ACOUSTICS 2025
JOONDALUP
SOUNDS OF THE SUNSET COAST

Evaluating Sound Pressure Levels of Noise Sources in Busy Noise Environments with Deep Learning.

Joshua Burwood (1), Clayton Sparke (1)

(1) Advitech Pty Ltd., Newcastle, Australia

ABSTRACT

Accurate and reliable estimation of sound pressure levels (SPL) for specific noise sources is critical for effective environmental monitoring, community noise management and even regulatory compliance. This study presents a novel deep learning–based framework that demonstrates the potential for high-confidence, low-error quantification of targeted source contributions in real-world settings. Leveraging a network of acoustic monitoring devices, skilled listeners reviewed data (recorded audio and 1/3 octave spectra) to classify and make determination of short-term (LAeq (energy average)) noise contributions for nine (9) noise source classes. Our method comprises three key components: (1) Expert-Annotated Event Analysis – acoustic specialists evaluate short-term LAeq contributions to each noise class; (2) Self-Supervised Pretraining, employing masked spectrogram reconstruction on unlabelled field recordings to learn robust feature representations; and (3) Supervised Regression, fine-tuning a hierarchical convolution-transformers (MaxViT) model on paired spectrogram–SPL data to predict source-specific A-weighted SPLs for a specific dataset. On an independent test set, our approach determined noise levels with a mean absolute error (MAE) of 0.8±2.1 dBA (MAE ± std). These results highlight the potential of deep learning methods for precise, source-specific SPL estimation. Integration into automated noise-assessment dashboards and mobile monitoring platforms can provide real-time decision support to environmental acousticians and regulatory agencies.

1 INTRODUCTION

A range of emerging technologies exist for assisted listening and automated classification of environmental sounds (Bansal & Garg, 2022). These technologies provide valuable support to human reviewers and environmental noise managers by automatically identifying events of interest in recorded or real-time measurement data, enabling re-direction of effort from event detection to event investigation.

These technologies yield a clear benefit, as they draw attention only to events of potential concern. However, it is typically the sound level of a source – not just its audibility – that is of concern to environmental noise managers. Accurately quantifying the contribution of a specific noise source within a mixed ambient environment can be a complex analytical task. While methods do exist to perform this analysis, they are typically manual, time-consuming, and require a high level of expertise. This productivity constraint is a limiting factor in the wider adoption of real-time and unattended noise monitoring, and use of these technologies as investigative or regulatory tools in complex noise environments.

The goal of this study is to introduce and explore the potential utility of deep learning as a solution to noise source identification and quantification challenges. We present an example workflow in which machine learning is used to both identify component sources, and estimate their sound pressure levels (SPL) in complex noise environments. Efforts are made to construct this workflow in a way that aligns with common regulatory standards (Environment Protection Agency, 2022)

1.1 Existing Machine Learning Methods

There has been recent, rapid development in machine learning (ML) and other artificial intelligence (AI) tools, which have achieved outstanding performances in a multitude of domains. Computer vision has been acknowledged to be equal or better than human level (Geirhos, et al., 2021). GraphCast (Lam, et al., 2023), a Graph

Neural Network proposed in 2023, achieved state-of-the-art performance in weather forecasting. Neural network based systems are already being deployed in real-time water quality monitoring systems (Zainurin, et al., 2022). There has been great success in the domain of environmental sound classification (ESC) (Bansal & Garg, 2022), which has shown high accuracies in the classification of common environmental sounds.

While autonomous environmental sound classifiers are increasingly performative (Alex, Ahmed, Mustafa, Awais, & Jackson, 2024), these classifiers do not typically quantify sound levels, or permit direct comparison with regulatory requirements (such as noise limit levels). A true 'noise' monitoring system would require the capacity to robustly measure and report SPLs of specific sources, enabling direct evaluation with regulated limits that aim to control noise pollution. Prior work (Sparke, 2018) has demonstrated the effectiveness of using machine learning for ESC to identify the salient noise source in rural receiving environments. However, this is limited to informing compliance only when the noises of interest are the salient noise sources, which is often not the case.

State-of-the-art ESC models, have revolutionized the accuracy of such models with transformer-based approaches (Halkon, et al., 2024). Furthermore, the Audio Spectrogram Transformers have been further optimized with hierarchical and multi axis adaptations (Alex, Ahmed, Mustafa, Awais, & Jackson, 2024), providing greater precision at lower latency.

Audio classification tasks have shown great accuracies (Bansal & Garg, 2022), similar to the performances observed in image classification tasks that have surpassed human capability (Geirhos, et al., 2021). Sound Event Localisation and Detection (SELD), has been proposed to derive temporal localisation. However, these methods are still insufficient for deriving noise characteristics from audio signals. Variational Autoencoders (VAEs) have been effective in suppressing background noise in human speech (Nogales, Caracuel-Cayuela, & García-Tejedor, 2024), and thereby isolating a particular sound source in a noisy environment. However, such reconstructions tend to be lossy, and prone to the caveats of generative AI methods, such as hallucinations.

This work aims to extend the successes of machine learning methods into the domain of noise monitoring systems, where there exists an apparent gap in the literature.

1.2 Aim of this study

The aim of this study is to present experimentation on a deep learning-based framework for the estimation of source-specific sound pressure levels (SPLs) in complex environmental noise environments. The goal is to contribute to accurate, real-time and autonomous data analysis that can assist environmental professionals with effective management of noise impacts.

2 METHODOLOGY

2.1 Input data

Environmental sound measurements were collected from acoustic monitoring devices deployed across rural New South Wales. Audio was recorded in mono at a 44.1 kHz sampling rate and encoded as MP3 at 32 kbps. Recordings were segmented into fixed 10-second intervals, following conventions in environmental sound classification literature, which also aligns the temporal granularity with the expert-attributed LAeq targets.

Each 10-second clip underwent the following preprocessing pipeline: mean normalization, conversion to a Mel-spectrogram with 128 Mel bins, a 25 ms window size, and a 10 ms hop size, and min–max normalization of the resulting magnitudes. These representations served as the input to the model.

Concurrently, 10-second average (LAeq) sound levels (including one-third octave spectra) were measured in accordance with prevailing industry standards (Approved methods for measurement and analysis of environmental noise (Environment Protection Agency, 2022)). This was achieved using Sound Level Meters (SLMs) satisfying the Class 1 requirements of AS/NZS IEC 61672.1. This ensured that each 10-second sample was represented by both precision measurement data (a single 10-second LAeq (and one-third octave spectra)) and a Mel-spectrogram derived from the 10-second audio recording. All data modalities (Mel-spectrograms, one-third octave spectra, and expert annotations) were time-aligned via their timestamps.

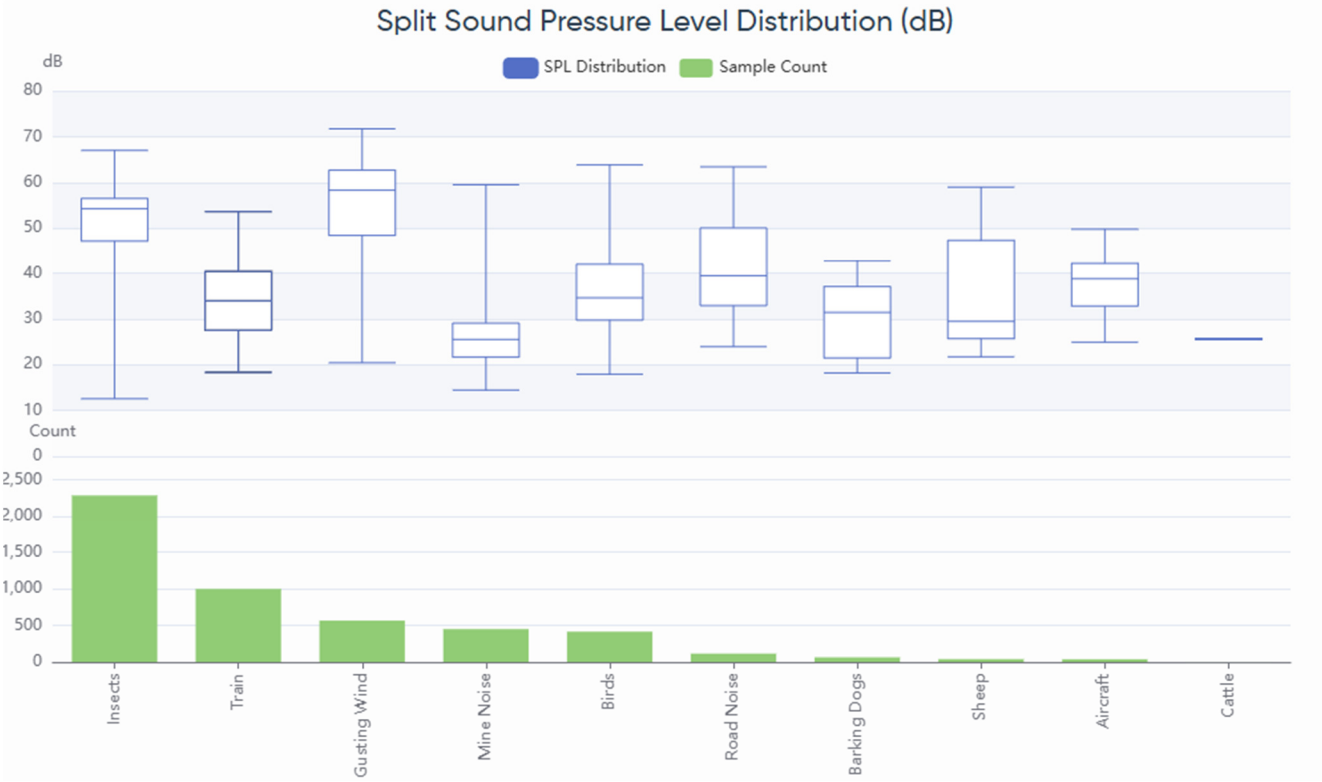
Ground truth sound pressure levels (SPLs) in A-weighted LAeq format were determined by acousticians using the recorded audio together with the available one-third octave spectra, following guidance in the NSW EPA "Approved methods for measurement and analysis of environmental noise". Each ground truthed sample was a vector containing expertly estimated SPLs for each source in the dataset. This was done to ensure that the summation of all identified sources was equal to the total measured sound pressure level (10-second LAeq) of the sample. An example of this annotated data is provided in Table 1.

Table 1: Example input data to model: vector of estimated SPLs by source. 0dBA annotations indicate that source was not audible in sample (subset of all sources is presented to ensure clarity in the table)

Datetime	Estimated Source SPL (as LAeq,10second, dBA)						
	Train	Aircraft	Insects	Mine Noise	Birds	Sheep	Total
4/3/2025 2:32:20	40.8	0	53.9	0	0	0	54.1
4/3/2025 2:32:30	43.1	0	53.9	0	0	0	54.3
4/3/2025 2:32:40	47.9	0	54.3	0	0	0	55.2

No formal uncertainty quantification was applied; however, notable sources of unquantified uncertainty included (a) situations where target sounds were very quiet relative to background, making them difficult to distinguish, and (b) adverse environmental conditions (e.g., strong wind) that degraded microphone fidelity. These considerations will be the objective of future work.

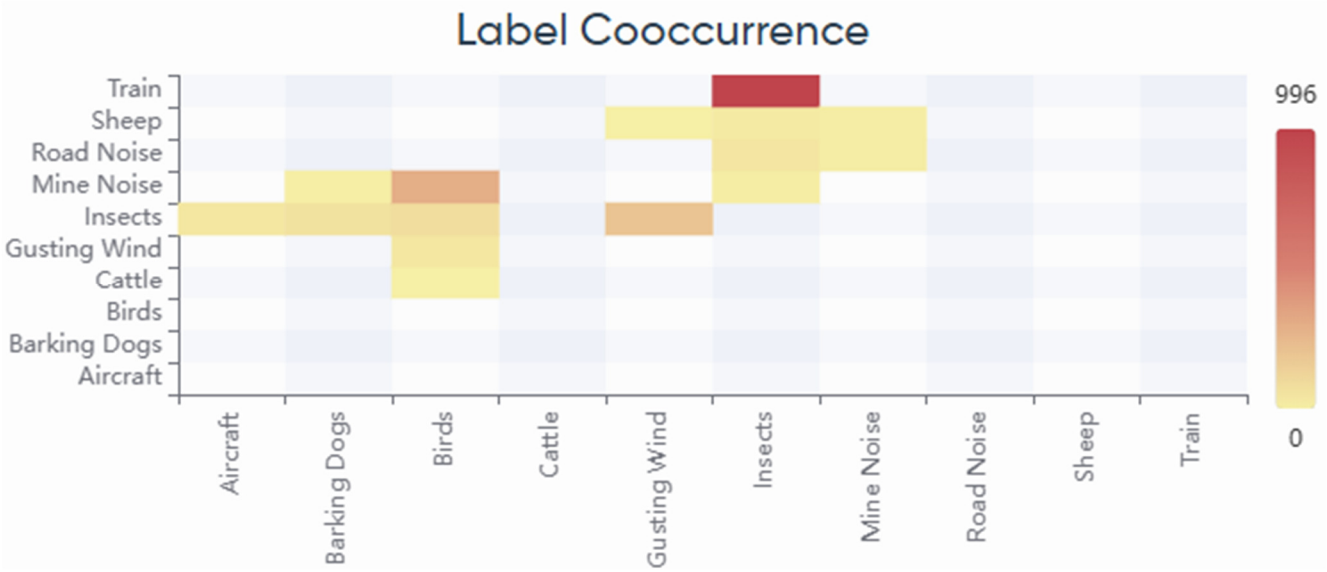
The ground-truth data was generated as a by-product of noise monitoring project work; the mix of sources identified in the dataset are thus unique to the monitoring environment, and are skewed towards the noise sources under investigation. The dataset included 4062 samples of 10-second mp3 audio and per class SPL ground truth vectors. Of this data, 9 classes were included: Insects, Trains, Gusting Wind, Mine / Quarry Noise, Birds, Road Noise, Barking Dogs, Sheep, Aircraft and Cattle. Data included in the class "Cattle" were ultimately omitted from experimentation due to very small sample size. This represents the totality of available ground-truth data at the time of manuscript preparation (4062 samples x 9-classes).



Source: (Burwood, 2025)

Figure 1: SPL distributions ($L_{Aeq,10seconds}$) and sample counts per class for user annotations for the training dataset

The distribution of SPLs and sample counts varied greatly between classes (see *Figure 1*). This skew also existed in the co-occurrence of sound sources (see *Figure 2*). This skew may have impacted results but was mitigated via weighted sampling. Growth in the ground-truth data set (in terms of the number of samples, variety of sources and occurrence of sources) is expected, and further experimentation will be reported as part of future work.



Source: (Burwood, 2025)
Figure 2: Cooccurrence of sources

2.2 Expert LAeq attribution protocol

Source-specific LAeq contributions were obtained through a manual attribution protocol conducted by three acoustic specialists. Each 10-second clip was assigned to exactly one acoustician—there was no overlapping annotation or inter-rater adjudication. Annotators had access to recorded audio, were informed of the environmental conditions at the time of the clip, and were aware of the plausible noise sources present in the deployment area. They were also provided with auxiliary visual context including mel-spectrograms and one-third octave spectra.

For each clip, the acoustician produced a per-class attribution in the form of a 10-second A-weighted LAeq value (in dBA) for each of the 9 predefined noise classes contained in the dataset. No mechanism was available for expressing uncertainty or confidence in individual attributions, and conflicting annotations did not arise due to the non-overlapping assignment strategy. The resulting LAeq values were min–max normalized and used directly as regression targets; no further filtering or reconciliation was applied. The lack of inter-rater variability analysis is acknowledged as a limitation and left for future work.

2.3 Model Architecture

The backbone model is based on MaxViT-L, adapted to operate on audio mel-spectrogram inputs, following the success of Max-AST in the ESC domain. Input spectrograms (as defined in Section 2.1) are treated analogously to image patches. The hierarchical architecture uses the following configuration: patch/window sizes of [(8,8), (8,8), (8,8), (8,4)] across stages; embedding dimensions of (96, 192, 384, 768); depths of (3, 3, 7, 3); and 32 attention heads with a head size of 512. These design choices preserve the nested convolution-transformer structure of MaxViT while scaling it appropriately for the spectral resolution of the audio inputs.

For self-supervised pretraining, a decoder head was attached to perform masked spectrogram reconstruction. The decoder consists of five ConvTranspose2d blocks, each with kernel size 4, stride 2, and padding 1, progressively upsampling to reconstruct the masked portions of the input Mel-spectrogram. For supervised regression, a separate head produces per-class SPL estimates—specifically, source-specific A-weighted LAeq predictions for each of the 9 noise classes.

2.4 Self-supervised pre-training

The model was pre-trained on over 600,000 unlabelled 10-second audio samples using a masked spectrogram reconstruction objective. Input Mel-spectrograms were randomly masked by sampling permutations with a masking ratio of 0.75; the model was tasked with reconstructing the original spectrogram over the masked regions. This step sought to promote model familiarity with Mel-spectrogram data prior to commencement of training on specific targets. The reconstruction loss was mean squared error (MSE) computed on the Mel-spectrogram magnitudes after min–max normalization. No additional data augmentations were applied during pretraining.

Optimization was performed with the Adam optimizer using a cosine annealing learning rate schedule. Pretraining was run for 100,000 steps, and the final checkpoint from the last epoch was selected for downstream fine-tuning.

2.5 Supervised regression fine-tuning

Fine-tuning was carried out on labelled data pairs consisting of the normalized Mel-spectrogram inputs and their corresponding expert-attributed source-specific A-weighted LAeq targets. The regression objective was the Huber loss with delta set to 1, chosen to balance robustness to outliers while maintaining sensitivity to small errors. To address the class imbalance among the 9 noise classes, we employed a weighted sampling strategy based on the empirical distribution of noise classes.

Fine-tuning used the Adam optimizer with a cosine decay schedule; private optimization hyperparameters are withheld for confidentiality. The dataset was split randomly into training, validation, and test sets with proportions of 70%, 15%, and 15%, respectively. No backbone layers were frozen during fine-tuning. Regularization included weight decay and layer decay, but no gradient clipping was applied. The model produced point estimates of per-class LAeq; no explicit uncertainty modelling (e.g., Monte Carlo dropout or ensembling) was incorporated during fine-tuning or inference.

2.6 Evaluation design

Model performance was evaluated on the held-out test set. The primary regression metric was Mean Absolute Error (MAE) of the predicted A-weighted LAeq versus the expert-provided ground truth, reported per noise class. Secondary evaluation employed an F1 score derived by framing detection as a binary problem: for each class, both the prediction and the ground truth were thresholded at 10 dBA (values ≥10 dBA considered “present”), and precision/recall/F1 were computed accordingly. This threshold was chosen arbitrarily but is below the minimum values observed in the dataset, effectively reflecting audibility.

Confidence bounds were constructed empirically from the distribution of absolute errors per class by computing multiple percentiles (90, 95, 99, 99.9, and 99.99). Thus, for example, the “95% confidence” bound corresponds to the 95th percentile of absolute error, i.e., the error value below which 95% of test-sample absolute errors fall. No formal ablation studies were performed; the reported results reflect the performance of the pre-trained MaxViT-L model fine-tuned as described. Additional analyses include class-wise error breakdowns to assess variability across noise types.

3 RESULTS

3.1 Error Statistics

Table 2 summarizes the primary performance metrics of the fine-tuned MaxViT-L model. The regression target was per-class A-weighted LAeq on 10-second clips. A secondary binarized detection task (presence if LAeq ≥ 10 dBA) yields the reported accuracy and F1 scores. The high accuracy score is skewed by the binarization of the classification task; targets generally have 1-3 sources, resulting in up to 11 classes being correctly identified as not being present in the noise source. As such, we have provided an F1 score, which provides a fairer representation of model precision. The Mean Absolute Error (MAE) is similarly skewed in this way. We also provide metrics for the Salient source, which is least impacted by the presence of other sounds. The difference between the Salient Accuracy and F1 score indicates that the model has greater difficulty identifying quieter sounds in the presence of louder sounds.

Table 2: Overall model performance on hold-out test set. "Salient" only considers the error and accuracy of the loudest sound present in the audio stream

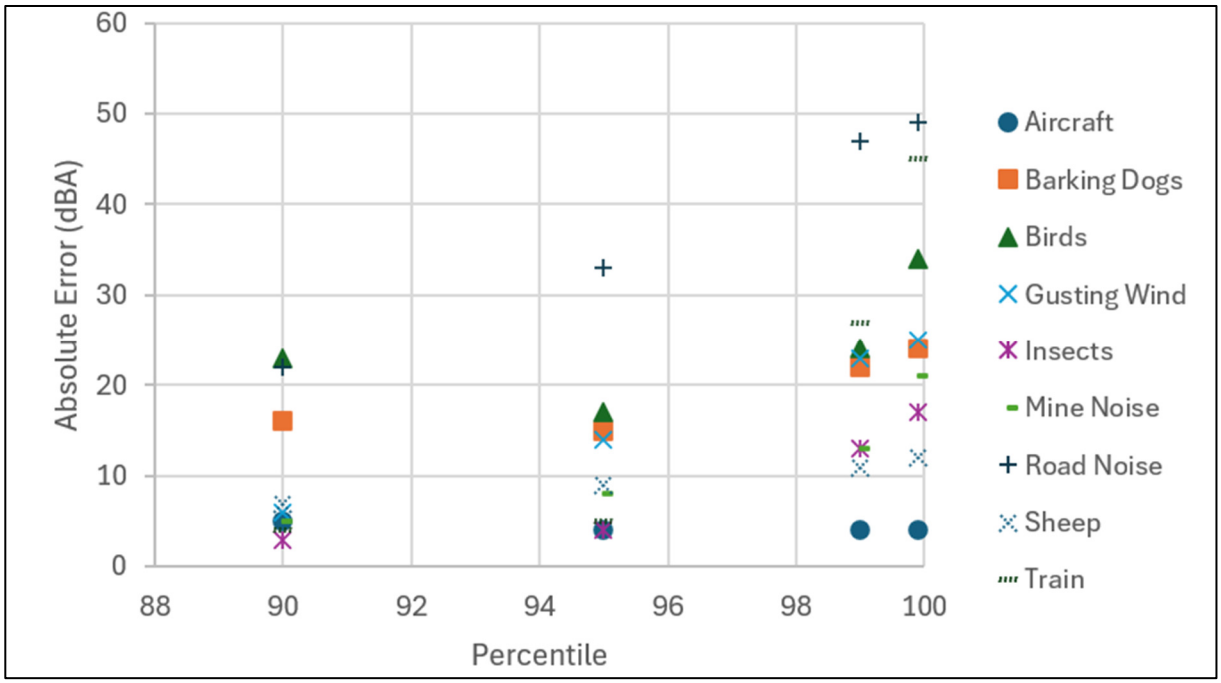
Metric	Value
Accuracy	99.19%
F1	0.86
MAE (dBA)	0.8
MAE Std (dBA)	2.1
Salient Accuracy	95.35
Salient MAE (dBA)	1.4
Salient Std (dBA	2.1

The Salient source MAE may be largely accounted for by the 95% accuracy. With the mean target value of 43 dBA and assuming a false classification output of 0 dBA, misclassifications of the Salient source should account for an MAE of at least 2.0 dBA. This suggests that the sounds being misclassified may have been correctly classified if a more optimal classification threshold was chosen, and perhaps the accuracies and F1 scores are higher than what have been reported. However, this would present a possible decrease in precision. Precision (as error tolerance) vs accuracy curves would be interesting, but we leave this for future work.

3.2 Class-Specific Confidence Intervals

We computed empirical error bounds per class by taking multiple percentiles (90, 95, 99, 99.9) of the absolute error distribution, with true negatives excluded which heavily skewed the results towards a MAE of 0 dB. This yields intuitive "confidence" statements; for example, the 95th percentile error is the value below which 95% of predictions fall.

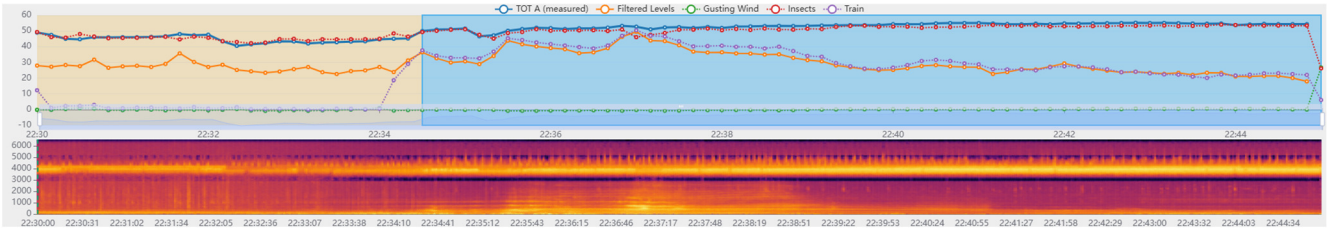
The sources of Sheep, Mine Noise, Train, and Insects exhibit relatively tight high-confidence error bounds, with 95th percentile absolute errors of 9.03, 7.00, 4.90, 3.32 dBA respectively, indicating stable reconstruction of their levels even in the presence of mixed sources. Other classes (Aircraft, Barking Dogs, Birds, Gusting Wind, Road Noise) show varying degrees of difficulty—potentially driven by their spectral overlap, environmental prevalence, or sample scarcity. Pivotally, the results indicate a relationship between sample size and percentile errors, and average LAeq's and percentile errors, highlighting the need for larger datasets with greater diversity of sound pressure level distributions.



Author: (Burwood, 2025)
Figure 3: Class-wise Absolute Errors vs percentile curves.

3.3 Example Implication

An example provided in *Figure 4* indicates a measurement data scenario with concurrence of 2 sound sources (insect sounds and the passage of a train). The figure exemplifies the concurrent use of two data modalities (precision measurement data (short-term LAeq) and spectrogram derived from recorded audio).



Author: (Burwood, 2025)

Figure 4: Example scenario with overlapping noise sources with a time-series plot of total LAeq alongside a band-limited (20-2000Hz) reference and model predictions. Spectrogram provided for reference.

In this scenario, the sources occupy different parts of the frequency domain, so can be reasonably differentiated using more traditional (i.e. bandpass) filtering. Expert ground-truthing indicates that insect noise contributions are well approximated by Total measured noise levels, while transient contributions from a passing train may be evaluated by a low-pass (≤630Hz) filter. In terms of model inference, the Mel-spectrogram is the input, and the output returned by the model are source-wise predictions of SPLs as 10-second LAeq. In this illustrative case, the model quantifies both the salient source (Insects) and the non-salient but concurrent source (Train) with reasonable fidelity.

The predicted per-class LAeq trajectories closely track the expert labels, and the decomposition of total noise demonstrates that the model can separate overlapping contributions that may otherwise require intensive manual analysis. This indicates the potential for using the model output to drive real-time source-aware monitoring dashboards, where both dominant and background contributions are important for regulatory or diagnostic decisions. We note that this is a trivial case, where both sources are reasonably separable in the time and frequency domains. However, the potential for use in scenarios with both time and frequency overlap is promising.

4 DISCUSSION

4.1 Model Performance and Reliability

The results demonstrate that a large hierarchical convolution-transformer backbone (MaxViT-L), when pretrained with a high masking ratio and fine-tuned on expert-attributed data, can deliver accurate and source-specific LAeq estimates in complex environmental sound mixtures. The combination of detection and regression objectives yields a model that not only reliably identifies the presence of noise sources but also quantifies their contribution with low error: the salient-source MAE of 1.4 dBA indicates that the dominant noise components are estimated with precision that is meaningful for environmental monitoring contexts.

4.2 Confidence Bounds and Data Dependence

Confidence bounds derived from empirical percentiles offer interpretable error envelopes for each class. Because these bounds are taken directly from the distribution of absolute errors (across the 90, 95, 99, 99.9, and 99.99 percentiles), stakeholders can make risk-aware decisions—for example, understanding that for certain classes like Insects or Train, 95% of predictions fall within a narrow error margin, whereas other classes exhibit wider tails. The observed variation in these bounds across classes correlates with inherent data characteristics: sources with more abundant training examples and lower average levels tended to have tighter error distributions, suggesting that both data quantity and signal salience materially influence reliability.

4.3 Source Decomposition in Overlapping Scenarios

An illustrative example with overlapping Train and Insect noise (*Figure 4*) underscores the model's practical capacity to decompose concurrent sources without explicit signal separation preprocessing. In this scenario, predicted LAeq trajectories for each source closely track expert ground truthing, demonstrating that the learned representations encapsulate both presence and level in a way that enables fine-grained, source-aware monitoring. Such capability could significantly improve the productivity of continuous and unattended monitoring, allowing analysis dashboards to autonomously surface both dominant and secondary contributors in real time.

4.4 Limitations

Despite these promising outcomes, several methodological limitations temper the generality of the conclusions. The expert annotation process assigned each clip to a single acoustician, preventing any assessment or mitigation of inter-rater variability; consequently, label noise and subjective bias remain unquantified. Environmental conditions such as low signal-to-noise situations or microphone degradation under wind introduce additional, unmodeled uncertainty in the ground truth that is not explicitly accounted for during training or evaluation. Furthermore, while empirical percentile-based confidence bounds provide practical reliability insights, the model does not incorporate formal uncertainty modelling mechanisms—such as Bayesian inference, ensembling, or predictive distributions—which would enable probabilistic calibration beyond observed error statistics.

The study also did not include ablation experiments, leaving the isolated contribution of self-supervised pretraining versus training from scratch unmeasured, nor were alternative architectures compared to understand architectural sensitivity. The weighted sampling strategy partly addresses class imbalance, but residual skew may still bias performance, particularly in the tails for underrepresented classes. Calibration between predicted and real-world level distributions was not performed, which could impact thresholded decision-making in regulatory contexts.

The limited size of the dataset (in terms of both source diversity and sample count) is also recognised as a constraint of this study. The model is likely to perform poorly on unseen sources (e.g., a chainsaw), but this would be a common limitation for any deep-learning based model exposed to unfamiliar inputs. The size of the training data set is comparable to contemporary studies in the Australian context (Halkon, et al., 2024), and expanding and balancing the ground-truthed training data will be a focus of future work.

4.5 Operational Implications

Despite these limitations, results indicate that the approach represents a viable path toward automation of context aware noise source quantification, and associated increases in productivity from continuous and unattended environmental noise monitoring. Low-error, source-specific LAeq estimates coupled with interpretable confidence bounds allow acoustic practitioners and regulators to prioritize attention, trigger alerts with quantified reliability, and present overlapping source contributions in interactive dashboards. The reliance on standard audio preprocessing (Mel-spectrograms) and the transferability of the model architecture make adaptation to similar rural or semi-rural deployments feasible with modest engineering effort.

4.6 Future Work

Looking ahead, addressing the identified limitations is a clear direction for future work. Incorporating overlapping annotations with inter-rater agreement analysis would help quantify and reduce label noise. Systematic ablation studies would clarify the benefits of pretraining and benchmark the architecture against lighter or alternative models. Enhancements to uncertainty quantification—through techniques like Monte Carlo dropout, deep ensembles, or heteroscedastic regression—would complement empirical bounds with richer probabilistic insights. Calibration methods could align model outputs more closely with operational thresholds, and domain adaptation or online updating would improve robustness in acoustically shifting environments. Together, these extensions would strengthen both the reliability and applicability of source-specific SPL estimation in real-world environmental noise management.

5 CONCLUSIONS

In this study, we demonstrated that deep learning-based methods can effectively estimate source-specific sound pressure levels (SPLs) in complex, real-world acoustic environments. The fine-tuned MaxViT model showed strong performance with low mean absolute error (MAE) and high accuracy in predicting SPLs for a range of noise classes. The model's ability to handle overlapping noise sources, providing precise contributions for both dominant and background sounds, suggests it is well-suited for use in operational environmental monitoring systems. Confidence intervals derived from empirical error distributions offer interpretable insights, helping stakeholders make informed, risk-aware decisions.

However, limitations such as the absence of inter-rater variability analysis in expert annotations, unaccounted environmental uncertainties, and the lack of formal uncertainty modelling prevent the framework from being fully operational in all settings. Despite these constraints, the approach holds significant promise for improving productivity of environmental noise analysis. Future work should focus on enhancing uncertainty quantification, conducting ablation studies, and improving the calibration of model outputs with real-world thresholds to further strengthen its applicability for regulatory use.

By addressing these limitations and further refining the framework, this study paves the way for scalable, automated noise assessment systems that can be integrated into regulatory and environmental management platforms to improve the accuracy, efficiency, and reliability of noise monitoring and compliance.

ACKNOWLEDGEMENTS
This work was completed at Advitech Pty Ltd. and funded by the company. The support provided by Advitech in terms of resources and funding was crucial in enabling the completion of this research.

The research builds upon the foundation laid in the Honours Thesis conducted at the University of Newcastle, which was supervised by Nasimul Noman. The guidance and expertise provided by Nasimul Noman were instrumental in shaping the direction of this work.

The contributions of the acoustic specialists involved in the expert annotation process are also gratefully acknowledged, as well as the valuable feedback and suggestions received from colleagues and peers throughout the course of the research.

REFERENCES

AS/NZS IEC 61672.1:2019 Electroacoustics: sound level meters – Part 1: Specifications

Alex, T., Ahmed, S., Mustafa, A., Awais, M., & Jackson, P. J. (2024). Max-AST: Combining Convolution, Local and Global Self-Attentions for Audio Event Classification. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1061-1065). Seoul: IEEE.

Bansal, A., & Garg, N. K. (2022). Environmental Sound Classification: A descriptive review of the literature. *Intelligent Systems with Applications, 16*, 200115.

Environment Protection Agency. (2022, 3 2). *Approved methods for the measurement and analysis of environmental noise in NSW*. Retrieved from NSW EPA: https://www.epa.nsw.gov.au/Your-environment/Noise/regulating-noise/approved-methods-for-measurement-and-analysis-of-environmental-noise

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*.

Halkon, B., Darroch, M., Cooper-Woolley, B., Zhao, S., Miller, A., Hanson, D., . . . Mifsud, S. (2024). Advancing AI-based Acoustic Classifiers for (Rail) Construction Noise – A Pilot Project. *Acoustics in the Sun.* Gold Coast: Australian Acoustical Society.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., . . . Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science, 382*(667), 1416-1421.

Nogales, A., Caracuel-Cayuela, J., & García-Tejedor, Á. J. (2024). Analyzing the Influence of Diverse Background Noises on Voice Transmission: A Deep Learning Approach to Noise Suppression. *AppliedSciences, 14*(2).

NSW Government. (2025). *eTendering - Transport ofr NSW (Transport Infrasctructure Projects*. Retrieved from buy NSW: https://buy.nsw.gov.au/notices/1FE26BB5-AB4E-AADE-AC768EF6DD491F33

NSW Government. (2025, March 2). *NSW legislation*. Retrieved from Protection of the Environment Operations Act 1997 No 156: https://legislation.nsw.gov.au/view/html/inforce/current/act-1997-156

Sparke, C. (2018). Environmental Noise Classification through Machine Learning. *Hear to Listen* (p. 85). Adeliade: Australian Acoustical Society.

Zainurin, S. N., Wan Ismail, W. Z., Mahamud, S. N., Ismail, I., Jamaludin, J., Ariffin, K. N., & Kamil, W. M. (2022). Advancements in Monitoring Water Quality Based on Various Sensing Methods: A Systematic Review. *International Journal of Environmental Research and Public Health, 19*, 14080.