ORIGINAL RESEARCH PAPER

Extracting spatially global and local attentive features for rolling bearing fault diagnosis in electrical machines using attention stream networks

Yannis L. Karnavas | **Spyridon Plakias** | **Ioannis D. Chasiotis**

Department of Electrical & Computer Engineering, Electrical Machines Laboratory, Democritus University of Thrace, Xanthi, Greece

Correspondence

Yannis L. Karnavas, Electric Machines Laboratory, Department of Electrical & Computer Engineering, Department of Electrical & Computer Engineering, School of Engineering, Office 0.21, Build. B, University, Campus, Kimmeria, GR-67100 Xanthi, Greece.
Email: karnavas@ee.duth.gr

Abstract

A health diagnosis mechanism of rolling element bearings is necessary since the most frequent faults in rotating electrical machines occur in the bearing parts. Recently, convolutional neural networks (CNNs) have redefined the state-of-the-art accuracy for bearing fault detection and identification, extracting location invariant feature mappings without the need for prior expert knowledge. With the use of convolution operations as the core of the process, CNNs consider the local spatial coherence of the input. However, one major drawback of the convolutional models is the weakness to capture global information about the input vector and to derive knowledge about the statistical properties of the latter. The authors propose a deep learning (DL) model that concatenates the features that are produced from two neural streams. Each consists of an attention mechanism that intends to learn different representations of the input vector, and so finally to produce a feature mapping that contains global and spatial locally information. Simulation results on two famous rolling element bearings fault detection benchmarks show the effectiveness of the method. In particular, the proposed DL model achieves 99.60% in the Case Western Reserve University bearing data set and 99.10% in the Paderborn University bearing data set.

1 | INTRODUCTION

The demand for effective fault detection and identification has been drastically increased due to the complexity and cost of modern industrial systems. Indeed, successful health diagnosis enhances security and reliability, preventing catastrophic unexpected downtimes and reducing the cost and time of repair operations. Rotating components are core elements of mechanical systems and their health condition ensures performance and stability. On the other hand, the most common (over 40%) failures of the rotating electrical machines occur in their rolling element bearings [1].

Therefore, efficient bearing fault diagnosis and identification is crucial in modern industrial applications. Vibration, acoustic, current and temperature measurements have been widely used in fault diagnosis techniques, preventing the losses of bearing element failures [2, 3]. Among them, vibration sensor measurements are extensively employed in industrial applications due to the following: (a) the appearance of bearing defects induces vital changes in the form of the vibration signals and (b) the recent advances in sensor technology have resulted in the efficient measurement and storage of vibration signals with a low signal-to-noise ratio. Moreover, signal processing techniques and machine learning classification approaches exploit the vibration measurements in applications of bearing fault diagnosis.

The signal processing techniques are mainly classified in time-domain and frequency-domain approaches [4, 5]. In the time-domain analysis, statistical indicators such as peak value, root mean square value, crest factor, kurtosis, and others are estimated from raw vibration signals. Alternatively, spectrum analysis investigates the energy in different frequency regions performing Fourier transform. Also, in cases of non-stationary vibration signals, the combination of time and frequency domain has been successfully used, applying techniques as a wavelet transform.

Traditional machine learning approaches require a vast amount of raw data and perform in three phases: initial pre-processing, extraction of hand-crafted features, and training of the conventional supervised algorithm. Artificial neural networks (ANNs) [6], support vector machines (SVMs) [7], Bayesian networks (BNs) [8], neuro-fuzzy inference logic (NFIL) [9–12], hidden Markov models (HMMs) [13], and NN-based multi-agent systems (MASs) [14] have been applied to fault diagnosis applications where each method has specific advantages and disadvantages. Therefore, ANNs are predisposed to over-fitting, while SVMs generalise well even with a limited amount of training data. HMMs are ideal for cases with unobserved states but their training is computationally expensive. Finally, BNs and FNIL-based models are relatively easy to interpret and can handle uncertainties and non-linearities.

Deep learning (DL) models achieve feature hierarchy with the processing of information through multiple non-linear neural layers. In that way, the extracted features acquire representation, abstraction, and discriminative capabilities. The DL models have restated the accuracy results in many research fields such as image recognition, natural language processing, recommendation systems, automatic speech recognition, and others [15]. Also, DL-based models have shown superior performance in the tasks of fault detection and diagnosis in the rotating components of electrical machines, stimulating the interest of the scientific community. Therefore, in the last few years, deep belief networks [16], deep Boltzmann machines [17], deep auto-encoders [18], deep recurrent models [19], and deep generative adversarial networks (GANs) [20] have been employed successfully in the task of bearings fault detection.

The most commonly used models of the DL framework in the fault detection and identification task are the convolutional neural networks (CNNs) [21] and their variants. One of the main advantages of CNNs over conventional neural networks is their ability to learn location-invariant features, exploiting the local spatial coherence of the input mapping. The CNN framework considers the input as a hierarchy of local regions and so each weight filter is moved across the input. In that way, convolutional models employ the weight-sharing property and have fewer learning parameters, enhancing the generalisation ability.

In [22], the raw vibration sensor signals are transformed into two-dimensional images and, in the sequel, a CNN model based on LeNet-5 architecture is used to detect and classify faults. In the same task, a deep CNN with wide first-layer kernels presents robustness to noise and domain adaptation abilities [23]. In another research effort [24], the original raw data signal is transformed into the time-frequency domain. Then, the resulting continuous wavelet transform scalogram acts as input in a CNN model. Also, a CNN model is employed successfully in an experimental environment for fault diagnosis of induction motor [25] and a broad convolutional neural architecture [26] achieves high performance, with the ability to adapt and so to include new abnormal cases. Finally, a comprehensive review of bearing fault detection applications using CNNs is presented in [27].

In the last few years, many advancements of CNNs have been proposed trying to improve their architecture and training. Thus, deep residual and densely connected CNNs present state-of-the-art accuracy in many tasks, exploiting the use of short connections between successive layers [28, 29]. In particular, the practice of skip connections reduces the impact of vanishing gradient and deals with over-fitting problems. Also, the feature mappings concatenation of subsequent layers in the densely connected neural framework reinforces the feature reuse and substantially reduces the number of parameters. In the current research effort, the authors employ the use of dense connections between the feature mappings of successive layers, to make use of their advantages.

Also, another influential DL advancement that recognises the dependencies of the sequential feature mappings is the attention mechanism. This mimics the human visual practice to concentrate and focus on the most relevant regions of visual information for inference. For example, such an attention mechanism is used in the task of natural language processing and especially in neural machine translation systems where it acts as an enhancement of the classical encoder–decoder framework, succeeding to identify long-range dependencies [30]. Furthermore, attention-based DL models have been used for speech recognition, document classification, image caption generation, and others [31, 32].

In the domain of bearing fault diagnosis, the attention mechanism is used in [33] to focus on the most informative regions of the data segments, corresponding to successive vibration signals, and thus to augment the representation strength of the extracted features. Also, an attention mechanism combined with a dense convolutional network is applied successfully for bearing fault identification in [34]. The latter model has fewer learning parameters than the conventional one and generalises better in cases of small training data sets. However, in both research efforts, the attention framework recognises the temporal dependencies of successive feature mappings, which correspond to subsequent segments of the input vibration signal. In contrast, these authors uses the attention mechanism to (i) reconstruct the input vibration signal, using dynamic convolutional layers, focussing on the most informative regions of the latter, and (ii) extract global features that correspond to the long-range interactions of the input. Finally, the extracted local and global attentive features are concatenated to identify the type of bearing fault.

One main drawback of convolution layers is the defect in the occupation of global knowledge about the input mappings since they operate only in a local region. Indeed, convolution operations extract location-invariant hierarchical representative features but do not recognise the global consistency of the input mappings. The work carried out in [35] deals with the above deficiency using a combination of self-attention and constitutional features for visual discrimination and achieving improving on the performance. In the current research effort, the authors propose the use of two independent neural streams, each integrating with spatially local and global information of the input. The feature mappings that are generated

from the neural streams are concatenated and inserted in the final discriminative layers of the network.

In more detail, the first neural stream identifies the spatial consistencies of the input mappings using as core convolution operations. Furthermore, the first block of layers in the first stream acts as a convolution-based local attention mechanism that transforms dynamically the input mapping emphasising its most valuable segments. The most important regions of the input are estimated with the use of a convolutional layer, exploiting its ability to recognise local patterns. On the contrary, the second neural stream examines the global consistency of the input signal, seeking to extract features that combine statistical and energy properties of the latter. Likewise, the first block of the stream provides an attentive mechanism based on a simple feed-forward neural network and concentrates on the most significant parts of the input mapping. In both spatially local and global attention mechanisms, a softmax layer is applied to capture the probabilistic contributions of the attention weights.

A summarisation of the contributions of the current research effort is as follows:

- A convolutional-based attention block of layers is proposed that recognises the spatial consistencies of the input vibration signal, focussing on the most informative region of the latter.
- Global information is identified by another attention mechanism considering the global context of the input mapping with the use of dense layers.
- The extracted features mapping combines spatially local and global information.
- The proposed model identifies successfully the fault classes, achieving state-of-the-art results in two well-known rolling element bearings data set benchmarks. Especially, in one of them, the DL model achieves recognising not only the type of bearing fault but also its severity.
- The robustness of the proposed model is demonstrated also (by simulation means) under the presence of additive noise.

In Section 2, a brief overview of the rolling element bearings and their corresponding failure causes is given. In Section 3, the components of the proposed attention steam network are described. The bearing data set benchmarks, the simulation experiments, the comparison with state-of-the-art literature, and the corresponding results are presented and thoroughly discussed in Section 4. Finally, in Section 5 the conclusions and future work are outlined.

2 | BRIEF OVERVIEW OF THE ROLLING ELEMENT BEARINGS AND THEIR FAULTS

There are many types of rolling element bearings, the most common of them being the ball, needle, and roller ones (Figure 1a–c). For example, ball bearings have spherical rolling elements and are mainly used in low to intermediate load applications, in contrast to roller bearings which use cylindrical rolling elements and are found in heavier load-carrying requirements. More manufacturing options also apply for example regarding the number of rows of the rolling elements so that they can be found as single, double, or multiple-row, the groove geometry (shallow or deep), etc. Without loss of generality, the basic elements of a rolling element bearing are shown in Figure 1d, where apart from the basic diameters which characterise the component, the seals, the inner and outer rings, along with their corresponding races, and the retainer which keeps the rolling elements (balls here) in place, are depicted. Moreover, many important factors should be taken into account during selection of the appropriate bearing such as the available space, the type of load, the rotational speed, the noise, precision and stiffness requirements, the operating environment, etc.

Despite the little attention or maintenance in service which may be required, it is apparent that all of the elements are potentially subject to failure. Figure 2 shows three fault types usually met in electrical machinery, that is inner-race fault, outer-race fault, and ball fault. There are several reasons for which bearings fail. The primary contributors to abnormal bearing signatures are possible imbalance, misalignment, rotary
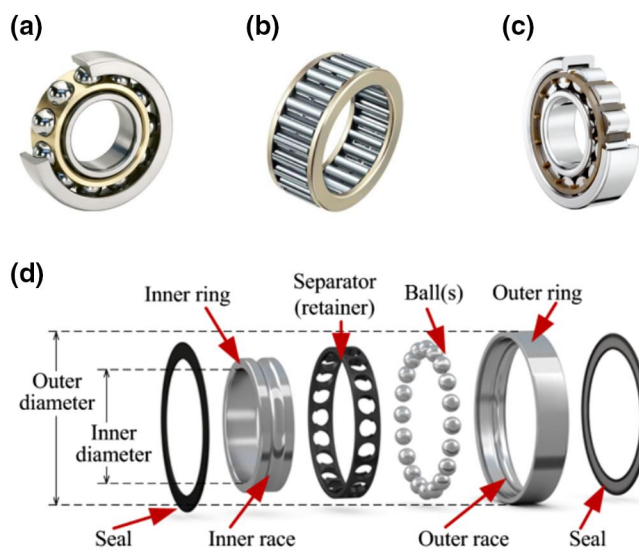

FIGURE 1 Common rolling element bearings types: (a) ball type, (b) needle type, (c) roller type, (d) ball bearing parts (exploded view)
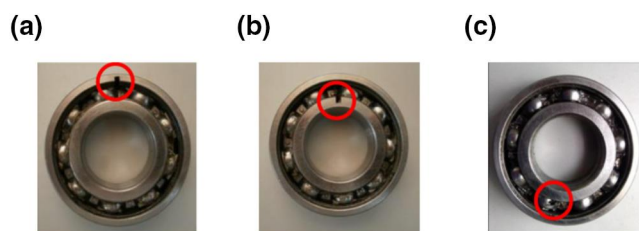

FIGURE 2 Typical bearing faults: (a) outer-race fault, (b) inner-race fault, (c) ball fault

instability, excessive or abnormal loads, and mechanical looseness [36]. The following describe the most common ones.

Removal of small material particles subsequent to a running surface breakage leads to the so-called *spalling* or fatigue failure. This failure can show up on the balls, or the inner or outer ring. Once initiated, it exhibits a progressive behaviour and further operation spreads it out. The further the operation, the more progressive the failure. Of course, the vibrations which accompany it mark an increment.

When the ring's material elastic limit is exceeded as a result of loading, we have the situation of *brinelling* (named after the Brinell scale of hardness). The marks in this case are shown in the raceways as permanent indentations. Actually, brinelling may be caused by any severe impact or static overload. Similarly to fatigue failure, vibration and resultant noise are increased.

Premature fatigue may be caused by excessive loads. Improper fits (tight or loose), improper preloading conditions, and brinelling can also lead to early fatigue failure. A solution to this case is a load reduction or the use of a greater capacity bearing.

Another cause relates to overheating. Discolouration (gold to blue) of the cages, balls, and rings constitutes a typical symptom. The ring and ball materials can also be annealed if a temperature of 400 °F or above is reached. The capacity of the bearing is reduced due to the loss in hardness and early failure appears. Element deformation may be noticed in extreme cases. The bearing lubricant will also be destroyed or at least degraded with the temperature rise.

Lubricant failure and overheating are strongly connected. A blue/brown discolouration in the balls and their tracks denotes bad lubricant, resulting in consequent excessive wear of balls and rings, which in turn leads to overheating and in turn catastrophic failures. The bearing's health is largely dependent on a very thin film of lubricant (millionths of an inch) with adequate viscosity, which should be continuously present between the races/balls and between the cage/rings/balls [37].

Contamination and corrosion are also among the leading causes of bearing failure. The usual contaminants such as sand, water, and dirt, but also corrosives and chemicals also cause failures. The former dilute the thin oil film and reduce the lubricant viscosity, while the latter corrode the bearing surfaces and thus create many abrasive particles. The denting of the bearing balls and raceways is then inevitable, resulting in high vibrations and wear. Discolouration of red/brown areas on the raceways and balls is a usual symptom of corrosion, while in some cases it can initiate fatigue failures.

Misalignment and improper fitting are also usual bearing failure causes. It is worth noting that although some, but not all, bearings tolerate minor misalignment between shaft and bearing housing, in most of them a 1/1000th of an inch per inch misalignment leads to abnormal temperature rise and wear of the ball retainer. Excessive vibration then is introduced.

In cases where loose fits are present, and the relative motion between mating rotating parts (shaft/inner ring) may be slight but continuous, fretting occurs. Fretting is the generation of fine metal particles which oxidise forming an abrasive material (leaving a distinctive brown colour), which, in turn, will aggravate the looseness. On the contrary, if a tight fit is indicated (e.g. the rotating interference fits exceed the radial clearance), a heavily loaded balls situation is encountered. The results include high load torque, temperature rise, and rapid wear and fatigue in continuous operation [38].

It is thus evident from the above that a reliable and very precise fault diagnosis method is of critical importance to avoid electrical machine damages. The current research is towards this direction.

3 | DESCRIPTION OF THE PROPOSED DL MODEL

The architecture of the proposed DL model is illustrated in Figure 3. The latter takes as input the raw vibration signal, provided by sensors, and outputs the class that corresponds to the normal state or the type of electrical motor bearing fault. With respect to Figure 3, a brief description of the layers of the DL model follows.

The convolutional layer tries to stimulate the structure of the biological visual cortex, where each cortical neuron operates only in a restricted region of the visual space (receptive field). The receptive fields of neighbouring neurons overlap to cover and so to represent the entire visual field. The artificial visual process is stimulated by the convolving of the input signal with a set of learnable kernel weights. The responding feature representations correspond to the kernel weights of the convolutional layer. The mathematical operation of the convolutional layer is given by the following equation

$$s_o = \sum_P K^l * s_i + b^l, \tag{1}$$

where the operator $(*)$ is the dot product, K^l denotes the l-th kernel filter of the convolutional layer which is shared across the local region P, b^l is the bias of the l-th filter, and s_i, s_o are the input and output of the convolutional layer, respectively. With the passing of the same kernel filter over the input mapping, the property of weight sharing and the extraction of location invariants features are achieved. The setting hyperparameters of the convolutional layer are the quantity of the kernel filters, their dimensions, the stride, and the padding. The reader may refer to [21], for a recent broad survey which describes and analyses in detail the improvements in CNNs on many and different aspects.

The batch normalisation (BN) layer enhances the learning process of DL models, eliminating the effect of covariate shift [39]. The latter is the variation in the distribution of the activations of the internal layers during training. To restrict the phenomenon of covariate shift, the activations of each layer are normalised to zero mean and unit variance. In the sequel, to maintain the representation capability of each layer, a transformation of the normalised mapping is applied using the learnable parameters $\gamma^{(k)}$ and $\beta^{(k)}$. Through the above stable conversion of the distribution of activations, larger values of
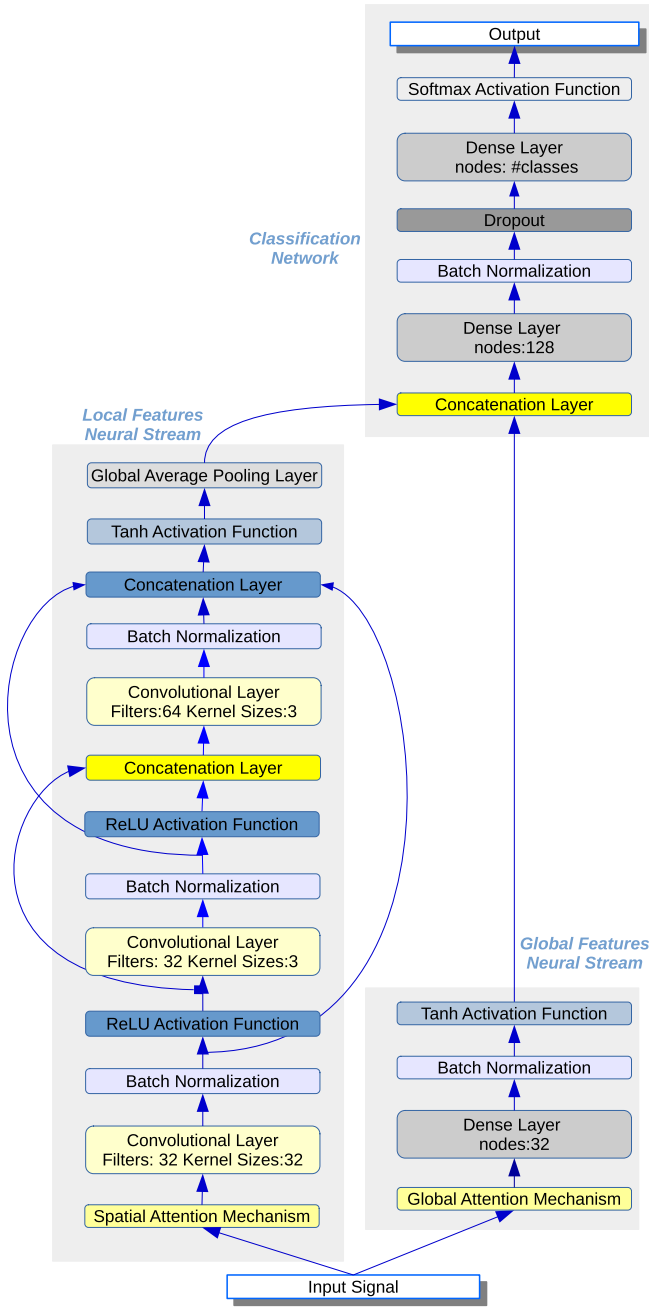
FIGURE 3 Architecture of the proposed deep learning model

Activation layers apply the non-linear transformation of the input mappings, improving the representation and discriminative capability of the DL model. The most typical activation function is the rectified linear unit $[ReLU(z) = \max(0, z)]$ since it addresses the vanishing gradient problem and is computationally efficient. In the current research effort, right before the concatenation of the features from both neural streams, we apply the hyperbolic tangent activation function $(tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}})$ to permit the dying ReLU problem and get representation balance. Also, in the last layer of the multi-class classifier, the authors use the *softmax* activation function which firstly normalises the activation vector into a probability distribution using exponential terms and thereinafter returns the position of the maximum value of the normalised vector.

The output mappings of convolutional layers provide information about the exact location of features in the input. This results in the sensitivity of the feature mappings in minor alterations of the input, like shifting, cropping, and rotation. To deal with the described problem, the authors employ pooling layers that apply down-sampling in the feature mappings with maximum or mean operations. In addition to the advantage of invariance to local translation, the down-sampling of the activations reduces the learning parameters of the network, preventing over-fitting and reducing computational cost. A global mean pooling layer is applied in the proposed model to extract robust locally spatial features provided by the convolutional neural stream.

Also, we use dense layers that implement fully connections between the input and output neurons of successive layers. The latter executes a multiplication between a learnable weight matrix and the vector mapping achieving the extraction of global features. Furthermore, the concatenation layer deploys the connection of the feature mappings $s_1 = [s_1^{(1)} s_1^{(2)} .. \ s_1^{(p)}]$ and $s_2 = [s_2^{(1)} s_2^{(2)} .. s_2^{(q)}]$ as $s_o = Concat(s_1, s_2) = [s_1^{(1)} s_1^{(2)} .. s_1^{(p)} s_2^{(1)} s_2^{(2)} .. s_2^{(q)}]$.

Finally, the authors practice the *dropout* technique to prevent the over-fitting phenomenon and enhance the regularisation capability of the DL model [40]. To achieve the latter, at each training iteration a set of randomly selected neurons is dropped out. Hence, a set of different neural models is trained to learn the pattern object, forming an ensemble of models, and producing better classification accuracy. Dropout effectiveness is limited in combination with convolutional layers and so the authors employ dropout at the final classification network where they operate dense layers.

It can be observed from Figure 3 that the DL model can be separated into three individual neural parts. Two are streams that generate features, which are concatenated and inserted into the final classification network. The architecture of the latter is simple and is commonly used in any DL classification model since it consists of two dense layers that are connected with a batch normalisation layer and a dropout one. Furthermore, Figure 3 provides information about the number of filters and the sizes of kernels in the convolutional layers and the number of nodes in the dense ones.

learning rate parameter can be applied improving the speed of training. The d-dimensional activation of the inner layer as $s_i = (s_i^{(1)}, s_i^{(2)}, .., s_i^{(d)})$ is now considered. Then, the BN transformation $s_o^{(k)}$ of the input mapping $s_i^{(k)}$ is described by the following mathematical formula

$$\widehat{s_i}^{(k)} = \frac{s_i^{(k)} - \mathbb{E}(s_i^{(k)})}{\sqrt{Var(s_i^{(k)})}} \tag{2}$$

$$s_o^{(k)} = \gamma^{(k)} \widehat{s_i}^{(k)} + \beta^{(k)}.$$

3.1 | Extraction of locally spatial features

The neural stream that extracts locally spatial features (left in Figure 3) consists of the attention mechanism followed by three convolutional blocks. Each convolutional block is made of the convolutional layer, the BN layer, and the activation layer. The produced feature mappings from successive convolutional blocks are concatenated to take advantage of the densely connected framework [29]. With the adoption of concatenation connections, the authors achieved extraction of features that correspond to different levels of hierarchical representation and reinforce the propagation of information through the network.

The locally spatial attention mechanism (Figure 4) reconstructs the input signal focussing on the most important and informative regions of the latter. To accomplish the transformation of the input, a convolutional layer is employed in the first place to recognise the feature patterns with discriminative strength. Therefore, each output vector of the convolutional layer c_i corresponds to the identification of an individual kernel filter f_i in the input signal s. In particular, $c_i^{(j)}$ indicates the presence of filter f_i in the neighbourhood of input s around the position j. Then, a global average pooling layer is applied to detect the recognition of each kernel filter in the input signal s as

$$a_i = \sum_j c_i^{(j)} \Big/ P, \tag{3}$$

where P is the length of the input signal and the convolutional feature mappings too, since the authors use zero padding to preserve the original input size. It is worth noticing that the values a_i intuitively indicate the appearance of kernel filter f_i in the input signal and so the discriminative importance of feature mapping c_i.

Afterwards, the softmax function is used to normalise the values a_i into a probability distribution using the following equation

$$e_i = \frac{e^{a_i}}{\sum_j e^{a_j}}. \tag{4}$$

Each convolutional feature vector c_i is multiplied with its normalized significance term ei and then an element-wise addition follows.

$$\alpha = \sum_{i=1}^{Q} e_i \times c_i. \tag{5}$$

In Equation (5), Q denotes the quantity of the kernel filters in the convolutional layer. Through the process, the attentive vector holds vital information about the regions of the input signal that have been activated through the convolutional process. Finally, the input signal s is dynamically adapted by its element-wise multiplication with the attentive vector α.

$$s_o = \alpha * s \tag{6}$$

3.2 | Extraction of global features

The assignment of the global attention mechanism (Figure 5) is to extract features considering the global context of the input signal. The convolutional operation acts locally and thus has no power to recognise the long-range interactions and the statistical properties of the input. To achieve the extraction of global information, the global attention mechanism is based on dense layers, performing linear transformations of the input vector by its multiplication with weight matrices.

Therefore, two dense neural blocks are employed to produce the global attentive vector. Each dense block consists of a dense layer, a batch normalisation layer, and an activation function. Firstly, as activation function the ReLU operation is used and subsequently the softmax one is used to normalise the attentive features into a probability distribution. The input signal is adapted by its point-wise multiplication with the global
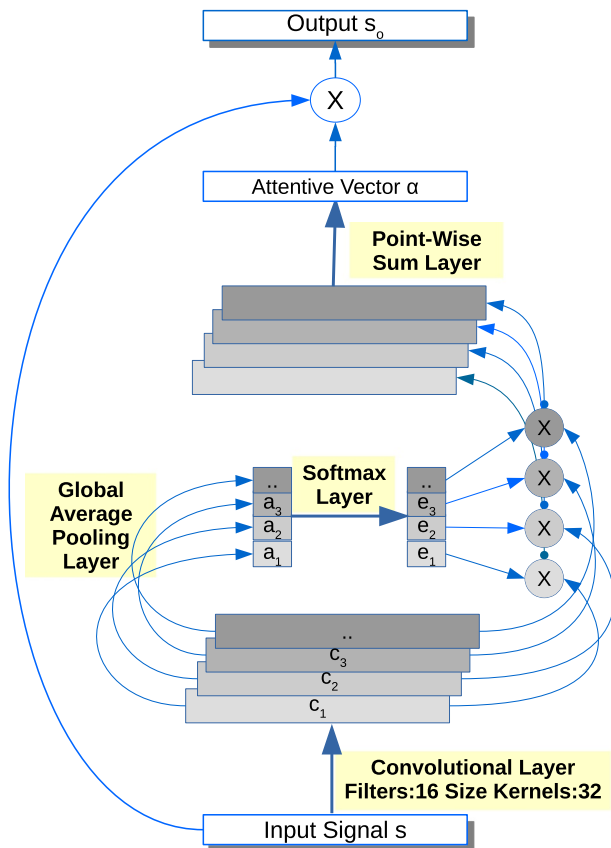

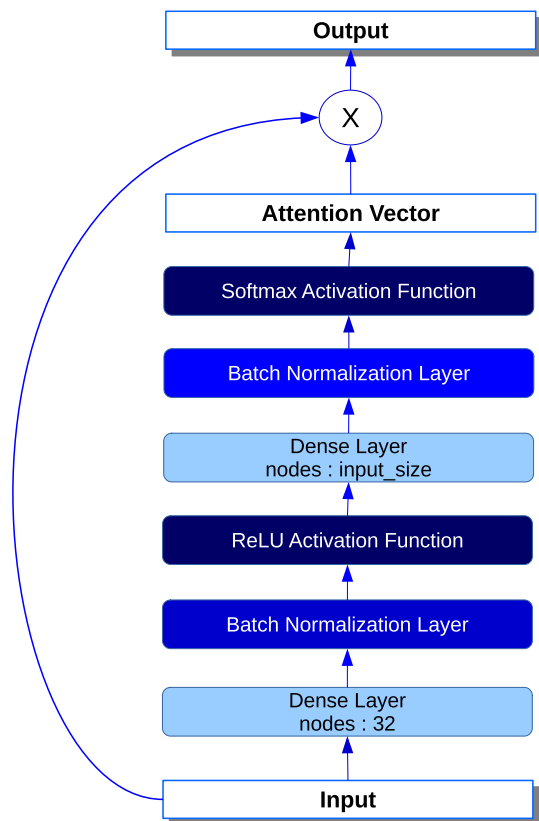**FIGURE 4** Architecture of spatial attention mechanism

FIGURE 5 Architecture of global attention mechanism

attentive features and finally, the neural stream applies another dense block to extract the final global feature mappings.

4 | SIMULATION EXPERIMENTS, RESULTS, AND STATE-OF-THE-ART COMPARISON ANALYSIS

To confirm the feasibility of the proposed DL model in the detection and identification of bearing faults, the authors apply simulation experiments with the employment of two well-known benchmark data sets: the Case Western Reserve data set and the Paderborn University fault detection data set. The overall simulation code is written in Python ver. 2.7 programming language and is based on the Keras (a highly modular neural networks library) library, running "on top" of Theano software library.

The objective function during the optimisation process is the categorical cross-entropy loss. Also, the stochastic gradient descent, with Nesterov momentum enabled, has been selected as the optimiser. The learning rate during the optimisation process, the decay, and the momentum parameters of Nesterov momentum were set to 10^{-3}, 0.0, and 0.9, respectively. The training was completed within 200 epochs and the batch size was 200.

Furthermore, a convolutional neural network (CNN) has been developed as a comparison tool for the specific bearing fault diagnosis task. The latter is made of three convolutional
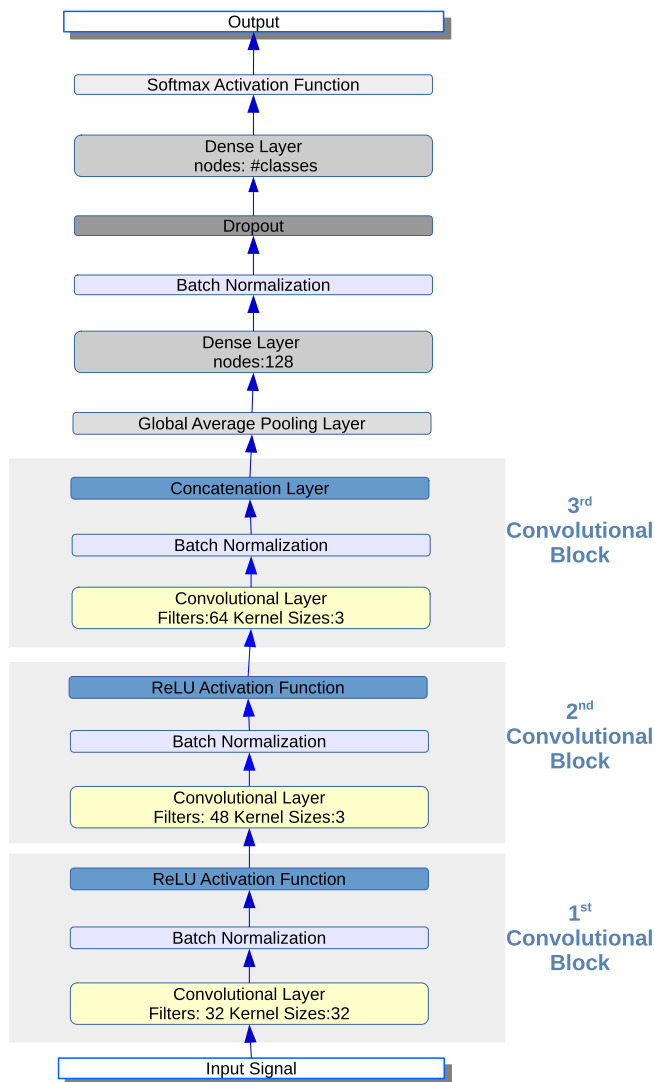

FIGURE 6 Architecture of the developed convolutional neural network (for comparison purposes)

blocks, each consisting of a convolutional layer, a batch normalisation layer, and a ReLU activation layer. The final classification part of the CNN is identical with the proposed model. Figure 6 displays the architecture of the CNN model with details about the hyper-parameter of each layer. Also, to bypass the effect of the stochastic initialisation of the neural weights, the authors run each experimental simulation 10 times and present the mean and standard deviation values of fault detection accuracies for both models and each simulation case.

4.1 | Simulation experiment 1: CWRU bearing data set benchmark

The Case Western Reserve University (CWRU) [41] fault detection and identification data set supplies vibration signal samples that correspond to normal and faulty bearings of a motor shaft. The experiments are conducted in a test rig that is

made of a 2 Hp Reliance Electric motor, a torque encoder, and a dynamometer (Figure 7). The single-point failures of the bearings were caused by electric discharge machining (EDM) processes with varied fault diameters. The vibration signals are collected from accelerometers attached to the housings with magnetic bases of the drive-end, the fan-end, and the motor-supporting base plate. The sampling rate during the acquisition of the digital vibration samples is 12 kHz. Also, the electric motor operated under four loading conditions, from 0 to 3 Hp and so the motor speed varied between 1792 and 1730 rpm.

In the current research effort, the drive-end bearing signals are used in the detection of five types of failures: ball, inner race, and three classes of outer race faults. The position of the outer race fault concerning the load zone of the bearing affects significantly the vibration signal. In the CWRU data set benchmark, data samples were acquired with the outer race faults placed at 3 o'clock (directly in the load zone), at 6 o'clock (orthogonal to the load zone), and at 12 o'clock. Also, the vibration signals correspond to different diameters of faults and, in more detail, there are faults of 0.007, 0.014, 0.021, and 0.028 inches.

Regarding the types of faults and their severity, a classification problem with 16 classes is built (Table 1). Furthermore, taking into account the sampling rate of 12 kHz and the motor speed of 1797 rpm, the number of sample points per revolution is estimated at approximately 400. Therefore, the input dimension of each vibration signal is 400 and during the creation of the training and the testing data set, 20 sampling points are used as an overlap between two successive samples. It is worth noting that the training and the testing data set contain samples that correspond to all the loading conditions of the electric motor. Finally, the data sets of Table 2 are built which are composed of a diverse amount of samples, to explore the efficiency of both neural models under comparison.

The simulation results of both examined DL models are shown in Table 3. The superiority of the proposed model is noticed, adopting the concatenation of the attentive neural streams, over CNN especially in cases where the amount of training samples is limited. Indeed, the proposed model presents 99.08% accuracy using training data set A with 100 samples per class. On the contrary, CNN performs 98.66% with the use of the same training data set. Finally, it is observed that an accuracy of 99.60% is achieved using 400 training samples from each class and s not only the type of fault but also its severity can be successfully identified.

Also, the confusion matrix of the performance of the proposed model trained with data set C is demonstrated in Table 4. It is observed that the normal condition is classified with success (100%) and so the proposed model detects the appearance of a fault with absolute accuracy. Also, it is considered that the state-of-the-art refers to identification levels of 98.5% and above, then it can be noticed that the system has 'difficulties' in identifying classes with ID 14 and 10. In particular, the recognition of classes 14 (outer@12.00 fault

FIGURE 7 Experimental test stand of Case Western Reserve University benchmark data set

TABLE 1 Description of classes of CWRU benchmark data set

Class ID	Condition	Fault diameter(inches)
1	normal	–
2	Ball	0.007
3	inner	0.007
4	outer@6.00	0.007
5	outer@3.00	0.007
6	outer@12.00	0.007
7	Ball	0.014
8	inner	0.014
9	outer@6.00	0.014
10	Ball	0.021
11	inner	0.021
12	outer@6.00	0.021
13	outer@3.00	0.021
14	outer@12.00	0.021
15	Ball	0.028
16	inner	0.028

Abbreviation: CWRU, Case Western Reserve University.

TABLE 2 Training and testing data sets of CWRU benchmark data set

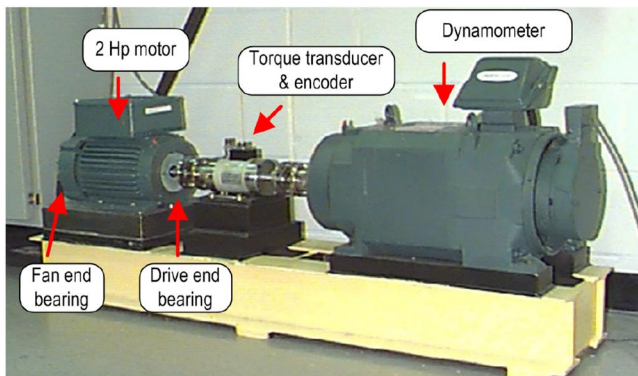| Data sets	# Samples per class	# Total samples
Training data Set A	100	1600
Training data Set B	200	3200
Training data Set C	400	6400
Testing data Set	400	6400

Abbreviation: CWRU, Case Western Reserve University.

TABLE 3 Performance (%) of compared neural models in the CWRU benchmark data set

# Training Samples	Convolutional NN	Attention Stream Net (Proposed)
1600	98.66 ± 0.4	99.08 ± 0.3
3200	99.12 ± 0.37	99.52 ± 0.28
6400	99.48 ± 0.13	99.60 ± 0.14

Abbreviation: CWRU, Case Western Reserve University.

TABLE 4 Confusion matrix (%) of Attention Stream Net trained with data set C in the CWRU benchmark data set

		Predicted Class ID														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Actual Class ID	1	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	0	0	0	99.9	0	0.025	0	0	0	0	0	0.075	0	0	0	0
	5	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	99.9	0	0.1	0	0	0	0	0	0	0	0
	7	0	0	0	0	0	0	99.825	0	0	0	0.175	0	0	0	0	0
	8	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
	9	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0
	10	0	1.7	0	0	0	0	0	0	0	98.3	0	0	0	0	0	0
	11	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0
	12	0	0	0	0	0	0.025	0	0	0	0	0	99.125	0	0.85	0	0
	13	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
	14	0	0	0	0	0	0.025	0	3.4	0	0	0	0	0	96.575	0	0
	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0.025	99.975	0
	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

with diameter 0.021 inches) and 10 (ball fault with diameter 0.021 inches) is not as successful as the others, since the proposed system classifies correctly with 96.575% and 98.3%, respectively. However, the system for most classes has an accuracy of approximately 100%.

Finally, Figure 8 shows a visualisation in 2D-space via the t-Distributed Stochastic Neighbor Embedding (t-SNE) method of the extracted features of the proposed DL model, produced from the concatenation of the feature mappings of both local and global attention neural streams. The t-SNE method is a stochastic non-linear dimensionality reduction technique that projects each high-dimensional sample point into a low-dimension space focussing on data visualisation [42]. The main advantage of the t-SNE algorithm is preservation of the local structure of the data, estimating conditional probabilities in higher and lower dimensions to match most similar neighbours. It is observed from Figure 8 that the produced clusters correspond to each class. Therefore, the extracted features hold representation and discrimination capability and the task of the classification mechanism of the proposed DL model becomes easier.

4.2 | Simulation experiment 2: Paderborn University bearing data set

The Paderborn University bearing benchmark data set [43] provides high-resolution vibration data, obtained from experiments conducted on six healthy and 26 damaged bearing sets. The test rig apparatus (Figure 9) consists of a permanent magnet synchronous motor, a torque measurement shaft, the test modules, and a synchronous servo motor acting as the load motor. With the use of a rolling element bearing module, changing testing bearings are applied under constant radial load and so the vibration signal of the inner housings is acquired and stored.

Herein, the authors used vibration signals from five healthy bearings and 10 faulty bearings obtained from accelerated life testing. Half of the faulty bearings correspond to a failure in the inner race, while the other half correspond to a failure in the outer race. Therefore, a fault detection problem with three classes was built. Also, the authors simulated the fault detection task by trying to create conditions that prevail in real-life situations. Therefore, they used healthy bearings of a different

strain with varying total sum of operating hours under different operating conditions and faulty bearings with varying characteristics of damage, different damage combinations (single, repetitive, multiple), and a different arrangement of damages (regular, random, none). In particular, the following bearing codes of the Paderborn data set benchmark are used: for the normal class (K001, K002, K003, K004, K005), for the outer race (KA04, KA15, KA16, KA22, KA30), and for the inner race (KI04, KI14, KI16, KI18, KI21). Also, the operation settings of a rotational speed of 1500 rpm, load torque of 0.7 Nm, and a radial force of 1000 N are employed in the current research effort.

Taking into account the sampling rate of 25 kHz, the authors estimate the number of sample points per revolution as 2560. Furthermore, to obtain sufficient information in the input of the model for one revolution, they set the input dimension to 2560. Similarly to the previous simulation case, they use training data sets with a different number of samples
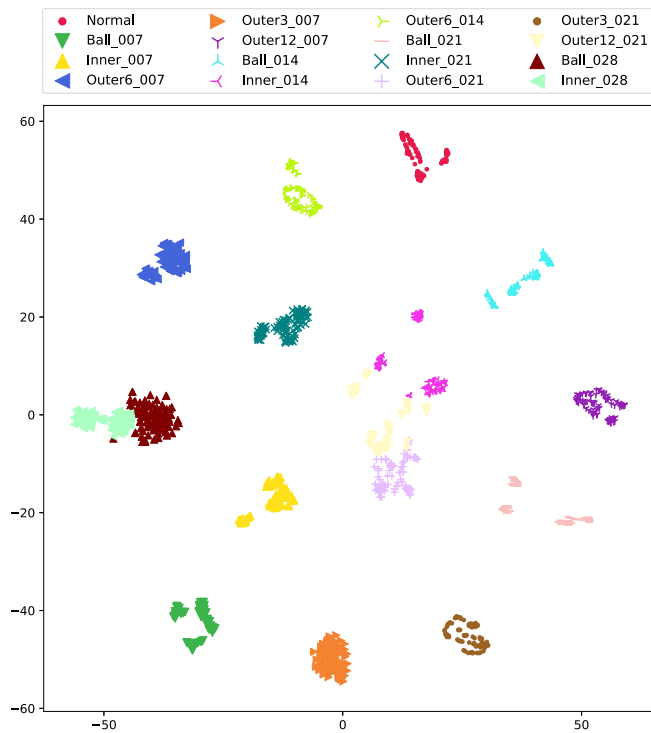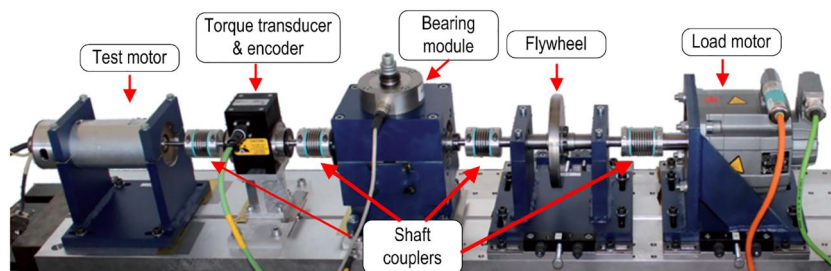
(Table 5). The simulation results are shown in Table 6, where they perceive the superiority of the proposed DL model over the CNN model. Indeed, the model that combines the attentive streams performs better regardless of the training data and achieves 99.10% classification accuracy trained with the 500 samples per class data set. Also, although it consists of more learnable unknown parameters, it generalises better and presents robustness to over-fitting. Furthermore, from the confusion matrix of Table 7 it is noticed that the normal class is recognised with absolute success.

The visualisation of the feature mappings using the t-SNE method is illustrated in Figure 10. It is noticed that the number of forming clusters is greater than the number of classes. The phenomenon of the greater number of clusters is explained by the fact that each class contains samples that correspond to different modes of operation. Indeed, observing Figure 10 15 clusters are counted, equal to the number of bearing codes of the data set. Also, it is noticed that each class is separated from the others since each class is made of a group of separated clusters. It is worth mentioning that distances between clusters and cluster sizes are meaningless in the t-SNE plots [44]. The lack of information about the global distances prohibits the extraction of conclusions from the position of the clusters. On the other hand, someone can obtain intuition knowledge based on the membership of the clusters. In any case, the t-SNE algorithm is a visualisation method for high-dimensional data and the use of clustering methods based on distance or density after its application is not appropriate.

4.3 | Effect of additive noise

In a real industrial environment, the vibration signals captured from accelerometers suffer from the presence of additive

FIGURE 8 Visualisation in 2D-space via t-SNE of the extracted features for the Case Western Reserve University data set

T A B L E 5 Training and testing data sets of the Paderborn benchmark data set

Data Sets	# Samples per class	# Total samples
Training Data Set A	300	900
Training Data Set B	400	1200
Training Data Set C	500	1500
Testing Data Set	500	1500

FIGURE 9 Experimental test stand of the Paderborn University bearing data set

TABLE 6 Performance (%) of compared neural models in the Paderborn benchmark data set

# Training samples	Convolutional NN	Attention stream net (Proposed)
900	97.95 ± 1.01	98.44 ± 0.41
1200	98.06 ± 0.37	98.95 ± 0.33
1500	98.39 ± 0.36	99.10 ± 0.27

noise. To explore the robustness of the compared DL models in the presence of noise, firstly the authors train the networks with the original vibration signals and then check their performance with noisy signals of varying signal-to-noise ratio (SNR) containing additive Gaussian noise. The training data set A of the CWRU bearing data set is chosen with 1600 samples (100 samples per class). Also, the balanced testing data set with 6400 total samples of vibration signals is modified to include additive noise. By choosing a limited number of training samples, the authors examine the robustness of the models in the most challenging case.

Figure 8 contains the simulation results under the appearance of noise for both models. It is perceived that for all examined cases of varying SNR, the proposed DL model overpowers the convolutional one. More particularly, the difference in fault identification accuracy is more noted for cases where the level of noise is higher (SNR 10 and 15), confirming the robustness of the proposed DL model (Table 8).

4.4 | Comparison with state-of-the-art methods using the CWRU benchmark data set

Finally, the proposed Attention Stream Net model is compared against other state-of-the-art models in Table 9 in the CWRU bearing data set. The table presents a description of the methods, the corresponding references, the accuracies during the identification task, the number of the classes, and the percentages of the used sample points concerning the whole amount of sampling points in the CWRU data set. Therefore, [45] generates synthetic frequency spectrums using the GAN framework and applies a stacking denoising auto-encoder to classify six faulty situations achieving 99.20% accuracy. In [46], a convolutional-based DL model detects six types of bearing faults with a performance of 99.63%. Wavelet kernel local fisher discriminant analysis is employed first and, in the sequel, support vector machines solve the fault identification task in [47]. Furthermore, in [34] the temporal coherence of the bearing signals is recognised, utilising a simple attention mechanism that is based on a feed-forward neural model. The inception framework is used in [48], building a multi-scale CNN for the identification of 10 classes. Also, the dependencies of successive bearing signals are recognised in [49], practising LSTM blocks over CNN ones.

In Table 9, it can be noticed that only [34] and the model proposed herein identify 16 classes trying to recognise not only the type of bearing faults but also their severities and so solving a more difficult task. Also, the authors observed that the proposed Attention Neural Stream and the models of [34, 47]

TABLE 7 Confusion matrix (%) of attention stream net trained with data set C in the Paderborn benchmark data set

		Predicted Class ID		
		1	**2**	**3**
Actual class ID	1	100	0	0
	2	0	98.54	1.46
	3	0	1.24	98.76
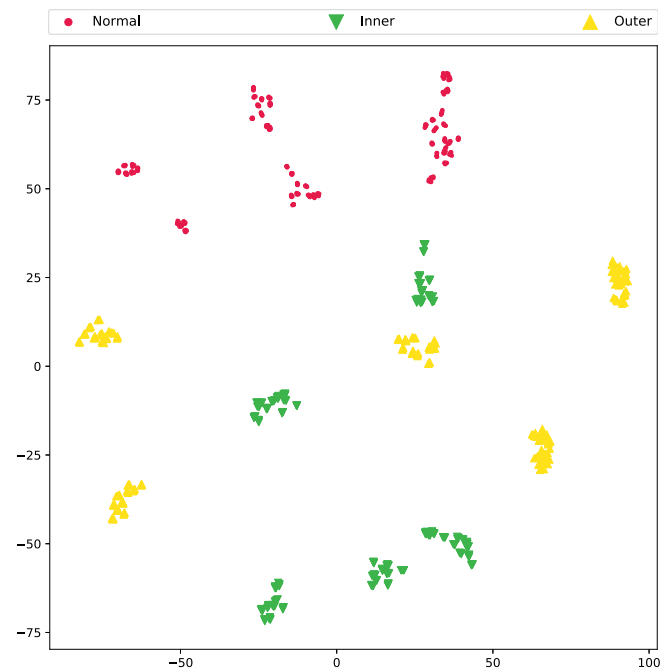
FIGURE 10 Visualisation in 2D-space via t-SNE of the extracted features for the Paderborn data set

TABLE 8 Performance (%) of compared neural models using training data set A of the CWRU bearing benchmark under the presence of additive noise

# Noise Level (SNR)	Convolutional NN	Attention Stream Net
10 dB	90.98 ± 3.69	92.74 ± 2.99
15 dB	92.06 ± 3.38	94.44 ± 2.84
20 dB	98.65 ± 0.42	98.96 ± 0.26

Abbreviation: CWRU, Case Western Reserve University.

TABLE 9 Comparison with state-of-the-art methods using the CWRU benchmark data set

Method	# Classes	Percentage of training samples	Accuracy (%)	Reference
GAN + SDAE	6	78	99.20	[45]
Lifting Net (CNN based)	6	56	99.63	[46]
Wavelet kernel local fisher discriminant analysis + Particle Swarm Optimisation + SVM	10	14	98.80	[47]
ADCNN	16	14	99.51	[34]
Multi-scale deep CNN	10	90	98.87	[48]
CNN-LSTM	10	83	99.60	[49]
Attention stream net (Proposed)	**16**	**18**	**99.60**	-

Abbreviations: CWRU, Case Western Reserve University; GAN, generative adversarial network; SVM, support vector machine.

use fewer training sampling points. Despite the above, the proposed model achieves an accuracy of 99.60%. From the models under comparison, [46, 49] have similar performance but both use a larger training data set and identify fewer fault classes.

5 | CONCLUSIONS AND FUTURE WORK

The current research effort introduces an artificial intelligence model based on deep learning (DL) techniques for the task of rolling element bearings fault detection and identification in electrical machines. The proposed Attention Stream Net (ASN) is based on the concatenation of two independent neural streams that in parallel produce features with different characteristics. The first stream explores the spatially local consistency of the vibration signal based on an attentive convolutional operation, while the second produces features that examine the global content of the input mapping. Simulation experiments on two famous bearing data set benchmarks strongly confirm the efficiency of the method with state-of-the-art accuracies. Also, experimental results with the presence of additive noise show the robustness of the proposed artificial intelligence method. Future work can be applied to considering the temporal coherence of the input vibration signals with the employment of an extra attention mechanism.

ORCID
Yannis L. Karnavas https://orcid.org/0000-0002-7390-3249

Ioannis D. Chasiotis https://orcid.org/0000-0002-2726-3690

REFERENCES
1. Lau, E.C.C., Ngan, H.W.: Detection of motor bearing outer raceway defect by wavelet packet transformed motor current signature analysis. IEEE Trans. Instrum. Meas. 59(10), 2683–2690 (2010)
2. Wei, Y., et al.: A review of early fault diagnosis approaches and their applications in rotating machinery. Entropy. 21(4), 409 (2019)
3. Tavner, P.J.: Review of condition monitoring of rotating electrical machines. electric power applications, IET, 2(4), 215-247. IET Electr. Power Appl. 2, 215–247 (2008)
4. Gupta, P., Pradhan, M.K.: Fault detection analysis in rolling element bearing: a review Part A. Mater Today: Proceedings. 4(2), 2085–2094.5th International Conference of Materials Processing and Characterization (ICMPC 2016). (2017)
5. Gustafsson, F.: Statistical signal processing approaches to fault detection. Annu. Rev. Contr. 31(1), 41–54 (2007)
6. Heo, S., Lee, J.H.: Fault detection and classification using artificial neural networks. IFAC-PapersOnLine. 51(18), 470–475 (2018). 10th IFAC Symposium on Advanced Control of Chemical Processes ADCHEM 2018
7. Konar, P., Chattopadhyay, P.: Bearing fault detection of induction motor using wavelet and support vector machines (SVMS). Appl. Soft Comput. 11(6), 4203–4211 (2011)
8. Zhang, N., et al.: Naive bayes bearing fault diagnosis based on enhanced independence of data. Sensors. 18, 463 (2018)
9. Abdelkrim, C., et al.: Detection and classification of bearing faults in industrial geared motors using temporal features and adaptive neuro-fuzzy inference system. Heliyon. 5, e02046 (2019)
10. Karnavas, Y.L., Chasiotis, I.D., Vrangas, A.: Fault diagnosis of squirrel-cage induction motor broken bars based on a model identification method with subtractive clustering. In: 2017 IEEE 11th International Symposium on Diagnostics for Electrical Machines, Power Electronics and Drives (SDEMPED), pp. 304–310. Tinos, Greece (2017). https://doi.org/10.1109/DEMPED.2017.8062372
11. Helmi, H., Forouzantabar, A.: Rolling bearing fault detection of electric motor using time domain and frequency domain features extraction and anfis. IET Electr. Power Appl. 13(5), 662–669 (2018)
12. Karatzinis, G., Boutalis, Y.S., Karnavas, Y.L.: Motor fault detection and diagnosis using fuzzy cognitive networks with functional weights. In: 2018 26th Mediterranean Conference on Control and automation (MED). Zadar, 709–714 (2018)
13. Ocak, H., Loparo, K.A.: A new bearing fault detection and diagnosis scheme based on hidden markov modelling of vibration signals. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221). Salt Lake City, 3141–3144 (2001)

14. Drakaki, M., et al.: Study on fault diagnosis of broken rotor bars in squirrel cage induction motors: a multi-agent system approach using intelligent classifiers. IET Electr. Power Appl. 14(2), 245–255 (2020)

15. Shrestha, A., Mahmood, A.: Review of deep learning algorithms and architectures. IEEE Access. 7, 53040–53065 (2019)

16. Tao, J., Liu, Y., Yang, D.: Bearing fault diagnosis based on deep belief network and multisensor information fusion. Shock Vib. 2016 1–9 (2016)

17. Li, C., et al.: Fault diagnosis for rotating machinery using vibration measurement deep statistical feature learning. Sensors. 16, 895 (2016)

18. Meng, Z., et al.: An enhancement denoising autoencoder for rolling bearing fault diagnosis. Measurement. 130, 448–454 (2018)

19. Xiao, D., et al.: Fault diagnosis of asynchronous motors based on lstm neural network. In: Proceedings of 2018 Annual Conference of the Prognostics and Health Management Society. Philadelphia, 540–545 (2018)

20. Plakias, S., Boutalis, Y.S.: Exploiting the generative adversarial framework for one-class multi-dimensional fault detection. Neurocomputing. 332, 396–405 (2019)

21. Gu, J., et al.: Recent advances in convolutional neural networks. Pattern Recogn. 77, 354–377 (2018)

22. Wen, L., et al.: A new convolutional neural network-based data-driven fault diagnosis method. IEEE Trans. Ind. Electron. 65(7), 5990–5998 (2018)

23. Zhang, W., et al.: A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. Sensors. 17(2), 425 (2017)

24. Guo, S., et al.: A novel fault diagnosis method for rotating machinery based on a convolutional neural network. Sensors. 18(5) (2018)

25. Lee, J.H., Pack, J.H., Lee, I.S.: Fault diagnosis of induction motor using convolutional neural network. Appl. Sci. 9(15), 2950 (2019). https://doi.org/10.3390/app9152950

26. Yu, W., Zhao, C.: Broad convolutional neural network based industrial process fault diagnosis with incremental learning capability. IEEE Trans. Ind. Electron. PP, 1–1 (2019)

27. Jiao, J., et al.: A comprehensive review on convolutional neural network in machine fault diagnosis. Neurocomputing. 417, 36–63 (2020)

28. He, K., et al.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, 770–778 (2016)

29. Huang, G., et al.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269. Honolulu, HI (2017)

30. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, San Diego (2015)

31. Chorowski, J., et al.: Attention-based models for speech recognition. In: 29th Conference on Neural Information Processing Systems, Montreal (2015)

32. Yang, Z., et al.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, 1480–1489 (2016)

33. Li, X., Zhang, W., Ding, Q.: Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism. Signal Process. 161, 136–154 (2019)

34. Plakias, S., Boutalis, Y.S.: Fault detection and identification of rolling element bearings with attentive dense CNN. Neurocomputing. 405, 208–217 (2020)

35. Bello, I., et al.: Attention augmented convolutional networks. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3285–3294. Seoul, Korea (South) (2019). https://doi.org/10.1109/ICCV.2019.00338

36. Guyer, R.A.: Rolling Bearings Handbook and Troubleshooting Guide. Chilton Book Company, Radnor (1996)

37. Lugt, P.M.: A review on grease lubrication in rolling bearings. Tribol. Trans. 52, 470–480 (2009)

38. Affonso, L.O.A.: Machinery failure analysis handbook. Gulf Publishing Company, Houston (2006)

39. Ioffe, S., Szegedy, C.: Batch normalisation: Accelerating deep network training by reducing internal covariate shift. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on machine learning. Vol. 37 of Proceedings of machine learning research, pp. 448–456.PMLR, Lille (2015)

40. Srivastava, N., et al.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn Res. 15, 1929–1958 (2014)

41. Smith, W., Randall, R.B.: Rolling element bearing diagnostics using the case western reserve university data: a benchmark study. Mech. Syst. Signal Process. Vol. 64–65, pp. 100–131 (2015). https://doi.org/10.1016/j.ymssp.2015.04.021

42. van der Maaten, L., Hinton, G.: Visualising data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605 (2008)

43. Lessmeier, C., et al.: Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: a benchmark data set for data-driven classification. In: Proceedings of Third European Conference of the Prognostics and Health Management Society 2016, Bilbao (2016)

44. Wattenberg, M., Viégas, F., Johnson, I.: How to use t-SNE effectively. Distill (2016)

45. Mao, W., et al.: Imbalanced fault diagnosis of rolling bearing based on generative adversarial network: a comparative study. IEEE Access. 7, 9515–9530 (2019). https://distill.pub/2016/misread-tsne/

46. Pan, J., et al.: Liftingnet: a novel deep learning network with layerwise feature learning from noisy mechanical data for fault classification. IEEE Trans. Ind. Electron. 65(6), 4973–4982 (2018)

47. Van, M., Kang, H.: Bearing defect classification based on individual wavelet local Fisher discriminant analysis with particle swarm optimization. IEEE Trans. Ind. Informat. 12(1), 124–135 (2016)

48. ZiLong, Z., Qin, W.: Intelligent fault diagnosis of rolling bearing using one-dimensional multi-scale deep convolutional neural network based health state classification. In: Proceedings of the IEEE 15th International Conference on Networking, Sensing and Control (ICNSC 2018), Zhuhai (2018)

49. Pan, H., et al.: An improved bearing fault diagnosis method using one-dimensional cnn and lstm. Stroj Vestn-J. Mech. Eng. 64(7–8), 443–452 (2018)