Advancing AI-based Acoustic Classifiers for (Rail) Construction Noise – A Pilot Project

Halkon, Benjamin (1), Darroch, Michael (2), Cooper-Woolley, Ben (2), Zhao, Sipei (1), Miller, Aaron (3), Hanson, David (3), Marrinan, Matthew (4), Hendy, Andrew (4) Parnell, Jeffrey (1,4) and Mifsud, Stephanie (5)

(1) Centre for Audio, Acoustics and Vibration, UTS, Ultimo, NSW 2076, Australia
(2) SiteHive, Surry Hills, NSW 2010, Australia
(3) Acoustic Studio, Stanmore, NSW 2048, Australia
(4) Sydney Metro, Sydney, NSW 2000, Australia
(5) Gamuda Laing O'Rourke Consortium, North Sydney, NSW 2060, Australia

Abstract – Recent advances in AI technology have lowered the barrier to entry and cost of ownership of Internet-of-Things (IoT) related sensing. This may present opportunities to supplement or replace the current approaches for construction noise management with dynamic, real-time systems which can provide direct feedback to site managers. Improvements in noise predictions and better real-time tools will also convey strong benefits to surrounding impacted communities. In this paper, a recent pilot project, which aimed to establish that AI is suited to airborne noise data analysis and predictive capability and can be improved with increasing training data in the context of construction noise, will be described. Existing technology developed by SiteHive, which can predict noise sources from audio recordings, was deployed on Sydney Metro construction sites with audio gathered from a number of typical activities. Using accepted measurements of AI model accuracy, it is demonstrated that the Audio Classifier (AC) was able to increase its predictive accuracy from 29% to 81%, within 14 shortlisted classes comprising 10 construction and four non-construction categories. A strong predictive capability and rapid learning with increased body of training data demonstrate the potential of the AC in this and other applications in environmental acoustics.

1 INTRODUCTION

1.1 Background and Motivation

Sydney Metro is delivering major construction works in close proximity to residents and other sensitive receivers. Some of these works, such as stations and tunnelling support facilities, require active construction sites for several years. Sydney Metro has a strong commitment to managing the noise impacts from construction and strives for best practice both internally, through policies and standards, and externally, through contractors and key stakeholders. Construction noise modelling is at the core of management and mitigation across Sydney Metro projects. The results are captured in documents called Detailed Noise and Vibration Impact Statements (DNVISs) (SLR Consulting Australia Pty Ltd, 2023), and Out of Hours Works Permits and, in turn, guide noise impact mitigation. Construction noise predictions are typically made with sophisticated noise modelling ray-tracing software that captures how sound levels change over distance and around obstacles to receivers. Predictions, however, are only as good as the information and assumptions used to create the models. Furthermore, while there are advanced techniques for use in the assessment phase, there are also opportunities to leverage AI in the subsequent monitoring process, as will be described herein.

Noise measurements are a useful tool to guide construction noise management, both in combination with predictions and in their own right. Most Sydney Metro sites feature unattended noise monitors that capture noise levels 24/7. These systems, however, cannot substitute for a human being physically present and making observations in combination with the noise measurements. Extraneous noise sources (passing traffic, planes, birds, etc.), particular pieces of plant that may be unusually loud or require maintenance, and incidental noises may not have been included in the noise model. It is, however, not feasible for human observers to always be on site, nor is it possible for the datasets from unattended systems to be interrogated by a human team.

The ability to automatically detect and classify sound scenes and events has major potential impact in a wide range of applications and has seen increased interest in the last decade (Plumbley, 2022). AI-based audio classification for construction noise is the capability to automatically and reliably identify which pieces of plant were operating and/or which activities were taking place. Beyond audio-based approaches, alternative classification based on acoustic data are options which may be more efficient, particularly in terms of storage and processing performance. However, with the advent of Cloud-based solutions, such data reduction constraints are increasingly less important. Of the three primarily applicable analysis systems available for audio samples (Virtanen, 2018), Audio Tagging (AT), has been previously identified as more appropriate to this particular application than Sound Scene Classification and Sound Event Detection (Xiao, 2023). AT, a type of “labelling” in the emergent Data Science discipline, involves manually labelling specific segments of an audio recording descriptive tags or classes (Salamon, 2014) to indicate the presence of particular sounds. Coupled with the use of the Audio Spectrogram Transformer (AST) (Gong, 2021) deep learning technique, AT allows audio samples to be labelled with one or more tags. Computational models can be trained on large datasets of labelled audio recordings, using techniques such as transfer learning (Xiao, 2023) and data augmentation (Braun, 2024) to reduce development time and improve model performance.

The over-arching objective of the project described was to demonstrate the suitability of employing Machine Learning to assist in the management of airborne noise impacts. Testing and updating existing technology improves, over time, the predictive capability of the solution specifically for the types and nature of sound sources which can be expected to occur on Sydney Metro sites. The Audio Classifier (AC) used in this work was previously developed for integration into commercially available hardware (Cooper-Woolley, 2022) and software based environmental monitoring and management platform (SiteHive, 2024) and is based on the AST implementation of AT (Xiao, 2023). In this context, classification means correctly identifying the plant and equipment present in sound recordings taken from Sydney Metro construction sites. This project built upon previous work in the design and development of the unattended monitoring device (Cooper-Woolley, 2022) and the AC itself (Xiao, 2023). This included iterative field trials to evolve the classification system and to test its applicability and accuracy in this specific context.

1.2 Audio Tagging with the Audio Spectrogram Transformer

The AST model is similar to transformer models used for natural language processing tasks and has been modified to work with audio data. Sampled audio data are conveniently transformed into spectrograms (2D images) that represent the frequency content and noise levels at discretised intervals in time. Image-based classification solutions have received considerable attention and there are, therefore, lots of benefits to be leveraged by utilising these capabilities for the classification of audio data (Xiao, 2023). The classification algorithm analyses these spectrogram images, assigning tags depending upon the nature of the acoustic features identified. There are several “hyperparameters” specified and adjusted for the conversion of the sampled audio into the spectrogram image, which have an impact on the performance of the model. This topic has been previously explored, with the hyperparameter specifications for optimal AST performance in the context of construction noise determined in advance (Xiao, 2023). An example six sec spectrogram is provided in Figure 1.

During model development, inputs to the AC were audio samples that had been “ground-truthed”. These samples are labelled by an expert human observer, with one or more labels, in accordance with the class ontology – the set of classifications from which the AC will identify sounds in all audio signals. The classifications are subjective and must be carefully determined to be sufficiently distinct from one another and meaningful to the end user. Depending on the class ontology, the input data/samples should represent a broad range of instances which capture as many scenarios as the AC is expected to face. This includes, but is not limited to, ‘clean’ samples, i.e. samples that include only a single class with a very clear signal, through to ‘messy’ samples i.e. those which

include multiple classes of interest, some of which may be quiet, distorted or masked by the prevalence of other sounds or background noise. The set of input data should be comprehensive and unbiased, i.e. including a feature 'balance' across the set of classes, which means relatively equal representation across the entire class ontology.

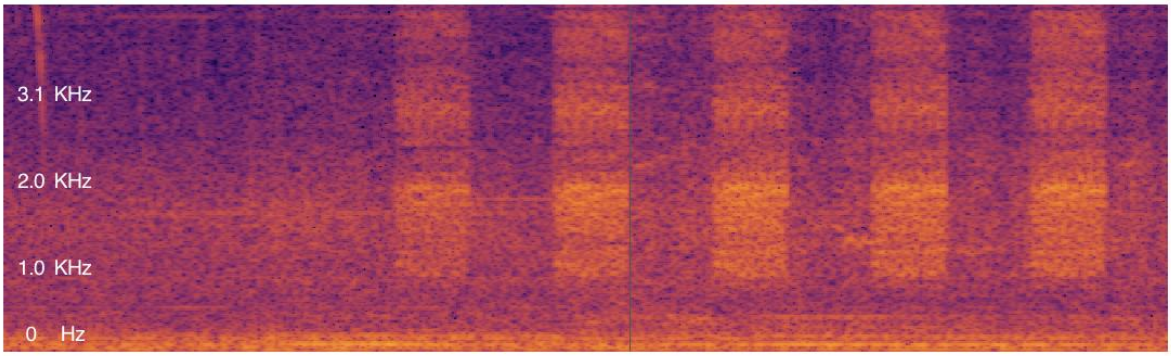


Figure 1 – Example spectrogram representation of a broadband or pneumatic movement alarm

The pool of labelled sample inputs were utilised in the AC by employing a k-fold cross-validation approach (Stone, 1974). A simple approach would be to split the input samples into a static 'training' and 'test' data set. This would, however, mean not every piece of data available is used to train the model. Alternatively, the k-fold cross-validation approach involves iteratively splitting the pool of samples into a majority group, or 'fold', of samples for training the AC and a minority fold used to automatically evaluate the model's own performance. The k-fold approach avoids the potential for biasing in the dataset that the simple approach risks by ultimately utilising the whole dataset for both training and for testing. The evaluation from the k-fold validation is used to automatically develop 'confidence thresholds'. When the AC attempts to identify one or more classes in a sample, it first reports a set of confidence *levels* for every class in the ontology. The confidence *level* is the percentage likelihood the AC predicts a class to be present in a given audio sample. Confidence *thresholds* are the percentage confidence levels above which the AC will make a positive identification. For example, if the confidence thresholds for **human-activity**, **animal-activity** and **weather-wind** are 80%, 75% and 55%, respectively, and the AC reports for a given sample is 95%, 13% and 60% for the same three classes, respectively, only **human-activity** and **weather-wind** will be identified.

2 PROJECT METHODOLOGY

The project was delivered through the completion of three phases: i) Preparation, ii) Data Collection and iii) Analysis and Reporting. The activities within each of these phases will be described here.

2.1 Preparation Phase

During the Preparation Phase, the computational, Cloud-based capability for the AC was prepared to enable rapid iteration and deployment. Firstly, the class ontology for the AC being was to be redefined, taking into account the Sydney Metro West - Western Tunnelling Package (WTP) DNVIS (SLR Consulting Australia Pty Ltd, 2023), and in consultation with domain specialists to better align it with the construction activities underway on Sydney Metro project sites. In parallel, an audio sample labelling tool was developed to allow (newly collected as well as existing) audio samples to be quickly and interactively (re-)classified by skilled practitioners, and specifically aligned with the revised class ontology. To enable more rapid model iterations to be undertaken and to track the model performance improvement following each subsequent Data Collection field trial, the existing AC was migrated from a desktop computational infrastructure into a Cloud-based environment. Lastly, construction sites onto which unattended monitoring devices were to be deployed to capture new audio samples were identified.

2.1.1 Class ontology revision

The pre-existing, proprietary class ontology – developed for more general construction noise – was revised to align better with Sydney Metro classifications used in existing noise modelling approaches, and hence to make them consistent with the DNVIS. This exercise was a collaborative process amongst the working group with feedback sought from stakeholders prior to class ontology finalisation. The class ontology must achieve the right level of detail to distinguish between plant in a way that a) supports effective classification, b) is reasonable given

the available training data, and c) is relevant to the noise impacts that the construction plant generates, as described in Table 1. The class ontology evolved throughout the project as more data were recorded and more samples covering a wider range of classes were labelled.

Table 1 – Example of defining an effective class ontology for a large, tracked excavator moving

Proposed classification	Suitability (Y/N)	Comments
Excavator	N	Too general – cannot distinguish between different types and sizes of excavator, nor the different operations of excavators, that result in different noise profiles.
Large Tracked Excavator Moving	Y	Just right – differentiates between plant at a level that supports classification (e.g. tracked vs. wheeled), represents the noise profile (e.g. large vs. small), and recognises the different operating modes (e.g. digging vs. moving) that generate different noises.
Caterpillar 345B L Hydraulic Excavator Digging Loose Clay	N	Too specific – would require a very large ontology and enormous amounts of training data to populate each classification to a level we could obtain accurate predictions from the model.

2.1.2 Labelling of (new) samples

Audio samples collected on site by the Hexanodes required labelling by experienced acoustic professionals familiar with construction noise sources to make them useful for effective training of the AC model, specifically for subsequent validation and testing. This involved an experienced practitioner listening to the audio samples and, in combination with knowledge about the equipment in use, labelling the sample with one or more of the classes in the class ontology. The project made use of an updated labelling tool, developed to use the new class ontology, to better handle evolutions to this new class ontology. This approach was more user-friendly to facilitate labelling by a wider group of experts in future stages of the project. A diagram showing the labelling approach is provided in Figure 2.

Figure 2 – Labelling approach employed during the project where effective class labels were a key component.

2.2 Data Collection Phase

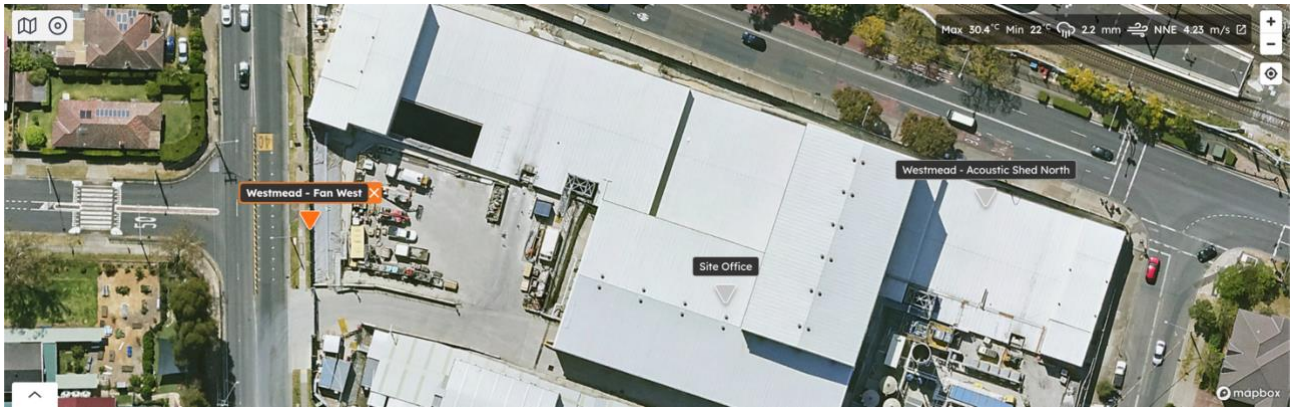
Collection of representative construction noise samples for subsequent labelling and training, validation and testing of the updated model was completed via two separate streams:

- Data collected from SiteHive Hexanode Noise unattended monitoring devices deployed to active construction sites, and
- Field trials involving attended monitoring by an acoustic expert alongside free-running data capture with an unattended monitoring device.

2.2.1 Unattended monitoring device deployments

Two unattended monitoring devices – configured to capture six second audio samples at 25 kHz Fs – were initially deployed at Westmead Metro station site in locations shown in Figure 3a. The sensitivity of the devices was increased to 75 dB(A) triggering to yield a greater range and higher volume of audio samples than would normally be collected for a construction noise monitoring application. The devices were deployed rather remotely from the construction activities to minimise the impact of their deployment on the construction works. It became apparent *after* the data collection period that the relative levels of the noise sources with respect to the background noise from fans and reverberation were too low to yield significant model improvements. Therefore, a further deployment at the Sydney Metro West Train Maintenance and Service Facility (MSF) at Clyde, as shown in Figure 3b, was assigned, where three monitoring devices were deployed.

The majority of samples captured in these locations included activities associated with construction of the Sydney Metro West Train MSF at Clyde, and the excavation of a station shaft at Westmead. These activities involved noise sources such as heavy plant idling, moving and operating, as well as, excavators, movement alarms and human activity – all major sources of construction noise impact. The ambient noise levels at all sites were determined prior to construction commencing and has been published in the DNVIS (SLR Consulting Australia Pty Ltd, 2023) and is typically in the 40-50 db(A) LAeq. Ambient noise levels determined from the monitoring devices on non-construction days, e.g. at midday on Jan 1st 2024, during the project were found to be in agreement with this for the outdoor location sand ~55 db(A) in the sheds.



a)

b)

Figure 3 – Unattended monitoring device deployment locations at a) Westmead metro station and b) MSF at Clyde trial sites

2.2.2 Attended monitoring field trials

The primary aim of the single field trial was to collect human-validated samples which would serve as ground truths for the model validation and testing. To realise this primary aim, on-site noise level measurements were made close to the noise sources. A secondary aim of the attended monitoring stage was to also make measurements at the receivers for comparison of these parameters against the DNVIS modelled predictions. However, this will need to be an objective of a subsequent study. The detailed, chronological and annotated with photographs field notes that were made to complement the attended monitoring audio samples were crucial to yield the most value from the >2000 samples were collected by the continuously recording unattended monitoring

devices. These ground-truth labelled datasets were of sufficiently high quality to be used in the AC training. One of the additional benefits realised from the attended monitoring exercise was that it enabled higher numbers of labelled samples to be collected for activities i) which only occur for a short duration, and ii) with low noise levels. In both cases, unattended monitoring would normally not yield high numbers of samples for such activities.

2.3 Analysis and Reporting Phase

Following the collection and labelling of a significant number of new samples, the AC model was iteratively re-run, and its performance evaluated. Instances of the model were generated in AWS Sagemaker (Amazon Web Services, 2024) and automatically pushed through proprietary processing pipelines to be deployed and tested through proprietary test infrastructure. The majority of data were reserved for model training and testing, as per the k-fold cross validation approach previously outlined with a smaller portion (arbitrarily 10% aim) reserved as a 'validation set'. The validation set in this pilot project is a pool of labelled samples comprised of a subset of classes of particular interest: 10 construction noise and four non-construction noise. All metrics for reporting are based on model performance against the validation set. Samples which were added to the validation set were never included in training and test to avoid 'cross pollination' of samples; instead they were embedded as the core of the test infrastructure. The results from the test infrastructure were analysed to determine areas for improvement and experimentation alongside targeted labelling efforts. After the final iteration, reporting was carried out.

2.3.1 Classification: True vs. False and Positive vs. Negative

In the context of the AC for construction noise classification and with reference to Table 2, the Confusion Matrix gives:

- A **true positive (TP)** is an outcome where the model correctly identifies the positive class, i.e. the construction activity in the sample aligns with the class in the ontology identified by the AC; conversely,
- A **true negative (TN)** is an outcome where the model correctly identifies the negative class, i.e. the model correctly identifies that a particular class in the ontology is *not* present in the noise sample;
- A **false positive (FP)** is an outcome where the model incorrectly identifies the positive class, i.e. a construction activity is *not* occurring in the noise sample, but the model indicates it is; meanwhile,
- A **false negative (FN)** is an outcome where the model incorrectly identifies the negative class, i.e. construction activity is occurring during the sample, but the model fails to identify it.

- *Table 2 – The Confusion Matrix – classification examples in the context of construction noise*

	Predicted: True	Predicted: False
Actual: True	**TP – Desired outcome** -Reality: Rock hammering on site. -Model classifies: "Rock hammering" with high confidence (above threshold of positive classification)	**FN – Undesired outcome** -Reality: Rock hammering on site. -Model classifies: "Rock hammering" with low confidence (below threshold of positive classification)
Actual: False	**FP – Undesired outcome** -Reality: no activity on site. -Model classifies: "Rock hammering" with high confidence (above threshold of positive classification)	**TN – Desired outcome** -Reality: no activity on site. -Model classifies: "Rock hammering" with low confidence (below threshold of positive classification)

2.3.2 Model Performance Assessment

For each model iteration, made after each successful data collection campaign, classifications were made against a set of test samples which were separately stored with their associated ground truth labels. The model classifications were compared with the labels to produce the following metrics: *Precision*, *Recall*, *Specificity* and *F1 Score*, in the context of the *Confusion Matrix*. *Precision* – preferred to *Accuracy* as a measure of the *quality* of the predictions – is the ratio of the true positive results to the total number of positive results predicted:

$$Precision = \frac{TP}{(TP+FP)} \quad (1)$$

Recall, meanwhile, is ratio of the true positive results to the number of samples that *should* have been identified as true:

$$Recall = \frac{TP}{(TP+FN)} \quad (2)$$

Specificity is the ratio of the correctly identified negatives with respect to all negatives:

$$Specificity = \frac{TN}{(FP+TN)} \quad (3)$$

F1 Score is the harmonic mean between *Recall* and *Precision* and is widely accepted as the preferred measure of model performance:

$$F1\ Score = \frac{2}{\left(\frac{1}{Recall}\right)+\left(\frac{1}{Precision}\right)} \quad (4)$$

Mean F1 Score across classes was the primary metric for assessing headline changes in the AC model performance. Analysis of Specificity for each class was also conducted to identify classes the model identifies poorly, informing where efforts were to be directed to improve the model, for example by obtaining more labelled audio samples in that class between model iterations. It should be noted, as the choice of classes to include in the classifier ontology evolved over time, the mean F1 Score also varied. It is not a fair test to compare the performance of two models with different class ontologies since they will yield different F1 Scores without any other adjustments to the training/validation/test data used. Therefore, some consideration was made to balance the aim of maximising model performance (i.e. achieving a high F1 Score) while maintaining desired class granularity. F1 Score also accounts for class imbalances in ground-truth labelled training data unlike simpler metrics like 'accuracy', which is appropriate given the nature of this pilot project. Iteratively comparing versions of the AC against a common test set provided a quantitative measure of model performance over time. This measure, combined with key insights, directed efforts towards improving the performance of the AC for individual classes and overall for all classes.

2.3.3 Model Adjustment and Improvement

Three primary vehicles for improving AC performance were identified in advance:

- increasing the number of quality labelled samples to train, validate and test,
- adjusting the AC class ontology to better represent the range of construction activities,
- optimising confidence thresholds.

However, it was later clear that increasing the number of quality labelled samples was the key driver of improvement, and optimising confidence thresholds is not effective in the scope of this project. Classes that presented poor F1 Scores in the performance evaluation were interrogated to determine the root cause of incorrect identifications. False positives indicated the model was mistakenly identifying a source was present in the sample. False negatives indicated the model was failing to recognise the source present within the sample. Increasing the number of quality labelled samples was prioritised. Targeting poorly performing classes by collecting more samples from various sources was a core strategy, as well as improving the quality and consistency of labelling of samples already in the dataset. In doing so, the AC evolved and its improved performance throughout the project. Altering the class ontology was a necessary but work-intensive model adjustment. Classes were added, merged or removed to improve model performance. Altering class labels involved remapping and even re-labelling samples, adjusting the labeller and remapping the test set.

3 PROJECT RESULTS

For brevity, three of the eleven model iterations completed during the project have been selected for this report to represent the key shifts in model performance. Model Run 1 was generated after the work of the Preparation Phase was completed. This iteration of the model included a number of the labelled samples from the Westmead devices deployment which was ongoing. Run 2 included the bulk of the Westmead labelled samples and an early number of those from the Clyde MSF deployment. By this stage, the 14 classes of particular interest to the project, based on the activities in the deployments, were largely identified. Run 3 model iteration was completed after the conclusion of the attended monitoring activity and the human-validated ground-truth samples were incorporated.

3.1 Final Label Counts

The final label count by the end of the project amounts to 5,473 labels within 3,461 separate audio files, of which 883 came from the attended monitoring in the field trial. Of particular interest is the increase in label counts in the 14 shortlisted classes. These increased from 1,669 in Run 1 to 2,937 in Run 3. In particular, of the 10 construction noise classes, four had the number of labels increase by approximately an order of three between Run 1 and 3.

3.2 Metrics Analysis

Figure 4 below shows the average metrics for each of the shortlisted classes across the three model runs.

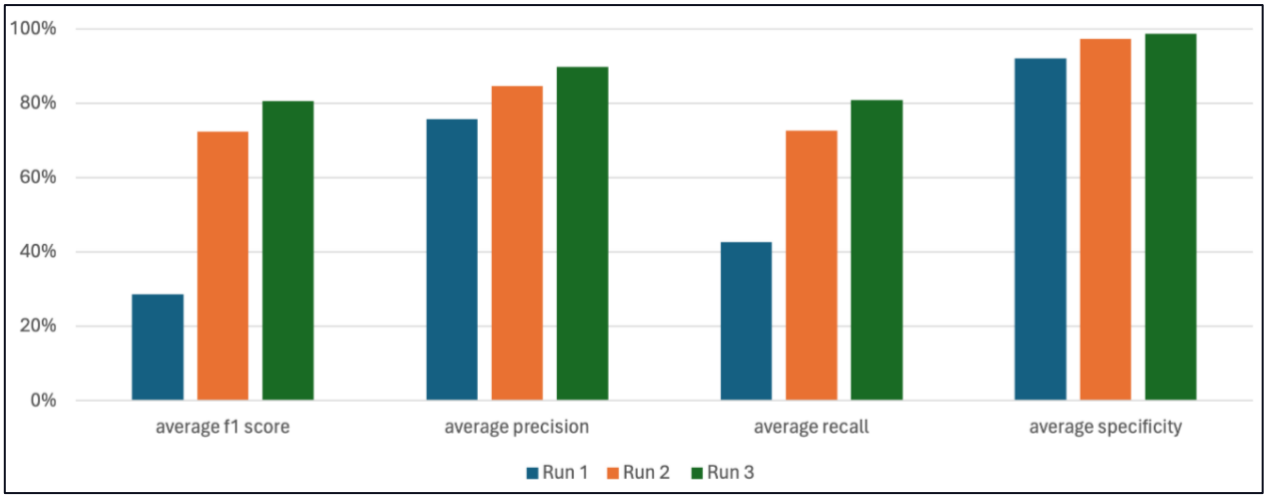


Figure 4 – AC Improvement in performance, expressed in terms of the metrics averaged across 14 shortlisted classes

The significant improvement observed in F1 Score can be attributed by analysing the Precision and Recall components. Precision was already high in Run 1 and, with more labelled samples added and the class ontology evolved, was further but only incrementally improved in Run 2 and 3. Recall, meanwhile, was substantially improved between Run 1 and 3. In Run 1, some classes contained no labelled samples such that the AC, untrained on that class, was unable to predict it returning, therefore, a Recall of zero. The compounded improvements in both Precision and Recall lead to the major improvement in the AC as measured by F1 Score.

These statistics paint a consistent picture. Across the 14 shortlisted classes against the ground-truthed test set, when the model was trained with fewer labelled samples for a given class, it was less likely to identify that class. As the number of labelled samples for a given class increased, the model was more likely to identify that class, and it tended to identify it correctly. This is demonstrated by the interplay between the Precision and Recall. Precision started high because when the model did identify a class it was likely to be correct, even from Run 1. However, Recall started lower because the model was not sufficiently well-trained to make an identification until it had been trained with more labelled samples from all classes of interest.

To further demonstrate the improvements in the AC model metrics shown above, Figure 5 below shows this interplay for an example single construction class over the three runs. Figure 5a shows the class increased from containing 41 labelled samples in Run 1 to 219 (100%) samples by Run 3. The increase in labelled samples resulted in a fluctuation in performance metrics. Figure 5b shows the class improved from a 0% F1-Score in Run 1 to 84% in Run 3 which mirrors the macro performance metrics shown in Figure 4 above. Interestingly, both

Specificity and Precision decreased significantly in Run 2, however this was offset by a vast improvement in Recall from 0 to 58%. As Precision and Specificity stabilized in Run 3 and Recall improved to 84% the F1 Score rose to a respectable 84%. It is acknowledged that performance should improve with more runs and further data, provided the new datasets follows the same pattern or characteristics of previous data. In this case, the outcomes shown in Figure 4 thereby confirm that the new data follow the pattern of those included in earlier Runs. Furthermore, knowledge of the numbers of samples in the datasets at various stages allow confirmation of the requirements for minimum sample numbers expected to achieve the performance improvements that are observed in the future.

It is hypothesised the decrease in Precision and Specificity in Run 2 was due to over-identification. The model was incorrectly identifying the selected class was present in samples when it was not. However, this over-identification meant the Recall metric was strong; when the class was actually in the ground-truth sample, i.e. the model identified it more often than not. This emphasizes the importance of having sufficient samples to train the AC as it both improves and stabilises the model performance.

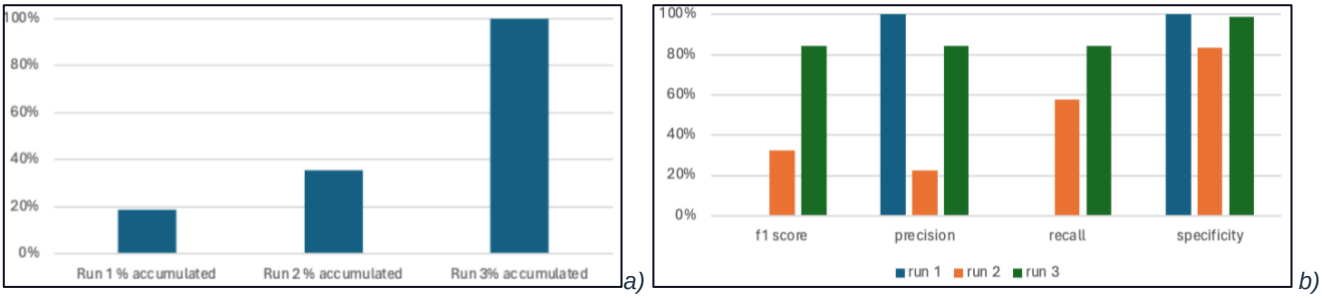


Figure 5 – Improvement in performance across the project for a single class; a) label increase, b) metric differences

4 CONCLUSIONS AND FURTHER WORK

With appropriate training data, the AC was developed with a high degree of predictive performance for the target construction noise sources on Sydney Metro construction sites, ultimately with an F1 Score of 81% for a shortlisted set of 14 classes, 10 construction and four non-construction noise. This classification capability was demonstrated in real-world scenarios with all the inherent complexity that this entails, such as multiple sources active at a time, and extraneous noise (e.g. from traffic and wind). It is likely the outcome could be further improved by increasing the number and variety of construction audio samples used to train the model. Using Audio Tagging in combination with the Audio Spectrogram Transformer technology employed by the AC was crucial to this outcome. It demonstrates construction site audio data is a suitable input into AI models and enables AI models to make useful predictions for the purpose of managing noise impacts.

Secondly, the predictive performance of the AC was improved by 52% with the addition of 1268 labelled audio files for shortlisted classes from the initial 1669 samples used in Run 1. This demonstrates a strong capability for the AC to learn, i.e. to increase its predictive performance with increasing training data. This supports the idea of further investment into building the body of training data would broaden the number of classes the model can make accurate prediction upon, and generally increase the accuracy of these predictions for all classes. With a more comprehensive ontology across the range of plant and equipment used on major projects such as Sydney Metro West, advances could be made immediately in terms of smarter notification systems and intelligent separation of background noise sources from construction noise sources.

Successfully employing construction noise spectrograms as an input into AI models responds to the fundamental question for further investigation: Is AI is a plausible supplement to current ray-tracing algorithms used by 3D modelling software for the predication of the impact of construction noise on sensitive receivers? Practically, further work will be required to develop additional techniques to build AI models that can make useful predictions based on construction scenarios as model inputs as opposed to real-time audio data. The AC will be an essential tool to accumulate a rich repository of training data to make this transition. In addition, despite the need for further work, some immediate applications of the AC may improve management approaches on the ground.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the 2023 Sydney Metro Innovation Challenge Program funding through which the work described in this paper was completed.

REFERENCES

Amazon Web Services. (2024). *Machine Learning Service - Amazon SageMaker*. Retrieved from https://aws.amazon.com/pm/sagemaker/?trk=ba84e666-9a76-470d-9c91-2864b6cfbb7a&sc_channel=ps&ef_id=Cj0KCQjw7Py4BhCbARIsAMMx-_KVLRdBjGriFiej_mUjSGrJegTUoe-_7LGiCpkALBKXtu877UpDGA4aAp27EALw_wcB:G:s&s_kwcid=AL!4422!3!532547883839!e!!g!!aws%20sagemaker!1153988

Braun, S. a. (2024, July 20). *Multi-label audio classification with a noisy zero-shot teacher.* Retrieved from arxiv.org: https://arxiv.org/pdf/2407.14712

Cooper-Woolley, B. D. (2022). Initial design, development & calibration of mems based sound level meter for real-time construction monitoring. *Proceedings of Acoustics 2022, The nature of acoustics.* Wellington: The Acoustical Society of New Zealand.

Gong, Y. C. (2021). Improving audio tagging with pretraining, sampling, labeling, and aggregation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, (pp. 29: 3292–3306).

Plumbley, M. a. (2022). Creating a new research community on detection and classification of acoustic scenes and events: Lessons from the first ten years of DCASE challenges and workshops. *INTER-NOISE and NOISE-CON Congress and Conference Proceedings* (p. https://api.semanticscholar.org/CorpusID:256854604). Glasgow: I-INCE.

Salamon, J. J. (2014). A dataset and taxonomy for urban sound research. *Proceedings of the 22nd ACM international conference on Multimedia*, (pp. 1041-1044).

SiteHive. (2024). *SiteHive home page*. Retrieved from https://sitehive.co

SLR Consulting Australia Pty Ltd. (2023, May). Sydney Metro West - Western Tunnelling Package (WTP) Detailed Noise and Vibration Impact Statement (DNVIS) - Westmead to Sydney Olympic Park. *SLR Ref No: 610.30644-R02-v5.0-20230530.docx*.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological), 36*(2), 111-147.

Virtanen, T. P. (2018). *Computational analysis of sound scenes and events* (Vols. ISBN: 978-3-319-63450-0). Springer Cham.

Xiao, T. H.-W. (2023). Environmental noise tagging via audio spectrogram transformer. *INTER-NOISE and NOISE-CON Congress and Conference Proceedings, InterNoise23* (pp. 7394-7400). Chiba, Japan: Institute of Noise Control Engineering.