Mechanical Systems and Signal Processing 224 (2025) 112164

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

journal homepage: www.elsevier.com/locate/ymssp

Invited for the Special Issue in Honor of Professor John Mottershead

Rate of change of torque for gear tooth damage detection

George Hunt-Pain [a], Ryan Walker [b,*], Ben Cahill [a], Alastair Clarke [a]

[a] School of Engineering, Cardiff University, Queens Buildings, 14-17 The Parade, Cardiff, CF243AA
[b] School of Engineering Science, Oxford University, Thom Building, Oxford, OX1 3PJ

ARTICLE INFO

Communicated by Pietro Borghesani

ABSTRACT

This paper presents an investigation into the applicability of a magnetoelastic rate of change of torque (ROC) sensor to condition monitoring of gearbox tooth damage, using a recirculating torque test gearbox test facility. ROC sensors have significant potential in condition monitoring of rotating machinery, due to direct signal paths between the sensors and the damaged region. To assess this, data has been collected across a wide range of operating conditions for tooth bends as small as 2.4 µm. A number of industry-standard analysis techniques have then been applied. This includes a variety of algorithms (based around time-synchronous averaging), wavelet transforms, neural network classification, shapely additive explanations (SHAP) and deep learning. It is concluded that the ROC sensor demonstrates significant potential in condition monitoring of rotating machinery, worthy of further study.

1. Introduction

Gearboxes are mechanical components that transmit power from source to destination and are comprised most critically of gears, shafts, and bearings, with one of the most common applications being the in the automobile. Modern gearboxes convert power with over 99 % efficiency and resultantly torque and speed are inversely proportional to each other [1]. Naturally, the result is that torque transfer is inherent to the operation of every gearbox. Faults in gearboxes have several key causes including overloading conditions, poor lubrication conditions, wear, and forms of corrosion [2]. In high-performance gearboxes in applications such as motorsport, the likelihood for failure shifts to plastic-deformation from transient overloading conditions [3]. This failure is more prevalent due to the desire for light-weight compact designs and high load cases. Other failure mechanisms are less prevalent, particularly wear, as gearboxes are replaced relatively frequently compared to other industrial machinery. For instance, in a popular high-performance application of motor-sport, Formula 1 gearboxes can be changed every 6 races [4]. In such environments, being able to extend the life of a gearbox reliably or carry out preventative maintenance can have significant implications to team outcomes.

A rate of change of torque sensor directly measures the electromotive force (EMF), in volts, of a coil that encircles a thin ferromagnetic ring which is rigidly mounted to a shaft transmitting a torque. The output voltage is directly proportional to the instantaneous Rate of Change of Torque (ROC). The direct measurement of the Rate of Change of Torque has advantages over the measurement of torque alone and then differentiating with respect to time, as it possesses greater immediacy and sensitivity to small and rapid torque changes, and inherently a less-noisy measurement than a differentiated torque signal. This enables data to be captured for a greater bandwidth with reduced noise than conventional magneto-elastic torque measuring methods [5].

The application of ROC sensors to condition monitoring of gearboxes and bearings is well-suited compared to typical contact-based

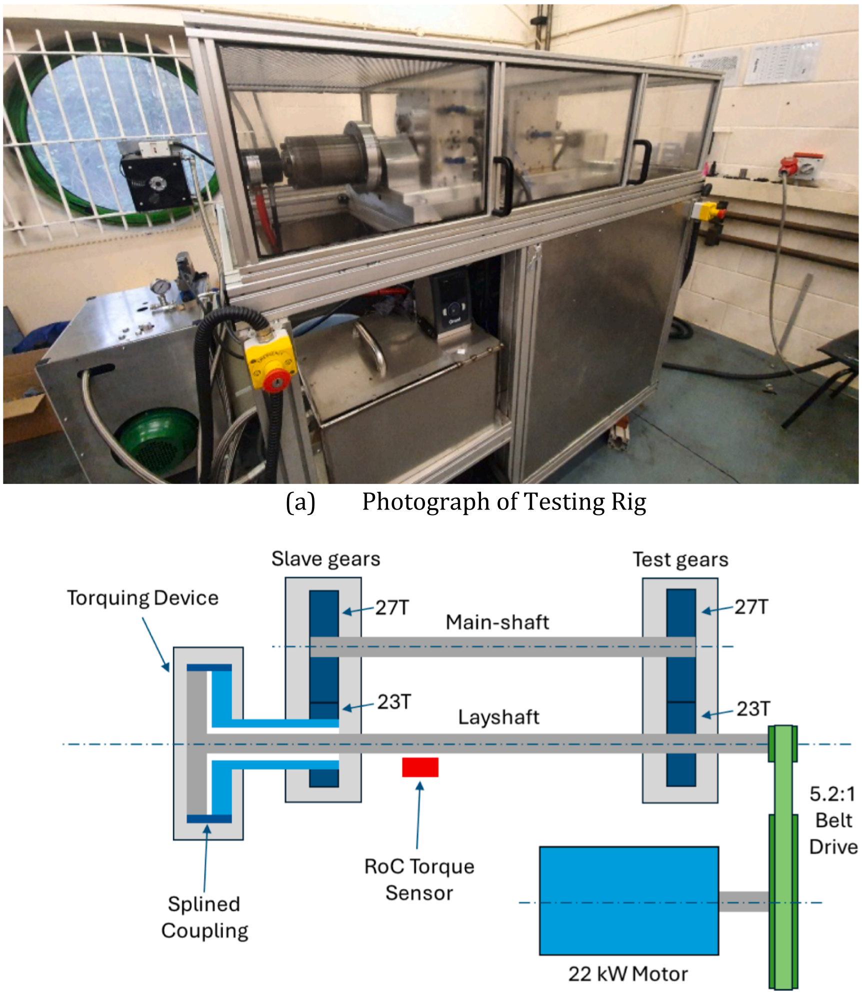
[*] Corresponding author.
E-mail address: ryan.walker@eng.ox.ac.uk (R. Walker).

https://doi.org/10.1016/j.ymssp.2024.112164
Received 15 February 2024; Received in revised form 26 September 2024; Accepted 14 November 2024
Available online 25 November 2024

vibration or acoustic emission (AE) techniques. The placement of these traditional sensors (typically accelerometers) is paramount to their clarity as signal detection can be hampered by the attenuation, inherent damping and frequency response of the surrounding structures [6]. Ideally these sensors would be mounted directly to the components that require monitoring. However, this is often not plausible in rotating components and a tightly packaged design assembly — meaning that often these sensors are displaced to casings or other structures which leads to the reduction of their sensitivity. The non-contact nature of ROC sensors allows relative ease of placement and the losses through the transmission of torque are low compared to the torque being transmitted. This allows the sensor to be placed anywhere within the torque path and still be sensitive to torque variations.

With this in mind, the work that follows is intended to investigate the applicability of a rate of change of torque sensor for condition monitoring in high-performance gearboxes. It should be noted that whilst the authors consider this as a novel contribution to the field, there is a huge body existing of work into condition monitoring for rotating machinery more generally (often, but not exclusively, using accelerometers as the sensing methods). Much of the recent focus of condition monitoring for rotating machinery is in the context of wind turbines. Comprehensive reviews on the topic have been performed by Wang et al and Tiboni et al [7,8]. More broadly, the work on condition monitoring covering transport applications and beyond summarized in the book by Randall is an invaluable resource [9], as is the original work by Bently [10]. Acoustic emission techniques have also been extensively used for condition monitoring, with promising results, however some similar challenges to purely vibration-based techniques [11–13]. Specifically on the topic of gear faults, a number of researchers across the globe continue to study in this area. Gaining detailed understandings of tooth mechanics can aid drivetrain efficiency in addition to the requirements for condition monitoring [14–16]. In the automotive sphere, interest has been renewed by the drive for electric vehicle efficiency [17–19], which is one potential application for the rate of change of torque sensor studied for this work.

Condition based monitoring of rotating machinery components by conventional methods requires continuous manual analysis and


(a) Photograph of Testing Rig

(b) Schematic of Testing Rig Internal components.

Fig. 1. Image and schematic diagram of the recirculating torque test rig.

re-configuration to ensure early detection of abnormal operating conditions and faults. This is due to the time varying characteristics caused by wear or maintenance [20]. As a result, these conventional methods are limited in their efficiency and scalability. These issues, in principle, can be addressed by machine learning as algorithms have been developed to produce generalising, robust, predictive models and perform automated feature extraction [21].

The rise of machine learning has been of particular interest in the world of condition monitoring for rotating machinery. A number of review papers are available which cover this in some depth [22–24]. Some particular work of relevance to this paper is the use of Shapely Additive exPlanations (SHAP), which have been used for a range of condition monitoring applications, including the recent work by Movsessian et al [25], Herwig [26] and Brito [27].

This paper presents an investigation into the novel application of ROC techniques to monitoring high performance gearbox components. It also compares a range of conventional and machine learning-based analysis methods.

2. Experimental methods

The test rig used in this experiment is a recirculating torque design and a schematic is shown in Fig. 1. Many of these types of test apparatus found in academia and industry follow the FZG approach. A manual torquing device was used to apply a load statically to the gearbox, where one shaft was locked in place and the other was twisted by a weighted loading arm, as shown by the “Torquing Device” in Fig. 1 (b). The shafts were then locked together in the loaded position with a splined coupling. Contrary to a typical gearbox where load is delivered from an engine via an input shaft and power exits via the final drive — in this rig the torque is recirculated with a reactive (slave) gear-pair, also shown in the Fig. 1 schematic, with torque being provided via the torque head with the drive motor only providing sufficient torque to overcome the parasitic (frictional and windage) losses. The ROC sensor was positioned within the torque loop – equivalent to being mounted on, for example, the input shaft of an automotive gearbox. The general specifications of the test rig were:

- Maximum achievable input shaft speed 15′000 RPM.
- Maximum sustainable torque of 600Nm, Maximum peak torque of 1000Nm.
- Controllable oil temperature.
- Individual flow control of bearing lubrication

The system is driven via a 22 kW 2-pole 3-phase AC electric motor with a 16 rib polyvee belt system with a 5.2:1 ratio allowing a maximum rig speed of 15,000 rpm. The pulley drives the layshaft of the gearbox that rotates in the direction of the driving torque. A gear ratio of 23/27 (layshaft/mainshaft) was used for all tests in this work.

Data was collected at *l*00 kHz using a data acquisition device from National Instruments at input speeds from l500 rpm to 15000 rpm in l500 rpm increments and torques from 50 Nm to 300 Nm at 50 Nm increments. The acquisition system was set to capture two bursts, of 50 rotations each, at every speed and torque level, resulting in 60 data points for each gear set.

4 gear-sets of varying damage were used in the production of this dataset. This includes a healthy set and three damaged sets. The damaged sets had one gear tooth on the main-gear bent a number of microns away from normal in a single event on a load machine. This results in a tooth spacing error in the direction that a plastic deformation would occur due to an overloading condition seen in a real world gear defect. The three damaged gears resultantly had tooth spacing errors of 2.4 μm, 8.7 μm and 14.7 μm respectively. Plots

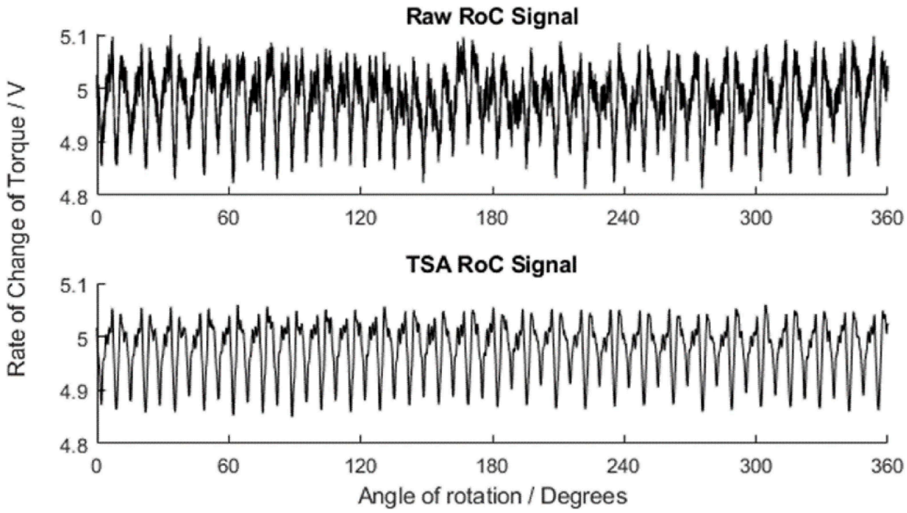


Fig. 2. Precision Klingelnberg gear measurement plots of tooth-to-tooth spacing for (a) Healthy, (b) 2.4 μm gear, (c) 8.7 μm Gear and (d) 14.7 μm Gear.

of the left flank tooth-to-tooth spacing from precision Klingelnberg Gear measurements are shown below in Fig. 2.

It should be noted that the 2.4 μm defect is within the allowable manufacturing tolerances of these gears, however often pitch spacing errors are less than this with the average <1 μm. Small amounts of elastic deformation is not uncommon, as gear teeth under load will experience some form of it routinely during standard operations. This is why tooth geometry modification such as tip-relief are routinely employed in gear design. However, plastic deformation caused by overload beyond the designed load capacity of the gears quickly results in efficiency loses and can lead to reciprocating cycles that can generate new damage as well as furthering the initial damage. If undetected this can lead to catastrophic gear failure. The main reason for concern when looking at tooth bends is their effect on the meshing cycle. As the tip of the gear tooth is no longer evenly spaced with those preceding and following it, the tooth no longer matches the meshing cycle exactly and will engage later than expected, with the following tooth prematurely engaging too. This has the same effect as suddenly introducing additional backlash into the system, where the system will try and 'catch up' across this new space. This causes shock loading to occur on the already deformed tooth as well its mating counterpart. The sudden impact caused by a late engagement results in a larger normal force applied to the gear tooth, this force will generate a higher contact pressure and bending stress.

The remainder of this work will be presented in three sections. Section 3 will look at traditional condition monitoring techniques including signal denoising, manual feature extraction and time–frequency domain representations. Section 4 will apply a range of Neural Network architectures to differentiate between damaged and healthy gears with the manually extracted features. Section 5 will apply Deep Learning and automated feature extraction to the ROC data.

3. Traditional condition monitoring

In the raw data recorded, there is a large amount of variation of the quality and noise in the time signal. This means no consistent trend could be seen in the meshing events. To combat this, simple Time Synchronous Averaging (TSA) was used, where the time signal was split into individual rotations and then averaged. This means the consistent events in the meshing cycle are maintained and the random noise variations are reduced. In these TSA signals, it is often possible to plainly distinguish the larger defects, however, the 2.4-μm bend can only be identified in a small number of cases, furthermore when all defects are detectable visually, they are not always differentiable from each other. A comparison of the raw data against the TSA signals for a healthy gear-set is shown below in Fig. 2.

From these TSA signals, it is clear to see individual meshing events in a saw tooth pattern, including two types of ROC spike, one increases rapidly, then declines more slowly, whilst the other slowly climbs to a spike then rapidly decreases. It is thought a pulse is generated at each tooth's loading and unloading events. Hence there are twice as many pulses in the pattern as there are teeth. One can easily imagine a torque pulse being generated as two teeth first mesh and the loading split between two teeth. This can also be extended as another torque pulse is generated when the tooth exits the meshing and torque loading changes from two teeth to one. The distinct pattern of these signals; equal spacing, magnitude and direction generated from torque pulses comes from the rate of change of torque characteristics of the sensor.

Table 1
Table of traditional metrics used in analysis.

Name	Description	Damage-Type	Equation
Root-Mean-Square (RMS)	Represents the amplitude and energy of a signal	General fault progression	$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x)}$
Kurtosis (K)	Fourth order normalised moment, representing leaked-ness of a signal	Breakage, wear	$K = \frac{N\sum_{i=1}^{N}(x_i - \overline{x})^4}{(\sum_{i=1}^{N}(x_i - \overline{x})^2)^2}$
Energy-Operator (EO)	Normalised Kurtosis calculated for a time averaged signal	Scuffing, severe pitting	$K = \frac{N\sum_{i=1}^{N}(re_i - \overline{re})^4}{(\sum_{i=1}^{N}(re_i - \overline{re})^2)^2}$
Zero Order Figure of Merit (FM0)	Ratio of peak-to-peak amplitudes and harmonics	Wear, scuffing, pitting and tooth bending	$FM0 = \frac{PP_x}{\sum^{N} P_N}$
M6A	Normalised sixth order moment of difference signal	Surface damage indicator	$K = \frac{N\sum_{i=1}^{N}(d_i - \overline{d})^6}{(\sum_{i=1}^{N}(d_i - \overline{d})^2)^3}$
M8A	Normalised eight order moment of difference signal	Surface damage indicator	$K = \frac{N\sum_{i=1}^{N}(d_i - \overline{d})^8}{(\sum_{i=1}^{N}(d_i - \overline{d})^2)^4}$

Cahill's critical analysis of this data [28] concluded there were no clear single value metrics (shown in Table 1) that gave a major distinction between healthy and damaged data, this is regarding both the raw ROC and TSA signals. However general trends in these metrics were seen – for instance feature metric values were higher on average for higher levels of damage. There are also obvious effects caused by the changes to speed and torque. These are clearly highlighted in the M6A feature plot in Fig. 3. An increased torque, means larger M6A values on average while an increased speed does not always. The effect of increased torque is to be expected as more torque in the system will generate correspondingly larger rates of change of torque. The speed effects meanwhile are somewhat unexpected as it was expected that the faster speeds hence see quicker changes in torque hence increasing the rate of change, hence M6A would be expected to increase with speed). Perhaps the fixed sampling rate is influential here as the comparisons of metric are statistically in-equivalent in terms of samples per event.

The metrics shown in Table 1 were originally developed for use with vibration data, so it is unsurprising that blindly implementing these metrics to ROC signals achieves unreliable results. As the ROC data is inherently peaked – it would make sense to look at metrics that capture this. M6A, shown in Fig. 3, shows there is a general trend of increased feature value for increasing levels of damage. This could mean the M6A is a predictor of damage in ROC signals, but there are many outliers to this rule and it would give many false negatives if this was used as a damage indicator alone.

A range of frequency based metrics were developed to try and capture the meshing disorder introduced by a damaged gear tooth. It is expected that a healthy gear-set would exhibit a consistent meshing frequency, resulting in a single clear peak in the frequency domain at that frequency and harmonics above that. So it follows that for a damaged gear tooth it could be expected that the meshing cycle is disrupted as a result of instantaneous backlash being introduced into the system. The disorder in meshing should then manifest in a disruption to that meshing frequency and show broader peaks in the frequency domain.

The frequency based metrics summed the normalised energy in octave bands calculated via Fast Fourier Transform (FFT) where the first octave is centered on the shaft frequency of the damaged gear and increasing bands were centered on multiples thereof. The FFT was calculated on the envelope signal from the TSA of the raw ROC data with a periodicity related to the shaft speed, thus reducing random noise and keeping persistent features related to gear tooth damage. The resulting features summed energy in 7 octaves and thus comprised 7 features, Q1, Q2, Q4, Q8, Q16, Q32 and Q64 – named after the multiples of shaft orders at which each octave band starts. Q8 hence spans from the 8th-16th shaft orders and the dominant meshing frequency, 27x the shaft frequency will fall in the Q16 band. An example of the octaves in the frequency domain is shown in Fig. 4 for all 4 gear-sets.Fig. 5.

The general trend seen in the octave bands is that as the damage increases, the key frequency spikes are more distributed across each octave as the individual particular frequencies become less distinct and more blurred. This adheres to the observation by Stewart [29] that gear tooth damage increases the amplitude of the side-bands about the regular meshing components. Thus summing this energy in each octave can give an measure analogous to this disorder. However, no trend holds true for all cases. It is also noted that this disruption seen occurs in larger magnitude at the lower frequency bands.

The limitation of this method is that, although the disorder in the meshing is visible in the frequency domain, capturing it entirely with a single metric is incredibly complex and many interacting components of the signal manifest in complex variations. The primary interaction of a gear defect is a very impulsive event with a high energy level that then influences the dynamic system and resulting harmonic interactions – so it is unsurprising that no clear metric or trend holds true for every sample.

A time–frequency representation of this signal allows a visual interpretation of these impulsive events caused by gear tooth bends. A Continuous Wavelet Trans- form (CWT) provides instantaneous time–frequency representations and will accurately capture impulsive events that span many frequencies. Wavelets are well suited to these non-stationary applications as they are short in time and can instantaneously capture frequency and time events – overcoming the traditional problems with Fourier Transforms. Fig. 6

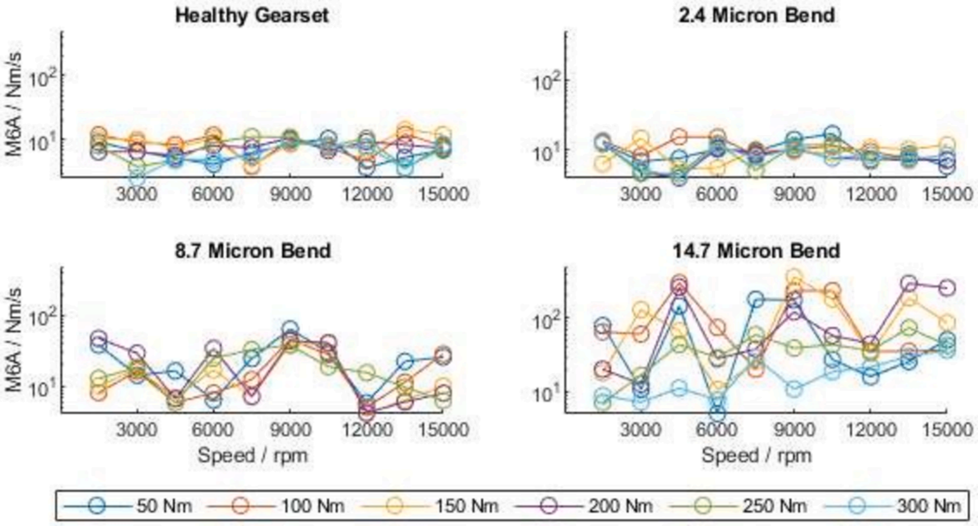


Fig. 3. M6A for all load and speed cases for the TSA across the range of defects.

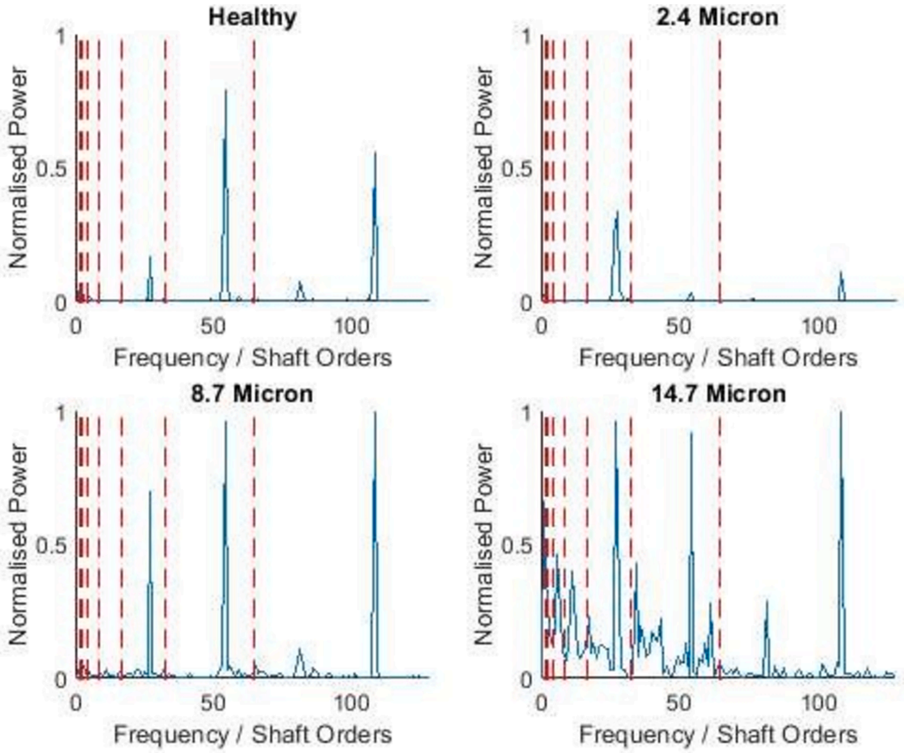


Fig. 4. FFT plot with octave metric bands for 4500 rpm, 200Nm.

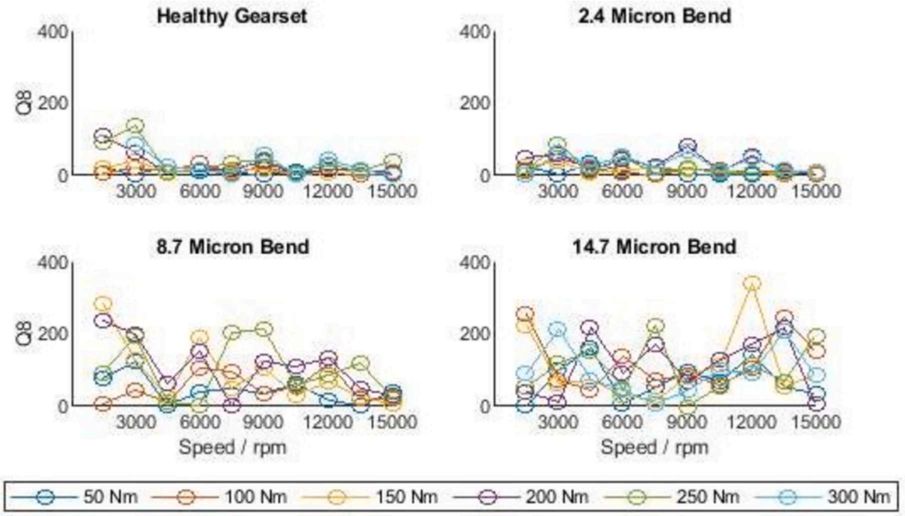


Fig. 5. Q8 Metric comparison for all damage levels.

shows a normalised CWT plot for the largest defect tested; a TSA signal was used to remove the random noise from the signal. In this example, the defect can be clearly seen at approximately 30 degrees and spans a broad range of frequencies. The CWT also shows individual meshing events of mostly lower intensity with 27 individual striations visible at intervals across the 360 degrees of the Mainshaft Gear.Fig. 7.

Since this is a novel dataset presented in this paper, there are no existing benchmarks for accuracy. Therefore, we used a human rater to provide a comprehensive comparison to see if the produced models matched or exceeded this initial benchmark. To obtain a human level baseline of accuracy we look at the CWT plots – the 4 levels of damage were assessed at each sample point and were placed by the authors into one of five categories, identified in Table 2. The highest category represents a human being able to clearly

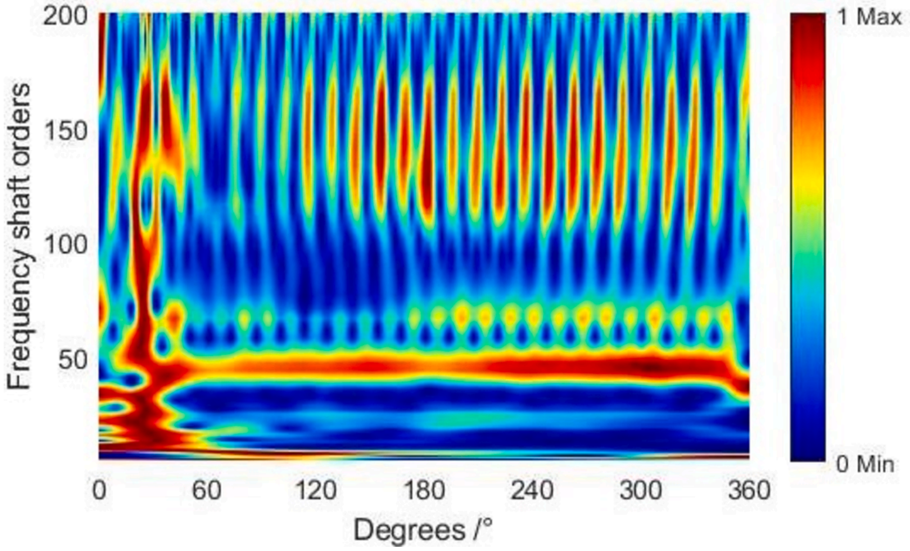


Fig. 6. Normalised Continuous Wavelet Transform (CWT) on TSA with a 14 μm bend.

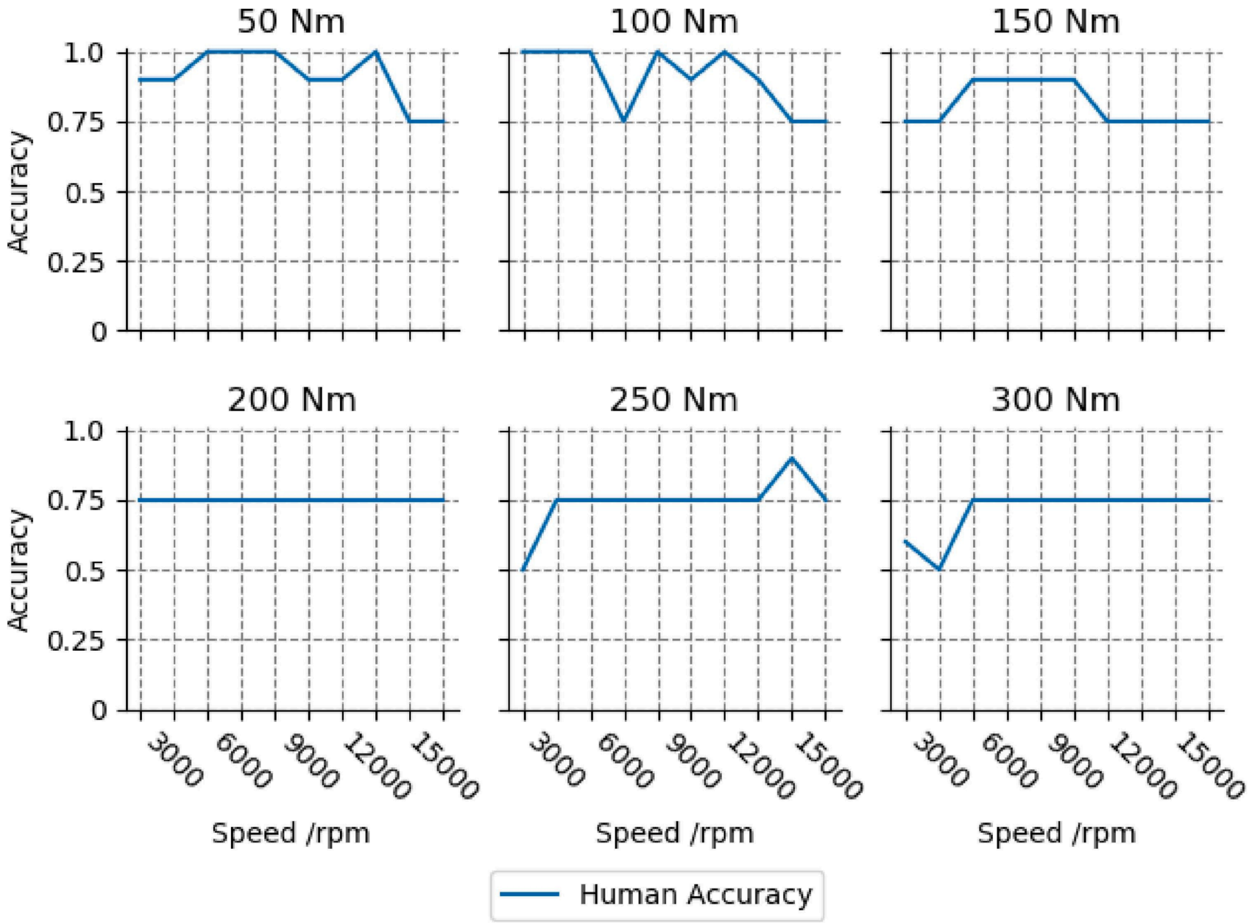


Fig. 7. Human level accuracy for manual identification of damage using CWT.

identify all damage and classify the levels of severity correctly too. The lowest category would then represent no damage being visible and none of the images are distinguishable from one another. The categories in between are thus incremental objective differences between these two extremes and are outlined below. Each category has also had a subjective accuracy assigned to it to allow some comparisons to be drawn between human and machine learning performance, the highest 100 % and lowest 50 % as randomly guessing will achieve 50 % accuracy.

Thus, typically, manual interpretation of the CWT data generally allowed the identification of the more severe defects only (8.7 and 14.7 μm) and even this was not always the case at higher torque levels. At lower torques, in some cases each damage level could be identified although this was inconsistent. Clearly, human interpretation of the ROC data using conventional analysis techniques alone is insufficient to produce a reliable condition monitoring system.

Some examples of the CWT plots the Human rater was presented at 3000 rpm are shown below for three speed cases that exhibit a range of accuracy; 100 Nm, 200 Nm, and 300 Nm, for 100 %, 75 % and 50 % accuracy respectively.

In the Fig. 8 below, we can clearly see that the 2 μm defect is substantially different to the healthy sample, with a single impulse of colour near the centre of the image on the second meshing harmonic. This impulse gets larger with each increase of damage to 8 and 14 μm bends. This results in clearly identifiable and distinguishable samples.

In Fig. 9 below we can again see very clearly identifiable samples for the 8 and 14 μm bends, but there is very little detail in the 2 μm sample that can be separated from the healthy, particularly given the contrasting signal level that is seen in the 100 Nm samples. As such, these sets of samples are given a 75 % objective accuracy.Fig. 10.

In the final figure of the CWT plots shown here, the 300Nm samples are less distinguishable again, with very little to distinguish between any of the plots. One could argue that the 14 μm CWT has some damage present, but comparing back to the 100Nm samples ranging from healthy to most damage, its challenging to find a confident category to split between. Further to this, the 2 μm defect appears more damaged than the 8 μm defect creating more confusion at this operating point.

It is thought that the 3000 rpm case exhibits a large range of variance in each sample's signal content due to the inherent resonances of this system. This explanation is derived as the machinery itself audible resonates significantly louder at 3000 rpm than other frequencies in this range.

4. Neural networks

Extending the work of Cahill, a neural network was developed to use the extracted metrics from TSA ROC signals. RMS, Kurtosis, Energy Operator, Zero Order figure of merit, M6A and M8A were employed as pure statistical features of the recorded signals. The frequency-based metrics discussed in the antecedent section were also utilised. Splitting the frequency bands into octaves enabled these metrics to capture the disorder in each harmonic of the gear tooth meshing frequency.

To increase the amount of data for learning and introduce some randomness to the information, the TSA signals were generated over a 5-rotation average and a single rotation from this 5 rotation TSA was randomly selected n —number of times from it and the metrics were calculated on this subsection of signal. This allowed an artificially much larger dataset to be generated than initially recorded whilst introducing a variability to the defect location within each sample.

Fig. 11 plots each normalised metrics' variance contribution to the first 5 Principal Components that capture 96.4 % of the total variance in the dataset. Interestingly, the frequency metrics have the most significant effect in general with M8A providing similar levels of variance. It should be noted that a feature providing variance to data does not necessarily mean it is a good predictor – however the more features showing more variance could mean that there is information for ML models to learn from. Summarily, it shows that the frequency metrics could provide good benefit to the models' learning capability.Fig. 12.

A test was created where models of varying architecture were trained on half of all data and tested on the other half. Then the process was repeated training and testing on opposing halves of the data. The test set was constructed to select alternating sample conditions from a matrix of all data-points and is illustrated in Table 3. For each model test, the data was trained on samples either labelled A or B, and then tested on the other.Table 4..

The models trained were asked to distinguish between damaged or healthy samples, not the severity of the damage. The importance of this distinction is to under- stand if the models have truly identified what damage means and therefore are well generalised to be able to predict accurately on completely unseen data. Breaking the problem down to a 4 category problem may allow some over fitting to the problem domain and reduce its effectiveness in a real-world application as it has only been trained on 4 defect sizes, if it is now presented a 5th defect size it is undefined how the model will predict that sample. Ergo, training for a binary classification of damaged or healthy will cover this uncertainty in the model's scope.

Splitting the test and training sets into a 50/50 distribution fairly provides completely unseen samples to the trained models across

Table 2
Quantifying damage levels into an objective accuracy.

Category	Description	Objective Accuracy
1	Each damage level is clearly detectable and severity is distinguishable	100 %
2	2.4 μm and healthy have some ambiguity	90 %
3	Only 14.7 and 8.7 μm defects are detectable	75 %
4	Some damage seen; no severity is distinguishable	62 %
5	No obvious damage seen	50 %

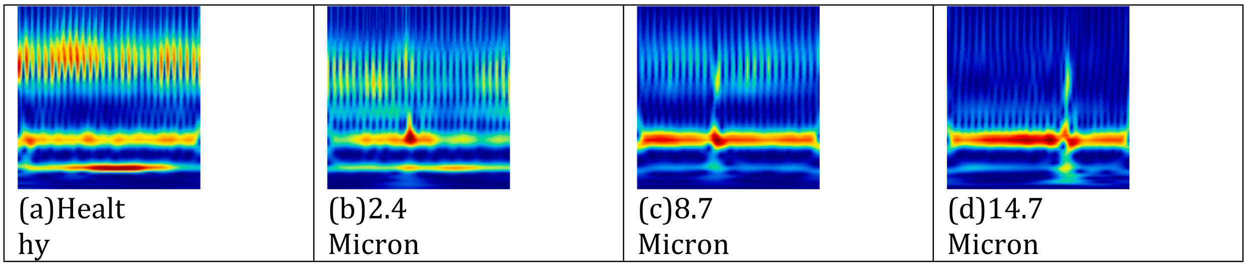

(a)Healthy (b)2.4 Micron (c)8.7 Micron (d)14.7 Micron

Fig. 8. 100Nm 3000 rpm.

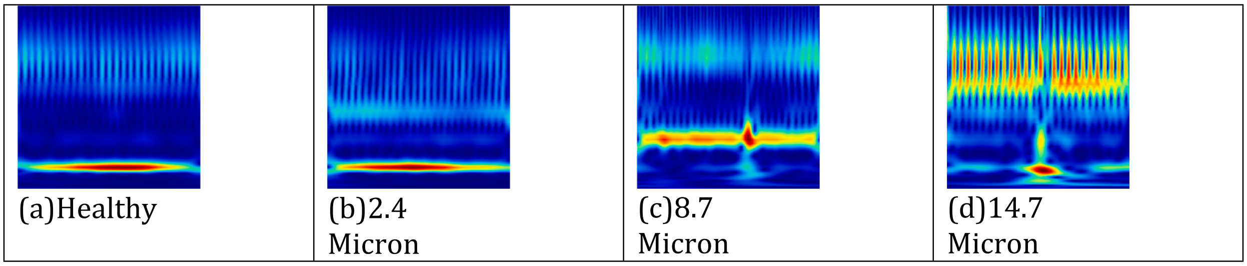

(a)Healthy (b)2.4 Micron (c)8.7 Micron (d)14.7 Micron

Fig. 9. 200Nm 3000 rpm.

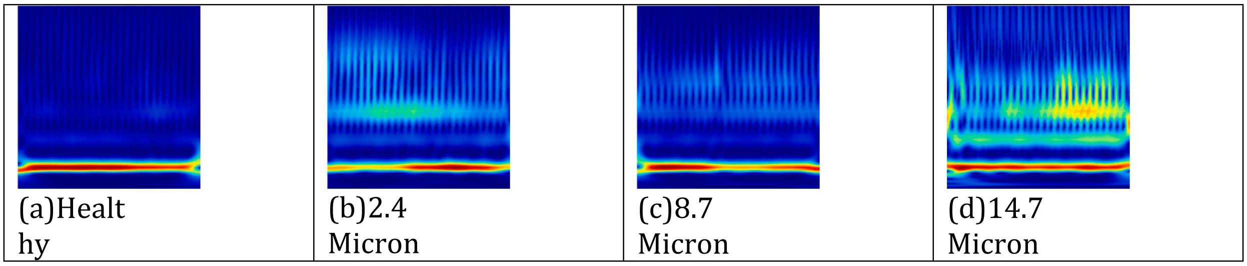

(a)Healthy (b)2.4 Micron (c)8.7 Micron (d)14.7 Micron

Fig. 10. 300Nm 3000 rpm.

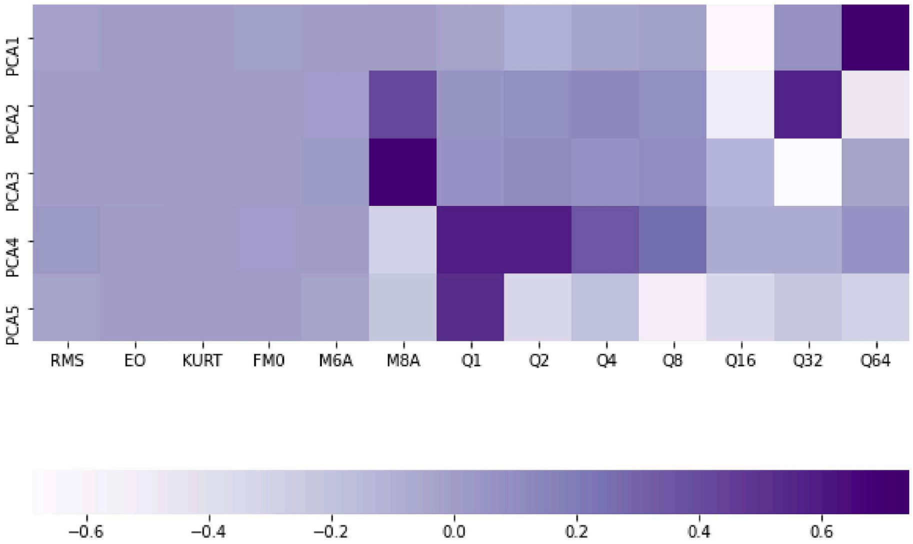


Fig. 11. Each feature’s contribution to the first 5 Principal Components.

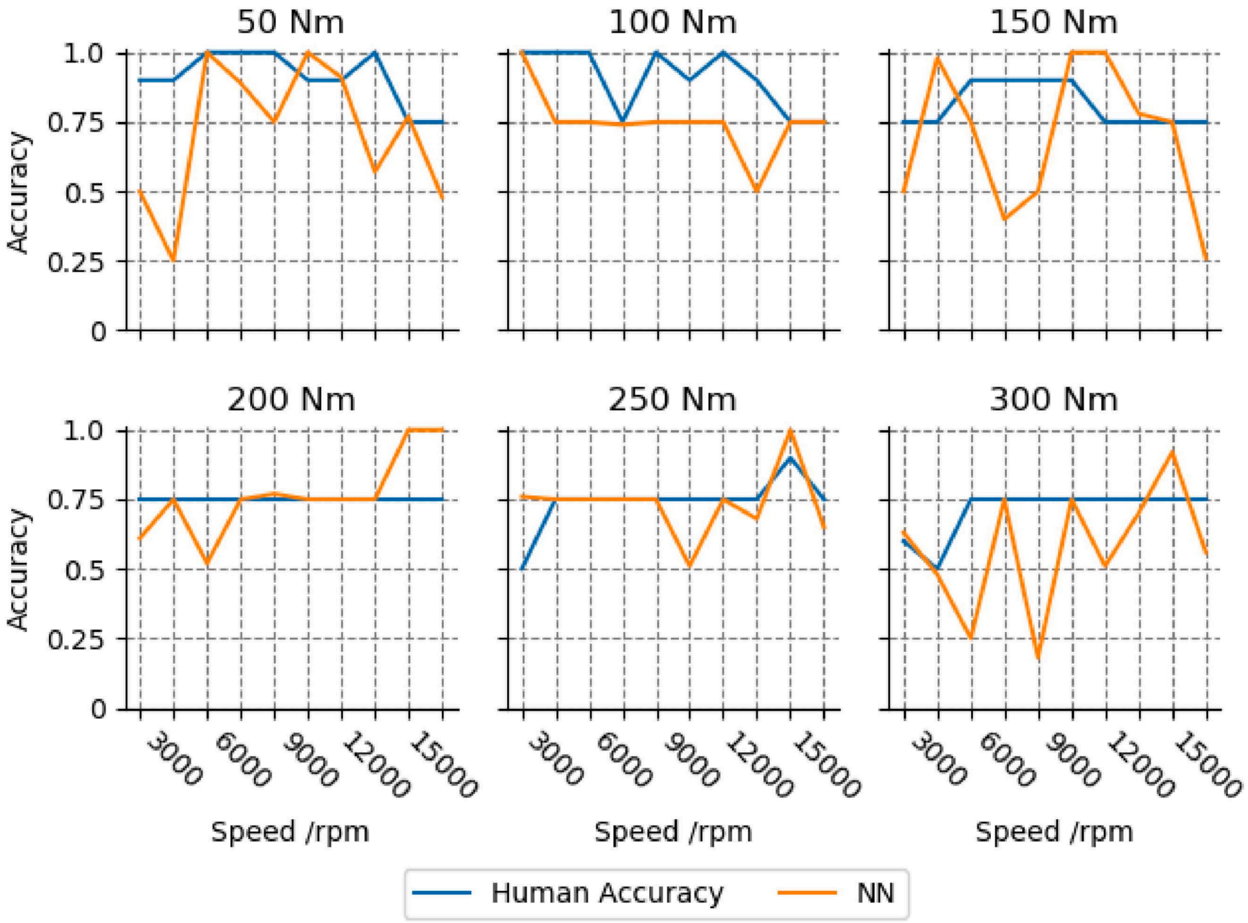


Fig. 12. Human vs Neural Network Radar Accuracy plot for classifying ROC Data.

Table 3
Train-test matrix.

Speeds/RPM											
Torque/Nm		1500	3000	4500	6000	7500	9000	10,500	12,000	13,500	15,000
	50	A	B	A	B	A	B	A	B	A	B
	100	B	A	B	A	B	A	B	A	B	A
	150	A	B	A	B	A	B	A	B	A	B
	200	B	A	B	A	B	A	B	A	B	A
	250	A	B	A	B	A	B	A	B	A	B
	300	B	A	B	A	B	A	B	A	B	A

Table 4
Neural Network parameter search results.

Architecture	Parameters	F1-Score	Accuracy	AUC	Kappa
64	1′154	0.792	0.718	0.T24	0.371
128	2′306	0.798	0.72	0.T09	0.357
256	4′610	0.801	0.727	0.724	0.378
64,64	4′610	0.778	0.706	0.73	0.366
128,128	18′818	0.815	0.743	0.733	0.404
256,256	18′818	0.832	0.757	0.714	0.395
64,64,64	9′346	0.794	0.716	0.706	0.35
128,128,128	35′074	0.829	0.755	0.722	0.402
256,256,256	135′682	0.829	0.753	0.708	0.383
128,256,512	167′682	**0.838**	**0.767**	**0.736**	**0.429**

the range of the data distribution. This gives a realistic idea of performance a model might achieve if a new out-of-sample piece of data were recorded and provided to the model — again the focus being on real-world application.

Each model was trained with L2 weight and LI bias regularisation of 0.01 and 0.03, respectively. The hidden layers used 'ReLu' activation and the output layer used 'softmax' activation. A dropout of 0.2 was used between each hidden layer. Adam was the optimizer with a 0.03 learning rate trained over 500 epochs with a binary cross-entropy loss function. Early stopping was used to avoid overfitting, training was stopped when the validation loss value did not improve for 15 consecutive training epochs and the weights that gave the best validation loss were restored at the end of training. The validation was split randomly from the training data at a 30/70 proportion. To correct the imbalance in the dataset's classes, SMOTE (Synthetic Minority Oversampling Technique) resampling was used to selectively re- sample the minority class based on a k-Nearest Neighbours algorithm. This allows the training data to be fed with a completely 50/50 split distribution. Each model was trained 5 times and the test metrics were averaged over the set and shown here.

Table 5 below shows the best performing architecture's classification report for test set B. Despite an average Fl-Score of 77 %, the classifier here has clear shortcomings in predicting the under-represented healthy class. Which is not unexpected, but a precision of 52 % means that when the classifier predicts a sample as "Healthy" it is only slightly better than randomly guessing. Of the samples that were healthy, only 66 % of them were classified correctly.

In the bigger picture with respect to high performance gearbox condition monitoring, accuracies of 80 % are not good enough for deployment to the real world. Potential for 1 in 5 predictions to be returned as false positives, or worse false negatives could have significant impacts on decision making and result in catastrophic failures.

The overarching goal of machine learning is to automate human tasks or improve upon them. Briefly comparing the Neural Networks' performance to the human classification line plots shows that generally, the neural network trained on this range of metrics performs poorly and thus reinforces the view that this methodology may not provide real-world benefit. However, it may provide some insight into the way the model has learned and where it occasionally outperforms humans.

4.1. Model explanation

Shapely Additive exPlanations (SHAP) provides a method to explain individual predictions and is based on game theoretically optimal Shapely Values [30]. It is demonstrated, by user studies, that SHAP is better aligned to human interpretation and more effectively discriminates among model output classes than existing methods. According to the original authors, SHAP provides a unified framework for interpreting predictions that unifies six existing methods, including LIME, DeepLIFT, Layer-wise Relevance Propagation and Classic Shapely Value Estimation [30]. The proposed SHAP method introduces the perspective that any explanation of a model's prediction can be viewed as a model itself, which is termed the explanation model. It further uses game theory to guarantee a unique solution that applies to the entire class of additive feature attribution methods and feature importance.

To interpret the SHAP plots, a greater SHAP value will mean a greater influence on the model prediction outcome. For example, with this binary problem, a high SHAP value will have a large influence on a damaged prediction through its partial contribution to the overall outcome.

Looking at the SHAP "summary plot" below for each class in Fig. 13a, it can be noted that the features are organised in order of importance according to their accumulated impact on the model based on the SHAP value. Each point on the plot represents a value of the feature for a single sample, thus there are *n—features x samples* total points on the plot. The points for each feature are colour coded with respect to their relative magnitude for that feature and are plotted with a 'jitter' type plot. The jitter signifies that when multiple points are overlapping, i.e., in spaces of high point density, the point's exact location is altered slightly within that area. Resultantly, the broadness of points within each feature gives a measure of value distributions for each feature and their respective impacts on the model output for that class.

The formulation of this data presents reasons for pure statistical metrics being unsuitable for this application. This may attribute an explanation for the frequency-based metrics consistently being more important in terms of their partial dependence of model outcome, determined by SHAP values. To elaborate, the test rig used to measure ROC signals from the gearbox consists of a reactive gear pair in the slave housing as well as the tested gear pair, meaning there is an echo of gear meshing signals occurring in the raw data. Varying the testing torque will mean a relative varying deflection of the shafts causing this echo to shift and super impose itself on the main ROC generated from the tested gear pair. Furthermore, there are two possible orientations to assemble the two gear housings as a result of the hunting ratio of the mating splines. These two orientations are approximately 4 degrees apart from one another, which is significant as this is in the same order of the contact lengths of each gear pair. These two factors make it difficult to attribute effects to specific causes and Cahill [28] established some of the peculiar behavior of these ROC signals as "dynamic gear effects".

The downfall of statistical metrics for this application is due to the consistent sampling frequency of l00kHz whilst the shaft speeds

Table 5
NN (128, 256, 512) Classification Report.

	Precision	Recall	F1-Score	Support
Healthy	0.52	0.66	0.58	2900
Damaged	0.88	0.80	0.84	8900
Accuracy			0.77	11,800
MacroAvg	0.70	0.73	0.71	11,800
WeightedAvg	0.79	0.77	0.78	11,800

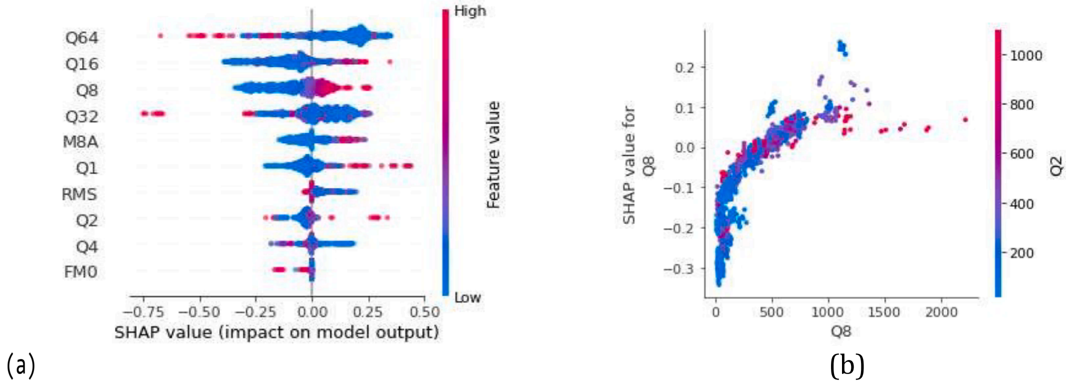


Fig. 13. SHAP Summary Plot (13a) showing overall feature importance and contribution on model outcome based on feature value. SHAP Dependence Plot (13b) shows the specific distribution and contribution of the Q8 metric on model outcome and its interaction with Q2 feature.

vary by a factor of 10. The result is that the number of samples recorded in each meshing event is approximately 148 at 1500 RPM and 14.8 at 15,000 RPM. This extreme variation in samples taken per event inevitably leads to a reduction in clarity/resolution of signals at higher speeds and thus is statistically inequivalent to compare between these features at different speeds even if upsampling techniques were used. Frequency metrics are not so dependent on the statistical properties of a signal and the number of samples taken for each event. For a stationary signal, the frequency domain is time and speed invariant, so the result that the frequency metrics outrank nearly all the purely statistical metrics is an unsurprising result.

Looking at the SHAP summary plot, the highest rated statistical metric is M8A. The purpose of this feature is to measure the peakedness of a signal. This means that when the M8A is high, the signal appears spikier and represents larger amounts of disorder in the signal [31]. We see that this is reflected in the feature value contribution to model outcome in the summary and interaction plots, where larger values contribute more to a damaged prediction outcome. Interestingly, this metric was shown to also have a high variance contribution to the dataset.

Many other metrics' feature values are not ordered entirely by their partial dependence on Shapley value, meaning that the information captured in each metric is not solely representative of a model outcome. For example, it cannot be inferred that high Q64 is indicative of a damaged prediction alone. Instead the models appear to be relying on a distributed representation of the metrics, where individual interactions between multiple metrics are what determines model outcome.

We see that frequency metric Q8 has the third highest feature importance ac- cording to the SHAP summary plot. Q8 is the metric associated with the eighth octave band and relates to half the tooth meshing frequency, or 27/2 times the shaft frequency. Q8 is one of the only features to exhibit a proportional contribution to model outcome with respect to its feature value, this can be identified by the visual order of the colour spectrum relating to the feature value on the summary plot. This means, alone, the Q8 feature may be capable of providing direct predictive power. Looking closer at this metric, in Fig. 13b (right), we can see the dependence plot which plots every sample's Q8 value against the SHAP value — a color scale is added for the most 'interacting' feature to attempt to highlight any further relationships that influence model outcome.

From the dependence plot, we see the proportional relationship to model outcome of Q8 is not linear, and any consistent relationship between Q2 and Q8 is not clearly visible. This means it is not by itself a perfect indicator for damage. Small changes to Q8 at low values may have a large impact on model outcome, and the variability in this 0–200 range can have very little meaning in true outcomes. From this range on-wards however the trend is more predictable. However, very high values of Q8 coupled with high Q2 values has a small attribution to a positive (damaged) prediction 0.05, but high values of Q8 (Approx. 1200) with low values of Q2 (Approx. 200) has a very high attribution to a damaged prediction. This relationship is not held consistent through the range of values seen on the dependence plot and does not relate to a clear interpretability of understanding from a human perspective.

As such, it is believed that the models have not learned what it means for the ROC signal to be damaged, perhaps because the information is not there in the metrics. Simply put, each metric is a proxy or a measure of some aspect of the signal, but damage itself is not exhibited truly in any one metric − if it were ML wouldn't be necessary. Thus each model optimises the loss of a network based on the relationships and interactions between each metric. This is qualitatively understandable as damage cannot be clearly seen by manual examination of any single metric. This suggests that this form of manual feature extraction is a sub-optimal solution for automating the condition monitoring with ROC signals and more advanced automatic feature extraction techniques will provide superior results. This is the general trend that has been apparent in the condition monitoring field since the relative explosion of Deep Learning in the 2010′s.

5. Deep learning

A number of pre-trained 2-D CNNs were trained to classify Continuous Wavelet Transform (CWT) images, in a technique known as Transfer Learning (TL). Transfer Learning initially uses pre-existing weights as a feature extractor to tune a top network of predictive layers. It then fine tunes the rest of the weights with an extremely low learning rate to better extract the individual features. TL is well suited to this domain as it allows some of the deepest available models to be trained on a limited dataset and reduce the overfitting which can occur when trying to train these networks from scratch. It hence reduces time to convergence and allows a good comparison between state-of-the-art techniques to provide a proof of concept for this type of model and problem domain.

It is expected that the time–frequency representation of the signal will better extract the key features that are representative of damage in an automated way. This makes sense as the most important metrics for the simple network formulations were frequency based. This will also overcome the problem of resampling/interpolating the raw signals as an image of fixed size will be produced from any CWT. That said, the inherent limitation of signal resolution will still exist because of a fixed sampling frequency. A range of randomly selected images from the training set are shown below, in Fig. 14.

In these CWT plots, each image is a time–frequency representation of the signal, that when plotted on two axes has time equal to one full rotation of the gear on the abscissa, with frequency up to 200 shaft orders on the ordinate axis. This effectively normalizes the speed dependent content of the signal to a consistent coordinate system for each load and speed case recorded. The frequency axis is plotted on a log scale as this was seen to represent the more significant components of damage in the plots and enforces the previous observation that the disruption in the signal from gear tooth bends influences the lower frequency bands more significantly than higher. This was also reinforced by resulting model accuracies, where the best results were achieved with the log scale when compared to a linear frequency with the same model architecture.

The CWT is originally an image size of 224 × 224 × 3 (width × height × channels), a common dimension for CNN computer vision problems. This will allow easier transferability to pre-trained models, specifically designed around the "Imagenet" dataset. The CWT was normalised so that the highest power in the spectrum was assigned the highest colour value in the colour palette. The result of this is that healthy samples with evenly distributed power have a broad stroke of red across the meshing frequency. Damaged samples have high disorder and periodic spikes in signal power result in localised spots of red on a normal background, typically this localised spot of power spans a broader range of frequency bands as the disorder in meshing introduces high-frequency impulse events. It should be further noted the natural ambiguity observed in these simple images, particularly between Healthy and 2 μm as a large number of these images have visually indistinguishable characteristics – as the damage levels increase the number of examples where this is also true reduces such that the largest defect of 14 μm is almost always obviously distinguishable to us.

The results from training a range of pre-trained network using the TL technique described above is shown here in Table 6. The ResNet50v2 network provided the best result, but all of the models tested here performed better than the best performing Neural Network training on metrics alone. The ResNet architecture implements an additive residual function that creates a direct path to propagate information in both the forward and backward pass in training across the entire network experiments showed [32] this formulation makes training easier and improves generalisation. As such it makes sense this architecture lends itself well to Transfer Learning.

Interestingly it is not the most complex models of each model family that perform the best. For instance, although the difference is marginal the largest EflicientNet model, B7, performs worse than the B6 variant with 65 % fewer parameters. It could make logical sense that when a model has saturated its learning ability with the training information, increasing depth and complexity will not increase generalising ability any further and only lead to inefficiency in the training, hence the plateau seen in these examples. It is also worth noting that the test-to-test variation of accuracy for each model could account for this difference in model ability, especially as the total accuracy variation between all EfficientNets varies only 3.8 %.

By default, a GlobalMaxPooling2D layer was used immediately after the base model, it transforms the 3-dimensional convolutional output to a single dimensional set of features. A top model was added beyond this global-pooling layer with the same parameters for each model run, with hidden layer architecture 512, 256, 64 (ReLu activation) before the output layer (softmax activation). Again,

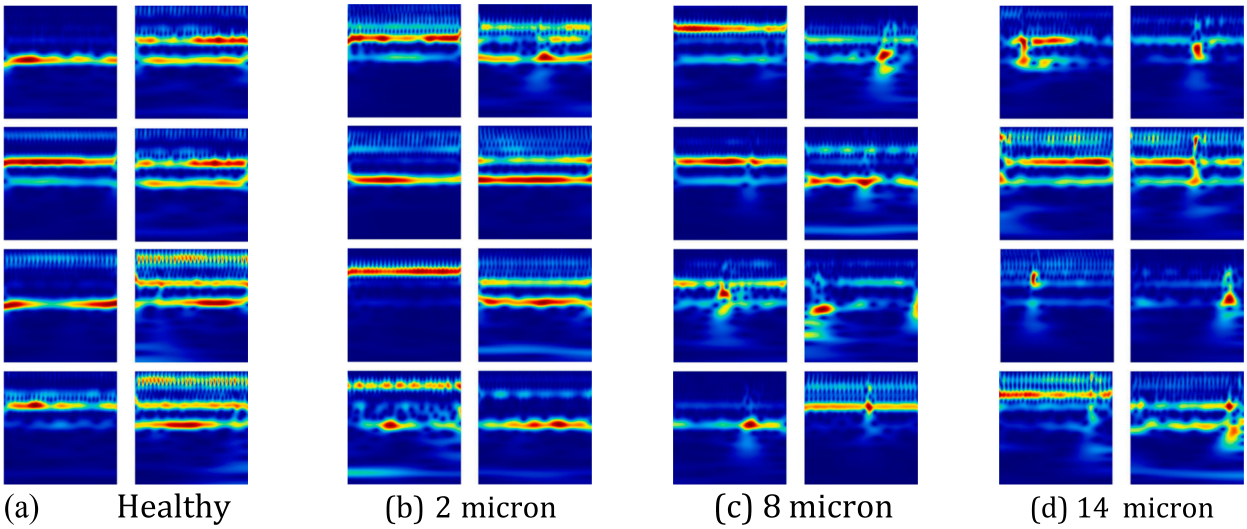


Fig. 14. Random selection of raw CWT Images used in training.

Table 6
Neural Network parameter search results.

Base Model	Parameters	N-Layers	Fscore	Accuracy	AUC	Kappa
DenseNetl2l	7,710,210	427	0.903	0.859	0.845	0.648
DenseNetl69	13,643,266	595	0.908	0.866	0.847	0.661
DenseNet20l	19,453,442	707	0.913	0.871	0.84	0.664
EfficientNet B0	4,853,349	237	0.905	0.862	0.838	0.649
EfficientNet B1	7,379,017	339	0.884	0.833	0.813	0.584
EfficientNetB2	8,637,883	339	0.894	0.845	0.814	0.602
EfficientNetB3	11,718,385	384	0.9	0.853	0.818	0.618
EfficientNet B4	18,739,745	474	0.908	0.865	0.844	0.658
EfficientNetB5	29,710,521	576	0.903	0.857	0.827	0.634
EfficientNetB6	42,288,209	666	0.915	0.874	0.842	0.672
EfficientNet B7	65,556,825	813	0.908	0.864	0.828	0.643
InceptionResNetV2	55,271,586	780	0.899	0.849	0.798	0.598
InceptionV3	22,999,778	311	0.891	0.837	0.786	0.57
MobileNet	3,901,570	87	0.914	0.873	0.844	0.67
MobileNetV2	3,061,762	155	0.909	0.864	0.818	0.638
NASNetMobile	4,958,806	769	0.878	0.816	0.751	0.505
ResNetl0l	43,855,170	345	0.914	0.871	0.833	0.659
ResNet50	24,784,706	175	0.901	0.855	0.831	0.632
ResNetl52	59,567,938	515	0.911	0.864	0.803	0.627
ResNet50V2	24,761,794	190	0.929	0.894	0.857	0.717
ResNetl0lV2	43,823,554	377	0.913	0.872	0.849	0.673
ResNetl52V2	59,528,642	564	0.914	0.872	0.829	0.658
VGG16	15,125,250	19	0.908	0.863	0.825	0.64
VGG19	20,434,946	22	0.902	0.854	0.82	0.62
Xception	22,058,474	132	0.904	0.858	0.822	0.627

each hidden layer had a dropout of 0.2 and L2 weight and L1 bias regularisation both of 0.07. For each initial training, the base model was frozen and only the top predictive layers were trained – this treated the base model as a standalone feature extractor that yielded a number of elements proportional to the model. For the fine-tuning portion of the training, the base model is unfrozen except that the bottom 20 % of layers were held frozen as these bottom layers are typically learn very simple and generic features that generalize to almost all types of images so do not need to be retrained.

Fig. 15 below shows the comparisons of this method vs the accuracy of a human when sorting through the CWT images by hand at each sample point.

Generally, the trend seems to indicate that lower loads provide a clearer signal with easier to predict features in the CWT. The speed does not have the same kind of effect as load. This suggests that the method of CWT generation overcomes the problems with fixed sampling frequency. Notably, comparing the Transfer Learning approach to a Human's ability to classify the CWTs shows very similar trend – both performing the worst at 300Nm, the CNN even outperforms the Human significantly in a number of samples for example, l2000rpm / l00Nm and 9000 rpm / 250Nm.

5.1. *Model explanation*

To try and further understand the models' predictions, we can again use the SHAP library. For the 2-D representation, a kernel explainer was used. The image was split into 200 segments using the skimage toolbox, with a compactness of 10 and sigma of 7. These parameters were decided through a generalised search of the parameters with the goal to segment the image into slices of the same shape and scale of the main features in each CWT. The use of segmentation and kernel explanation allowed compatibility between all modules and a significantly lower computational cost. The result is that segments that attribute positively to the prediction are highlighted red, and the segments which contribute negatively to the prediction are highlighted blue. The level of shading and transparency in each segment is therefore a representation of magnitude of SHAP value.

Taking some key points of this test set to look at, we may understand further the model's accuracy and inaccuracy. Looking at a high accuracy region first – 9000 rpm 50 Nm had 100 % accuracy. There is obvious distinction between these plots shown in Fig. 16 and the larger damages exhibit a more localised power intensity where the fault occurs in the rotation as expected. This is a perfect representation of the abilities of ROC as the instantaneous measurement at a gear defect manifests as a clearly interpretable impulse signal. The smallest defect (2.4 μm) has some differentiable characteristics from the Healthy sample with a larger intensity band at a lower harmonic, but does not exhibit the same clear distinctions as the larger two defects.

The SHAP image plot, in Fig. 17, for the 14 μm sample shows that the model is identifying this same feature that we would use to determine this damage. There is a highlighted red ring around the localised energy peak on the CWT caused by a damaged gear tooth, this shows that this area influenced a damaged prediction heavily. Since it is the darkest red of the plot, we see that it had the most significant single impact on model prediction outcome.Fig. 18.

We now consider 4 samples from a less successful region of 64 % accuracy, 7500 rpm / 300Nm. The same characteristics are still present as before but with significantly lower power intensity around the fault. Perhaps, given the scale of disruption to the severely damaged samples in this plot, it may be an indication of other unknown damage present in the 'healthy' components of the test rig. The

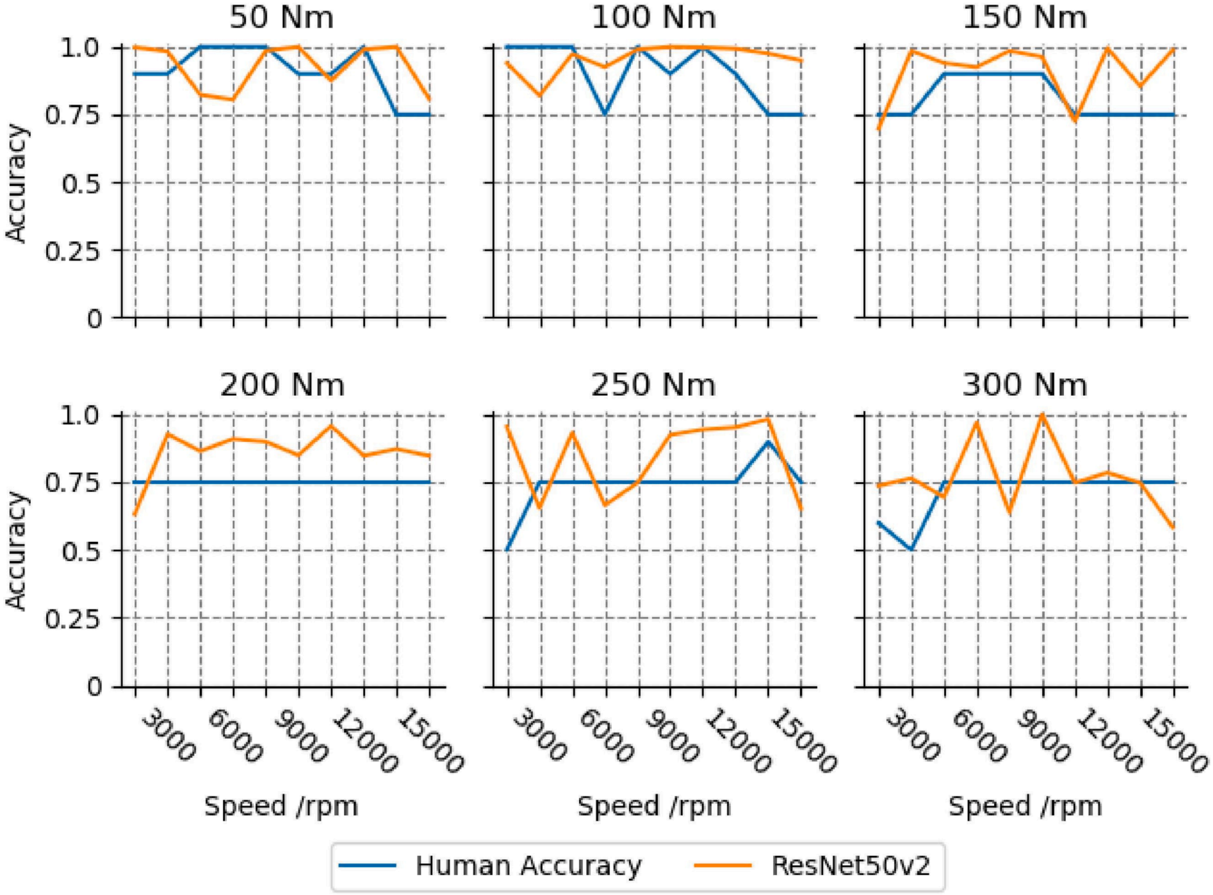


Fig. 15. ResNet50v2 vs Human Accuracy Line Plot.

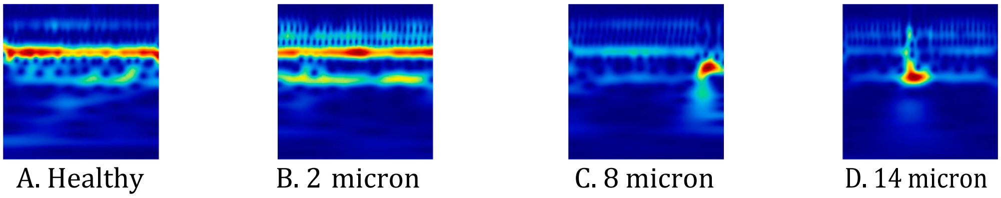


Fig. 16. Random images at 9000 rpm 50Nm all damage levels.

unevenness of the frequency bands is broad and doesn't span multiple frequencies like an impulsive gear defect might produce. So, this may even mean that there is some misalignment in the shafts and bearings, or perhaps some eccentricity of shafts in relation to the sensors.

The SHAP image plot, shown in Fig. 19, highlights an important case where the model has misattributed a damaged prediction as Healthy. The broad blue horizontal band between the two red frequencies on the original CWT has been highlighted red in the SHAP plot and hence the model is using that to most significantly to influence a Healthy prediction. This appears to be suggesting the lack of anything in this area is important, as more damaged predictions have a disorder spanning multiple frequencies and breaking up these two bands. The overall CWT image is hard for us to distinguish as damaged in this case and the lower accuracy is understandable and the explanation given by the SHAP model provides useful and sound reasoning to the model's decision.Fig. 20.

To examine a speed and load scenario where human raters had high confidence but the models performed less effectively, we present the figure below. At 600 RPM and 50 Nm, the rater identified these samples as having distinguishable severity. From left to right, the samples represent Healthy, 2.4-μm, 8.7-μm, and 14.7-μm conditions. The Healthy sample, despite some transients during rotation, shows a broad power distribution across most of the image. Minor factors, such as potential misalignment or subtle tooth-spacing errors, may contribute to a light banding pattern at a lower gear harmonic in the latter half of the image. The 2.4-μm defect exhibits a consistent power band at the lower meshing harmonic, with a small impulse at the higher harmonic near the center of

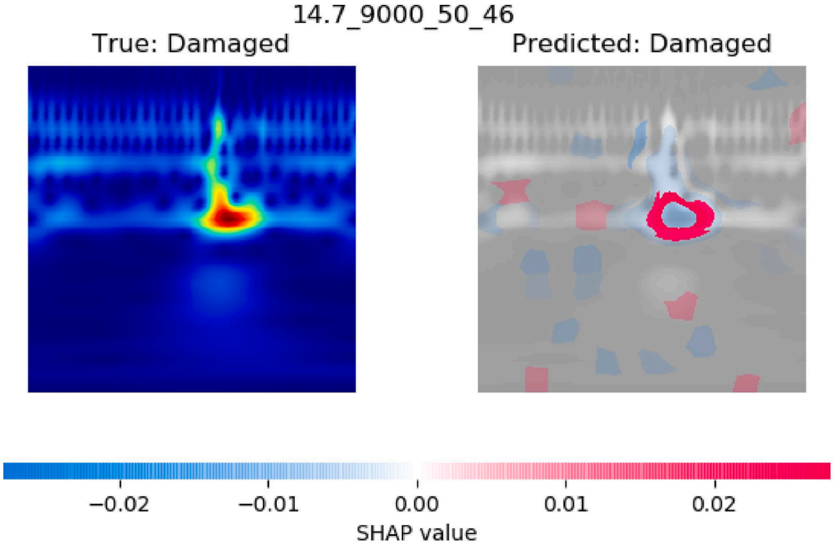


Fig. 17. 14 μm 9000 rpm 50Nm SHAP image plot showing successful prediction.

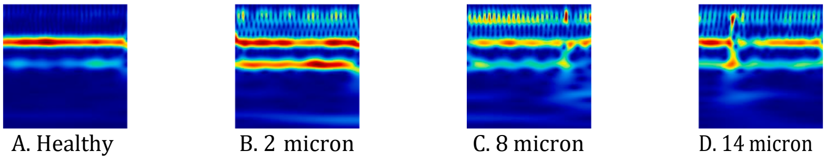


Fig. 18. Random images at 7500 rpm 300Nm all damage levels.

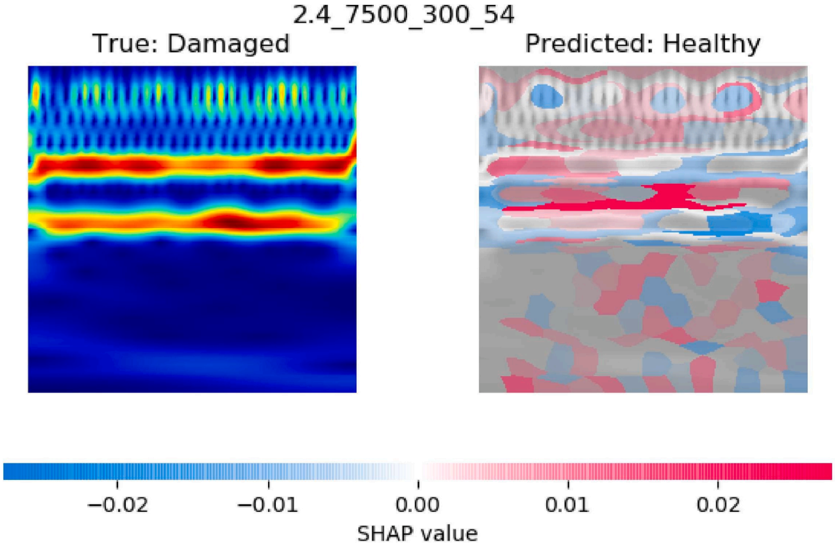


Fig. 19. SHAP image plot for 2.4 μm defect incorrect prediction at 7500 rpm 300Nm.

the image. Both the 8.7- and 14.7-μm defects display significant power in the higher meshing harmonic, but there is a noticeable disruption between the two that spans multiple frequencies, with the 14.7-μm defect causing a larger disruption compared to the 8.7-μm defect.

While a rater can distinguish between these samples and understand their intricacies when presented within the context of a specific load and speed case, these patterns differ notably from the more prevalent trends observed in the rest of the dataset discussed thus far. Typically, the Healthy sample shows more consistent and broad meshing harmonics with significantly fewer transient features

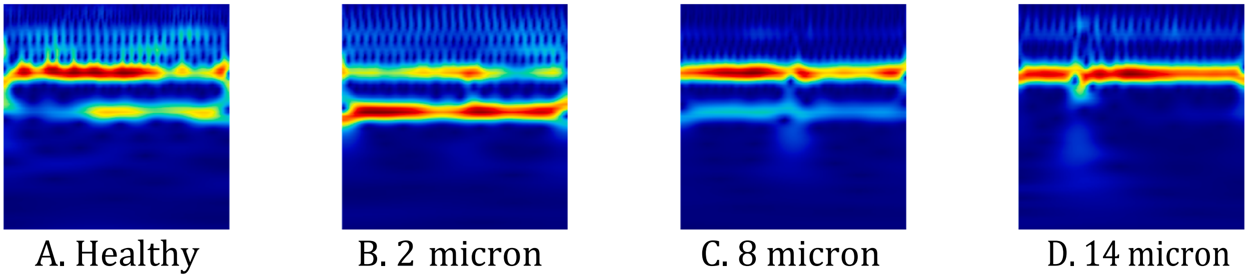


Fig. 20. Random images at 6000 rpm 50Nm all damage levels.

across the image. The 8.7- and 14.7-μm defects also deviate from the expected norm, where the typical defect pattern involves a large red striation that disrupts the main meshing harmonic for that speed case, spanning and linking with other visible meshing harmonics. In contrast, these samples lack such large red patches of impulsive events and are instead characterized by the absence of such features, with light blue regions and breaks in the meshing components.

To further explore the disparity between human judgment and the machine learning model, we can use additional SHAP plots to illustrate which image features are being leveraged to produce the performance outcomes shown in Fig. 15. Looking at this first plot below, Fig. 19, we see that the model has incorrectly attributed the Healthy sample as damaged. Its clearly highlighted the main inconsistent portion or red across the image as contributing strongly to this prediction outcome.

In the case of the 2-μm sample shown in Fig. 21, the model incorrectly classified it as Healthy instead of damaged. A key point to note with these SHAP plots is that a positive SHAP value (red) indicates a contribution to the positive prediction, which in this context means a Healthy classification—not necessarily an indication of damage. Here, the red areas indicate that the model is using these portions of the image to strongly contribute to the Healthy prediction. Specifically, the large, consistent band of blue across the image—a trait typically associated with Healthy samples—is a significant factor.Fig. 22.

The 8-μm sample shows a similar pattern, with the incorrect Healthy prediction driven by image characteristics typically associated with the Healthy class. This likely because these types of images are underrepresented in the dataset, preventing the model from learning the subtle patterns that a human rater might identify when comparing this specific load and speed case (Fig. 23).

To improve accuracy on this dataset, one potential approach is to incorporate speed and load variables into the model during training, enabling it to learn new internal representations for these nuanced cases. However, this introduces the risk of overfitting, which could negatively impact generalization performance. The current training process employed a relatively light augmentation strategy and strong dropout to enhance generalization. In the future, using more robust augmentations—such as significant color shifts, translations, or masking—could encourage the model to detect more subtle features associated with these nuanced damage patterns. Additionally, future directions for refining the problem might include applying label smoothing to handle ambiguous damage cases or even creating a separate class for instances with unclear levels of damage.

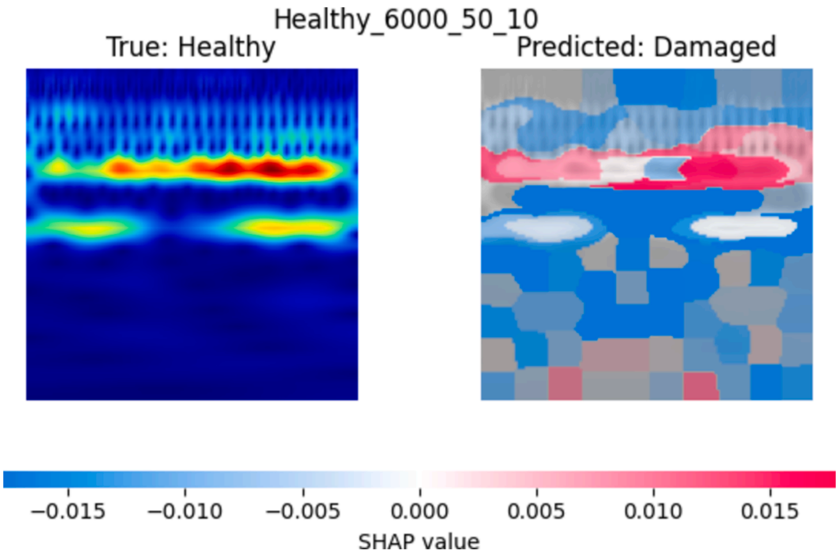


Fig. 21. SHAP image plot for Healthy gear incorrect prediction at 6000 rpm 50Nm.

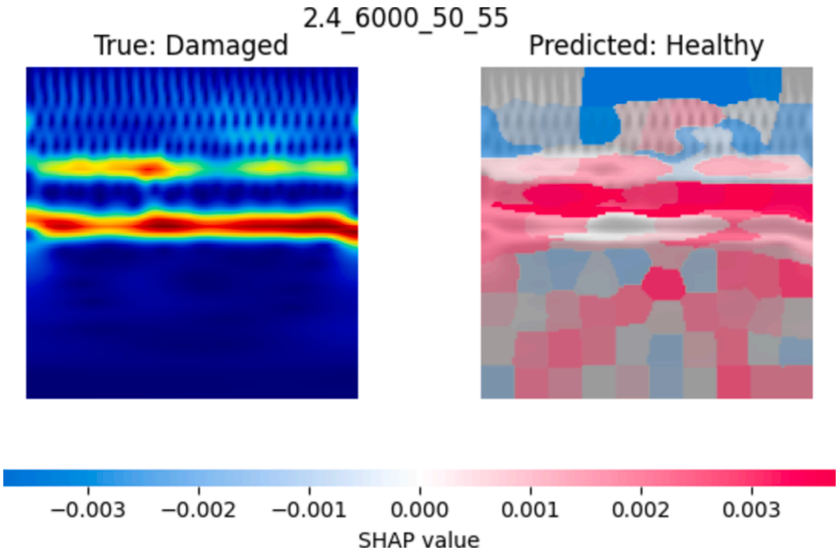


Fig. 22. SHAP image plot for 2.4 μm defect gear incorrect prediction at 6000 rpm 50Nm.

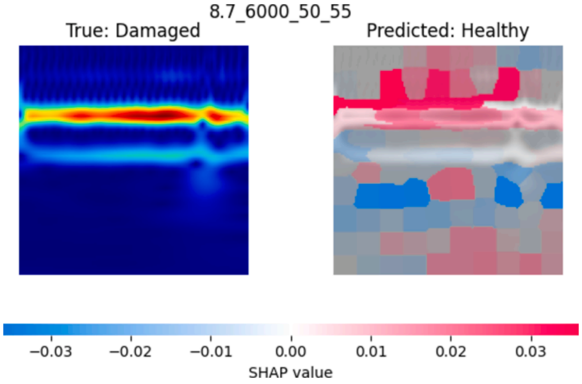


Fig. 23. SHAP image plot for 8 μm defect gear incorrect prediction at 6000 rpm 50Nm.

6. Conclusions

In this paper, a magnetoelastic rate of change of torque sensing device has been used successfully to identify gear tooth bend damage as small as 2.4 μm from normal. The sensing device is capable of identifying individual meshing events and excels at capturing the impulsive nature of gear-tooth bends.

A range of frequency based metrics have been developed and are presented in this paper. These metrics are directly applicable to traditional metric based condition monitoring, and a trend is shown between these metrics and the damage severity. A combination of all these metrics provides a predictive power to a neural network but summarising the complex waveform into 12 single value metrics ultimately reduces the information content of the signals and leads to a poorly-generalising network.

A range of pretrained (with the ImageNet dataset) Convolutional Neural Net- works (CNN) were trained using a technique known as Transfer Learning to classify damaged samples from a Time-Frequency representation of the signal, using normalised Continuous Wavelet Transforms (CWT). Firstly, as a feature extractor, freezing the main network and only training the top predictive layers, then secondly fine-tuning the entire network with a lower learning rate. Ultimately, TL provided a better generalising solution with a faster time to convergence than the manual feature extraction technique. The benefits resulted from the amount of available data and network complexity.

The general trend seen in the CNN development shows that an increasing network size and complexity results in better generalising performance up to a limit – where the model's number of parameters and complexity becomes redundant. Within network families such as the DenseNets and EfficientNets, this certainly holds true. However, it is the lightest ResNetV2 model that achieves the best results. Furthermore, there is no relationship across all models that can link model depth and number of parameters to the generalising accuracy. Instead it is specific architectures and their nature to pass information easily through the layers with a limited number of training samples that achieve the highest generalising accuracies. All of the trained networks with this technique achieved better results than the manual feature extraction technique and matched or surpassed human level interpretations.

As a result, the application of machine learning to data from the Rate of Change of Torque sensor has demonstrated the efficacy of automating condition monitoring in this application.

CRediT authorship contribution statement

George Hunt-Pain: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Ryan Walker:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Ben Cahill:** Resources, Data curation. **Alastair Clarke:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Alastair Clarke reports financial support was provided by Mercedes-Benz Grand Prix Ltd. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

References

[1] Jing Li, Chang-Chun Li, Jing Huang, Transmission efficiency analysis and experiment research of gear box. 3rd Annual International Conference on Mechanics and Mechanical Engineering (MME 2016), Atlantis Press, 2016.

[2] X. Liang, M.J. Zuo, Z. Feng, Dynamic modeling of gearbox faults: A review, Mech. Syst. Sig. Process. 98 (2018) 852–876.

[3] C. Spitas, V. Spitas, Calculation of overloads induced by indexing errors in spur gearboxes using multi-degree-of-freedom dynamical simulation, Proc. Inst. Mech. Eng., Part K: J. Multi-Body Dyn. 220 (4) (2006) 273–282.

[4] F. I. de L'Automobile, "2020 F1 Sporting Regulations 2020 Formula One Sporting Regulations," no. July 2018, pp. 1–70, 2018.

[5] I.J. Garshelis, R.J. Kari, S.P.L. Tollens, A rate of change of torque sensor, IEEE Trans. Magn. 43 (6) (2007) 2388–2390.

[6] I. Howard, A Review Of Rolling Element Bearing Vibration Detection, Diagnosis And Prognosis (AR-008-399), Department of Defence, Defence Science and Technology Organization, 1994.

[7] T. Wang, et al., Vibration based condition monitoring and fault diagnosis of wind turbine planetary gearbox: A review, Mech. Syst. Sig. Process. 126 (2019) 662–685.

[8] M. Tiboni, et al., A review on vibration-based condition monitoring of rotating machinery, Appl. Sci. 12 (3) (2022) 972.

[9] R.B. Randall, Vibration-Based Condition Monitoring: Industrial, Automotive And Aerospace Applications, John Wiley & Sons, 2021.

[10] D.E. Bently, T. Hatch'Charles, Fundamentals of rotating machinery diagnostics, Mech. Eng.-CIME 125 (12) (2003) 53–54.

[11] M.S. Raghav, R.B. Sharma, A review on fault diagnosis and condition monitoring of gearboxes by using AE technique, Arch. Comput. Meth. Eng. 28 (4) (2021) 2845–2859.

[12] Y. He, et al., An overview of acoustic emission inspection and monitoring technology in the key components of renewable energy systems, Mech. Syst. Sig. Process. 148 (2021) 107146.

[13] E. Caso, et al., Monitoring of misalignment in low speed geared shafts with acoustic emission sensors, Appl. Acoust. 159 (2020) 107092.

[14] X. Yang, et al., Analysis of spur gearbox dynamics considering tooth lubrication and tooth crack severity progression, Tribol. Int. 178 (2023) 108027.

[15] A. Singh, D.R. Houser, S. Vijayakar, Detecting gear tooth breakage using acoustic emission: a feasibility and sensor placement study, ASME. J. Mech. Des. 121 (4) (1999) 587–593.

[16] A.B. Novoa, C. Molina Vicuña, New aspects concerning the generation of acoustic emissions in spur gears, the influence of operating conditions and gear defects in planetary gearboxes, Insight-Non-Destruct. Test. Cond. Monitor. 58 (1) (2016) 18–27.

[17] N. Ullah, et al., Influence of optimal tooth modifications on dynamic characteristics of a vehicle gearbox, Internat. J. Automot. Mech. Eng. 16 (1) (2019) 6319–6331.

[18] D. Hanumanna, S. Narayanan, S. Krishnamurthy, Bending fatigue testing of gear teeth under random loading, Proc. Inst. Mech. Eng. C J. Mech. Eng. Sci. 215 (7) (2001) 773–784.

[19] M. Bozca, Optimisation of effective design parameters for an automotive transmission gearbox to reduce tooth bending stress, Modern Mech. Eng. 7 (02) (2017) 35–56.

[20] L. Hong, J.S. Dhupia, A time domain approach to diagnose gearbox fault based on measured vibration signals, J. Sound Vib. 333 (7) (2014) 2164–2180.

[21] T. Praveenkumar, et al., Fault diagnosis of automobile gearbox based on machine learning techniques, Proc. Eng. 97 (2014) 2092–2098.

[22] A. Stetco, et al., Machine learning methods for wind turbine condition monitoring: A review, Renew. Energy 133 (2019) 620–635.

[23] A. Kumar, et al., Latest developments in gear defect diagnosis and prognosis: A review, Measurement 158 (2020) 107735.

[24] R. Zhao, et al., Deep learning and its applications to machine health monitoring, Mech. Syst. Sig. Process. 115 (2019) 213–237.

[25] A. Movsessian, D.G. Cava, D. Tcherniak, Interpretable machine learning in damage detection using Shapley Additive Explanations, ASCE-ASME J. Risk Uncert. Eng. Syst. Part B: Mech. Eng. 8 (2) (2022) 021101.

[26] N. Herwig, P. Borghesani, Explaining deep neural networks processing raw diagnostic signals, Mech. Syst. Sig. Process. 200 (2023) 110584.

[27] L.C. Brito, et al., An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery, Mech. Syst. Sig. Process. 163 (2022) 108105.

[28] B. Cahill, Use of Rate of Change of Torque to Detect Damage in Gear Systems, Cardiff University, 2018. PhD Thesis.

[29] R.M. Stewart, Some useful data analysis techniques for gearbox diagnostics. Technical Report Paper MHM/R/10/77, Institute of Sound and Vibration, Southampton, 1977.

[30] S.M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions Scott, Nips 16 (3) (2012) 426–430.

[31] P.D. Samuel, D.J. Pines, A review of vibration-based techniques for helicopter transmission diagnostics, J. Sound Vib. 282 (1-2) (2005) 475–508.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Identity mappings in deep residual networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016.