25-29 August 2024
inter.noise
NANTES FRANCE

A comparative study of noise event identification using AI in unattended monitoring

Karl Henrik Ejdfors[1]
Norsonic AS
Gunnersbråtan 2, 3409 Tranby, Norway

Naru Sato[2]
Norsonic AS
Gunnersbråtan 2, 3409 Tranby, Norway

Lars Andreas Sæle[3]
Norconsult Norge AS
Vestfjordgaten 4, 1338 Sandvika, Norway

ABSTRACT

This paper explores the use of multi-microphone devices and artificial intelligence (AI) for identifying noise events in unattended noise monitoring. The primary focus is to assess the reliability of a machine learning model initially trained on a dataset representing a particular soundscape. We evaluate the performance of this model when applied to diverse datasets collected from similar yet distinct soundscapes, encompassing various environmental conditions and noise profiles. Through comparative analysis, we determine the model's adaptability and potential limitations. The findings of this study offer insights into how well AI-based noise event identification models can work in different situations. This lays the groundwork for enhancing their applicability in diverse real-world settings and improving how well unattended noise monitoring systems function.

1. INTRODUCTION

In recent years, machine learning (ML) and artificial intelligence (AI) have become increasingly popular, and have been applied to a wide range of fields. This popularity has led to extensive research and development, and it requires a large amount of labeled data for training ML models [1].

In domains such as unattended noise monitoring, acquiring vast amounts of unlabeled sound data is feasible. However, annotating this data can be excessively time-consuming. To address this challenge, a strategy can involve initially training a model on a large dataset from a different domain and subsequently fine-tuning it with a smaller annotated dataset from the target domain.

[1]kejdfors@norsonic.com

[2]n-satou@norsonic.com

[3]lars.andreas.saele@norconsult.com

This paper presents a comparative study of noise event identification utilizing AI in unattended sound monitoring. Specifically, we construct an ML model based on data from one unattended noise monitoring station and evaluate its performance on data from another monitoring station. We begin by providing relevant background theory, followed by outlining an approach for comparing the performance of various ML models. Subsequently, we present the results of our study and discuss their implications for enhancing unattended noise monitoring systems.

2. BACKGROUND

A soundscape describes the acoustic environment as perceived by humans within a given context. Each soundscape is unique, and is comprised of multiple sounds originating from various sources such as traffic, construction activities, natural elements like flowing rivers, birds, and human interactions. To differentiate between distinct auditory phenomena within a soundscape, the term "sound event" is employed. A sound event represents a perceivable entity characterized by its acoustic attributes, which can be further described by temporal, spectral, and spatial properties. Temporal characteristics encompass parameters such as duration and onset, spectral properties involve the frequency content, while spatial properties relate to the direction and distance of the sound source.

Despite the semantic similarity of sounds across different soundscapes, their acoustic characteristics often vary significantly. For instance, the sound of a car horn in an urban environment differs from its counterpart in a rural setting. In urban locales, car horn sounds are typically embedded within other urban noises, whereas in rural settings, car horns resonate more distinctly and directly. These divergent acoustic profiles illustrate how environmental context influences the perception of sound [2].

Humans possess an ability to distinguish between different sounds and discern their sources effortlessly. Whether identifying the material composition of a falling object or recognizing the nature of a distant sound, humans excel in sound recognition [3]. However, for a machine, this is a difficult task. Machine-based sound recognition requires training algorithms to discern between various sound classes using ML techniques. In a supervised learning approach, the system undergoes training on labeled datasets, and learns to recognize the different sounds. The approach requires a set of classes describing the sounds, defined by a system developer, and a sufficient amount of labeled data for each class. Subsequently, the system's performance is assessed using separate labeled datasets, resulting in an evaluation of its classification. Once trained, the system can classify unlabeled data within a given soundscape [2].

When evaluating the performance of a ML model, one can employ various metrics to assess its classification accuracy. Common metrics include accuracy, precision, recall, and F1-score. Accuracy measures the proportion of correctly classified instances among all instances, following the formula:

$$\text{Accuracy} = \frac{\text{No. of correct predictions}}{\text{Total no. of predictions}} \tag{1}$$

Precision measures the proportion of true positive predictions out of all positive predictions, following the formula:

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \tag{2}$$

Recall, on the other hand, measures the proportion of correctly classified instances among all instances that are actually positive, following the formula:

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \tag{3}$$

The F1-score is the harmonic mean of precision and recall, and provides a single metric to evaluate the model's performance, following the formula:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

In addition to these metrics, one can also use the confusion matrix to visualize the model's performance. The confusion matrix displays the number of true positives, true negatives, false positives, and false negatives, and provides a comprehensive overview of the model's classification accuracy.

When designing a classification system, also called a taxonomy, it's important to account for both the complexity of the soundscape and the range of sound events. A well-designed taxonomy should be clear and comprehensive, and it should balance the granularity of the classes to ensure that each category is distinct and meaningful within the context of the soundscape [4]. It should be designed to accommodate a sufficient number of instances for each class to enable effective model training and evaluation. Furthermore, the taxonomy should be flexible and adaptable to accommodate new classes or categories as needed.

3. METHOD AND RESULTS

Construction activities are significant contributors to urban noise pollution, often causing annoyance for both residents and workers. To assess the performance of a ML model trained on data from one domain and applied to another, we collected data from two distinct construction sites. The first site, referred to as "Location A", pertains to the construction of a new hospital in Drammen, Norway. Meanwhile, the second site, referred to as "Location B", involves the renewal of an existing hospital in Oslo, Norway. Despite both sites being situated within complex urban soundscapes, they exhibit different sound profiles attributed to various sources. However, both datasets are annotated using the same taxonomy, facilitating direct comparison.

This section outlines the methodology employed for conducting the study and presents the results. Firstly, we introduce the taxonomy utilized for annotating the collected data. Then, we present the performance results of ML models trained independently on datasets from Location A and Location B. Following this, we detail the outcomes of fine-tuning the ML model trained on Location A with a smaller annotated dataset from Location B. Lastly, we present the findings of the comparative study, which evaluates the efficacy of the ML model across both construction sites.

3.1. Taxonomy

To ensure consistency in evaluating the performance of the ML model across both datasets, we employ a standardized taxonomy for annotation. This taxonomy describes various sound sources that are common for construction sites and is defined by the system developer. Table 1 presents the taxonomy used for annotation, categorizing sound events into distinct classes based on their origin and characteristics.

The taxonomy encompasses a comprehensive range of sound sources typically encountered in construction environments. However, it's important to acknowledge instances where certain sounds may not neatly fit into predefined categories. In dynamic environments like construction sites, ambient noise levels can fluctuate significantly, often resulting in the emergence of sounds that defy easy classification.

3.2. Location A: Drammen hospital construction site

The dataset from Location A serves as the primary training dataset for the ML model employed in our analysis. Collected over a span from 2019 to 2023, the data originates from an unattended noise monitoring system (Norsonic Nor1545 with Noise Compass). Situated within an urban environment, the construction site presents a complex soundscape characterized by diverse

Table 1: Taxonomy for construction noise.

Category	**Description**
Rattling	Sounds such as piling
Circ. saw	Cutting materials with a circular saw
Dump crush	Dumping or crushing materials
Shovel	Shovel digging
Metal	metal dropping
Rev. signal	Beeping noise when a truck reverses
Helicopter	Helicopters flying over the monitoring station

sound sources, including nearby railway and highway traffic, as well as industrial activities from neighboring facilities. For visual context, Figure 1 illustrates the layout of the construction site in Drammen, Norway.

Figure 1: Location A, the construction site for the new hospital in Drammen, Norway [5].

The data utilized to train the ML model was randomly sampled from the entire measurement period at Location A. Each data sample was annotated to align with a taxonomy comprising seven distinct classes, as depicted in Figure 2. To ensure the robustness of our training dataset, recordings were included only if they met a minimum threshold of 35 annotated instances per class. This criterion serves to maintain a balanced representation of sound events across all categories.

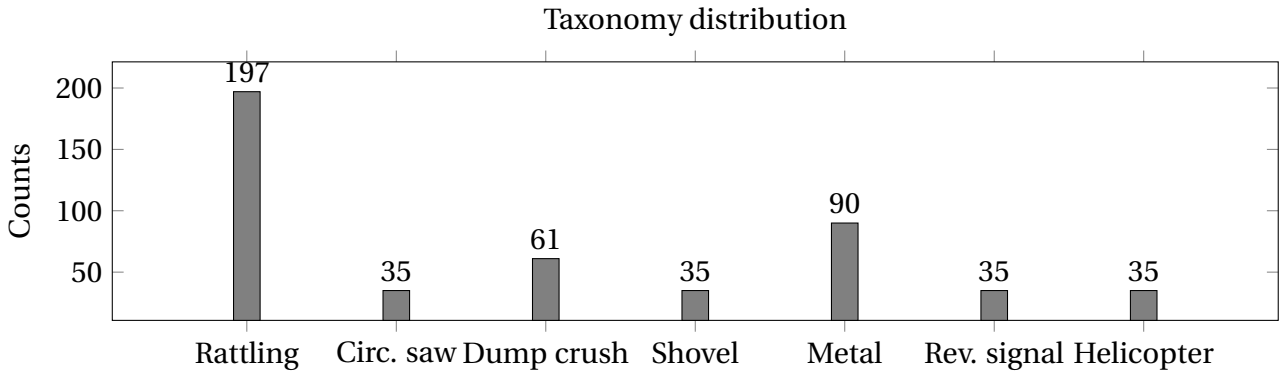

Figure 2: Location A taxonomy distribuiton.

Upon evaluation on the test dataset from Location A, the resulting ML model attained performance metrics summarized in Table 2. Additionally, the confusion matrix, providing a breakdown of classification outcomes, is presented in Table 3.

Table 2: Performance metrics for Location A model.

Metric	Accuracy	Precision	Recall	F1-score
Value	0.99	0.95	0.96	0.95

Table 3: Location A confusion matrix.

		Predicted						
		Rattling	Circ. saw	Dump crush	Shovel	Metal	Rev. signal	Helicopter
Actual	Rattling	35	0	0	0	0	0	0
	Circ. saw	0	2	0	0	0	0	0
	Dump crush	0	0	10	0	0	0	0
	Shovel	0	0	0	2	0	0	0
	Metal	0	0	2	0	14	0	0
	Rev. signal	0	1	0	0	0	3	0
	Helicopter	0	0	0	0	0	0	2

3.3. Location B: Oslo hospital construction site

The test dataset obtained from Location B was acquired using the same unattended noise monitoring system utilized at Location A, spanning a duration of three days. Situated within a comparable yet different urban environment, the construction site at Location B features distinct sound sources including a nearby tram station, a helipad, and proximity to a highway, as depicted in Figure 3.

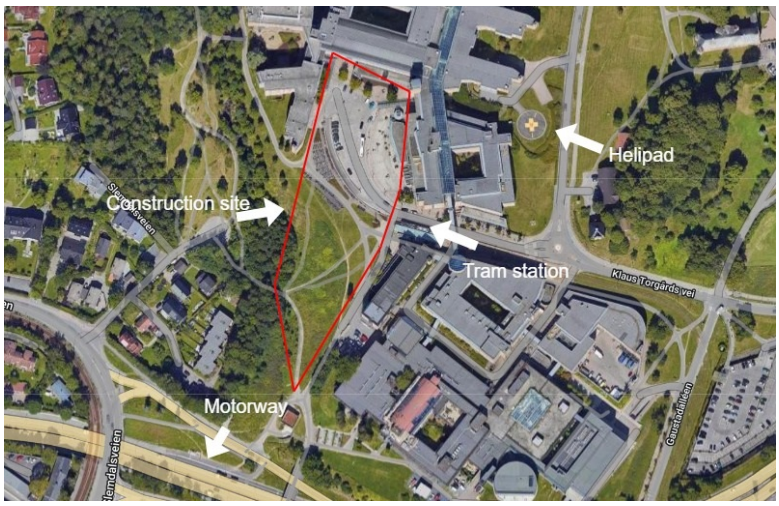

Figure 3: Location B, the construction site for the new hospital in Oslo, Norway.

The test dataset was selected from a specific period of construction activity spanning the three-day measurement period. This selection process ensures that the dataset captures a

representative sample of sound events encountered during active construction work at Location B.

Furthermore, the class distribution within the training dataset is illustrated in Figure 4. To maintain consistency and robustness, only recordings containing a minimum of 35 annotated instances per class were included in the training set.

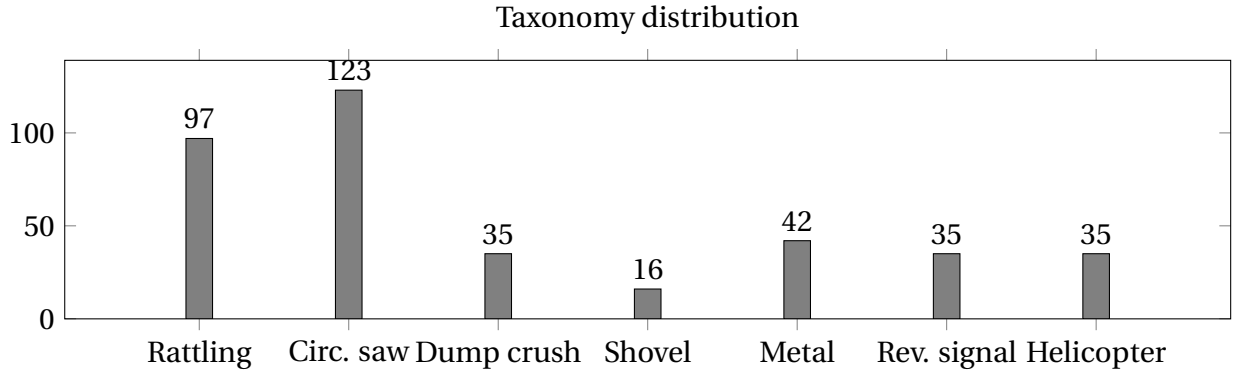

Figure 4: Location B taxonomy distribuiton.

Upon evaluation on the test dataset from Location B, the resulting ML model attained performance metrics summarized in Table 4. Additionally, the confusion matrix is presented in Table 5.

Table 4: Performance metrics for Location B model.

Metric	Accuracy	Precision	Recall	F1-score
Value	0.93	0.81	0.85	0.82

Table 5: Location B confusion matrix.

		Predicted					
		Rattling	Circ. saw	Dump crush	Metal	Rev. signal	Helicopter
Actual	Rattling	27	0	0	0	0	0
	Circ. saw	6	29	0	0	0	0
	Dump crush	3	0	5	0	0	0
	Metal	2	0	0	4	0	0
	Rev. signal	1	0	0	0	0	0
	Helicopter	0	0	0	0	0	6

3.4. Fine-tuning the model

When evaluating the ML model trained on Location A using the dataset from Location B, it was observed that the model's performance did not match its performance on the original dataset. The performance metrics are detailed in Table 6, with a comprehensive breakdown provided in the confusion matrix in Table 7.

To enhance the model's performance on test data from Location B, we fine-tuned the model using a subset of annotated data from Location B. This process involved retraining the model on

Table 6: Performance metrics for Location A model tested on Location B dataset.

Metric	Accuracy	Precision	Recall	F1-score
Value	0.81	0.33	0.44	0.31

Table 7: Confusion matrix for Location A model tested on Location B dataset.

		Predicted						
		Rattling	Circ. saw	Dump crush	Shovel	Metal	Rev. signal	Helicopter
Actual	Rattling	51	101	0	23	0	19	1
	Circ. saw	26	193	0	3	0	11	0
	Dump crush	21	12	0	21	0	2	0
	Shovel	0	0	0	10	0	0	0
	Metal	26	8	0	43	0	7	0
	Rev. signal	5	1	0	1	0	20	0
	Helicopter	17	1	0	1	0	3	17

the dataset to adapt its parameters to the new environment. The class distribution of the fine-tuned model is illustrated in Figure 5, showing the inclusion of data from both locations.

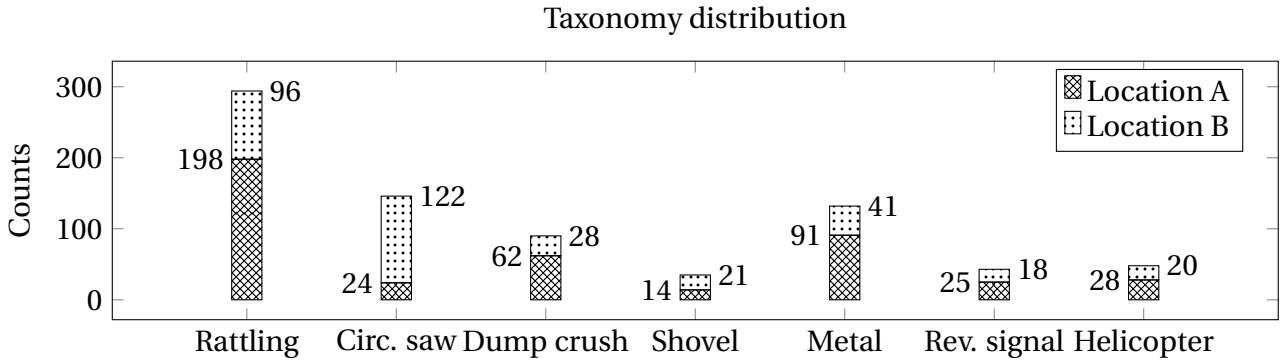

Figure 5: Combined model from Location A and Location B.

Upon evaluation on the remaining test dataset from Location B, the fine-tuned ML model achieved performance metrics summarized in Table 8. Additionally, the confusion matrix is presented in Table 9.

Table 8: Performance metrics for fine-tuned model based on datset from Location A and B, tested on Location B dataset

Metric	Accuracy	Precision	Recall	F1-score
Value	0.95	0.86	0.80	0.82

In the following section, we will look further into the results of our study and explore their implications.

Table 9: Confusion matrix for Location A model tested on Location B dataset

		Predicted						
		Rattling	Circ. saw	Dump crush	Shovel	Metal	Rev. signal	Helicopter
Actual	Rattling	86	2	0	5	4	0	1
	Circ. saw	8	111	0	0	1	2	0
	Dump crush	1	2	19	1	5	0	0
	Shovel	0	0	0	8	2	0	0
	Metal	4	0	4	5	29	0	0
	Rev. signal	3	1	1	1	3	7	1
	Helicopter	5	0	0	0	0	1	16

4. DISCUSSION

In this chapter, we delve into the findings from our study, examining the performance of the ML models trained and fine-tuned on datasets from Location A and Location B. We also explore the implications of our results, particularly regarding the need for additional annotated data from Location B and the challenges posed by the complex urban soundscape.

Our evaluation of the ML models used performance metrics including accuracy, precision, recall, and F1-score. Upon analysis, the ML model trained on the dataset from Location A showed high performance on its original dataset, achieving an accuracy of 0.99 and an F1-score of 0.95, indicating its efficacy in classifying sound events within that environment. However, when tested on the dataset from Location B, the model's performance notably declined, to a precision of 0.33 and a recall of 0.44. These metrics underscore the model's struggle with false positives and false negatives. This struggle could be a result of the model's overfitting to the specific characteristics of Location A. Consequently, the model shows poor generalization when applied to a different environment. This emphasizes the importance of considering the unique characteristics of each environment during both training and evaluation phases.

The ML model trained on the dataset from Location B showed lower performance on its original dataset compared to the fine-tuned model. While achieving an accuracy of 0.93 and an F1-score of 0.82, this model's performance improved following fine-tuning, resulting in an accuracy of 0.95 and an F1-score of 0.82. While these improvements are not groundbreaking, they underscore the utility of fine-tuning, especially when dealing with limited annotated data. This finding suggests that fine-tuning serves as a strategy for enhancing model performance under such constraints.

Annotating sound data into a set of predefined classes presented several challenges, primarily due to the complexity of the urban soundscape. Many recordings contained multiple sound sources, making it difficult to attribute specific events to individual classes accurately. This variability in the soundscape contributed to instances of misclassification by the fine-tuned model.

These challenges underscored the necessity for fine-tuning with data from Location B to achieve satisfactory model performance. Despite similarities between construction sites at Locations A and B, the distinct sound profiles and environmental factors at Location B requires a tailored approach to model training. The subset selected for fine-tuning was guided by two primary criteria: firstly, the availability of data from Location B, which was constrained by measurement duration, and secondly, the objective to maintain a balanced representation of each category within the taxonomy across both the training and fine-tuning datasets.

To apply a ML model for a new monitoring site, one must consider the soundscape's unique characteristics and the availability of annotated data. The results of our study suggest that a model trained on data from one location may not perform optimally when applied to another location. However, by annotating a smaller subset of data from the new location and fine-tuning the model, one can achieve improved performance. This approach is particularly relevant in unattended monitoring scenarios where acquiring big amounts of annotated data is challenging.

5. CONCLUSION

We conducted a study to evaluate the performance of ML models trained and fine-tuned using datasets collected from two different construction site locations: Location A and Location B. Our evaluation utilized performance metrics, including accuracy, precision, recall, and F1-score, to measure the models' performance.

The ML model trained on data from Location A achieved high performance on its original dataset, achieving an accuracy of 0.99 and an F1-score of 0.95. However, when subjected to testing on the dataset from Location B, the model's performance notably declined, with precision dropping to 0.33 and recall to 0.44. This difference underscores the importance of accounting for the characteristics of each environment during model development and evaluation.

Despite the similarities between the construction sites at Locations A and B, the distinct sound profiles and environmental factors at Location B required a customized approach to model training. Through a fine-tuning process, which involved retraining the model using a subset of annotated data from Location B, yielded significant improvements in the model's performance. This outcome highlights the utility of fine-tuning methodologies, particularly in scenarios where annotated data is limited.

ACKNOWLEDGEMENTS

We would like to express our gratitude to Norconsult Norge AS and Oslo University Hospital for their cooperation in facilitating the construction site for our study.

REFERENCES

1. Shikun Zhang, Omid Jafari, and Parth Nagarkar. A survey on machine learning techniques for auto labeling of video, audio, and text data. *arXiv preprint arXiv:2109.03784*, 2021.
2. Toni Heittola, Emre Çakır, and Tuomas Virtanen. The machine learning approach for analysis of sound scenes and events. *Computational Analysis of Sound Scenes and Events*, pages 13–40, 2018.
3. Robert A Lutfi. Human sound source identification. In *Auditory perception of sound sources*, pages 13–42. Springer, 2008.
4. Tuomas Virtanen, Mark D Plumbley, and Dan Ellis. *Computational analysis of sound scenes and events*. Springer, 2018.
5. Daniela Toledo Helboe and Erlend Fasting. Automatic detection of source direction and exclusion of irrelevant sounds in unattended noise monitoring systems. *Proceedings of INTER-NOISE 23*, pages 1131–1142, August 2023.