

Soroll-IA Temporal: A Temporally-Structured Audio Dataset for Real-World Industrial Port Monitoring

Javier Naranjo-Alcazar*, Jordi Grau-Haro*, Annamaria Mesaros[†], Tuomas Virtanen[†], Pedro Zuccarello*

*Instituto Tecnológico de Informatica (ITI), Valencia, Spain

[†]Audio Research Group, Tampere University, Tampere, Finland

Abstract—Current sound event classification benchmarks commonly rely on random splits or do not provide chronological metadata about when the audio was recorded. In operational monitoring scenarios, such protocols may introduce temporal data leakage by mixing segments from the same recording sessions across training and testing splits, artificially overestimating model performance. To address these limitations, we introduce and publicly release *Soroll-IA Temporal*, a curated subset of the Soroll-IA corpus specifically designed to validate acoustic models under realistic industrial port monitoring constraints. The dataset comprises 6,139 audio segments structurally organized into five sequential, chronological folds with no temporal overlap. Crucially, following the original Soroll-IA dataset, this dataset is released with two distinct ground truth targets to account for label consensus variations: Non-CV, which includes any annotated label, and CV, which requires a strict two-thirds annotator consensus. Unlike conventional class-wise random partitioning, this strict chronological split explicitly captures non-stationary distribution shifts, evolving background soundscapes, and the dynamic sampling gaps inherent to active learning query strategies. To establish a transparent baseline, a vanilla PANNs-based CNN14 classifier was trained from scratch, achieving a conservative yet realistic macro F1-score of 0.531 ± 0.032 under the Non-CV setup, which drops to 0.498 ± 0.033 under CV. These results reflect the impact of evaluating models under a non-overlapping partition designed to emulate actual field deployment conditions. Consequently, *Soroll-IA Temporal* provides a structured framework to evaluate model behavior under chronological constraints, reflecting the operational challenges of real-world deployment.

Index Terms—Dataset, Audio Tagging

I. INTRODUCTION

Environmental sound classification and audio event recognition have experienced significant progress in recent years due to the availability of large-scale datasets and advances in deep learning architectures [1], [2]. Public benchmarks such as AudioSet [3] and FSD50K [4] have enabled the development of robust audio tagging [5], [6] and sound event detection [7], [8] systems across diverse acoustic scenarios. In parallel, audio-based monitoring has become increasingly relevant in industrial and urban environments [9], where passive acoustic sensing provides a non-invasive mechanism for detecting

operational events, anomalous activities, and environmental conditions [10], [11].

Despite these advancements, most existing audio classification benchmarks rely on random train/test partitioning strategies where the handling of temporal dependencies remains ambiguous or entirely unaddressed [12]–[14]. In practice, recordings acquired close in time often share similar acoustic conditions, background characteristics, and contextual patterns [15], [16]. Randomly splitting audio segments by class introduces temporal leakage of localized environmental and background acoustics between training and evaluation subsets. This data overlap introduces a selection and memorization bias [17], leading to overly optimistic performance estimates and reduced reproducibility under real-world deployment conditions [18]. This oversight is critical because machine learning models deployed in dynamic environments suffer severe performance degradation when evaluated across long-term chronological horizons [19], [20]. While standard evaluation practices assume that data distributions remain stationary, real-world deployment challenges models with macro-temporal domain shifts, such as seasonal variations and evolving background acoustics.

The consequences of ignoring this temporal dimension become particularly severe in highly dynamic deployment sites. Ports [21] constitute exceptionally challenging acoustic ecosystems due to the continuous coexistence of heterogeneous sound sources that evolve according to non-stationary operational schedules, weather variations, and shifting maintenance cycles. Deploying audio event classification systems in these settings demands architectures that can generalize under severe distribution shifts over time [7], [22], [23]. Consequently, evaluating models via standard random splitting in such environments masks their failure to adapt to future chronological periods.

To address this methodological bottleneck within a highly demanding deployment scenario, we leverage the original Soroll-IA dataset [21]. Initially introduced to provide the research community with a public, real-world repository of industrial port audio, Soroll-IA was designed to foster the development of robust sound event classification systems capable of handling complex maritime soundscapes. Building upon this foundational resource, we present Soroll-IA Temporal, a specialized subset that shifts the focus toward practical deployment readiness. By introducing strict temporal structuring, this new benchmark explicitly accounts for chronological

This work was partially funded by the Valencian Institute for Business Competitiveness (IVACE). This work has been supported by Grants PID2025-174060OB-C31 and PID2022-137048OA-C44 funded by MICIU/AEI/10.13039/501100011033 and by “ERDF/EU”. The work carried out by Javier Naranjo-Alcazar was also partially funded by the *Ministerio de Ciencia, Innovación y Universidades* (Spain) through the José Castillejo mobility grant for young doctors 2024 (CAS24/00150), supporting a research stay at a foreign higher education and research institution.

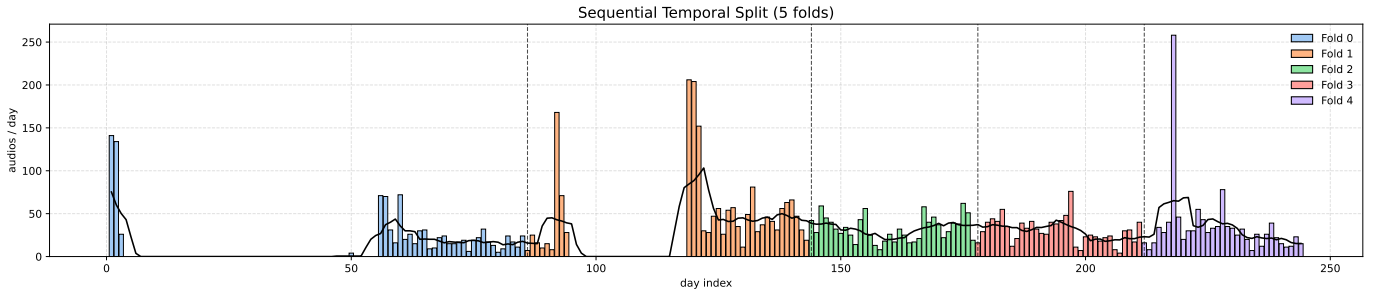


Fig. 1. Number of labeled audio files per day; gaps represent days unselected by the Active Learning loop. The black line denotes the 7-day rolling average.

variability, enabling a leakage-free evaluation protocol that accurately replicates long-term operational conditions.

The proposed dataset preserves temporal continuity across recordings and organizes data into sequential folds derived from recording dates. This enables evaluation under realistic temporal conditions, where training and test data belong to distinct chronological periods. This approach provides a more reliable prediction of the system’s performance once deployed, better reflecting practical scenarios in long-term passive acoustic monitoring.

II. SOROLL-IA

Soroll-IA is a weakly labeled environmental audio dataset designed for real-world industrial port monitoring. The dataset was acquired in two operational areas of the Port of Valencia. The corpus was originally introduced in [21]¹.

The dataset comprises various sound events that capture highly heterogeneous conditions characteristic of industrial maritime environments, including road traffic, heavy machinery, alarms, metallic impacts, and biological sounds. The annotation process was conducted by five independent human labelers; the complete active learning (AL) querying strategy, annotation consensus, and the subsequent cross-validation validation pipeline are detailed in our previous work [24].

The original dataset contains 7,396 audio recordings. Audio clips were acquired using autonomous sensing nodes distributed throughout the port infrastructure, enabling large-scale passive acoustic monitoring under real operational conditions [25]. The Soroll-IA dataset provides a balanced partitioning strategy suitable for the development and evaluation of audio classification systems in industrial port environments. Specifically, the original dataset is configured into a 5-fold cross-validation structure, where the splitting was performed directly at the level of the individual 10-second audio segments using a class-wise balancing strategy. Its primary objective was to facilitate research on the recognition of acoustic events under realistic weak-label conditions, demonstrating the capability of constructing robust audio tagging frameworks. In this context, the original dataset organization was designed to ensure representative class distributions and stable benchmarking conditions for supervised learning approaches.

¹<https://www.kaggle.com/datasets/itiresearch/soroll-ia-weakly-labeled-audio-port-monitoring>

III. SOROLL-IA TEMPORAL

In contrast with Soroll-IA, Soroll-IA Temporal addresses a different and more deployment-oriented research scenario. Rather than focusing exclusively on class-balanced evaluation, the proposed temporal configuration aims to emulate realistic operational conditions encountered in long-term acoustic monitoring systems. In practical deployments, computer audition systems are typically trained using recordings acquired during a given monitoring campaign and subsequently evaluated on audio captured during future time periods under potentially evolving environmental conditions.

To support this research direction, Soroll-IA Temporal introduces a temporally-structured organization of the dataset based on recording chronology. Recordings are partitioned into sequential temporal folds derived from acquisition dates, enabling the study of temporal robustness, chronological generalization, and realistic deployment behavior in industrial acoustic monitoring systems (see Fig. 1). The dataset is released (as Soroll-IA) with two distinct ground truth targets to account for label consensus variations: Non-CV, which includes any annotated label, and CV, which requires a strict two-thirds annotator consensus.

A key contribution of Soroll-IA Temporal is the incorporation of additional temporal metadata associated with each recording. Specifically, each audio sample is assigned an anonymized temporal index ranging from the first to the last available recording day. This representation preserves the chronological structure of the monitoring campaign while avoiding the disclosure of absolute recording dates. The proposed metadata facilitates temporal analysis, sequential partitioning strategies, and reproducible evaluation protocols for future research.

Furthermore, Soroll-IA Temporal corresponds to recordings acquired from a single sensing node within the original Soroll-IA infrastructure. This node was selected because it contained the largest amount of available audio data and the most temporally continuous acquisition history among the deployed recording devices. Consequently, the resulting subset provides sufficient temporal coverage to support meaningful chronological evaluation and long-term monitoring analysis.

To facilitate reproducible research and benchmarking, we officially release the dataset along with a pre-defined 5-fold split. We constructed these folds by aggregating all days

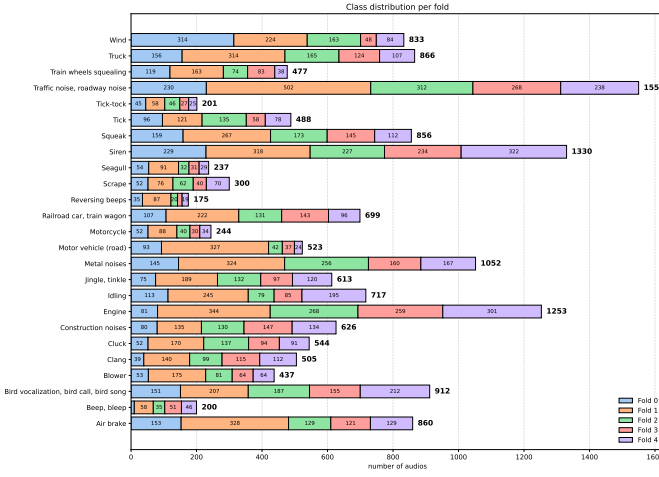


Fig. 2. Detailed sample distribution per acoustic class across the 5-fold split. The stacked bars illustrate the real-world class imbalance and the non-uniform distribution across folds resulting from our leak-free chronological partitioning.

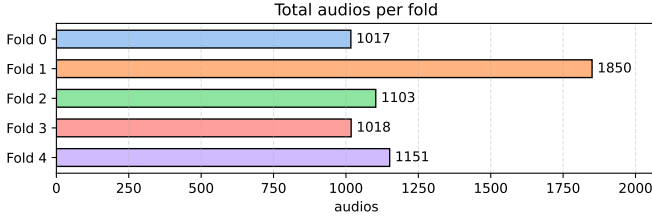


Fig. 3. Total audio samples allocated per validation fold ($N = 6, 139$). The size asymmetry (e.g., Fold 1) reflects the temporal-level grouping constraints applied to ensure rigorous evaluation.

containing annotated audio segments and partitioning them chronologically into five sequential blocks, each comprising an equal number of recording days. The dataset comprises 25 distinct classes of sound events. Fig. 2 illustrates the comprehensive distribution of audio labels, representing the total count of 10-second audio files associated with each specific label across the five folds in the Non-CV ground-truth. As shown in the distribution profile (see Fig 2), the dataset deliberately reflects a real-world long-tailed distribution with a pronounced class imbalance. The sample volume ranges from highly prevalent categories—such as Traffic noise, roadway noise ($N = 1,550$), Siren ($N = 1,330$), and Engine ($N = 1,253$)—to rarer, sparse acoustic events like Reversing beeps ($N = 175$) and Beep, bleep ($N = 200$).

Consequently, this temporal-based splitting strategy leads to an uneven distribution of total audio segments across the folds, as illustrated in Fig. 3. While Folds 0, 2, 3, and 4 maintain a relatively uniform size (ranging from 1,017 to 1,151 samples), Fold 1 contains a significantly larger subset of 1,850 samples. The annotations were collected via an iterative active learning (AL) pipeline that selectively queried informative acoustic segments [24]. Because the cross-validation folds cover completely disjoint time frames, this AL-driven sample

concentration introduces a severe data imbalance across loops, directly testing the model’s capacity to generalize under drastic shifts in data volume and acoustic context.

To ensure reproducibility and support future benchmarking, the dataset is publicly available in this link². The audio files follow a strict chronological nomenclature structured as `day_a_b.flac`, where *a* represents the specific recording day index and *b* denotes the chronological sequential order of the segment within that day. For instance, `day_0001_0001.flac` corresponds to the first audio segment recorded on day 1, while `day_0001_0002.flac` identifies the immediately subsequent segment from the same session. The day indices (*a*) map directly to the timeline presented in Fig. 1, allowing researchers to easily identify and analyze the temporal gaps between the active recording days included in the final subset.

IV. BENCHMARK

A. Baseline Model

To establish a rigorous performance benchmark, we implemented a baseline classifier using the CNN14 architecture from PANNs [5], [26]–[28]. To evaluate both the raw learning capacity of the architecture on our specific data distribution and the transferability of general acoustic features, the network was trained under two paradigms across the 5 folds: from scratch (serving as the baseline) and fine-tuned using pre-trained AudioSet weights [3]. Audio files were standardized to a sampling rate of 32,000 Hz to match the native specifications of the PANNs framework. All other hyperparameters, architectural configurations (e.g., mel-spectrogram features, window size, and hop length), and optimization settings were kept at their default values³. To provide an unmanipulated baseline for future benchmarks, no class-balancing or augmentation techniques—such as oversampling, undersampling, mixup or class-weighted loss functions—were applied. This choice ensures the baseline reflects the model’s vanilla performance on the dataset’s inherent long-tailed distribution, serving as a realistic reference point for future studies exploring imbalance-rectification strategies.

B. Performance Evaluation of Non-CV vs. CV Ground Truth Configurations

The baseline evaluation of the CNN14 architecture across the five temporal folds reveals the operational challenges inherent to real-world industrial environments. As summarized in Table I, the model achieves a global mAP of 0.590 ± 0.036 and a macro F1-score of 0.531 ± 0.032 under the Non-CV setup, demonstrating moderate but realistic overall performance. Under the CV condition, the metrics exhibit slightly lower absolute values, with a mAP of 0.557 ± 0.039 and a macro F1-score of 0.498 ± 0.033 . Rather than indicating a drop in model capability, this slight variance suggests that the

²<https://www.kaggle.com/datasets/tiresearch/soroll-ia-temporal/>

³https://github.com/qiuqiangkong/audioset_tagging_cnn/

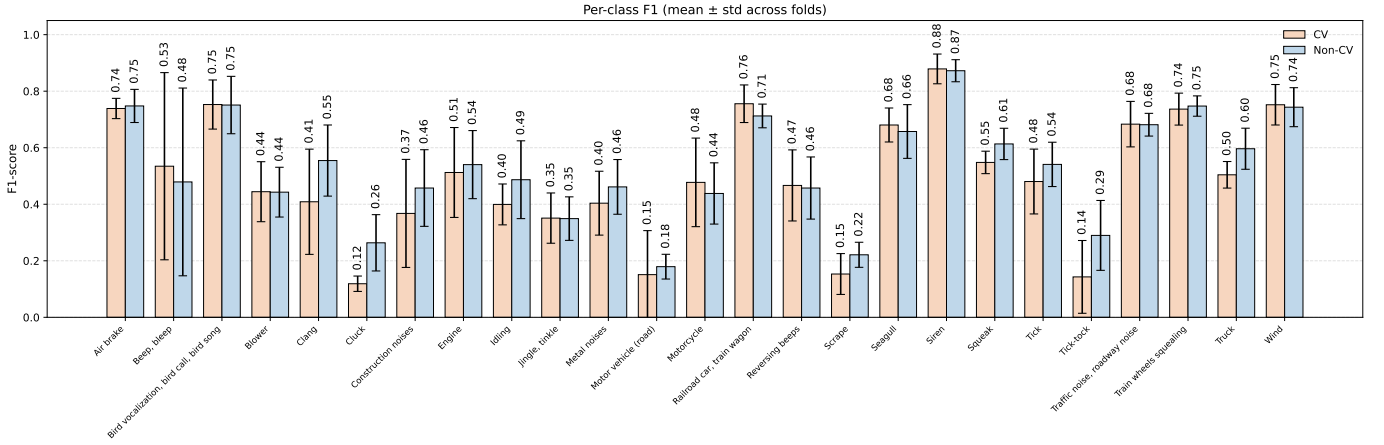


Fig. 4. Per-class F1-score comparison between Non-CV and CV ground truth configurations. Bars indicate the mean performance across the five temporal folds, while error bars denote the standard deviation ($\pm$ SD), highlighting the high performance volatility and campaign-dependent variance.

TABLE I
GLOBAL PERFORMANCE COMPARISON ACROSS GROUND TRUTH CONFIGURATIONS AND DATA SPLITTING STRATEGIES (MEAN $\pm$ STD OVER 5 FOLDS).

Ground Truth	Split Protocol	mAP	Macro-F1
Non-CV	Temporal	0.590 ± 0.036	0.531 ± 0.032
	Random	0.638 ± 0.009	0.581 ± 0.007
CV	Temporal	0.557 ± 0.039	0.498 ± 0.033
	Random	0.608 ± 0.009	0.560 ± 0.013

CV ground truth offers a more stringent and faithful representation of operational performance by incorporating label consensus filters. It is worth noting that these numerical results cannot be directly compared to the performance reported on the original Soroll-IA dataset [21], as *Soroll-IA Temporal* represents a smaller, chronologically isolated subset focused on a single deployment location. However, while absolute metrics differ, this setup isolates the temporal axis as the sole independent variable, allowing for a strict analysis of how a purely chronological partitioning alters baseline model stability under non-stationary shifts.

The standard deviation across folds is highly class-dependent, revealing performance variations in the per-class breakdown (Fig. 4) that are masked by the global macro averages. Specific categories show a distinct sensitivity to the evaluated split; for instance, sporadic or event-driven classes like *Beep, bleep* (Non-CV: 0.48 ± 0.33 , CV: 0.53 ± 0.33) and *Motorcycle* (Non-CV: 0.44 ± 0.11 , CV: 0.48 ± 0.15) display wide error bars, confirming that their presence and acoustic properties vary across temporal folds. The per-class variance in this temporal subset is higher than that reported in the original dataset; categories such as *Reversing beeps* show a higher inter-fold dispersion here, which reflects the structural impact of chronological splitting. Furthermore, these results differ from the trend observed in our previous work, where Non-CV annotations tend to yield lower standard deviations than the

CV consensus in the majority of categories. In this subset, the choice of ground truth does not govern the variance. Instead, each class behaves independently, indicating that the physical content of each temporal window introduces a sensitivity that outweighs the influence of the labeling framework. Conversely, pervasive categories such as *Siren* (Non-CV: 0.87 ± 0.04 , CV: 0.88 ± 0.05) and *Wind* (0.74 ± 0.07 vs. 0.75 ± 0.07) remain consistent across splits, showing that continuous environmental soundscapes are less vulnerable to campaign-specific shifts.

Finally, the comparison between ground truths reveals a performance drop under CV targets for specific categories, such as *Clang* ($0.55 \rightarrow 0.41$), *Cluck* ($0.26 \rightarrow 0.12$), and *Tick-tock* ($0.29 \rightarrow 0.14$). This phenomenon differs from the behavior observed in the original work [21], where the transition to CV did not cause such localized performance drops. In this temporal layout, a specific chronological fold may contain the majority of samples for a rare class; when evaluating via cross-validation, the model lacks sufficient historical anchors across the other splits, causing the absolute metric to fall. While the original dataset’s class-wise partitioning strategy insulated the cross-validation framework from these localized gaps, this non-overlapping temporal partition exposes the model’s sensitivity to chronological data distribution. Consequently, these findings illustrate that while traditional class-balanced or random splits are optimal for benchmarking core architectural capabilities under stable data distributions, they do not account for chronological variance. Instead, this non-overlapping temporal partition serves a complementary purpose by preventing session-specific memorization, establishing a more stringent benchmark specifically tailored for cross-campaign evaluation.

C. Impact of Data Splitting Protocols: Temporal vs. Random Split

To quantify the evaluation bias and artificial performance gains introduced by conventional validation schemes, we complement our proposed benchmark with a standard 5-fold random cross-validation split. Evaluating both ground-truth

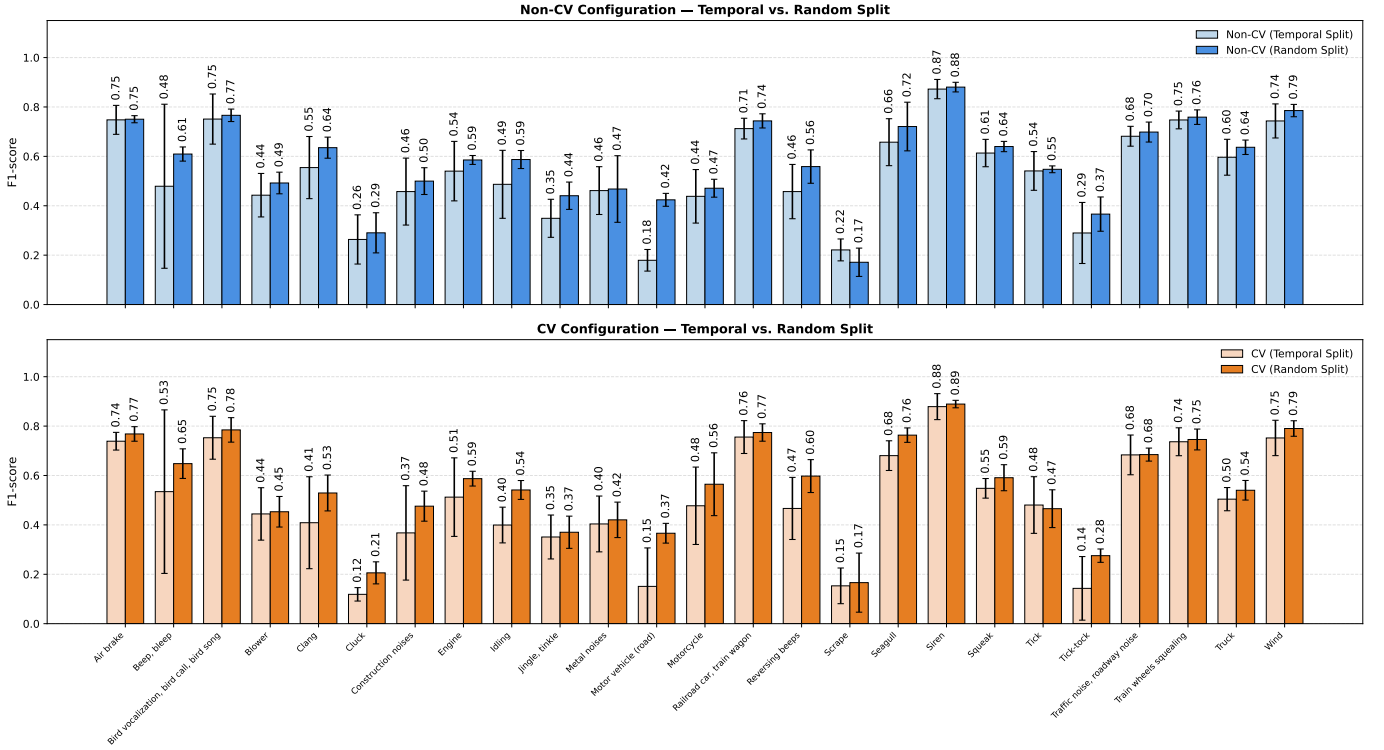


Fig. 5. Per-class F1-score comparison between Temporal Split and Random Split evaluation protocols for Non-CV (top) and CV (bottom) ground truth configurations. Bars indicate the mean performance across the five folds, while error bars denote the standard deviation ($\pm$ SD), highlighting the performance overestimation and reduced volatility of standard random splits compared to temporal evaluation.

configurations (Non-CV and CV) under both the released temporal protocol and this random partitioning allows us to explicitly measure the impact of temporal feature leakage on model generalization. Table I presents the overall performance comparison across all configurations. The results reveal a drop in performance when moving from a random to a temporal split across both ground truth settings. Specifically, for the Non-CV configuration, the temporal split leads to an absolute drop of 0.048 in mAP and 0.050 in macro F1-score. Similarly, for the CV configuration, performance decreases by 0.051 in mAP and 0.062 in macro F1-score. Furthermore, the temporal split exhibits significantly higher standard deviations across folds (e.g., ± 0.032 vs. ± 0.007 in macro F1-score for Non-CV), reflecting the natural domain shifts and acoustic variability over time.

Fig. 5 illustrates the per-class F1-score comparison between both evaluation strategies. The performance drop is particularly prominent in continuous sound sources (e.g., *Beep, bleep, Motorcycle, Motor vehicle (road)*), where random splitting benefits from temporal correlation and feature leakage between contiguous frames. Crucially, random partitioning yields unrealistically small standard deviations across folds, creating a false perception of model stability and substantially overestimating generalization. In contrast, the high variance observed across temporal folds in classes such as *Beep, bleep, Engine*, or *Reversing beeps* reveals that classifier performance is highly campaign-dependent, fluctuating significantly as local

acoustic environments evolve. These findings demonstrate that standard random validation masks real-world temporal shifts, reinforcing the necessity of strict chronological splitting protocols for environmental sound monitoring datasets.

D. Impact of Pre-trained Weights and Domain Shift: From-Scratch vs. Linear Probing Fine-Tuning

To evaluate the generalization capability of audio representations learned from large-scale general domain datasets, we compare a CNN14 architecture trained from scratch against a fine-tuned variant initialized with pre-trained AudioSet weights under a linear probing scheme, where backbone weights remain frozen and only the final classification head is updated. As summarized in Table II, models trained from scratch consistently outperform the fine-tuned linear probing strategy across both evaluation metrics and annotation paradigms. While mAP exhibits a moderate reduction of approximately 0.044 to 0.054 points (0.590 ± 0.036 vs. 0.536 ± 0.031 for Non-CV, and 0.557 ± 0.039 vs. 0.513 ± 0.031 for CV), the impact on macro F1-score is severe, dropping by more than 50% across both setups (0.531 ± 0.032 falling to 0.260 ± 0.021 in Non-CV, and 0.498 ± 0.033 dropping to 0.240 ± 0.022 in CV).

This disparity between ranking metrics (mAP) and decision-level metrics (F1-score) is further illuminated by the class-wise performance breakdown in Figure 6. The frozen AudioSet [3] feature extractor produces uncalibrated represen-

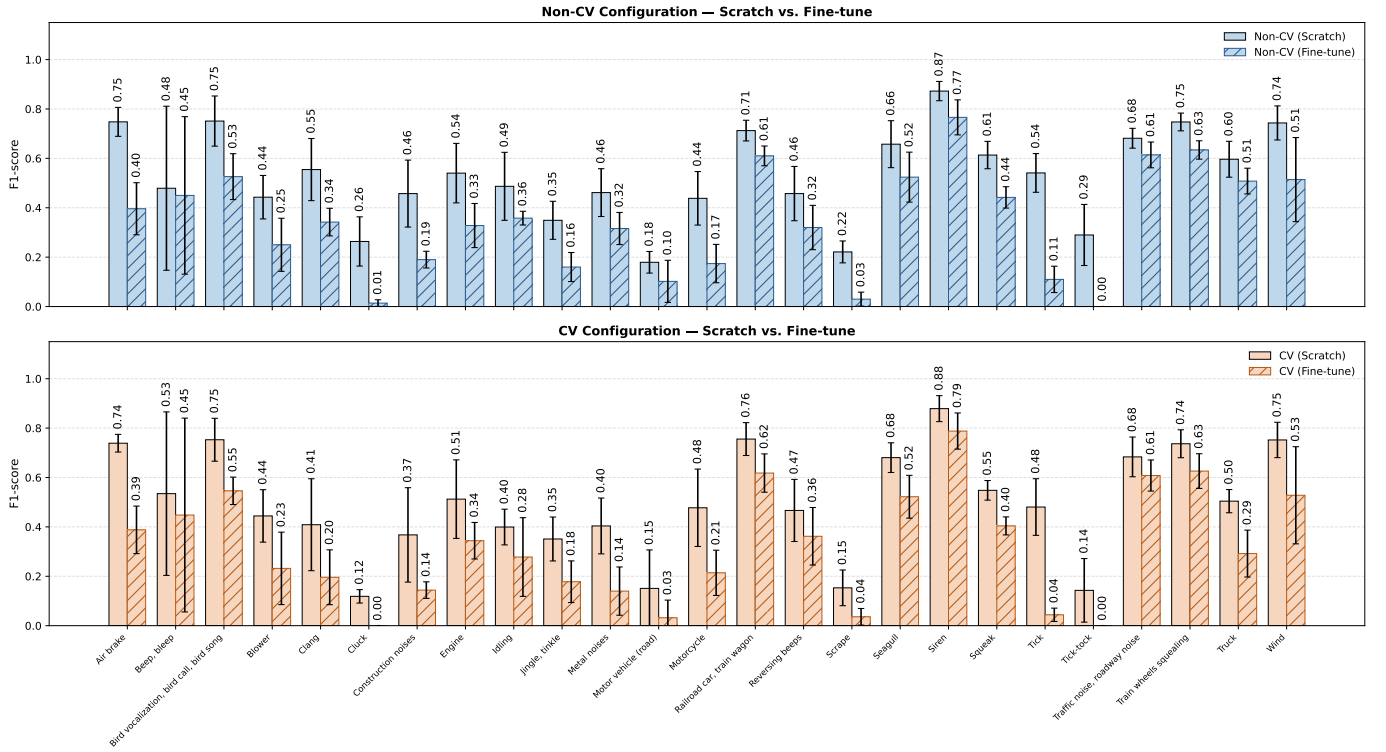


Fig. 6. Per-class F1-score comparison between models trained from scratch and fine-tuned via linear probing for Non-CV (top) and CV (bottom) ground truth configurations under the temporal split evaluation protocol. Bars indicate the mean performance across the five folds, while error bars denote the standard deviation ($\pm$ SD), highlighting the performance degradation and representation misalignment of AudioSet pre-trained weights compared to training from scratch.

TABLE II
GLOBAL PERFORMANCE COMPARISON ACROSS GROUND TRUTH CONFIGURATIONS AND TRAINING STRATEGIES (MEAN $\pm$ STD OVER 5 FOLDS).

Ground Truth	Training Protocol	mAP	Macro-F1
Non-CV	Scratch	0.590 ± 0.036	0.531 ± 0.032
	Fine-tune	0.536 ± 0.031	0.260 ± 0.021
CV	Scratch	0.557 ± 0.039	0.498 ± 0.033
	Fine-tune	0.513 ± 0.031	0.240 ± 0.022

tations for continuous urban acoustic monitoring environments, leading to a critical precision-recall mismatch under fixed decision thresholds. Consequently, fine-tuned models experience severe degradation in minority and short-duration acoustic classes—such as *Cluck*, *Tick-tock*, *Scrape*, and *Construction noises*, which collapse to near 0.00 F1-score—while only prominent, continuous sounds maintain moderate performance. This outcome highlights a significant domain shift between AudioSet’s general, high-signal-to-noise ratio clips and real-world urban continuous monitoring data characterized by heavy background noise and reverberation, demonstrating that linear probing without full backbone adaptation is insufficient to capture fine-grained urban sound patterns.

V. CONCLUSION

In this work, we introduced and officially released a temporally-structured, 25-class acoustic dataset for industrial port environments with a pre-defined 5-fold chronological split. While the original framework focuses on stable audio tagging, this subset specifically targets system validation for field deployment by preventing temporal leakage and accounting for non-stationary distribution shifts. As a baseline, a CNN14 classifier trained from scratch achieved a global Macro F1-score of 0.531 ± 0.032 under the Non-CV setup, and 0.498 ± 0.033 under cross-validation. This chronological split yields a more conservative performance than class-wise partitions due to its disjoint time frames. This closer alignment with non-overlapping acoustic distributions establishes a rigorous benchmark to develop and test model generalization under true operational constraints.

Furthermore, comparative evaluations against a standard random split confirm that chronological partitions provide a significantly more demanding benchmark, preventing the artificial performance overestimation typically caused by temporal leakage. In parallel, our investigation into pre-trained weights reveals that fine-tuning via linear probing with AudioSet initialization suffers a severe degradation in macro F1-score due to a pronounced acoustic domain shift. This highlights that general-purpose audio representations cannot easily adapt to complex, continuous urban monitoring environments without

full backbone adaptation, further reinforcing the value of this dataset as a challenging operational testbed.

REFERENCES

- [1] A. Bansal and N. K. Garg, "Environmental sound classification: A descriptive review of the literature," *Intelligent Systems with Applications*, vol. 16, p. 200115, 2022.
- [2] K. Zaman, M. Sah, C. Direkoglu, and M. Unoki, "A survey of audio classification using deep learning," *IEEE access*, vol. 11, pp. 106620–106649, 2023.
- [3] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780, IEEE, 2017.
- [4] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.
- [5] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.
- [6] A. Singh, H. Liu, and M. D. Plumbley, "E-panns: sound recognition using efficient pre-trained audio neural networks," in *Inter-Noise and Noise-Con Congress and Conference Proceedings*, vol. 268, pp. 7220–7228, Institute of Noise Control Engineering, 2023.
- [7] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.
- [8] Y. Miao, J. Zhu, and Y. Li, "Dynamic attention-asymmetric perceptron network for overlapping sound event detection," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 636–649, 2026.
- [9] S. Nsumba, T. Muhanguzi, E. N. Ouma, I. Sekalala, E. Bainomugisha, E. Mwebaze, and J. Quinn, "Noise mapping and ambient sound recordings of the urban environment in uganda," *Scientific Data*, 2026.
- [10] C. Mydlarz, M. Sharma, Y. Lockerman, B. Steers, C. Silva, and J. P. Bello, "The life of a new york city noise sensor network," *Sensors*, vol. 19, no. 6, p. 1415, 2019.
- [11] P. Gairí, T. Pallejà, and M. Tresanchez, "Environmental sound recognition on embedded devices using deep learning: a review," *Artificial Intelligence Review*, vol. 58, no. 6, p. 163, 2025.
- [12] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044, 2014.
- [13] M. Fuentes, B. Steers, P. Zinemanas, M. Rocamora, L. Bondi, J. Wilkins, Q. Shi, Y. Hou, S. Das, X. Serra, *et al.*, "Urban sound & sight: Dataset and benchmark for audio-visual urban scene understanding," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 141–145, IEEE, 2022.
- [14] K. Ooi, K. N. Watcharasupat, S. Peksi, F. A. Karnapi, Z.-T. Ong, D. Chua, H.-W. Leow, L.-L. Kwok, X.-L. Ng, Z.-A. Loh, *et al.*, "A strongly-labelled polyphonic dataset of urban sounds with spatiotemporal context," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 982–988, IEEE, 2021.
- [15] A. Mesaros, T. Heittola, and T. Virtanen, "Acoustic scene classification: An overview of DCASE 2017 challenge entries," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 411–415, September 2018.
- [16] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, pp. 85–92, November 2017.
- [17] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.
- [18] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, 2023.
- [19] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.
- [20] M. A. Lones, "Avoiding common machine learning pitfalls," *Patterns*, vol. 5, no. 10, 2024.
- [21] J. Naranjo-Alcazar, J. Grau-Haro, R. Ribes-Serrano, M. Garcia-Ballesteros, and P. Zuccarello, "Soroll-IA: A weakly labeled audio dataset for real-world industrial port monitoring," *arXiv preprint arXiv:2606.26195*, June 2026.
- [22] Y. Lin, Y. Li, W. Wu, S. Sun, F. Shi, and J. Shao, "An improved prototype learning method for domain generalized industrial machine anomaly sound detection," *IEEE Access*, 2026.
- [23] A. I. Mezza, E. A. Habets, M. Müller, and A. Sarti, "Unsupervised domain adaptation for acoustic scene classification using band-wise statistics matching," in *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 11–15, IEEE, 2021.
- [24] J. Naranjo-Alcazar, J. Grau-Haro, R. Ribes-Serrano, and P. Zuccarello, "A data-centric framework for machine listening projects: Addressing large-scale data acquisition and labeling through active learning," in *Advances in Information and Communication* (K. Arai, ed.), (Cham), pp. 647–659, Springer Nature Switzerland, 2025.
- [25] J. Naranjo-Alcazar, J. Grau-Haro, R. Ribes-Serrano, and P. Zuccarello, "Real-time audio monitoring pipeline with edge inference and iot supervision via thingsboard," in *AES International Conference on Machine Learning and Artificial Intelligence for Audio*, (London, UK), Audio Engineering Society, september 2025. Late Breaking Demo Paper.
- [26] A. Koh, X. Fuzhao, and C. E. Siong, "Automated audio captioning using transfer learning and reconstruction latent space similarity regularization," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7722–7726, IEEE, 2022.
- [27] Q. Kong, K. Chen, H. Liu, X. Du, T. Berg-Kirkpatrick, S. Dubnov, and M. D. Plumbley, "Universal source separation with weakly labelled data," *arXiv preprint arXiv:2305.07447*, 2023.
- [28] K. Wilkinghoff and F. Fritz, "On using pre-trained embeddings for detecting anomalous sounds with limited training data," in *2023 31st European Signal Processing Conference (EUSIPCO)*, pp. 186–190, IEEE, 2023.