Received: 30 September 2022 | Revised: 5 May 2023 | Accepted: 6 July 2023

DOI: 10.1002/we.2857

RESEARCH ARTICLE

WILEY

Convolutional neural network framework for wind turbine electromechanical fault detection

Emilie Stone[1] | Stefano Giani[1] | Donatella Zappalá[2] | Christopher Crabtree[1]

[1]Department of Engineering, Durham University, Durham, UK

[2]Faculty of Aerospace Engineering, Delft University of Technology, Delft, The Netherlands

Correspondence
Stefano Giani, Department of Engineering, Durham University, South Road, Durham DH1 3LE, UK.
Email: stefano.giani@durham.ac.uk

Funding information
Engineering and Physical Sciences Research Council, Grant/Award Number: EP/L014106/1

Abstract

Effective and timely health monitoring of wind turbine gearboxes and generators is essential to reduce the costs of operations and maintenance activities, especially offshore. This paper presents a scalable and lightweight convolutional neural network (CNN) framework using high-dimensional raw condition monitoring data for the automatic detection of multiple wind turbine electromechanical faults. The proposed approach leverages the potential of combining information from a variety of signals to learn features and to discriminate the types of fault and their severity. As a result of the CNN layers used to extract features from the signals, this architecture works in the time domain and can digest high-resolution multi-sensor data streams in real-time. To overcome the inherent black-box nature of AI models, this research proposes two interpretability techniques, multidimensional scaling and layer-wise relevance propagation, to analyse the proposed model's inner-working and identify the signal features relevant for fault classification. Experimental results show high performance and classification accuracies above 99.9% for all fault cases tested, demonstrating the efficacy of the proposed fault-detection system.

KEYWORDS
condition monitoring, convolutional neural network, deep learning, fault detection, gearbox, generator, multi-sensor data

1 | INTRODUCTION

The energy sector accounts for more than three-quarters of the EU's greenhouse gas emissions.[1] Taking energy efficiency measures and scaling up renewable energy are keys to decarbonizing the EU's energy system. Wind energy will be the pillar of a resilient, cost-effective and climate-neutral energy system by 2050.[2] Technological and commercial developments as well as industry maturation have already made onshore wind one of the most cost-efficient forms of power generation across Europe. Offshore wind cost decline has also accelerated over the past decade,[3] and experts predict substantial further cost reductions by 2050.[4] The rapid growth of wind turbine capacity, especially in remote and harsh offshore environments, poses the challenge of improving the reliability and cost-effectiveness of operations and maintenance (O&M) strategies to increase the profitability and availability of assets. While the cost of wind energy has reduced, mainly due to smaller upfront costs and improved

Abbreviations: AI, artificial intelligence; CNN, convolutional neural network; LRP, layer-wise relevance propagation; MDS, multidimensional scaling; NI, National Instruments; O&M, operations and maintenance; SCADA, supervisory control and data acquisition; WRIG, wound rotor induction generator.

performance, the cost of O&M activities still represents a major contributor to total expenditure, especially offshore where it is estimated to account for up to 35% of the total cost of wind energy.[5]

Unplanned downtime, associated with unexpected failures and extended overhauls or replacements of major components, represents one of the main cost drivers of O&M activities. Understanding when and how a component fails is crucial for minimizing unplanned outages and developing profitable preventive condition-based maintenance approaches.[6] A wind turbine's gearbox and generator are reliability critical as they have the greatest share of downtime among all other components.[7] Hence, their optimal monitoring and maintenance are key factors for cost-effective O&M.

In recent years, the health monitoring of these components has been the subject of intense research by the scientific community with most existing studies focusing on signal processing or physics-based numerical models as reviewed in previous works.[8–12] Since 2015, there has been a rapid interest and growth in applications of supervised and unsupervised artificial intelligence (AI) models, mainly relying on the use of low-frequency (usually 10-min averaged) Supervisory Control and Data Acquisition (SCADA) signals for data-driven decision-making.[12–15] Although recent literature[13,16–20] clearly shows the strong potential and promising results of deep learning techniques for wind turbine fault detection, there are still key challenges which limit their industry implementation.[12]

The low dimensionality of SCADA data has been identified as a limiting factor for the full exploitation of the potential of deep learning techniques.[13] High-frequency signals contain richer information than SCADA data as they can better capture short-term dynamic wind turbine behaviours. SCADA data aggregation over a period of 10-min results in a loss of information, potentially reducing its capabilities for detecting wind turbine component health deviations. Recent works[21,22] have shown improved performance and detection ability, and higher sensitivity to deviations, for AI models using high-frequency signals. Convolutional neural networks (CNNs) work very well with high-dimensional data. The real-world applications of CNNs are wide-ranging and include self-driving cars, advertising, cancer detection, auto-translation between different languages, document classification and facial recognition. Deep neural network (NN) applications for wind turbine fault detection can benefit from the more fine-grained view of data that is provided by a higher temporal resolution.[13] AI models have been applied to high-frequency vibration data in a few works; however, these have usually first applied classical signal processing approaches to generate spectra, such as in the approach proposed by Bach-Andersen et al,[20] or to extract low-dimensional features based on theoretical and semi-empirical fault models, as proposed by Strączkiewicz and Barszcz.[23]

Due to their structure, CNNs take advantage of local spatial coherence in the input and are therefore very good feature extractors. Computationally, CNNs are efficient, especially on GPUs, because they leverage two important ideas: sparse interaction and parameter sharing. Sparse interaction is accomplished by making the kernel smaller than the input, while parameter sharing improves the speed of the network by letting the kernels slide across the data. This allows CNNs to be successfully applied for many different tasks: image classification,[24] 3D object classification,[25] denoising[26] and data segmentation.[27,28]

Another limitation of the deep learning techniques proposed so far is that each mainly focuses on one turbine component only, such as the gearbox,[29] the generator[30] or the converter,[31] and this might limit their implementation in the field. While they show high fault prediction performance, the studies proposed in the literature so far do not generally provide rationales and transparency in their decisions, and this may make turbine operators reluctant to adopt such approaches.[12] Models are becoming increasingly complex, and it is sometimes unclear, even to experts, how they produce specific outputs. This poses an increasing concern about their trustworthiness. Currently, the accuracy of deep learning techniques is very high, but their social acceptance is very low. The main reason for this is the lack of interpretability and transparency of the models making their integration into our daily lives challenging. It is very hard to trust techniques that have a significant impact on our lives and future without understanding them. In layman's terms, interpretability is the level to which a human can understand the process used by a machine to reach a certain decision. Several interpretability techniques have been proposed in the literature and applied to different domains with the aim of understanding the rationale behind the decisions made by a model. Some of the most common techniques are activation maximization,[32] sensitivity analysis[33] and backward propagation.[32]

This paper proposes a modular CNN framework for real-time automatic classification of turbine component conditions without requiring additional signal processing and expert domain knowledge, which are both labour-intensive and time-consuming. The performance and effectiveness of the CNN framework are evaluated using experimental data from a condition monitoring test rig simulating the variable operating conditions of a wind turbine drive train. The proposed research aims to alleviate the limitations of previous works in the field by:

- leveraging higher dimensional raw condition monitoring data collected from operational wind turbines, such as vibration, speed and power signals, while also exploring the potential of using other signals not commonly recorded in the field, such as mechanical torque and shaft displacement;
- tailoring the model's parallel convolutional branches to tackle the most critical turbine components within a holistic fault-detection architecture;
- combining high-resolution monitoring signals for fault type (electrical or mechanical) and severity level (from early stages to the most severe damage conditions) classification;
- adopting two interpretability techniques to overcome the inherent black-box nature of AI models.

The rest of this paper is organized as follows. The experimental set-up and the condition monitoring dataset used in this work are presented in Section 2, the architecture of the proposed modular CNN framework for wind turbine drive train fault detection is detailed in Section 3, the

results of the case study are presented in Section 4, the interpretability of the results is presented and discussed in Section 5 and conclusions are drawn in Section 6.

2 | EXPERIMENTAL TEST RIG AND DATASET

This work aims to develop a CNN framework for wind turbine electromechanical drive train fault detection using multiple high-frequency monitoring signals. Data collected from the wind turbine condition monitoring test rig developed at Durham University (UK)[34,35] were used to explore and demonstrate the fault-detection capability of the proposed model across different operational regimes in the drive train. This allowed the authors to perform experiments and train the model under normal and different electrical and mechanical seeded-fault severity conditions in a controllable and reproducible environment, enabling validation under deterministic circumstances.

The test rig, shown in Figure 1, was designed to act as a model of a wind turbine drive train with the purpose of producing signals comparable to those encountered on operational turbines. A 54 kW DC motor, which simulates the wind turbine rotor input, drives a four-pole, 30 kW wound rotor induction generator (WRIG) through a two-stage gearbox with a gear ratio of 1:5. The test rig is controlled via a National Instruments (NI) LabVIEW environment and can be driven under either constant or variable speed conditions. Variable speed testing is performed using driving data derived from a highly detailed 2 MW exemplar turbine model operating under torque control developed by the University of Strathclyde as part of the SUPERGEN Wind Energy Technologies Consortium.[36] The model incorporates the properties of natural wind and simulates the mechanical behaviour of a 2 MW variable speed turbine operating under closed-loop conditions both below and above the rated wind speed. Two wind speeds and turbulence intensities (TI) are used in this work: a low mean wind speed with low turbulence (7.5 m/s mean, 6% TI), representative of the turbine operating at or below rated wind speed under generator torque control and a high mean wind speed with high turbulence (15 m/s mean, 20% TI), representative of the turbine operating above-rated wind speed under blade pitch control. These wind speeds and turbulence levels have been selected because they are representative of two significantly different operating conditions and allow the authors to provide an indication of the model's robustness. Wind profiles were input to the 2 MW model and the experimental driving conditions implemented in LabVIEW were obtained by scaling the model generator speed data to the test rig, as described in detail in Crabtree.[34]

During the experimental campaign, healthy conditions and two types of faults were investigated through seeded-fault testing performed on two different test rig configurations:

1. Gearbox tooth damage on the high-speed pinion, representing the effect of a mechanical fault in the wind turbine gearbox. Using the test rig configuration shown in Figure 2, progressive damage to the leading contact edge of one tooth of the gearbox pinion was introduced by manual filing during the experiments, based upon photographs from actual damage to wind turbine gear teeth.[34] The fault was progressively moved across the tooth face until the entire tooth was removed except for a small amount of the root. The healthy pinion and the four faults of increasing severity investigated during the experiments are shown in Figure 3.
2. Generator rotor electrical asymmetry, representing the effect of a winding fault, brush imbalance or air gap eccentricity in the wind turbine generator. In this case, experiments were performed using the same rig configuration but with the gearbox removed, as shown in Figure 4, in order to allow the generator to carry a higher applied load. The rig WRIG was operated with the rotor circuit coupled via slip rings to an adjustable three-phase resistive load bank. This configuration allowed to increase the rotor balanced phase resistances for a wider generator speed variation, and to introduce asymmetry into one phase of the rotor circuit under controllable conditions. Two seeded-fault levels were investigated on the test rig rotor by increasing the resistance of Phase 1 of the rotor of 0.3 and 0.6 Ω, corresponding to 21% and 43% rotor electrical asymmetry, respectively.[35] These asymmetry levels are comparable to those used in previous studies.[37–39]

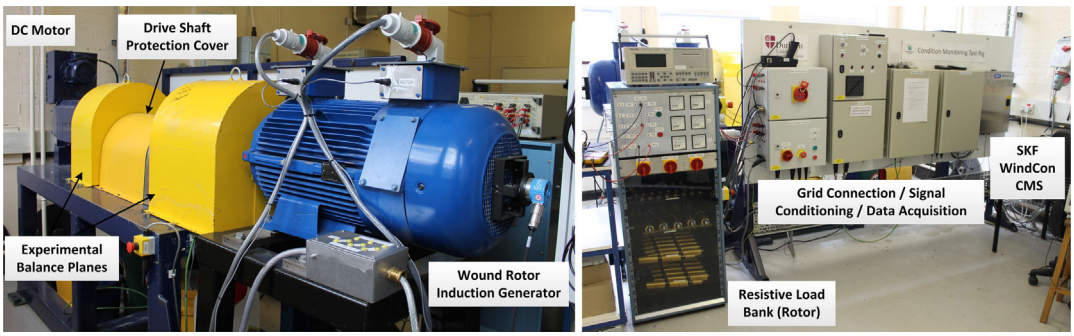

FIGURE 1 Wind turbine condition monitoring test rig equipped with the gearbox: main components (left), instrumentation and control systems (right).[34,35]

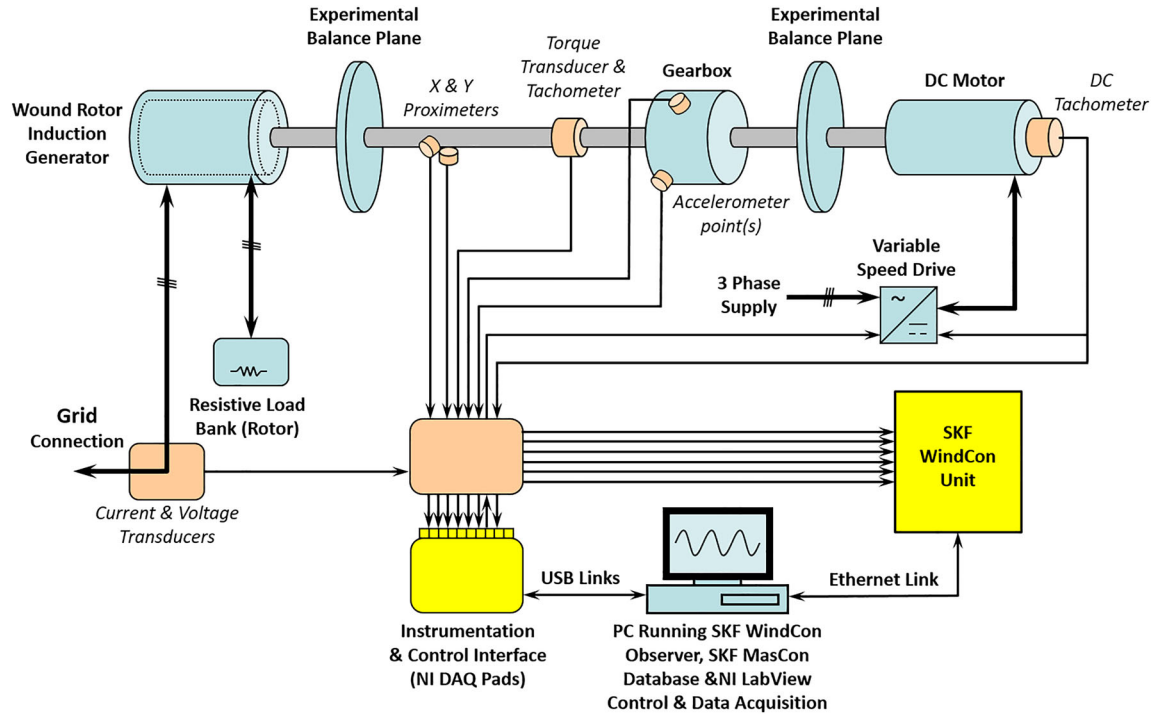

FIGURE 2 Schematic diagram of the test rig equipped with the gearbox.[34]

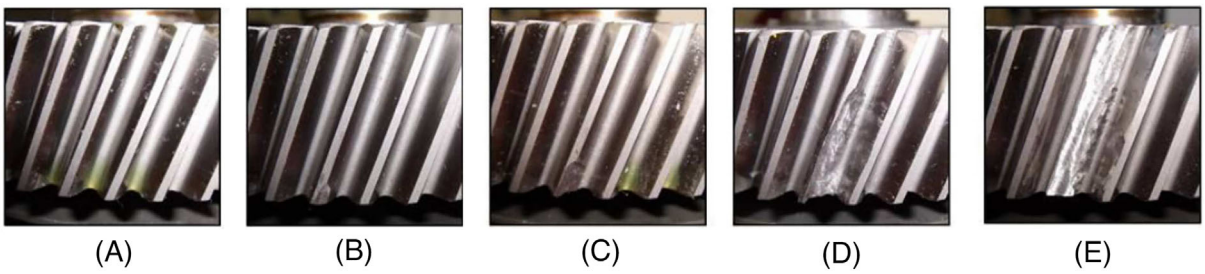

FIGURE 3 Gearbox pinion conditions investigated during the seeded-fault tests: (A) healthy (H), (B) 3 mm × 2 mm chip (F1 Gear), (C) 7 mm × 5 mm chip (F2 Gear), (D) 18 mm × 6 mm chip (F3 Gear) and (E) missing tooth (F4 Gear).[34]

During both sets of experiments, the following signals were acquired at a sampling frequency of 5 kHz using two NI 6015 data acquisition pads connected to the NI LabVIEW environment:

- Low-speed shaft speed (hereafter denoted as LSTacho): measured with a Eurotherm Drives tachometer fitted to the DC machine no-load side end-plate.
- High-speed shaft speed (hereafter denoted as HSTacho) and torque: measured using a Magtrol TMB 313/431 torque transducer, with a rated torque of 500 Nm and sensitivity of 10^{-2} V/Nm, mounted on the rig high-speed shaft. The transducer outputs 60 pulses per revolution, and it is also used as a high-speed shaft pulse tachometer.
- Two vibration signals (hereafter denoted as Acc1 and Acc2): measured using two Brüel&Kjær 4513-002 Deltatron piezoelectric shear accelerometers with integral electronics, a sensitivity of 500 mV/g, a frequency range of 1 to 10 kHz and 10 g peak dynamic measuring range. In the test rig configured with the gearbox (Figure 2), the accelerometers were located on the gearbox casing and on the high-speed end, respectively. In the test rig configuration where the WRIG was driven without the gearbox (Figure 4), the two accelerometers were fitted vertically and horizontally to the generator load side end-plate, at positions radial to the shaft.
- Horizontal and vertical shaft displacement (hereafter denoted as HShD and VShD): measured using two Kaman KD2300-4S1 eddy current proximeters located at 90° to each other on the rig high-speed shaft at the generator.

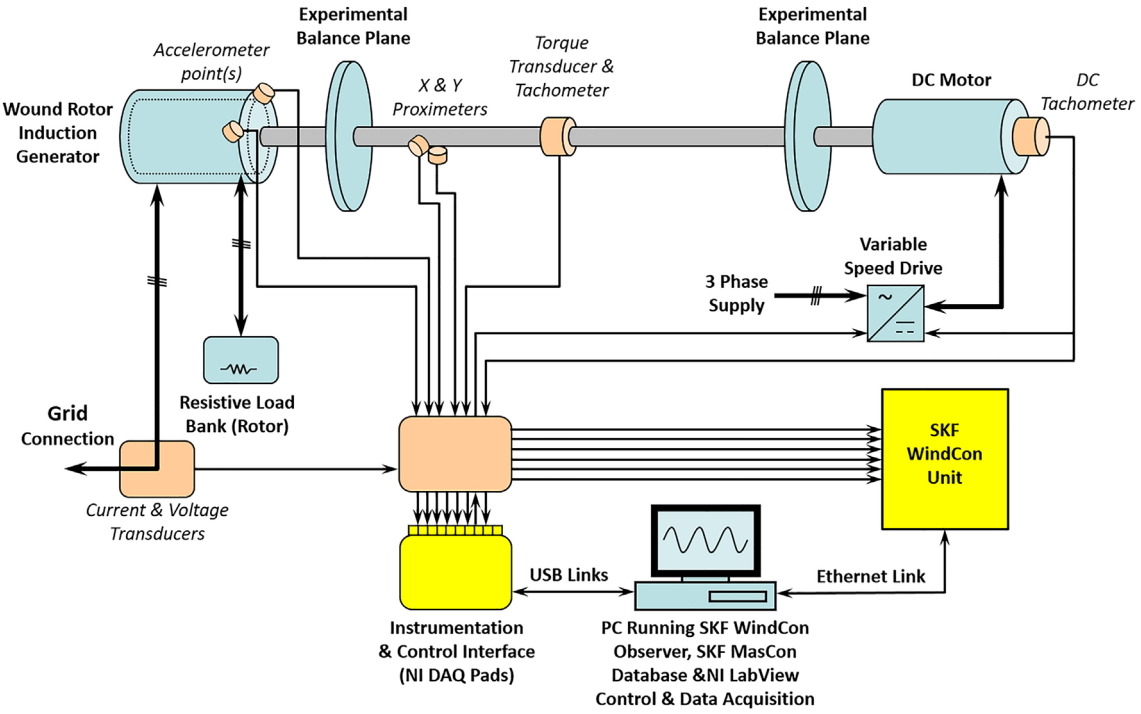

FIGURE 4 Schematic diagram of the test rig without the gearbox.[35]

TABLE 1 Summary of the experimental work performed on the wind turbine condition monitoring test.

Seeded fault	Test rig configuration	Fault level	Test speed	Test duration
High-speed pinion tooth damage	WRIG driven through the gearbox (Figure 2)	Healthy (H)	Constant,	150 s
		3 × 2 mm chip (F1 Gear)	1550 rpm	
		7 × 5 mm chip (F2 Gear)	Variable,	150 s
		18 × 6 mm chip (F3 Gear)	7.5 m/s, 6% TI	
		Missing tooth (F4 Gear)	15 m/s, 20% TI	
Rotor electrical asymmetry	WRIG driven without gearbox (Figure 4)	Balanced Rotor (H)	Constant,	300 s
		21% asymmetry (F1 Rotor)	1540 rpm	
		43% asymmetry (F2 Rotor)	1580 rpm	
			1590 rpm	
			Variable	450 s
			7.5 m/s, 6% TI	

- Three-phase generator stator terminal voltages (hereafter denoted as VoltagePh1, VoltagePh2 and VoltagePh3) and currents (hereafter denoted as CurrentPh1, CurrentPh2 and CurrentPh3): measured using transducer boards with a bandwidth of DC-100K Hz.
- Generator power signal: calculated with the two-wattmeter method, using the voltage and current measurements.

A summary of the experimental datasets collected from the Durham wind turbine condition monitoring test rig and used in this work is given in Table 1.

3 | CNN MODEL ARCHITECTURE

The proposed multi-input CNN model can be seen in Figure 5. The system is divided into feature extraction branches and the classifier head. The architecture is scalable for any number *N* of input signals and any number *M* of output classes. The inputs are *N* windows of data. Each window is

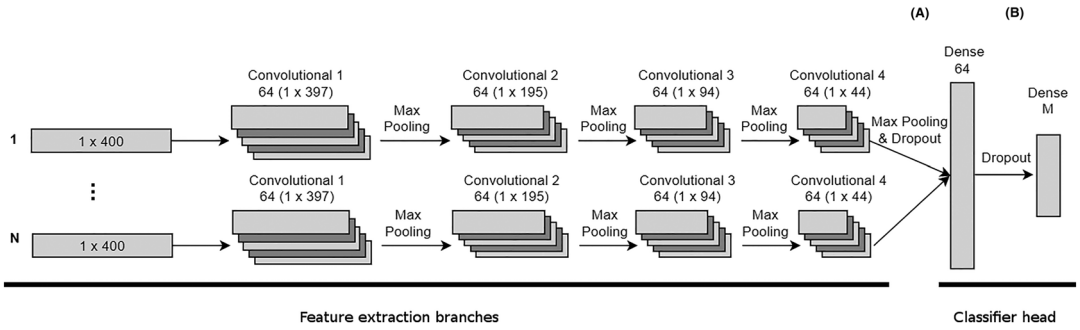

FIGURE 5 The proposed six-layer multi-input 1D CNN model structure. *N* separate convolutional branches consist of four 1D convolutional layers followed by max-pooling layers. Two fully connected dense layers of 64 and *M* units are used for classification. (A) denotes the place in the network where the outputs from the feature extraction branches are passed to the classifier head. (B) indicates the place in the network where the output from the first dense layer is passed to the neurons representing the different classes.

a 1D array containing a portion of a signal. Each window is normalized independently. The outputs are *M* values representing the confidence of the network in each class for the given inputs.

Feature extraction is performed by *N* separate convolutional branches allowing for *N* different signals to be analysed simultaneously. The benefit of using separate convolutional branches instead of one branch with multiple channels is that each branch can be customized for a specific signal type. For example, a different number of filters can be used in different branches. It can also be possible to vary the size of the inputs in different branches. This may be useful in case the sampling rate is not the same for all signals or for any other reason to decide to vary the size of the input window. In this study, it was found that the size of the input window of 400 points works well with all the analysed signals. However, the multi-branches architecture was kept to allow flexibility in future studies. The layout of the branches for the analysed signals consists of four 1D convolutional layers followed by max-pooling layers. Each convolutional layer has 64 kernels with a kernel size of 4, a stride of 1 and no padding. The ReLU activation function is used for all convolutional layers. In recent years, ReLU is considered the de facto choice for activation functions since it introduces non-linear behaviour in the NN improving performances.[40,41] Max pooling, with a pool size of 2, down-samples the data between the convolutional layers. Such a layout has been achieved through trial and error. To pass the final convolutional output to the dense layer, the *N* feature space matrices are flattened and concatenated to form one dense input vector.

CNN layers exploit a particular form of parameter sharing causing the layers to be equivariance to translation.[41] In other words, since the same kernels slide over the inputs to produce the output, if a shift is applied to the input, also the outputs are shifted. This is particularly useful in analysing images or time series because the kernels search everywhere in the inputs for patterns. On the other hand, every value in the output of CNN layers is processed using local information only.

Classification is achieved in the classifier head by two fully connected layers of 64 and *M* units. ReLU[42,43] activation functions are used in the first dense layer. The hyperparameter *M* allows the number of output classes to vary, this is necessary as different datasets may contain different numbers of health classes. Sigmoid activation functions are used for the final dense layer output. A dropout[44] rate of 0.5 is implemented before both dense layers to minimize overfitting. Dropout is a regularization technique that removes signals randomly during training. This forces the classifier head to rely on multiple features and patterns to classify the input windows making the decision process more robust.[41] The activation of the neurons in the dense layers is regulated by the equation

$$y = f\left(\sum_{j=1}^{m} w_j x_j\right), \tag{1}$$

where y is the activation value, f is the activation function, w_j are the weights learned during training and x_j are the input values from the previous layer. Both the vector of the weights and the input vector have size m that corresponds to the size of the output of the previous layer. In contrast to CNNs whose neurons only receive a portion of the inputs at any given time, neurons in dense layers have a complete view of the inputs. Each neuron in the first layer in the classifier head can see all the outputs from the feature extraction branches containing the strength of all the features extracted from the input window; as shown in (A) in Figure 5. Equation (1) allows those neurons to weight the different features in accordance with the learned weights for the classification. For the neurons in the last dense layer, as shown in (B) in Figure 5, equation (1) is used to identify the class of the input.

The Adam optimization algorithm[45] was used to train model weights, with exponential decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and $\epsilon = 1e-7$. The metric used during training is accuracy. The training was stopped when the validation accuracy had not improved for 10 epochs. The learning

rate was automatically adjusted during training starting from the value 10^{-3} and reduced 10 times every time a plateau was encountered. For the detection of plateaus, *patience* and a *cooldown* period of three epochs were used while any variation in the accuracy smaller than 10^{-4} was ignored. The categorical cross-entropy loss function was used during training:

$$\text{Loss} = \sum_{i=1}^{M} y_i \log \hat{y}_i,$$

where $\hat{y}_i$ is the ith scalar value in the model output and y_i is the corresponding target value.

4 | FAULT DETECTION—RESULTS

4.1 | Performance of single-input models

The performance of the proposed framework has been first estimated by running it as a single-input model. For each of the two faults investigated, the data collected during the tests has been split into three datasets for model training, testing and validation. It is common practice to split the data into 80% for training and 20% for testing.[41] However, a third set never seen during training is needed here for comparing different models and for hyperparameter optimization. Therefore, we split the data into three sets.[41] Those datasets contain 70%, 20% and 10% of the data, respectively. The training dataset is used to train the network, and the testing is used to evaluate the training after each epoch. The accuracy of the testing dataset is used to decide when to stop the training. The validation dataset is used to measure the accuracy of trained models. Since the validation dataset is not used during training, it can be seen as a measurement of the generalization capability of the network.

Tables 2 and 3 summarize the single-model results for the gearbox and generator rotor, respectively. For each signal, the data is split 10 times into training, testing and validation datasets using 10-fold cross-validation. Then for each split, the model was trained 10 times from randomly initialized weights. Overall, the model was trained 100 times on each signal. Tables 2 and 3 show the mean and the standard deviation values of accuracy on the validation sets over the 100 training. Accuracy is defined as the ratio of correct classifications over the total number of predictions. In both tables, the signals have been sorted in descending order with respect to the mean accuracy of their single-input models.

For both investigated seeded faults, overall, the model shows a very good performance for the case of the Power, Torque and HShD signals. This can also be observed in the case of the HSTacho signal for the gearbox dataset. The differences in the mean and standard deviation values obtained for the three current and voltage signals as well as for the two acceleration and displacement measurements can be explained as a consequence of slight variations in the intrinsic sensitivity of the sensors and in their locations within the test rig. In the case of the rotor dataset, the mean accuracy value for HShD is much higher than VShD suggesting that the horizontal shaft displacement is a more reliable signal for fault classification. It is also clear that, apart from the Power, in both cases, the model is unable to identify the correct health classes from the other

TABLE 2 Single-input model results for every considered signal from the gearbox dataset.

Signal	Mean	Std
HSTacho	0.999933	0.000203
Power	0.999564	0.000316
Torque	0.997062	0.000612
HShD	0.990858	0.003282
VShD	0.977920	0.003826
Acc2	0.968218	0.006072
Acc1	0.948413	0.003733
LSTacho	0.944520	0.003450
CurrentPh3	0.233271	0.145414
CurrentPh1	0.207778	0.077602
VoltagePh1	0.200227	0.001280
CurrentPh2	0.200067	0.001058
VoltagePh2	0.199951	0.000939
VoltagePh3	0.199933	0.000780

Note: The signals are sorted in descending order with respect to the mean accuracy.

TABLE 3 Single-input model results for every considered signal from the generator rotor dataset.

Signal	Mean	Std
Power	0.999294	0.000450
Torque	0.980498	0.003058
HShD	0.942658	0.004542
LSTacho	0.623923	0.007295
HSTacho	0.610000	0.225658
VShD	0.596594	0.195940
Acc2	0.524488	0.029262
Acc1	0.353846	0.031857
CurrentPh3	0.333606	0.001363
CurrentPh1	0.333408	0.000541
VoltagePh3	0.333394	0.000602
VoltagePh1	0.333335	0.000105
CurrentPh2	0.333308	0.000487
VoltagePh2	0.333304	0.000683

Note: The signals are sorted in descending order with respect to the mean accuracy.

TABLE 4 Four-input model results for the gearbox seeded-fault dataset.

Signals	Mean	Std
HSTacho Power Torque Acc2	0.999996	0.000044
HSTacho Power Torque HShD	0.999484	0.000164

Note: The signals are sorted in descending order with respect to the mean accuracy.

electrical signals, showing very low values of the mean accuracy. This is particularly striking considering that in the case of the gearbox, there are five classes so an accuracy of around 0.2 is a clue that the model is randomly guessing the classes. Similarly, in the case of the generator rotor, the model shows a very low accuracy of around 0.3.

4.2 | Performance of four-input models

Following the analysis of the single-input models, multiple-input models were tested to investigate whether performance improvement in fault detectability could be obtained from combining different signal information. Given the results from Section 4.1, two four-input models have been considered for the gearbox and generator fault scenarios. In the first case, the model uses the HSTacho, Power, Torque and HShD signals, while in the second case, the model incorporates the Power, Torque, HSTacho and Acc2 signals. HSTacho, Power, Torque and HShD were chosen because of their highest performance in the single-input model analysis. Although the Acc2 model shows a lower performance compared to those signals, it was included in the four-model analysis because vibration signals are usually monitored and available from commercial wind turbine condition monitoring systems.

For each quartet of signals, data has been split 10 times into training, testing and validation datasets using 10-fold cross-validation. Then, for each split, each model was trained 10 times from randomly initialized weights. Overall, the models were trained 100 times on each quartet. Tables 4 and 5 summarize the performance of the two models in terms of the mean and the standard deviation values of accuracy on the validation sets over the 100 training.

The results show that the selected four-input models perform similarly and, in some cases, outperform the single-input models referring to the same signals. In all instances, the four-input models have very low standard deviations showing very robust and reliable learning.

Tables 2 and 3 show that the accuracy for Acc2 is lower than the other signals. However, the four-input model containing Acc2 has an accuracy similar to the model using HSTacho, Power, Torque and HShD. This suggests that the model understands that some signals are more reliable than others and this is exploited in the signal analysis.

Finally, the capability of the proposed multi-high-resolution signal CNN framework to discriminate between electrical and mechanical drive train faults and their severity was investigated. The four-input model using HSTacho, Power, Torque and HShD was trained on the mixed dataset

combining the gearbox and rotor data. In this case, the model has to discern between seven classes: the healthy class (H), resulting from the union of the healthy classes from the gearbox and rotor dataset, two classes of rotor faults (F1 Rotor and F2 Rotor) and four classes of gearbox faults (F1 Gear, F2 Gear, F3 Gear and F4 Gear). As before, data were split 10 times into training, testing and validation datasets using 10-fold cross-validation. Then, on each split, the model was trained 10 times from randomly initialized weights. Overall, the model was trained 100 times. Table 6 shows the model results in the considered scenario in terms of mean and standard deviation values of accuracy on the validation sets over the 100 training.

When analysing the mixed dataset, the combination of the HSTacho, Power, Torque and Acc2 inputs was not been used because, unlike the other sensors, the accelerometers were located in different positions during the two sets of experiments, as described in Section 2.

In Table 7, the confusion matrix containing the average accuracy per class over 100 training is presented. When used for single-component fault detection the four signals achieved already high accuracies (Tables 4 and 5). However, results in Table 6 show that the model still performs very well when tested on the mixed dataset with even higher accuracy and lower standard deviation, implying that increasing the number of input

TABLE 5 Four-input model results for the generator seeded-fault dataset.

Signals	Mean	Std
HSTacho Power Torque HShD	0.999583	0.000303
HSTacho Power Torque Acc2	0.999358	0.000333

Note: The signals are sorted in descending order with respect to the mean accuracy.

TABLE 6 Four-input model results for the mixed dataset.

Signals	Mean	Std
HSTacho Power Torque HShD	0.999794	0.000123

TABLE 7 Confusion matrix for the four-input model on the mixed dataset.

		Actual Classes						
		H	F1 Rotor	F2 Rotor	F1 Gear	F2 Gear	F3 Gear	F4 Gear
Predicted Classes	H	1000.0‰	0.0‰	0.0‰	0.0‰	0.0‰	0.0‰	0.0‰
	F1 Rotor	0.2‰	999.8‰	0.0‰	0.0‰	0.0‰	0.0‰	0.0‰
	F2 Rotor	0.0‰	0.2‰	999.8‰	0.0‰	0.0‰	0.0‰	0.0‰
	F1 Gear	0.0‰	0.0‰	0.0‰	1000.0‰	0.0‰	0.0‰	0.0‰
	F2 Gear	0.0‰	0.0‰	0.0‰	0.0‰	1000.0‰	0.0‰	0.0‰
	F3 Gear	0.0‰	0.0‰	0.0‰	0.0‰	0.0‰	999.9‰	0.1‰
	F4 Gear	0.0‰	0.0‰	0.0‰	0.0‰	0.0‰	0.0‰	1000.0‰

Note: Accuracies are reported per mille.

signals is able to make the model more robust. The confusion matrix in Table 7 shows a clear ability to discriminate very well between the two different types of fault and their severity levels. There are only a few misclassifications occurring in neighbouring fault classes for the case of F1 Rotor, F2 Rotor and F3 Gear.

In terms of computational costs, the trained four-input model has very modest requirements making it possible to embed the necessary hardware inside conventional wind turbine condition monitoring systems. For example, on a Jetson Xavier NX module, the trained four-input model evaluated the entire validation dataset in about 3.3 s. Considering the sampling frequency of 5 kHz, the entire validation dataset contains 7050 windows of 400 points and is equivalent to 564 s of data. Clearly, just a Jetson Xavier NX module is more than enough to run the model in real-time. It could be possible to run several models on the same module to monitor a larger variety of faults or to run models on overlapping windows.

5 | INTERPRETABILITY OF RESULTS

CNNs are achieving high accuracies in many problem domains, but as the model architectures become more complex, so does interpreting the model's decisions. This leads to a black-box issue as the model functionality is hidden, which can result in a lack of trust in the outputs. Furthermore, in real wind turbine operating conditions, identifying which part of the signal the model is diagnosing as faulty can be critical for further study and potential maintenance action. To overcome the black-box issue, this section discusses two interpretability techniques to analyse the inner work of the proposed model. The model is constituted of two parts: the feature extraction branches and the classifier head. The feature extraction branches are formed by a stack of CNN layers. CNN layers only work on the local portion of the signals, but stacking more CNN layers causes the portion of the inputs used to compute the values in the output of the branches to increase, as shown in (A) in Figure 5. In the proposed model, any value in the feature space at the end of each branch is computed from a window of 61 consecutive values from the input. Consequently, no features larger than this window can be extracted from the input. This also forces the network to learn localized relevant features to detect faults leveraging the feature detection capabilities of CNNs. Since the 61-point window is shorter than the revolution of the machine, periodic features in the signal cannot be directly detected. This is very important to keep in mind when trying to interpret the inner work of the model. Commonly, periodicity is what humans tend to exploit when analysing signals. For this reason, the fast Fourier transform (FFT) is a very common approach for the analysis of high-frequency monitoring data to detect fault-specific frequencies in geared transmission systems and electrical machines and has proved a valuable tool for fault detection and diagnosis.[46–48] However, conventional FFT analysis does not handle well non-stationary waveform signals, and this may yield to the identification of inaccurate and unclear component features. The main differences that characterize wind turbine operation compared to other conventional industries are variable speed operation and the stochastic characteristics of the aerodynamic load. This makes it difficult to use traditional frequency domain signal processing techniques, such as FFT, and to develop effective algorithms for early fault detection and diagnosis, due to the non-stationary signals involved.[49] The majority of commercially available wind turbine condition monitoring systems usually require experienced trained engineers able to compare spectra at specific speeds and loads and to make a diagnosis of potential machine problems.[9] The approach proposed in this work is based on analysing localized patterns in the signals to detect faults. Hence, also the interpretation of the model must be based on studying local features and not frequencies or periodic features.

In this section, we focus on two types of interpretability methods: feature summary statistics, showing in statistical form how each feature affects the model predictions, and relevance back-propagation, computing the distribution of relevance values for the inputs. In presenting the results, the approach of *contrastive explanations* is used to argue why a certain classification was made instead of another classification.

5.1 | Multidimensional scaling (MDS)

MDS is a way to visualize the level of similarity of individual samples of a dataset. When MDS is applied to the signal windows, the similarity is computed with respect to the l^2 Euclidean norm. Each point in a MDS plot represents a signal window and the absolute position of the points is irrelevant. What is important is the point relative position which relates to the sample similarity. Two points are close if the corresponding windows have a high level of similarity. On the contrary, two points are far apart if the corresponding windows have a low level of similarity. Figure 6 shows a sample of 2000 windows of the HSTacho signal from the validation set of the gearbox dataset. Different colours represent different classes. As can be seen, often points from different classes are very close to each other. This suggests that windows from different classes are very similar to the naked eye, making the classification of the fault levels from the time domain signals a difficult task for humans. Figure 7 shows the MDS plot of the output values of the first dense layer for the same 2000 samples. The output of that layer is a vector of size 64. In this case, the points belonging to the same class are well-clustered and separated from those belonging to other classes. This is a visual representation of the fact that the model is able to determine correctly the classes of the samples, confirming the high level of accuracy shown in Table 2. This also

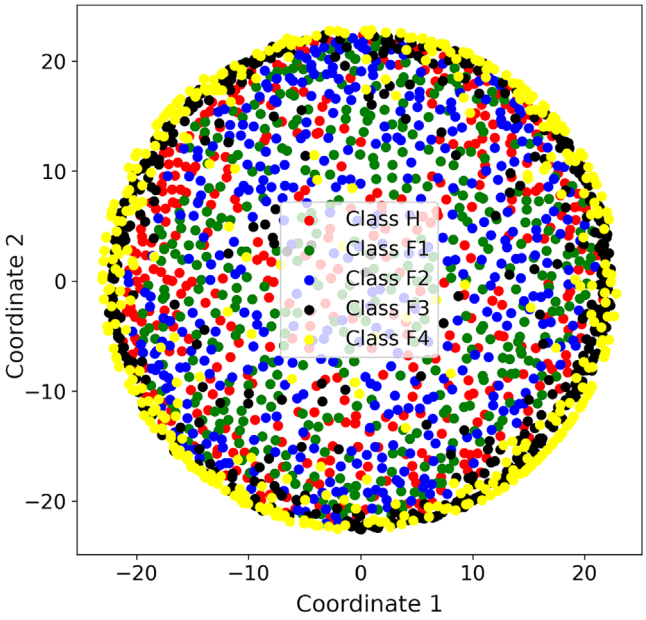

FIGURE 6 Multidimensional scaling plot of the HSTacho signal for a portion of the gearbox seeded-fault dataset validation set.

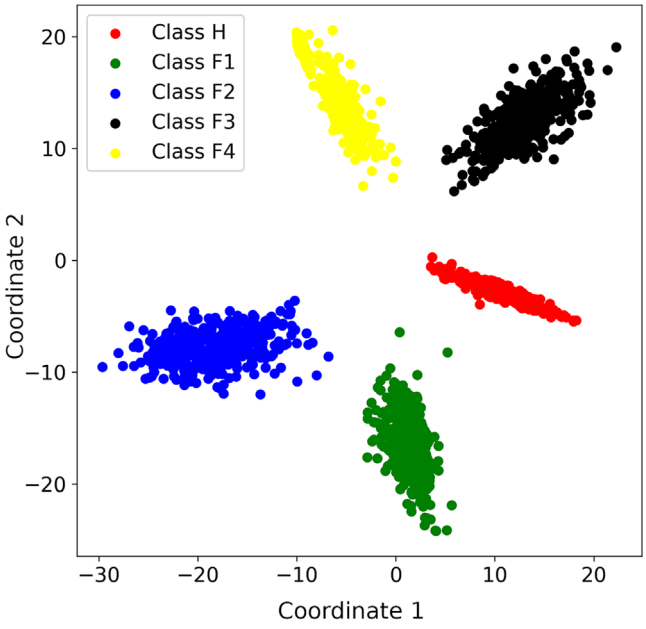

FIGURE 7 Multidimensional scaling plot of the output from the first dense layer in the classifier head, see Figure 5B, for the same windows analysed in Figure 6.

means that in the internal representation of the NN, samples from different classes are very different and easy to recognize, suggesting that the model is able to identify features in the signals of each class. Further, the gaps between clusters are a visual representation of the *confidence* of the model in determining the classes. A wide gap means that there are no ambiguous cases.

This is also confirmed by looking at the mean activation values for the 64 neurons in the first dense layer for different classes. In Figures 8 and 9, the mean activation values for the 64 neurons normalized by the maximum norm are reported for class H and F4. More precisely, the 10 neurons most activated by the class H are 5, 4, 31, 1, 25, 34, 24, 16, 27 and 41, while the 10 neurons most activated by the class F4 are 31, 13, 24, 10, 40, 41, 52, 11, 62 and 20. Clearly, the distribution of activations is different for different classes, and this explains how the last dense layer can positively determine the classes.

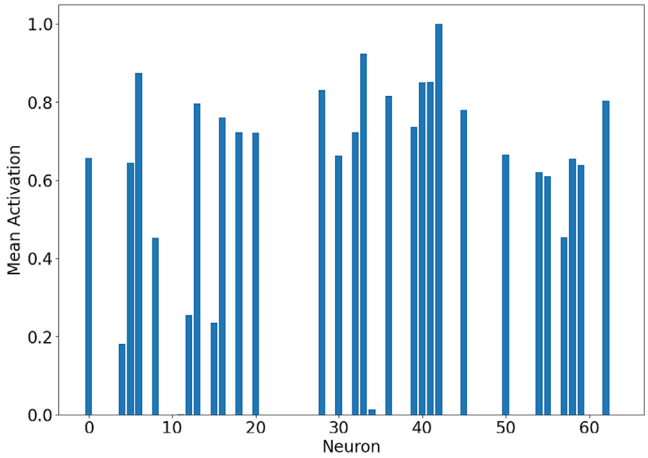

FIGURE 8 Normalized mean activation values of the first dense layer in the classifier head, shown in (B) in Figure 5, for the samples of class H in the validation set of the gearbox seeded-fault dataset for the HSTacho signal.

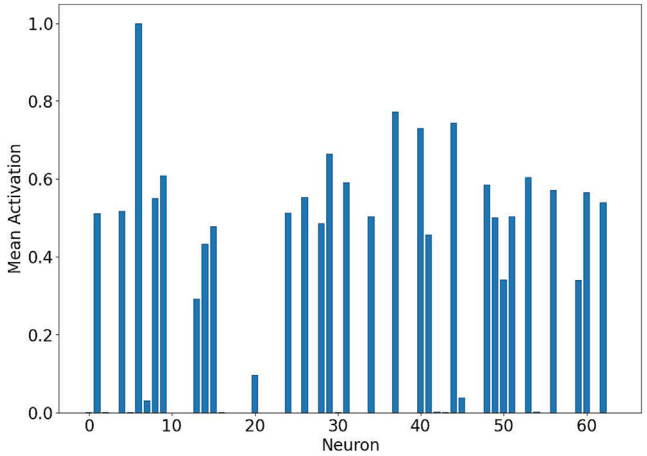

FIGURE 9 Normalized mean activation values for the first dense layer in the classifier head, shown in (B) in Figure 5, for the samples of class F4 in the validation set of the gearbox seeded-fault dataset for the HSTacho signal.

5.2 | Layer-wise relevance propagation (LRP)

Besides predicting the turbine component health state with high accuracy, it is also desirable to understand how and why the classification process works. This can be done by tracing the classification decision back to individual signal samples using the LRP technique.[50] LRP is a back-propagation technique that has been adopted in this work to provide an explanation of the output of the single-input models and to highlight which input features the models used to support the prediction. The result is a map identifying portions of the original signal window contributing to the fault level classification and to what extent. The magnitude of the contribution of each data sample is called relevance.

LRP works by propagating the prediction backwards in the NN back to the inputs. This is accomplished by implementing rules to reverse the action of the different types of layers. The propagation rules are subject to a conservation property that preserves the overall relevance. In other words, summing the relevance values for any input window, the result is always 1. This works like a normalization making it possible to compare relevance distributions from different input windows. Different sets of rules are available in the literature, we used the rules for the LRP-$\alpha_1\beta_0$ which identify positive relevance. This is the appropriate set of rules in this context since the aim is to highlight the portion of the inputs responsible for the classification.

The LRP analysis for the single-input model using the Power signal for the gearbox health state classification is shown in Figure 10. In each case, the top figure shows the signal window with grey areas pinpointing the significant features that contribute to the model decision with an LRP relevance equal to/higher than 0.01, while the bottom figure shows the LRP relevance associated with each sample of the window. Figure 10 shows that the portions of the signal that contribute the most to discriminate between the healthy and faulty gearbox conditions are mainly localized around the signal peaks. These faulty features are, in most cases, characterized by a higher relevance compared to the healthy gearbox case.

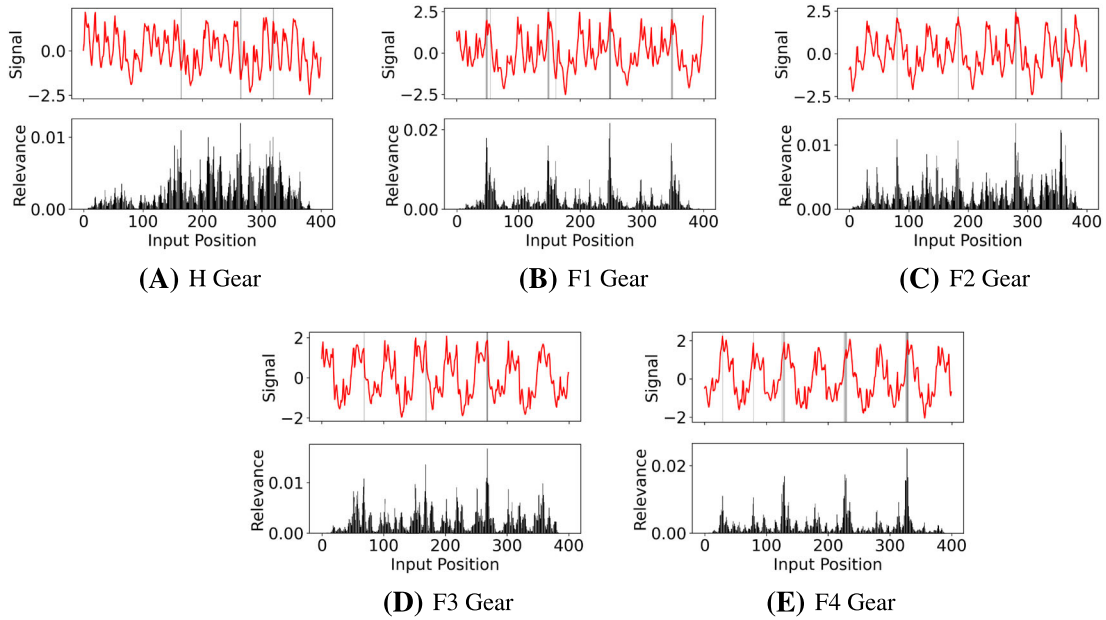

FIGURE 10 CNN model using the Power signal—for each gearbox health state: LRP map (top) and relevance plot (bottom).

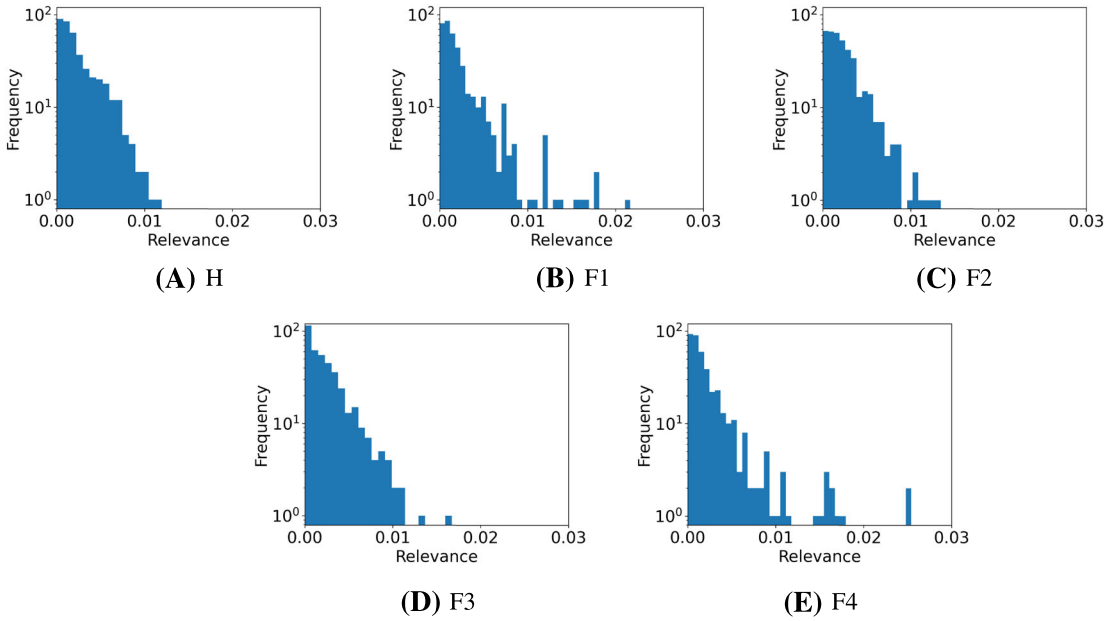

FIGURE 11 Distributions of the relevance for the five cases shown in Figure 10.

The comparison of the relevance plots confirms this behaviour as the healthy gearbox state (Figure 10A) is characterized by a more spread distribution within the signal window with no relevant faulty features pinpointed due to the high noise floor. In comparison when the gearbox is faulty, a more concentrated distribution is observed around the signal peaks with an overall lower noise floor and a more consistent pattern. For the fault conditions, the LRP plot features a clear increase in the number of peaks associated with a relevance equal to/higher than 0.01, with the highest relevance observed in the most severe case (F4 Gear) (Figure 10E). To quantify the consistent patterns observed for the four faulty classes, the distributions of their relevance values are plotted in Figure 11. The *y*-axis is logarithmic to better appreciate the distribution in the tails. The profile of class H decays quickly with a very thin tail. In comparison, all the other classes present longer tails with isolated higher relevance values. The fact that these peaks are mainly isolated from the surrounding relevance values confirms that, compared to the healthy case, the relevance profiles in case of a fault present sharper and more prominent peaks that stand out from the surrounding noise floor.

Similar behaviour has been observed for HSTacho and Torque, where the faulty features characterized by the highest relevance are mainly localized around the local peaks and dips of the signal window. On the other side, for ACC2 and HShD, the LRP analysis has not proved effective in the identification of remarkable portions of the signals contributing most to the output of the model. This might be due to the higher noise associated with these signals which could mask relevant signal faulty features and their human interpretability.

6 | CONCLUSIONS

This paper proposes a real-time automatic and modular CNN framework using high-dimensional raw condition monitoring signals for wind turbine drive train monitoring under complex fault scenarios. From the analysis of the presented NNs in Section 5.2, it is clear that very localized portions of the inputs are important for classification. Such localized regions would be lost in data with lower sampling frequencies, such as SCADA data. The authors believe that the reason why the NN accuracy is low for acceleration signals is that the sampling rate is not high enough to accurately describe the waveform. The proposed architecture has been validated experimentally on a wind turbine drive train test rig with four gearbox fault levels and two rotor electrical fault levels under both constant and variable speed driving conditions, representative of wind turbine operation. The main conclusions from the work are as follows:

- Unlike previous studies focusing on low-frequency SCADA data and the detection of faults in one component only, the proposed multi-input approach is more robust and has broader applications. In particular, it exploits the potential of combining information from a variety of signals and can be tailored to discriminate faults and their severity on multiple reliability-critical drive train components.
- The modular architecture of the CNN framework allows the incorporation of independent multi-sensor data streams characterized by different time resolutions (low-/high-frequency signals) which have traditionally been used independently. This allows users to make inferences that would not be possible based on single-dimensional signal resolution and to improve accuracy and confidence in decision-making capabilities when compared to that provided by individual diagnostic tools.
- The automatic CNN framework uses raw monitoring signals and provides an end-to-end approach which significantly reduces the large degree of manual analysis and expert knowledge currently required for reliable fault detection in traditional monitoring approaches. The procedure described in Section 4 can be applied as a general framework for designing systems for detecting faults. Once a set of faults to be targeted by the AI model has been identified, experts can compile a list of signals that might be relevant in detecting them. Single-input models can then be trained and tested to determine the best signals for the different faults. Finally, the top-performing signals can be combined in a multi-input fault-detection model. In this framework, domain knowledge is only needed to select the appropriate signals. The best signals can be identified using the NN.
- The method proposed in this work is versatile, and its modularity allows easy adaptability to other fault-detection scenarios, such as different failure modes, stages of failure progression and other wind turbine components.
- The architecture of the NN is very light making it possible to embed the necessary hardware inside conventional wind turbine condition monitoring systems at very modest costs. As seen in Section 4.2, the four-input model can run on a Jetson Xavier NX module at such a speed that it has no problems analysing the data in real-time. If new data with a different generating distribution[41] become available, the model can be retrained offline on a desktop machine and the new weights uploaded to the Jetson Xavier NX module, minimizing the downtime of the NN.
- Results show that increasing the number of input signals improves the model's robustness and outperforms single-input architectures showing that the classifier head can learn to have more or less confidence in different signals to achieve the classification. When more than one fault is considered, combining signals that are not traditionally used for condition monitoring of wind turbine gearboxes and generators provides valuable information on the component state with accuracies above 99.9% for all fault classes. It is interesting to observe that the architecture and the number of weights for the networks used for single and multiple fault detection are the same.
- MDS results show the ability of the model to classify correctly the input data streams with high confidence in the network to distinguish different classes. The LRP results show that even if changes in the input signals are almost invisible by visual inspection, the model is able to extract relevant diagnostic features from them and provide reliable detection accuracy. The LRP analysis also shows that the network relies on very localized portions of the signals to determine the fault class raising the possibility of fault detection in the time domain.

The proposed CNN framework was only trained and validated on experimental data from a small-scale test rig, and this is recognized as a limitation of this work. Although the results are promising, further investigation would be required to implement and validate the proposed model on field wind turbine data.

ACKNOWLEDGEMENTS

This work was supported by the EPSRC SUPERGEN Wind Hub (Grant EP/L014106/1).

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/we.2857.

DATA AVAILABILITY STATEMENT

A repository containing one of the trained single-input models and one of the validation sets for the power signal for the gearbox dataset is available at doi:10.15128/r25712m659m. A Python notebook is included in the repository to show how to load and run the model on the provided validations set. Any interested party can contact the authors to request access to the data and the rest of the models.

ORCID

Stefano Giani https://orcid.org/0000-0002-8190-9958

REFERENCES

1. European Commission. A European Green Deal. Striving to be the first climate-neutral continent. 2022. https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en
2. ETIPWind and Wind Europe. Getting fit for 55 and set for 2050 electrifying Europe with wind energy. 2022. https://windeurope.org/intelligence-platform/product/getting-fit-for-55-and-set-for-2050/
3. Jansen M, Staffell I, Kitzing L, Quoilin S, Wiggelinkhuizen E, Müsgens F. Offshore wind competitiveness in mature markets without subsidy. *Nat Energy*. 2020;5:614-622.
4. Wiser R, Rand J, Seel J, et al. Expert elicitation survey predicts 37% to 49% declines in wind energy costs by 2050. *Nat Energy*. 2021;6:555-565.
5. Stehly T, Beiter P, Duffy P. 2019 Cost of Wind Energy Review, National Renewable Energy Laboratory (NREL); 2020. Tech. rep.
6. Rinaldi G, Thies PR, Johanning L. Current status and future trends in the operation and maintenance of offshore wind turbines: a review. *Energies*. 2021;14(9):2484. https://www.mdpi.com/1996-1073/14/9/2484
7. Zappalá D, Tavner PJ. Wind turbine reliability—maintenance strategies. In: Letcher TM, ed. *Comprehensive Renewable Energy*, Vol. 2: Elsevier; 2022: 353-370.
8. Yang W, Tavner PJ, Crabtree CJ, Feng Y, Qiu Y. Wind turbine condition monitoring: technical and commercial challenges. *Wind Energy*. 2014;17(5): 673-693. https://onlinelibrary.wiley.com/doi/abs/10.1002/we.1508
9. Crabtree C, Zappalá D, Tavner PJ. Survey of commercially available condition monitoring systems for wind turbines, Durham University and the SUPERGEN Wind Energy Technologies Consortium; 2014. Tech. rep.
10. Crabtree CJ, Zappalá D, Hogg SI. Wind energy: UK experiences and offshore operational challenges. *Proc Instit Mechan Eng Part A: J Power Energy*. 2015;229(7):727-746. https://doi.org/10.1177/0957650915597560
11. Hossain ML, Abu-Siada A, Muyeen SM. Methods for advanced wind turbine condition monitoring and early diagnosis: a literature review. *Energies*. 2018;11(5):1309. https://www.mdpi.com/1996-1073/11/5/1309
12. Chatterjee J, Dethlefs N. Scientometric review of artificial intelligence for operations & maintenance of wind turbines: the past, present and future. *Renew Sustain Energy Rev*. 2021;144:111051. https://www.sciencedirect.com/science/article/pii/S1364032121003403
13. Helbing G, Ritter M. Deep learning for fault detection in wind turbines. *Renew Sustain Energy Rev*. 2018;98:189-198. https://www.sciencedirect.com/science/article/pii/S1364032118306610
14. Stetco A, Dinmohammadi F, Zhao X, et al. Machine learning methods for wind turbine condition monitoring: a review. *Renew Energy*. 2019;133:620-635. https://www.sciencedirect.com/science/article/pii/S096014811831231X
15. Maldonado-Correa J, Martín-Martínez S, Artigao E, Gómez-Lázaro E. Using SCADA data for wind turbine condition monitoring: a systematic literature review. *Energies*. 2020;13(12):3132. https://www.mdpi.com/1996-1073/13/12/3132
16. Jiang G, He H, Yan J, Xie P. Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox. *IEEE Trans Industr Electron*. 2019; 66(4):3196-3207.
17. Wang Z, Zheng L, Du W, et al. A novel method for intelligent fault diagnosis of bearing based on capsule neural network. *Complexity*. 2019;2019: 6943234.
18. Kong Y, Wang T, Feng Z, Chu F. Discriminative dictionary learning based sparse representation classification for intelligent fault identification of planet bearings in wind turbine. *Renew Energy*. 2020;152:754-769. https://www.sciencedirect.com/science/article/pii/S0960148120301154
19. Xiang L, Wang P, Yang X, Hu A, Su H. Fault detection of wind turbine based on SCADA data analysis using CNN and LSTM with attention mechanism. *Measurement*. 2021;175:109094. https://www.sciencedirect.com/science/article/pii/S026322412100124X
20. Bach-Andersen M, Rømer-Odgaard B, Winther O. Deep learning for automated drivetrain fault detection. *Wind Energy*. 2018;21(1):29-41.
21. Gonzalez E, Stephen B, Infield D, Melero J. Using high-frequency SCADA data for wind turbine performance monitoring: a sensitivity study. *Renew Energy*. 2019;131:841-853. https://www.mdpi.com/2076-3417/11/3/1280
22. Verma A, Zappalá D, Sheng S, Watson S. Wind turbine gearbox fault prognosis using high-frequency SCADA data. *J Phys: Conf Ser*. 2022;2265:32067.
23. Strączkiewicz M, Barszcz T. Application of artificial neural network for damage detection in planetary gearbox of wind turbine. *Shock Vibr*. 2016;2016: 4086324.
24. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM*. 2017;60(6):84-90. http://dl.acm.org/citation.cfm?doid%3D3098997.3065386
25. Muzahid AAM, Wan W, Hou L. A new volumetric CNN for 3D object classification based on joint multiscale feature and subvolume supervised learning approaches. *Comput Intell Neurosci*. 2020;2020:1-17. https://www.hindawi.com/journals/cin/2020/5851465/
26. Zhang K, Zuo W, Chen Y, Meng D, Zhang L. Beyond a gaussian denoiser: residual learning of deep CNN for image denoising. *IEEE Trans Image Process*. 2017;26(7):3142-3155. http://arxiv.org/abs/1608.03981
27. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. arXiv:150504597 [cs]; 2015.

28. Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. arXiv:151100561 [cs]; 2016.
29. Saufi SR, Ahmad ZAB, Leong MS, Lim MH. Gearbox fault diagnosis using a deep learning model with limited data sample. *IEEE Trans Industr Inform*. 2020;16(10):6263-6271.
30. Chen P, Li Y, Wang K, Zuo MJ, Heyns PS, Baggeröhr S. A threshold self-setting condition monitoring scheme for wind turbine generator bearings based on deep convolutional generative adversarial networks. *Measurement*. 2021;167:108234. https://www.sciencedirect.com/science/article/pii/S0263224120307739
31. Xiao C, Liu Z, Zhang T, Zhang X. Deep learning method for fault detection of wind turbine converter. *Appl Sci*. 2021;11(3):1280. https://www.mdpi.com/2076-3417/11/3/1280
32. Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. *Digital Signal Process*. 2018;73:1-15. https://linkinghub.elsevier.com/retrieve/pii/S1051200417302385
33. Shu H, Zhu H. Sensitivity analysis of deep neural networks. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI-19. Association for the Advancement of Artificial Intelligence; 2019:4943-4950.
34. Crabtree C. Condition Monitoring Techniques for Wind Turbines. *Ph.D. Thesis*. Durham University; 2011.
35. Zappalá D. Advanced Algorithms for Automatic Wind Turbine Condition Monitoring. *Ph.D. Thesis*. Durham University; 2015.
36. Chatzopoulos A-P, Leithead W. Reducing tower fatigue loads by a co-ordinated control of the Supergen 2MW exemplar wind turbine. In: Torque 2010: The Science of Making Torque from Wind, Proceedings of 3rd European Academy of Wind Energy (EAWE) Conference; 2010:667-674.
37. Stefani A, Yazidi A, Rossi C, Filippetti F, Casadei D, Capolino G-A. Doubly fed induction machines diagnosis based on signature analysis of rotor modulating signals. *IEEE Trans Industry Appl*. 2008;44(6):1711-1721.
38. Djurovic S, Crabtree CJ, Tavner PJ, Smith AC. Condition monitoring of wind turbine induction generators with rotor electrical asymmetry. *IET Renew Power Gen*. 2012;6(4):207-216.
39. Gritli Y, Rossi C, Casadei D, Filippetti F. Square current space-vector signature analysis for rotor fault detection in wound-rotor induction machine. In: 2016 XXII International Conference on Electrical Machines (ICEM). IEEE; 2016:2894-2898.
40. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. PMLR; 2011:315-323.
41. Goodfellow IJ, Bengio Y, Courville A. *Deep Learning*. MIT Press; 2016. http://www.deeplearningbook.org
42. Jarrett K, Kavukcuoglu K, Ranzato M, LeCun Y. What is the best multi-stage architecture for object recognition? In: 2009 IEEE 12th International Conference on Computer Vision. IEEE; 2009:2146-2153.
43. Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10. Omnipress; 2010:807-814.
44. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Machine Learn Res*. 2014;15(56):1929-1958. http://jmlr.org/papers/v15/srivastava14a.html
45. Kingma D, Ba J. ADAM: a method for stochastic optimization. *Int Conf Learn Represent*. 2014. https://arxiv.org/abs/1412.6980
46. Randall RB. *Vibration Signals from Rotating and Reciprocating Machines*. John Wiley & Sons, Ltd; 2011.
47. Tavner P, Ran L, Penman J, Sedding H. *Condition Monitoring of Rotating Electrical Machines*, Energy Engineering: Institution of Engineering and Technology; 2008. https://digital-library.theiet.org/content/books/po/pbpo056e
48. Tavner P. *Offshore Wind Power, Reliability, Availability and Maintenance*, Energy Engineering. Institution of Engineering and Technology; 2021.
49. Zappalá D, Tavner PJ, Crabtree CJ, Sheng S. Side-band algorithm for automatic wind turbine gearbox fault detection and diagnosis. *IET Renew Power Gen*. 2014;8(4):380-389. https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-rpg.2013.0177
50. Montavon G, Binder A, Lapuschkin S, Samek W, Müller K-R. Layer-wise relevance propagation: an overview. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer International Publishing; 2019:193-209.

How to cite this article: Stone E, Giani S, Zappalá D, Crabtree C. Convolutional neural network framework for wind turbine electromechanical fault detection. *Wind Energy*. 2023;26(10):1082-1097. doi:10.1002/we.2857