Perceptual Detection of Packet Loss-Induced Audio Artifacts in Black-Box Wireless Music Systems

Victória Guimarães[1,2], *Luiz Bentes*[1], *Ana Pires*[1], *Rosiane de Freitas*[2]

[1]Center for Advanced Studies and Systems of Recife (CESAR), Recife, Brazil
[2]Federal University of Amazonas (UFAM), Institute of Computing (ICOMP), Manaus, Brazil

***Abstract*—Audible degradations introduced by wireless audio transmission, such as clicks, dropouts, and glitches, can significantly compromise the perceived quality of music. These artifacts are typically caused by packet loss and often occur as short, perceptually salient events. In this work, we approach their detection as a binary sound event classification task based solely on perceptual analysis of the audio signal. The study focuses on black-box scenarios, where only the resulting audio is available for analysis, without any access to packet-level metadata, network diagnostics, or internal system information. We introduce BlueData, a dataset of music recordings labeled as clean or degraded. Degradation labels were assigned through listening tests under controlled Bluetooth transmission impairments, reflecting the presence of perceptual artifacts. A range of classical machine learning classifiers were trained using handcrafted acoustic features. Among them, models such as XGBoost and CatBoost achieved AUC scores close to 0.97, while K-Nearest Neighbors (KNN) reached the highest recall for the degraded class, with 85.09%. These results demonstrate the effectiveness of lightweight and interpretable models in identifying transmission-induced perceptual degradations directly from the audio signal and position BlueData as a relevant dataset for research in perceptual quality monitoring under black-box conditions.**

***Index Terms*—Black-box audio analysis, packet loss artifacts, perceptual audio degradation, sound event classification.**

1. INTRODUCTION

Sound event classification (SEC) is a fundamental task in signal processing, supporting applications such as environmental monitoring, speech recognition, and music information retrieval [1], [2]. Within the field of acoustic scenes and events, detection and classification tasks focus on identifying and categorizing distinct sound events in complex auditory environments [3], [4]. A sound event is typically defined as a segment of audio associated with a distinctive concept [5]. Traditional SEC tasks target structured events such as speech utterances, alarms, or instrument sounds [6], while recent studies have expanded this scope to include anomalous or transient acoustic phenomena [7].

In wireless audio transmission, degradations such as packet loss can introduce audible artifacts such as clicks, dropouts, and distortions, which, although brief, are perceptually disruptive [8]. These artifacts degrade the quality of speech or music [9] and present detection challenges due to their varying spectral and temporal characteristics. Analyzing such perceptual degradations is especially relevant in scenarios where only the final audio output is available. In these cases, internal metadata about the transmission or packet handling is inaccessible, and all quality judgments must be made solely based on the resulting audio signal.

This challenge is significant in audio quality assurance (QA) workflows for mobile and embedded systems, where human testers often rely on subjective listening to identify degradations [10]. Automating this perceptual evaluation process can reduce subjectivity, improve scalability, and enable continuous monitoring in real-world applications. By framing the detection of audible artifacts as a sound event classification problem, we aim to provide a signal-based, interpretable solution for analyzing audio integrity in black-box conditions, where only the audio signal is available to the end user.

To support this investigation, we introduce BlueData [11], a dataset of music recordings labeled as clean or degraded, based on perceptual annotation. The audio was recorded under controlled Bluetooth transmission with induced packet loss, simulating realistic degradation scenarios. BlueData is already publicly available to support reproducibility and further research. We benchmark several classical machine learning classifiers using handcrafted acoustic features to assess their effectiveness in identifying perceptual degradations as sound events.

2. RELATED WORK

SEC for anomaly detection has received considerable attention in recent years, particularly in scenarios involving machinery faults, urban environments, and industrial monitoring [12], [13], [14]. One of the key challenges in this domain is the creation of representative datasets, as capturing and annotating transient acoustic events depends heavily on the context and the nature of the sound source. In the audio domain, [15] proposed a dataset for the classification of audio artifacts such as “clicks” and “glitches” within .mp3 files. Their approach involves artificially inserting transient faults to simulate digital degradation. Similarly, the MIMII dataset [16] targets anomaly detection in industrial equipment sounds, including valves, fans, and pumps, providing normal and faulty audio samples for benchmarking.

For general-purpose sound event recognition, [17] developed an open dataset covering a variety of real-world sounds. Although not focused on music, it provides a useful benchmark for training and evaluating SEC models. These datasets, however, are not designed to capture perceptual audio degradations caused by transmission errors, such as those observed in wireless music streaming. In terms of methodology, ensemble learning and early-event detection have been explored to improve the efficiency and accuracy of SEC [18], [19]. However, most existing datasets either focus on synthetic noise artifacts or on environmental sounds unrelated to music or perceptual quality issues. Moreover, they do not account for the challenges of detecting signal-level degradations in black-box scenarios, where metadata such as transmission logs is unavailable. To support this investigation, we present BlueData, a dataset focused on perceptual artifacts caused by packet loss during wireless audio transmission. These artifacts appear as short, localized acoustic events that are difficult to characterize, especially in black-box scenarios. The dataset provides a controlled setting for exploring signal-based detection approaches and can serve as a benchmark for evaluating audio classification systems under realistic transmission conditions.

3. DATASET FOR WIRELESS AUDIO TRANSMISSION

To investigate perceptual degradations in wireless audio, we developed BlueData, a dataset comprising two classes: clean and degraded.

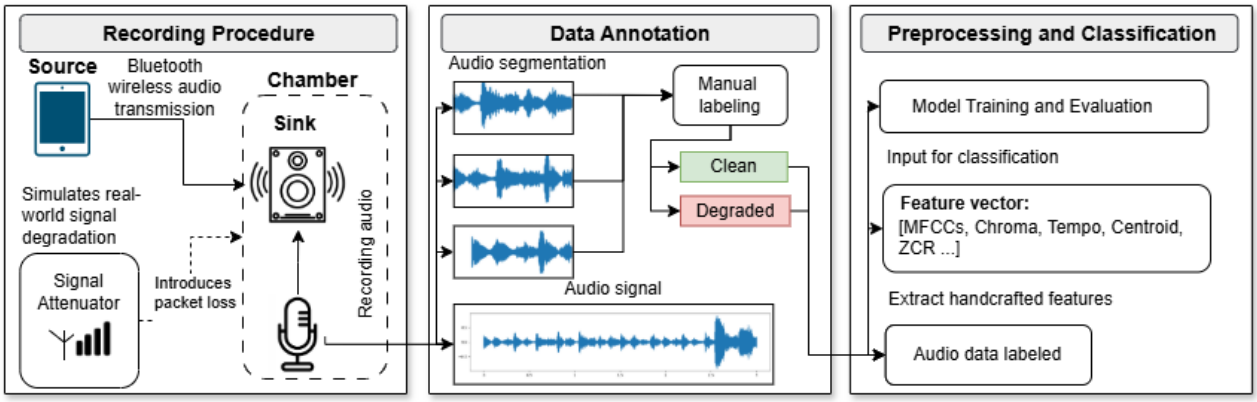

Fig. 1: Recording and preprocessing process: (1) Bluetooth audio playback through controlled attenuation; (2) acoustic recording; (3) segmentation and labeling; (4) feature extraction.

The degraded class includes perceptually salient audio events such as clicks, dropouts, and distortions caused by packet loss. These artifact types are not labeled separately, as the focus is on detecting the perceptual presence of degradation rather than categorizing its specific form. Tracks were selected from the Free Music Archive (FMA) [20], covering six musical genres: Rock, Blues, Pop, Hip Hop, Classical, and Electronic, and totaling 10.76 hours of audio. Genre balance in the train/test split was not controlled, as the dataset is intended for perceptual degradation detection rather than genre classification.

3.1. Recording Setup

The recording setup was designed to simulate realistic wireless transmission conditions while maintaining control over induced degradation. Audio playback was streamed via Bluetooth from a mobile phone placed outside an acoustically isolated chamber to a speaker inside. A microphone recorded the audio output from the speaker. To simulate transmission impairments, a Vauxin digital attenuator [21] was introduced in the signal path, allowing remote control of signal strength and inducing controlled levels of packet loss and attenuation. This process generated perceptual artifacts such as dropouts, clicks, and glitches, which are common in unstable Bluetooth connections. The general setup is illustrated in Figure 1.

This procedure enabled the creation of a dataset reflecting perceptual events caused by wireless audio degradation. All audio was stored in WAV format, mono, 44.1 kHz. Tracks were converted from MP3 (typically 192–320 kbps) to WAV to ensure a standardized, uncompressed format for feature extraction, avoiding further quality degradation during preprocessing.

3.2. Segmentation and Annotation

The recordings were segmented into 2-second windows using Librosa. This duration was selected to provide enough temporal context for detecting packet loss artifacts while maintaining temporal precision, consistent with standard SEC practices. While some artifacts may occur over very short intervals, the 2-second window captures their perceptual effect in context, which is crucial for human listeners and perceptual labeling. The full dataset comprises 19,397 labeled segments. For the experiments reported in this work, we used 15,517 labeled segments: 11,760 clean and 3,757 degraded. This natural class imbalance reflects the sporadic nature of transmission-induced artifacts in real-world scenarios. As such, the evaluation emphasizes recall and the F1-score for the degraded class.

Annotation was conducted manually through a two-pass auditory inspection by certified QA testers. Initially, one tester labeled each segment based on perceptual evaluation. A second tester then reviewed all labels independently. Disagreements were resolved through consensus to ensure labeling consistency. The dataset is publicly available [11]. Full details on file structure and access instructions are provided on the dataset page.

4. ACOUSTIC FEATURE EXTRACTION AND CLASSIFICATION

To classify clean and perceptually degraded audio, we implemented a traditional classification pipeline composed of segmentation, handcrafted feature extraction, model selection, and evaluation via cross-validation. Instead of relying on complex feature learning, we adopted interpretable acoustic descriptors and lightweight models, which are suitable for real-time or resource-constrained environments.

4.1. Data Preprocessing

Each 2-second segment was converted to mono and standardized to a sample rate of 22,050 Hz. We extracted 38 handcrafted features using the Librosa library, including zero-crossing rate, tempo (via beat tracking), 20 mel-frequency cepstral coefficients (MFCCs), 12 chroma features, spectral centroid, spectral bandwidth, and spectral rolloff. These features were selected for their ability to capture both spectral and temporal characteristics, such as harmonic structure, energy distribution, and rhythmic content.

All features were averaged over time, producing fixed-length 38-dimensional vectors. The vectors were then normalized using z-score standardization (zero mean and unit variance). No filtering or data augmentation was applied, preserving the perceptual integrity of the original segments.

4.2. Classification Models and Parameters

We trained eight classical supervised models, each representing a distinct learning paradigm. All classifiers were implemented using scikit-learn, XGBoost, or CatBoost libraries, and trained with their respective default hyperparameters, unless otherwise required. This decision ensures that comparisons focus on the feature space itself, without introducing biases from hyperparameter tuning.

XGBoost was used with `use_label_encoder=False` and `eval_metric='logloss'` to support binary classification. CatBoost was configured with `verbose=0`. The Random Forest classifier was trained with `n_estimators=100` and `criterion='gini'`. SVM used a radial basis function kernel with `C=1.0` and `gamma='scale'`. K-Nearest Neighbors was configured with `n_neighbors=5` and `weights='uniform'`.

The Decision Tree model used `criterion='gini'`. Gaussian Naive Bayes applied the standard `var_smoothing=1e-9`, and Perceptron used `max_iter=1000` and `eta0=1.0`. All hyperparameter values were confirmed using the `.get_params()` method and are fully reproducible.

4.3. Training and Evaluation

We used 5-fold stratified cross-validation to evaluate model performance. In each fold, 80% of the data was used for training and 20% for validation, ensuring balanced class distribution across splits. This procedure provides robust generalization estimates while minimizing variance from data partitioning.

We computed standard classification metrics: accuracy, precision, recall, and F1-score. Accuracy measures overall correctness, precision penalizes false positives, recall emphasizes sensitivity to degradations, and F1-score balances both. We report all metrics, but F1-score was used as the main criterion for comparing classifiers. This evaluation pipeline allows consistent, interpretable comparisons across models, highlighting their effectiveness in detecting audio degradations from handcrafted acoustic features.

5. RESULTS

This section evaluates the detection of perceptual degradations in black-box scenarios using handcrafted features, framing the task as segment-level classification of short, salient audio events. Table 1 presents the average performance of each classifier across all folds. Among the evaluation metrics, both F1-score and recall are considered primary indicators of performance. The F1-score provides a balanced measure by combining precision and recall. However, recall plays a particularly critical role in this context, as it reflects the model's ability to correctly identify degraded audio segments. Given that the objective of this task is to detect degradation events, achieving a high recall is essential to ensure that as many degraded instances as possible are accurately detected.

The best overall results were achieved by the XGBoost and CatBoost models, both with accuracy values close to 95%. In terms of F1-score, XGBoost slightly outperformed all other models, reaching 92.84%, closely followed by CatBoost (92.67%) and KNN (92.45%). Notably, KNN achieved the highest recall among all models, with 91.37%, making it particularly effective at detecting degraded segments.

Table 1: Overall classification metrics (%) for each model (average across folds).

Model	Accuracy	Precision	Recall	F1-score
XGBoost	94.99	95.22	90.94	92.84
CatBoost	94.90	95.47	90.51	92.67
RandomForest	93.35	95.44	86.59	90.02
KNN	94.60	93.68	**91.37**	92.45
SVM	94.27	94.67	89.51	91.73
DecisionTree	88.92	84.73	85.35	85.03
GaussianNB	87.36	83.05	82.03	82.51
Perceptron	85.07	79.73	81.35	80.33

To better understand the behavior of the models regarding different audio conditions, Figure 2 shows the F1-score per class (clean and degraded) for each classifier. Analyzing the F1 score by class is essential because general metrics can obscure performance discrepancies, especially when there is a class imbalance or when one class is more challenging to detect. This is particularly important in scenarios involving an audible artifact, where degraded audio can significantly impact classification accuracy.

The results show that XGBoost and CatBoost maintain strong performance in both conditions, with F1-scores greater than 96, 70% for the clean class and greater than 88% for the degraded class. Furthermore, KNN also performs well, achieving 96.48% for clean audio and 88.41% for degraded audio. These results highlight the importance of evaluating each class individually to identify models that are robust to audio degradations and capable of preserving classification performance under adverse conditions.

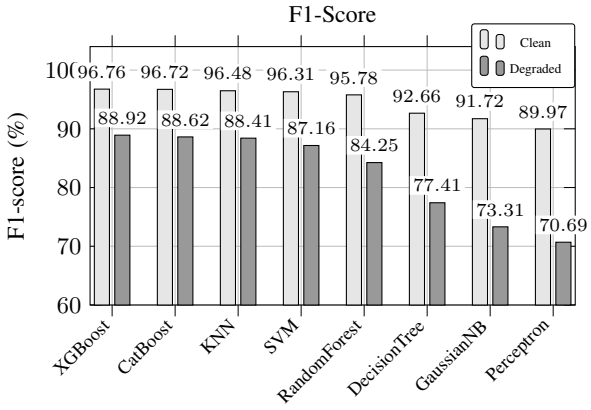

Fig. 2: Comparison of F1-scores for clean and degraded classes across classification models.

On the other hand, traditional classifiers such as Decision Tree, GaussianNB, and Perceptron exhibited a larger performance drop in the degraded class, with F1-scores below 75%, highlighting their reduced capability to generalize under noisy conditions. In the context of this task, recall plays a critical role, as it measures the model's ability to correctly identify degraded audio segments. Since the main objective is to detect subtle acoustic events associated with perceptual degradation, a high recall for the degraded class is important to minimize undetected faulty segments, which may compromise system reliability. Precision, on the other hand, measures how often the positive predictions made by the model are correct. A high precision for the clean class indicates that the model rarely misclassifies degraded segments as clean. In this context, however, the main objective is to capture as many degraded instances as possible. Thus, recall for the degraded class becomes an important aspect to consider in the overall evaluation. Figure 3 presents the recall obtained for each class across all evaluated models.

Figure 3 illustrates the recall scores for both clean and degraded classes across all evaluated classifiers. The results highlight the KNN model as the most effective in identifying degraded segments, achieving the highest recall for the degraded class with 85.09%. This demonstrates that KNN was the most sensitive model in detecting audio artifacts caused by packet loss, which is essential for ensuring system reliability.

Although boosting-based models such as XGBoost and CatBoost also performed well in terms of overall metrics and recall, their degraded-class recall was slightly lower than that of KNN, with values of 83.10% and 81.98%, respectively. This suggests that while boosting methods offer strong overall performance, KNN was the most effective specifically in capturing degraded segments.

On the other hand, Gaussian Naive Bayes exhibited the lowest recall for the degraded class, at 71.71%, indicating that a consider-

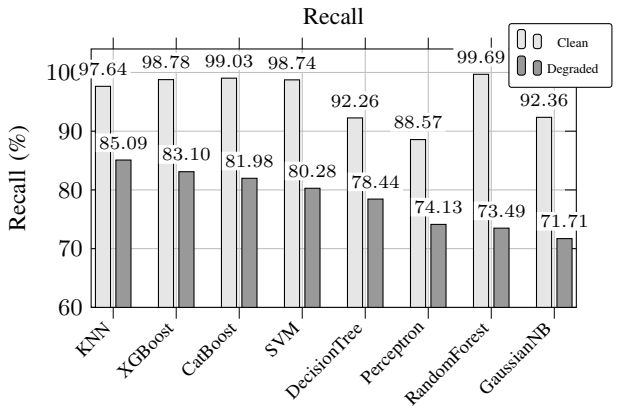

Fig. 3: Comparison of recall scores for clean and degraded classes across classification models, sorted by degraded class performance.

able number of degraded segments were not identified by the model. Despite its acceptable recall for the clean class, the tendency of the model to misclassify degraded segments limits its reliability in practical scenarios where the detection of faulty audio is critical. Figure 4 presents the confusion matrix of the KNN model, enabling a detailed evaluation of its classification performance, particularly regarding true positives and false negatives.

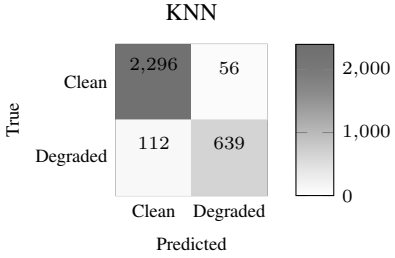

Fig. 4: Confusion matrix of the KNN model averaged across cross-validation folds.

The confusion matrix further confirms that KNN correctly classified most degraded segments, with a relatively low number of false negatives (i.e., degraded segments predicted as clean). This behavior aligns with its high recall score and reinforces its applicability in scenarios that require reliable detection of audio degradation. Nevertheless, despite the overall strong performance, the results indicate that certain degraded instances remain challenging to classify. Future work should investigate more specialized or ensemble-based classifiers capable of improving discrimination in borderline cases of degradation.

While the confusion matrix provides a detailed view of model behavior in terms of classification errors, it offers limited insight into model performance across varying thresholds. To address this, we present in Figure 5 the ROC curves, which capture the trade-off between true positive and false positive rates and offer a threshold-independent view of classifier performance.

The ROC analysis confirms that traditional machine learning models, when combined with handcrafted acoustic features, can effectively distinguish between clean and degraded audio in wireless transmission. The degraded class was defined as the positive class, so the true positive rate reflects correctly identified degradations, and the false positive rate corresponds to clean segments misclassified as degraded. Boosting-based classifiers such as XGBoost and CatBoost

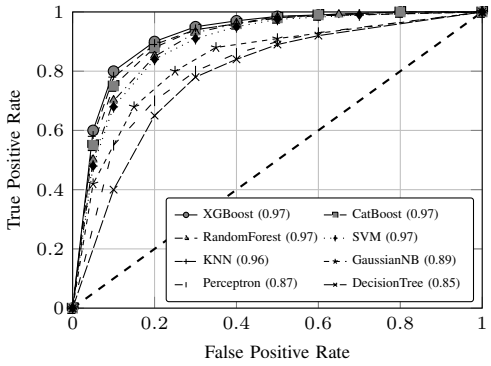

Fig. 5: ROC curve for the evaluated classifiers.

achieved the highest AUC values (both ≈ 0.97), demonstrating robust performance across decision thresholds, while KNN stood out for its sensitivity to degraded segments. This is supported by the recall results in Figure 3, where KNN achieved the highest recall for the degraded class (85.09%) In contrast, simpler models like DecisionTree and Perceptron showed more limited discrimination capabilities. These results confirm the potential of lightweight models in identifying perceptual degradations caused by packet loss and underscore the reliability of BlueData as a benchmark dataset for training and evaluating models in perceptual audio quality assessment under black-box conditions.

6. CONCLUSION

This study presented a binary sound event detection framework to identify perceptual audio degradations caused by packet loss in wireless music transmission. The task is framed as a classification problem applied to short audio segments, but the degradations exhibit event-like characteristics, as they are short, sparse, and perceptually salient. Our focus lies on black-box scenarios, where no access to transmission metadata or internal system information is possible, and only the resulting audio signal can be analyzed.

We introduced BlueData, a dataset designed to simulate realistic degradation conditions, and benchmarked a set of classical machine learning models trained on handcrafted acoustic features. Models such as K-Nearest Neighbors, XGBoost, and CatBoost demonstrated strong performance, achieving recall scores above 85% for the degraded class. These results confirm the viability of detecting packet loss-induced audio events using traditional classifiers, even without any access to system-level information. This work emphasizes the value of interpretable, low-complexity approaches for sound event detection in constrained environments. Future research will focus on improving feature extraction, refining decision thresholds for monitoring, and integrating the approach into automated wireless audio testing, where minimizing false negatives is critical.

7. ACKNOWLEDGMENT

This work was developed through a collaboration between the Federal University of Amazonas (UFAM), the Center for Advanced Systems of Recife (CESAR), and MOTOROLA, with the support of projects developed in the Manaus Free Trade Zone. According to SUFRAMA regulations, the author acknowledges this support and states compliance with Lei Federal nº 8.387/1991. This work was also partially supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES-PROEX, Finance Code 001), the National/Brazilian Council for Scientific and Technological Development (CNPq), and the Amazonas State Research Support Foundation (FAPEAM), through the POSGRAD project 2024/2025.

REFERENCES

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational analysis of sound scenes and events*. Springer, 2018, vol. 9.

[2] 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.

[3] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," *arXiv preprint arXiv:1807.10501*, 2018.

[4] S. Adavanne, "Sound event localization, detection, and tracking by deep neural networks," Ph.D. dissertation, Doctoral Thesis, Tampere University, 2020.

[5] R. Takeda and K. Komatani, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 603–609.

[6] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[7] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional lstm neural networks," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 1996–2000.

[8] S. Godsill, P. Rayner, and O. Cappé, *Digital audio restoration*. Springer, 2002.

[9] F. Toole, *Sound reproduction: the acoustics and psychoacoustics of loudspeakers and rooms*. Routledge, 2017.

[10] X. Min, G. Zhai, J. Zhou, M. C. Farias, and A. C. Bovik, "Study of subjective and objective quality assessment of audio-visual signals," *IEEE Transactions on Image Processing*, vol. 29, pp. 6054–6068, 2020.

[11] V. de Souza Guimarães, "Bluedata: A dataset of perceptual audio degradations caused by wireless packet loss," 2025. [Online]. Available: https://doi.org/10.5281/zenodo.15801604

[12] A. Abbasi, A. R. R. Javed, A. Yasin, Z. Jalil, N. Kryvinska, and U. Tariq, "A large-scale benchmark dataset for anomaly detection and rare event classification for audio forensics," *IEEE Access*, vol. 10, pp. 38 885–38 894, 2022.

[13] E. Rushe and B. Mac Namee, "Anomaly detection in raw audio using deep autoregressive networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3597–3601.

[14] L. Turchet, M. Lagrange, C. Rottondi, G. Fazekas, N. Peters, J. Østergaard, F. Font, T. Bäckström, and C. Fischione, "The internet of sounds: Convergent trends, insights, and future directions," *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11 264–11 292, 2023.

[15] D. Wolff, R. Mignot, and A. Roebel, "Audio defect detection in music with deep networks," *arXiv preprint arXiv:2202.05718*, 2022.

[16] H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "Mimii dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," *arXiv preprint arXiv:1909.09347*, 2019.

[17] 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, 2022.

[18] J. Liao, F. Yang, and X. Lu, "An enhanced contrastive ensemble learning method for anomaly sound detection." *Applied Sciences (2076-3417)*, vol. 15, no. 3, 2025.

[19] H. Phan, P. Koch, I. McLoughlin, and A. Mertins, "Enabling early audio event detection with neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 141–145.

[20] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "Fma: A dataset for music analysis," *arXiv preprint arXiv:1612.01840*, 2016.

[21] K. Staniec and K. Staniec, "Performance measurements methodology for ltn iot systems," *Radio Interfaces in the Internet of Things Systems: Performance studies*, pp. 119–135, 2020.