IJAAR PUBLISHING®
Scholarly Journal of Science and Technology Research & Development | ISSN: 2955-0807
Vol. 5, Issue 1 (2026) | www.ijaar.org/sjstrd

RULE-BASED ACOUSTIC FEATURE FUSION FOR REAL-TIME VIOLENCE DETECTION IN PUBLIC BUSES: A CONTROLLED EVALUATION USING PRE-RECORDED AUDIO SAMPLES

Anyalewechi C.J[1], Opara F.K[1] Achuga A.C[1], Isaac O.E[1], Nwosu N.D[1] and Ejeh S.O[1]

[1]Department of Electrical and Electronic Engineering, Federal University of Technology, Owerri, Nigeria

ABSTRACT

There are major security concerns in the transportation sector of developing countries, and this is due to the lack of adequate infrastructure put in place to combat this issue. This paper presents a controlled evaluation of a rule-based system implemented on a Raspberry Pi 5 platform for analysing audio signals and detecting violence in real-time. For this exercise, four acoustic features were evaluated. These were Root Mean Square (RMS) energy, Zero Crossing Rate (ZCR), High-Frequency Ratio, and Spectral Centroid. The evaluation was conducted using 40 pre-recorded audio samples gotten online. These audio samples comprised of gunshot sounds, human (male and female) screaming, and sounds of normal daily activities. The results gotten from the exercise showed a detection accuracy of 92.5%, precision of 100%, recall (sensitivity) of 85%, specificity of 100%, and an F1-score of 91.9%. The screaming samples achieved 90% detection, while the gunshot samples achieved 80%, with missed detections resulting due to the short-duration of the acoustic signatures they possessed. It was also observed that the first time of running the system (cold start) had a significant delay of about 3-8 seconds each time it was booted. Subsequently, the reruns had a reduced delay of about 1-2 seconds as it had already familiarized itself with the activities. From the findings gotten, it has been confirmed that rule-based feature fusion delivers very competitive detection performance without the extra computational demands of different machine learning approaches. This makes it a practical and affordable solution for violence detection in resource-constrained public environments.

Keywords: Acoustic feature extraction, rule-based classification, violence detection, public transportation security, embedded IoT systems, real-time audio processing

Anyalewechi, C. J., Opara, F. K., Achuga, A. C., Isaac, O. E., Nwosu, N. D., & Ejeh, S. O. (2026). Rule-based acoustic feature fusion for real-time violence detection in public buses: A controlled evaluation using pre-recorded audio samples. *Scholarly Journal of Science and Technology Research & Development, 5*(1), 35-42. https://doi.org/10.46654/9rzpnr10

1. INTRODUCTION

Transportation by bus is used heavily by daily commuters, yet safety in these vehicles is not guaranteed [1]. Various problems like physical violence, armed robbery, and threatening behaviour continue to occur within these public buses, without any timely detection or intervention from authorities. The absence of smart automated monitoring systems means that the issues are only addressed after the harm has already been done [2].

Already existing security measures such as CCTV cameras heavily depend on human operators who are meant to monitor multiple feeds at once. This is both costly and very prone to error. Apart from the financial burden of installing and maintaining such devices, video-based systems come with privacy concerns among passengers and have constantly shown that they are very limited when it comes to being reactive as opposed to being preventive [3].

Sound however offers a practical and privacy-friendly alternative for detecting violence. Aggressive events like gunshots and screaming produce acoustic patterns that can be differentiated from everyday background noise. This makes audio analysis a viable option for automated threat detection [4]. Rule-based systems that classify audio using predefined feature thresholds are very well suited to this task. This is so because they do not require large training datasets and can run efficiently on low-cost embedded hardware [5].

This paper presents a controlled evaluation of a rule-based acoustic feature fusion system tested on a Raspberry Pi 5 embedded platform using 40 pre-recorded audio samples. The evaluation examines the four features mentioned previously, assessing both their individual and combined contribution to violence detection accuracy.

2.0 LITERATURE REVIEW

Security Challenges in Public Transportation

Research in automated transportation security has been ongoing for a very long time. Valera and Velastin [1] identified that real-time automated response capability and cost-effectiveness are the most critical requirements in existing transportation security solutions. These observations still remain very valid today. Collins et al. [2] demonstrated that video surveillance systems always underperform during real time operation due to their dependence on continuous human attention, but are useful for investigation after an incident has occurred. Senior et al. [3] established that visual surveillance systems take record of important information, which in turn raise ethical and regulatory concerns, thereby making it difficult to adopt at a large scale in a real-world setting.

2. Acoustic Feature Extraction

Using acoustic features for classification of sound has been widely accepted in various research. Davis and Mermelstein [4] showed that spectral representations which are gotten from short-term audio analysis could reliably distinguish between different sound categories. Peeters et al. [15] further demonstrated that features including spectral centroid and energy-based measures captured complementary aspects of sound that individual features could not represent on their own.

3. Audio-Based Surveillance Systems

Rabaoui et al. [6] demonstrated that support vector machines combined with wavelet-based spectral features could achieve reliable acoustic threat detection. Their work established that feature selection significantly influences how reliable detection across different acoustic environments could be. Ntalampiras et al. [7] extended this by investigating probabilistic novelty detection under real-world conditions, revealing that systems that performed well in controlled settings always had a significant drop in accuracy when they were finally deployed in real-world settings. Stowell et al. [11] subsequently confirmed that no single feature could consistently outperform a carefully designed combination of features across different sound environments.

4. Machine Learning Versus Rule-Based Approaches

Recent research has increasingly explored the impact of deep learning in audio classification. Hershey et al. [8] showed that convolutional neural networks could achieve high accuracy in classification across large and diverse audio datasets. Salamon and Bello [9] demonstrated that data augmentation techniques substantially improved generalisation in CNN-based environmental sound classifiers, while Piczak [10] confirmed CNN effectiveness specifically for environmental sound recognition tasks. Despite all this, deep learning approaches carry major computational demands that render them impractical for deployment on low-power embedded hardware [14]. Rule-based classification systems, as described by Hayes-Roth [5], offer a more practical alternative for resource-constrained environments. This is so because they provide transparent, adjustable, and computationally efficient detection without the need for large labelled training datasets.

5. Research Gap

Despite the existing work already discussed, few studies have actually conducted systematic controlled evaluations of acoustic feature fusion specifically within public bus environments using embedded IoT hardware. Most evaluations rely on server-grade computing infrastructure unsuitable for practical vehicle-based deployment.

METHODOLOGY

1. System Overview

This evaluation was conducted using a rule-based audio detection system carried out on a Raspberry Pi 5 computer with a Raspberry Pi USB microphone for audio capture. The system continuously processes incoming audio streams using Python's signal processing libraries including NumPy [12] and SciPy [13], extracting acoustic features from fixed-length audio chunks for real-time classification. The Raspberry Pi 5 platform was selected due to its low cost, low power consumption, and suitability for embedded IoT deployment in resource-constrained environments [14].

2. Audio Dataset

A controlled evaluation dataset of 40 pre-recorded audio samples was assembled. These comprised of three categories:

- 10 gunshot sound samples

- 10 human screaming samples
- 20 normal daily activity samples, which including conversation, market sounds, ambient bus noise, etc.

The samples were played back through a speaker under controlled conditions and the system detection responses were observed and recorded for each sample.

3. Feature Extraction

The four acoustic features that were measured, analysed and configured were:

Root Mean Square (RMS) Energy: This measures the signal power and loudness. It is calculated as the square root of the mean of squared sample values. A threshold of 0.08 was applied for binary classification.

Zero Crossing Rate (ZCR): This measures the signal frequency content by counting how frequently the waveform crosses zero per unit time. A threshold of 0.15 was applied.

High-Frequency Ratio: This measures the proportion of spectral energy above 3000 Hz relative to total energy, capturing high-pitched distress sounds. A threshold of 0.30 was applied.

Spectral Centroid: This represents the centre of mass of the frequency spectrum, indicating overall sound brightness. A threshold of 2000 Hz was applied.

4. Classification Approach

The rule-based fusion system assigned weighted scores to each feature threshold crossing. A cumulative score threshold was used to determine the final violent or non-violent classification.

5. Evaluation Metrics

The system performance was evaluated using standard binary classification metrics derived from the confusion matrix: accuracy, precision, recall, specificity and F1-score. The response latency of the system was also observed across cold start and warm state operating conditions.

RESULTS AND DISCUSSION

1. Detection Performance

The table below represents the confusion matrix obtained from evaluating all 40 audio samples:

Category	**Predicted Violent**	**Predicted Non-Violent**
Actually Violent	17 (True Positive - TP)	3 (False Negative - FN)
Actually Non-Violent	0 (False Positive - FP)	20 (True Negative - TN)

From these values the following classification metrics were calculated:

Accuracy: This measures the proportion of total correct predictions: (TP + TN) 100/ Total = (17 + 20) 100/ 40 = 3700/40 = **92.5%**

Precision: This measures how many predicted violent events were genuinely violent: TP / (TP + FP)100 = 17 00/ (17 + 0) = 1700/17 = **100%**

Recall (Sensitivity): This measures how many actual violent events were successfully detected:

TP / (TP + FN)100= 1700/ (17 + 3) = 1700/20 = **85%**

Specificity: This measures how well the system correctly identified non-violent samples: TN / (TN + FP)100 = 2000/ (20 + 0) = 2000/20 = **100%**

F1-Score: This is the harmonic mean of precision and recall(sensitivity) providing balanced performance measure:

2 × (Precision × Recall)100 / (Precision + Recall) = 200 × (1.00 × 0.85) / (1.00 + 0.85) = 170/1.85 =

91.9%

2. Performance by Sample Category

The table presents detection outcomes broken down by audio sample category:

Sample Category	Total Samples	Correctly Classified	Missed
Gunshot sounds	10	8	2
Screaming sounds	10	9	1
Normal daily activity	20	20	0

The screaming samples achieved the highest detection rate at 90%, which is consistent with the findings by Rabaoui et al. [6] who established that sustained high-energy vocalisations produce strongly discriminative spectral characteristics. The gunshot samples achieved 80% detection, with the two missed detections attributed to shorter duration acoustic signatures that did not sustain across consecutive audio chunks long enough to satisfy the temporal validation requirement. The normal daily activity samples recorded zero false positives across all 20 samples. This demonstrated that the temporal validation mechanism effectively filtered brief acoustic anomalies from triggering unnecessary alerts [5].

3. Feature Fusion Discussion

While individual feature accuracy values were not independently logged during this evaluation, existing literature supports the advantage of feature fusion over single-feature classification [7][11][15]. RMS energy is widely established as the strongest individual discriminator for violent audio events, this is due to its direct measurement of acoustic energy elevation [4]. ZCR complements RMS by capturing high-frequency content characteristic of

screaming and impact sounds, while High-Frequency Ratio specifically targets the elevated spectral energy above 3000 Hz associated with distress vocalisations [15]. Spectral Centroid contributes primarily through false positive reduction. This distinguishes high-energy but low-brightness sounds such as bass-heavy music from genuine threats. The zero false positive result observed in this evaluation suggests the combined fusion of these four features successfully addressed the overlapping acoustic characteristics that challenge single-feature approaches [7].

4. Response Latency

Two distinct latency behaviours were observed during evaluation. Cold start detection, defined as the first detection event following system initialisation, exhibited noticeably higher latency estimated between 3 and 8 seconds. This behaviour is attributable to system warm-up processes including audio buffer initialisation and Python runtime optimisation on the embedded platform [14]. From the second detection event onwards response latency stabilised substantially, with detections occurring within approximately 1 to 2 seconds of violent audio onset. This warm state performance represents a meaningful improvement over human monitoring response times typically reported between 10 and 30 seconds [1][2].

5. Comparative Discussion

The 92.5% accuracy achieved in this evaluation compares favourably with earlier rule-based audio surveillance systems. Rabaoui et al. [6] reported approximately 85% detection rates under controlled conditions using SVM-based approaches, while Ntalampiras et al. [7] demonstrated that real-world noise conditions consistently degraded system performance below laboratory benchmarks. The zero false positive result of this evaluation is particularly notable given that false alarm rates represent one of the most commonly cited barriers to operator confidence in automated security systems [7]. While deep learning approaches reported by Hershey et al. [8] and Salamon and Bello [9] achieve higher accuracy on large datasets, their computational requirements remain incompatible with the low-cost embedded deployment demonstrated here [14].

4. CONCLUSION

This paper presents a controlled evaluation of a rule-based acoustic feature fusion system for real-time violence detection in public buses, tested using 40 pre-recorded audio samples on a Raspberry Pi 5 embedded platform. The system demonstrated an overall accuracy of 92.5%, precision of 100%, recall of 85%, specificity of 100%, and an F1-score of 91.9%.

The zero false positive result across all 20 non-violent samples represents the most practically significant finding of this evaluation. Excessive false alarms remain one of the primary obstacles to real-world adoption of automated security systems, and the temporal validation mechanism employed here proved effective in eliminating them entirely within the test conditions [5][7]. Screaming sounds were detected more reliably than gunshots, with missed gunshot detections attributed to short acoustic duration signatures that did not satisfy the temporal validation threshold across consecutive audio chunks.

Response latency analysis revealed a meaningful distinction between cold start and warm state operating conditions. While initial detection after system startup exhibited delays of up to 8 seconds, subsequent detections stabilised within 1 to 2 seconds, demonstrating that the system is capable of practical real-time performance once fully initialised [14].

Comparison with existing literature confirms that rule-based feature fusion delivers competitive detection accuracy relative to more computationally demanding approaches [6][7][8][9], while remaining deployable on affordable embedded hardware accessible to transit operators in developing nations. This combination of strong detection performance, zero false positives, and low computational cost positions rule-based acoustic feature fusion as a genuinely viable foundation for large-scale public transportation security systems.

Future work should investigate adaptive threshold mechanisms that respond dynamically to changing ambient noise levels, evaluation across larger and more diverse audio datasets, and integration with complementary sensing technologies to address missed detections arising from short-duration acoustic events.

REFERENCES

1. M. Valera and S. A. Velastin, "Intelligent Distributed Surveillance Systems: A Review," *IEE Proceedings - Vision, Image and Signal Processing*, vol. 152, no. 2, pp. 192-204, 2005.

2. R. T. Collins et al., "A System for Video Surveillance and Monitoring," *Carnegie Mellon University Technical Report*, CMU-RI-TR-00-12, 2000.

3. A. Senior et al., "Enabling Video Privacy through Computer Vision," *IEEE Security and Privacy*, vol. 3, no. 3, pp. 50-57, 2005.

4. S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366, 1980.

5. F. Hayes-Roth, "Rule-Based Systems," *Communications of the ACM*, vol. 28, no. 9, pp. 921-932, 1985.

6. A. Rabaoui, M. Davy, S. Rossignol, and N. Ellouze, "Using One-Class SVMs and Wavelets for Audio Surveillance," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 4, pp. 763-775, 2008.

7. S. Ntalampiras, I. Potamitis, and N. Fakotakis, "Probabilistic Novelty Detection for Acoustic Surveillance Under Real-World Conditions," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 713-719, 2011.

8. S. Hershey et al., "CNN Architectures for Large-Scale Audio Classification," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 131-135, 2017.

9. J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279-283, 2017.

10. K. J. Piczak, "Environmental Sound Classification with Convolutional Neural Networks," *IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1-6, 2015.

11. 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.

12. C. R. Harris et al., "Array Programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357-362, 2020.

13. P. Virtanen et al., "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261-272, 2020.

14. W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge Computing: Vision and Challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637-646, 2016.

15. G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The Timbre Toolbox: Extracting Audio Descriptors from Musical Signals," *Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2902-2916, 2011.