Acoustics 2019
Sound Decisions: Moving forward with Acoustics

Use of a Deep Convolutional Neural Network and Beamforming for Localisation and Diagnosis of Industry Noise Sources

G. Capon (1), H. Chen (1), C. Bao (1,2), H. Sun (1), D. Matthews (1,2) and J. Pan (1)

(1) Department of Mechanical Engineering, University of Western Australia, WA, Australia
(2) Defence Science and Technology, Australia

ABSTRACT

Many industries rely on the continual and unimpeded operation of turbines, pumps, pulleys, fans, motors, gearboxes, and other associated fixed and mobile plant equipment. As such, reliable and remote condition monitoring and fault localisation can improve safety, prevent unnecessary downtime and reduce maintenance costs. Current methods of condition monitoring such as acoustic emissions (AE) testing can prove difficult to automate and require careful analysis by a trained analyst. This research investigates the use of adaptive beamforming for source localisation and signal extraction in conjunction with a convolutional neural network classification system based on spectrogram plots. Furthermore, it tests the effects of reducing the number of microphones in the microphone array on the deep network classification accuracy. This technology has been investigated with the use of 12-volt computer fans as an analogue for rotating machinery, with the primary challenge of reliably separating and classifying the unique spectral signal of each fan. The outcome of this research from over 450 test samples demonstrates damage detection accuracy consistently above 97% based on available data when adequate beamforming resolution and array gain are achieved. This technology shows promise for use in an automated monitoring system for industrial applications, with available scope for further refinements.

1 INTRODUCTION

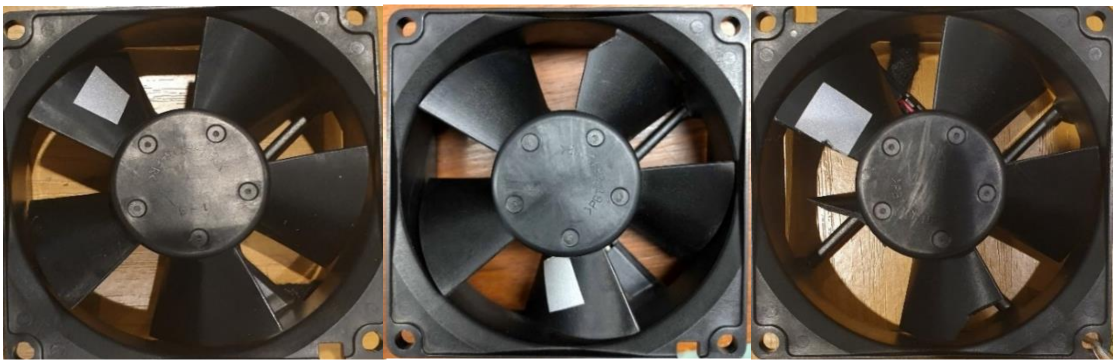
The capacity to continually and remotely monitor the performance and health of the machinery can reduce downtime, save maintenance costs, improve safety and improve productivity by finding a fault before it causes significant interruptions to production or poses any health and safety risks to personnel. For instance, maintenance costs in the mining industry can represent as much as 35% of total operating costs (Dhillon 2008). Subsequently, for the oil and gas sector, real-time condition monitoring is essential for immediate response to prevent loss of production, environmental damage and human life. All mechanical systems generate distinct noise and the sound signal can indicate the health of the system (Ravetta, Muract and Burdisso 2007). Current technologies for automated condition monitoring and damage localisation of industry noise sources such as acoustic emissions (AE) testing with accelerometers are often cumbersome prone to errors (Grabowski, et al. 2014). Furthermore, when AE testing is un-automated it requires detailed analysis by an expert technician.

In order to implement beamforming for source localisation and signal extraction, an array of microphones must be utilised. A microphone array comprises a set of microphones arranged such that spatial information can be captured (Benesty, Chen and Huang 2008). The spatial information captured by the microphone array makes the problematic task of isolating different sound sources a reality in combination with beamforming algorithms.

A deep convolutional neural network (CNN) is a class of neural networks for processing array-based data, such as an image. The primary benefit to a CNN is that it does not require handcrafted feature extraction (Yamashita, et al. 2018), it is autodidactic and learns data features from a training data set. Furthermore, complex representations can be formed for images from simple building blocks that feed-forward through the neural network, making CNN’s scale comparatively better than conventional fully connected neural networks while being less prone to overfitting. For this research, we have utilised a CNN to capture time and frequency domain information from spectrograms. The aim of this paper is to show the potential for a deep CNN to be utilised in combination with beamforming for signal extraction to perform real-time condition monitoring.

2 INDUSTRY NOISE SOURCE ANALOGUE

Due to a lack of access to real industry equipment to test, 12-volt computer fans were utilised as a cheap and easily repeatable analogue for rotating equipment. Three test fans were used to train the deep network. One is healthy, one slightly damaged (to pose a challenge for the deep CNN) and one heavily damaged. These fans are shown in Figure 1 and are denoted by a number. This number will be used to refer to the fan throughout this paper and is also the classification label used by the deep CNN.



(a) (b) (c)

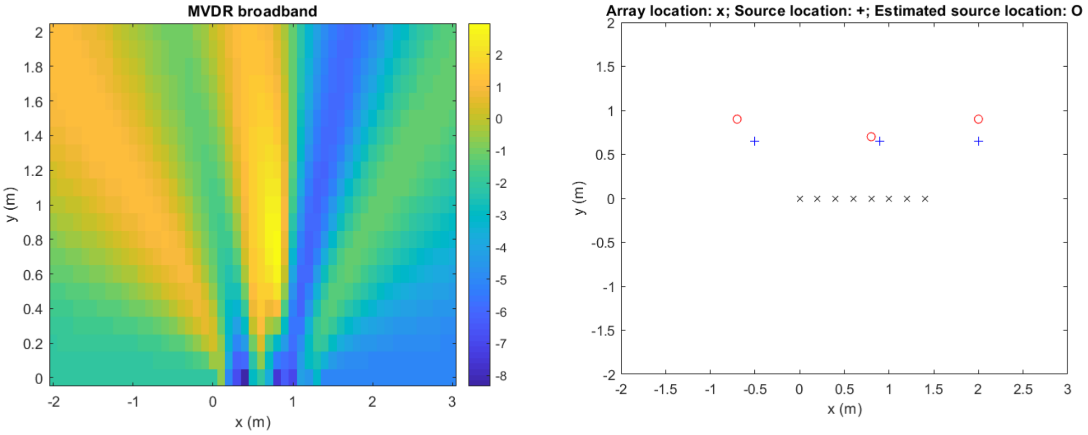
Figure 1: Test Fans: (a) Healthy (Fan#1) (b) Slightly Damaged (Fan#2) (c) Heavily Damaged (Fan#3)

3 BEAMFORMING

For this stage of the research, we assume two-dimensional beamforming to be appropriate with a uniform linear array of microphones (ULA). In our investigation, we divide the beamforming task for condition monitoring into two phases. In Phase 1 we scan the whole area of a selected site to find exact locations of sound sources of interest. This phase is termed the source localisation in this paper. In Phase 2 we extract the signal for a given sound source by beamforming only to the source location which is obtained in Phase 1. This phase is termed signal extraction. Once Phase 1 and 2 are completed we compute the spectrogram for that signal for classification.

3.1 Beamforming Algorithm

Source localisation is performed by investigating the intensity map of the area concerned, which is obtained through beamforming. For our application, frequency domain adaptive beamforming (ABF) algorithms are a better choice because of their superior performance. In this study, the well-known algorithm of MVDR (Minimum Variance Distortionless Response) with diagonal loading is used (Van Trees 2002). Figure 2 shows the typical source localisation intensity plot. The array used consists of eight microphones with an inter-element spacing of 0.2m. Given the primary focus of this research is the pairing of Beamforming and a CNN for condition monitoring, eight microphones were used so that there was an adequate resolution to locate the sound sources so that the deep network classification system wasn't hindered by the beamforming phases in any way.



(a) (b)

Figure 2: Typical Beamforming Plots: (a) MVDR Broadband (b) Source Localisation

The beamforming requirements for signal extraction are much simpler than for source localisation. Primarily, there is no need to beamform the entire area. The estimated signal location coordinates can be beamformed directly.

3.2 Spectrogram Formation

In the testing, the fans are sampled for segments of 60s length to achieve an excess of resolution in the time and frequency domains for spectrogram formation, such that the data collection would not limit the performance of the CNN. The resolution in frequency is 12 Hz and in time is approximately 0.67 seconds. Testing has been performed under constant RPM and variable RPM conditions to demonstrate the performance of the model under start-up and shut down conditions and continual operation.

3.2.1 Spectrogram characteristics constant RPM

Figures 3-5 show the typical spectrogram for each fan under constant RPM conditions. We note that high power frequencies are mostly stable with small fluctuations to the frequency and to the power. Furthermore, under these conditions, there is minimal difference between the spectrograms of the healthy and slightly damaged fan. Due to this fact, it is expected that the primary source of confusion for the deep network will stem from correctly classifying these two fans despite their similar features.

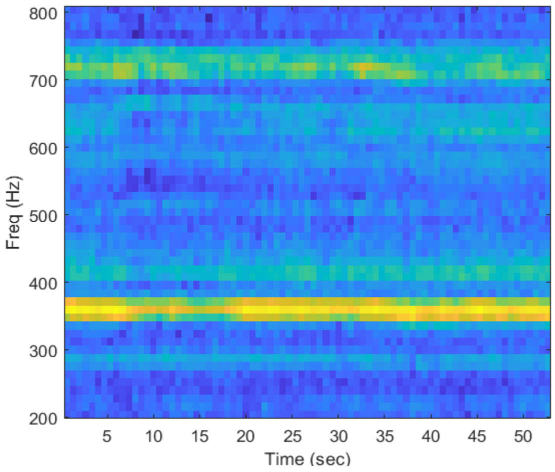


Figure 3: Spectrogram Fan#1 Constant RPM

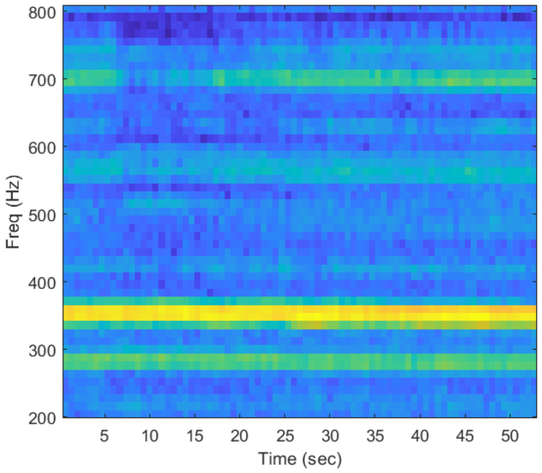


Figure 4: Spectrogram Fan#2 Constant RPM

3.2.2 Spectrogram characteristics variable RPM

In order to replicate the ramping nature of machines up until their operating speed and then their eventual shutdown, we have performed tests which ramp up the fans to their maximum voltage over the first 30s of the test and then ramp down for the remaining 30s. Figures 6-8 show the typical spectrogram for each fan under variable RPM conditions. From an observation of the data, characteristic features that indicate the fan is damaged are clearly accentuated by the fluctuating RPM. Theoretically, this should make it simpler for the CNN to correctly classify the fans.

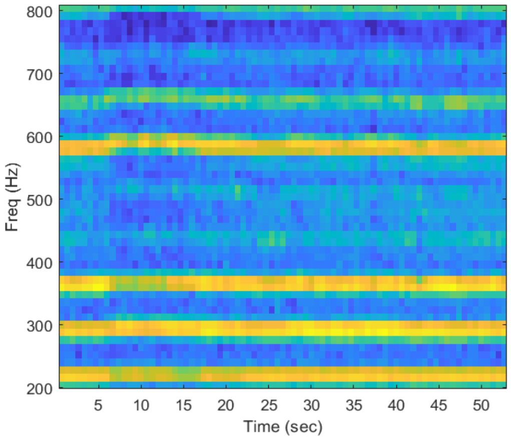


Figure 5: Spectrogram Fan#3 Constant RPM

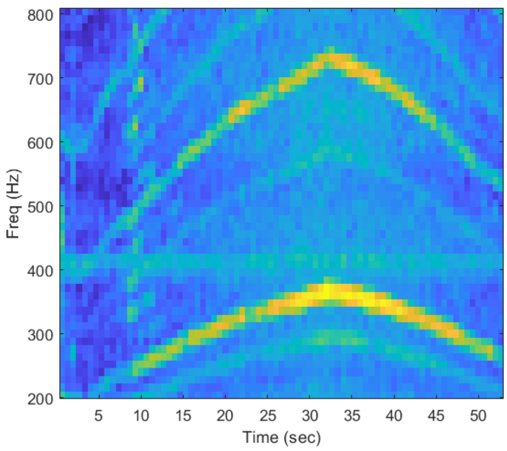


Figure 6: Spectrogram Fan#1 Variable RPM

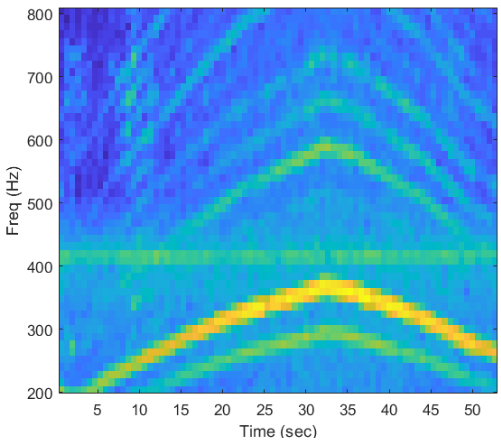


Figure 7: Spectrogram Fan#2 Variable RPM

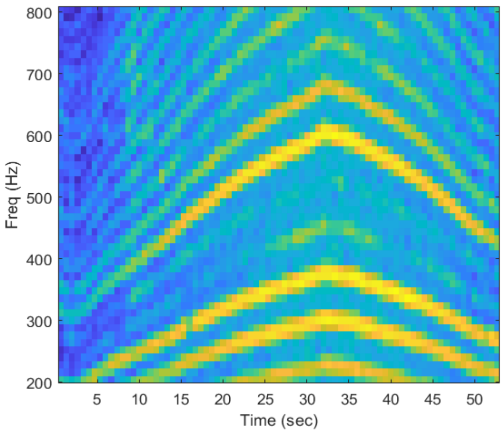


Figure 8: Spectrogram Fan#3 Variable

4 DEEP CONVOLUTIONAL NEURAL NETWORK

Due to the simplicity of the spectrogram plots, a simple 15 layer CNN has been developed for this research paper based loosely off the AlexNet architecture (Krizhevsky, Sutskever and Hinton 2012) with some distinct simplifications and implementation of newer neural network techniques. In particular, batch normalisation has been used to normalise the inputs to specific layers to increase tolerance to higher learning rates and to provide some regularisation. Furthermore, ELU activation layers have been used instead of ReLU layers to avoid the dying ReLU phenomenon. Lastly, a big kernel size was used in the first convolution layer with a large stride due to the need to down-sample the spectrogram blocks, which contain many pixels in a similar configuration. The deep network structure is shown in Figure 9. This architecture was utilised due to its high performance while retaining a simpler structure with low model complexity when compared to alternative CNN's. Our inputs to the network are 427x479 pixel images with three output classes (one for each fan).

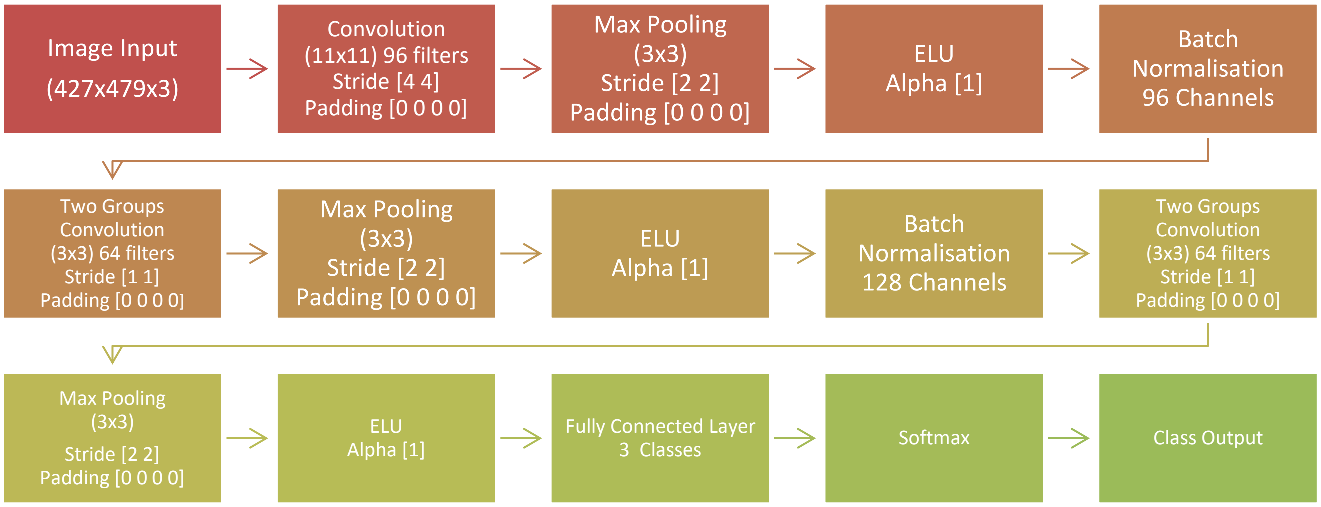


Figure 9: CNN Architecture

4.1 Overfitting Considerations

A primary concern when constructing a neural network is ensuring that the network has the ability to generalise. If the network overfits to the training data, the results of the training may appear promising, but under changing testing scenarios, the accuracy will drop substantially. The primary method to prevent overfitting used in AlexNet was to include dropout layers that are assigned a probability to temporarily exclude a connection in the neural network. This method forces the CNN to update the weights of alternate nodes resulting in a more robust network. Though research has shown that in many circumstances, batch normalisation can eliminate the need for dropout layers by allowing larger learning rates, leading to better convergence and superior network generalisation (Ioffe and Szegedy 2015). Though the regularisation effect provided by batch normalisation is a secondary effect of its primary purpose and is generally not as strong as dropout regularisation. The primary purpose of batch normalisation is to normalise the inputs to each hidden layer to have a constant mean and variance to reduce their sensitivity to changing inputs and to allow each layer to train more independently of the neighbouring layers. The regularisation occurs by the noise introduced when each mini-batch is scaled by the mean and variance computed on that specific mini-batch. The added noise ensures that downstream elements of the network are not overly reliant on any previous elements in the network. Since the application of this network is very focused and it is not needed to map complex data sets, it was deemed that the regularisation of batch normalisation alone was sufficient and a small mini-batch size of 32 was selected to maximise the regularisation. Further to this, our data is partitioned when training the neural network with 70% allocated to training and 30% to validation. The validation data pool provides a good unbiased reference for the model's performance while training the hyperparameters in the network and given we do not see a divergence of the training accuracy and the validation accuracy we can be confident the generalisation is sufficient for the purposes of this application.

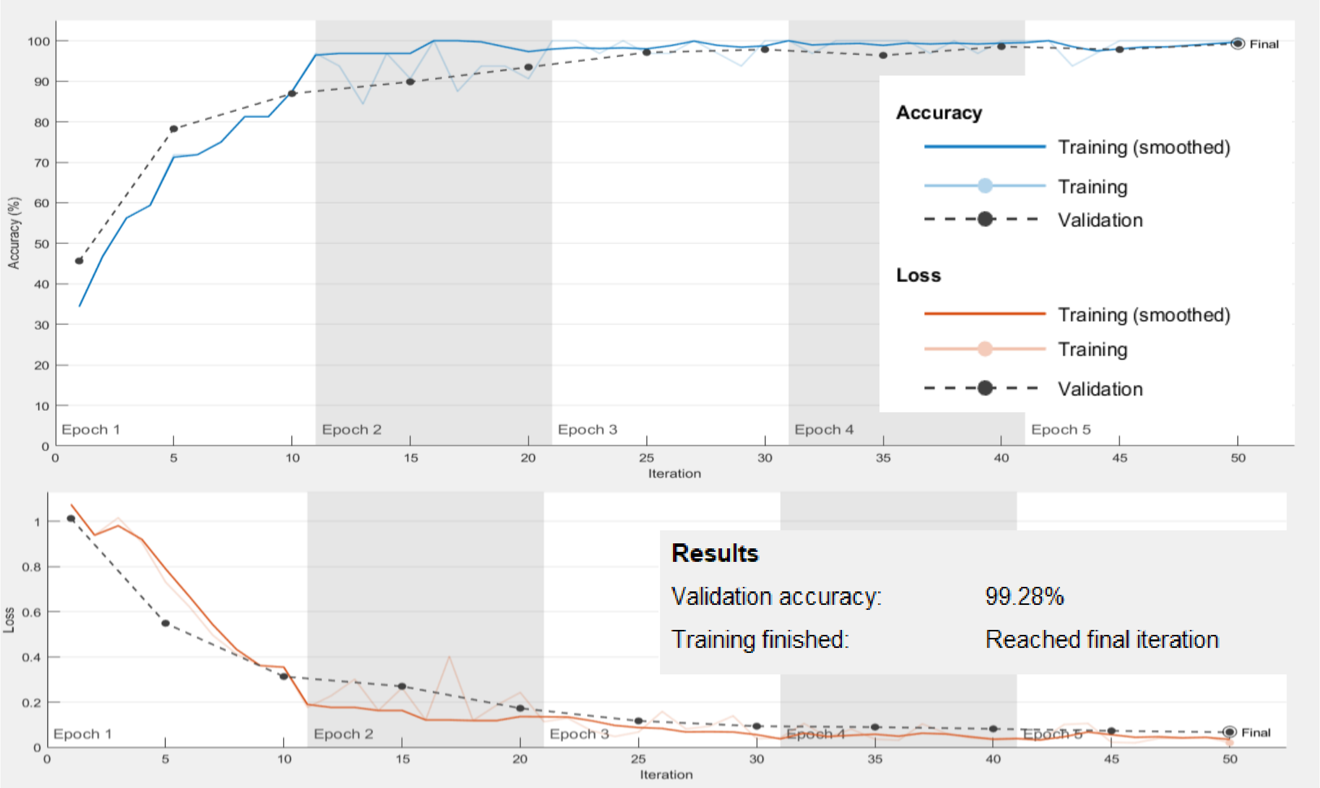
4.2 Training and Validation

Many different arrangements of the fans were used when collecting data to form a set of 460 unique spectrograms. Furthermore, an array of eight microphones with good beamforming resolution was used to ensure the spectrograms are not distorted by alternate signals. The spectrograms were then split into 70% for training data and 30% for validation data. The model was trained with inputs shown in Table 1.

Table 1: Deep Network Training Parameters

Parameter	Value
Batch Size	32
Epochs	5
Initial Learning rate	5e-4
Validation Frequency	5

Training this model has yielded high accuracy. Typically, the validation accuracy achieved is consistently greater than 97% for both constant and variable RPM conditions. Furthermore, when conducting further tests, most false classifications observed were attributable to a failure to separate the sound signals of two separate fans. These cases were primarily due to close proximity of fans, side lobes during beamforming or poor beamforming resolution. An example of the training progress and loss minimisation is shown in Figure 10.



Results

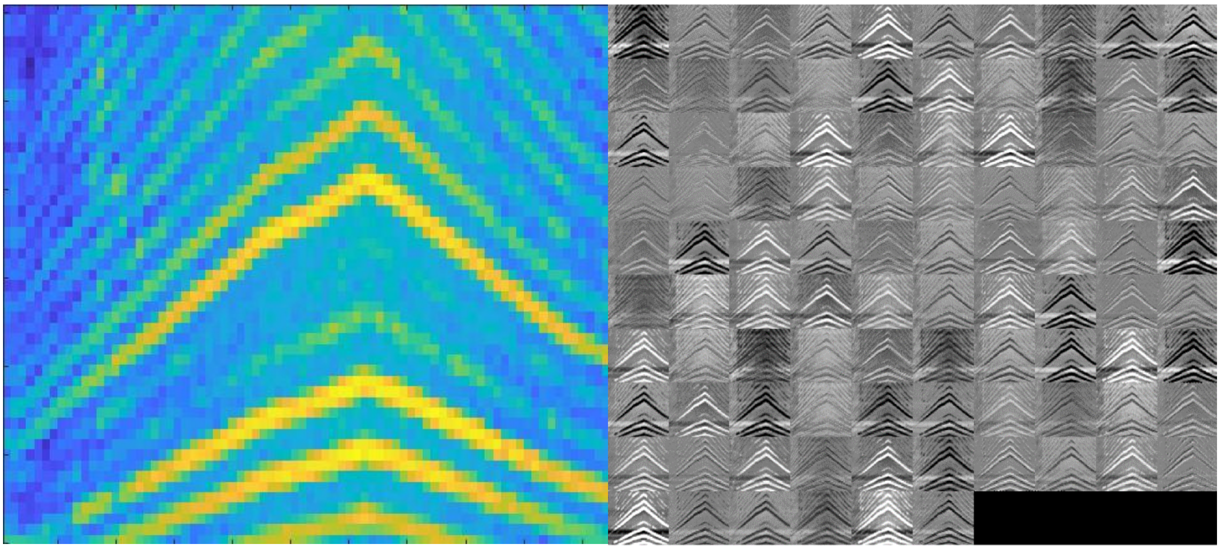
Validation accuracy: 99.28%

Training finished: Reached final iteration

Figure 10: Training Plots (Accuracy and Loss)

4.3 Network Feature Visualisation

We can see the activation response of the deep network to an input image. Figure 11 demonstrates a spectrogram generated from the damaged fan under fluctuating RPM conditions and the respective response from the first convolutional layer of the deep network. Key features and patterns that match the input image are activated and light up.



(a) (b)

Figure 11: Network Layer Activation: (a) Spectrogram of Fan (b) Network Layer Activation

4.4 Effects of Reducing Array Size

As the number of microphones required decreases, the practicality of the technology for use in industry improves. In this research project, the effects of reducing the array size without retraining the network and with retraining the network were both tested. By not retraining the network, the effects to the deep network accuracy as the beam resolution worsens and the array gain decreases are tested, simulating a situation where faulty microphone signals are discarded within the array. Secondly, by testing the model accuracy by retraining the network to compensate for the poorer resolution and array gain, the ability to adapt to a restricted setup or sub-optimal conditions can be investigated. Under this analysis, we retain a similar spacing of the fans, but change their order and place them in slightly different locations relative to the array to ensure the data collected is not too similar. Though care was taken to keep them in front of the array, rather than off at an extreme angle.

4.4.1 Effects of reducing array size without retraining the network

As discussed prior, data collection and training of the neural network was performed with a ULA of microphones consisting of 8 elements, for adequate beamforming resolution. As we reduce the array size, the array gain decreases, and the beamforming resolution degrades. This has a detrimental effect when trying to extract subtle features that indicate the condition of the fan. As the resolution worsens, spectral features from neighbouring sources will begin to appear in the extracted spectrogram for a specific fan. This presents great confusion for the CNN and can lead to incorrect classification due to problems with signal extraction rather than network accuracy. Furthermore, the decreased array gain makes the data more susceptible to noise and can reduce the visibility of subtle features in the spectrogram.

It was discovered that classification accuracy does not degrade significantly for the damaged fan when decreasing the number of microphones as it is the dominant signal and minor spectral features from the neighbouring fans do not greatly alter its spectrogram output. Secondly, for the healthy fan, accuracy began to suffer when the number of microphones was less than four, due to the aforementioned resolution issues. Lastly, for the slightly damaged fan, the accuracy suffered when less than seven microphones were used. This is most probably due to the reduced array gain failing to reveal its subtle damage. It then dropped once again when the number of microphones was less than four, due to poorer resolution. These observations are shown graphically in Figure 12.

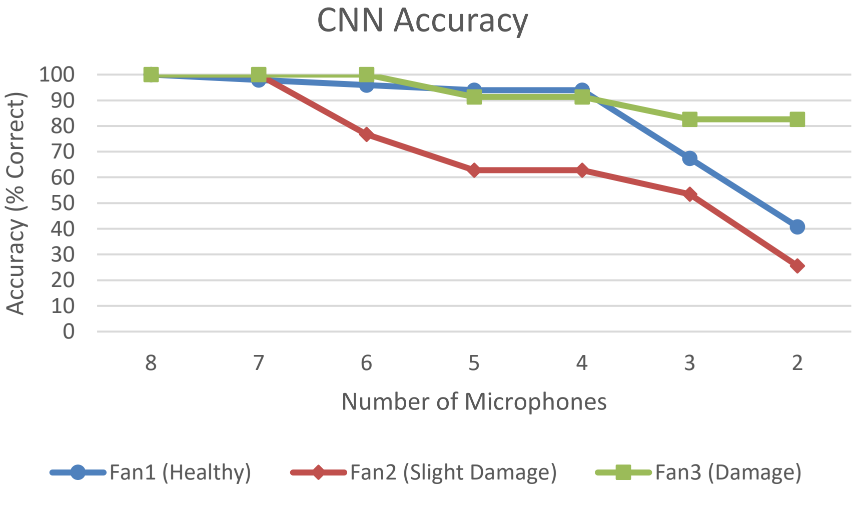


Figure 12: CNN Accuracy with Decreasing Microphones

4.4.2 Retraining the network for smaller array sizes

This research has also tested if it is possible to mitigate the effects of poorer beamforming resolution and a reduction to array gain by retraining the network for a specific number of microphones. It can be seen in Figure 13 that retraining the network can compensate for these factors, though the accuracy is notable lower than a properly set up array with adequate resolution and array gain. With retraining the network for the smaller array size, accuracy only appreciably degrades when under four microphones are used. Retraining the network ensures that despite the likelihood of any given spectrogram having faint spectral features of its neighbouring fans, the network can be made accustomed to this fact and more robust to it. When dealing with quite a low number of microphones and poor resolution, retraining the network simply adapts it to a very specific testing setup for an exact number of microphones and is not very transferable between testing scenarios. This approach is generally not best practice and furthermore, source localisation can become more troublesome when the beamforming resolution is inadequate. Ideally, it is best to use an array with enough resolution based on a given test setup for proper localisation and to completely separate the sound source signals when extracting data for training and classification. This would provide the greatest utility and reliability over a vast number of potential applications or testing scenarios. For instance, the minimum optimal array size based on the testing in this research would be an array size of seven microphones, and to train the neural network based on data collected from this array. This conclusion can be drawn by examining Figure 12, where the accuracy of the network trained on excellent resolution beamforming data and high array gain only begins to degrade after dropping below seven microphones for the typical fan spacing. Thus, we can be confident that under our testing scenario, an array size of seven or greater microphones provides enough resolution and array gain. Furthermore, we can see in Figure 13, high accuracy can be achieved with a seven-element array once it is retrained to this data.

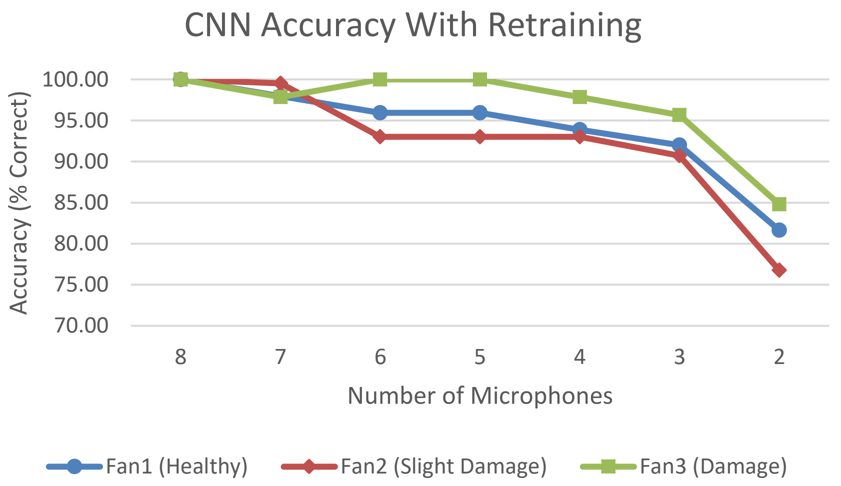


Figure 13: CNN Accuracy with Decreasing Microphones and Retraining the Network

4.5 Comparison of CNN Accuracy to Simple Statistic-Based Classifier

A simple statistics-based classifier was also developed within this research to compare the relative benefits of the CNN and the differences in performance. Initially, many algorithms and features were selected and tested. Then systematically, the number of useful features was reduced without compromising on accuracy. Cross-validation folds were also used to provide a good measure of accuracy. Even before detailed testing and comparison began, two significant benefits to the CNN approach was identified over a simpler statistics-based classification model. If it is required to monitor machines that have a variable RPM nature such as a gas turbine that will ramp up to speed, the CNN is not hindered in any way. In fact, the classification problem becomes simpler as discussed in Section 3.2.2. Though under these conditions, the simple statistics approach becomes limited due to the fact that the important characteristics vary as the RPM changes. For example, under constant RPM conditions it is possible to examine the power at the blade passing frequency of the fan, but under variable RPM conditions the blade pass frequency is always changing. Therefore, one cannot rely on this parameter for classification under variable RPM conditions. This is also observed across other parameters as well. Secondly, since different statistics are needed for the variable RPM case, a second model will be needed, whereas the CNN model can handle both constant and variable conditions in a standalone fashion. In the following sections, we will separate constant and variable RPM conditions and consider the potential accuracy under these two different scenarios. For conciseness, only the performance under adequate resolution and array gain with eight microphones will be shown.

4.5.1 Constant RPM statistic-based classifier

A Subspace KNN algorithm was found to be the most accurate for this model. The statistical parameters used to classify the data from each fan for constant RPM conditions are shown in Table 2.

Table 2: Classifier Parameters – Constant RPM

Statistics
Number of Prominent Peaks in the Power Spectrum
Blade Pass Frequency Power
Mean Frequency
Median Frequency
3db Bandwidth of Blade Pass Frequency Peak
99% Occupied Bandwidth for Power Spectrum

Based on the same signal data that was used to generate the spectrograms for the CNN model, the best accuracy that could be obtained was 92%. This accuracy is much lower than that achieved by the simple CNN. For the statistics model, practically all the false classifications were seen to be between the healthy and slightly damaged fans due to their very subtle differences. This is shown in Figure 14 where the healthy fan is class 1, the slightly damaged fan is class 2 and the damaged fan is class 3. It is difficult to separate these two fans based on statistics derived from their signals and as a result, the model predicts a healthy fan as being damaged at a significant rate. The CNN model does not exhibit this same concerning behaviour which would lead to many false alarms.

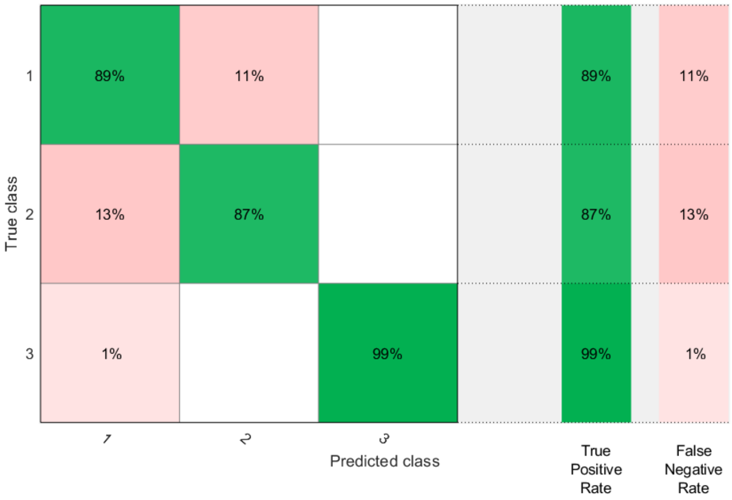


Figure 14: Subspace KNN Model Accuracy for Constant RPM

4.5.2 Variable RPM statistic-based classifier

A Cubic SVM algorithm was found to be the most accurate for the parameters used for the variable RPM case. The parameters for the variable RPM conditions are shown in Table 3.

Table 3: Classifier Parameters – Variable RPM

Statistics
Skewness of Power Spectrum
Standard deviation of audio data
Mean Frequency
Median Frequency
Mean absolute deviation of audio data
99% Occupied Bandwidth for Power Spectrum

The best accuracy that could be obtained was 83% and it required the addition of a few time-domain statistics, which were not beneficial for the constant RPM case. This model also struggles to differentiate between the healthy and slightly damaged fan and significantly underperforms when compared to both the model used for constant RPM and the CNN. It can be concluded that this condition monitoring approach is not suitable for variable RPM machines. The model performance is shown in Figure 15 where the healthy fan is class 1, the slightly damaged fan is class 2, and the damaged fan is class 3.

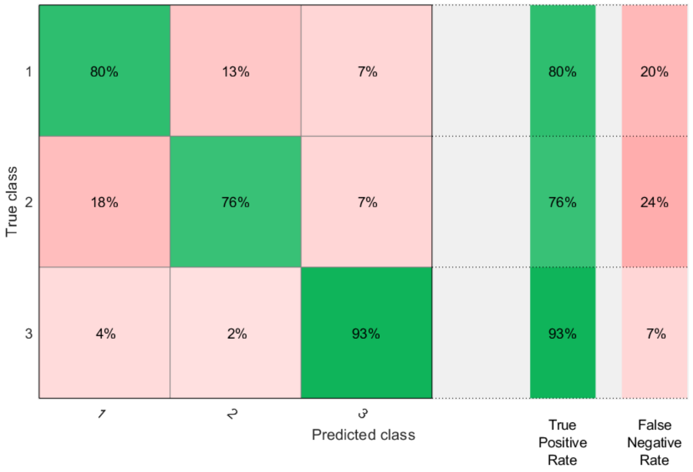


Figure 15: Cubic SVM Model Accuracy for Variable RPM

5 CONCLUSIONS

This research demonstrates that it is feasible to use a combination of a microphone array and a deep CNN to locate sound sources and perform condition monitoring with damage detection accuracy over 97%, even when considering subtle damage. Though it is apparent that adequate beamforming resolution and sufficient array gain are required to achieve these levels of accuracy in a repeatable way. Furthermore, a basic CNN model has been shown to significantly outperform other simple classification techniques, while being easily adaptable to both constant and variable speed machines. This technology shows promise for use in an automated monitoring system for industrial applications, with available scope for further refinements and improvements to the CNN model and microphone array to achieve a more robust system.

REFERENCES

Benesty, Jacob, Jingdong Chen, and Yiteng Huang. 2008. Microphone Array Signal Processing. Berlin: Springer.

Dhillon, B.S. 2008. Mining Equipment Reliability, Maintainability, and Safety. London: Springer.

Grabowski, Krzysztof, Mateusz Gawroński, Wiesław Jerzy Staszewski, Tadeusz Uhl, Ireneusz Baran, Wojciech Spychalski, and Paweł Paćko. 2014. “Acoustic Emission Source Localization in Thin Plates through a Dispersion Removal Approach.” In Proceedings of 31st Conference of the European Working Group on Acoustic Emission. EWGAE.

Ioffe, Sergey, and Christian Szegedy. 2015. “Batch Normalisation: Accelerating Deep Network Training by Reducing Internal Covariate Shift.” In Proceedings of ICML 2015. Lille, France. 448-456.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. Toronto, Canada: University of Toronto.

Ravetta, Patricio, Jorge Muract, and Ricardo Burdisso. 2007. “Feasibility Study of Microphone Phased Array Based Machinery Health Monitoring.” In Proceedings of Mecanica Computacional, Vol XXVI. Córdoba. 23-37.

Van Trees, Harry L. 2002. Optimum Array Processing. Wiley.

Yamashita, Rikiya, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. 2018. “Convolutional neural networks: an overview and application in radiology.” Insights into Imaging 611-629.