Environmental Noise Classification through Machine Learning

Clayton Sparke (1)

(1) Advitech Pty Ltd, Mayfield West NSW, Australia

ABSTRACT

Machine Learning is often cited as an effective means of analysing complex datasets. This paper presents a case study that considers the ability of Machine Learnt models to identify common sources of environmental and industrial noise in rural receiving environments. Preliminary assessment indicates that Machine Learnt models are at least as effective as a human listener at identifying certain sources of environmental noise. Furthermore, the results of source contribution assessment using classified models are consistent with manually derived contribution results. These classification frameworks may present a means for automatically screening extraneous noise contributions from environmental monitoring data.

1 INTRODUCTION

Over the past decade noise monitoring practices have evolved in concert with advances in computing and communications technologies. Vast amounts of monitoring data are now collected, often via containuous monitoring systems. These systems acquire data from remote locations, process the results and deliver relevant information to noise managers in real time.

While various methods exist to measure and report on levels of ambient *sound*, reliable assessment of *noise* contributions (ie sound from a particular source) still relies heavily on the intervention of appropriate qualified people. This raises questions about the capacity for continuous systems to return relevant information in a timely manner.

Despite the use of modern, real-time monitoring systems, the task of evaluating contributions from a specific source (ie mine noise) still requires some cognitive effort. This typically involves combining quantitative (ie measured sound pressure levels) and qualitative measurement data (ie field observations, recorded audio) to discriminate *noise* from ambient sounds. The required cognitive load is typically served by suitably skilled people (acousticians or others with specific training). Thus, effective noise monitoring is restricted to those circumstances that permit an operator to be in attendance, or is available to interpret remotely acquired measurement data.

This limits the duration and timing of noise monitoring campaigns, and typically limits experimental design to those methods that can be tended by suitably skilled people. While long term sound measurements can be undertaken without significant problem, the durations of these campaigns are often restricted by a capacity for cost-effective interpretation of noise levels within the measurement data.

Analysis methods within the domain of Machine Learning (ML) were identified as candidates for development of semi-supervised tools that may address these constraints; models trained by acousticians may enable application of their skills and experience, without need for their physical or cognitive presence.

Machine Learning frameworks are viewed as an opportunity to automate existing methods, increasing the volume and scope of analysis that is achievable; rather than a replacement for acousticians. These improvements may permit noise assessment at times when a skilled operator is (or was not) present (i.e. assessment of past events subject to complaints), or promote efficient analysis of long-term measurement data, enabling evaluation of more complex results than typically derived. A case study is presented in which analysis of historical monitoring data was undertaken to evaluate the utility that Machine Learnt classification systems mayhave in automating historically manual tasks.

2 METHODOLOGY

An assessment methodology was developed to address 2 broad objectives:

- can ML be used to identify salient noise sources in a manner consistent with the expected response of a human reviewer?; and
- can the same classification data be used to identify and exclude extraneous source contributions, and thus enable contributions from target sources to be quantified?

2.1 Data Preparation

The case study presents an application of a supervised classification model. Supervised classification methods require the preparation and training of a model; this training process enables a nominated algorithm to map a function between a set of inputs (measurement data) and outputs (a noise source). Two sets of data are required for this process:

- a training set, which is comprised of examples of measurement data from known sources. Each record in the training set is annotated with a tag, identifying the actual source of noise represented by the measurement data;
- a validation set. Fundamentally the same as the training data, but this data is withheld during training. This data (previously unseen by the model) is used to evaluate (validate) the performance of the classifier.

2.1.1 Validation Dataset

The validation set was prepared through review of historical monitoring data. This review sought to identify a period representative of a typical 24-hour period (comprising day / evening / night) in a receiving environment adjacent to an open cut coal mine. The validation set was manually prepared by reviewing recorded audio and identifying the predominant source of noise at each time step (10 seconds) in the monitoring record.

Following manual classification of salient noise sources at each time step, a method was developed to evaluate mine noise contributions at 15 minute intervals. The objective of this step is to return a result equivalent to the Specific Sound or Rating Level (defined in AS1055.1-1997), or the Component Noise Level (*Queensland Noise Measurement Manual*, 2013). This method simply involved calculating the log-average of any $L_{Aeq,10second}$ results that were identified during audio review as mining noise, at 15 minute intervals. This method is considered consistent with techniques routinely cited for removal of extraneous noise in measurement data (NPfI, 2017). The analysis returned a population of assessed $L_{Amine,15minute}$ results.

This process also yielded a list of the predominant noise sources active at the assessment location.

2.1.2 Training Dataset

Review of prior project work was undertaken to identify examples of the predominant noise sources identified during preparation of the validation dataset. These measurement data (with previously documented source identifications) were aggregated to create a training dataset.

2.2 Training the Classification Model

The model was constructed and trained using Scikit-learn, a library written in the Python programming language. The library provides implementations for a range of machine learning algorithms (Pedegrosa et al, 2011). A series of models were constructed and trained using various classification algorithms. Examples of tested algorithms include: Support Vector Machines (SVM), generalised linear models (logistic regression), decision trees, nearest neighbour algorithms, neural networks and ensemble methods (such as random forest). The accuracy of each algorithm was evaluated, and the best performing models were retained.

Model performance was initially carried out by the common ML method of a train / test split. This process splits the training data into to subsets; the training split (typically 75% of the full training set) is used to train the model, while the remaining 25% (test split) is withheld and used to evaluate the accuracy of the model predictions. Once a preferred alogirthm has been identified, the split is removed and the model trained on 100% of the data.

2.3 Evaluating Performance

A functional definition of accuracy is provided to assist in evaluating performance of the classifier in line with the objectives of the case study:

- how frequently does the classifier identify the same type of noise as the manual review process?;
- when the output of the classifier is used to quantify the mine noise contribution (ie $L_{Amine,15minute}$), is this result consistent with the contribution manually derived by an expert reviewer (ie an acoustician)?

To evaluate performance in this way, the trained classification model was used to predict the salient noise source for each 10s time step in the measurement record of the validation data set. These predictions were compared with the salient sources identified by manual review, and the percentage of matching source identifications reported as the accuracy.

The salient source predictions were then used as inputs to the mine contribution assessment method. The $L_{Amine,15minute}$ contributions returned via the manually and automatically classified processed were brought together for comparison.

3 RESULTS

3.1 Manually Classified Validation Data

Manual review of the validation dataset identified a mix of mining, transportation and environmental noise sources at the assessment location. All records in the validation data set were identified as being representative of one of 11 noise source types.

Identified source types included: Aircraft (n=148), Birds (n=623), Gusting Wind (n=812), Mine & Birds (n=1,639), Mine Noise (n=2,239), Mixed Environmental (n=497), Mixed Mining & Env (n=353), Road Noise (n=269), Trains (n=1,140), Train & Birds (n=331) and Train Horn (n=15). In some cases it was necessary to identify multiple sources in a single sample (ie Mine & Birds) due to sources occupying different parts of the frequency spectrum.

$L_{Amine,15minute}$ results were calculated, and results less than 25dB(A) excluded from further analysis; manual classification of sources in this range was challenging given decreasing signal to noise ratios. Results at these low levels are also less likely to be of interest to the end user of the data. The assessed $L_{Amine,15minute}$ levels ranged between 26dB(A) and 39dB(A), and were present in 53 x 15 minute periods.

3.2 Preparation of Training Data

The training data set was comprised of approximately 14,000 examples of 10s average 1/3 octave measurement results. The training dataset contained examples representative of sources across the 10 (of the 11) class types identified in the validation data set. No prior examples of Train & Bird noise were available, so the model could not be trained on this source type.

Imbalance was observed in the training data, with obvious bias towards sources of specific interest to the end user (ie the primary target: mine noise). Examples of mine noise dominated the training data (approximately 40% of the dataset). The mixed mining sources (Mixed Mining & Env (20%) and Mine & Birds (approximately 10%)) were reasonably well represented. Examples of transportation noise (trains and road noise) comprised approximately 15%, with the remainder representing the various sources of environmental noise (gusting wind, birds etc).

While model accuracy typically benefits from larger pools of training data, review of the effects of imbalance (Haibo and Garcia, 2009) indicates that models can learn to accuractely classify outputs that are relatively rare within the data domain. The skew between naturally abundant and rare source types is referred to as relative imbalance in the classification problems. This acknowledges that some sources (ie aircraft noise) are intrinsically more rare than other sources (ie train noise).

While a well–rounded classifier (that has similar accuracy across abundant and rare sources) would be the ideal outcome for the first objective of the case study, the second objective favours a classifier that has good accuracy on a particular target source. On this basis, and given the effort required to prepare more training data, a decision was made to proceed and acknowledge the imbalance as a potential limitation.

3.3 Training the Classification Model

Several variations of early models were developed to identify the best performing classification algorithms. Preliminary assessment of classifier accuracy was achieved by splitting the training data into training and test subsets. The training subset (approximately 75% of the training data set) was used to train the model, while the remaining 25% (test subset) was used to evaluate the accuracy of the predictions. Following this assessment, the training and test sets were re-merged, and the final model retrained on 100% of the training data.

The kNeighbours algorithm consistently returned the highest performing classifier (F1 score: 92%), and was ultimately adopted for development of the final model. After training, the model was used to make predictions of salient sources for each record in the validation dataset.

3.4 Assessment of Classifier Performance

3.4.1 Agreement Between Model and Manual Classifications

Comparison of the salient source classifications was undertaken to evaluate whether agreement may be expected between the model and manual classification methods. This analysis is presented in the form of a confusion matrix, provided in Figure 1.

		Salient Source: Predicted by Classifier										
		Mine Noise	Mine & Birds	Train	Gusting Wind	Birds	Mixed Environmental	Mixed Mining & Env	Train & Birds	Road Noise	Aircraft	Train Horn
Salient Source: Manually Identified	Mine Noise	88%	0%	6%	0%	0%	0%	0%	0%	6%	0%	0%
	Mine & Birds	3%	82%	7%	0%	3%	1%	2%	0%	1%	0%	0%
	Train	10%	3%	76%	1%	0%	2%	7%	0%	2%	0%	0%
	Gusting Wind	3%	9%	38%	18%	0%	6%	25%	0%	2%	0%	0%
	Birds	12%	33%	3%	3%	13%	20%	11%	0%	6%	0%	0%
	Mixed Environmental	17%	41%	3%	0%	10%	3%	2%	0%	24%	0%	0%
	Mixed Mining & Env	26%	4%	3%	0%	0%	0%	0%	0%	67%	0%	0%
	Train & Birds	1%	28%	41%	1%	0%	4%	24%	0%	2%	0%	0%
	Road Noise	14%	7%	9%	0%	0%	14%	12%	0%	44%	0%	0%
	Aircraft	37%	16%	12%	0%	19%	3%	3%	0%	9%	1%	0%
	Train Horn	13%	13%	60%	0%	0%	7%	7%	0%	0%	0%	0%

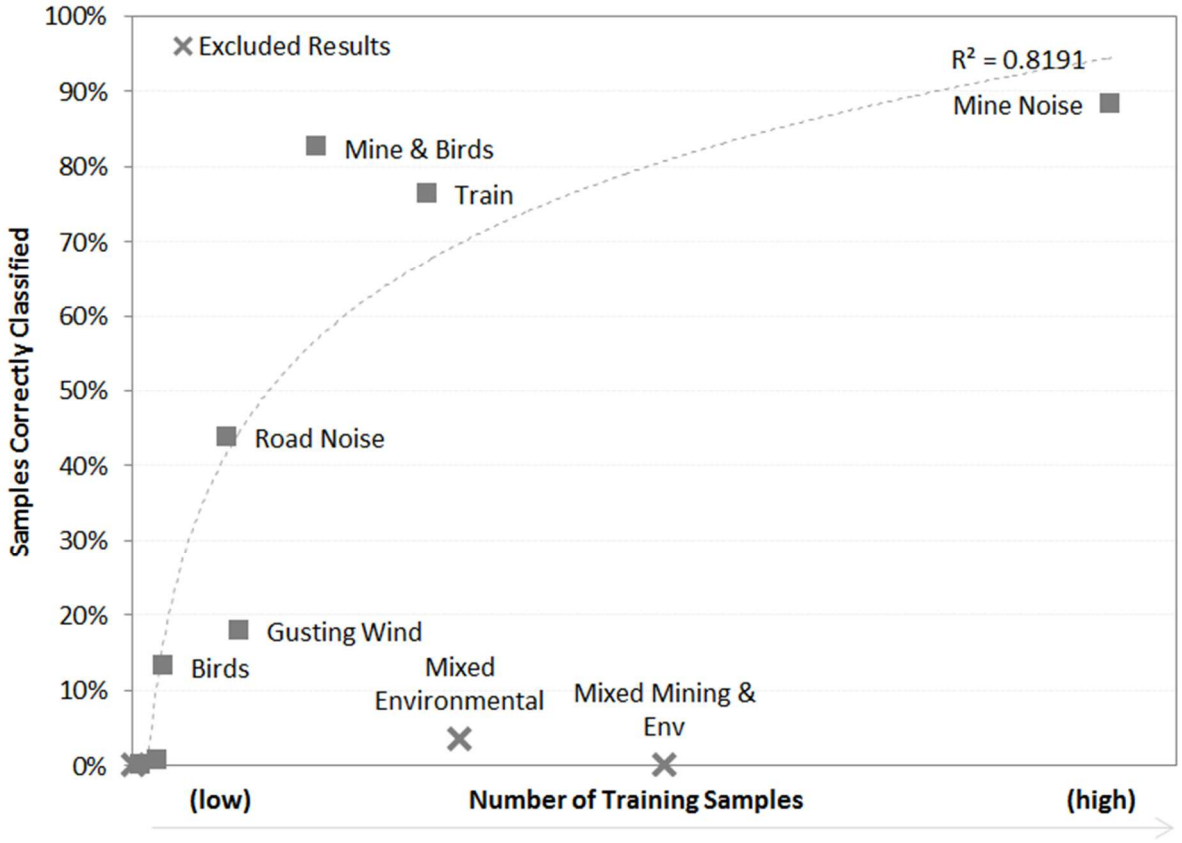
Source (Clayton Sparke, 2018)
Figure 1: Comparison of sources identified by the model and manual classification

The confusion matrix aids in evaluating the performance of the model by source type (rather than overall average accuracies). Results read along the diagonal of the matrix represent the proportion of agreement between the model and manually classified sources. Other results represent the proportion of misclassified data, and the source type to which is was incorrectly attributed. For example, the matrix shows that the model correctly classified Mine Noise (ie agreed with manuall classification) in 88% of cases; where misclassified, Mine Noise was incorrectly identified as either Road Noise (6% of cases) or Train noise (6%).

This presentation method allows for some interpretation of whether the classifier may fail in a material way; this is discussed in terms of the classifier performance on Mixed Mining & Env noise. During manual classification, this source type was identified where a mix of environmental (typically bird or road) noise occupied the foreground, with background (but audible) contributions from mining sources.

While the classifer performs quite poorly in identifying the Mixed Mining & Env source type, it did identify the measurement data as representative of either mine noise, road noise or birds in 97% of cases. Thus, while failing to correctly match with the manually identified source, these results suggest the classifier has learned the features representative of those individual sources that comprise the mixed source types.

Further analysis presented in Figure 2 explores the relationship between size of the training data set and classifier performance for each source type. When results for 'mixed' (and extremely poorly performing (<5% agreement)) classes are excluded, analysis suggests correlation between the accuracy of the classifier and size of the training data set. Thus, while some source types (such as aircraft noise or train horns) may be naturally rare, the performance of the classifier may be improved by training the model on more examples of these source types.



Source (Clayton Sparke, 2018)
Figure 2: Comparison of sources identified by the model and manual classification

Despite the relatively poor performance for some source types, good agreement (>80%) was observed for source types that are likely to be of particular interest for users of this application. While influenced by the type and complexity of the noise source under review, work on human classification (Piczak, 2015) suggests that an average classification accuracy of 80% may represent a reasonable benchmark for environmental sounds. On this basis, the classification model may be performing similarly to a human observer.

3.4.2 Agreement Between Model and Manual Classifications

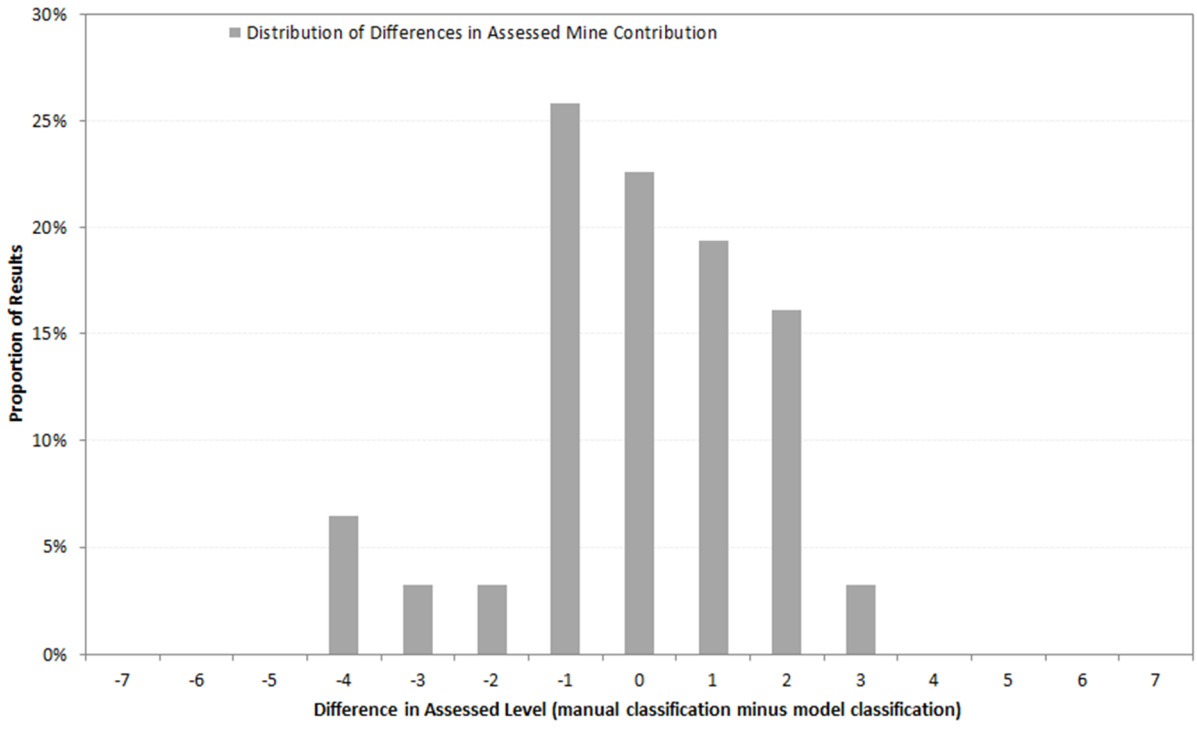
Further analysis evaluates whether the classification data may be used to quantify contributions from target sources. While the accuracy of salient source classification was variable, the classifier performed relatively well on the target source (mine noise). Analysis involved comparison of $L_{Amine,15minute}$ results derived from the model classifications and manual assessment. The results are presented in Table 1.

Table 1: Difference between evaluated mine noise contributions: model and manually classified data

Difference in Assessed Mine Contribution, dB(A)	Proportion of Data (n=53)
0	23%
+/-1	68%
+/-2	87%
+/-3	94%
+/-4	100%
+/-5	100%

While some differences were observed, analysis indicates more than half of results derived from the model classifications were +/-1dB(A) of the manually assessed result; almost 95% of results were within 3dB(A).

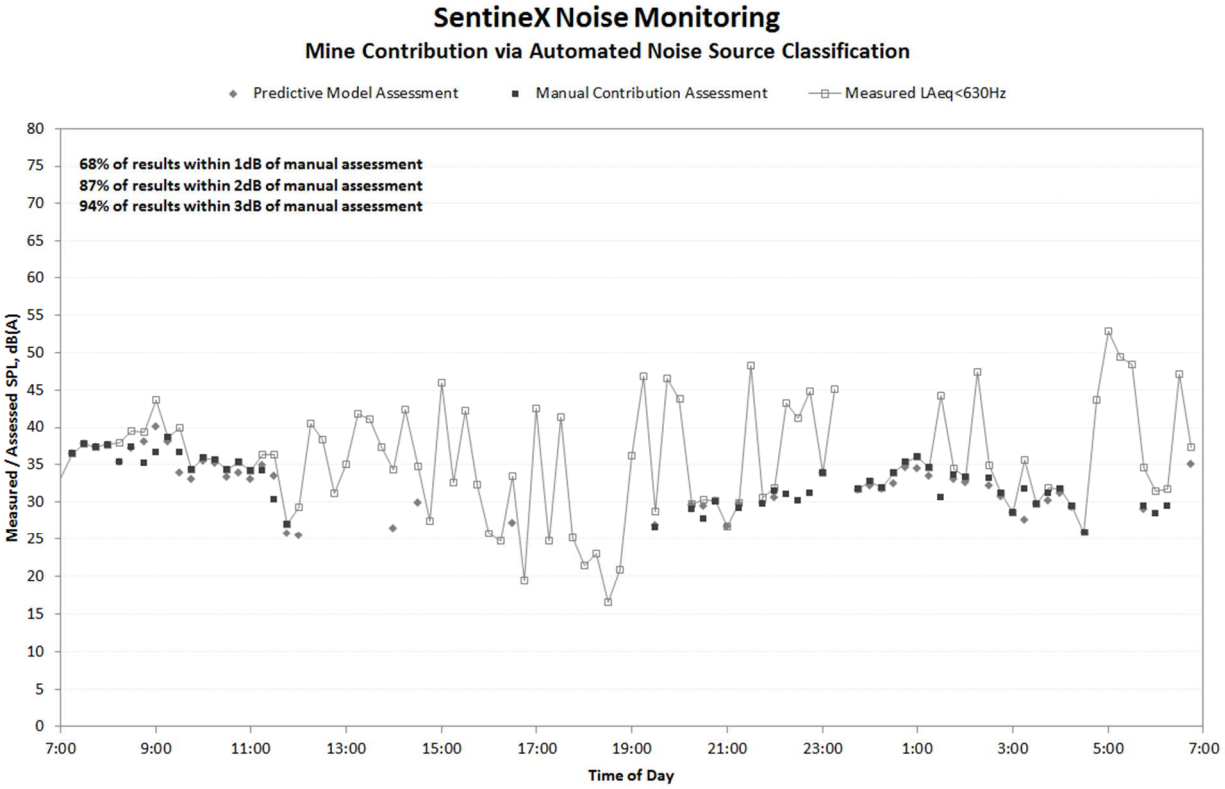
While formal benchmarks for accuracy of this task have not been not identified within the literature, the results are consistent with reported uncertainties in other aspects of environmental noise assessment. Bies and Hansen (2009) suggest that outdoor sound propogation is predictable within +/-3dB. Kerry and Waddington (2005) cite results indicating that the repeatability of measured noise levels (assessment of the same source repeated at short intervals using the same equipment by the sample operator) may lie in the range +/-1dB (at 95% confidence level). The same authors indicate that reproducibility of results (assessment of the same source using the same method applied by different operators) may be associated with uncertainties of +/-3-4dB.



Source (Clayton Sparke, 2018)
Figure 3: Comparison of sources identified by the model and manual classification

Finally, Figure 4 offers a direct comparison of mine noise contributions, evaluated using three different methods. The $L_{Aeq<630Hz}$ series applies a low pass filter to measurement data, in an effort to exclude extraneous environmental noise contributions and thus resolve underlying lower frequency contributions (such as mine noise). The potential benefits of lowpass filters (such as the $L_{Aeq<630Hz}$) have been previously demonstrated with regards to mining noise management (Parnell, 2015). Notwithstanding, these descriptors may also to respond to other sources of ambient low frequency noise such as road traffic, aircraft noise, trains, livestock and barking dogs.

The other series presented on the chart (Predictive Model Assessment and Manual Contribution Assessment) show the evaluated mining noise contributions at 15 minute intervals. While difference are observed, the results indicate that the manually classified and modelled classified results appear to follow a similar trend. The agreement between manually and model classified results differs over the course of the day; assessment suggests that agreement improves when exposed to less 'noisy' data (ie periods with fewer active sources), typically observed during the evening and night.



Source (Clayton Sparke, 2018)
Figure 4: Comparison of sources identified by the model and manual classification

4 CONCLUSIONS

A case study was presented to explore the potential utility of Machine Learning in environmental noise monitoring. An assessment methodology was developed to address the following questions:

1. Can ML be used to identify salient noise sources in a manner consistent with the expected response of a human reviewer?;
2. Can the same classification data be used to identify and exclude extraneous source contributions, and thus enable contributions from target sources to be quantified?

Results suggest that Machine Learning processes can be used effectively in the training of classification models. While the evaluated classification model clearly performs better on a subset of source types, analysis suggests that improvements would be available by refining methods used to train the model. This is expected to lead to the creation of more robust models, capable of higher accuracy across more source types.

The model was shown to express recall accuracies comparable with that of a typical human listener, for those sources targeted by the classification framework. On this basis, the assessed classifier is considered reasonably successful at replicating the response of a human reviewer when exposed to mine and train noise.

This finding is supported by the assessment of mine noise contributions. Comparison of methods indicates the classification model typically evaluated the mining noise contribution within 3dB(A) of the level manually derived by a skilled operator. Furthermore, better correlation was observed between the mine noise contributions derived from some form of salient noise identification (whether modelled or manual), than contributions evauated by lowpass filtering alone.

These findings suggest that further efforts to automate data classification are likely to be worthwhile, given the potential for classification models to return results consistent with the observations of a skilled operator. These classification frameworks may present a means for automatically screening extraneous noise contributions from environmental monitoring data. Several benefits may flow, including:

- reduction of the manual assessment burden, allowing skilled labour resources to be re-invested in more complex aspects of noise management;
- the return of consistent assessment. Exposed to the same input data, the classification model should always return the same prediction of sound class. This may reduce some uncertainty associated with different observers making different interpreations of salient noise sources;
- the ability for classification models to operate continuously, free of the of external factors such as fatigue.

The ability to effectively quantify contributions from a target source within a complex ambient noise environment will be of significant significant advantage to noise managers, and maximise the utility of continuous envionmental noise monitoring.

FURTHER WORK

This paper presents limited assessment of factors that may improve the performance of classification models. A thorough assessment of key factors influencing classifier performance is beyond the scope of this paper, and will form the focus of future work. Future assessment may consider factors such as training data preparation, size and balance of training data sets, transferability of classifiers, selection of classification algorithms, tuning of hyper-parameters and integration of non-acoustic (ie meteorological) monitoring data into the classification domain.

REFERENCES

AS1055.1-1997: *Acoustics – Description and Measurement of Environmental Noise, Part 1: General Procedures*;

Bies, DA and Hansen, CH. 2009. Engineering Noise Control, Theory and Practice. 4th ed. Spon Press, London;

Haibo, H and Garcia, EA. 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284;

Kerry, G and Waddington, DC. 2005. Considering Uncertainty When Performing Environmental Noise Measurements. *Presented in the Proceedings of the Institute of Acoustics*, Oxford, 2005; http://usir.salford.ac.uk/19519/

New South Wales Environment Protection Authority. 2017. Noise Policy for Industry. Environment Protection Authority, Sydney, Australia.

Parnell, J (2015). Acoustic Signature of Open Cut Coal Mines, *Proceedings of Acoustics 2015. Hunter Valley, Australia, November 15-18.*

Piczak, KJ. 2015. ESC: Dataset for Environmental Sound Classification, paper presented at *MM'15, Brisbane, Australia, October 26–30*.

Queensland Departement of Environment and Heritage Protection. 2014. Queensland Noise Measurement Manual (ver4).