# A New Comprehensive Monitoring and Diagnostic Approach for Early Detection of Mechanical Degradation in Helicopter Transmission Systems

Jessica Leoni[(a)], Mara Tanelli[(a,b)], Andrea Palman [(c)]

*(a) Dipartimento di Elettronica Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy*
*(b) Istituto di Elettronica e Ingegneria dell'Informazione e delle Telecomunicazioni - IEIIT CNR, Corso Duca degli Abruzzi 24, Torino, Italy*
*(c) Electrical and Avionics Systems, Leonardo Helicopters Division, Cascina Costa di Samarate, Italy*
*Email: jessica.leoni@polimi.it, mara.tanelli@polimi.it, andrea.palman@leonardo.com*

[*] *Corresponding author*

**Abstract**

Helicopters vulnerabilities specifically lie in single-load-path critical parts that transmit the engine's power to the rotors. A fault in even one single transmission's gear component may compromise the whole helicopter, yielding high maintenance costs and safety hazards. In this work, we present an effective diagnosis and monitoring system for the early detection of the mechanical degradation in such components, also capable of providing insights on the damage's causes. The classification task is performed by an ensemble of two learners: a convolutional autoencoder and a distance&density-based unsupervised classifier that use as regressors specific Health Indexes (HIs) and flight parameters. The proposed approach employs the autoencoder reconstruction error information to infer the most probable cause of each detected fault, and enacts post-processing filtering policies that effectively reduce the number of false alarms. Extensive experimental validation witnesses the good performances and the robustness of the proposed approach.

*Keywords:* Helicopter transmission; Fault detection; Time-frequency analysis; Machine-learning; Predictive maintenance; Autoencoder; Vibrations monitoring

## 1. Introduction and Motivation

Propulsion, lift, and flight maneuvering in helicopters are made possible by the cooperation of multiple single-load-critical parts. Since helicopters' operation is closely linked to every component's health status, their degradation is the leading cause of accidents, second only to human factors,[1]. To address this issue, during the North Sea operations conducted in the second half of the '80s, the UK Government promoted Health Usage Monitoring Systems (HUMS), which were ad-hoc designed for diagnostic, monitoring, and predictive maintenance purposes [2], [3]. Since then, these systems have been continually refined. Nowadays HUMS effectively monitor the integrity of the gearboxes [4, 5, 6] and other critical components as bearings [7], shafts [8], and rotors [9]. One of the core functionalities embedded in the HUMS is transmission vibration monitoring. According to this approach, failures are detected by analyzing over time health indicators extracted by the vibration signature of the transmission's components, as damages alter their characteristic patterns [10]. Vibrations are usually monitored, resorting to a pool of accelerometers ad-hoc placed along the transmission housing. To manage the huge data size coming from the high-frequency accelerometer measurements, the time series are usually processed on board and condensed in specific Health Indexes (HIs), appropriately designed to emphasize the presence of possible anomalies in the vibrational signature of the monitored components. Such HIs are stored on board and then downloaded to the ground station. On-ground, they are analyzed, and the components' health status is assessed.

### *1.1. Related Works*

As reported in [11], two are the main approaches to monitor helicopters components' health status. The first one aims at estimating the remaining useful life for a component by resorting to data-driven or physics-based models.

However, the transmission subsystem is characterized by complex nonlinear dynamics and is also affected by external factors that prevent estimating accurate models [12]. The other approach mostly relies on the HIs and requires defining a set of static thresholds, one for HI that defines the healthy distribution's boundary [13]. Accordingly, when an index deviate exceeds its hard threshold, an alert is triggered, and a maintenance intervention is planned to inspect the nature of the reported failure. This paradigm is the most employed when designing HUMS, despite false alarms still representing a limitation.

Therefore, in the last decade, research has been conducted to produce new HIs that lead to more accurate and robust results [14, 15]. In more details, indexes based on vibrations spectral kurtosis [16], envelope analysis [17], and cyclic spectral coherence [18] lead prove to be particularly effective. However, even the most promising indexes lead to false alarms, as the causes are related to the thresholds' calibration procedure and the HIs analysis process. Indeed, most helicopter companies fine-tune thresholds according to the domain experts' knowledge, which is costly in terms of time and money, and may also lead to human errors [19]. Moreover, univariate indexes analysis does not account for the correlations between different HIs, and between HIs and the helicopter operating conditions, thus reducing the capability of inference of critical conditions while increasing the risk of false alarms. Indeed, evidence in the literature is reported that parameters such as temperature, pressure, flow rates, and rotational speed of the helicopter are strongly non-linearly correlated with the computed HIs, and may cause their scattering without being related to an actual fault [20]. Therefore, attempts were provide to produce multivariate indicators for anomaly detection, which consider both HIs and flight operating conditions, see *e.g.*, [21], [22].

In addition, the recent development of machine- and deep-learning techniques provide an effective solution to the human-based fine-tuning procedure, revealing outstanding results in aircraft diagnostics and prognostics [23, 24, 25]. Indeed, these approaches are designed to infer the optimal separation hyperplane that allows for distinguishing instances belonging to different distribu-

tions. However, as transmission faults are rare, common supervised learning approaches are not directly applicable, as the imbalance between classes masks the behavior of the minority class. To overcome this issue, several literature approaches suggest combining under and over-sampling techniques to balance the dataset before applying standard supervised learning techniques, [26]. However, this entails neglecting part of the collected data and/or generating synthetic samples [27]. Instead, devising a one-class classification setup allows training a classifier to recognize the expected behavior from a representative number of healthy observations, distinguishing such behavior from anything that deviates from it. This approach does not demand any data manipulation, and it is also flexible, as the classifier is not trained to recognize a specific anomalous class but any behavior that does not conform to the expected vibration signature. Accordingly, applications relying on one-class classification algorithms, as autoencoders, achieve high performances and are less prone to generate false alarms [28, 29].

Works as [30, 31, 32, 33] propose relevant approaches, capable of exploiting the HIs information with machine- and deep-learning algorithms to infer an effective separation hyperplane that separates healthy instances from failures. One of the state-of-the-art approaches considering helicopters is the one presented in [34], where a convolutional autoencoder is applied to recognize vibration signatures referred to a faulty condition. Despite the benefits of introducing machine- and deep-learning techniques in HUMS, several false alarms are still triggered. One of the leading causes is that most of them still neglect the information referred to the helicopter operating conditions during the HIs computation. To minimize false positives, ad-hoc post-processing filtering policies should be designed. Moreover, as these approaches rely on black-box models, techniques should also be produced to interpret the results. An example approach lacks that provides information and localizes the faulty component identified in the transmission system. This information can be extremely useful to improve maintenance effectiveness and reduce operating costs.

*1.2. Contributions*

In this paper, we propose an effective diagnosis and monitoring system to ensure an early detection and localization of mechanical degradations of the transmission's components, i.e., gearboxes and bearings, shafts, and rotors. To enhance the evidence of specific deviations in the vibration signature pattern of the monitored components, ad-hoc HIs' are extracted by applying statistical and signal processing techniques in the time, frequency, and quefrency domains [35]. Our system also considers as predictors the flight regimes and the environmental operating conditions recorded at acquisition time, so as to reduce the false alarms due to HIs scattering caused by the variations of the aircraft operating condition. All the features are finally projected onto a one-dimensional anomaly score, estimated by an ensemble of two off-board classification algorithms: a semi-supervised convolutional autoenconcoder (CAE) and a Distance& Density-Based Unsupervised Classifier (DBUC), given by the combination of four learners, *i.e.,* k-Nearest Neighbours (kNN), Angle-Based Outlier Detection (ABOD), isolation Forest (iForest), and Local Outlier Factor (LOF).

A relevant novel feature of the proposed approach is its interpretability, which is fundamental to help domain experts investigating the outputs of the machine learning system. This result is achieved leveraging the AE reconstruction error to produce a ranking of the features used in the anomaly score determination, which is based on their displacement from the respective baseline. Since each HI is appropriately designed to enhance a specific damage, the one associated with the larger anomaly score is reported to the user as the most probable cause of the detected damage, and its combination with the observed components allows us to perform fault isolation, also suggesting which is the most probable component experiencing the fault. To this end, the tool is endowed with a mapping algorithm capable of reporting to the user the predictions at different aggregation levels: sensor and component. Indeed, a single component may be monitored by more than one accelerometer, and the tool merges all the accelerometer-level predictions to yield a component-level one. Another

innovative contribution is further given by dedicated post-processing filtering policies, which further reduce false alarms.

The proposed approach is tested on data collected by a pool of single-axis accelerometers over a 4-year-period on three helicopters that are part of the usual in-service fleet: two of them experience a single fault during the whole observation period, assessed by visual inspection of domain experts, while the third one is faultless. The proposed monitoring tool detects all the faults, also with the predictive capability of raising the alarm a few days before the damage was actually observed. Thus, the proposed solution goes beyond state of the art in different directions, namely:

1. Accuracy, as specific filtering policies have been designed to remove conceivable false damage conditions;
2. Applicability for predictive maintenance purposes, identifying anomalous behavior before the actual fault occurs;
3. Interpretability, providing insights about the source of the damage. This is crucial to guide troubleshooting and maintenance procedures leveraging fault isolation.
4. Hierarchy, allowing the user to investigate the aircraft status at different levels, *i.e.*, component or single accelerometer-wise.

The rest of this paper is organized as follows: Section 2 describes the dataset structure, along with its construction process. Then, Section 3 illustrates the structure of the classification and the post-processing phase. Section 4 presents and discusses the experimental results.

## 2. Dataset Description and pre-processing

This section presents the experimental data used to design the anomaly detection system and assess its performance. Useful insights are also provided on the data acquisition and on the computation of the HIs, so as to allow the reader to understand the complex dataset structure.

*2.1. Dataset Description*

The considered data were collected on three servicing helicopters from January 2015 to December 2018, from a pool of single-axis accelerometers mounted radially on the transmission housing. Such sensors allow to monitor the critical rotating components and to acquire their vibration signatures during each flight.

More specifically, the considered HUMS acquisition system is composed of 23 accelerometers placed along the whole transmission, which monitors 88 components overall. The piezoelectric accelerometers are specifically designed to measure vibrations in structures and objects. Of those 23 sensors, 18 are mono-axial, 2 bi-axial, and 3 tri-axial, as company experience suggests that some components should be monitored along more than one direction. Each accelerometer operates at 40kHz and has a peak-to-peak scale range of $\pm 500g$ within a temperature range of $-54$-$+150°$; the sensitivity of the sensors is $10mV/g$. The accelerometers' arrangement is the same for each considered helicopter as it is the standard employed in the industrialized platforms of the Agusta Westland family used in this work, and the sensors position was kept the same throughout the entire study. Each accelerometer monitors the vibrational behavior of the surrounding components. Thus, each sensor can actually monitor more than one component at a time. Therefore, to uniquely identify each accelerometer-component pair, an incremental acquisition ID is used, yielding a total of 180 acquisition IDs. Specifically, the $i_{th}$ acquisition ID corresponds to the pair of the $x_{th}$ accelerometer and the $y_{th}$ component, where $x$ and $y$ are integers ranging between 1 to 23 (the overall number of accelerometers) and 1 to 88 (the overall number of components in the transmission), respectively.

According to manual inspections of the machines and to the observation of the HIs behaviours, domain experts assessed that one of the three considered helicopter never experienced a fault during the study period, while the other two underwent one damage each. For the first faulty helicopter, the reported fault was detected on January $29^{th}$ 2017, while for the second one the faults was recorded on May $25^{th}$ 2018.

According to the indications provided by the domain experts, we label as anomalous all those observations whose acquisition IDs contain information on the damaged component starting one week before the fault day.

Note that, as reported in Table 1 in both cases, the anomaly-related observations are just a few samples in the huge dataset. Indeed, in the first aircraft, the potentially faulty observations constitute 0.1% of those collected in 2017 and 0.02% of the whole dataset. In the second helicopter, the potentially faulty observations constitute 0.2% of the 2018 ones and 0.18% in the whole dataset. The scarce representation of the anomalous conditions calls for managing the fault detection problem leveraging approaches that can deal with rare event investigation.

Table 1: Dataset Summary

| | Helicopter 1 | | Helicopter 2 | | Helicopter 3 | |
|---|---|---|---|---|---|---|
| **Year** | **Healthy (%)** | **Faulty (%)** | **Healthy (%)** | **Faulty (%)** | **Healthy (%)** | **Faulty (%)** |
| 2015 | 3914 (3.6%) | 0 (0%) | 0 (0.0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| 2016 | 34057 (31.1%) | 0 (0%) | 9795 (7.0%) | 0 (0%) | 1702 (1.8%) | 0 (0%) |
| 2017 | 26648 (24.3%) | 23 (<0.1%) | 36853 (26.3%) | 0 (0%) | 37465 (40.0%) | 0 (0%) |
| 2018 | 44951 (41.0%) | 0 (0%) | 93233 (66.2%) | 349 (0.2%) | 54431 (58.2%) | 0 (0%) |
| Total | 109593 | | 140230 | | 93598 | |

### *2.2. Data Acquisition and Computation of the Health Indexes*

The acquisition of the sensors time-series and the computation and storage of the HIs take place on-board the helicopter. In the literature, well-known statistical techniques are known to extract effective HIs, mainly based on processing methods applied to the raw signals in both time domain and frequency domain, [36]. The specific implementation of HIs considered in this work also extracted HIs leveraging methods based on spectral Kurtosis, phase demodulation, and signal enhancement. These techniques allow for robust fault detection even when intense masking noise affects the measured data, [37]. Aware of the possible false alarms triggered by HIs scattering due to variations in the acquisition condition, the flight regimes and the aircraft's operating conditions are

also stored, represented by airspeed, torques, pitch and roll attitudes, gearbox oil temperature and pressure. Since the HUMS is supposed to monitor not only the gearboxes, but also bearings, shafts, and rotors, different acquisition modes are available. The extracted HI strongly depends on the acquisition mode, as each if them is designed to highlight specific vibration signature frequencies. In more detail, a memory buffer is associated with each acquisition ID, which resides in the onboard control unit, *i.e.,* the so-called Aircraft Mission and Management Computer (AMMC). When the acquisition is triggered, manually or automatically, each buffer is cyclically filled with the vibration signature measured by the accelerometer corresponding to the associated acquisition ID, and, according to its specific sensor/component pair that such ID represents, the stored time series undergoes one out of four available pre-processing phases, namely:

- Time average acquisition mode: it emphasizes the presence of a localized damaged tooth in shafts and gearboxes, see also e.g., [38]. In this mode, the HIs are extracted by the accelerometer signal through:
  - Temporal analysis, to highlight the presence of improper meshing and tooth surface degradation;
  - Phase demodulation, to highlight the presence of gear/shaft fatigue cracks;
  - Spectral analysis, to highlight the presence of distributed defects on gear teeth (anomalous wear, pitting) or gear/shaft fatigue cracks in one or more teeth;
  - Enhancement analysis, to highlight the presence of shaft cracks.
- Envelope acquisition mode: it is particularly suitable for monitoring the bearings' health status. The corresponding HIs consider:
  - Inner and outer race energy components, which allows identifying localized pits, spalls, cracks, and debris over the inner and outer surface of bearings races, respectively;

  - Rolling element energy components, which allows identifying localized pits, spalls, cracks, and debris over the surface of rolling elements;
  - Cage energy components, which allows the detection of cage failures.

- Time-history acquisition mode: it is used to investigate the health status of gears and bearings. To this extent, the related HIs rely on enhancement, temporal, and residual analysis to highlight the presence of localized or distributed pits, spalls, or cracks over their surface.

- Auto spectrum average acquisition mode: it computes the vibrations cepstra and considers as HIs the set of its first coefficients. As discussed in e.g., [39], this acquisition mode allows highlighting localized pits, spalls, cracks, and debris over the outer bearing elements or gears.

According to the specific pre-processing technique, up to 18 HIs are extracted from the time series stored in each buffer of the AMMC, along with the operating condition characterizing the acquisition of the original vibration. When the helicopter returns to the hangar, the collected HIs and the parameters are downloaded, providing the dataset considered as input in our work.

## 3. Architecture of the Active Monitoring System

This section presents the architecture of the proposed anomaly-detection system.

As shown in Figure 1, the proposed pipeline consists of several phases. First, the collected dataset is split into 180 sub-datasets, each associated to a single acquisition, as discussed in the previous section. Further, each sub-dataset undergoes cleaning and pre-processing steps, where the corrupted instances are removed, and the input samples are extracted. The pre-processing is followed by the anomaly prediction phase, in which each sample is associated with an anomaly score, and such score is complemented with a features' important ranking, which indicates the features that most contribute in determining the produced score,

enhancing explainability and interpretability of the inference engine that yields the scores. Then, the anomaly scores that have been computed at the acquisition ID-level are aggregated at accelerometer level and then at component-level, thus enabling not only the detection of faults but also their isolation and association with the most probable components that are being damaged. Finally, the post-processing filtering policies are used to get the final results on the detected anomalies, if any.

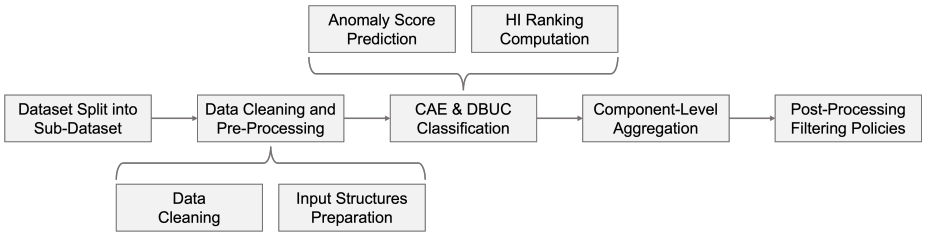



Figure 1: Overall architecture of the anomaly-detection pipeline.

### *3.1. Data Cleaning and Pre-Processing*

Each dataset comes in an aggregated form containing HIs, operating conditions and boolean flags indicating whether single data are consistent or not, coming from on-board evaluations of the signal ranges. Table 2 reports an overview of the structure of the inital datasets.

Table 2: Dataset Structure

| **SN** | **Acq ID** | **Date** | **HI1** | **HI1L** | **...** | **HI18** | **HI18L** | **Pitch** | **...** | **Roll** | **Flag1** | **...** | **Flag16** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ... | 1 | 06/10/2015 | ... | CA1 | ... | ... | CA18 | ... | ... | ... | False | ... | False |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| ... | 192 | 18/05/2018 | ... | TAVP2P | ... | NaN | NaN | ... | ... | ... | False | ... | False |

As mentioned in Section 2, the mathematical definition of each HI varies according to the acquisition mode. So, each dataset column assigned to an HI stores different information based on the specific HI instance. Therefore, an HIiL column is associated with each $i_{th}$ HI, indicating the name of the corresponding statistical metric used to compute it.

For this reason, the dataset was finally split into 180 sub-datasets, each one collecting only the instances referred to a single acquisition ID, so that the corresponding metric is consistent for all the instances. The sub-datasets containing less than 20 instances were discarded, since no robust model can be trained on such few data. As last pre-processing step, the data in each sub-dataset was normalized between 0 and 1 to prevent biases in the classification due to scaling reasons.

Table 3 reports the size of each dataset before and after the data cleaning and pre-processing phases, along with its average size.

Table 3: Dataset and Sub-dataset Sizes.

| **HC** | **Global Dataset** | | **Sub-Datasets** | | |
|---|---|---|---|---|---|
| | **Original** | **Cleaned** | **Min** | **Avg** | **Max** |
| 1 | 152536 | 109593 | 249 | 571 | 1088 |
| 2 | 260912 | 140579 | 326 | 1519 | 732 |
| 3 | 177320 | 93598 | 212 | 909 | 490 |

The assumption underlying usual anomaly detection paradigms is that the probability distribution of those instances that are considered anomalous differs from the standard one in terms of mean value, standard deviation, or both. In some cases, the anomaly distribution may also follow a different distribution. To investigate whether such a difference can be appreciated in our data we employed Andrews Plot, a Fourier series-based method to visualize high-dimensional data, [40]. Figure 2 shows the Andrews Functions' trend obtained from the considered data, each of which corresponds to a single instance in the dataset $x_i = \{x_{1_i}, x_{2_i}, ..., x_{n_i}\}$. Each Andrews Function (AF) is calculated as

$$AF_i(\vartheta) = \frac{x_{i1}}{\sqrt{2}} + x_{i2} \cdot \sin(\vartheta) + x_{i3} \cdot \cos(\vartheta) + x_{i4} \cdot \sin(2\vartheta) + x_{i5} \cdot \sin(2\vartheta) + ..., \quad (1)$$

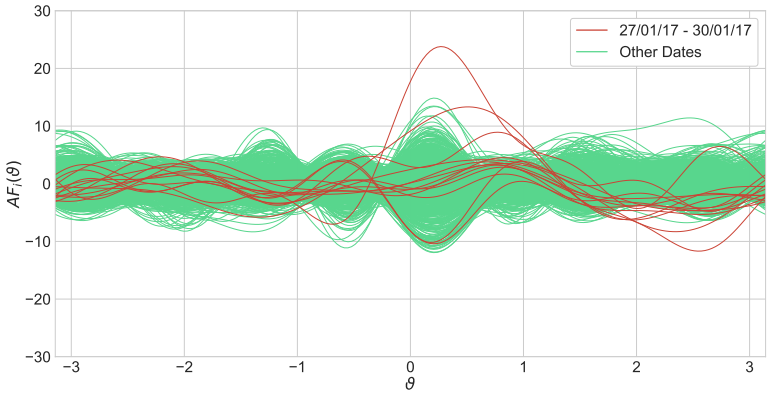
where $i = 1, ..., n$ identifies the considered instance. The number of terms sums up to the number of features, and the periodic functions are then plotted with respect to $-\pi \leq \vartheta \geq \pi$. According to AFs' definition, the anomalous instances

are easily identifiable, as they are characterized by curves that deviate from the prevailing trend shared by the AFs corresponding to the healthy instances. Figure 2 shows, as an example, two Andrews Plots obtained considering data from the first helicopter. Figure 2a is referred to data far from the reported damage, while Figure 2 reports the trend of those instances collected for an acquisition ID that is referred to the component that experiences the fault. It turns out that, in 2a all the AFs share the same behavior, while in Figure 2 the instances recorded during flights performed in the same week of the reported fault, *i.e.,* the red ones, behave differently. These results are consistent on the other datasets, confirming that there is statistical evidence that fault-related instances can be recognized in principle. Moreover, they are compliant with the expert domain suggestion to consider as anomalous all the instances referred to the component that experience the damage which are collected in the week preceding the fault.

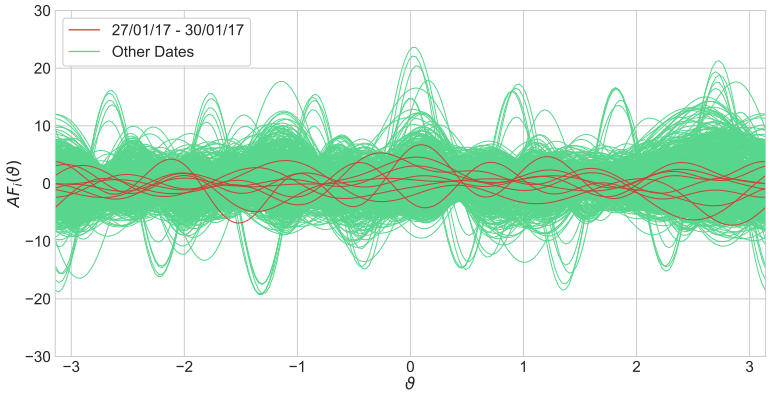
### *3.2. Design of the Ensemble Classifier*

The preliminary analysis supports the assumption that damages in an helicopter component do alter its vibration signature. The proposed classifier to detect the corresponding fault is given by the ensemble of a convolutional autoencoder (CAE) and a Distance&Density-Based Unsupervised Classifier (DBUC), which work to diagnose components' status considering all the provided features and produce a concise 1-dimensional anomaly score, along with the respective HIs ranking. This ranking is produced for each instance, based on the deviation of each HI trend from its healthy baseline. This information provides relevant insights into each detected fault's nature since each HI is related to a specific anomalous condition.

The ensemble architecture is depicted in Figure 3, and it shows that the two learners, *i.e.*, the CAE and the DBUC provide two anomaly scores that are merged, for each instance, to yield the final 1-dimensional anomaly score, obtained as the average of those independently estimated by each learner, after a standard normalization step.




(a) Healthy ID: AcqID 1 - Acc 1




(b) Faulty ID: AcqID 26 - Acc 14

Figure 2: Andrews Plots for a faultless (top) and faulty (bottom) dataset.

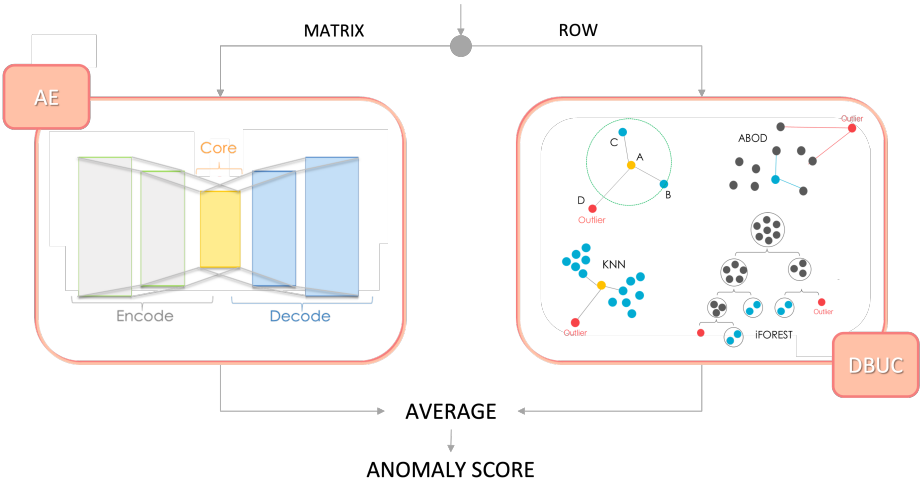



Figure 3: Overall Ensemble Architecture.

*3.2.1. Input Data Structure*

To optimally train the CAE and the DBUC, the input data must be correctly prepared to match the different characteristics of the two learners. In particular, the main difference between the implementation of the two models is in the input data structure. Indeed, as reported in Figure 4, CAE optimally manages 2-dimensional input samples, while DBUC processes 1-dimensional inputs. Thus, the CAE input data is prepared for each sub-dataset by extracting 2D the input matrices. In fact, each CAE input sample consists of a $nxn$ matrix, whose rows and columns host the HIs and the operating conditions collected during subsequent acquisitions. This structure allows the CAE to account for both features' dependencies and temporal evolution, and this optimizes the overall performance. The DBUC input, instead, is inherently 1-dimensional, so that its input consists of the HIs and the operating condition collected at each time instant.

The extracted samples are then split into train and test set. Accordingly, each model is trained on the data collected on all provided data, except those referred to the year in which the damage was experienced. For the helicopter that did not experience any fault, a year was used to test the models and the

other ones to train them. Please consider that one ensemble model composed of a CAE and a DBUC is trained for each of the 180 sub-datasets. As reported in Table 3, the number of instances differs based on the acquisition ID and the helicopter that are considered, and so do the training and test set sizes.

Before providing the samples as inputs to the ensemble DBUC classifier, normalization is also applied. This step is key, especially considering the distance-based learners included in the DBUC. Indeed, without normalization, these algorithms might attribute more importance to those samples that are larger in magnitude, regardless of their actual discriminating contribution to fault recognition. In the literature, several normalization techniques are presented, such as the min-max and the standard scaling. Therefore, we compare the performances provided by both of them for CAE and DBUC. It turned out that CAE performs better considering min-max normalized samples, while DBUC provides more reliable results with standard normalization.

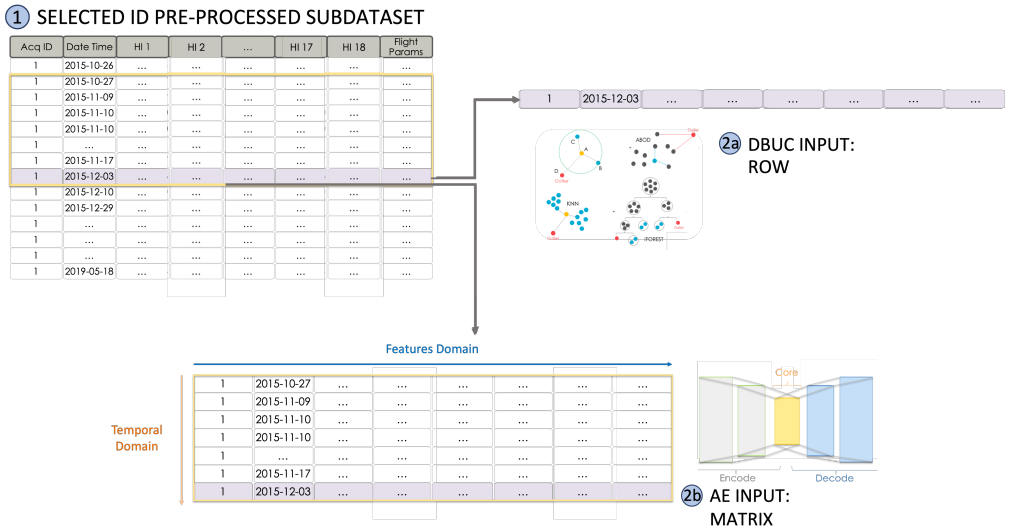



Figure 4: DBUC and CAE Input Data Structure.

*3.2.2. Convolutional Autoencoder Design*

Semi-supervised learning is a paradigm in which the classifier learns in the presence of both labeled and unlabeled samples, [41]. In particular, a CAE consists of a neural network where the input and output layers share the same number of units. It aims at reconstructing with the minimum amount of distortion the samples on which it is trained, see e.g., [42]. The amount of introduced distortion is evaluated using a loss function that computes the so-called reconstruction error.

In more detail, an AE is composed of encoding layers that progressively reduce the dimensionality of the input data until achieving a compressed representation at the bottleneck layer level, the core of the network, whose number of hidden units defines the maximum number of representative features allowed. This operation can be compared to non-linear principal component analysis in which the number of main components considered is equal to the number of hidden units in the bottleneck layer. The AE then tries to reconstruct the input data based on the minimum set of representative features through the decoding layers, whose extraction rules were learned during the training phase. Thus, the training set observations must belong to a known health condition for the monitored components, while the test set ones may belong to both healthy and anomalous distribution.

Provided that the classifier learned the characteristic pattern correctly for the healthy observations, the global reconstruction error must be small for healthy samples and increase for the anomalous ones. Therefore, as the reconstruction error is directly proportional to instances' outlyingness, so it can be interpreted as an anomaly score. Also, the reconstruction error can be investigated for each feature. This allows us quantifying the deviation of each HI from its normal behavior, and it is critical to provide insights on the most probable source of the reported damages.

As previously mentioned, a CAE is trained for each sub-dataset, referring to each acquisition ID. Even if the networks' weights and biases depend on the specific ID they refer to, the overall architecture, *i.e.,* the number of layers and hidden nodes, the activation function, and the learning rate are the same for each helicopter and each of the 180 related CAEs. The most suitable parameters set was identified according to a fine-tuning process, which considers the average network reconstruction performances as cost function. The choice of sharing the same CAE structure for each of the 180 networks is also supported by the fact that the acquisition ID-based extracted sub-datasets share the same semantic. The final encoder structure consists of three convolutional layers, composed of 16, 8, and 8 hidden units, respectively, and connected by MaxPooling2D layers. The decoder consists of the same structure but is mirrored. UpSampling layers are employed instead of the MaxPooling2D ones. Each convolutional layer has a 3x3 kernel and the ReLu as the activation function. Further detail about the autoencoder activation function and its layers can be found in [43].

The reconstruction error is computed as the Mean Average Error (MAE). In the scientific literature, approaches rely on both MAE and RMSE to assess the models performances, but evidence suggests that RMSE is less reliable and can be affected by the distribution of the errors magnitude, see e.g., [44, 45], which is a critical issue in anomaly detection systems, where of course specific errors can be very larger than others.

### *3.2.3. HIs Ranking for Fault Isolation*

The overall CAE's anomaly score for each instance is computed applying MAE, and considering the whole input and output matrices. Using this approach for each column allows us obtaining a feature-based anomaly score that increases as the associated feature deviates more from its expected behavior. For each processed instance, it is possible to rank the predictors basing on this quantity. This information is key:

- For the user. As HIs are specially designed to highlight certain damages, the damage associated with the HI characterized by the largest reconstruc-

tion error will be the most likely cause of a reported fault.

- For designing an ad-hoc filtering policy. Indeed, instances detected as anomalous by the system, but whose most deviant feature is an operating condition or a flight regime, can be neglected, thus providing an essential means to reduce the number of false alarms.

*3.2.4. Distance&Density-Based Unsupervised Classifier*

The unsupervised learning paradigm includes all those techniques to identify homogeneous groups in a set of observations, basing on their intrinsic patterns. The belonging of observation to a group can be assessed according to statistical, density-based, or distance-based methods. In our design, we decided to complement the AE with four learners that work with an unsupervised approach, yielding the DBUC classifier. In detail, the selected unsupervised techniques are the following:

- Angle-Based Outlier Detection (ABOD), a distance-based technique that compares the angles between pairs of distance vectors with respect to other points to distinguish inliers from outliers [46];
- Local Outlier Factor (LOF), a density-based technique that compares the local reachability densities of a sample and its neighbors [47];
- Isolation Forest (iFOREST), a density-based technique that iteratively splits the observation set by randomly selecting a feature an then selecting a split value within its range [48];
- K-Nearest Neighbors (KNN), a distance-based technique in which each observation is iteratively assigned to the group composed of the most similar k ones [49].

Each learner receives the 1-dimensional input instances described before, whose features includes the HIs and the referred operating conditions. Therefore, while CAE accounts for both features dependencies and temporal evolution, DBUC is focused on the patterns that characterizes the features collected

during a single acquisition phase. As output, each learner returns an anomaly score $s$ for each instance. To make the scores comparable, they are first regularized between $[0, +\inf)$, such that their regularized value $Reg(s) \approx 0$ for healthy samples and $Reg(s) >> 0$ for anomalous ones, and then normalized within the same range. Finally, the four scores given by each single learner are averaged, obtaining the final DBUC anomaly score. An overview of the unsupervised pipeline is provided in Figure 5.

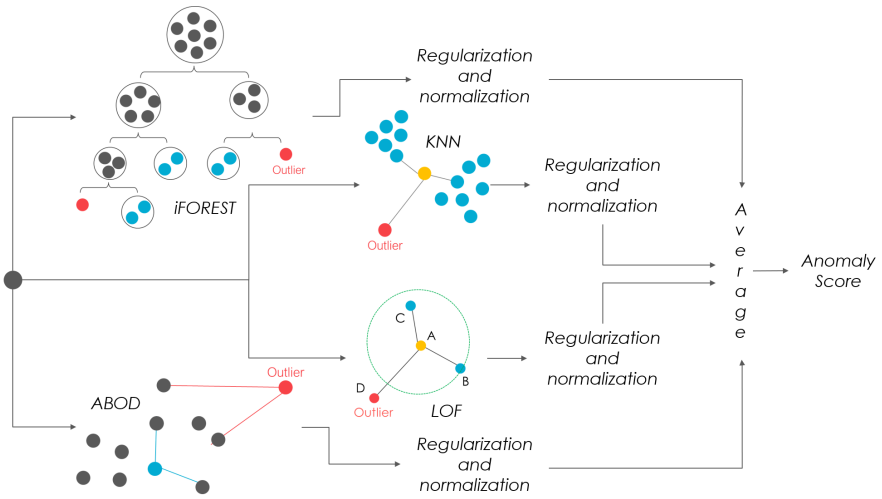



Figure 5: DBUC Pipeline: the unsupervised classifiers estimate separate anomaly scores, which are regularized and normalized before computing the output score as their average.

### *3.3. Component-Level Aggregation*

Each ensemble learner returns an anomaly score for each instance that quantifies its deviation from the expected behavior. This information is at the acquisition ID-level, while the system's primary focus should be to report the information at component-level, which is where maintenance should intervene. It follows that the first post-processing step is to locate the damage by mapping the acquisition IDs to the respective components. As explained in Section 2, an ID uniquely identifies an accelerometer-component pair. This relation allows the mapping algorithm to link each acquisition ID to the corresponding

component. Also, scores referred to the same component in the same flight and computed by different sub-dataset IDs are combined to provide an aggregated component-level score.

### *3.4. Post-Processing Filtering Policies*

The last post-processing step consists of applying the designed filtering policies to reduce false alarms. In detail, two filtering policies are used:

1. *HI as the most anomalous feature.* Recalling that often false alarms are due to HIs scattering caused by deviations in the operating condition and/or flight regime, the first filtering policy removes the scores whose most deviant feature is not an HI, but rather a flight regime or an operating condition.
2. *More than one in consecutive flights.* Assuming that the damage spread, from its inception on, is a continuous process, we assume that all vibration signatures related to the faults will deviate from the normal one from a certain moment on. Thus, the second policy is specifically designed to filter out sporadic outliers. For this purpose, a seven-flights-long moving average is applied to the anomaly scores related to the same component. An alert is triggered only if at least two of these scores exceed a given threshold, which, following the adopted normalization, is set to the $3\sigma$ interval of the anomaly score for all helicopters.

## 4. Discussion of the Obtained Results

This section discusses the performance of the proposed active monitoring system for the detection of anomalies in the helicopter's transmission. The output predictions are at first presented at acquisition ID-level, and then aggregated at component-level. Finally, HIs ranking based on the CAE reconstruction error is also presented to demonstrate the capability of the system to offer interpretable insights about the damage's causes.

### *4.1. Definition of the Evaluation Metrics*

Defining effective metrics of HUMS performance assessment is a critical task, see [50], [51]. In our context, the most crucial one is undoubtedly the *true positive* rate, as it quantifies the system's capability to identify the actual faults. It is also worth considering the *false positive* rate, which detects the number of classified anomalous instances that are not linked to an actual fault. Another evaluation criterion we aim at determining is the predictive capability of our system. It corresponds to the time interval between the actual fault and the first anomalous instance reported for the same component. This metric is critical to assess the effectiveness of our system for predictive maintenance purposes. Finally, we are also interested in evaluating the alert intensity, defined as the distance of an anomalous score from the normality threshold. Specifically, the intensity referred to the $i-th$ instance is computed as

$$Intensity(i) = 1 - \frac{th}{as(i)} \tag{2}$$

where $th$ represents the threshold value, and $as(i)$ the anomaly score of the $i_{th}$ instance.

### *4.2. Experimental Results*

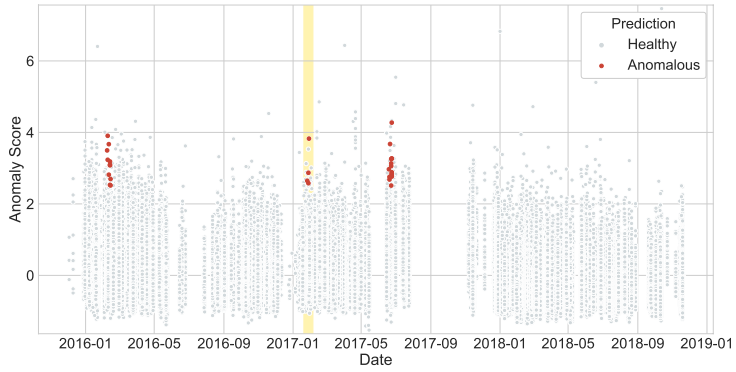
As explained in Section 3, the anomaly scores predicted analyzing helicopter data can be investigated at different aggregation levels. Considering the acquisition ID point of view, the predicted scores are reported in Figure 6. First, we can notice that no false alarm is triggered for the healthy helicopter, whose predictions are reported in Figure 6c. The first group of anomalous scores, referred to instances recorded on October $11^{th}$ 2017, is filtered out by the *HI as the most anomalous feature* policy. We can thus deduce that the scattering of the anomaly scores was due to a variation in the acquisition condition and not to an actual fault. Also, considering the healthy helicopter's results, it is possible to underline the effectiveness of the *More than one in consecutive flights* filtering policy since the remaining scores above the threshold do not trigger any alarm. This means that they are outliers which are either isolated or not referred to

the same component.

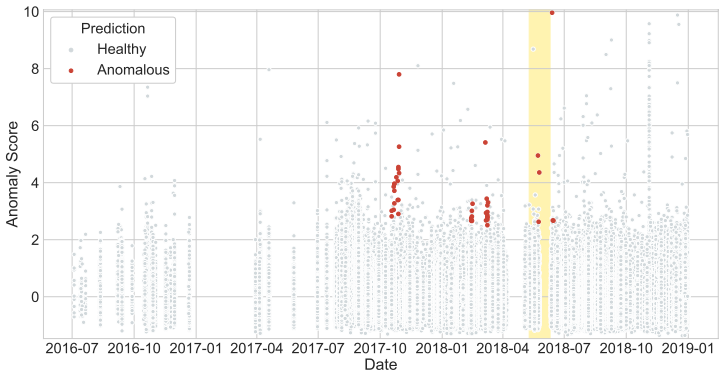
Considering Figure 6a and Figure 6b, showing the system's output referred to the helicopters that experienced a fault, the importance of the *More than one in consecutive flights* filtering policy is even more compelling. Indeed, it allows us isolating only three and four groups of alarms over the whole 4-year observation period, for a total of 27 and 41 instances reported as anomalous, respectively. The anomaly scores reported as fault-related that fall into the yellow area, which highlights the week preceding the damage inspection, correspond to the true faults. The other groups of anomalous instances, instead, are supposed to represent false positives. Therefore, the false positive rate results were 0.02% for the first helicopter and 0.03% for the second one, a very low value which confirms the robustness of the approach. The true positive rate is 100% for both the helicopters that experienced the faults.

Figure 7 shows the scores aggregated at component-level for the two helicopters experiencing a fault during the observation period. No component-aggregated view is reported for the third helicopter, as no fault was experienced.

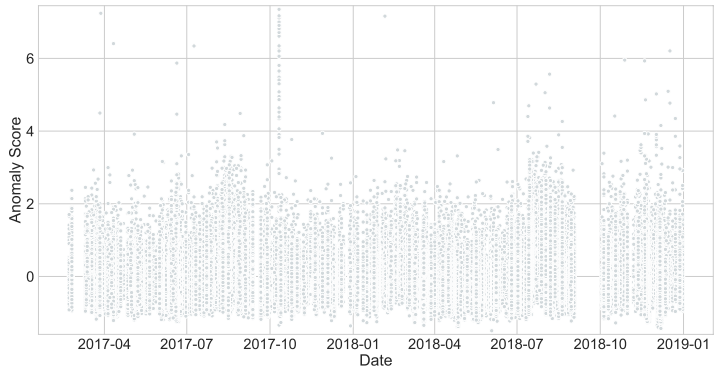
Only the scores referred to the faulty component are reported in Table 4. In detail, the first helicopter was affected by a swashplate fault occurred on January $30^{\text{th}}$ 2017, while the second one experienced a gear bearing fault on May $25^{\text{th}}$ 2018. These results show that the only reported trigger for the damaged component over the whole observation period corresponds to the actual fault. Therefore, no false positives are present at component level. Considering the predictive capability, the first alert in the first helicopter is triggered two days before the actual fault, while in the second one, it occurs three days before. The anomaly score reported in the actual fault day is 3.83 for the first helicopter, giving an intensity of 21.7%, and 3.96 for the second one, giving an intensity of 24.2%. Thus, the score computation make the anomalous samples emerge from the healthy one with good resolution.




(a) First Helicopter

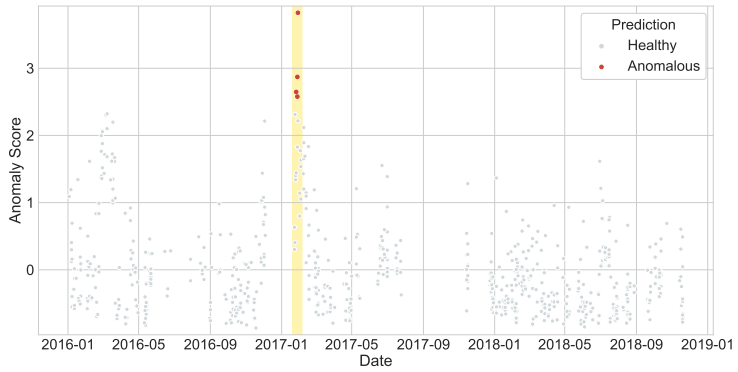



(b) Second Helicopter

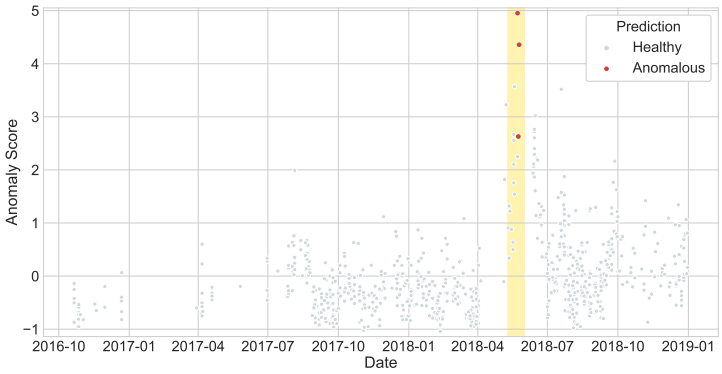



(c) Third Helicopter

Figure 6: Acquisition ID-Level Anomaly Scores. Anomalous instances are represented by red dots.




(a) First Helicopter - Swashplate




(b) Second Helicopter - Gear Bearing

Figure 7: Component-Level Anomaly Scores: the figure shows the anomaly scores predicted for all the vibration signatures related to the fault component of the first helicopter (left side) and of the second one (right side). The red dots show the anomalies triggered by our system.

Table 4: Performance Evaluation of the Overall Active Monitoring System.

| HC | True Positives Rate | | False Positive Rate | | Predictive Capability | Intensity |
|---|---|---|---|---|---|---|
| | **Whole HC** | **Component** | **Whole HC** | **Component** | | |
| 1 | 100% | 100% | 0.02% | 0% | 3 Days | 21.7% |
| 2 | 100% | 100% | 0.03% | 0% | 2 Days | 24.2% |

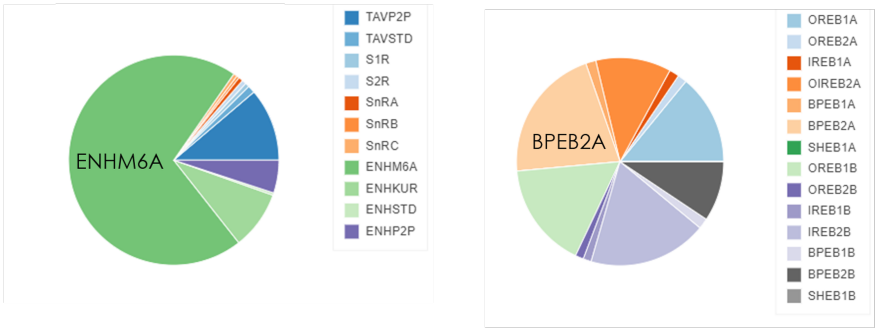
### *4.3. Investigation of the Triggered Alerts*

As explained in Section 3, it is possible to obtain a ranking for each instance based on the CAE's reconstruction error. However, this information is relevant only for those instances identified as anomalous by the system. Indeed, investigating the reconstruction error in HIs reconstruction allows us to establish the most probable cause of a detected fault. Indeed, the filtering *HI as the most anomalous feature* policy ensures that the reported anomalies are due to real HIs anomalies and not to variations in the acquisition conditions.

Figure 8 shows the HIs ranking referred to the detected faults. Considering the whole reconstruction error committed by the AE for the HIs values of the most anomalous instance identified analyzing the first helicopter, our framework suggests that the HI that most deviates from its expected behavior is ENHM6A. This HI is suitably defined to enhance the presence of pits, spalls or cracks starting from a gear surface. Therefore, this failure can be considered as the most probable cause of damage for the transmission's component referred to the considered instance. Considering the second helicopter, the most probable cause of the reported fault is the presence of localized pits, spalls, cracks or debris over the surface of the bearings, since those are the damages that the BPEB2A HI is supposed to enhance.

## 5. Concluding Remarks and Future Works

This work presented a diagnostic and monitoring system for the early detection of mechanical degradation in helicopters' transmission components. The main innovative contribution is given by the system capability to report, along with the detected faulty condition, its most probable cause in an interpretable




(a) First Helicopter (b) Second Helicopter

Figure 8: Percentage error in the reconstruction of each HIs value with respect to the whole reconstruction error of all HIs. These results are shown for the first helicopter's fault (left) and the second one's fault (right) and provide insights of the most probable cause of the reported damage.

manner, relating it to a specific component. When tested on real data, the proposed tool proves to have very good and consistent performance and significant predictive capabilities. This is achieved thanks to an ensemble of both supervised and unsupervised methods, optimally blended and paired with effective filtering policies that further minimize false positives. Ongoing work is being devoted to further test the model performance by applying it on a wider ranges of different machines.

## References

[1] A. Safety, Security program, the helicopter accident analysis team, Final Report of the Helicopter Accident Analysis Team (1998).

[2] C. A. Authority, Review of helicopter airworthiness, report of the helicopter airworthiness review panel (harp), CAP419 (1985).

[3] J. E. Land, Hums-the benefits-past, present and future, in: 2001 IEEE Aerospace Conference Proceedings (Cat. No. 01TH8542), Vol. 6, IEEE, 2001, pp. 3083–3094.

[4] L. Zhou, F. Duan, M. Corsar, F. Elasha, D. Mba, A study on helicopter main gearbox planetary bearing fault diagnosis, Applied Acoustics 147 (2019) 4–14.

[5] I. Manarikkal, F. Elasha, D. Mba, Diagnostics and prognostics of planetary gearbox using cwt, auto regression (ar) and k-means algorithm, Applied Acoustics 184 (2021) 1–16.

[6] Y. Kong, F. Chu, Z. Qin, Q. Han, Sparse learning based classification framework for planetary bearing health diagnostics, Mechanism and Machine Theory 173 (2022) 1–25.

[7] F. Elasha, X. Li, D. Mba, A. Ogundare, S. Ojolo, A novel condition indicator for bearing fault detection within helicopter transmission, Journal of Vibration Engineering & Technologies 9 (2) (2021) 215–224.

[8] M. A. Hassan, M. R. Habib, A. M. Bayoumi, Detection and classification of helicopter drive shaft faults using neuro-fuzzy based on wavelet power spectrum algorithm, in: Advances in Asset Management and Condition Monitoring, Springer, 2020, pp. 437–450.

[9] J. Hu, N. Hu, Y. Yang, L. Zhang, G. Shen, Nonlinear dynamic modeling and analysis of a helicopter planetary gear set for tooth crack diagnosis, Measurement (2022) 1–17.

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

[11] C. Li, L. Ru, Prognostics and health management techniques for integrated avionics systems, in: 2019 Prognostics and System Health Management Conference (PHM-Qingdao), IEEE, 2019, pp. 1–5.

[12] R. Wang, L. Yu, C. Wei, S. Ma, W. Zhang, Z. Chen, L. Yu, Aerodynamic noise separation of helicopter main and tail rotors using a cascade filter with

vold-kalman filter and cyclic wiener filter, Applied Acoustics 192 (2022) 1–15.

[13] A. Mauricio, W. Wang, J. Antoni, K. Gryllias, Advanced signal processing techniques for helicopter's gearbox monitoring, Aerospace Science and Technology 1909 (1) (2021) 1–8.

[14] Q. Ni, J. Ji, K. Feng, Data-driven prognostic scheme for bearings based on a novel health indicator and gated recurrent unit network, IEEE Transactions on Industrial Informatics (2022).

[15] W. Tang, Y. Chen, M. J. Zuo, Health index development for a planetary gearbox, Procedia Manufacturing 49 (2020) 155–159.

[16] F. Elasha, C. Ruiz-Carcel, D. Mba, P. Chandra, A comparative study of the effectiveness of adaptive filter algorithms, spectral kurtosis and linear prediction in detection of a naturally degraded bearing in a gearbox, Journal of Failure Analysis and Prevention 14 (5) (2014) 623–636.

[17] L. Zhou, F. Duan, D. Mba, W. Wang, S. Ojolo, Using frequency domain analysis techniques for diagnosis of planetary bearing defect in a ch-46e helicopter aft gearbox, Engineering Failure Analysis 92 (2018) 71–83.

[18] L. Zhou, F. Duan, M. Corsar, F. Elasha, D. Mba, A study on helicopter main gearbox planetary bearing fault diagnosis, Applied Acoustics 147 (2019) 4–14.

[19] P. J. Dempsey, D. G. Lewicki, D. D. Le, Investigation of current methods to identify helicopter gear health, in: 2007 IEEE Aerospace Conference, IEEE, 2007, pp. 1–13.

[20] M. Mosher, E. M. Huff, E. Barszcz, Analysis of in-flight measurements from helicopter transmissions, in: American Helicopter Society 60th Annual Forum, Citeseer, 2004, pp. 1–14.

[21] M. L. Mimnagh, W. Hardman, J. Sheaffer, Helicopter drive system diagnostics through multivariate statistical process control, in: 2000 IEEE Aerospace Conference. Proceedings (Cat. No. 00TH8484), Vol. 6, IEEE, 2000, pp. 381–415.

[22] C. A. Authority, Intelligent management of helicopter vibration health monitoring report, CAA Paper 1 (2011) 2011.

[23] V. Camerini, G. Coppotelli, S. Bendisch, Fault detection in operating helicopter drivetrain components based on support vector data description, Aerospace Science and Technology 73 (2018) 48–60.

[24] T. Li, Z. Zhao, C. Sun, R. Yan, X. Chen, Adaptive channel weighted cnn with multisensor fusion for condition monitoring of helicopter transmission system, IEEE Sensors Journal 20 (15) (2020) 8364–8373.

[25] S. Ferreiro, A. Arnaiz, B. Sierra, I. Irigoien, Application of bayesian networks in prognostics for a new integrated vehicle health management concept, Expert Systems with Applications 39 (7) (2012) 6402–6418.

[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.

[27] S. Sharma, A. Gosain, S. Jain, A review of the oversampling techniques in class imbalance problem, in: International Conference on Innovative Computing and Communications, Springer, 2022, pp. 459–472.

[28] J. U. Ko, K. Na, J.-S. Oh, J. Kim, B. D. Youn, A new auto-encoder-based dynamic threshold to reduce false alarm rate for anomaly detection of steam turbines, Expert Systems with Applications 189 (2022) 1–20.

[29] B. X. Yong, A. Brintrup, Bayesian autoencoders with uncertainty quantification: Towards trustworthy anomaly detection, Expert Systems with Applications (2022) 1–54.

[30] T. H. Mohamad, A. Abbasi, E. Kim, C. Nataraj, Application of deep cnn-lstm network to gear fault diagnostics, in: 2021 IEEE International Conference on Prognostics and Health Management (ICPHM), IEEE, 2021, pp. 1–6.

[31] Y. Kong, Z. Qin, T. Wang, M. Rao, Z. Feng, F. Chu, Data-driven dictionary design–based sparse classification method for intelligent fault diagnosis of planet bearings, Structural Health Monitoring 21 (4) (2022) 1313–1328.

[32] C. Sun, Y. Liu, L. Huang, Helicopter planetary gear crack fault identification utilizing multidomain stacked contractive autoencoders based deep learning framework, in: 2021 Global Reliability and Prognostics and Health Management (PHM-Nanjing), 2021, pp. 1–8.

[33] R. Singh, B. Bhushan, Fault classification using support vectors for unmanned helicopters, Computational Methods and Data Engineering (2021) 369–384.

[34] V. Malviya, I. Mukherjee, S. Tallur, Edge-compatible convolutional autoencoder implemented on fpga for anomaly detection in vibration condition-based monitoring, IEEE Sensors Letters 6 (4) (2022) 1–4.

[35] B. P. Bogert, The quefrency alanysis of time series for echoes; cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking, Time series analysis (1963) 209–243.

[36] W. Wang, P. McFadden, Early detection of gear failure by vibration analysis i. calculation of the time-frequency distribution, Mechanical Systems and Signal Processing 7 (3) (1993) 193–203.

[37] J. Antoni, R. Randall, The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines, Mechanical systems and signal processing 20 (2) (2006) 308–331.

[38] G. Dalpiaz, A. Rivola, R. Rubini, Effectiveness and sensitivity of vibration processing techniques for local fault detection in gears, Mechanical systems and signal processing 14 (3) (2000) 387–412.

[39] D. G. Childers, D. P. Skinner, R. C. Kemerait, The cepstrum: A guide to processing, Proceedings of the IEEE 65 (10) (1977) 1428–1443.

[40] D. F. Andrews, Plots of high-dimensional data, Biometrics (1972) 125–136.

[41] X. Zhu, A. B. Goldberg, Introduction to semi-supervised learning (synthesis lectures on artificial intelligence and machine learning), Morgan and Claypool Publishers 14 (2009).

[42] P. Baldi, Autoencoders, unsupervised learning, and deep architectures, in: Proceedings of ICML workshop on unsupervised and transfer learning, 2012, pp. 37–49.

[43] A. Gulli, S. Pal, Deep learning with Keras, Packt Publishing Ltd, 2017.

[44] C. J. Willmott, K. Matsuura, Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance, Climate research 30 (1) (2005) 79–82.

[45] Y. Ouyang, W. Liu, W. Rong, Z. Xiong, Autoencoder-based collaborative filtering, in: International conference on neural information processing, Springer, 2014, pp. 284–291.

[46] H.-P. Kriegel, M. Schubert, A. Zimek, Angle-based outlier detection in high-dimensional data, in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 444–452.

[47] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander, Lof: identifying density-based local outliers, in: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 2000, pp. 93–104.

[48] F. T. Liu, K. M. Ting, Z.-H. Zhou, Isolation forest, in: 2008 Eighth IEEE International Conference on Data Mining, IEEE, 2008, pp. 413–422.

[49] N. S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, The American Statistician 46 (3) (1992) 175–185.

[50] E. Bechhoefer, A. P. Bernhard, Setting hums condition indicator thresholds by modeling aircraft and torque band variance, in: 2004 IEEE Aerospace Conference Proceedings (IEEE Cat. No. 04TH8720), Vol. 6, IEEE, 2004, pp. 3590–3595.

[51] E. Bechhoefer, A. P. Bernhard, A generalized process for optimal threshold setting in hums, in: 2007 IEEE Aerospace Conference, IEEE, 2007, pp. 1–9.