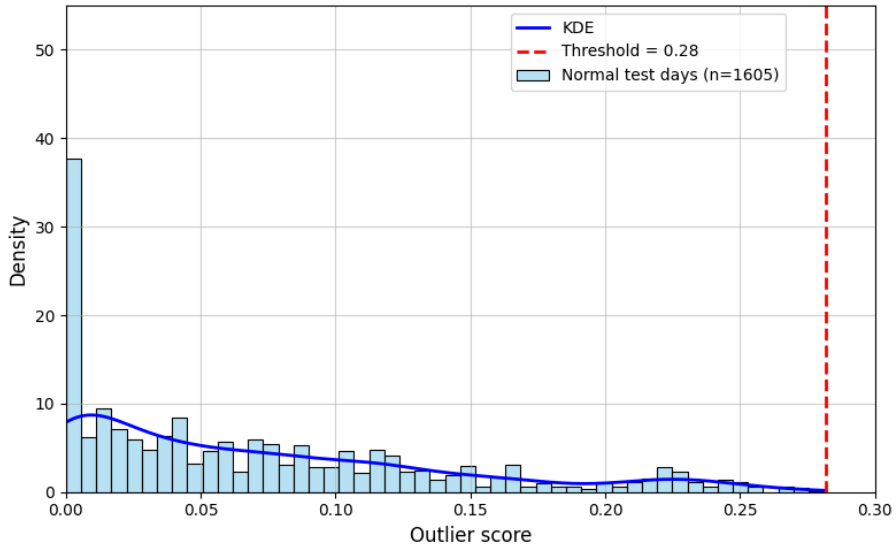
Figure 4.8 shows a right-skewed distribution of the anomalous training and test days. The two distribution have a similar shape, but Figure 4.8a displays a higher density of days close to the threshold. Additionally, Figure 4.8b shows that the test set has some extreme anomalies when the RE is close to 50. The smaller KDE peak in Figure 4.8b is can be caused by these extreme anomalies.

4.2.3 Outlier score histogram distribution

In this section, the OS distribution will be visualised for the AE. The distribution of the normal training and test days are shown in Figure 4.9:



(a)

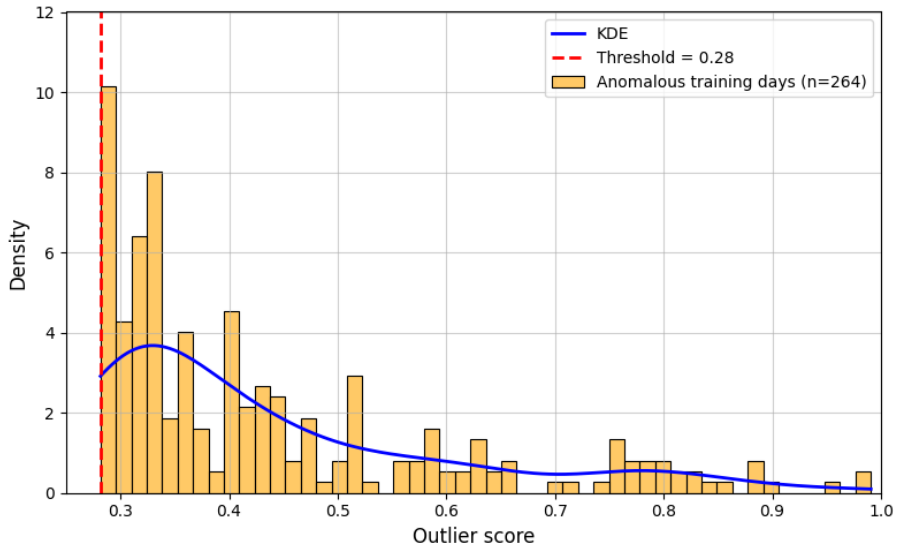


(b)

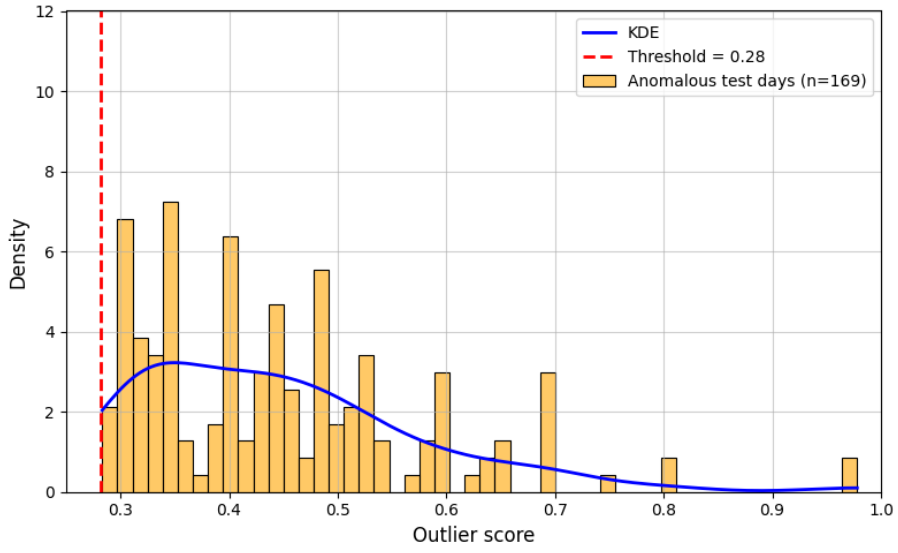
Figure 4.9: Histogram distribution plots of the anomalous training days and test days for the AE. The dashed red line represents the threshold calculated from the training set. The solid blue line is the KDE.

Figure 4.9a shows a clear right-skewed distribution. The distribution in Figure 4.9b is also right-skewed, but the shape is different. The first bin is tall in Figure 4.9b, followed by much smaller and slowly decreasing bins. This is different compared to Figure 4.9a, which shows a more drastic decrease in the length of the histograms. This is an indication that the AE OS pipeline interprets unseen data somewhat similar but more spread out with higher OS for more days.

Next, the distribution of the anomalous days are shown in Figure 4.10:



(a)



(b)

Figure 4.10: Histogram distribution plots of the anomalous training days and test days for the AE. The dashed red line represents the threshold calculated from the training set. The solid blue line is the KDE.

Figure 4.10a shows a right-skewed distribution of the anomalous training days. Figure 4.10b shows a more spread out distribution with larger densities of days as the OS increases, compared to the tail in Figure 4.10a. One can hint a right skewed distribution in Figure 4.10b, in particular due to the KDE, which shows similarity in both subfigures.

4.3 Variational autoencoder results

Here, the results from the VAE will be presented. For comparison, similar figures have been used as for the previous models. An overview of the training data set is visualised in Figure 4.11.

4.3.1 Comparison of the anomaly detection methods

Figure 4.11 shows the training days and their classification according to the two anomaly detection methods.

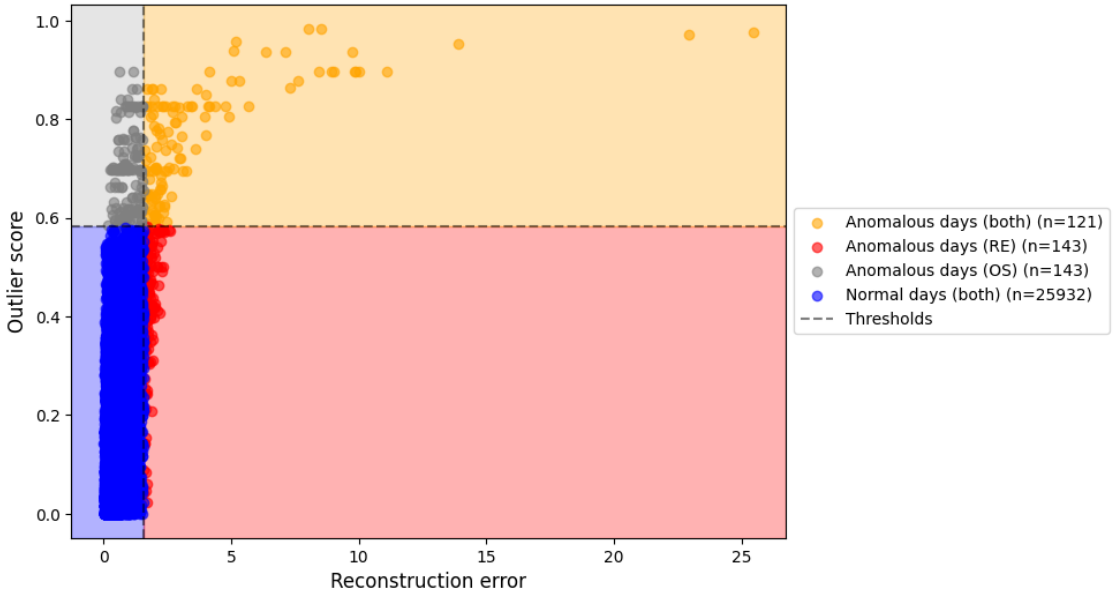
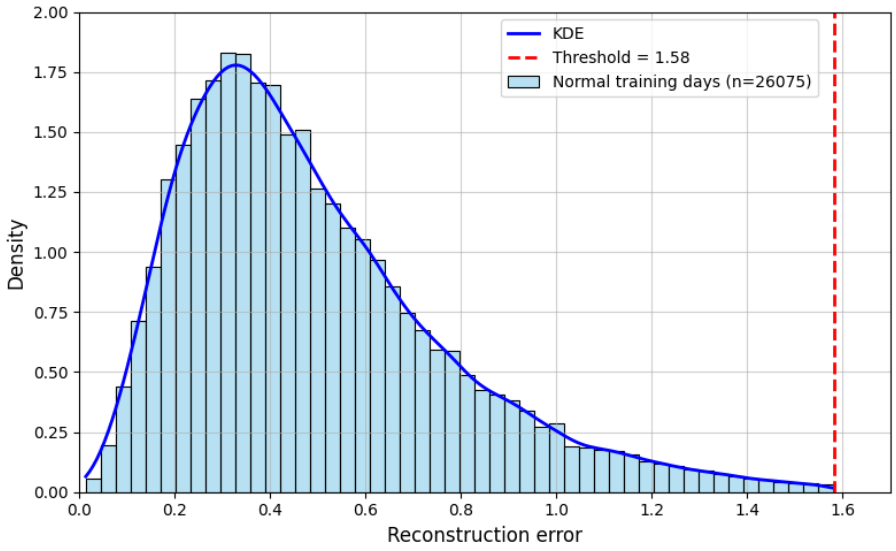


Figure 4.11: RE of the VAE plotted against the OS of the VAE for the training days. Each point represent one day. The figure is divided into four colour coded zones, by the two thresholds. The zones represent if a day is considered anomalous by both, none, or one of the scores.

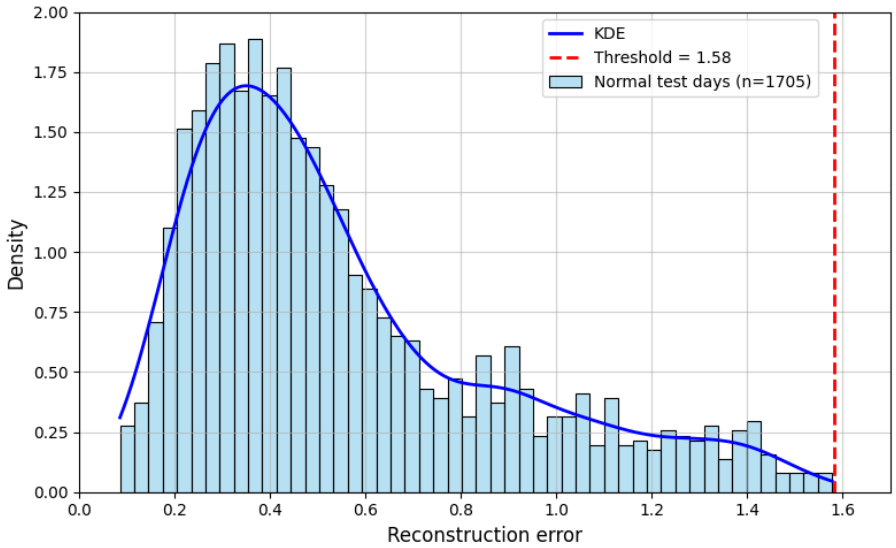
Figure 4.11 shows that the majority of days are considered normal by both methods. The RE is low for most days, while OS is more spread out. Some correlation can be identified for days with high RE ($8 \leq \text{RE} \leq 25$) and high OS ($0.9 \leq \text{OS} \leq 1.0$). This indicates that the pipelines agree on extreme anomalous events.

4.3.2 Reconstruction error histogram distribution

The distribution of the RE will be analysed in this section. Figure 4.12 shows the distribution of the normal days in the training and test set.



(a)

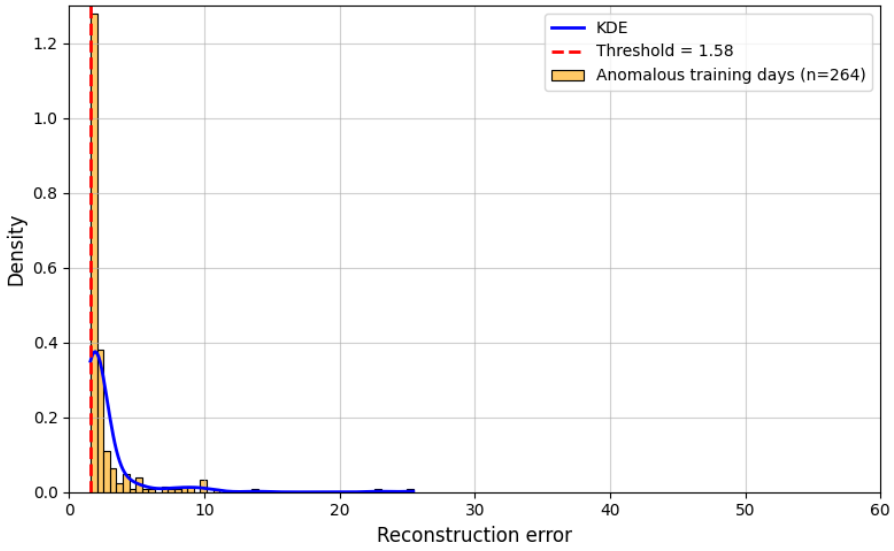


(b)

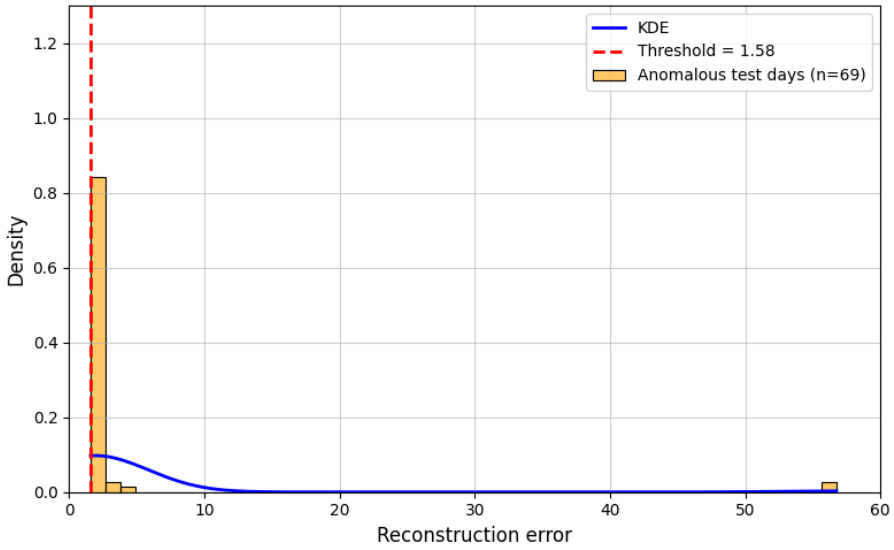
Figure 4.12: Histogram distribution plots of the training days and test days for the VAE. The dashed red line represents the threshold calculated from the training set. The solid blue line is the KDE.

Figure 4.12 show similar shape, size and KDE for the training and test days, with Figure 4.12b displaying a bit noisier distribution and larger densities of days with large RE. Nevertheless, Figure 4.12 indicates correlation between the normal days of the training and test set.

Moreover, the distribution of the anomalous training and test days are shown in Figure 4.13:



(a)



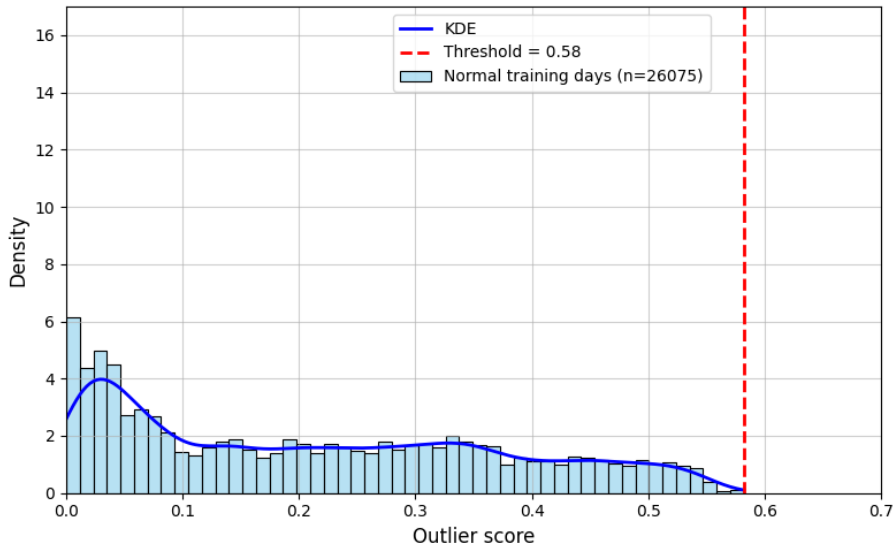
(b)

Figure 4.13: Histogram distribution plots of the anomalous training days and test days for the VAE. The dashed red line represents the threshold calculated from the training set. The solid blue line is the KDE.

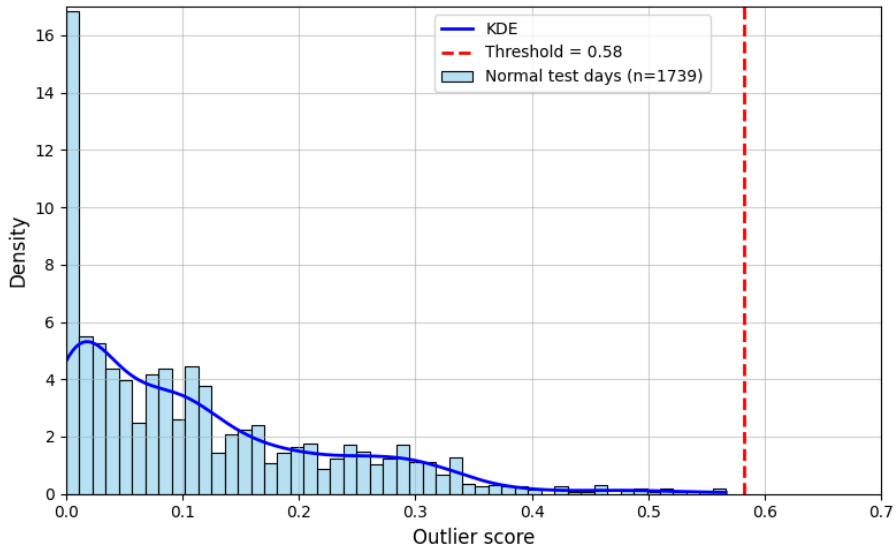
Figure 4.13 shows similarity between the anomalous training and test days. Both sub-figures display that the majority of anomalous days are close to the threshold. Figure 4.13a has a taller first bin, and consequently a larger KDE peak. This can be explained by the density of days with a RE close to 60 in Figure 4.13b. In short, the shape is similar, but the large RE values of the test set contribute to a more spread out distribution.

4.3.3 Outlier score histogram distribution

In this section, the distribution of the OS is investigated. Figure 4.14 shows the distribution of the normal training and test days.



(a)

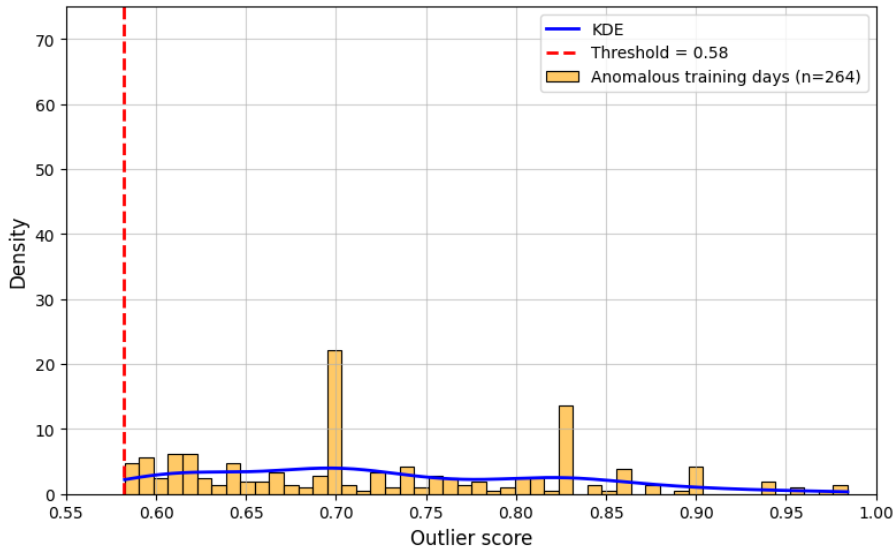


(b)

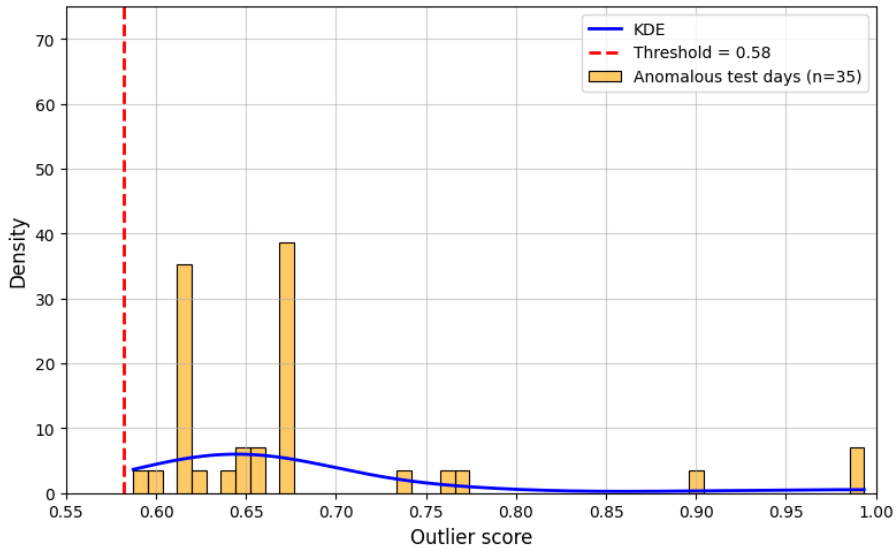
Figure 4.14: Histogram distribution plots of the normal training days and test days for the VAE. The dashed red line represents the threshold calculated from the training set. The solid blue line is the KDE.

Figure 4.14 shows a bit of agreement between the training and test set. Figure 4.14a shows that the normal training days are almost uniformly distributed. On the other hand, Figure 4.14b displays a right skewed distribution. This indicates that the VAE OS pipeline does not generalise well for unseen data.

Furthermore, the distributions of the anomalous days are visualised in Figure 4.15.



(a)



(b)

Figure 4.15: Histogram distribution plots of the anomalous training days and test days for the VAE. The dashed red line represents the threshold calculated from the training set. The solid blue line is the KDE. There is a small density of days with OS close to 1 in Figure 4.15a.

Figure 4.15 show little agreement. Figure 4.15a indicates a rather uniform distribution with density peaks when OS is around 0.70 and 0.83. Moreover, Figure 4.15b show a more spread out distribution with the data concentrated at lower OS values, with a few outlier. This can be due to a small sample size of anomalous test days, or that the VAE OS pipeline fails to generalise to unseen data.

4.4 Overlap and agreement between models

It is of interest to analyse how much the models agree on what days are anomalous. Table 4.1 is presented to show overlap between the models for the QE/RE pipeline. The overlap here is defined as the Jaccard index.

Table 4.1: Table over the Jaccard index comparison of error (RE for the autoencoder variants and QE for the SOM) for the training set between the models used. The cells in the lower left part of the table are coloured black, since they provide the same information as the upper triangle.

Jaccard Index (Error %)	**AE**	**VAE**	**SOM**
AE	100.0	26.32	24.53
VAE		100.0	19.19
SOM			100.0

Table 4.1 shows that the overlap between the models are quite similar, around 19-26%. In other words, the models agree on about a quarter of identified anomalous days, and disagree on the rest.

Next, the overlap between the models for the OS pipeline is shown in Table 4.2.

Table 4.2: Table over the Jaccard index comparison of OS for the training set between the models used, shown as a percentage. The cells in the lower left part of the table are coloured black, since they provide the same information as the upper triangle.

Jaccard Index (OS %)	**AE**	**VAE**	**SOM**
AE	100.0	22.22	5.81
VAE		100.0	4.55
SOM			100.0

Table 4.2 shows that the SOM has a low overlap (around 5%) with the autoencoder variants. The autoencoder variants show a bit more agreement (around 22%).

5
Discussion

In this chapter, the results, method and implications of the work will be discussed. Firstly, a general discussion regarding observations of all models and pipelines will be presented. Next, the models and pipelines are discussed in more detail one by one. This is then followed by a discussion regarding the methodology of the thesis, comparisons with related work, ethical impacts, the impact of the limitations on the results, and finally suggestions for future work.

5.1 General discussion

In general, all models were able to identify anomalous days based on their respective scoring mechanisms (QE/RE and OS). A relatively high degree of consistency was observed within each anomaly detection pipeline, particularly for the RE and QE approaches. However, agreement between different models was lower, which is expected for several reasons.

First, the models have inherent stochasticity due to random initialisation and training procedures (e.g., weight initialisation and stochastic optimisation). As a result, repeated training on the same dataset may yield slightly different representations and, consequently, different anomaly scores. While this variability is typically limited, it contributes to discrepancies between model outputs.

Second, the SOM and AE variants rely on fundamentally different principles for detecting anomalies. In the SOM, anomalies are identified through the QE, which measures the distance between an input sample and its best matching unit BMU. This reflects how well the input is represented within a discrete prototype-based mapping of the data space. In contrast, AE-based models define anomalies in terms of RE, where anomalous samples are those that the model is unable to accurately reconstruct. This corresponds to deviations from a learned continuous data manifold. Although both QE and RE can be interpreted as distance-based measures and are therefore comparable, they operate in fundamentally different representation spaces. Discrete prototype space for SOMs and continuous latent manifold space for AEs reflects distinct notions of what constitutes an anomaly.

Finally, the OS pipelines apply a common anomaly detection method to transformed representations of the data. However, these representations differ substantially between models. For the SOM, OS is computed on discrete mappings within

a predefined grid structure, whereas for AE-based models it is applied to continuous latent representations in an eight-dimensional space. Since the geometry and density structure of these representations differ. Consequently, variation in detected anomalies across models is expected.

A fundamental challenge in unsupervised learning is the absence of labeled data, which makes it difficult to systematically identify false positives and false negatives. This limitation is particularly relevant in the context of the QE, RE, and OS pipelines, where anomaly detection is based solely on model-specific representations and scoring mechanisms.

For instance, high QE in the SOM or high RE in the AE variants may indicate anomalous behaviour, but may also correspond to rare yet valid patterns in the data. Similarly, the OS, which is based on density in the transformed feature space, may incorrectly flag low-density normal observations as anomalies or fail to detect anomalies that form dense clusters.

Consequently, without access to ground truth labels, it is not possible to definitively assess whether detected anomalies correspond to true deviations or model-specific artifacts. This introduces uncertainty in the interpretation of results and highlights the importance of combining quantitative analysis with domain knowledge when evaluating anomaly detection performance. However, the findings of this thesis is not that the anomalies detected are absolute ground truth. Instead, the findings suggest that the models reliably detect distinct and infrequent patterns across multiple anomaly criteria, while simultaneously exhibiting consistent and stable performance when applied to unseen data.

5.1.1 Anomaly metric comparison

Overall, the RE/QE pipeline provided more consistent identification of anomalous days than the OS pipeline. This is based on the histogram plots showing more consistency for the RE/QE pipelines. But also that manual inspection, based on visible deviations, suggested that RE- and QE-based detections more often corresponded to clear anomalies. In contrast, the OS approach produced less similar distributions, as well as a higher proportion of cases without obvious PQ deviations. This indicates a greater tendency toward false positives in this setting.

This difference can be explained by the nature of the anomaly scores. RE and QE are distance-based measures that directly quantify deviation from learned normal behaviour. This means that the QE/RE pipelines can identify points that are globally separated, but in dense local clusters as anomalous. On the other hand, the density-based OS identifies anomalies based on local sparsity rather than deviation from a learned data distribution. While effective for detecting isolated points, they can struggle when similar anomalies cluster together, as such points may not appear locally sparse. Conversely, rare but valid patterns may be incorrectly flagged as anomalies due to their low local density, even if they are close to the data distri-

bution in a global sense.

Additionally, the OS pipeline is sensitive to the underlying data representation. Applying HDBSCAN's OS to latent spaces introduces dependence on how well models such as SOM, AE, and VAE preserve meaningful geometric and density relationships. Distortions in these representations can reduce detection consistency. This issue is less pronounced in the RE/QE pipeline, where anomalies are evaluated directly using model-specific error measures.

Overall, these results suggest that distance-based methods are better aligned with the characteristics of PQ data. If anomalies are defined as deviations from expected behaviour. However, density-based approaches remain relevant in contexts where anomalies are rare and unstructured. This highlights the importance of selecting methods that match both the data properties and the domain-specific definition of anomalies. Moreover, the limited agreement between the methods should not be viewed as a limitation, but rather as an indication that they capture complementary aspects of anomalous behaviour. Combining distance-based and density-based approaches may therefore provide a more comprehensive anomaly detection framework.

5.1.2 RE/QE histogram distribution plots

For all RE/QE pipelines, the histogram distributions show consistent unimodal, right-skewed behaviour between normal training and test days, while showing a decaying density distribution for anomalous training and test days. The test set exhibits higher QE/RE values, reflected in both an increased mean and greater density at larger errors. This is expected, as models are optimised on training data and typically produce larger errors on unseen samples. Importantly, the overall distributional shapes are preserved despite this shift, indicating that the models generalise well. However, a substantial change in shape, rather than a simple shift, would instead suggest poor generalisation or dataset mismatch, which is not observed here.

As mentioned above, the anomalous days according to all QE/RE pipelines are concentrated around the threshold. This is followed by a decay in densities for larger errors. A reason for this can be that there are some extreme anomalies such as blackouts or long-lasting spikes/drops in the parameters for certain days. These rare anomalous events produce large errors that stretch the distributions. The majority of anomalous days are less prominent anomalies, as indicated by the concentrated density around the threshold value.

On the other hand, the OS pipelines show less similarity between models, as well as less consistency between the training and test sets. Particularly for the AE variants. This may indicate weaker generalisation performance and is further examined later in the chapter when each model is examined individually.

5.2 Self-organising map performance

In this section, the results of the SOM are discussed in more detail.

5.2.1 Self-organising map grid

Figure 4.1a shows that anomalies identified by the OS pipeline are concentrated in the top right region of the SOM. This suggests that while the method separates anomalous observations from normal data, it does not meaningfully distinguish between different types of anomalies. Instead, anomalous points are mapped to a common region due to their shared dissimilarity from the majority of the data. This behaviour is characteristic of density-based methods such as HDBSCAN, which identify anomalies as points located in low-density regions of the data space. A SOM must fit a continuous data ditribution onto a finite two-dimensional grid. As a consequence, it allocates most of its neurons to dense regions, forcing sparse regions to boundaries. Hence, low-density regions are typically mapped to contiguous areas of the grid, often near its boundaries or corners, resulting in a localised cluster of anomalies. This is therefore consistent with the structure of the SOM and the density-based nature of the outlier score.

In contrast, Figure 4.1b shows that anomalies detected using QE are more spatially distributed, with many located along the boundaries of the map. This difference arises because the QE measures the distance between an observation and its BMU, and therefore captures how well the SOM represents the input rather than the local data density. As a result, QE identifies observations that are poorly represented by the learned prototypes, regardless of their position in the data distribution, leading to a more distributed set of anomalies across the map.

Additionally, the presence of many QE anomalies at the edges and corners of the SOM can be explained by both the data distribution and the topology of the grid. Boundary neurons typically represent regions of lower data density and receive fewer updates during training due to having fewer neighbouring nodes. Consequently, these neurons may provide a poorer representation of the data, resulting in higher QE for observations mapped to these regions. Since each input is always assigned to its BMU, even highly dissimilar or extreme points are forced onto the grid, and are therefore often projected onto edge or corner regions.

A notable cluster of anomalies appears around $(x, y) \approx (6, 16)$ in Figure 4.1b. Manual inspection reveals that these days share a common pattern, characterised by a large spike in VUF while THD and RMS values remain stable. This indicates that the SOM is capable of grouping structurally similar anomalies, reflecting its ability to preserve local similarity in the input space.

The presence of both clustered anomalies and edge-located anomalies suggests varying degrees of similarity. Some anomalous observations share common structure and are mapped to the same region, while more extreme or isolated anomalies are pushed

toward the boundaries. This highlights that the QE-based SOM pipeline not only detects anomalies but can also differentiate between distinct types of anomalous behaviour.

Overall, these findings indicate that QE captures both globally poorly represented observations and boundary effects of the SOM, while the OS pipeline primarily reflects underlying density structures in the data. The differing spatial distributions of anomalies are therefore expected and demonstrate that the two methods identify complementary types of anomalous behaviour.

Figure 4.1 illustrates a key advantage of the SOM: its inherent interpretability. By projecting high-dimensional data onto a two-dimensional grid, the SOM provides a direct visual representation of data structure. In contrast, AE-based models learn latent representations that are not inherently constrained to low dimensions. While this allows them to capture more complex relationships, it reduces interpretability, as the latent space is not directly visualisable without additional processing.

Although AE representations can be visualised by restricting the latent space or applying dimensionality reduction, such approaches may limit model expressiveness or introduce additional transformations. The SOM therefore offers a natural trade-off for simplicity and interpretability.

5.2.2 Quantisation error vs outlier score

Figure 4.2 shows limited agreement between the two anomaly detection methods. Although two observations are identified as highly anomalous by both methods (QE around 120 and OS around 0.5), there is otherwise little correspondence. In general, QE appears more conservative in assigning high anomaly scores compared to OS.

This discrepancy can be explained by the fundamentally different definitions of anomaly underlying the two metrics. As explained in section 5.1.1, QE measures the distance to the BMU and therefore captures how poorly an observation is represented by the SOM, regardless of how many similar observations exist. In contrast, OS is based on local density and assigns high scores primarily to points that are isolated from their surroundings.

As a result, observations that deviate from typical patterns but occur in groups may yield high QE but low OS, since they form local structures in the data space. Conversely, points in sparse regions may receive high OS despite appearing qualitatively normal. This is supported by manual inspection, where several observations with high OS did not exhibit clearly anomalous behaviour from a domain perspective.

This highlights an important distinction between statistical and practical definitions of anomalies. While OS captures relative isolation, it does not necessarily correspond to meaningful deviations in the data. QE, on the other hand, was found

to align more closely with qualitative assessments, as it directly reflects deviation from learned normal behaviour.

These results suggest that the two methods detect different types of anomalies. QE is more sensitive to systematic deviations from expected patterns, whereas OS primarily identifies isolated or rare observations. This indicates that relevant anomalies often manifest as structured or repeated deviations rather than isolated points, and QE may provide a representation that is more consistent with domain-specific interpretations.

5.2.3 Histogram distribution plots

Figure 4.3–4.5 show consistent behaviour between training and test data, indicating that both pipelines generalise well to unseen observations. However, the two anomaly scores exhibit markedly different distributions due to the distinct principles on which they are based. As mentioned in subsection 5.1.2, the QE distribution is unimodal with a moderate right skew, meaning that the majority of days that represent normal operating behaviour has a non-zero distance to the BMU. This may be because each SOM neuron represents a prototype summarising a local region of the feature space, rather than an exact representation of every observation assigned to it. Consequently, natural variation among observations mapped to the same neuron gives rise to a characteristic spread of quantisation errors. Combined with the finite resolution of the SOM, this produces a distribution centred around a non-zero typical error scale rather than one concentrated near zero.

In contrast, the OS distribution is highly skewed, with most values concentrated close to zero and a rapid decay toward higher scores. This follows from its density-based formulation. The majority of observations are located in dense regions of the feature space and therefore receive low scores, while only a small proportion of observations situated in sparse regions are assigned elevated outlier scores. Consequently, QE provides an absolute measure of deviation from the learned representation, whereas OS captures the relative degree of isolation of an observation with respect to the overall data distribution.

Figure 4.5 shows a strong similarity in OS distributions between normal training and test days. This can be explained by the structural properties of the SOM. The SOM defines a discrete and fixed representation of the data space, where each neuron corresponds to a prototype, and all samples are mapped to their BMU. The OS is computed at the level of neurons rather than individual data points, meaning that each sample inherits the OS of its corresponding BMU. Consequently, unseen data is projected onto the existing topology, and the density structure underlying the OS is not recomputed at the sample level but instead preserved from the training. As a result, new data points can only take on OS values already present in the trained SOM, which explains the observed similarity in the distributions. Furthermore, since the neuron-wise OS values are fixed after training, no new variations in the distribution can emerge when evaluating unseen data. The histograms can

only increase or decrease, no new histograms can emerge. This behaviour contrasts with continuous latent representations, where unseen data can alter the underlying density structure and thereby affect the resulting outlier score distribution.

The threshold derived from the training data in the OS pipeline does not identify any anomalous days in the test set. This may indicate that the test data does not contain sufficiently isolated days under the model's definition of anomaly. However, it may also reflect an overly conservative threshold. Since the threshold is based on the training distribution, variability or outliers in the training data can inflate it, reducing sensitivity to moderate anomalies in the test set.

This highlights a limitation of fixed, training-based thresholds. They do not adapt to shifts in score distributions between datasets. As a result, alternative approaches such as adaptive thresholds or validation-based calibration may improve detection sensitivity while maintaining robustness. Overall, these results emphasise the importance of jointly considering score distributions and threshold selection when evaluating anomaly detection performance.

5.3 Autoencoder performance

In this section, the results for the AE will be discussed in more detail.

5.3.1 Dimensionality choice

A dimensionality of eight meant that the model could find similarities within the normal days, days with little to no deviating behaviour, and could generate a proper reconstruction. Days with deviating behaviour, such as spikes in the RMS, VUF, sudden amplitude changes in the THD were problematic for the model to reconstruct and therefore generated a larger RE value. This means that the intrinsic dimension of eight was enough to represent the data without losing too much structure, in this case normal behaviour. There was a little dilemma when choosing the dimensionality of the latent space. A latent dimension 16 was also optimal. However, the model was able to reconstruct more days with anomalous behaviour. A smaller latent dimension caused the model to be worse at reconstructing anomalous days, which was the better in terms of this study for detecting anomalous behaviour in PQ data. However, a lower dimensionality than eight caused the model to overfit and became too sensitive to small deviations in the data. Hence, a latent dimension of eight was deemed appropriate for the study.

5.3.2 Reconstruction error vs outlier score

Figure 4.6 shows some limited agreement between the two anomaly detection methods. The days are fairly spread out in the different regions but the blue region. The days considered highly anomalous by both methods had RE ranged between 12-25 and OS ranged between 0.8-1.0. Generally speaking, RE is a bit more conservative in assigning days with high anomaly scores compared to OS.

The inconsistency between the two methods could be explained by the difference in metric handling. As mentioned in subsection 5.1.1, RE is a distance error based metric, whereas OS detects anomalies based on local sparsity, meaning that points far from local densities receive a high anomaly score. This creates two very different approaches to detecting anomalies.

As can be seen in Figure 4.6, a day that received a high RE score could receive a low OS. The same could happen for days given a high OS score but a low RE score. However, some days with a high OS score were upon observed manually, considered to be of normal behaviour, whereas most of the days with a high RE score were upon observed manually considered to be of anomalous behaviour. This supports what is stated in subsection 5.1.1 concerning the compatibility of distance based methods for PQ data. This implies that RE (in the case for AE) is more sensitive to deviating behaviour from expected patterns/normal behaviour, whereas OS identifies rarely observed or isolated behaviours. This suggest that RE may provide a better suited representation of the domain.

5.3.3 Histogram distribution plots

Figure 4.7-4.8 show a similar behaviour between training and test data. The test set exhibits higher RE values, as expected since the test set is unseen data which may cause a greater error. As mentioned in subsection 5.1.2, the histograms show a right skewed, unimodal distribution, with a non-zero mean for the normal days, while the anomalous days show a more decaying distribution towards larger RE.

The reason for this may be that for AEs, the RE reflects how well the model can compress and reconstruct an input given its learned latent representation. Even for typical observations, the reconstruction is not exact due to several factors such as limited model capacity, regularisation constraints, and the need to represent multiple similar observations using shared latent features. This leads to differences between the input and its reconstruction, analogous to the BMU data points variability observed in SOMs. As a result, RE tend to concentrate around a non-zero characteristic scale, producing a unimodal distribution with moderate skew rather than a sharp peak near zero for normal days.

As for the OS, Figure 4.9 shows that both the normal training and test datasets are strongly concentrated around zero, with the selected threshold located at the upper tail of the normal data distribution. The test data exhibits a more dispersed density, with a larger proportion of days receiving higher OS values, which can be attributed to the fact that these observations are unseen during training and therefore deviate more from the learned data structure.

It is expected that the AE-based approach produces a different distribution compared to the SOM-based method. In the AE OS pipeline, the OS is computed independently for each observation, resulting in a continuous and less constrained

distribution of scores. In contrast, in the SOM-based approach, each observation is mapped to a discrete neuron and inherits that neuron's pre-computed OS. This mapping introduces a quantisation effect, where the resulting OS values are restricted to those associated with the SOM neurons, leading to a more structured and more discrete distribution.

However, an important distinction is that AE variants can, in some cases, achieve very low RE for large portions of the data, particularly if the model has high capacity. In such settings, the distribution may shift closer to zero or even become sharply peaked. Despite this, the fundamental difference remains. RE measures how well the model represents the data, while density-based scores, OS, measure how rare or isolated the data are. This difference may be the primary reason for the contrasting distributional shapes.

In general, the results shows that the AE effectively captures patterns of normal behaviour and assigns elevated anomaly scores to deviations from the normal.

5.4 Variational autoencoder performance

In this section, the results for the VAE will be discussed in more detail.

5.4.1 Reconstruction error vs outlier score

Figure 4.11 shows a stronger agreement between RE and OS, in the sense that days with high RE also tend to exhibit high OS, compared to Figure 4.2 and Figure 4.6. This behaviour can be attributed to the probabilistic formulation of the VAE. During training, the latent representations are regularised to follow a predefined prior distribution (typically Gaussian), while simultaneously minimising reconstruction loss. This results in a structured and continuous latent space with meaningful density properties.

Consequently, samples that are difficult to reconstruct are often mapped to regions of lower latent density. Since the OS is based on local density estimation in the latent space, this leads to a natural alignment between RE and OS in the VAE. This suggests that reconstruction quality and latent density are interrelated in this setting.

In contrast, a standard AE does not impose constraints on the latent distribution, allowing representations to form more freely. While some structure may still emerge, there is no explicit mechanism enforcing a correspondence between latent density and reconstruction quality. As a result, the relationship between RE and OS is generally weaker.

Similarly, for the SOM, the QE is based on the distance to the best matching unit, whereas the OS is derived from density properties of the mapped data. Since these measures rely on different aspects of the data representation, their agreement

is expected to be more limited.

5.4.2 Histogram distribution plots

Figure 4.12-4.13 show similar differences to the distribution plots for the other models: Unimodal distribution with a right skew, and increased mean and errors for the test data, as well as the anomalous days being concentrated around the threshold with decaying density towards higher RE. Possible explanations for this are mentioned in subsection 5.1.2 and subsection 5.3.2. While this behaviour is also observed in the SOM-QE and AE-RE pipelines, the similarity of the distributions are the more similar here, indicating that the VAE-RE pipeline may be the most generalisable model.

A notable difference compared to the other models can be observed in Figure 4.14 and Figure 4.15. For the normal test days, the distribution exhibits a smaller mean and a more rapidly decaying tail toward higher OS values compared to the normal training-day distribution. This differs from the behaviour observed for the SOM and AE, where the training set generally produced smaller anomaly scores.

This behaviour can be attributed to the regularised structure of the VAE latent space. Unlike a standard AE, the VAE imposes a prior distribution through the KLD term, encouraging latent representations to occupy smooth and compact regions of the latent space. Consequently, unseen normal samples may be mapped closer to central latent regions than some training samples, resulting in lower average OS values and fewer high-score samples. This behaviour is also observed for the anomalous training and test days in Figure 4.15, where the anomalous training days observe a more uniform distribution while the anomalous test days are more concentrated toward the threshold. This may explain the difference observed compared to the OS distribution in the regular AE in Figure 4.10.

While the latent-space OS behaviour can largely be explained by the regularised latent structure, RE follows a different mechanism. Reconstruction quality depends on how well the decoder has learned to represent specific input patterns. Since training samples are directly optimised during training, they are generally reconstructed more accurately, including mildly anomalous samples present in the training data. In contrast, unseen anomalies deviate from the learned manifold and therefore tend to produce larger reconstruction errors. Consequently, reconstruction error reflects the degree to which a sample matches patterns observed during training, whereas latent-space outlier scores reflect conformity to the imposed prior distribution.

5.5 Overlap between the models

Table 4.1 shows that the models agree on approximately 19–26% of anomalous days in the RE/QE pipeline, indicating limited overlap in detected anomalies. Rather than identifying a single consistent set of outliers, the models considers different days as anomalous.

The highest agreement is observed between the AE and VAE, which is expected given their shared AE framework. Both models learn compressed representations by minimising RE, leading to similar notions of normality. The agreement between the SOM and AE can likewise be attributed to their reliance on distance-based representations. The AE measures deviation from a learned manifold, while the SOM evaluates distance to the nearest prototype.

In contrast, the lowest agreement occurs between the VAE and SOM, reflecting fundamental differences in methodology. The SOM is a topology-preserving, distance-based model operating on a discrete grid, whereas the VAE models a continuous latent distribution. Consequently, the VAE is sensitive to deviations from the global data distribution, while the SOM emphasises local geometric representation.

Another potential explanation for the discrepancy between the models lies in the choice of threshold. As indicated by the QE and RE distribution plots, there is a high density of observations near the decision boundary. Consequently, samples that lie just above or below the threshold may be classified as anomalous or normal due to minor numerical variations, rather than reflecting a clear and meaningful distinction between the two classes. This sharp boundary presents a risk of false positives and negatives emerging, affecting the overlap between the models.

Table 4.2 reveals substantially lower agreement in the OS pipeline, with overlap between the SOM and the AE variants dropping to 4–6%, compared to around 20% in the RE/QE case. The AE and VAE, however, maintain a higher agreement of approximately 22%. This difference may arise because the OS pipeline introduces an additional dependency on representation. Anomaly detection is performed via a shared density-based method applied to each model's embedding. As a result, differences in representation directly translate into differences in detected outliers.

The AE and VAE produce continuous latent spaces with similar geometric and density properties, leading to comparable clustering structure and OS. In contrast, the SOM's discrete prototype representation introduces effects that alter local neighbourhood structure, resulting in density patterns that differ significantly from those of the AE variants.

Overall, the low agreement between models reflects the inherently model-dependent nature of anomaly detection. Each method encodes a distinct notion of normality, and the results highlight the importance of aligning model choice with the structure of the data and the intended application. Combining multiple approaches may therefore provide a more robust and comprehensive identification of anomalous behaviour.

5.6 Method discussion

Here a discussion concerning choices in the method will be conducted.

5.6.1 Threshold choice

The threshold was set to be the 99th percentile based on previous research in the field, as well as a good balance between finding anomalies and reduce the number of false positives. It is possible that the two anomaly detection methods should have different percentile thresholds, or that the threshold should be adaptable. One example of this could be calculating a mean of the anomaly metric, and define an anomaly as a certain number of standard deviations away from the mean. However, we claim that the important finding here is that the idea of setting a threshold seems to work for the purpose of this thesis. How to optimise this threshold could be a subject for further research.

5.6.2 Statistical representation of the data

The use of statistical features to represent daily PQ data provides a compact and computationally efficient alternative to processing raw time series. By summarising each day using measures of central tendency, variability, and range, the approach captures key characteristics relevant for many types of anomalies, while reducing the dimensionality of the problem. This is particularly beneficial for ML models that may otherwise struggle with high-dimensional inputs. Moreover, it is also a well established data processing method for ML tasks.

However, this representation also introduces important limitations. In particular, the transformation from time series to summary statistics results in a loss of temporal information. As a consequence, short-duration events, transient disturbances, and temporal patterns within a day may not be adequately captured. Two days with significantly different temporal behaviours may therefore appear similar when represented through statistical features alone.

While the selected features (mean, standard deviation, extrema, and robust range) capture several important aspects of PQ behaviour, they do not provide a complete description of the data. Additional features, such as higher-order statistical moments or frequency-domain characteristics, could further enhance the representation. Alternatively, the use of raw time series data would enable the detection of temporal anomalies, but at the cost of increased model complexity and possibly reduced interpretability.

Overall, the chosen representation reflects a trade-off between computational efficiency, robustness, and the level of detail captured. The results of this study should therefore be interpreted in the context of this abstraction, as the anomaly detection is limited to deviations that are observable at the level of daily statistical summaries.

5.6.3 Normalisation method

The normalisation method used was per-meter normalisation. This choice was made to remove systematic, meter-specific characteristics such as reference voltages, cali-

bration offsets, or sensor drift. As a result, the models are discouraged from learning meter-identity cues and instead focus on detecting anomalous behaviour that is independent of the specific meter.

A limitation of this approach is that per-meter normalisation relies on statistics (mean and standard deviation) computed from historical training data for each individual meter and each parameter. Consequently, the pipeline depends on historical data not only for model training but also for input preprocessing. This creates a practical challenge when deploying the system on previously unseen meters, for which no prior statistics are available to perform normalisation.

Additionally, care must be taken to avoid data leakage. Data leakage can occur when information other than the training data leaks into the training of the models. One such risk is when normalising test data on parameters derived from the test set. While computing normalisation statistics from training data alone is valid, using statistics derived from test data or future observations would introduce information from outside the training distribution. This makes the models prone to optimistic theoretical results that could not agree with real world applications.

This limitation motivates the consideration of alternative normalisation strategies, such as global normalisation, or other robust statistics that do not require meter-specific historical data. Another alternative would be to gather data during a calibration period of specific length, and then normalise based on the statistics of this calibration period. This would assume that the data during the calibration period is well-behaved.

5.7 Societal and ethical impacts

Integration of AI for analysing PQ data could have several societal and ethical impacts. First, minimising manual labour of the actual analysing part. As humans we are not as efficient at analysing large quantities of data, compared to computationally powered products. This would reduce the analysis time enormously and thus aid the users of the AI. One problem that could arise from the aid of AI is partial deskilling of the engineers who could turn to only start to relay on AI. However, the reduction in manual labour also comes with a possible reduction in jobs. As AI would do the same job as a couple of workers in a fraction of the time, in addition to doing multiple jobs simultaneously, this could possibly remove working positions for people. Maybe not in an early phase, but later when implemented more vividly.

Second, integration of AI could reduce the response time for problem discovered to problem solved. As mentioned above, AI is capable of analysing large amounts of data quickly and would therefore be able to in an early phase discover a potential problem and send an alert. An early discovery of a problem could save large amount of money as losses would be mitigated, a much desired aspect corporately. Additionally, increased efficiency in response time would improve the reliability of the electrical grid.

Third, an increased integration of AI would increase the dependence of automated decision making. As mentioned above, this would improve effective response time and reliability, but also reduce the backend simplicity of the system. Automated decision making could give rise to questions such as; Who bares responsibility in case of an error? Who is ethically responsible? to name a few.

Finally, as an integration of AI for monitoring PQ data could allow for easier implementation of renewable energy sources, this would have a positive impact on the environment. This could further reduce the need of fossil fuels and thus mitigate the carbon dioxide emissions. Example, AI could make the electrical grid more robust when more inverters are introduced, electrical vehicle chargers.

5.8 Impact of the limitations on the results

The results of this study must be interpreted in light of several limitations related to the data, methodology, and evaluation framework, all of which influence both the reliability and generalisability of the findings.

First, the dataset is derived from PQ meters located in only two geographical regions, Landskrona and Ulricehamn. While this provides some variability in operating conditions, it nevertheless constrains the diversity of the data. PQ characteristics are influenced by factors such as grid topology, load composition, and environmental conditions, which may differ significantly across regions. Consequently, the models developed in this study may capture location-specific patterns rather than universally applicable characteristics, limiting their transferability to other networks or geographical locations.

Second, the use of a fixed one-day time window may influence the types of anomalies detected. While a daily aggregation provides a manageable and interpretable unit of analysis, it may obscure hourly variations that are relevant for certain PQ disturbances. Conversely, longer-term trends or gradual degradation processes may not be fully captured within isolated daily segments. As a result, the findings primarily reflect deviations at the daily level and may not generalise to anomaly detection at finer or coarser resolutions.

A more fundamental limitation is the absence of ground truth labels. Without verified annotations of true PQ events, it is not possible to directly assess the accuracy or validity of the detected anomalies. This restricts the evaluation to indirect measures, such as distributional analysis and qualitative inspection, which can indicate consistency but not correctness. Furthermore, the lack of labelled data makes it difficult to determine whether discrepancies between models reflect differences in sensitivity or merely the identification of false positives.

In addition, the anomaly detection approach employed in this study is inherently descriptive rather than diagnostic. The models are designed to identify deviations

from learned patterns, but they do not provide explanations for why such deviations occur. This limits the practical utility of the results in operational settings, where understanding the root cause of anomalies is often essential for decision-making. Consequently, the detected anomalies should be interpreted as indicators of unusual behaviour rather than confirmed events, and further domain-specific analysis is required to establish relevance.

Taken together, these limitations imply that the results should be viewed as exploratory rather than definitive. The identified anomalies and model comparisons provide insights into the behaviour of different detection approaches and highlight their relative sensitivities, but they do not constitute a validated detection system. Future work should aim to incorporate labelled data, extend the analysis to a broader range of locations and conditions, and integrate interpretability or diagnostic tools to better link detected anomalies to underlying physical causes. Such developments would strengthen both the interpretability and practical applicability of the proposed methods.

5.9 Future work

There are a few different topics that are relevant in future work of the study. First, optimising the threshold for detection of anomalous days. For the study, the 99th percentile was chosen and deemed appropriate as it captures days with very deviating behaviour. The top 1 % with the highest RE/OS captures large anomalies and some anomalies that seem to have an increasing anomalous behaviour, such as increase in RMS, VUF or a sudden amplitude change in THD. However, this threshold might serve a better purpose with a smaller or bigger percentile, capturing different needs and sensitivity. Additionally, different zones for different percentiles could be added and be used as an indicator of when a deviating behaviour is emerging.

Second, filtering out anomalous days, remove them from the dataset and retrain the model on the normal days. This would make the model more sensitive towards anomalies and could make the models even better at detecting anomalous trends at an early stage. However, one could also use found anomalous days from the current model, label them, extract the anomalous days and use the labelled anomalous data to train supervised models. This would allow for classification of not only that the day is anomalous, but also what type of anomaly that might be occurring. This would allow users to not only categorise vasts amount of data and remove unnecessary manual labour, but possibly be able to predict when a problem would occur, according to some statistical suggestion of what the emerging problem could be. This would allow the users to work more effectively and proactively.

Third, the number of parameters and dependence of parameters. For the scope of the project, to investigate whether an AI could be used to cluster/detect anomalies in PQ data, RMS, VUF and THD was chosen. These parameters are some of the fundamental parameters used to monitor PQ in the electrical grid. However, one could include more parameters that are used when motoring PQ. This could allow

the models to be more precise and acquire better results in finding more anomalous days and trends. Additionally, choosing specific parameters normally observed in anomalous behaviour, and use these parameters to train an AI to detect when these anomalous behaviours are emerging or are occurring. One example could be the detection of earth fault, by using parts of the VUF. Additionally, one could repeat the study using other parameters, i.e current, to see if the model would find the same anomalous days as for the current parameters.

Fourth, since the test set was not that large, only 7 % of the training data, this could lead to some inaccuracies when validating the performance of the models. But since there do not exist any ground truth labels for the data and the models use unsupervised learning to interpret the data, this does not constitute a particularly significant issue. From what can be seen in the QE/RE density Figures 4.4, 4.8, 4.13, the models do learn to interpret unseen data well even though the test data set was not very large.

Fifthly, as briefly discussed in subsection 5.6.3, a key challenge concerns the integration of newly introduced meters that lack historical data. Since the normalisation of input data relies on statistics derived from the training data of existing meters, it becomes necessary to develop a more generalised normalisation framework prior to model deployment.

One potential solution is to group meters into predefined categories, such as reference voltage, and apply a category-specific normalisation scheme. In this approach, all meters within a given category would share a common normalisation strategy, thereby enabling immediate inclusion of new meters. However, this method may introduce new limitations. Specifically, when considerable variation exists among meters within the same category, stemming from geographical or technical factors, such as ageing infrastructure or extreme environmental conditions. This presents a risk that smaller fluctuations in the data may be obscured, thereby potentially reducing the sensitivity of the models to detect anomalies.

An alternative approach involves implementing a dedicated data collection phase for each new meter. This would enable the derivation of meter-specific normalisation parameters, thereby preserving sensitivity to fine-grained variations in individual data streams. Nevertheless, such a strategy entails a delay in the operational deployment of the models, as sufficient data must first be collected. Additionally, it would necessitate retraining the model whenever a new meter is introduced. Despite these drawbacks, this approach offers the advantage of improved detection performance by tailoring the normalisation process to each individual meter.

6
Conclusion

In this chapter, a conclusion of the results will be drawn and how well the results answered the research questions.

The models were all able to distinguish anomalies from normal data. However, they differed somewhat in which days they identified as anomalous and in how strongly anomalous those days were considered. The models also showed consistency between training and test distributions, particularly for the QE and RE pipelines, suggesting that they have learned stable and generalisable representations of the data.

Individually, SOM is fast and visually interpretable, grouping days into clusters through its mapping onto neurons. However, it enforces a discrete projection into a two-dimensional space, which may lead to information loss. AEs, in contrast, are better suited for capturing complex relationships in high-dimensional data and offer greater flexibility. A drawback is that their latent representations can be difficult to interpret.

The VAE-RE pipeline was the most effective at identifying deviations. This may be due to the VAE's constraint of the latent space to a smooth distribution, which makes it more difficult to accurately represent anomalous days. As a result, these anomalies tend to produce higher REs, which is desirable for anomaly detection.

Bibliography

[1] M. H. J. Bollen, "What is power quality?" *Electric Power Systems Research*, vol. 66, no. 1, pp. 5–14, Jul. 2003, doi: https://doi.org/10.1016/S0378-7796(03)00067-1.

[2] Z. Olczykowski and Z. Łukasik, "Evaluation of flicker of light generated by arc furnaces," *Energies*, vol. 14, no. 3901, pp. 1–23, June. 2021, doi: https://doi.org/10.3390/en14133901.

[3] *Energimarknadsinspektionens föreskrifter och allmänna råd om krav som ska vara uppfyllda för att överföringen av el ska vara av god kvalitet*, EIFS 2023:3, Energimarknadsinspektionen, Eskilstuna, Sweden, 2023. Available: https://ei.se/download/18.214b931418a26108737f2f0/1694668077847/EIFS-2023-3-om-krav-som-ska-vara-uppfyllda-f

[4] *Testing and Measurement Techniques - Power Quality Measurement Methods*, IEC 61000-4-30, International Electrotechnical Commission, Geneva, Switzerland, 2015, Edition 3. Available: https://cdn.standards.iteh.ai/samples/20084/2fec037f20904f299e95ceceeda936dd/IEC-61000-4-30-2015.pdf.

[5] *Voltage characteristics of electricity supplied by public distribution systems*, EN 50160, European Committee for Electrotechnical Standardization (CENELEC), Brussels, Belgium, 2010, Edition 3. Available: https://elstandard.se/standard/3009201.

[6] A. Laouafi, "Towards efficient solutions for automatic recognition of complex power quality disturbances," *Expert Systems with Applications*, vol. 286, p. 128077, Aug. 2025, doi: https://doi.org/10.1016/j.eswa.2025.128077.

[7] M. Srivastava, S. K. Goyal, A. Saraswat, R. S. Shekhawat, and G. Gangil, "A review on power quality problems, causes and mitigation techniques," in *Proceedings of the 2022 1st International Conference on Sustainable Technology for Power and Energy Systems (STPES)*, Jaipur, India, 2022, pp. 1–6, [Online]. Available: https://doi.org/10.1109/STPES54845.2022.10006587, Accessed on: 2026-03-25.

[8] R. C. Dugan, S. Santoso, M. F. McGranaghan, and H. W. Beaty, *Electrical Power Systems Quality*, 2nd ed. New York, NY, USA: McGraw-Hill Professional, 2002.

[9] P. H. Gadde and S. Brahma, "Unbalance compensation using single-phase inverters in inverter-dominated microgrids," in *2022 IEEE Industry Applications Society Annual Meeting (IAS)*, pp. 1–7, [Online]. Available: https://doi.org/10.1109/IAS54023.2022.9939799 Accessed on: 2026-04-13.

[10] M. Senol, I. S. Bayram, L. Hunter, K. Sevdari, C. McGarry, D. C. Gaona, O. Gehrke, and S. Galloway, "Harmonics measurement, analysis, and impact

assessment of electric vehicle smart charging," *IEEE Open Journal of Vehicular Technology*, vol. 6, pp. 109–127, Jan. 2025, doi: https://doi.org/10.1109/OJVT.2024.3505778.

[11] J. E. Caicedo, D. Agudelo-Martínez, E. Rivas-Trujillo, and J. Meyer, "A systematic review of real-time detection and classification of power quality disturbances," *Protection and Control of Modern Power Systems*, vol. 8, no. 3, Jan. 2023, doi: https://doi.org/10.1186/s41601-023-00277-y.

[12] G. Devadasu and M. Sushama, "Identification of voltage quality problems under different types of sag/swell faults with fast fourier transform analysis," in *Proceedings of the 2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, Karnataka, India, 2016, pp. 464–469, [Online]. Available: https://doi.org/10.1109/AEEICB.2016.7538332, Accessed on: 2026-03-25.

[13] A. Hildén, P. Pakonen, and P. Koponen, "Power quality monitoring data management and analysis for distribution networks," in *Proceedings of the CIRED 2021 Conference on Electricity Distribution*, Geneva, Switzerland, 2021, pp. 1–5, [Online]. Available: https://doi.org/10.1049/icp.2021.1833, Accessed on: 2026-03-25.

[14] M. He *et al.*, "Power quality mitigation in modern distribution grids: A comprehensive review of emerging technologies and future pathways," *Processes*, vol. 13, no. 8, Aug. 2025, doi: https://doi.org/10.3390/pr13082615.

[15] J. Kaur and S. K. Bath, "Harmonic distortion in power systems due to electronic control and renewable energy integration: a comprehensive review," *Discover Electronics*, vol. 2, no. 1, pp. 1–16, Aug. 2025, doi: https://doi.org/10.1007/s44291-025-00111-9.

[16] A. Furlani Bastos and S. Santoso, "Optimization techniques for mining power quality data and processing unbalanced datasets in machine learning applications," *Energies*, vol. 14, no. 2, Jan. 2021, doi: https://doi.org/10.3390/en14020463.

[17] D. K. Nishad, A. N. Tiwari, S. Khalid, S. Gupta, and A. Shukla, "Ai-based hybrid power quality control system for electrical railway using single phase pv-upqc with lyapunov optimization," *Scientific reports*, vol. 15, no. 1, Jan. 2025, doi: https://doi.org/10.1038/s41598-025-85393-5.

[18] S. Jain, A. Satsangi, R. Kumar, D. Panwar, and M. Amir, "Intelligent assessment of power quality disturbances: A comprehensive review on machine learning and deep learning solutions," *Computers and Electrical Engineering*, vol. 123, pp. 1–31, Apr. 2025, doi: https://doi.org/10.1016/j.compeleceng.2025.110275.

[19] I. S. Samanta, S. Panda, P. K. Rout, M. Bajaj, M. Piecha, V. Blazek, and L. Prokop, "A comprehensive review of deep-learning applications to power quality analysis," *Energies*, vol. 16, no. 11, May 2023, doi: https://doi.org/10.3390/en16114406.

[20] Unipower AB, "About unipower – your power quality partner," 2026, [Online]. Available: https://www.unipower.se/about-unipower/ (accessed on: 2026-03-01).

[21] "Ieee recommended practice for monitoring electric power quality," *IEEE Std 1159-2019 (Revision of IEEE Std 1159-2009)*, pp. 1–98, Aug. 2019, doi: https://doi.org/10.1109/IEEESTD.2019.8796486.

[22] W. Liu *et al.*, "Power quality assessment in shipboard microgrids under unbalanced and harmonic ac bus voltage," *IEEE Transactions on Industry Applications*, vol. 55, no. 1, pp. 765–775, Aug. 2019, doi: https://doi.org/10.1109/TIA.2018.2867330.

[23] A. Ostrowska *et al.*, "Power quality assessment in a real microgrid-statistical assessment of different long-term working conditions," *Energies*, vol. 15, no. 21, Oct. 2022, doi: https://doi.org/10.3390/en15218089.

[24] J. L. Kirtley Jr., "Introduction to symmetrical components," MIT OpenCourseWare, 6.061 Introduction to Electric Power Systems, 2011, [Online]. Available: https://ocw.mit.edu/courses/6-061-introduction-to-electric-power-systems-spring-2011/7603169f5d8514b769a46fa900b79a57_MIT6_061S11_ch4.pdf.

[25] P. Pillay and M. Manyage, "Definitions of voltage unbalance," *IEEE Power Engineering Review*, vol. 21, no. 5, pp. 50–51, May 2001, doi: https://doi.org/10.1109/39.920965.

[26] Britannica Authors, "Root-mean-square voltage," in *Encyclopedia Britannica*. Chicago, IL, USA: Encyclopedia Britannica Inc, 2025, [Online]. Available: https://www.britannica.com/technology/root-mean-square-voltage Accessed on: 2026-03-02.

[27] "Ieee standard for harmonic control in electric power systems," *IEEE Std 519-2022 (Revision of IEEE Std 519-2014)*, pp. 1–31, 2022, doi: https://doi.org/10.1109/IEEESTD.2022.9848440.

[28] R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander, "Hierarchical density estimates for data clustering, visualization, and outlier detection," *ACM Trans. Knowl. Discov. Data*, vol. 10, no. 1, Jul. 2015, doi: https://doi.org/10.1145/2733381.

[29] K. Ghosh, M. C. Naldi, J. Sander, and E. Choo, "Unsupervised parameter-free outlier detection using hdbscan* outlier profiles," in *2024 IEEE International Conference on Big Data (BigData)*, Washington, DC, USA, 2024, pp. 7021–7030, [Online]. Availible: https://doi.org/10.1109/BigData62323.2024.10825917, Accessed on: 2026-03-19.

[30] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, Jan. 1982, doi: https://doi.org/10.1007/BF00337288.

[31] ——, *Self-Organizing Maps*, 3rd ed. Heidelberg, Germany: Springer, 2001.

[32] Y. Sun, "On quantization error of self-organizing map network," *Neurocomputing*, vol. 34, no. 1, pp. 169–193, Sep. 2000, doi: https://doi.org/10.1016/S0925-2312(00)00292-7.

[33] E. de Bodt, M. Cottrell, P. Letremy, and M. Verleysen, "On the use of self-organizing maps to accelerate vector quantization," *Neurocomputing*, vol. 56, p. 187–203, Jan. 2004, doi: https://doi.org/10.1016/j.neucom.2003.09.009.

[34] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[35] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, p. 504–507, 2006. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1127647

[36] I. D. Mienye and T. G. Swart, "Deep autoencoder neural networks: A comprehensive review and new perspectives," *Archives of Computational Methods in Engineering*, vol. 32, pp. 3981–4000, Oct. 2025, doi: https://doi.org/10.1007/s11831-025-10260-5.

[37] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, Banff, Canada, 2014, pp. 1–12, [Online]. Available: https://doi.org/10.48550/arXiv.1312.6114, Accessed on: 2026-03-15.

[38] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–86, Mar. 1951.

[39] R. Yao, C. Liu, L. Zhang, and P. Peng, "Unsupervised anomaly detection using variational auto-encoder based feature extraction," in *2019 IEEE International Conference on Prognostics and Health Management (ICPHM)*, San Francisco, CA, USA, 2019, pp. 1–7, [Online]. Available: https://doi.org/10.1109/ICPHM.2019.8819434, Accessed on: 2026-02-18.

[40] S. Fletcher and M. Islam, "Comparing sets of patterns with the jaccard index," *Australasian Journal of Information Systems*, vol. 22, pp. 1–17, Mar. 2018, doi: https://doi.org/10.3127/ajis.v22i0.1538.

[41] N. C. Chung, B. Miasojedow, M. Startek, and A. Gambin, "Jaccard/tanimoto similarity test and estimation methods for biological presence-absence data," *BMC Bioinformatics*, vol. 20, no. 23, pp. 1–12, Dec. 2019, doi: https://doi.org/10.1186/s12859-019-3118-5.

[42] T. Hastie, R. Tibshirani, and J. Friedman, "Kernel smoothing methods," in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009, pp. 208–210. [Online]. Available: https://doi.org/10.1007/978-0-387-84858-7

[43] I. S. Samanta *et al.*, "A comprehensive review of feature extraction techniques for power quality event detection with novel approaches for enhanced classification in smart grids," *Electrical Engineering*, vol. 107, pp. 12 227–12 259, May 2025, doi: https://doi.org/10.1007/s00202-025-03149-w.

[44] D. H. Chiam, K. H. Lim, and K. H. Law, "Lstm power quality disturbance classification with wavelets and attention mechanism," *Electrical Engineering*, vol. 105, pp. 259–266, Feb. 2023, doi: https://doi.org/10.1007/s00202-022-01667-5.

[45] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, Nov. 2011, [Online], Available: http://jmlr.org/papers/v12/pedregosa11a.html.

[46] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, Jun. 2007, doi: https://doi.org/10.1109/MCSE.2007.55.

[47] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," pp. 1–12, Dec. 2019, doi: https://doi.org/10.48550/arXiv.1912.01703.

[48] G. Vettigli, "Minisom: minimalistic and numpy-based implementation of the self organizing map," 2018. [Online]. Available: https://github.com/JustGlowing/minisom/

[49] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *The Journal of Open Source Software*, vol. 2, no. 11, mar 2017, doi: https://doi.org/10.21105%2Fjoss.00205.

[50] The pandas development team, "pandas-dev/pandas: Pandas," Feb. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3509134

[51] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: https://doi.org/10.1038/s41586-020-2649-2.

[52] M. U. Ndubuaku, A. Anjum, and A. Liotta, "Unsupervised anomaly thresholding from reconstruction errors," in *Proceedings of the European Conference of the Prognostics and Health Management Society*, 2019, pp. 123—-129, [Online]. Available: https://doi.org/10.1007/978-3-030-34914-1_12, Accessed on: 2026-04-02.

[53] B. Eiteneuer, N. Hranisavljevic, and O. Niggemann, "Dimensionality reduction and anomaly detection for cpps data using autoencoder," in *2019 IEEE International Conference on Industrial Technology (ICIT)*, Melbourne, VIC, Australia, 2019, pp. 1286–1292, [Online]. Available: https://doi.org/10.1109/ICIT.2019.8755116, Accessed on: 2026-04-03.

[54] T. K. Spohn *et al.*, "A machine learning approach using autoencoders to perform quality control on meteorological data," *Environmental Data Science*, vol. 5, pp. 1–18, Jan. 2026, doi: https://doi.org/10.1017/eds.2026.10030.

DEPARTMENT OF ELECTRICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY