Advances in Science and Technology Research Journal

Advances in Science and Technology Research Journal, 2026, 20(2), 388–401
https://doi.org/10.12913/22998624/211974
ISSN 2299-8624, License CC-BY 4.0

Received: 2025.10.15
Accepted: 2025.18.09
Published: 2025.12.23

Detection of selected gear tooth defects using deep neural networks

Piotr Bojarczak[1*], Edyta Osuch-Słomka[2], Sebastian Stanisławek[2], Bartosz Lucedarski[2,3]

[1] Faculty of Transport, Electrical Engineering and Computer Science, Casimir Pulaski Radom University, Malczewskiego 29, Radom, Poland
Lukasiewicz Research Network–Institute for Sustainable Technologies, Pułaskiego 6/10, 26-600 Radom, Poland
[3] School of Management, Technische Universität München, Trivastrasse 23, 80637 München c/o Noori, Germany
* Corresponding author's e-mail: p.bojarczak@urad.edu.pl

ABSTRACT

Gear transmissions, due to demanding operating conditions, are particularly susceptible to wear. Surface fatigue wear (pitting) is among the most difficult to predict and, at the same time, the most critical forms of damage to the working surfaces of cylindrical gears, alongside scuffing. Pitting is classified as an unacceptable tribological degradation process leading to catastrophic wear, the propagation of which ultimately results in tooth fracture and consequently the failure of the entire gear system. For this reason, detecting them is an important task. This paper presented the algorithms for classifying pitting defects using deep learning networks. Classification was based on recorded images of defects. Three types of feature extractors (convolutional network, vision transformer, and hybrid model) used in the classifier were tested. A neural network and a vision language model were used as the classifier. The best accuracy of 86% was achieved for the vision transformer together with the neural network. The novel elements are a comparison of three types of feature extractors with respect to their application to the detection (classification) of gear tooth pitting damage and the use of a multimodal language model for the detection of gear tooth pitting damage.

Keywords: tribology, gear faults, image processing, machine learning.

INTRODUCTION

The operation of gear transmissions under high loads and variable stresses promotes the development of tribological wear. Among the possible forms of degradation of the working surfaces of cylindrical gears is pitting, which results from material fatigue. This phenomenon manifests itself as pits on the tooth surface, which occur as a result of exceeding the fatigue strength under contact stresses. The process is progressive in nature and results from the accumulation of energy within the surface layer of the material. Pitting is determined by the physicochemical properties of the tribological system, including the interacting materials, the lubricant, and the surrounding environment. Once a certain threshold of accumulated energy is exceeded, fragments of the tooth surface break off, typically near the tooth root, where the highest contact stresses occur (Figure 1). Pitting is classified as an unacceptable tribological degradation process leading to catastrophic wear, the propagation of which ultimately results in tooth fracture and consequently the failure of the entire gear system.

The greatest challenge associated with this phenomenon lies in the difficulty of early detection and prediction. Manufacturers of drive gears intended for operation under harsh environmental conditions, such as coal mines – where the transmission oil may be contaminated with solid

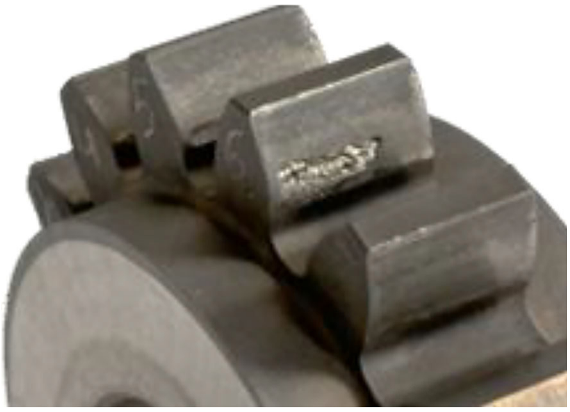

Figure 1. Pitting on the working surface of the gear tooth

particles – as well as producers of agricultural, construction, and forestry machinery, demand high durability and reliability of delivered components [1–10]. These elements must be manufactured from the materials capable of meeting severe operational requirements, thereby minimizing the risk of machine failures and reducing operating costs. Therefore, effective diagnostics and the identification of surface damage on gear teeth are of critical importance for ensuring the durability of gear transmissions.

The literature extensively discusses various types of gears, such as spur, helical, bevel, and planetary gears, along with their influence on gearbox performance characteristics [11]. Particular attention is given to the analysis of vibrations and acoustic emissions as carriers of information about the technical condition of gear systems and their potential for early damage detection during operation [12, 13]. The studies conducted by Kuczaj *et al.* [14] highlighted the importance of vibroactivity analysis of gear transmissions with flexible metal couplings under variable load conditions, which enables a better understanding of the dynamic behavior of gear systems. Diagnostic methods based on vibration signal analysis are being extensively developed and tested under various operating conditions. Łazarz *et al.* [15] compared the effectiveness of selected vibration metrics in diagnosing complex gear damage, while Zhang *et al.* [16] provided a review of modern vibration signal processing techniques in the context of data-driven gear diagnostics. In diagnostic research, multidimensional data interpretation methods are gaining increasing importance, as they enable the monitoring of gearbox conditions under demanding operating environments [17]. The application of machine learning and soft computing methods in the diagnosis of planetary gear systems is also being actively investigated [18]. The issues of wear and fatigue durability of gear teeth have been addressed, among others, by Wersa *et al.* [19], who emphasized the influence of vibrations and overloads on the development of microcracks and surface degradation of gear teeth. Proper lubrication of the gearbox also plays a crucial role, as highlighted by Wieczorek [20,21] in the context of maintaining appropriate operating conditions.

In summary, the literature highlights the importance of a comprehensive analysis of the dynamic operating parameters of gear transmissions and the application of advanced diagnostic methods, which enable early damage detection and extend the service life of gear systems [22–24].

In recent years, there has been rapid advancement in both camera technology and GPU performance. This has enabled the development and implementation of increasingly complex computer vision algorithms. The advancement of machine learning, and in particular the algorithms based on deep learning, has significantly improved the effectiveness of vision-based methods.

Previously, the vision methods for feature extraction relied on simple filters combined with HOG algorithms [25, 26]. The effectiveness of this approach was limited by two factors:

- the manual selection of features describing the classified object,
- the proper determination of the size of the feature set from which features were selected for classification.

The publication [27] represents a milestone in the development of vision-based classification (detection) algorithms. In this approach, the extraction of characteristic features describing the classified object is carried out automatically, based on the previously collected training data. In the case of [27], a trained convolutional neural network is responsible for feature extraction. Another important property of modern vision algorithm based on deep learning is the ability to integrate both the feature extraction stage and the construction of the classifier itself into a single training process. This avoids the aforementioned factors affecting the quality of the algorithms [28, 29].

In the past few years, there has also been rapid progress in translators, including large language models (LLMs) [30]. In these models, the transformer plays a key role. Its purpose is to encode a single word in such a way that its representation depends on all other words appearing in the sentence. This concept has been adopted in vision algorithms for feature extraction [31] – the so-called vision transformer. In addition to convolutional neural networks and vision transformers, hybrid models are also used for feature extraction [32]. Such models combine the aforementioned two approaches.

The significant progress in deep learning algorithms has enabled their application in such domains as autonomous transport [33], diagnostics [34], and satellite image analysis [35]. In this study, the authors attempted to apply classifiers based on three types of feature extractors to the detection (classification) of gear tooth pitting damage. According to the authors, the novel contribution lies in the following aspects:

- a comparison of three types of feature extractors with respect to their application to the detection (classification) of gear tooth pitting damage,
- the use of a multimodal language model for the detection of gear tooth pitting damage.

PROPOSED APPROACHES

This section presents a method for the classification of pitting-type damage using classifiers of different architectures. The following classifiers were considered: a convolutional network combined with a single fully connected layer, a vision transformer, a hybrid network (a combination of a convolutional network and a vision transformer), and a vision–language model.

Detection based on the classifier

Tooth damage detection can be treated as a classification process, in which tooth images are assigned to predefined groups based on their characteristic features. An essential element of classification, which directly determines its quality, is the proper selection of the features being classified. Figure 2 presents the structure of the classifier used for tooth damage classification. It consists of a feature extraction block and a single layer neural network or Vision Language Model. In the case of tooth damage image classification, feature extraction can be realized using one of three available methods:

- based on a convolutional neural network,
- based on a vision transformer,
- based on a hybrid model (a combination of a transformer and a convolutional network).

A convolutional neural network extracts local features (lines, curves) describing the classified object. In contrast, the transformer relies on global features for extraction. The hybrid model combines both approaches, utilizing local, as well as global features. The authors tested all of the above-mentioned methods for the purpose of tooth damage classification.

CNN backbone

A convolutional neural network (CNN) consists of filter banks [27]. Its structure is illustrated in Figure 3. Each filter is represented by a single neuron, with weights adjusted during the training process. The hierarchical design of the network allows it to capture simple patterns in lower layers and progressively more complex representations in higher layers.

Each filter extracts a single feature of the object, represented by lines or curves in the feature map. The resulting feature maps are aggregated, and their size is reduced by half through downsampling. The obtained output is subsequently processed by the following filter banks (in Figure 3, a filter bank is represented by a single network layer). These operations are repeated until the final feature map reaches the intended size. Due to the filter frame size of 3 × 3, the extracted features are of a local nature, which limits the ability to capture global information describing the classified object.

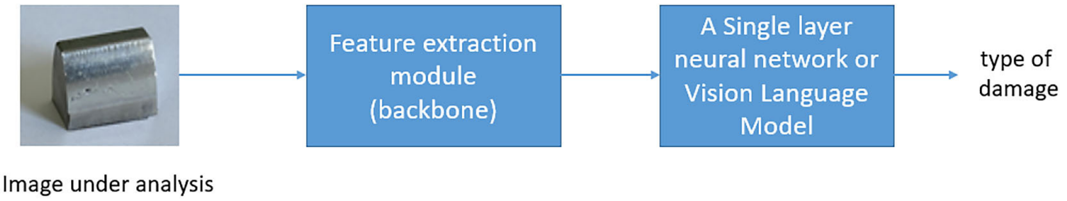
Feature extraction module (backbone)
A Single layer neural network or Vision Language Model
type of damage
Image under analysis

Figure 2. Structure of the tooth damage classifier

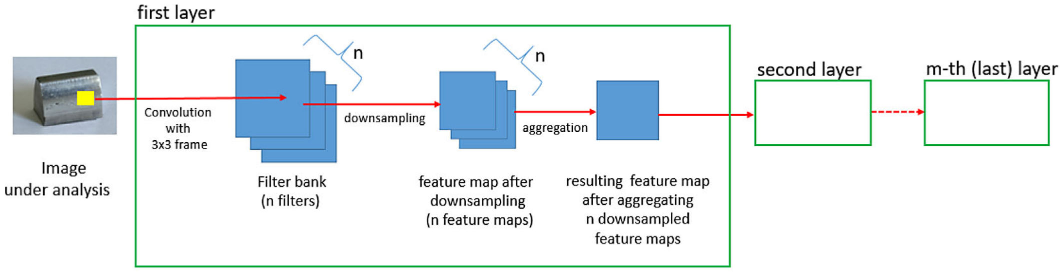

Figure 3. Structure of a convolutional neural network

Vision transformer backbone

This backbone leverages the characteristic features of language models originally applied in machine translation [31]. In sentence translation, the semantic meaning of an individual word depends on all or a subset of the other words present in the sentence. The degree of dependency of each word on the remaining sentence elements is determined by so-called “weights,” which are calculated using the Multi-Head Attention component [36]. This approach has been adapted for image classification, where the analyzed image is divided into a predefined number of patches. In this formulation, each patch can be treated analogously to a word, while the image corresponds to a sentence. Consequently, the global features describing the classified image can be effectively captured.

Hybrid backbone

The hybrid model integrates a convolutional neural network with a vision transformer [37]. In this model, the part represented by the vision transformer is preceded by a convolutional network. In the first stage, the convolutional network, using a filter bank, performs the extraction of local features present in the analyzed image. These features, represented as lines or curves, often do not allow for defining the relationships that describe complex objects, particularly when their meaning also depends on the surrounding background. For this reason, in the second stage, the local features obtained from the convolutional network are combined with other parts of the image (the background). This process is carried out through the use of the Multi-Head Attention mechanism implemented in the vision transformer. This approach allows for the combination of advantages from both architectures: the extraction of local features through the convolutional network and their aggregation via the vision transformer.

Detection based on multimodal LLM

Vision language models (VLMs) are multimodal generative models [38] that process and ‘understand’ textual data and images simultaneously. For both encoding and decoding, VLMs utilize neural networks and Vision Transformer Models. For the construction and decoding process, the two models need to be ‘glued’ with the networks that either project in one of two requirements the data into the VLMs, or utilize cross ‘attentional’ processes to bind them so that both models share the same VLM. For classification reasoning, VLMs are able to ‘see’ and ‘speak’ by processing and associating textual descriptions to images, with one of the major outputs being that they are able to label the image in question through image embeddings. This exceptional capability of VLMs permits them to perform tasks without any prior learning or instruction. Aside from classification tasks, VLMs are able to carry out tasks, such as question answering, generating captions, reasoning about images, and understanding spatial relationships. VLMs possess a rich variety of competencies in image understanding and reasoning.

EXPERIMENTAL RESULTS

This section presents the practical verification of the proposed classifier models. It consisted of the following stages:

- conducting a tribological test using the T-12U device developed at Łukasiewicz – Institute for Sustainable Technologies (Ł-ITEE),
- acquiring a series of images of pitting damage, which served as the training dataset for the classifier models,
- performing the training process of the classifier models,
- testing the trained models on test datasets,

- evaluating model performance based on effectiveness metrics.

Recording images of gear teeth after tribological tests

The analysis of the working surface condition of gear teeth was carried out on the surfaces of cylindrical gears from drive transmissions. The tribological tests concerning surface fatigue durability (pitting) were performed within the framework of the project financed by the National Centre for Research and Development (Poland): "Development of an innovative technology for the manufacture of toothed components with hybrid surface layers with a nanostructure base for the drive units of conveyors designed to be used in extreme operating conditions" (No. POIR.04.01.04-00-0064/15).

The tribological tests were conducted on the T-12U gear test rig at Łukasiewicz – Institute for Sustainable Technologies (Ł-ITEE), Poland (Figure 4).

The PT C/10/90 method was applied, based on FVA document No. 2/IV from 1997 [39]. Figure 5 presents the FZG C-PT test gears and the working surface of a tooth sectioned after fatigue testing. The test assembly consisted of a pair of cylindrical FZG C-PT gears. The tooth width of both gears was 14 mm. The pinion had 16 teeth, while the larger gear had 24 teeth.

The experiment involved the application of a specific lubricant to the gear pair operating under the conditions specified in Table 1, at a constant rotational speed, constant load, and a controlled oil temperature. The test continued until fatigue pitting occurred, which was evaluated visually by measuring the spalling area on the most damaged tooth of the pinion.

The tests continued until pitting appeared on the working surface of the pinion teeth. Damage assessment was performed visually, by measuring the spalling area on the most damaged tooth. The critical threshold was defined as 4.0% of the tooth working surface, corresponding to 5.0 mm². The test was terminated upon reaching this damage level or after 300 hours of rig operation, which corresponds to approximately 40 million fatigue cycles (~40 × 10^6 cycles).

In order to document the observed wear mechanisms on the working surfaces of the gear teeth, photographs were taken using an Apple iPhone 15 smartphone, equipped with an advanced camera capable of capturing high-resolution images. The photographs were obtained with the main camera of the device, featuring a 48 MP sensor, a 26 mm focal length, an *f*/1.6 aperture, and an optical image stabilization (OIS) system with automatic sensor stabilization. The use of this camera enabled the acquisition of highly detailed images, essential for the analysis and classification of damage occurring on the working surfaces of the teeth.

A review of the relevant literature confirmed the timeliness and necessity of research on the application of machine learning methods in tribological studies. Previous works have focused mainly on laboratory applications, including lubricant formulation, composite material development, and experimental design [40–45].

In the process of gear damage identification, several hundred recorded images of the working

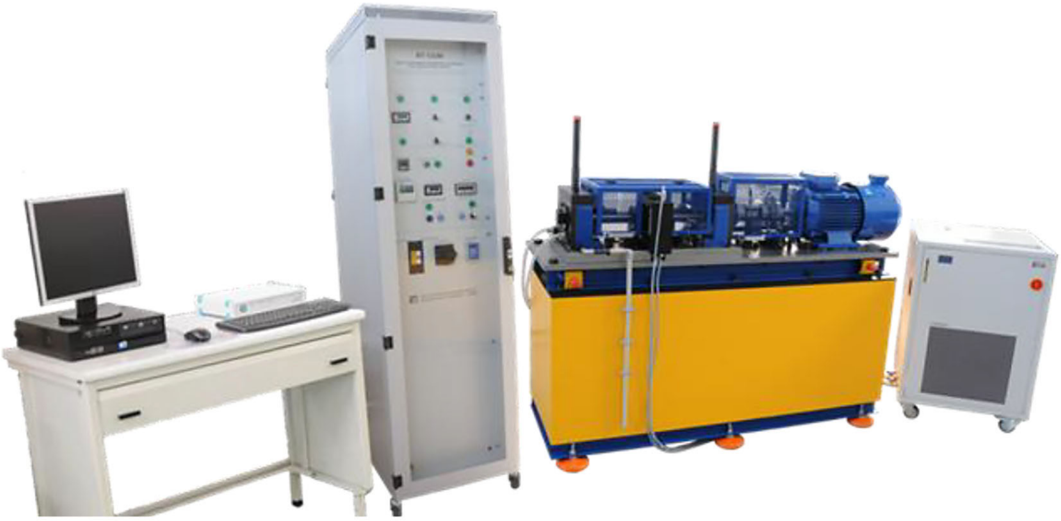

Figure 4. T-12U face-to-face gear test rig

Table 1. Test conditions for the PT C/10/90 method

Parameter	Value
Gear type	C-PT, tooth width of both gears (pinion and wheel): 14 mm
Motor speed	1450 rpm
Rolling circumferential speed	8.3 m/s
Lubrication	Immersion (oil quantity: 1.5 L)
Working-in conditions	
Working time	2 h
Load stage	6
Torque load	135.5 Nm
Main test conditions	
Working time	7 h
Load stage	10
Torque load	372.6 Nm
Maximum Hertzian pressure	1.8 GPa
Test oil temperature	90 °C
Maximum number of fatigue cycles	40 × 10^6 (corresponding to 300 h of operation)

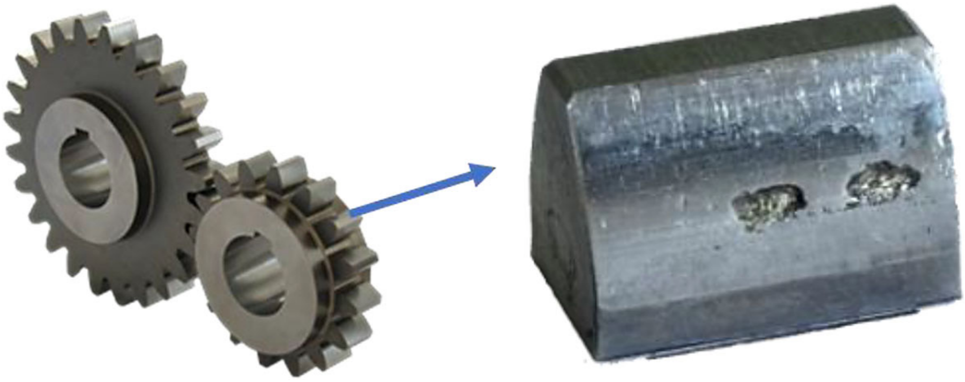

Figure 5. FZG C-PT test gear pair for pitting investigations and the working surface of a tooth sectioned after fatigue testing

tooth surfaces were analyzed and used as input data for the applied ML models. The collected image data were classified into subcategories corresponding to different surface conditions, distinguished on the basis of the dominant mechanism of fatigue crack initiation:

- A – No damage,
- B – Micropitting,
- C – Pitting (damage covering less than 4% of the tooth working surface),
- D – Macropitting (damage covering more than 4% of the tooth working surface).

Each subcategory contained several dozen images. The images obtained in this way were divided into six datasets using the 6-fold cross-validation method. Each dataset consisted of three subsets:

- *train_set*, used for model training,
- *valid_set*, used for model validation during training,
- *test_set*, used for testing the trained model.

Detection with classifier

Each of the models (containing a convolutional network, a vision transformer, or a hybrid model) was trained and tested on the same six datasets. To evaluate the effectiveness of the proposed models, the following metrics were used: accuracy, precision, recall, and F1-score.
The accuracy metric is defined as:

$$accuracy = \frac{correctly\ classified\ images}{total\ number\ of\ available\ images} \quad (1)$$

It specifies the percentage of correctly classified data in the total number of available test data.

The precision metric is defined separately for each class as follows:

$$precision = \frac{TP}{TP + FP} \tag{2}$$

where: *TP* (true positive) denotes the number of correctly classified images of a given class, and *FP* (false positive) denotes the number of images not belonging to that class which were classified by the model as belonging to it. This metric specifies the ratio of correctly classified images of a given class to the total number of images classified by the model as belonging to that class.

The recall metric is also defined separately for each class:

$$recall = \frac{TP}{TP + FN} \tag{3}$$

where: *TP* is defined in the same way as in (2), while *FN* (false negative) denotes the number of images belonging to a given class that were classified by the model as not belonging to that class. This metric specifies the ratio of correctly classified images of a given class to the total number of images of that class.

In addition to the above-mentioned metrics, the F1-score is also used, defined separately for each class:

$$F1score = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4}$$

This metric corresponds to the harmonic mean, which is an average that gives more weight to the smaller of the two values. This means that to achieve a high F1 score, both precision and recall must be high.

In the case of non-binary classifiers applied to multiple data classes (in this study, four classes), the following averaged values (macro-averaging) of precision, recall, and F1-score are also used:

$$precission_{aver} = \frac{\sum_{i=1}^{N} precission_i}{N} \tag{5}$$

$$recall_{aver} = \frac{\sum_{i=1}^{N} recall_i}{N} \tag{6}$$

$$F1score_{aver} = \frac{\sum_{i=1}^{N} F1_i}{N} \tag{7}$$

where: precision$_i$ denotes the precision, metric defined by (2) for the i-th class, recall$_i$ denotes the recall metric defined by (3) for the i-th class, $F1_i$ denotes the *F1-score* metric defined by (4) for the *i*-th class, and *N* denotes the number of available classes.

To compare the quality of the feature extractors, all models were trained with the same parameters:
- number of epochs = 300,
- learning rate (lr) = 0.0001,
- batch size = 4.

Classifier with ResNet18 backbone

For this experiment, the ResNet18 architecture was used as the backbone. To improve performance and speed up training, the authors have used a pretrained ResNet18 model that had been trained on the ImageNet dataset [46], applying transfer learning to adapt it to the task at hand. The core concept of ResNet relies on residual connections, which mitigate the vanishing gradient problem and allow for the effective training of deeper networks.

During the classifier construction process, the parameters of the ResNet18 backbone were frozen, and only the classification head was fine-tuned. This approach so-called transfer learning [47] reduced the computational costs of training and improved classification performance under conditions of limited dataset size.

The training procedure (train_set and valid_set subsets) and testing (test_set subset) were carried out separately for each of the six datasets. Table 2 presents the obtained performance metrics for the six test subsets (test_set), while Table 3 summarizes the averaged results across all test sets.

Classifier with vision Transformer backbone

Due to the complex structure of the backbone, its pretrained version – google/vit-base-patch16-384, trained on the ImageNet dataset [48] – was used. The training process of the classifier involved adjusting the parameters of the classification component (see Figure 3), while keeping the parameters of the pretrained backbone frozen. Training (subsets train_set and valid_set) and

Table 2. Performance metrics for the classifier with the convolutional network

No. set	Class	Precision	Avg. precision	Recall	Avg. recall	F1-score	Avg. F1-score
1	A	1.0	0.83	1.0	0.78	1.0	0.79
	B	0.64		0.78		0.70	
	C	0.67		0.67		0.67	
	D	1.0		0.67		0.80	
2	A	1.00	0.85	0.80	0.80	0.89	0.81
	B	0.69		1.00		0.82	
	C	0.90		0.75		0.82	
	D	0.80		0.67		0.73	
3	A	0.83	0.73	1.00	0.76	0.91	0.74
	B	0.78		0.78		0.78	
	C	0.64		0.58		0.61	
	D	0.67		0.67		0.67	
4	A	0.71	0.88	1.00	0.86	0.83	0.83
	B	1.00		0.44		0.62	
	C	0.80		1.00		0.89	
	D	1.00		1.00		1.00	
5	A	1.00	0.87	0.80	0.83	0.89	0.85
	B	0.88		0.78		0.82	
	C	0.79		0.92		0.85	
	D	0.83		0.83		0.83	
6	A	0.62	0.82	1.00	0.82	0.77	0.79
	B	0.64		0.78		0.70	
	C	1.00		0.67		0.80	
	D	1.00		0.83		0.91	

Table 3. Average performance metrics for the convolutional network model, calculated based on six test datasets

Precision	Recall	F1-score	Accuracy
0.83	0.808	0.802	0.792

testing (test_set subset) were carried out separately for each of the six datasets. Table 4 presents the performance metrics obtained for the six test datasets (test_set), while Table 5 shows their averaged values calculated on the basis of the aforementioned test datasets.

Classifier with hybrid backbone

Similar to the vision transformer model, a pretrained version of the hybrid model – faster_vit_0_224, trained on the ImageNet dataset [49] – was used. The training process of the classifier involved adjusting the parameters of the classification component (see Figure 3), while keeping the parameters of the pretrained hybrid model frozen. Table 6 presents the performance metrics obtained for the six test datasets (test_set), while Table 7 shows their averaged values calculated on the basis of the aforementioned test datasets.

Detection with multimodal LLM

Bootstrapping language-image pretraining (BLIP) is a multimodal model [50] that consists of Vision Transformer (ViT) for image encoding and a large language model (LLM) for text encoding as well as generation, and maps both modalities into a common vector space of representations. The modality and representation mapping in BLIP allow applied models for image–text matching, contrastive learning and captioning, which allows for the integration of classification and generative tasks. With regards to functionality, the noteworthy aspect of BLIP is the bootstrapping mechanism, which outputs candidate captions iteratively, maintains the most consistent candidates, and filters out lower-quality data, mitigating annotation noise and enhancing effectiveness of multimodal pretraining.

Table 4. Performance metrics for the classifier with the vision transformer

No. set	Class	Precision	Avg. precision	Recall	Avg. recall	F1-score	Avg. F1-score
1	A	0.83	0.80	0.71	0.78	0.77	0.79
	B	0.75		0.60		0.67	
	C	0.75		0.92		0.83	
	D	0.88		0.88		0.88	
2	A	1.00	0.87	1.00	0.84	1.00	0.83
	B	1.00		0.50		0.67	
	C	0.69		0.85		0.76	
	D	0.80		1.00		0.89	
3	A	1.00	0.92	1.00	0.91	1.00	0.91
	B	0.88		0.70		0.78	
	C	0.80		0.92		0.86	
	D	1.00		1.00		1.00	
4	A	0.78	0.87	1.00	0.89	0.88	0.87
	B	0.89		0.80		0.84	
	C	1.00		0.77		0.87	
	D	0.80		1.00		0.89	
5	A	0.86	0.90	0.86	0.90	0.86	0.90
	B	0.89		0.80		0.84	
	C	0.86		0.92		0.89	
	D	1.00		1.00		1.00	
6	A	1.00	0.94	1.00	0.89	1.00	0.91
	B	1.00		0.80		0.89	
	C	0.76		1.00		0.87	
	D	1.00		0.75		0.86	

Table 5. Average performance metrics for the vision transformer model, calculated based on six test datasets

Precision	Recall	F1-score	Accuracy
0.883	0.868	0.868	0.859

In the image classification training, BLIP uses prompt-based supervision, and associates with each image a textual template or prompt corresponding to its class. For the four classes in this study, images labeled as "A- No Damage", "B-Micropitting", " C-Pitting (small pits)", and " D-Pitting (large pits)" were respectively aligned with prompts such as "Clean metal surface with no visible damage or spots", "Metal surface with many tiny bright dots and micro-cracks scattered across a smooth area, forming a fine speckled texture", "Metal surface with a few small to medium pits scattered in a limited area, slightly shiny and contrasting with the smooth surrounding surface", and "Metal surface with a large, rough damaged area featuring deep, irregular pits and exposed texture that strongly contrasts with the surrounding material". The language model embeds these prompts and compares them to the visual embeddings in a common space, ensuring the visual and label align; since BLIP is capable of image–text matching but not reasoning. Thus, these prompts need to be relatively basic and unambiguous to allow BLIP to effectively learn. Table 8 presents the performance metrics obtained for the six test datasets (test_set), while Table 9 shows their average values calculated on the basis of the aforementioned test datasets.

COMPARISION OF A MODEL PERFORMANCE AND DISCUSSION

On the basis of the average performance metrics presented in Tables 3, 5, 7, and 9, it can be observed that the least effective model was the one with a convolutional network–based feature extractor, while the most effective was the model with a vision transformer. The lower performance of the convolutional model may result from the fact that

Table 6. Performance metrics for the classifier with the hybrid model.

No. set	Class	Precision	Avg. precision	Recall	Avg. recall	F1-score	Avg. F1-score
1	A	0.67	0.79	0.86	0.80	0.75	0.77
	B	0.80		0.40		0.53	
	C	0.80		0.92		0.86	
	D	0.89		1.00		0.94	
2	A	1.00	0.86	0.86	0.83	0.92	0.83
	B	1.00		0.70		0.82	
	C	0.71		0.77		0.74	
	D	0.73		1.00		0.84	
3	A	0.88	0.83	1.00	0.84	0.93	0.82
	B	0.83		0.50		0.62	
	C	0.73		0.85		0.79	
	D	0.89		1.00		0.94	
4	A	0.88	0.92	1.00	0.94	1.00	0.94
	B	0.90		0.90		0.90	
	C	1.00		0.85		0.85	
	D	0.89		1.00		1.00	
5	A	0.86	0.89	0.86	0.86	0.86	0.87
	B	0.90		0.90		0.90	
	C	0.80		0.92		0.86	
	D	1.00		0.75		0.86	
6	A	1.00	0.80	0.71	0.77	0.83	0.77
	B	0.73		0.80		0.76	
	C	0.75		0.69		0.72	
	D	0.70		0.88		0.78	

Table 7. Average performance metrics for the hybrid model, calculated based on six test datasets

Precision	Recall	F1-score	Accuracy
0.848	0.84	0.833	0.828

the extracted features describing the classified objects (defects) are local in nature. This is caused by the small filter frame size (3 × 3) used in the convolutional network. As a result, the extracted local features do not allow for encoding the complex dependencies present in the classified objects, which may translate into reduced algorithm effectiveness.

In contrast, the model employing a vision transformer for feature extraction, through the use of the multi-head attention mechanism and patch-based encoding analogous to that applied in translators, allows for the extraction and aggregation of global features of the detected objects. This contributes to the higher effectiveness of this algorithm compared to the convolutional model. On average, this difference amounts to approximately 6%.

The study also used a hybrid model, which is combining a vision transformer with a convolutional network. The performance obtained in this case lies between the results achieved with the vision transformer alone and the convolutional network alone.

The last model evaluated was the multimodal LLM. In this case, feature extraction was also based on a vision transformer, while the role of the classifier was performed not by a classical single-layer neural network, but by an LLM supported by prompts. The results obtained were only slightly inferior to those of the model consisting of the vision transformer with a classical single-layer neural classifier.

The optical analysis of the damage to the working surfaces of gear teeth after testing on the FZG test rig, performed using artificial intelligence algorithms for image damage processing, represents an alternative to traditional diagnostic methods. It enables an objective and repeatable assessment of the technical condition of gearbox components. The currently used method for assessing the degree

Table 8. Performance metrics for the model based on a multimodal language model

No. set	Class	Precision	Avg. precision	Recall	Avg. recall	F1-score	Avg. F1-score
1	A	1.00	0.81	1.00	0.80	1.00	0.80
	B	0.88		0.78		0.82	
	C	0.69		0.75		0.72	
	D	0.67		0.67		0.67	
2	A	0.80	0.89	0.80	0.90	0.80	0.89
	B	0.89		0.89		0.89	
	C	1.00		0.92		0.96	
	D	0.86		1.00		0.92	
3	A	1.00	0.80	1.00	0.81	1.00	0.80
	B	0.80		0.89		0.84	
	C	0.73		0.67		0.70	
	D	0.67		0.67		0.67	
4	A	1.00	0.89	0.80	0.86	0.89	0.87
	B	0.67		0.89		0.76	
	C	0.90		0.75		0.82	
	D	1.00		1.00		1.00	
5	A	1.00	0.90	0.80	0.83	0.89	0.85
	B	0.86		0.67		0.75	
	C	0.75		1.00		0.86	
	D	1.00		0.83		0.91	
6	A	0.71	0.79	1.00	0.83	0.83	0.78
	B	0.73		0.89		0.80	
	C	1.00		0.58		0.74	
	D	0.71		0.83		0.77	

Table 9. Average performance metrics for the model with a multimodal language model, calculated based on six test datasets

Precision	Recall	F1-score	Accuracy
0.847	0.838	0.832	0.823

of pitting is based primarily on manual visual inspection through a magnifying glass. Unfortunately, this approach is inefficient, time-consuming, and lacks precision. Moreover, it is characterized by subjectivity and depends on the experience of the person performing the evaluation.

Therefore, the development of an accurate, fast, and quantitative method for pitting detection using optical inspection supported by AI has a significant practical value. The application of machine learning and image analysis serves two main purposes:

1. Standardization of wear assessment: Automatic damage classification based on images eliminates the human factor, increases measurement repeatability and accuracy, as well as enables the standardization of test result documentation.
2. Extension of the testing device functionality: The T-12U device, developed and offered by Łukasiewicz–ITeE, is dedicated to users with varying levels of tribological expertise. The objective is to enable them to independently assess test results. For this reason, a database of damage images and algorithms for their automatic detection and classification are being developed.

In terms of practical applications under real operating conditions, the described method is not intended to compete with non-invasive diagnostic techniques (e.g., vibration analysis, acoustic emission, or MCSA), but rather to complement them in laboratory and testing environments, where disassembly is part of the experimental procedure – such as in the evaluation of lubricants, protective coatings, or pinion materials. In these applications, optical surface inspection after the completion of the test constitutes a key stage of fatigue assessment rather than merely the detection of damage occurrence.

The collected damage images and operational data will be used in subsequent stages of the research to develop predictive wear models, which may also be applied in online systems (e.g., as reference data for vibration sensor readings).

CONCLUSIONS

The article presented and compared four classifier models in terms of their application to the classification of pitting damage on the working surface of cylindrical gear teeth. The first three models used, respectively, a convolutional neural network, a vision transformer, and a hybrid model for feature extraction, while a single-layer neural network was used for classification. The fourth model utilized a vision transformer for feature extraction and a multimodal large language model (LLM) for classification.

The obtained results indicate that the best performance was achieved by the model employing the vision transformer together with a neural classifier, with precision equals 0.883, recall amounts to 0.868, and accuracy is equal to 0.859. Transfer learning was applied in model training, which improved defect classification performance despite the limited volume of available defect data. The acquired results demonstrate great potential and provide a foundation for the continuation as well as expansion of research in the classification and prediction of damage on the working surface of gear teeth following tribological tests.

In summary, the conducted research contributes to the following:

- the classification and quantification of gear tooth surface wear based on image analysis,
- the automation of the gearbox technical condition assessment process,

The establishment of a foundation for the further development of non-contact damage prediction methods using image data and machine learning techniques.

Acknowledgements

The authors would like to thank Prof. A. N. Wieczorek from the Faculty of Mining, Safety Engineering and Industrial Automation, Silesian University of Technology, and Prof. R. Michalczewski from the Tribology Centre, Łukasiewicz Research Network – Institute for Sustainable Technologies, for providing the gears manufactured within the framework of research projects, No. TECHMATSTRATEG-III/0028/2019 and No. POIR.04.01.04-00-0064/15 ,as well as M. Gibała from the Tribology Centre Ł-ITeE for conducting the fatigue tests using the FZG method.

REFERENCES

1. Tuszyński, W.; Michalczewski, R.; Piekoszewski, W.; Szczerek, M. Effect of ageing automotive gear oils on scuffing and pitting. Tribol. Int. 2008, 41, 875–888. https://doi.org/10.1016/j.triboint.2007.12.010
2. Szczerek M. et al. The correlated selection of a thin coating and gear oil to increase the resistance of 18CrNiMo7-6 gears to pitting-Part 1 Proceedings of Asia International Conference on Tribology 2018, 293–295. Malaysian Tribology, Society, https://www.mitc2020.mytribos.org/resources/Proceeding/Asiatrib2018/Asiatrib-part-1.pdf
3. Michalczewski, R.; Kalbarczyk, M.; Mańkowska-Snopczyńska, A.; Osuch-Słomka, E.; Piekoszewski, W.; Snarski-Adamski, A.; Szczerek, M.; Tuszyński, W.; Wulczyński, J.; Wieczorek, A. The effect of a gear oil on abrasion, scuffing, and pitting of the DLC-coated 18CrNiMo7-6 steel. Coatings 2019, 9, 2. https://doi.org/10.3390/coatings9010002
4. Barth, Y.J.; Sagraloff, N.; Egger, G.; Tobie, T.; Stahl, K. Investigations on Ways to Improve the Scuffing and Wear Behavior of Oil-Free Water-Based Lubricants for Gear Applications. J. Tribol. 2024, 146, 054601. [CrossRef]
5. Jao T.C. et al.: Influence of Surface roughness on gear pitting behaviour. Gear Technology 2006, 31–38.
6. Amir G. Mustafayev, Chingiz R. Nasirov.: A study of factors affecting wear and destruction of teeth in gear mechanisms. Nafta-Gaz 2023, 9, 604–610, https://doi.org/10.18668/NG.2023.09.06
7. Hoehn B.-R., Oster P., Schedl U.: Pitting load capacity test on the FZG gear test rig with load-spectra and one-stage investigations. Tribotest Journal 1999, 5, 417–430.
8. Kalin M., Vižintin J. The Tribological Performance of DLC-Coated Gears Lubricated with Biodegradable Oil in Various Pinion/Gear Material Combinations. Wear 2005, 259, 1270–1280. https://doi.org/10.1016/j.wear.2005.02.028
9. Hoehn B.-R., Michaelis K: Influence of oil temperature on gear failures. Tribology International 2004, 37, 103–109. https://doi.org/10.1016/S0301-679X(03)00047-1
10. Michaelis K., Hoehn B.-R., Oster P.: Influence of lubricant on gear failures – test method and application

to gearboxes in practice. Tribotest Journal 2004, 11, 43–56. https://doi.org/10.1002/tt.3020110105

11. Davis J.R. Gear materilas, properties and manufacture. 2005, ASM International Materials Park, OH 44073-0002, USA.
12. Fan Q., et al: Gear tooth surface damage diagnosis based on analyzing the vibration signal of an individual gear tooth. Advances in Mechanical Engineering, 2017, 9(6), 1–14, https://doi.org/10.1177/1687814017704356
13. Zieja M., Golda P., Zokowski M., Majewski P.: Vibroacoustic technique for the fault diagnosis in agear transmission of a military helicopter. International Journal of Vibroengineering, 2017, 19, https://doi.org/10.21595/jve.2017.18401
14. 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 2023, 23, 287. https://doi.org/10.3390/s23010287
15. Łazarz B., Wojnar G., Figlus T.: Comparison of the efficiency of selected vibration measures used in the diagnosis of complex cases of tooth gear damage. Diagnostyka, 2007, 4, 11–18.
16. Zhang S., Zhou J., Wang E., et al.: State of the art on vibration signal processing towards data-driven gear fault diagnosis. IET Collab. Intell. Manuf. 2022, 4, 249–266, https://doi.org/10.1049/cim2.12064
17. Wojnar G., Burdzik R., Wieczorek A.N., Konieczny Ł.: Multidimensional data interpretation of vibration signals registered in different locations for system condition monitoring of a three-stage gear transmission operating under difficult conditions. Sensors 2021, 21, 7808. https://doi.org/10.3390/s21237808
18. Jabłonski A., Dworakowski Z., Dziedziech K., Chaari F.: Vibration-based diagnostics of epicyclic gearboxes – From classical to soft-computing methods. Measurement, 2019, 147, 106811. https://doi.org/10.1016/j.measurement.2019.07.039
19. Wersa, E., Rak, Z., Seweryn, A. Badania trwałości zmęczeniowej kół zębatych. Rozdział w książce: Seweryn, A. (red.), VII Międzynarodowe Sympozjum Mechaniki Materiałów i Konstrukcji, Augustów, 3–6 czerwca 2013. Białystok: Oficyna Wydawnicza Politechniki Białostockiej, 2013, 123–130.
20. Wieczorek N.A. Effect of construction changes in the teeth of a gear transmission on acoustic properties. International Journal of Occupational and Ergonomics. 2012, 18, 499–507, https://doi.org/10.1080/10803548.2012.11076956
21. Wieczorek, A.N. Ensuring appropriate conditions for lubrication of gear transmissions as a priority for maintenance services in industrial transport. Scientific Journal of Silesian University of Technology. Series Transport. 2021, 111, 193-204. https://doi.org/10.20858/sjsutst.2021.111.17
22. Fidali M.: Metody diagnostyki maszyn i urządzeń w predykcyjnym utrzymaniu ruchu. 2020r, Wyd. Elamed Media Group.
23. Zimroz R.: Diagnozowanie przekładni zębatych w układach napędowych przenosników tasmowych – ocean stanu współpracy kół zębatych w warunkach zmiennego obciążenia. Transport Prezmysłowy. 2007, 4, 10–16.
24. Noga, S., Markowski, T. Analiza drgań własnych przekładni zębatej małej mocy. Advances in Mechanical and Materials Engineering. Zeszyty Naukowe Politechniki Rzeszowskiej. RUTMech, t. XXXIV, z. 89 (4/17), 2017, 517–528.
25. Abhishree T. M., Latha J, Manikantan K, Ramachandran S, Face recognition using gabor filter based feature extraction with anisotropic diffusion as a pre-processing technique, Procedia Computer Science 2015, 45, 312–321.
26. Takumi Kobayashi, BoF meets HOG: Feature Extraction based on Histograms of Oriented p.d.f Gradients for Image Classification, 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA.
27. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with Deep Convolutional Neural Networks. Adv. Neural Inf. Process. Syst. 2012, 14, 1097–1105.
28. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. J. Big Data 2021, 8, 1–74.
29. Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C.J.; Müller, K.R. Explaining deep neural networks and beyond: a review of methods and applications. Proc. IEEE 2021, 109, 247–278.
30. Shahzad, T., Mazhar, T., Tariq, M.U. et al. A comprehensive review of large language models: issues and solutions in learning environments. Discov Sustain 2025, 6, 27. https://doi.org/10.1007/s43621-025-00815-8
31. Dosovitskiy, A., et al. “An image is worth 16 × 16 words: Transformers for image recognition at scale.” arXiv preprint arXiv:2010.11929 (2020).
32. Khan, Asifullah, et al. “A survey of the vision transformers and their CNN-transformer based variants.” Artificial Intelligence Review 56. Suppl 3. 2023, 2917–2970.
33. Kiran, B. Ravi, et al. “Deep reinforcement learning for autonomous driving: A survey.” IEEE transactions on intelligent transportation systems 2021, 23(6), 4909–4926.

34. Bojarczak, P.; Nowakowski, W. Application of deep learning networks to segmentation of surface of railway tracks. Sensors 2021, 21, 4065. https://doi.org/10.3390/s21124065

35. Pritt, Mark, and Gary Chern. "Satellite image classification with deep learning." 2017 IEEE applied imagery pattern recognition workshop (AIPR). IEEE, 2017.

36. Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 2017, 30.

37. Hatamizadeh, Ali, et al. "Fastervit: Fast vision transformers with hierarchical attention." arXiv preprint arXiv:2306.06189. 2023.

38. Noyan M., Beeching E., Vision language models explained. Hugging Face Blog 2024, https://huggingface.co/blog/vlms

39. FVA Information Sheet. Influence of lubricant on the pitting capacity of case carburized gear in load-spectra and single-stage-investigation. No 2/IV, 1997.

40. Abhishek T. Sose, et al., A review of recent advances and applications of machine learning in tribology. Phys. Chem. Chem. Phys., 2023, 25, 4408, https://doi.org/10.1039/d2cp03692d

41. Maheshwera U., Paturi R., The role of machine learning in tribology. A systematic review. Archives of Computational Methods in Engineering 2023, 30, 1345–1397, https://doi.org/10.1007/s11831-022-09841-5

42. Raj Shah et al. Artificial Intelligence and Machine Learning in Tribology: Selected Case Studies and Overall Potential. Advanced Engineering Materials, 2025, article 2401944, https://doi.org/10.1002/adem.202401944

43. Yin N., Yang P., Liu S.,Pan S., Zhang Z., AI for tribology: Present and future. Friction 2024, 12(6), 1060–1097, https://doi.org/10.1007/s40544-024-0879-2

44. Tremmel S., Marian M., Machine learning in tribology – more than buzzwords? Lubricants 2022, 10(4), 68; https://doi.org/10.3390/lubricants10040068

45. Rosenkranz A., Marian M., Profito J.F., et al. The use of artificial intelligence in tribology—a perspective. Lubricants 2021, 9(1), 2. https://doi.org/10.3390/lubricants9010002

46. ResNet-18 from Deep Residual Learning for Image Recognition. Available at: https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet18, 2017 (accessed 8 September 2025).

47. Iman, M.; Arabnia, H.R.; Rasheed, K. A review of deep transfer learning and recent advancements. Technologies 2023, 11, 40.

48. Vision Transformer ViT model. Available at: https://www.huggingface.co/google/vit-base-patch16-384, 2021. (accessed 8 September 2025).

49. FasterViT: Fast Vision Transformers with Hierarchical Attention. Available at: https://www.huggingface.co/nvidia/FasterViT, 2022. (accessed 8 September 2025).

50. Li J., Hoi S., BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. Salesforce Blog. Available at: https://www.salesforce.com/blog/blip-bootstrapping-language-image-pretraining/, 2022. (accessed 8 September 2025).