

Article

Gear Fault Detection Method Based on Convex Hull Clustering of Autoencoder's Latent Space

Michał Batsch ^{1,*}  and Bartłomiej Kiczek ²¹ Department of Mechanical Engineering, Faculty of Mechanical Engineering and Aeronautics, Rzeszów University of Technology, Al. Powstańców Warszawy 8, 35-959 Rzeszów, Poland² Data Center of Excellence, Bunge, 1391 Timberlake Manor Parkway, Chesterfield, MO 63017, USA; bartlomiej.kiczek@gmail.com

* Correspondence: mbatsch@prz.edu.pl

Abstract: This paper presents a method of pitting failure detection in toothed gears based on the reconstruction of the gear case vibrational signal. The effectiveness of the proposed method was tested in an experiment on a power circulation test stand. The autoencoder deep neural network architecture, semi-supervised training, and validation, along with the latent data convex hull-based clustering, are presented. The proposed method offers high efficiency (0.99 F1-measure) in gear state prediction (100% in failure detection, 98.9% in normal state prediction) and provides more capabilities in terms of generalization in comparison with linear machine learning techniques such as principal component analysis and nonlinear like the generative adversarial network. Moreover, it is distinguished by high sensitivity while also being able to detect even slight surface damage (initial pitting). These findings will be of particular relevance to a range of scientists and practitioners working with gear drives who are willing to implement machine learning in signal processing and diagnosis.

Keywords: gear fault diagnosis; semi-supervised deep learning; autoencoder; generative adversarial network; vibrational signal; anomaly detection



Citation: Batsch, M.; Kiczek, B. Gear Fault Detection Method Based on Convex Hull Clustering of Autoencoder's Latent Space. *Appl. Sci.* **2024**, *14*, 5282. <https://doi.org/10.3390/app14125282>

Academic Editors: Stefano Marchesiello and Dario Anastasio

Received: 21 May 2024
Revised: 11 June 2024
Accepted: 17 June 2024
Published: 18 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Neural networks constitute one of the most powerful tools in the domain of machine learning that are suitable for solving multidimensional complex problems. These are naturally tasks such as classification and prediction, but also more sophisticated generative applications. In the context of engineering, they can be used in various situations for image and pattern recognition [1,2], robot control [3], tracking systems [4], smart buildings [5], signal processing [6], economic analyses [7], fake news detection [8], and many more. One of the key applications of deep neural networks in mechanical engineering is to predict the failure of machine components in real-time based on the measurement of its vibrational signal. This approach states the alternative to conventional signal filtering and analysis in the time or frequency domain [9,10]. Failure should be detected at the earliest stage possible to prevent damage to the whole mechanical system, as in the case of helicopter transmissions [11].

Usually, the neural network is trained with supervision to classify the damage of the components. In [6], the authors describe a deep neural network for planetary gear fault predictions based on two-channel vibration measurement. The network was trained with signal features extracted from the data of mechanically-induced teeth failure. The major disadvantage of this method is the use of the supervised learning paradigm itself. However, it was proven to have good efficiency in failure recognition. Another application for fault detection is the use of reinforcement learning in bearing state prediction [12]. The authors proposed a novel network architecture called a reinforcement learning unit, matching the recurrent neural network. Supervised learning can be difficult to perform since it

requires data from the worn mechanical component and proper signal feature selection, which provides information about the relevant faults. In the case of high-power geared transmissions, extensive fatigue testing is required to obtain reliable data for training. Another aspect is that such neural networks are trained only to predict predefined failures such as pitting or tooth fracture and may not cover all possible failures occurring during real operation. To overcome the above-mentioned disadvantages, a specific neural network architecture known as an autoencoder may be used [13]. An autoencoder is a type of deep neural network that can reconstruct its input. Usually, they are used for feature extraction [14–17], however, they can also work as anomaly detectors [11,14,18]. For this purpose, the autoencoder neural network is trained with the aid of unlabeled data in which there are no or very few abnormal states. These kinds of data are usually easy to collect because they refer to the normal operation of the system. As a measure of fault (abnormal state), the training loss function is selected. If the autoencoder fails to reconstruct the input with sufficient accuracy (obtained while training), it means that an abnormal state (failure) has occurred.

Another technique that can be used for unsupervised anomaly detection is the generative adversarial network (GAN), widely used for image generation [19]. This type of network is trained to generate data, making use of the game between the generator and discriminator. If the network is trained on data for a normal state, the probability output space of the discriminator calculated for abnormal data is then used to predict an abnormal state. GANs have been used for anomaly detection in time series data [20], image and tabular data [21], and fault detection in unbalanced datasets [22,23].

Several papers related to the predictive maintenance of gears and generally of rotating machinery based on its vibrations have adopted a semi-supervised or unsupervised learning approach. The main reason for using such techniques is the problem of heavily imbalanced datasets occurring in industrial practice. Usually, there are no extensive test results to construct a proper dataset with a uniform distribution of each feature, but there are a lot of data related to the proper operation of the mechanical system. For such imbalanced datasets, unsupervised or semi-supervised learning approaches are suitable. In [14], the authors presented an unsupervised approach based on an autoencoder to monitor and detect anomalies in industrial machines. The authors compared their models with supervised machine learning ones and achieved 91% accuracy, 98% precision, and 83% recall. Planetary gear trains were also within the interests of the researchers [15], who proposed the supervised method based on a stacked autoencoder. Model training and validation were performed on a dataset with artificially induced gear teeth failures. The recognition rates varied depending on gear state from 95% to 97%. The authors of [16] used stacked autoencoders with a relatively simple architecture (one hidden layer) in order to extract the features of faulty planetary gear vibrational signals. In [11], the complex expert system was proposed to monitor the condition of helicopter transmissions. The dataset consisted of measurements from 23 accelerometers mounted in various places within the helicopter driveline, which provided 80 time signals. The authors used a convolutional autoencoder and unsupervised classifier that combined K-nearest neighbors, isolation forest, angle-based detection, and the local outlier factor. The anomaly detection was based on averaging the outputs from these two models. The method was highly efficient, giving 100% of the true positive rate and 0.03% of the false positive rate, but was limited in terms of generalization on the type of anomaly. In [18], the authors investigated spur gear train vibrations collected for different artificially-induced teeth failure types. They used sparse stacked autoencoders for feature extraction and supervised learning, which resulted in 97% accuracy. Researchers have also used GANs as anomaly detectors [23] where the bearing and gear teeth were under investigation. To detect an abnormal state, the anomaly score was used on the discriminator output, which resulted in accuracy from 95% to 98% and an F1-score from 97% to 99%, depending on the test case. The use of GANs has also been proposed for general time series anomaly detection [20]. The authors compared three models: cumulative sum chart, MAD-GAN (multivariate anomaly detection with genera-

tive adversarial networks) and TAnoGAN (time series anomaly detection with generative adversarial networks). Depending on the model, the authors achieved a 90% accuracy, 72% recall, and 61% precision.

It may be seen, based on the literature review presented, that there are many supervised methods that might not be suitable for imbalanced datasets. Moreover, many models have been trained and validated on artificially-induced teeth failures, and are rarely focused on the pitting type of failure. Taking the above into account, the novelty of the work as well as its aim is to:

- Construct a dataset as a result of extensive fatigue tests of gears with real gear pitting failure of different severity rates;
- Propose a robust semi- or unsupervised model for failure detection;
- Ensure the high-performance metrics of the model;
- Construct a model that gives a general perspective on gear pitting wear prediction.

In this paper, we explore the experimental data from a helical gear setup using semi-supervised machine learning techniques. In Section 2.1, we describe the setup and data collection, and in Section 2.2, we discuss the data preprocessing for the algorithms. In Section 2.3, we commence with a simple analytic solution, based on principal component analysis (PCA). Finally, in Section 3, we present the three methods for pitting failure detection in gearboxes, namely, the autoencoder as a basic anomaly detector (Section 3.1), autoencoder and convex hull-based classification in its latent layer (Section 3.2), and a generative adversarial network for anomaly detection (Section 3.3). All models based on neural networks used in this study were subjected to neural ablation tests in order to choose a model architecture with the maximum performance and reasonable learning time. A representative example of the ablation test is given in Appendix A. We discuss and conclude the results in Sections 4 and 5, respectively. Based on the performed calculations, we may state that:

- The proposed methods based on autoencoders were efficient and could detect even the initial state of pitting formation, which may be difficult with the aid of signal analysis in the time and frequency domain;
- The best model (AE+CH) showed a very high effectiveness (100%) in failure detection (true positive) and 98.9% in normal state prediction (true negative), which resulted in a very high F1-measure (0.99);
- The latent space analysis revealed a generalized perspective on the gear wear—the measurements drifted in a specific direction in the latent space with the progress of gearbox damage;
- The autoencoder outperformed the generative adversarial network in terms of generalization on wear prediction.

2. Materials and Methods

2.1. Experiment Setup

The experiment was conducted on a power circulation test stand for cylindrical gear fatigue testing (Figure 1).

The test stand consisted of two gear pairs—the tested (item no. 1) and the stand (item no. 2). These were connected to each other using shafts (items no. 3 and 4). The torque was provided by preload applied by the load clutch (item no. 5). The electric motor (item no. 6), with the aid of belt transmission (item no. 7), covered only the power losses and provided rotational speed. The stand gear pair was designed in such a way as to have a much greater service life than the tested gears to maintain durability over all tests. A piezoelectric accelerometer (item no. 8) mounted on the case of the tested gear was used for the vibration measurements. The tested gear pairs (Figure 2) were the helical gears, and the data are listed in Table 1. The gear teeth were manufactured with a worm-shaped tool on a Koepfer EMAG 200 CNC (Koepfer gruppe, Furtwangen, Germany) gear hobbing machine from 42CrMo4 steel. The teeth surface profile was measured on a Mahr MarSurf GD 120 measuring machine (Mahr Federal Inc., Providence, RI, USA). Measured roughness

parameters were $Ra0.4$ for pinion and $Ra0.6$ for gear. Gears were classified into the seventh accuracy grade based on measurements using the Klingelnberg P40 coordinate measuring machine.

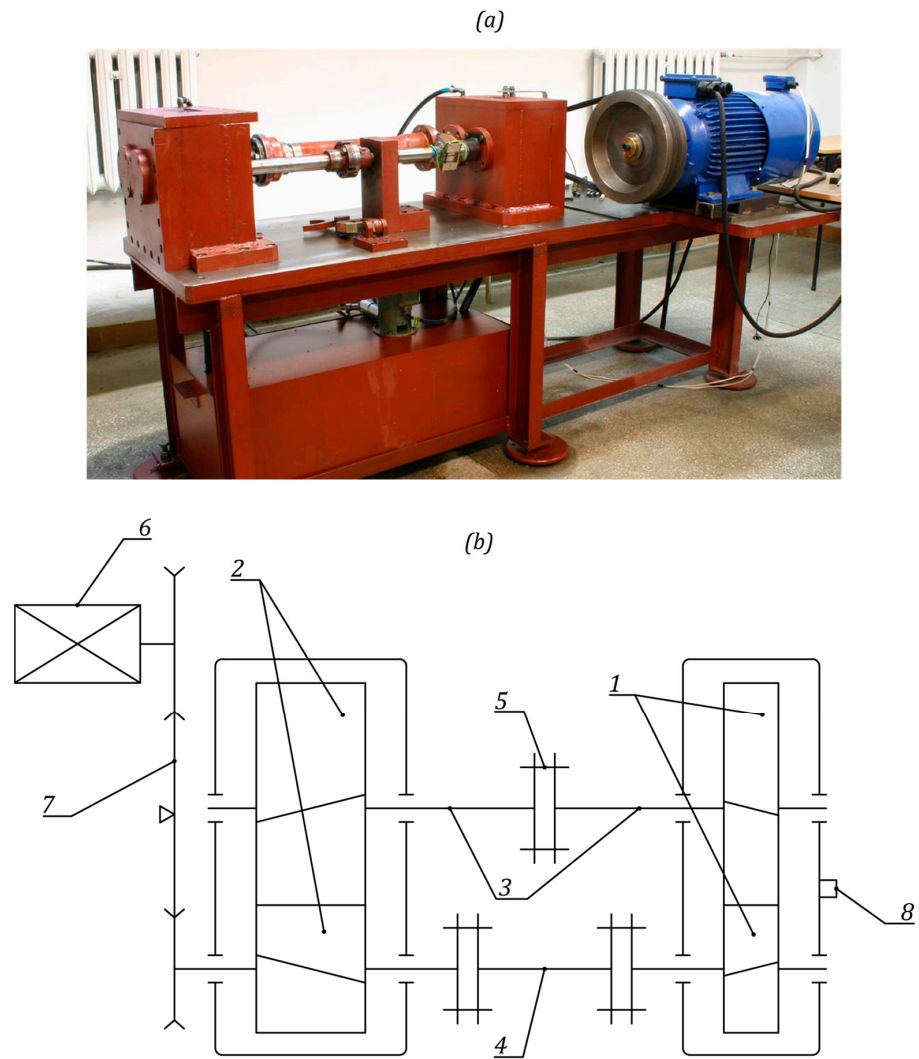


Figure 1. Power circulation test rig: (a) view; (b) scheme.

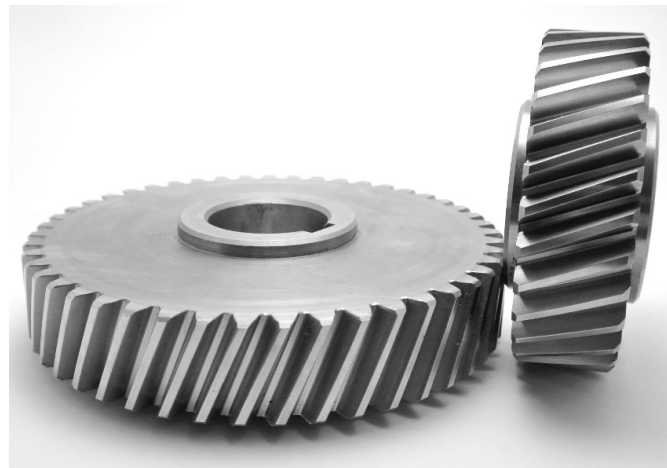


Figure 2. Gear pair for testing.

Table 1. Data of the tested gear pairs.

Parameter	Pinion		Gear
Geometry			
Normal module, mm		$m_n = 3$	
Number of teeth	$z_1 = 30$		$z_2 = 47$
Face width, mm		$b = 30$	
Helix angle, °		$\beta = 22.482$	
Normal pressure angle, °		$\alpha_n = 20$	
Profile shift		$x = 0$	
Axes distance, mm		$a = 125$	
Pitch diameter, mm	$d_1 = 97.40$		$d_2 = 152.59$
Tip diameter, mm	$da_1 = 103.40$		$da_2 = 158.59$
Root diameter, mm	$df_1 = 89.90$		$df_2 = 145.09$
Accuracy			
Total profile deviation $F_\alpha, \mu\text{m}$	13.5		15.1
Profile form deviation $f_{fa}, \mu\text{m}$	11.6		8.4
Profile slope deviation $f_{H\alpha}, \mu\text{m}$	5.3		−12.3
Total helix deviation $F_\beta, \mu\text{m}$	12.3		14.1
Helix form deviation $f_{f\beta}, \mu\text{m}$	7.0		12.7
Helix slope deviation $f_{H\beta}, \mu\text{m}$	9.5		17.9

In the experiment, two sets of gears were used. The first set, whose teeth had been gas-nitrided (surface hardness 600–750 HV), provided the necessary data for training and testing the neural network in normal operational conditions (without damage). The second gear pair, which was only quenched and tempered (surface hardness 28–30 HRC), was used to test the ability of the neural network to detect a failure. Table 2 provides the details of the load applied and the duration time of each load stage.

Table 2. Load stages for the tested gear pairs.

Gear Pair I (Quenched and Tempered + Gas-Nitrided)				
Load Stage	Pinion Torque, Nm	Pinion Revolutions, rev/min	Number of Pinion Load Cycles	Load Stage Duration Time
0	42	2500	$1.5 \cdot 10^5$	1 h
1	455	2500	$1.5 \cdot 10^6$	10 h
2	455	2500	$1.5 \cdot 10^6$	10 h
3	455	2500	$1.5 \cdot 10^6$	10 h
4	455	2500	$1.5 \cdot 10^6$	10 h
5	455	2500	$1.5 \cdot 10^6$	10 h
6	455	2500	$1.5 \cdot 10^6$	10 h
7	455	2500	$1.5 \cdot 10^6$	10 h
8	455	2500	$1.5 \cdot 10^6$	10 h
Gear Pair II (Quenched and Tempered Only)				
Load Stage	Pinion Torque, Nm	Pinion Revolutions, rev/min	Number of Pinion Load Cycles	Load Stage Duration Time
0	42	2500	$1.5 \cdot 10^5$	1 h
1	138	2500	$2.5 \cdot 10^6$	16 h 40 min
2	244	2500	$2.5 \cdot 10^6$	16 h 40 min
3	342	2500	$2.5 \cdot 10^6$	16 h 40 min
4	455	2500	$2.5 \cdot 10^6$	16 h 40 min
5	455	2500	$2.5 \cdot 10^6$	16 h 40 min

There are various methods for measuring the pitting damage for quantitative evaluation such as the loss of mass of gears [24], tooth profile measurements [25], measuring the

particles in oil [26], or the use of image processing [27] and vision measurements [28]. In this paper, after each load, the stage gears were inspected for any sign of damage. If failure occurred, it was captured as a photograph, and then the image was processed to evaluate the percentage area of pitting damage. The details of the image processing algorithm for the calculation of the percentage area of pitting are given in [29].

For gear pair I, no sign of damage was detected. A macro photograph of the tooth surfaces after the eighth load stage is shown in Figure 3.

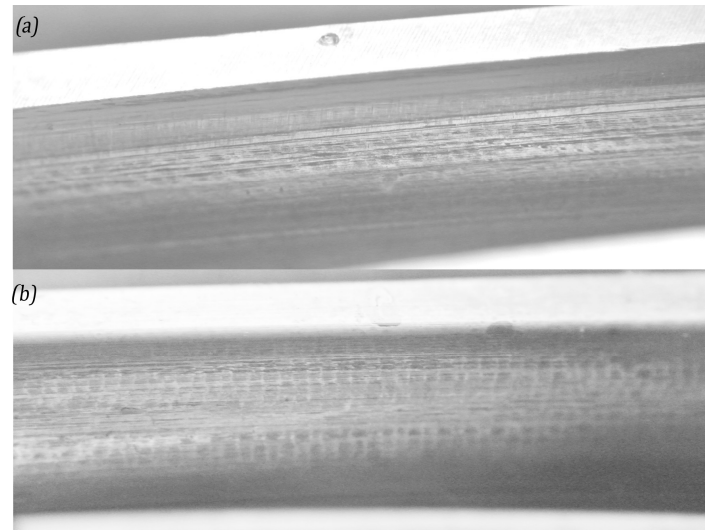


Figure 3. Macro photograph of tooth surfaces of the gas-nitrided gear pair (gear pair I) after tests: (a) pinion tooth surface; (b) gear tooth surface.

The pitting rate for the quenched and tempered gear pair (gear pair II) after each load cycle is presented in Figure 4.

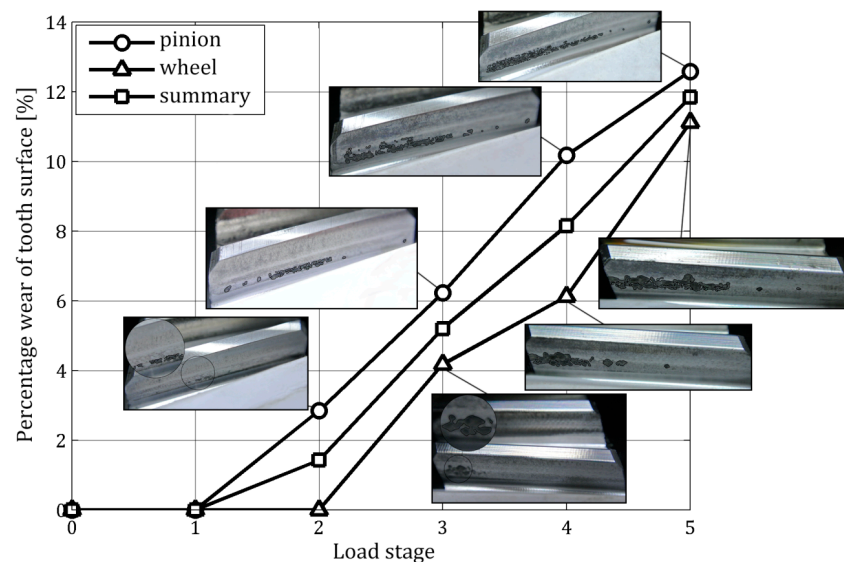


Figure 4. Percentage damage of tooth surface after each load stage for gear pair II (quenched and tempered only).

The first sign of failure occurred after the second load stage on the pinion tooth surfaces. Next, the pitting progressively increased, leading to heavy damage to surfaces after the fifth load stage (Figure 5).

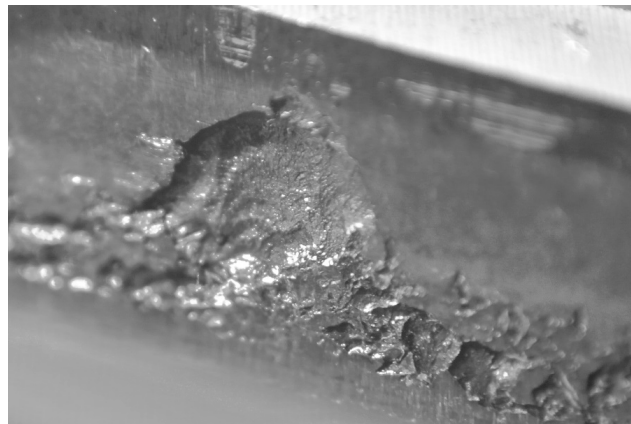


Figure 5. Macro photograph of pitting in gear pair II after the tests.

More information about this kind of damage and the nature of pitting formation can be found in [29]. Moreover, at the beginning and end of each load cycle, the horizontal (perpendicular to the plane of the shaft) vibrations of the tested gear case were acquired with the aid of a 3-axis piezoelectric sensor (PCB Piezoelectric ICP 356B21) and data acquisition module (National Instruments NI9234). The sampling frequency was 25 kHz while the signal duration was 1 s, which gives 25,000 samples for each measurement.

2.2. Data Preparation

Before applying analytical machinery, the data have to be prepared adequately, which is described in this section. First, the measured and acquired vibrational signal (Figure 6a) is split into 25 segments (Figure 6b).

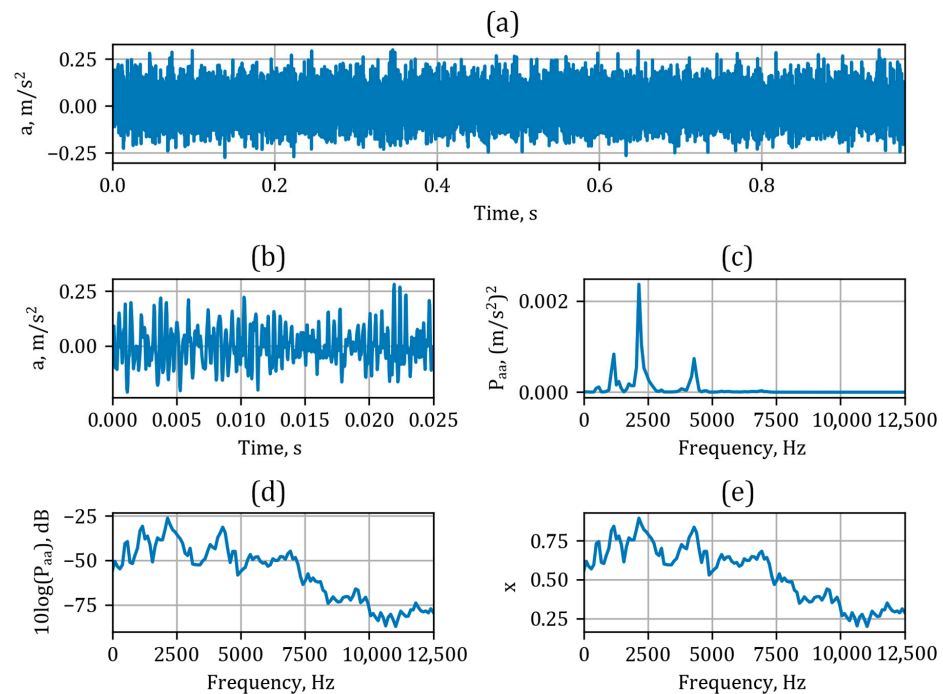


Figure 6. Signal processing: (a) acquired acceleration signal; (b) segment of signal; (c) power spectral density; (d) logarithmic power spectral density; (e) normalized power spectral density.

In the next step, a segment of the signal is transformed into the frequency domain with the aid of power spectral density (P_{aa}) estimation according to Welch's approach [30] (Figure 6c) with the Hann window function. Furthermore, the power spectral density is presented in dB scale by $10\log(P_{aa})$ (Figure 6d). Finally, this logarithmic power spectral

density is normalized (x at Figure 6e). The above-mentioned procedure performed for all measurements resulted in 880 normalized periodograms in the case of normal operation (gear pair I, load stages from 1 to 8), and 560 in the case of damaged gears (gear pair II, load stages from 2 to 5). Figure 7 shows the distribution of selected parameters of the obtained signals for two populations.

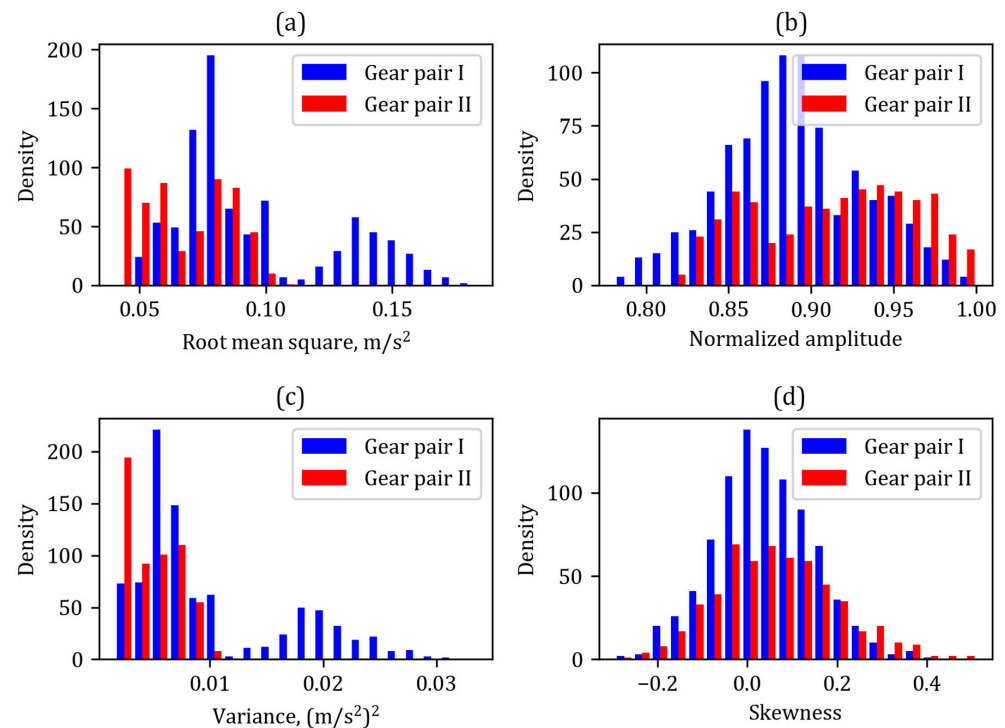


Figure 7. Distribution of selected parameters of vibrational signal: (a) root mean square value; (b) maximum normalized amplitude over 900 Hz; (c) variance; (d) skewness.

It can be seen that simple analysis within the time domain (Figure 7a,b,d) does not provide sufficient information about the state of the gearbox operation. The density distribution of selected parameters varied over the two populations, however, failure could not be predicted since the parameter values overlapped. The same conclusions could be drawn in the case of simple analysis within the frequency domain. There was no significant difference between the amplitude of harmonics corresponding to the frequency of meshing (>900 Hz) for normal operation and failure (Figure 7b).

After the applied preprocessing, we had two datasets, to clarify, one for normally operating gears and the other for malfunctioning. The idea was to expose the models only to the valid data and observe the reaction of the trained models to the faulty gear measurements. In this way, we did not want to impose a simple classification (on the particular type of damage), but rather see a more generic behavior of the relevant features extracted by the semi-supervised algorithms. Therefore, we split the first dataset into training and test subsets with a 70–30 ratio.

2.3. Principal Component Analysis

Principal component analysis (PCA) is a method used for dimensionality reduction in multi-featured data. It relies on the sequence of linear transformations of input variables. As a result, one obtains a set of new features, being linear combinations of the input, which maximizes the explanation of the dataset variance. Therefore, the first PCA variable explains the most variance, and hence the following subsequent features less. In our case, the PCA fit was performed on the healthy gear training dataset. All data were projected into a two-component space, as presented in Figure 8.

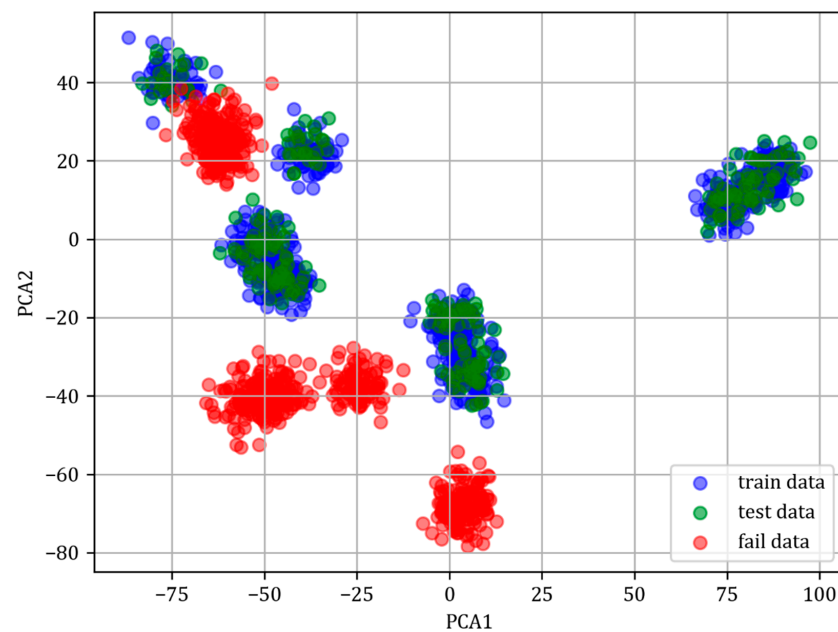


Figure 8. Reduced dimensionality data with PCA.

The data formed five clusters for normal operation and four clusters for the failure state. Each cluster corresponded to the different measurement series, conducted at different times and states of gear wear. As can be noted, there was no clear separation between the points representing the measurements of the operating and failing gears. First, two principal components explained no more than 61.3% and 9.9% of the data variance. This led us to the conclusion that more robust, nonlinear treatment was needed.

3. Calculation

3.1. Autoencoder as a Simple Anomaly Detector

The autoencoder (AE) was designed as a deep neural network with five dense hidden layers (Figure 9). As input, the network takes normalized periodogram x (vector of length 129). This periodogram is next coded by the encoder to a vector h . The encoded vector h is the input layer for the decoder, which reconstructs it to vector r (output layer).

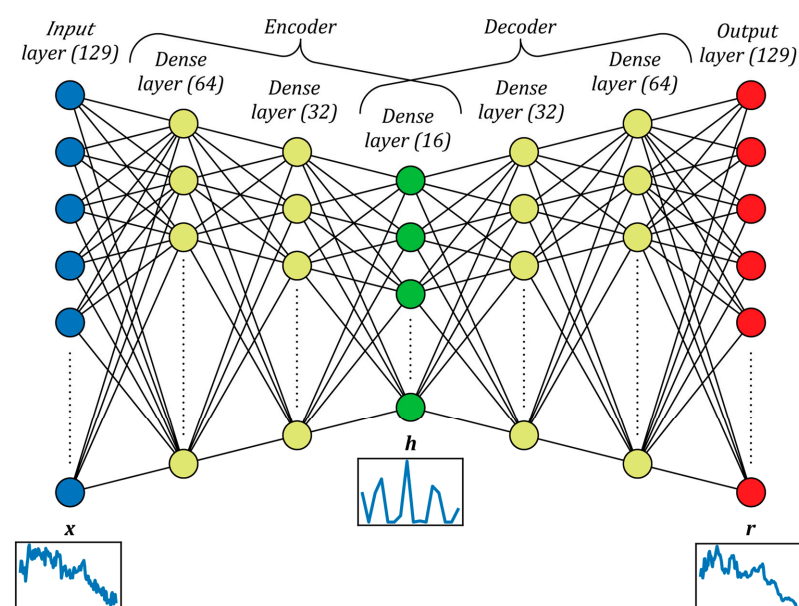


Figure 9. Autoencoder with deep neural network architecture.

For all hidden layers, the rectified linear unit (ReLU) activation function was used. The network was trained with the aid of the randomly selected 70% of periodograms captured for normal operation of the gearbox (gear pair I), which resulted in 616 training samples. The other 30% (264 samples) was used for validation. As the measure of loss, the mean absolute error was applied (Figure 10).

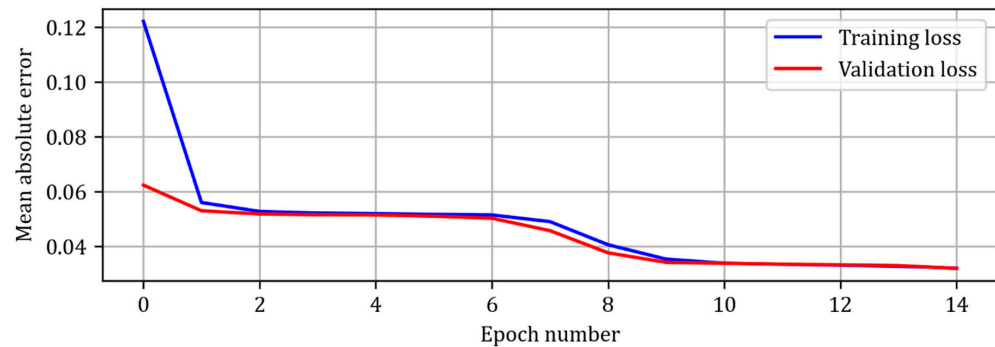


Figure 10. Training and validation loss.

To understand whether the observed measurement came from a healthy or damaged gear set, we looked at its reconstruction. We fed our model with different measurements and verified the error of the reconstruction. The threshold value, above which the state of the gearbox will be classified as an anomaly (failure), was calculated with Formula (1):

$$TH = \frac{1}{n} \sum_{i=1}^n MAE_i + 1.5 \sqrt{\frac{1}{n} \sum_{i=1}^n \left(MAE_i - \frac{1}{n} \sum_{i=1}^n MAE_i \right)^2} = 0.0408, \quad (1)$$

where MAE_i is the i -th mean absolute error for training, and $n = 616$ is the number of training data. In other words, the gear pair will be classified as faulty if the reconstruction error is at least one and a half standard deviations above the mean training error. Figure 11 presents the classification results for the testing data (gear pair I without failure).

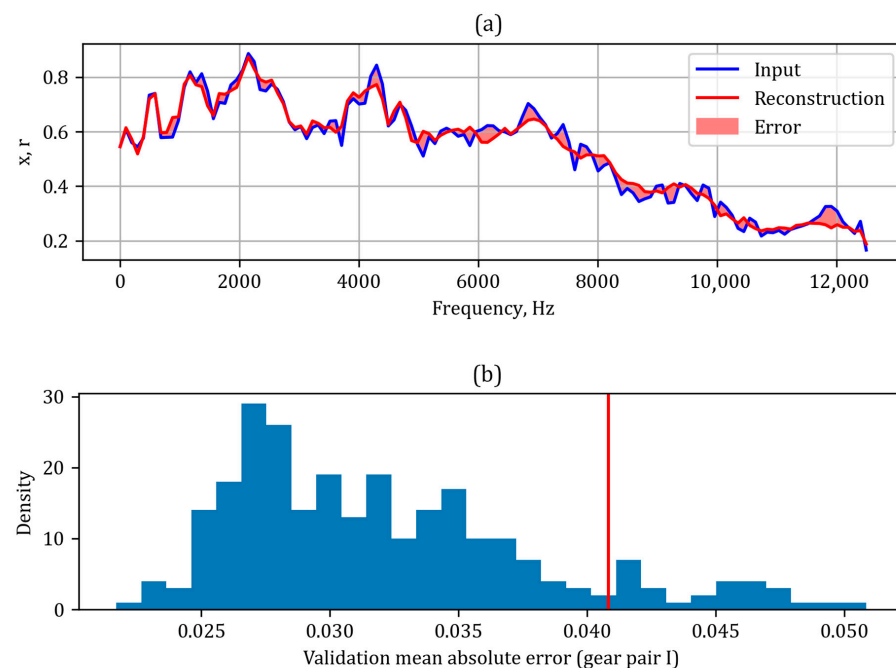


Figure 11. Results of validation: (a) example of input x , reconstruction r and error; (b) distribution of reconstruction error for testing the data of gear pair I.

It can be seen that the reconstruction matched the input (Figure 11a), however, 10.2% of the test data were incorrectly classified as an anomaly, while in most cases (89.8%), the normal state of operation was predicted correctly (Figure 11b).

The main goal of the designed approach is the ability of the neural network to detect a failure—even the smallest initial pitting (Figure 2, second load stage). To check this ability, the periodograms of gear pair II were used as inputs. Figure 12 shows the results of the state prediction in the case of damaged teeth.

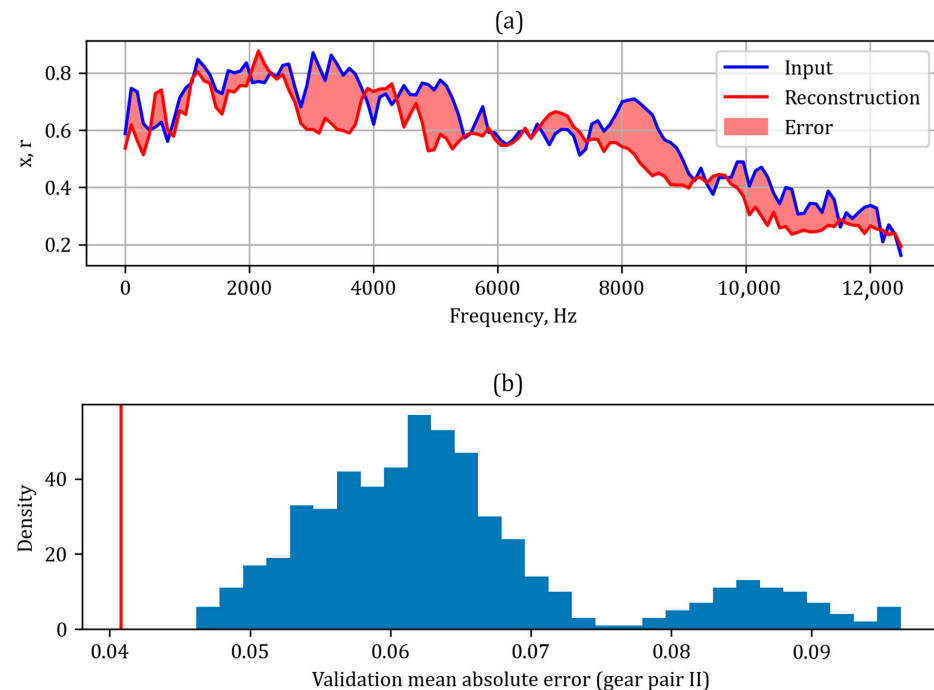


Figure 12. Results of failure detection: (a) example of input x , reconstruction r , and error; (b) distribution of reconstruction error for the testing data of gear pair II.

It can be seen that the reconstruction error was large (Figure 12a). The greatest differences occurred within the frequency range between 2000 and 4000 Hz. Such results (Figure 12b) suggest that signs of pitting might be observed as the increase in the amplitude of the third- and fourth-order harmonics of transmission error (meshing frequency is 1250 Hz). This increase, however, was too subtle to predict a failure with just simple analysis in the frequency domain (Figure 7b). The neural network correctly predicted the state of failure in all cases (Figure 12). Figure 13 shows the confusion matrix that summarizes the results of the state prediction experiment (Figures 11b and 12b).

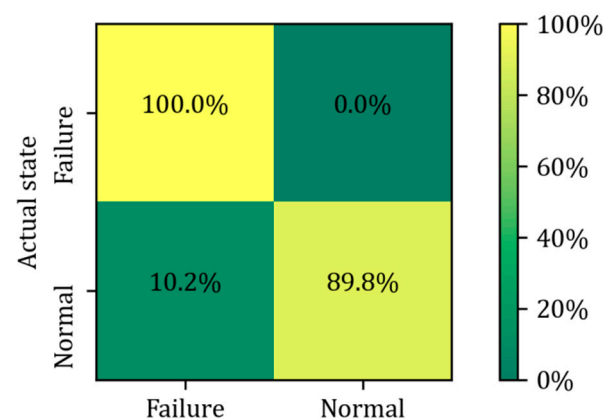


Figure 13. Confusion matrix for basic anomaly detection.

Neural networks are very effective (100%) in failure detection in the case of true positive classification. This type of classification is crucial for any accountable mechanical system, since every failure should be detected quickly to prevent further damage. The method was able to properly predict 89.8% of normal states in true negative classification. The network misclassified only 10.2% (false positive) of states where the gearbox was actually without damage. The above results give a high precision (0.90), F1-score (0.95), and recall (1.00). Based on the performed experiments, it can be seen (Figure 12) that the threshold could be shifted to 0.045. This could lead to an increase in correctly classified true negative cases.

3.2. Autoencoder and Convex Hull-Based Clustering in Latent Layer

The key feature of autoencoders is their ability to perform a nonlinear feature reduction. To use its power, we introduced a small modification to the model from the previous section. The autoencoder architecture for latent layer exploration was the same as previously, however, only the two neurons in the latent layer (with linear activation function) were used (Figure 14).

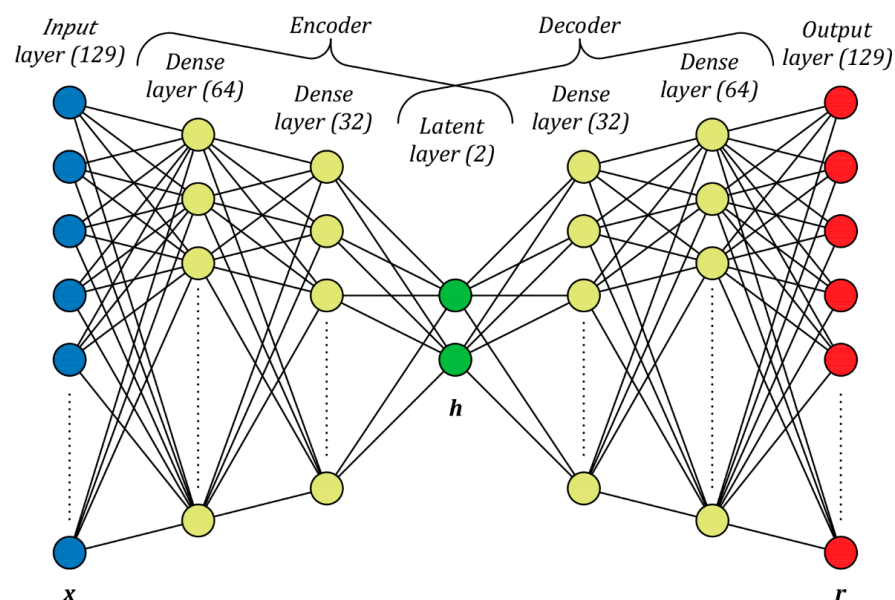


Figure 14. Autoencoder architecture for latent layer exploration.

The network was trained with the same data and methods described in Section 3.1. Within the latent space, clear clusters could be observed (Figure 15).

Gear case acceleration normalized periodograms for normal operation in latent space (Figure 15 green and blue points) formed five clusters, which represent the time of operation (number of load cycles taken) and the state of the natural wear of teeth. The red points (Figure 15) formed three clusters that could be clearly distinguished from the clusters referring to a normal operation, which indicated the abnormal state of the gearbox (failure). For the prediction of gear state, a convex hull-based classification was proposed. This method calculates a convex hull (CH) [16] of the training data in latent space. Whenever the checked point lies within the convex hull, the gearbox is considered to be in the normal state of operation, while if an outlier is detected (the point lies outside the convex hull), gearbox failure is predicted. Point $h = [h_1 \ h_2]^T$ lies within the convex hull if for every facet it satisfies the condition $h \cdot n + w < 0$, where n is the vector unit normal to the facet and w is the offset of the facet from the origin. The results of the convex hull-based classification are shown in Figure 16.

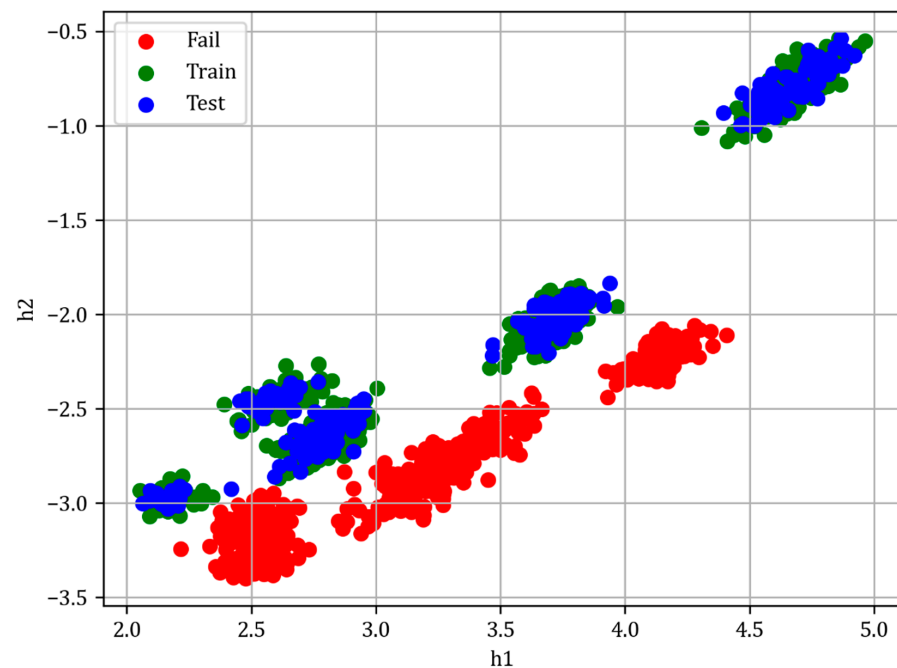


Figure 15. Data in the latent space of the autoencoder.

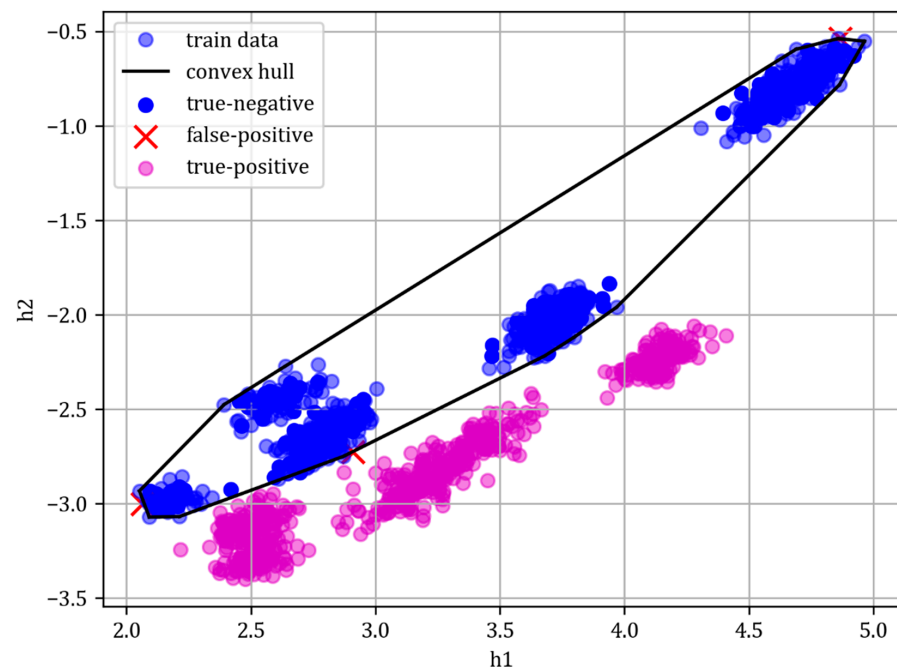


Figure 16. Results of the convex hull-based classification in latent space.

The resulting confusion matrix is presented in Figure 17.

Similar to the case of the autoencoder in Section 5, the proposed approach was very effective (100%) in failure detection in the case of true positive classification. Moreover, the number of normal state predictions in the case of true negative classification was increased to 98.9%. The method misclassified only 1.1% (false positive) of states where the gearbox was actually without damage. The above results showed a very high precision (0.98), F1-score (0.99), and recall (1.00).

To show the nature of the latent data separation and compare the results with semi-supervised methods, we conducted supervised training of the support vector machine models. For this purpose, we reorganized the dataset and granted labels to normal and

failure subsets. The labeled data were used with a 70–30% training–test split. Results for the linear and nonlinear support vector machine (SVM) models are given in Figure 18.

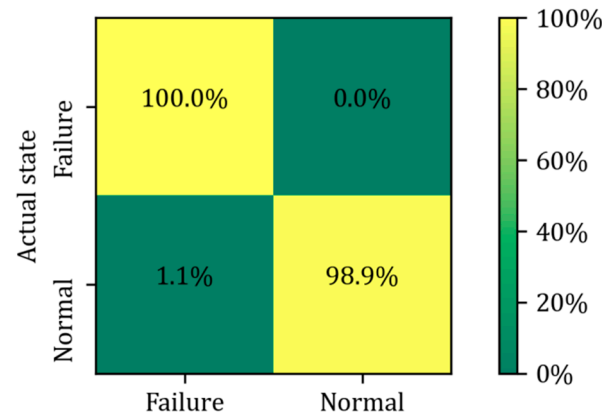


Figure 17. Confusion matrix for convex hull classification of latent data.

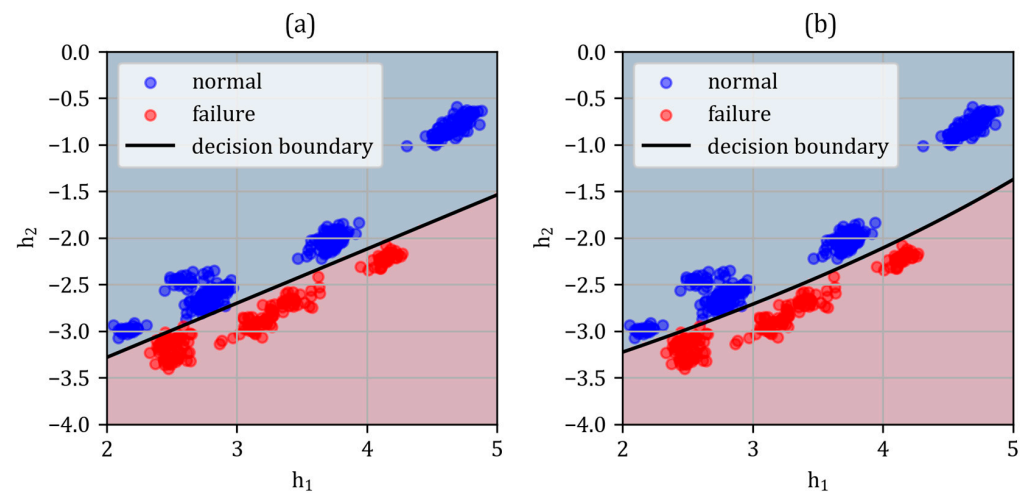


Figure 18. Decision boundaries and test data for: (a) linear SVM; (b) SVM with nonlinear kernel.

In the case of the linear SVM model, the decision boundary takes the form of a straight line $h_2 = mh_1 + n$ where $m = 0.581$ and $n = -4.444$. For nonlinear SVM, as a kernel, the polynomial of the fourth-order was used, resulting in a decision boundary that can be approximated by equation $h_2 = ah_1^4 + bh_1^3 + ch_1^2 + dh_1 + e$, where $a = 0.0012$, $b = -0.0107$, $c = 0.0712$, $d = 0.2795$, and $e = -4.0032$. The resulting confusion matrices are given in Figure 19.

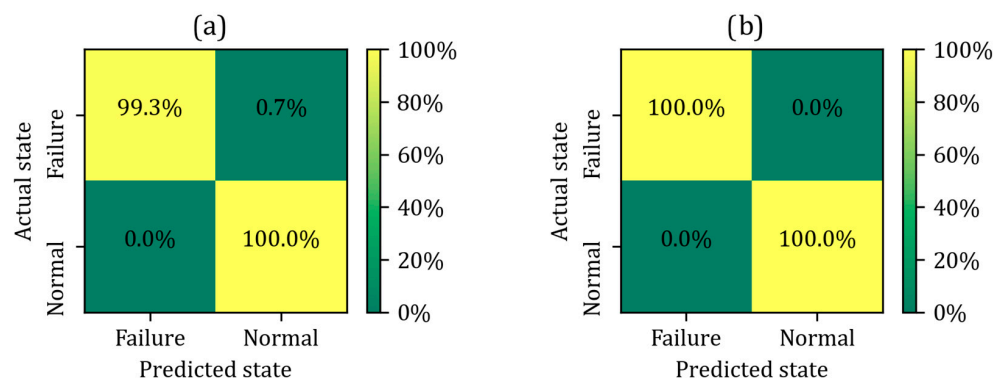


Figure 19. Confusion matrices for SVM classification: (a) linear SVM; (b) nonlinear SVM.

The linear SVM model performed well, achieving high precision (1.00), recall (0.99) and F1-score (0.99), while SVM with the polynomial kernel separated the data perfectly (Figure 19b). The nonlinearity of latent data separation can be reflected by the R-squared measure between the linear and nonlinear decision boundary line. For the analyzed case, it was $R^2 = 0.98$, so the data separation was highly linear, making it easily distinguishable.

3.3. Generative Adversarial Network and Convex Hull-Based Clustering of Discriminator Output

The generative adversarial network for anomaly detection used in this study consisted of two deep neural networks: the generator and the discriminator (Figure 20).

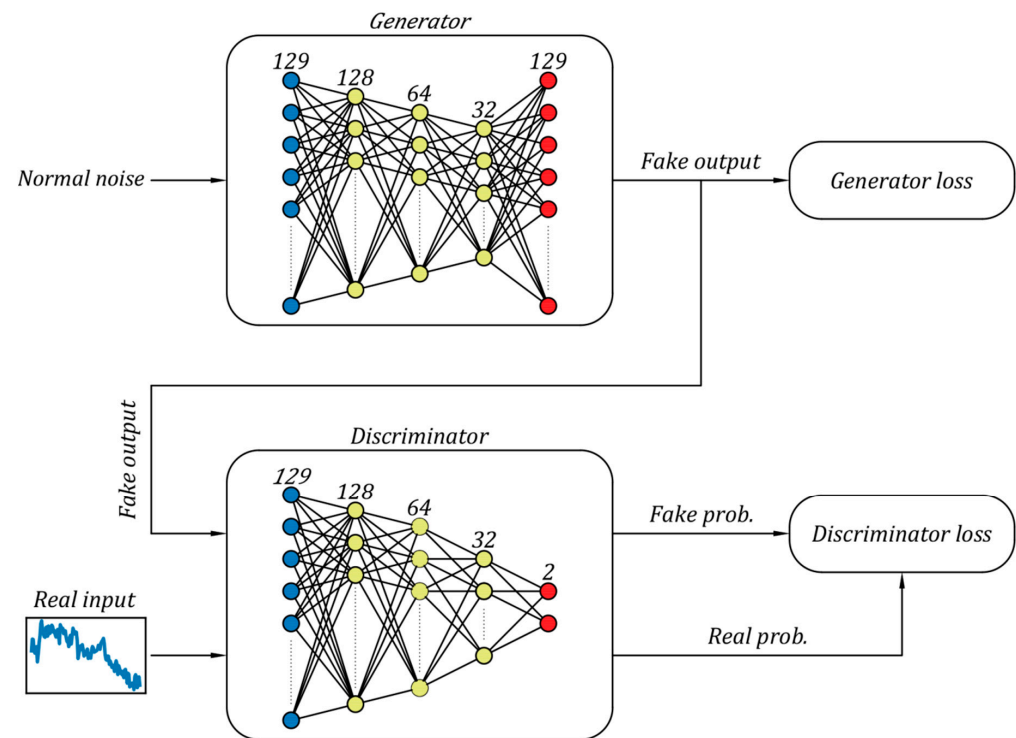


Figure 20. Generative adversarial network for failure prediction.

The generator consisted of three hidden dense layers with a rectified linear unit activation function and aims to generate the fake output based on real input data (periodogram). The discriminator predicts whether the input is real or fake by mapping the input into two-dimensional probability space. As in the case of the generator, the discriminator was designed as a deep neural network with three hidden dense layers with rectified linear unit activation function. As a loss function for the generator and discriminator while training, the binary cross entropy was used. The network was trained only on periodograms for a normal state of operation. The probability space (output) of the discriminator, along with convex hull-based clustering, is shown in Figure 21.

The data points for both the normal and abnormal states were spread across a straight line. The training data formed a clear cluster at the top. Based on this, convex hull-based clustering was performed similarly, as in Section 3.2. The resulting confusion matrix is plotted in Figure 22.

The use of GAN for anomaly detection resulted in relatively high precision (0.98) and F1-measure (0.95). The recall was slightly lower and reached the value of 0.93.

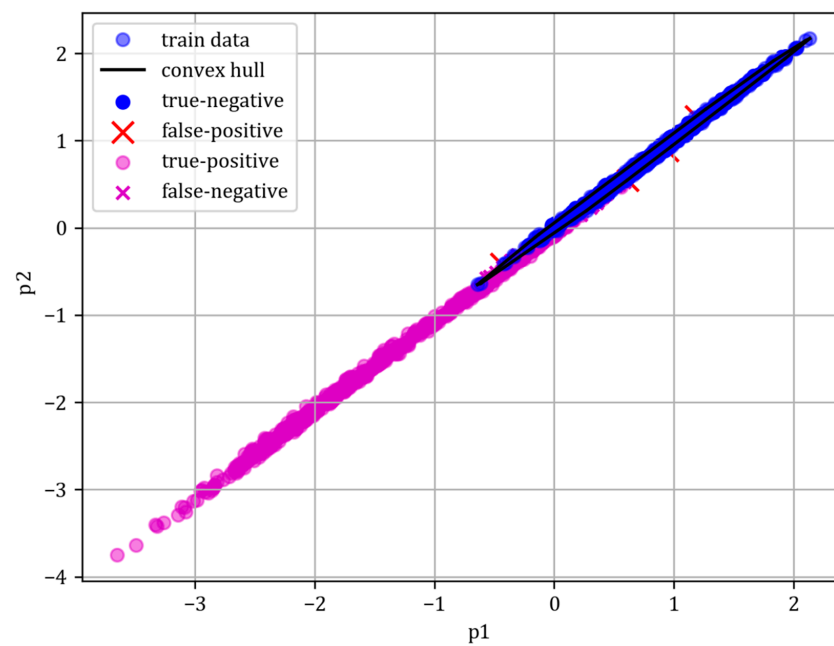


Figure 21. Discriminator output.

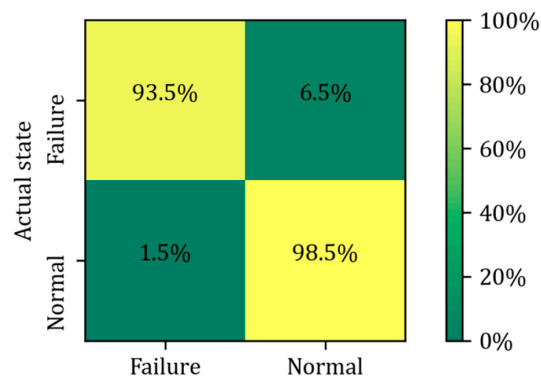


Figure 22. Confusion matrix for convex hull classification of discriminator output.

4. Results and Discussion

To understand the nature of autoencoder and GAN behavior and evaluate the possibility of wear severity prediction, we present the labelled latent data of an autoencoder (Figure 23) and the discriminator output of the GAN (Figure 24).

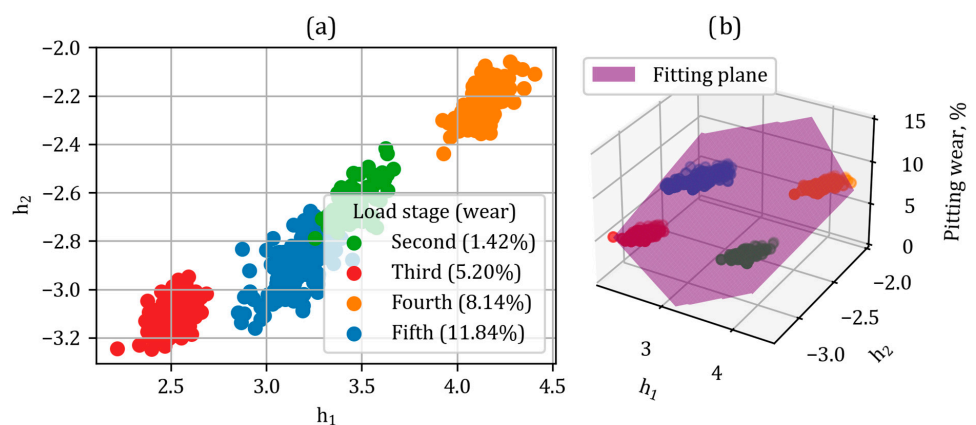


Figure 23. Labelled latent data of an autoencoder: (a) latent space; (b) percentage pitting wear.

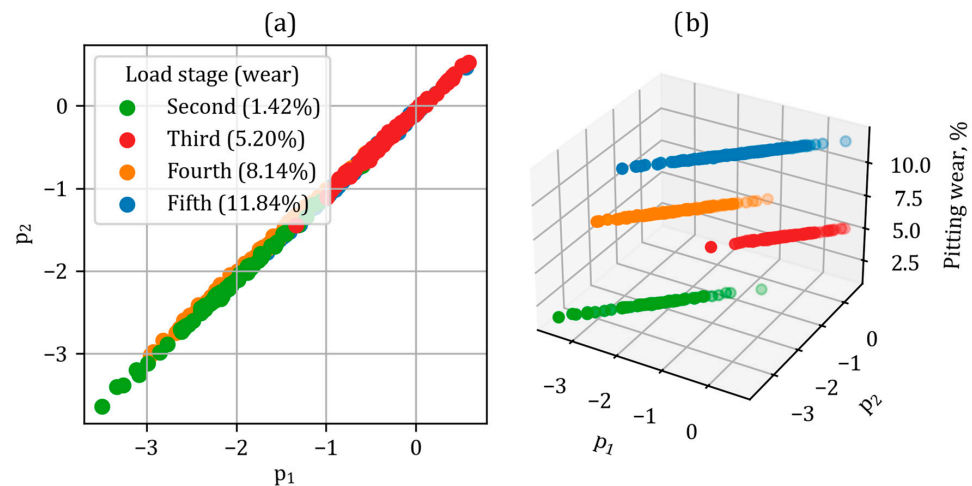


Figure 24. Labelled discriminator output for the GAN: (a) probability space; (b) percentage pitting wear.

It can be seen that the latent data, mapped into three-dimensional space by adding the pitting wear coordinate (Figure 23b), allowed us to evaluate the pitting rate based on latent coordinates h_1 h_2 . The plane was fitted with a 0.26 regression score. The prediction of wear based on linear regression was not accurate in this case, but it gave some idea of how the data points were spaced. A more sophisticated nonlinear regression should be applied like a decision tree or polynomial regressor, which would perform well in supervised learning within the latent space.

The opposite situation might be observed within the output space of the discriminator (Figure 24). Data points formed nearly straight lines within the probability space. Mapping them into three-dimensional space by adding wear coordinates (Figure 24b) did not help, since wear took non-unique values for the same point on the probability plane. It can also be observed as an overlapping of clusters in Figure 24a.

A comparison of the performance of the proposed models is given in Table 3.

Table 3. Comparison of results.

Model	Precision	Recall	F1-Score
Autoencoder	0.90	1.00	0.95
Autoencoder + convex hull classification	0.98	1.00	0.99
Autoencoder + linear SVM	1.00	0.99	0.99
Autoencoder + nonlinear SVM	1.00	1.00	1.00
GAN + convex hull classification	0.98	0.93	0.95

Recall was very high for both the autoencoder (AE) and autoencoder combined with convex hull classification of the latent data (AE+CH), while the lowest value was noted for GAN + convex hull. The best overall performance (highest precision, recall, and F1-score) was observed for the proposed AE+CH model. The conventional unsupervised machine learning approach based on principal component analysis (PCA) did not show a clear separation between normal and failure operations within a space of reduced dimensionality, making it difficult to generalize the model and provide proper classification of the fault. Reduced dimensionality by the autoencoder (AE+CH) provided an advantage over the PCA and GAN, that is, an almost linear separation between the two states of gearbox operation (Figures 15 and 16). This allows us, apart from failure detection, to monitor the condition of the gearbox in both normal and failure operations. It is worth mentioning that the proposed semi-supervised learning approach (AE+CH) achieved similar results to the supervised methods (AE + linear SVM and AE + nonlinear SVM).

5. Conclusions

This paper presented a semi-supervised learning approach for fault detection and state prediction in gearboxes. The experimental setup, measurements, and data preparation were presented. Moreover, the architecture of the proposed models was shown along with semi-supervised training and validation. Based on the performed experiment, the following conclusions can be drawn:

- The proposed methods based on autoencoders were efficient and could detect even the initial state of pitting formation, which may be difficult with the aid of signal analysis in the time and frequency domain;
- The best model (AE+CH) showed very high effectiveness (100%) in failure detection (true positive) and 98.9% in normal state prediction (true negative), which resulted in a very high F1-measure (0.99);
- Vibration excitation due to pitting was observed in a higher order of harmonics;
- The latent space analysis revealed a generalized perspective on the gear wear—the measurements drifted in a specific direction in the latent space with the progress of the gearbox damage;
- Autoencoder outperformed the generative adversarial network in terms of generalization on wear prediction.

Future work could focus on the use of some interesting autoencoder abilities to extract the features of vibrational signals that may be beneficial in detecting failure based on the frequency spectrum. This can therefore lead to a better understanding of the influence of surface damage on gearbox vibration and noise excitation.

Author Contributions: Conceptualization, M.B.; Methodology, M.B. and B.K.; Software, M.B.; Validation, M.B.; Formal analysis, M.B. and B.K.; Investigation, M.B.; Resources, M.B.; Data curation, B.K.; Writing—original draft preparation, M.B. and B.K.; Writing—review and editing, M.B. and B.K.; Visualization, M.B.; Supervision, B.K.; Project administration, M.B.; Funding acquisition, M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding and the APC was funded by the Dean of the Faculty of Mechanical Engineering and Aeronautics of Rzeszów University of Technology.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

This section presents the methodology of the neural ablation testing of models used in this paper. Every model was tested to maximize performance while maintaining a reasonable learning time. The methodology is shown here based on a representative example of an autoencoder and convex hull clustering in its latent layer (Figure 14). Neural ablation testing is based on deactivating the selected neurons and evaluating its impact on the model performance. In the present study, we symmetrically deactivated each of the hidden layers of an autoencoder according to the test scheme shown in Figure A1.

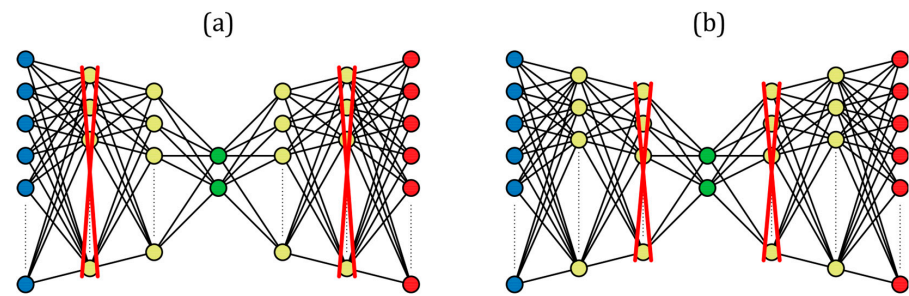


Figure A1. Model architectures used in the neural ablation tests: (a) test no. 1; (b) test no. 2 (blue—input layer, yellow—hidden layer, green—latent layer, red—output layer).

The results of the convex hull classification of the ablated models are shown in Figure A2.

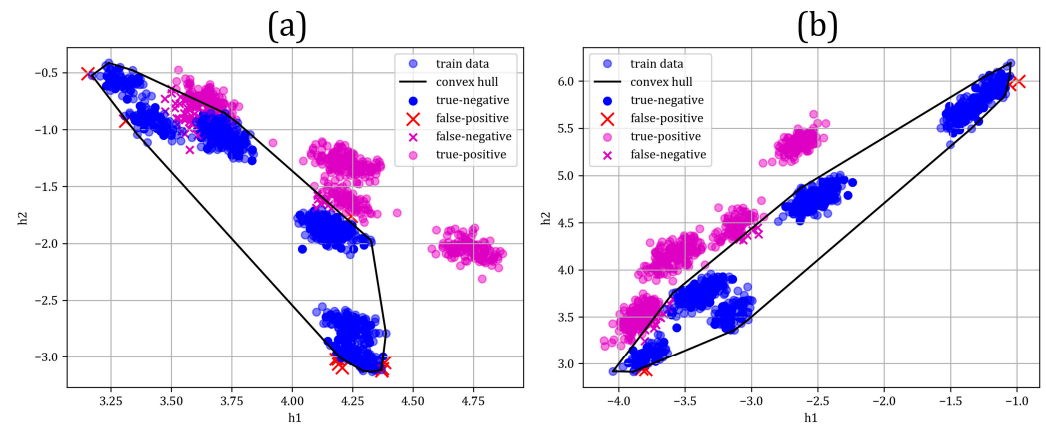


Figure A2. Results of the convex hull classification of the ablated models: (a) test no. 1; (b) test no. 2.

It can be noted that there was no clear separation between the two gear states. Corresponding confusion matrices are shown in Figure A3.

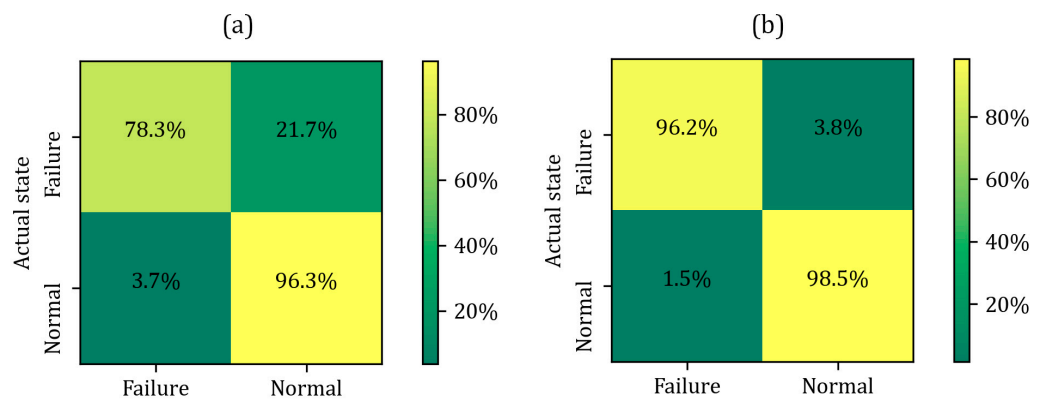


Figure A3. Confusion matrices for the ablated models: (a) test no. 1; (b) test no. 2.

The results of the ablation tests show the need to provide the model with more capabilities in terms of reflecting nonlinearities within the data. Therefore, the model took on the architecture shown within the paper in Figure 14.

Similar procedures were used to evaluate the autoencoder as a basic anomaly detector (Figure 9) and GAN for anomaly detection (Figure 20).

References

1. Yan, W.; Shabaz, M.; Rakhra, M. Research on Nonlinear Distorted Image Recognition Based on Artificial Neural Network Algorithm. *J. Interconnect. Netw.* **2022**, *22*, 2148002. [\[CrossRef\]](#)
2. Rykała, Ł. Application of Hybrid Neural Network System in Image Processing. *SLW* **2019**, *51*, 141–153. [\[CrossRef\]](#)
3. Burghardt, A.; Gierlak, P. Robotic Grinding Process of Turboprop Engine Compressor Blades with Active Selection of Contact Force. *Teh. Vjesn.* **2022**, *29*, 15–22. [\[CrossRef\]](#)
4. Styła, M.; Kiczek, B.; Adamkiewicz, P. Image Reconstruction Using Radio Tomography and Artificial Intelligence in Tracking and Navigation Systems for Indoor Applications. In Proceedings of the 2023 International Interdisciplinary PhD Workshop (IIPhDW), Wismar, Germany, 3–5 May 2023; IEEE: Wismar, Germany, 2023; pp. 1–4.
5. Styła, M.; Kiczek, B.; Kłosowski, G.; Rymarczyk, T.; Adamkiewicz, P.; Wójcik, D.; Cieplak, T. Machine Learning-Enhanced Radio Tomographic Device for Energy Optimization in Smart Buildings. *Energies* **2022**, *16*, 275. [\[CrossRef\]](#)
6. Chen, H.; Hu, N.; Cheng, Z.; Zhang, L.; Zhang, Y. A Deep Convolutional Neural Network Based Fusion Method of Two-Direction Vibration Signal Data for Health State Identification of Planetary Gearboxes. *Measurement* **2019**, *146*, 268–278. [\[CrossRef\]](#)
7. Rykała, M.; Rykała, Ł. Economic Analysis of a Transport Company in the Aspect of Car Vehicle Operation. *Sustainability* **2021**, *13*, 427. [\[CrossRef\]](#)
8. Bhardwaj, P.; Yadav, K.; Alsharif, H.; Aboalela, R.A. GAN-Based Unsupervised Learning Approach to Generate and Detect Fake News. In *International Conference on Cyber Security, Privacy and Networking (ICSPN 2022)*; Nedjah, N., Martínez Pérez, G., Gupta, B.B., Eds.; Lecture Notes in Networks and Systems; Springer International Publishing: Cham, Switzerland, 2023; Volume 599, pp. 384–396. ISBN 978-3-031-22017-3.
9. Gierlak, P.; Szybicki, D.; Kurc, K.; Burghardt, A.; Wydrzyński, D.; Sitek, R.; Goczał, M. Design and Dynamic Testing of a Roller Coaster Running Wheel with a Passive Vibration Damping System. *J. Vibroeng.* **2018**, *20*, 1129–1143. [\[CrossRef\]](#)
10. Kuczaj, M.; Wieczorek, A.N.; Konieczny, Ł.; Burdzik, R.; Wojnar, G.; Filipowicz, K.; Głuszek, G. Research on Vibroactivity of Toothed Gears with Highly Flexible Metal Clutch under Variable Load Conditions. *Sensors* **2022**, *23*, 287. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Leoni, J.; Tanelli, M.; Palman, A. A New Comprehensive Monitoring and Diagnostic Approach for Early Detection of Mechanical Degradation in Helicopter Transmission Systems. *Expert Syst. Appl.* **2022**, *210*, 118412. [\[CrossRef\]](#)
12. Li, F.; Chen, Y.; Wang, J.; Zhou, X.; Tang, B. A Reinforcement Learning Unit Matching Recurrent Neural Network for the State Trend Prediction of Rolling Bearings. *Measurement* **2019**, *145*, 191–203. [\[CrossRef\]](#)
13. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
14. Ahmed, I.; Ahmad, M.; Chehri, A.; Jeon, G. A Smart-Anomaly-Detection System for Industrial Machines Based on Feature Autoencoder and Deep Learning. *Micromachines* **2023**, *14*, 154. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Chen, X.; Ji, A.; Cheng, G. A Novel Deep Feature Learning Method Based on the Fused-Stacked AEs for Planetary Gear Fault Diagnosis. *Energies* **2019**, *12*, 4522. [\[CrossRef\]](#)
16. Liu, S.; Liu, Y.; Gu, Y.; Xu, X. Method of Extracting Gear Fault Feature Based on Stacked Autoencoder. *J. Eng.* **2019**, *2019*, 8765–8769. [\[CrossRef\]](#)
17. He, G.; Li, J.; Ding, K.; Zhang, Z. Feature Extraction of Gear and Bearing Compound Faults Based on Vibration Signal Sparse Decomposition. *Appl. Acoust.* **2022**, *189*, 108604. [\[CrossRef\]](#)
18. Nguyen, C.D.; Prosvirin, A.E.; Kim, C.H.; Kim, J.-M. Construction of a Sensitive and Speed Invariant Gearbox Fault Diagnosis Model Using an Incorporated Utilizing Adaptive Noise Control and a Stacked Sparse Autoencoder-Based Deep Neural Network. *Sensors* **2020**, *21*, 18. [\[CrossRef\]](#)
19. Han, C.; Hayashi, H.; Rundo, L.; Araki, R.; Shimoda, W.; Muramatsu, S.; Furukawa, Y.; Mauri, G.; Nakayama, H. GAN-Based Synthetic Brain MR Image Generation. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; IEEE: Washington, DC, USA, 2018; pp. 734–738.
20. Lee, C.-K.; Cheon, Y.-J.; Hwang, W.-Y. Studies on the GAN-Based Anomaly Detection Methods for the Time Series Data. *IEEE Access* **2021**, *9*, 73201–73215. [\[CrossRef\]](#)
21. Zenati, H.; Romain, M.; Foo, C.-S.; Lecouat, B.; Chandrasekhar, V. Adversarially Learned Anomaly Detection. In Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), Singapore, 17–20 November 2018; IEEE: Singapore, 2018; pp. 727–736.
22. Su, Y.; Meng, L.; Kong, X.; Xu, T.; Lan, X.; Li, Y. Generative Adversarial Networks for Gearbox of Wind Turbine With Unbalanced Data Sets in Fault Diagnosis. *IEEE Sens. J.* **2022**, *22*, 13285–13298. [\[CrossRef\]](#)
23. Li, J.; Liu, Y.; Wang, Q.; Xing, Z.; Zeng, F. Rotating Machinery Anomaly Detection Using Data Reconstruction Generative Adversarial Networks with Vibration Energy Analysis. *AIP Adv.* **2022**, *12*, 035221. [\[CrossRef\]](#)
24. Niemann, G.; Winter, H.; Bergsträsser, M.; Dietrich, G.; Thomas, W.; Richter, W.; Rettig, H.; Cameron, A.; Blok, H.; Brugger, H.; et al. *Zahnräder Zahnradgetriebe: Vorträge und Diskussionsbeiträge Fachtagung “Antriebs-elemente”, Essen 1954*; Vieweg+Teubner Verlag: Wiesbaden, Germany, 1955; ISBN 978-3-663-06703-0.
25. Lin, J.; Li, H.; Wang, P.; Li, N.; Shi, Z.; Olofsson, U. Compensation of Mounting Error in In-Situ Wear Measurement during Gear Pitting Test. *Measurement* **2022**, *191*, 110808. [\[CrossRef\]](#)
26. Kattelus, J.; Miettinen, J.; Lehtovaara, A. Detection of Gear Pitting Failure Progression with On-Line Particle Monitoring. *Tribol. Int.* **2018**, *118*, 458–464. [\[CrossRef\]](#)
27. Žák, P.; Dynybyl, V. Innovative Analysis and Documentation of Gear Test Results. *Gear Technol.* **2008**, *9*, 64–70.

28. Wang, Z.; Qin, Y.; Chen, W. Vision Measurement of Gear Pitting Based on DCGAN and U-Net. *J. Mech. Sci. Technol.* **2021**, *35*, 2771–2779. [[CrossRef](#)]
29. Batsch, M.; Markowski, T. Comparative Fatigue Testing of Gears with Involute and Convexo-Concave Teeth Profiles. *Adv. Manuf. Sci. Technol.* **2016**, *40*, 5–25.
30. Welch, P. The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging over Short, Modified Periodograms. *IEEE Trans. Audio Electroacoust.* **1967**, *15*, 70–73. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.