A Comparative Analysis of Denoising Methods for Deep Learning-Based Audio Event Detection in Noisy Agricultural Environments*

André Moreira Souza[1], **Guilherme Augusto Moreira**[1], **Lucas Eduardo Gulka Pulcinelli**[1]

[1] Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP) – Avenida Trabalhador São-carlense, 400
Centro – CEP: 13566-590 – São Carlos – SP – Brazil

{andre.moreira.souza,guilherme_augusto,lucasegp}@usp.br,

Abstract. *This study investigates whether traditional denoising improves deep learning for Audio Event Detection (AED) in real-world, noisy environments like livestock farms. We evaluated three denoising algorithms — Spectral Subtraction, Adaptive Kalman Filter, and SD-ROM — on the* aSwine *dataset using state-of-the-art models, including Pretrained Audio Neural Networks (PANNs) and the Audio Spectrogram Transformer (AST). Contrary to conventional wisdom, all denoising methods proved detrimental, with a worst-case scenario showing a* 65.8% *mean Average Precision (mAP) decrease. We conclude that powerful models learn to be inherently noise-robust, making robust architectures a superior strategy to noise preprocessing.*

1. Introduction

Audio Event Detection (AED) has emerged as a critical technology in a variety of applied fields, from urban planning to wildlife conservation. One of the most promising domains for AED is Precision Livestock Farming (PLF), where non-invasive acoustic monitoring offers a scalable means to assess animal health, welfare, and behavior [Lostanlen et al. 2019]. In commercial swine farms, for instance, acoustic sensors can detect events like pig coughs, which are early indicators of respiratory disease, or screams, which can signify stress or pain. However, the practical application of AED in real-world settings like swine barns is severely hampered by noise. Acoustic clutter from ventilation systems and overlapping sounds is often non-stationary, degrading the Signal-to-Noise Ratio (SNR) and masking target event features, fundamentally challenging model performance [Azarang and Kehtarnavaz 2020].

Building on the work of Souza et al. (2025), which established the promise of deep learning on the raw, noisy *aSwine* dataset, this study addresses a key limitation they identified. Their use of unprocessed audio leaves an open research question: *Can the performance of these sophisticated models be improved by applying traditional denoising algorithms as a preprocessing step?* While conventional wisdom suggests enhancing the SNR should aid classification, the process can introduce artifacts or remove vital acoustic cues, potentially harming models that have implicitly learned to

*__Acknowledgements__: This research was supported by FAPESP (grants 2016/17078-0, and 2020/07200-9) and CAPES (grants PROEX-9778985/M and 88887.893807/2023-00)

be noise-robust [Souza et al. 2025, Hardjanto and Wahyono 2025]. To systematically investigate this trade-off, we provide a data-driven analysis comparing state-of-the-art AED models on raw audio versus audio preprocessed with several traditional denoising algorithms, clarifying whether this step is beneficial, neutral, or detrimental in a complex acoustic environment.

The paper is organized as follows: Section 2 reviews the main concepts and related work on AED and denoising; Section 3 details the experimental framework; Section 4 presents the results and discusses their implications; and Section 5 concludes the paper and suggests future work.

2. Background and Related Work

To contextualize this study's central question, it is necessary to review three intersecting research domains: the application of AED in acoustically complex environments, the theoretical underpinnings and known limitations of traditional denoising algorithms, and the paradigm shift introduced by noise-robust, end-to-end deep learning architectures.

The shift of AED from controlled laboratory conditions to real-world environments like swine barns presents significant acoustic challenges [Pan et al. 2024, Lostanlen et al. 2019], which we detail in Section 3.1. To combat performance degradation in noisy environments, signal enhancement via denoising has long been a standard preprocessing step. Traditional methods aim to estimate and remove the noise component from a signal, thereby improving the SNR of the target event. This study evaluates two classic and one specialized algorithm, each with theoretical assumptions that are critical to understanding their performance in the swine barn context. First, Spectral Subtraction (SS) subtracts an estimated noise spectrum from the signal. Its primary flaw is the assumption of stationary noise, which can introduce audible "musical noise" artifacts in non-stationary environments [Boll 1979]. Second, the Adaptive Kalman Filter (KF) is a recursive algorithm that can track non-stationary noise by modeling the signal's evolution over time; however, its efficacy depends on accurate parameter estimation from the noisy signal itself, making its benefit uncertain [Permana Putra et al. 2024]. Finally, the Single Dependent Rank-Ordered Mean (SD-ROM) filter is designed to remove short, impulsive noise. Given the continuous nature of the noise in this study, it is included as a negative control to test the impact of a theoretically mismatched filter [Rao et al. 2021].

The limitations of traditional, assumption-based signal processing have motivated a shift towards end-to-end deep learning models for audio tasks. Architectures like Convolutional Neural Networks (CNNs) and Transformers, originally developed for computer vision, have been successfully adapted to treat log-mel spectrograms as "images", learning hierarchical spectro-temporal features directly from the data [Ding et al. 2024]. A key distinction between these modern approaches and traditional denoising lies in how they handle noise. Traditional methods explicitly model and attempt to remove the noise. In contrast, data-driven deep learning models learn to extract salient features from the signal, effectively learning to be robust to the specific noise profiles present in their training data [Chen et al. 2024, Gong et al. 2021]. They do not rely on explicit assumptions about noise stationarity. This inherent robustness suggests that a separate preprocessing step may be redundant

or even counterproductive. Some research has explored using deep learning itself for preprocessing, for instance, by training a sound separation model to isolate target sources before classification [Turpault et al. 2020].

3. Materials and Methods

The core of our methodology is a systematic evaluation of traditional denoising as a preprocessing step for AED. We introduce denoising as the primary experimental variable and expand the model evaluation to include a comprehensive suite of modern architectures: the attention-based AST [Gong et al. 2021] and several models from the Pretrained Audio Neural Networks (PANNs) family [Kong et al. 2020]. To isolate the impact of preprocessing, our framework builds upon a baseline study [Souza et al. 2025], using the same *aSwine* dataset, training protocol, and evaluation metrics. The source code for this study is available in a public GitHub repository [1].

3.1. The *aSwine* Audio Dataset

We use the *aSwine* dataset[2], which consists of ∼15 hours of annotated, single-channel audio ($16\ kHz$) from a commercial swine barn. It provides a demanding testbed due to its challenging acoustic properties. The environment features: (1) High, continuous background noise; (2) A non-stationary noise profile, where the intensity and spectral content vary unpredictably; (3) Frequent overlapping events, with concurrent sounds from multiple animal and human sources; and (4) Weak labels at the 5-second level with severe class imbalance. The dataset's seven annotated classes include sounds critical for PLF, such as pig coughs (indicative of respiratory illness), pig screams (indicative of stress or pain), other animal events, and human activity-related sound events such as speech and cleaning.

3.2. Denoising Implementation and Parameterization

Each algorithm was applied to the raw audio waveforms before feature extraction. Key parameters for each denoiser were optimized via the process described in Section 3.3. First, our Spectral Subtraction (SS) pipeline, based on Boll's algorithm [Boll 1979], estimates a noise spectrum from an initial, noise-only segment of each file and subtracts it from the entire signal. Tuned hyperparameters included STFT settings, noise window duration, and the noise reduction factor. Second, using a first-order state-space model, we implemented the KF [Permana Putra et al. 2024] as a time-domain, sample-by-sample recursive filter. Unlike SS, it does not operate on the frequency spectrum but instead predicts the clean signal one sample at a time. Its behavior is governed by the process and measurement noise variances and the adaptation frequency, which were its primary tuned hyperparameters. Finally, following its design as an impulse noise filter [Ferahtia et al. 2009], the SD-ROM uses a sliding window to detect and replace samples that deviate significantly from their neighbors. The key tuned parameters were the window size and the impulse detection thresholds.

[1] Available at `https://github.com/andremsouza/denoise-aed`
[2] Available at `https://github.com/andremsouza/aswine`

3.3. Audio Event Detection Experimental Framework

The experimental setup is almost identical to the baseline study, ensuring performance differences are attributable only to the denoising step. Further, we assessed the pipeline performance on the aforementioned PANNs to provide a deeper comparative analysis.

Feature Extraction. For each of the four conditions (Baseline, SS, KF, SD-ROM), audio was converted to log-mel spectrograms for the CNN-based and AST models. We used 1-second clips, $25ms$ windows, $10ms$ hops, and 64 mel bins, creating 100×64 input patches. Each dataset was Z-normalized based on its training split. The models designed for raw waveform input (DaiNet19 and LeeNet24) naturally bypassed this step, operating directly on the 1-second audio clips.

Evaluated Deep Learning Models. We evaluated two prominent architectural families: First, the Pretrained Audio Neural Networks (PANNs) [Kong et al. 2020], a family of CNN models, including CNN14, ResNet38, MobileNetV1, and MobileNetV2 with spectrograms, plus DaiNet19 and LeeNet24 with raw waveforms. Second, the Audio Spectrogram Transformer (AST) [Gong et al. 2021], a Transformer-based model pre-trained on AudioSet that processes spectrograms as sequences of patches.

Training and Evaluation Protocol. The dataset was split 80/20 for training/testing. We trained all models using the AdamW optimizer and *BCEWithLogitsLoss*. An early stopping criterion halted training after four epochs without validation loss improvement, and a scheduler reduced the learning rate by a factor of 10 after two epochs of plateaued validation loss. Finally, we saved and evaluated the model checkpoint with the best validation weighted-average AUC.

Hyperparameter Optimization. Employing Optuna with a Tree-Structured Parzen Estimator (TPE), we conducted optimization studies for each model-denoiser pairing, allowing up to 20 optimization trials to maximize the validation weighted-average AUC. Across all model-denoiser combinations, we trained and evaluated 560 models, each with varied hyperparameters. The tuning process encompassed training parameters, including batch size, learning rate, weight decay, and denoiser parameters such as STFT settings, noise reduction factors, noise variances, and impulse thresholds, all within viable search spaces. Additionally, we implemented a median stopping criterion to terminate ineffective trials. Ultimately, we employed the optimal configuration for each pairing for the final evaluation.

Performance Metrics. We assessed performance using two primary metrics from the baseline study, both weighted by class support to account for imbalances. First, a weighted-average Area Under the ROC Curve (AUC), which measures discrimination ability independent of a classification threshold. Second, a weighted-average mean Average Precision (mAP), which evaluates the ranking of predictions and is more sensitive to false positives.

4. Experimental Evaluation

This section presents the quantitative outcomes of our experiments, focusing on the overall impact of each denoising method on model performance. The results demonstrate

that applying traditional denoising as a preprocessing step for this dataset and task is detrimental.

Figure 1 provides a clear visual summary of this core finding. The bar charts compare the mean AUC and mAP for each of the seven evaluated models across the four experimental conditions. It is visually apparent that the "No Denoiser" baseline consistently outperforms the three denoising methods. This trend holds true for both evaluation metrics, indicating that the models' ability to discriminate between classes and correctly rank predictions is highest when trained and evaluated on the raw, noisy audio.

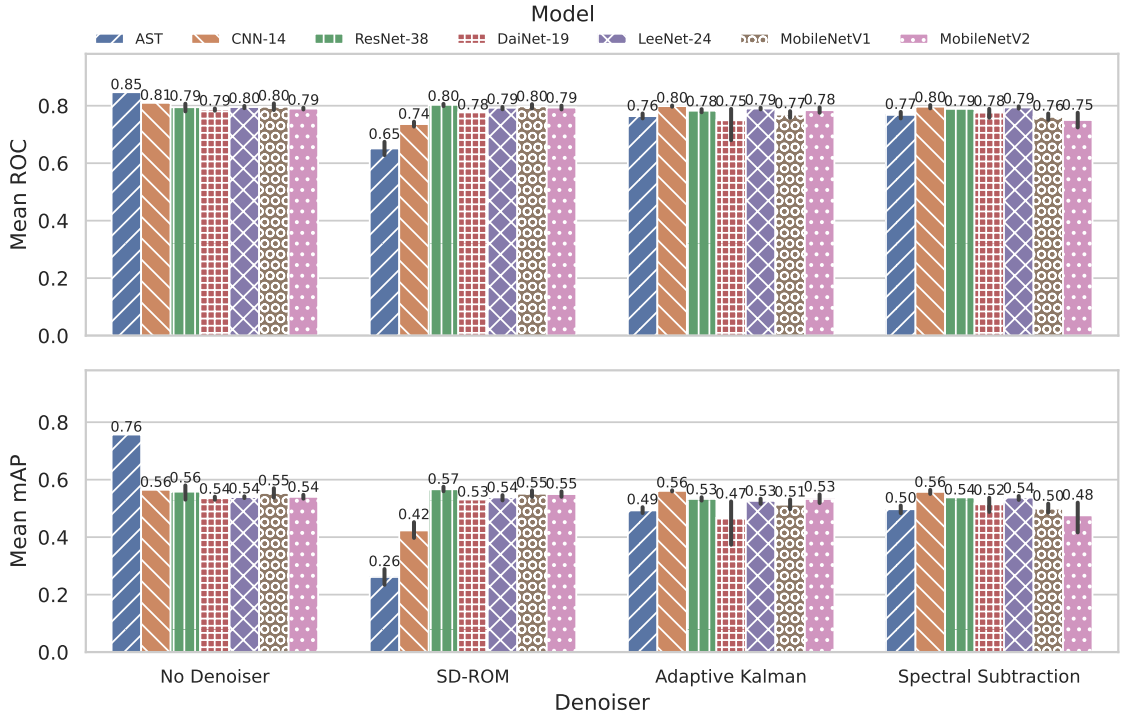

Figure 1. Mean AUC and mAP across all optimization runs for each model-denoiser combination. The "No Denoiser" baseline consistently achieves the highest performance. Error bars represent the standard deviation across runs.

Table 1 reinforces this conclusion with aggregated numerical data. The "No Denoising" condition achieved the highest mean AUC of $\mathbf{0.795 \pm 0.016}$ and the highest mean mAP of $\mathbf{0.553 \pm 0.042}$ when averaged across all models. The best single model run, an AST model without denoising, reached an AUC of $\mathbf{0.847}$. In contrast, all three denoising methods resulted in a statistically significant performance drop. The SD-ROM filter, included as a negative control due to its theoretical mismatch with the noise profile, performed as poorly as expected. More surprisingly, both the classic Spectral Subtraction and the more sophisticated Adaptive Kalman Filter also failed to provide any benefit, yielding the lowest average scores. This suggests that the signal distortion introduced by these methods, even when optimized, was more damaging than the original background noise.

The primary takeaway from these results is that the deep learning models, particularly the AST and PANNs, possess a high degree of inherent robustness to the

Table 1. Overall performance metrics aggregated across all evaluated models. Values are presented as mean ± standard deviation for the mean metrics and the best score achieved during the optimization runs.

Denoising Method	Mean AUC	Mean mAP	Best AUC	Best mAP
No Denoising	$\mathbf{0.795 \pm 0.016}$	$\mathbf{0.553 \pm 0.042}$	$\mathbf{0.847}$	$\mathbf{0.757}$
SD-ROM	0.780 ± 0.041	0.518 ± 0.082	0.811	0.581
Adaptive Kalman	0.775 ± 0.026	0.516 ± 0.039	0.807	0.564
Spectral Subtraction	0.771 ± 0.024	0.508 ± 0.040	0.802	0.565

specific, non-stationary noise found in the *aSwine* dataset. The training process on the raw spectrograms likely enables the models to learn discriminative features directly from the noisy signal, effectively treating the consistent background noise as part of the acoustic scene to be ignored. The detrimental effect of applying an external denoising algorithm is starkly illustrated in the worst-evaluated scenario: using the SD-ROM filter on the otherwise top-performing AST model caused the mean AUC to plummet from 0.85 to 0.65, a relative drop of approximately 23.5%. The impact on mean mAP was even more dramatic, collapsing from 0.76 to 0.26, representing a 65.8% decrease in performance. This demonstrates that altering the input features with an imperfect filter can disrupt the models' learned robustness, leading to a significant performance decline.

5. Discussion and Conclusion

This study demonstrates that traditional denoising is detrimental to modern AED models in this real-world environment. Our findings suggest the core issue is not a failure of noise reduction but a problem of feature distortion. The consistent, non-stationary background sounds may not be "noise" to be discarded but rather a contextual baseline that deep models use to identify target events. Imperfect filtering corrupts this baseline and disrupts the models' learned representations.

These results align with the known theoretical flaws of the chosen algorithms, such as SS creating 'musical noise' artifacts [Boll 1979] or the KF oversmoothing the signal [Permana Putra et al. 2024]. In contrast, the baseline models performed a type of intrinsic "attentive filtering", learning to ignore irrelevant patterns without explicit preprocessing. This implies a key methodological warning: applying traditional filters, developed for feature-engineered systems, can conflict with the learned invariances of end-to-end models. For modern architectures, it is often better to embrace the noise and build models that learn to hear through it, rather than risk corrupting the input with a mismatched preprocessing pipeline.

To address generalizability, future work should evaluate this pipeline on public datasets with diverse noise profiles to test where traditional filters might still be beneficial. Research should also move beyond traditional methods to explore deep learning-based denoisers, which could be co-trained with the AED model to preserve salient features. Finally, this analysis could be formalized into a framework for automatically assessing the utility of any preprocessing step for a given dataset and model.

References

Azarang, A. and Kehtarnavaz, N. (2020). A review of multi-objective deep learning speech denoising methods. *Speech Communication*, 122:1–10.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120.

Chen, R., Ghobakhlou, A., and Narayanan, A. (2024). Interpreting CNN models for musical instrument recognition using multi-spectrogram heatmap analysis: a preliminary study. *Frontiers in Artificial Intelligence*, 7:1499913. Publisher: Frontiers.

Ding, B., Zhang, T., Wang, C., Liu, G., Liang, J., Hu, R., Wu, Y., and Guo, D. (2024). Acoustic scene classification: A comprehensive survey. *Expert Systems with Applications*, 238:121902.

Ferahtia, J., Djarfour, N., Baddari, K., and Guérin, R. (2009). Application of signal dependent rank-order mean filter to the removal of noise spikes from 2D electrical resistivity imaging data. *Near Surface Geophysics*, 7(3):159–169.

Gong, Y., Chung, Y.-A., and Glass, J. (2021). AST: Audio Spectrogram Transformer. In *Interspeech 2021*, pages 571–575. ISCA.

Hardjanto, V. L. and Wahyono, . (2025). Audio Enhancement for Gamelan Instrument Recognition using Spectral Subtraction. *Engineering, Technology & Applied Science Research*, 15(2):22042–22048.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. (2020). PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S., and Bello, J. P. (2019). Robust sound event detection in bioacoustic sensor networks. *PLOS ONE*, 14(10):e0214168.

Pan, W., Li, H., Zhou, X., Jiao, J., Zhu, C., and Zhang, Q. (2024). Research on Pig Sound Recognition Based on Deep Neural Network and Hidden Markov Models. *Sensors*, 24(4):1269.

Permana Putra, F., Kartika, K., Sitti Nurfebruary, N., Misriana, M., G. S, K., and Siregar, R. H. (2024). Matlab Simulation Using Kalman Filter Algorithm to Reduce Noise in Voice Signals. *Journal of Renewable Energy, Electrical, and Computer Engineering*, 4(1):23–31.

Rao, G., Babu, D. R., Kanth, P. K., Vinay, B., and Nikhil, V. (2021). Reduction of Impulsive Noise from Speech and Audio Signals by using Sd Rom Algorithm. *International Journal of Recent Technology and Engineering (IJRTE)*, 10(1):265–268.

Souza, A. M., Kobayashi, L. L., Tassoni, L. A., Garbossa, C. A. P., Ventura, R. V., and Machado De Sousa, E. P. (2025). Deep learning solutions for audio event detection in a swine barn using environmental audio and weak labels. *Applied Intelligence*, 55(7):668.

Turpault, N., Wisdom, S., Erdogan, H., Hershey, J. R., Serizel, R., Fonseca, E., Seetharaman, P., and Salamon, J. (2020). Improving Sound Event Detection In Domestic Environments Using Sound Separation. In *DCASE Workshop 2020 - Detection and Classification of Acoustic Scenes and Events*, Tokyo / Virtual, Japan.