Expert Systems With Applications 326 (2026) 132549

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Mel-based feature extraction for acoustic event detection and classification: A systematic review and research roadmap

Vahid Hajihashemi [a], Abdorreza Alavi Gharahbagh [a], José J.M. Machado [b], João Manuel R.S. Tavares [b,*]

[a] Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, s/n, 4200-465, Porto, Portugal
[b] Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, s/n, 4200-465, Porto, Portugal

ARTICLE INFO

Keywords:
Mel-based features
Mel spectrogram
Feature extraction methods
Sound event detection and classification

ABSTRACT

Acoustic Event Detection and Classification (AEDC) systems are artificial intelligence technologies widely used in modern audio-based monitoring and decision-support applications. These systems may operate independently or function as perceptual modules within intelligent decision-making frameworks and expert systems. AEDC has been applied in diverse domains, including security monitoring, healthcare diagnostics, environmental surveillance, and industrial maintenance. Although Mel-based features are extensively employed in both traditional machine learning and deep learning AEDC approaches, no comprehensive review has systematically analysed their usage patterns and parameter configurations. This systematic review examines 373 AEDC studies published between 2020 and 2025 that employ Mel-based feature extraction. Five principal feature types are analysed: Log Mel Spectrogram (237 studies), Mel Spectrogram (61 studies), Log Mel Band Energies (34 studies), Mel-Frequency Cepstral Coefficients (36 studies), and Mel Band Energies (5 studies). Four critical Mel-related parameters-number of Mel filters, frame length, overlap percentage, and window type-are evaluated to identify dominant configurations, feature-specific patterns, and temporal trends. Classification architectures are systematically mapped across studies to analyse the relationship between feature selection and model design. The results reveal substantial variability in parameter choices across feature types and application contexts, as well as dominant trends aligned with deep learning frameworks. The review further identifies methodological gaps and outlines research directions to improve the design, reporting, and optimisation of Mel-based AEDC systems.

1. Introduction

Acoustic/Audio Event Detection and Classification (AEDC) has grown into an active area of signal processing research. Applications in this field are rapidly expanding because detecting acoustic events is effective both independently and when combined with video-based systems. AEDC systems understand their environment through sound analysis. Their popularity covers diverse applications from security systems to smart homes and medical systems. Sound and audio signals provide substantial environmental information. AEDC systems increasingly function as sensory components within larger intelligent systems. Modern expert system architectures utilise AEDC for situation assessment, anomaly detection, and context-aware decision-making across diverse domains. Combined with visual processing, sound and audio create a more comprehensive understanding of situations. The following items clarify the distinction between sound, audio, and acoustic signals.

Sound refers to vibrations travelling through media (air, water, solids) as mechanical waves, including:

- Human audible range (20 Hz to 20 kHz);
- Infrasound (below 20 Hz): earthquake vibrations, whale communications;
- Ultrasound (above 20 kHz): medical imaging, bat echolocation;
- Structural mechanical vibrations (even when not perceived as sound).

Acoustic signals relate to sound properties, behaviour, and study within specific environments (air, water). Acoustics represents a specialised field encompassing:

- Architectural acoustics: room and building sound behaviour;
- Underwater acoustics: sonar applications;
- Aeroacoustics airflow: generated sounds;

* Corresponding author.
E-mail addresses: up201912327@edu.fe.up.pt (V. Hajihashemi), up202003516@fe.up.pt (A.A. Gharahbagh), jjmm@fe.up.pt (J.J.M. Machado), tavares@fe.up.pt (J.M.R.S. Tavares).

https://doi.org/10.1016/j.eswa.2026.132549
Received 1 November 2025; Received in revised form 28 February 2026; Accepted 20 April 2026
Available online 7 May 2026
0957-4174/© 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

- Bioacoustics: animal sound production and perception.

Audio signals are signals that microphones acquire and supply to devices such as radios, televisions, and phones, which encode, transmit, and decode.

- In summary, audio signals represent the broadcastable form of sound and serve as a bridge between the physical and digital worlds.

List of key acronyms

Acronym	Definition
AED	Acoustic/Audio Event Detection
AEDC	Acoustic/Audio Event Detection and Classification
CNN	Convolutional neural network
Conformer	Convolution-Augmented Transformer
CRNN	Convolutional Recurrent Neural Network
DNN	Deep Neural Network
EINV2	Event-Independent Network V2
FCN	Fully Convolutional Network
FDY	Frequency Dynamic Convolution
GMM	Gaussian Mixture Model
GRU	Gated Recurrent Unit
HMM	Hidden Markov Model
JSD	Jensen Shannon divergence
LMS	Log Mel Spectrogram
LMBE	Log Mel Band Energies
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficients
MIL	Multiple Instance Learning
MS	Mel Spectrogram
NF	Number of Mel filters
OP	Overlap Percentage
RNN	Recurrent Neural Network
SED	Sound Event Detection
SVM	Support Vector Machine

Note: This table contains only the most frequently used acronyms in this article. For the complete alphabetical list of all acronyms, see Appendix A.

Sound signals span the widest spectrum, from microscopic, undetectable vibrations to massive, low-frequency waves travelling through the Earth during seismic events. The term “sound signal” is equivalent to all three categories throughout this study. Sound signals have several practical advantages over video. They cost less to set up, they cope better with varying environmental conditions, and they can pick up events that fall outside the camera’s view. Using microphones, a space can be fully covered; sound is less sensitive to weather conditions and light exposure than video systems. Unlike video, physical obstacles cannot eliminate or block sound signals. Many events, such as gunshots or screams, can be more accurately and efficiently understood through sound. Methods for sound data are less complex than those for video data. In contrast, video covers a wider range of events and contains more information than sound signals. Additionally, more complex events can be understood using visual signals.

An overview of a Mel-based AEDC system, including details of different steps, is shown in Fig. 1. AEDC divides into two categories: frame-level and segment-level approaches. At the frame level, sound is analysed with very high temporal precision, typically in short frames such as every 20 milliseconds (ms). This analysis provides highly accurate, time-specific predictions, making the frame-level AEDC suitable for detecting both short- and long-term events. Due to the short frame length, this category is more sensitive to background noise than segment-level approaches. The segment-level approach divides sound into larger frames, such as several seconds, and examines event occurrence across the entire frame. This method has lower accuracy in determining precise event timing but is stronger against noise and computationally lighter. Frame-level detection performs better for tasks requiring precise timing, while the segment-level approach performs better for classifying temporally extended events. Some applications of AEDC include:

- **Environmental monitoring systems**: Systems equipped with sensors and AEDC algorithms for monitoring and analysing sounds of wildlife and various species in natural habitats (Juodakis & Marsland, 2022; Nolasco et al., 2023).
- **Urban monitoring systems**: In urban areas or smart cities, the aim of using AEDC is to improve security, detect anomalies (Harimi et al., 2022), and assess noise level (Hajihashemi et al., 2022; Shanmukha & Deepak, 2024).
- **Industrial environment monitoring systems**: Monitoring various events in industrial environments, workers’ conditions, noise level assessment, and even workforce management (Akbalık et al., 2024; Boztas et al., 2024; Wißbrock et al., 2024).
- **Traffic systems**: Monitoring traffic anomalies and rapid detection of accidents (Jiang et al., 2024; Podda et al., 2024).
- **Vehicle safety and driver assistance systems**: In combination with video systems and sensors aimed at enhancing driving safety and assisting drivers (Castorena et al., 2024).
- **Healthcare monitoring systems**: Monitoring health-related sounds, such as coughing, snoring, breathing patterns, for telemedicine, sleep analysis, monitoring patients with respiratory conditions, or detecting signs of distress. These systems also have applications in hospital environments (Baronetto et al., 2025; Fernando et al., 2022).
- **Security and surveillance systems**: Technologies for detecting and classifying sounds related to security threats or unusual events, such as glass breaking, gunshots, or car alarms, and even analysing surveillance camera footage for advanced security monitoring (Ciaburro, 2025).
- **Smart home automation**: The integration of AEDC capabilities into smart home systems enables task automation, for instance, by alerting homeowners when acoustic events occur anywhere in the house (Krstulović, 2017).
- **Industrial equipment monitoring and predictive maintenance**: Systems embedded in industrial equipment for monitoring machinery and equipment in industrial environments by analysing sounds to detect anomalies, predict failures, schedule maintenance, and optimise performance to minimise breakdowns and improve operational efficiency (Tagawa et al., 2021; Ueda et al., 2024).
- **Entertainment and gaming**: Integrating AEDC algorithms into gaming and entertainment systems for creating more immersive and responsive experiences (Park et al., 2022).

Fig. 2 shows the prisma diagram of the current review. The systematic search was conducted across four major academic databases: IEEE Xplore, Springer, Elsevier (ScienceDirect), and MDPI. The initial search targeted article titles containing the terms “Acoustic”, “Sound”, or “Audio” combined with “Event” and either “Detection” or “Classification”, which returned 2163 records. After removing 136 duplicates and 42 non-English publications, 1985 records remained. A time window filter restricted the results to studies published after 2019, reducing the set to 1325 records. A second inclusion step then searched the keywords, abstracts, and full text of these records for the terms “Mel” and “Mel-base” to ensure that only studies employing Mel-based features were retained, leaving 1229 studies. Two rounds of qualitative screening followed: a clarity assessment checked whether each study reported sufficient methodological detail to extract the Mel parameters of interest, reducing the set to 826, and a quality assessment evaluated the rigour of the experimental design and result reporting, leaving 517 studies. A final overall assessment selected the 373 studies that met all inclusion criteria and are analysed in this review. Despite advances and extensive research in AEDC, due to the number of applications and rising expectations for artificial intelligence-based systems, many challenges remain.

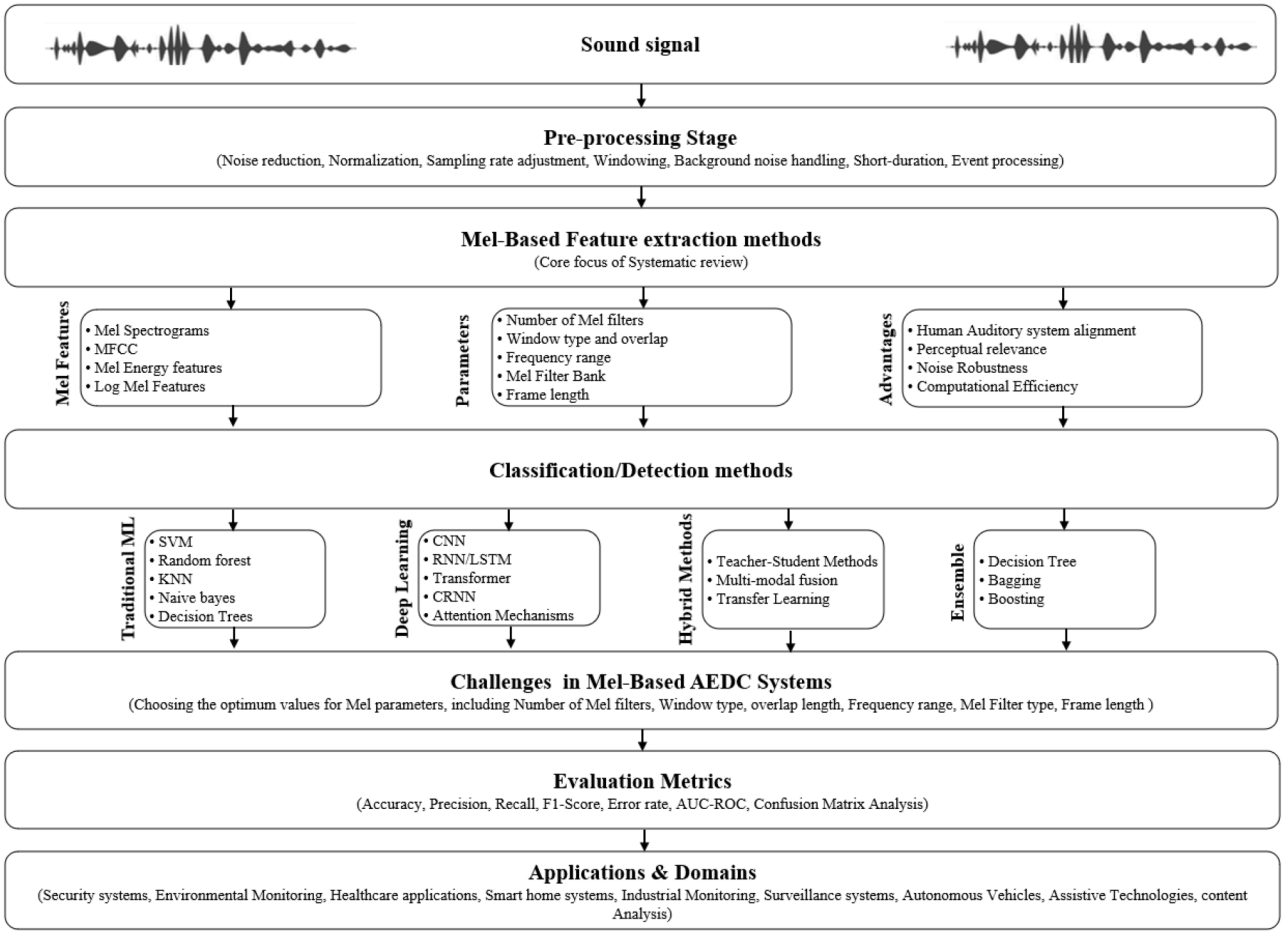


Fig. 1. An overview of a Mel-based AEDC system.

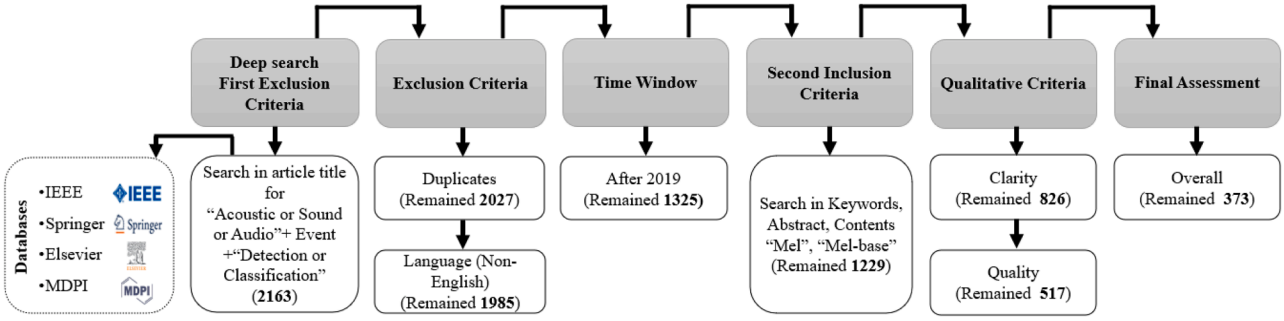


Fig. 2. PRISMA flow diagram of the article selection process.

When multiple acoustic events co-occur, the environment is noisy, the signal-to-noise ratio is low, or a fast response is needed within a short time, there are still serious concerns in the field of AEDC. Among the three areas of preprocessing, feature extraction, and classification algorithms, the present review focuses on feature extraction.

Various features have been used in AEDC. The most widely used features are mel-based features such as LMS, MS, MFCCs, and LMBE. This review examines methodological trends in Mel-based feature extraction of AEDC systems rather than performance benchmarking. Direct quantitative comparison of reported accuracies or F1-scores across studies is not feasible because researchers employ different evaluation datasets (UrbanSound8K, DCASE challenges, AudioSet, domain-specific collections), address different problem complexities (binary detection versus 50+ class classification), and operate under varying noise conditions.

This review identifies the configurations that were widely used for different acoustic event types, guiding parameter selection based on event characteristics and application constraints. Some deep learning-based approaches or pre-trained models, such as VGGish, also perform feature extraction from raw sound signals using Mel-based features. Methods such as the Gammatone Filterbank, Auditory Spectrograms, wavelet transform, statistical features, zero-crossing rate, energy distribution, pitch-related information, and harmonic/timbral features have also been considered in AEDC research. Review articles conducted in the field of AEDC, from 2020 to 2025, are listed in Table A.4. Mel-based feature extraction has been addressed as a subsection only once, which was merely explanatory. In other reviews, Mel-based features have been mentioned only as standard features, and no categorisation of detection and classification methods applied to this feature in the AEDC field has been provided. All reviews have mentioned the Mel-based

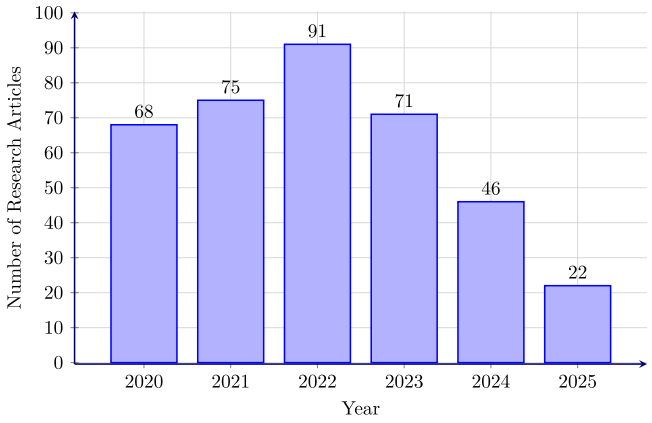


Fig. 3. Number of MEL-based AEDC research studies identified from 2020 to Spring 2025.

features as one of the most critical feature extraction methods in AEDC systems. Still, none have given special attention to the Mel-based feature extraction methods as the dominant approach.

The advantage of the current review over these previous works is its dedicated focus on extracting and comparing Mel parameters across a large number of studies, enabling the identification of patterns and trends that would not be visible in a broader review that treats all feature types equally. The analytical framework used here, combining statistical parameter analysis with temporal trend measurement through JSD and domain-specific discussion, provides a more complete picture of how Mel features are actually used in practice. A limitation of this approach is that the review focuses on methodological trends in parameter selection rather than direct performance comparison. Since different studies use different datasets, class counts, and noise conditions, comparing reported accuracy or F1 scores across studies would not produce meaningful conclusions. The review covers only Mel-based features and does not examine alternative feature extraction methods, such as Gammatone filter banks or wavelet transforms, which would require a separate analysis. Fig. 3 shows the number of AEDC research studies using Mel-based features from 2020 to spring 2025. The high number of studies conducted in recent years demonstrates the importance of Mel-based features in AEDC research. According to Fig. 3, the highest number of research was conducted in 2022. After that, the number of studies has been declining, possibly due to the emergence of deep learning-based models that directly use sound signals.

The main contributions of this review are as follows. First, it provides the first systematic review dedicated entirely to Mel-based feature extraction in AEDC, covering 373 studies from 2020 to 2025 across five feature types. Second, it extracts and analyses four critical Mel parameters (NF, frame length, OP, and window type) from each study, revealing distinct parameter selection patterns across different features. Third, it quantifies temporal trends in parameter choices using Jensen-Shannon divergence, showing how the field has evolved over time. Fourth, it provides analytical reasoning for why certain parameter ranges suit specific acoustic conditions, moving beyond descriptive statistics to explain the signal processing rationale behind observed patterns. Fifth, it offers practical guidance for researchers selecting Mel parameters based on their target application domain and acoustic environment. Table A.4 shows that previous AEDC reviews have mentioned Mel features only briefly, and none have provided this level of focused analysis on the parameters that directly affect system performance. This review proceeds in four main sections. Section 2 covers Mel scale theory and the five features examined: LMS, MS, LMBE, MBE, and MFCC. The key parameters are explained, including NF, frame length, OP, and window type. Section 3 analyses 373 AEDC studies from 2020 through 2025, organised by feature type. Each feature is analyzed separately, with charts showing temporal trends in parameter selection, whilst the detailed tables listing individual studies and their parameters are provided in Appendix B. Section 4 then discusses the analytical reasoning behind observed parameter patterns, including how different application domains influence parameter selection (Section 4.3.7). Section 4 examines findings through three complementary approaches: overall parameter comparison across all features, temporal pattern analysis using Jensen-Shannon divergence to quantify, and detailed explanations of why specific parameters suit different acoustic conditions. This section also identifies current research limitations and future directions. Section 5 summarises key contributions and provides practical guidance for AEDC system development.

2. Mel fundamentals

This section presents the theoretical foundation of the Mel scale and describes the complete feature extraction process used in AEDC systems. The key parameters analysed throughout this review (NF, frame length, OP, and window type) are introduced here, and a three-phase algorithmic framework is provided to show how raw audio signals are transformed into Mel-based features. The Mel scale represents how people hear different pitches. It is built on the idea that listeners can judge when sounds are equally spaced apart in pitch perception. The pitch in acoustics means the perceptual quality that lets one order sounds from low to high by frequency; therefore, it is what the human mind does when processing different tones.

2.1. Mel theory

The conversion from the Hz scale to the Mel scale is given by:

$$m = 1127 \ln\left(1 + \frac{f}{700}\right), \tag{1}$$

where f represents the frequency in Hz and m denotes the corresponding value on the Mel scale. The inverse transformation, from the Mel scale to Hz, is given by the following equation, which corresponds to the inverse of Eq. (1):

$$f = 700\left(e^{\frac{m}{1127}} - 1\right), \tag{2}$$

where m is the Mel value and f becomes the frequency in Hz.
Fig. 4 illustrates the process flow of the standard Mel-based feature extraction scheme. In the first step, the sound signal is segmented into shorter frames using an appropriate window function. Proper windowing mitigates undesirable effects in the frequency domain, such as distortion and spectral leakage.

The frame length affects all Mel outputs. The input signal is multiplied point-by-point by the window type. The two common window types for extracting Mel-based features in AEDC systems are Hamming and Hanning. Other types, such as Blackman, are so rare. The Hamming window is one of the most commonly used window functions, as it provides a good trade-off between reduced sidelobe levels and a relatively flat main lobe response. Its discrete-time form is given by:

$$w[n] = 0.54 - 0.46 \cdot \cos\left(\frac{2\pi n}{N-1}\right), \tag{3}$$

where n represents each sample index in the window (ranging from 0 to $N-1$), N is the total number of samples in the window, and $w[n]$ is the window function. The Hanning window works similarly to the Hamming window but produces a smoother transition. Many researchers prefer this one when acoustic events need a sharper transient band. Eq. (4) shows the Hanning window discrete function.

$$w[n] = 0.5 \cdot \left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right), \tag{4}$$

where $w[n]$ is the weight for sample n, n goes from 0 (zero) to $N-1$, and N represents the window length. The Blackman window offers the

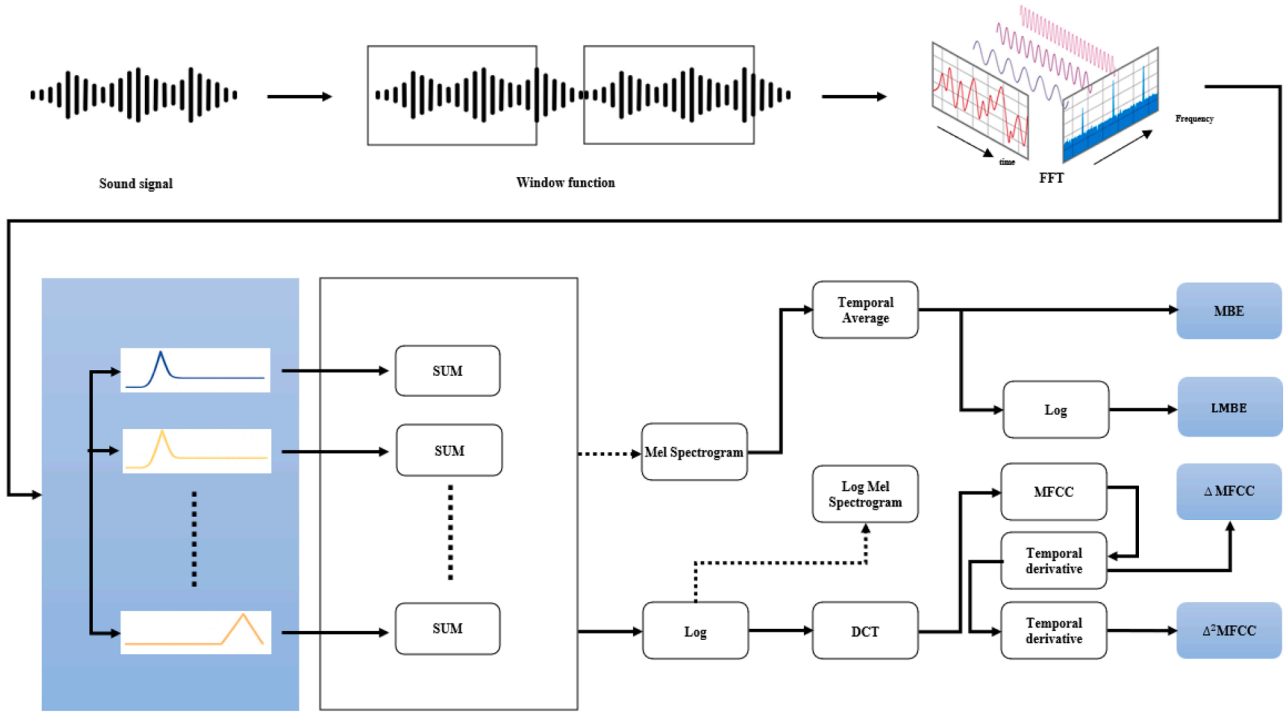


Fig. 4. Flowchart of the conventional Mel feature extraction process.

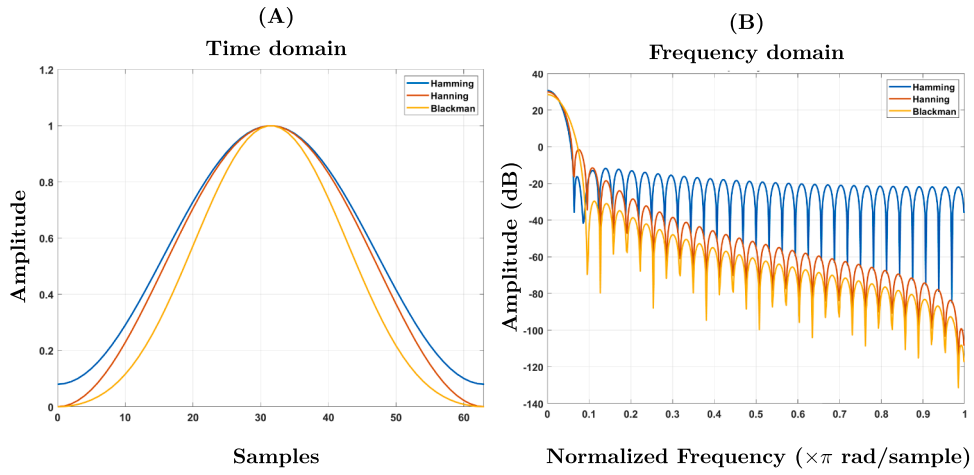


Fig. 5. (A) time domain and (B) frequency domain window types.

lowest maximum stopband ripple, making it better suited when precision in the frequency band matters most. The Blackman window has parameters similar to those of the Hamming and Hann windows, where n denotes the sample index, N represents the total window length, and $w[n]$ corresponds to the window function:

$$w[n] = 0.42 - 0.5 \cdot \cos\left(\frac{2\pi n}{N-1}\right) + 0.08 \cdot \cos\left(\frac{4\pi n}{N-1}\right). \quad (5)$$

The time patterns of these functions (Fig. 5) look similar to each other, but the fundamental differences are in the frequency domain (Fig. 5). As shown in Fig. 5, the Hamming function gives you decent frequency resolution in the passband but drops off gradually in the transient band. The Hanning function behaves similarly to the Hamming function in the passband. It drops off more slowly than Hamming in the transition band, but its side lobes are lower in the transient band. The Blackman function has the slowest drop-off in the passband and a wider passband than the other two functions. However, its side lobes have much lower amplitudes in the transition and stopbands than those of the previous two functions.In the ideal windowing, the main lobe in the passband should be a rectangle, and the amplitude of the lobes in both the transition and stop bands should tend to 0 (zero). Other window types, such as the Povey and Blackman Harris windows, have appeared occasionally in AEDC studies but remain rare, each accounting for fewer than 2% of the reviewed works. After the signal passes through the window, it is transferred to the frequency domain using the Fourier transform. At this stage, the number of points in the Fourier transform depends on how fast you sample the signal and how long each frame is. The Fourier transform output is applied to the Mel filter bank. The Mel filter bank is the main component for extracting Mel-based features. In this step, several parameters must be selected, including the NF (number of filters), the maximum and minimum filter bank frequencies, and the type of Mel filter (e.g., triangular or Gaussian). By selecting these three parameters, the Filter's characteristics can be obtained. After calculating the centre frequency f_c and the upper and lower filter frequencies F_{m+1}

and F_{m-1} for each filter on the Mel scale, the Mel filter is defined. The frequency response for the m-th filter in triangular form is given by:

$$H_m(k) = \begin{cases} 0 & k \le f_{m-1}, \\ \frac{k - f_{m-1}}{f_m - f_{m-1}} & f_{m-1} \le k \le f_m, \\ \frac{f_{m+1} - k}{f_{m+1} - f_m} & f_m \le k \le f_{m+1}, \\ 0 & k \ge f_{m+1}, \end{cases} \tag{6}$$

where k represents the frequency bin index, f_{m-1}, f_m, and f_{m+1} are the triangular filter's lower, centre, and upper frequencies, respectively. The filter creates a triangular shape where it starts at 0 (zero), rises linearly to its peak at the centre frequency, and then drops to 0 (zero). The transition between the triangular filters happens linearly. Triangular filters are perfect for real-time applications where speed and simplicity are important. On the other hand, Gaussian filters create bell-shaped curves, and the frequency overlap between filters is smoother. Gaussian filters give an excellent approximation of how the human ear works because their frequency response is similar to that of the human cochlea. Gaussian filters require more processing power. The choice between triangular and Gaussian filters often depends on whether the AEDC system prioritises computational efficiency or biological accuracy. After the Mel filter bank collects outputs from each filter, the outputs are summed in the frequency domain. The final form shows the spectral energy contents of the input frame, defined as MS, as the first feature of the Mel system. The logarithm of these values formed the LMS feature. After a DCT operation, the result will be MFCC. The first and second time derivatives of MFCC (Δ MFCC and Δ^2 MFCC) are among other Mel-based features.

2.2. Mel-based feature extraction algorithm

Mel-based feature extraction can be modelled as a systematic three-phase algorithmic framework, where each phase builds upon the outputs of its predecessor. The complete algorithm comprises: **Phase 1** (Algorithm 1) preprocessing raw waveforms through temporal segmentation and windowing; **Phase 2** (Algorithm 2) transforming windowed frames into perceptually-scaled frequency representations via Mel filter banks; and **Phase 3** (Algorithm 3) computing final features with optional temporal dynamics and normalisation. The windowed frames $\{\mathbf{x}w^{(t)}\}$ produced by Phase 1 serve as direct inputs to Phase 2, while Phase2's Mel filter bank outputs $\{\mathbf{S}^{(t)}\}$ feed into Phase3's feature computation, creating a continuous processing pipeline from time-domain audio to feature vectors $\mathbf{F}final$.

Algorithm 1: Signal preprocessing and windowing.

Input: Raw audio signal $\mathbf{x}[n]$, sampling rate f_s, frame length N_{frame}, hop size H, window type W_{type}, pre-emphasis coefficient $\alpha = 0.97$
Output: Windowed frames $\{\mathbf{x}_w^{(t)}\}_{t=0}^{T-1}$

```
// Step 1.1: Apply Pre-emphasis Filter
x_pre[0] ← x[0]
for n ← 1 to N_total − 1 do
    x_pre[n] ← x[n] − α x[n − 1]
// Step 1.2: Frame Segmentation
T ← ⌊(N_total − N_frame)/H⌋ + 1
for t ← 0 to T − 1 do
    x^(t) ← x_pre[ t · H : t · H + N_frame − 1 ]
// Step 1.3: Window Function Application
for t ← 0 to T − 1 do
    switch W_type do
        case Hamming do
            w[n] ← 0.54 − 0.46 cos( 2πn / (N_frame − 1) )
        case Hanning do
            w[n] ← 0.5 ( 1 − cos( 2πn / (N_frame − 1) ) )
        case Blackman do
            w[n] ← 0.42 − 0.5 cos( 2πn / (N_frame − 1) ) + 0.08 cos( 4πn / (N_frame − 1) )
    x_w^(t)[n] ← x^(t)[n] w[n]
```

2.2.1. Phase 1: signal preprocessing

The initial phase prepares the raw audio signal for frequency analysis (Algorithm 1). This preprocessing step is essential because raw audio signals contain spectral imbalances that can affect the quality of extracted features. The windowing process then segments the continuous signal into manageable frames while minimising spectral distortions. The pre-emphasis filter in Step 1.1 of Algorithm 1 balances the spectral distribution of speech and audio signals. Natural audio signals typically have more energy at lower frequencies, which can mask critical high-frequency components. The pre-emphasis filter applies a simple high-pass characteristic to compensate for this imbalance, making subsequent spectral analysis more effective. The choice of window type in Step 1.3 significantly impacts the quality of spectral analysis. The Hamming window balances frequency resolution and spectral leakage, making it suitable for most AEDC applications. The Hanning window offers smoother spectral transitions, while the Blackman window provides the lowest spectral leakage at the cost of reduced frequency resolution.

2.2.2. Phase 2: spectral mapping

The second phase transforms the windowed frames into the frequency domain and applies the Mel filter bank (Algorithm 2). This phase sits at the heart of the Mel approach. It converts the linear frequency scale to the Mel scale, which better matches how the human ear processes sound.

The power spectrum computation in Step 2.1 of Algorithm 2 shows how energy is spread across the frequency range. Different acoustic events tend to have their own distinct energy patterns, which is what makes this step useful for telling them apart. The Mel scale conversion in Step 2.2 transforms the linear frequency scale into the Mel scale, closely matching human auditory perception. This transformation ensures that frequency bands equally crucial for human hearing receive equal representation in the feature space. The filter bank application in Step 2.3 applies the constructed Mel filters to the power spectrum. The resulting filter bank outputs represent the energy content in each Mel frequency band, forming the foundation for all subsequent Mel-based features.

2.2.3. Phase 3: feature extraction

The final phase converts the filter bank outputs into the desired feature representation. This phase offers three distinct feature types: MBE, LMBE, and MFCC, each with specific advantages for different AEDC applications. The phase also includes optional enhancements such as temporal derivatives and normalisation.

2.2.4. Temporal dynamics and normalisation

The temporal derivative computation in Step 3.2 (Algorithm 3) adds dynamic information to the feature set. First-order derivatives (Δ features) capture how spectral characteristics change over time, which is crucial for distinguishing between events with similar spectral content but different temporal patterns. For instance, a door closing and a book dropping might have similar frequency content, but the door closing typically shows a more gradual energy decay. Second-order derivatives (Δ^2 features) capture the acceleration of spectral changes, providing information about how quickly the rate of change is changing. The normalisation step in Step 3.3 (Algorithm 3) brings feature values onto the same scale regardless of recording conditions. This matters because without it, differences in microphone quality, recording distance, and room acoustics can throw off the results. For instance, a system trained on close-microphone recordings might fail when tested with distant-microphone recordings, even if the acoustic events are the same.

Algorithm 2: Spectral analysis and Mel filter bank construction.

Input: Windowed frames $\{\mathbf{x}_w^{(t)}\}_{t=0}^{T-1}$, NF M, frequency bounds $[f_{\min}, f_{\max}]$, filter type F_{type}
Output: Mel filter bank outputs $\{\mathbf{S}^{(t)}\}_{t=0}^{T-1}$

```
// Step 2.1: Fourier Transform and Power Spectrum
```
for $t \leftarrow 0$ **to** $T-1$ **do**
- $\mathbf{X}^{(t)} \leftarrow \mathrm{FFT}(\mathbf{x}_w^{(t)})$
- $P^{(t)}[k] \leftarrow \frac{|\mathbf{X}^{(t)}[k]|^2}{N_{\text{frame}}}$

```
// Step 2.2: Mel Scale Conversion
```
$m_{\min} \leftarrow 1127 \ln\left(1 + \frac{f_{\min}}{700}\right)$
$m_{\max} \leftarrow 1127 \ln\left(1 + \frac{f_{\max}}{700}\right)$
for $m \leftarrow 0$ **to** $M+1$ **do**
- $mel[m] \leftarrow m_{\min} + \frac{m}{M+1}(m_{\max} - m_{\min})$
- $f_{center}[m] \leftarrow 700\left(e^{\frac{mel[m]}{1127}} - 1\right)$
- $bin[m] \leftarrow \lfloor \frac{f_{center}[m]\, N_{\text{frame}}}{f_s} \rfloor$

```
// Step 2.3: Filter Bank Application
```
for $m \leftarrow 1$ **to** M **do**
- **for** $t \leftarrow 0$ **to** $T-1$ **do**
  - $S^{(t)}[m] \leftarrow 0$ **for** $k \leftarrow 0$ **to** $\frac{N_{\text{frame}}}{2}$ **do**
    - **if** F_{type} = Triangular **then**
      - **if** $bin[m-1] \le k \le bin[m]$ **then**
        - $H_m[k] \leftarrow \frac{k - bin[m-1]}{bin[m] - bin[m-1]}$
      - **else if** $bin[m] \le k \le bin[m+1]$ **then**
        - $H_m[k] \leftarrow \frac{bin[m+1] - k}{bin[m+1] - bin[m]}$
      - **else**
        - $H_m[k] \leftarrow 0$
    - **else if** F_{type} = Gaussian **then**
      - $H_m[k] \leftarrow \exp\left(-\frac{(k - bin[m])^2}{2\sigma^2}\right)$
    - $S^{(t)}[m] \mathrel{+}= P^{(t)}[k] \cdot H_m[k]$

The AEDC performance depends on several key parameters requiring careful tuning:

- **Frame length** controls the trade-off between temporal and frequency resolution; longer frames provide better frequency resolution for harmonic analysis, while shorter frames capture rapid temporal changes.
- **OP** determines temporal smoothness between frames; higher values yield smoother transitions, but increase computational cost and redundancy.
- **NF** affects spectral resolution: more filters provide finer detail at an increased dimensionality cost.
- **Window type** influences spectral leakage, while **filter type** determines frequency response: triangular filters offer efficiency and sharp separation, whereas Gaussian filters provide smoother transitions.
- **Feature type** selection fundamentally affects representation characteristics and downstream processing requirements.

From the above parameters, feature type, NF, frame length, OP, and window type are considered in the current review as they directly impact AEDC performance across different acoustic conditions. The computational complexity of the complete Mel algorithm is $O(T \cdot N_{frame} \log N_{frame} + T \cdot M \cdot N_{frame}/2)$, where the first term corresponds to the FFT operations and the second term represents the filter bank computations (all parameters were defined in Algorithms 1–3). This complexity makes the algorithm suitable for real-time AEDC applications while maintaining high feature quality. The linear relationship with M indicates that increasing spectral resolution has a predictable computational cost, allowing system designers to balance performance requirements with available processing power.

Algorithm 3: Feature computation and enhancement.

Input: Filter bank outputs $\{\mathbf{S}^{(t)}\}_{t=0}^{T-1}$, feature type $Feat_{type}$, number of cepstral coefficients C, normalization flag, small constant ϵ
Output: Final feature matrix $\mathbf{F}_{final}$

```
// Step 3.1: Primary Feature Extraction
```
switch $Feat_{type}$ **do**
- **case** *MBE* **do**
  - **for** $m \leftarrow 1$ **to** M **do**
    - $MBE[m] \leftarrow \frac{1}{T}\sum_{t=0}^{T-1} S^{(t)}[m]$ `// Band energy averaging`
  - $\mathbf{F} \leftarrow [MBE[1], \ldots, MBE[M]]^\top$
- **case** *LMBE* **do**
  - **for** $m \leftarrow 1$ **to** M **do**
    - $\overline{E}_m \leftarrow \frac{1}{T}\sum_{t=0}^{T-1} S^{(t)}[m]$
    - $LMBE[m] \leftarrow \ln(\overline{E}_m + \epsilon)$ `// Log of band energy`
  - $\mathbf{F} \leftarrow [LMBE[1], \ldots, LMBE[M]]^\top$
- **case** *MFCC* **do**
  - **for** $t \leftarrow 0$ **to** $T-1$ **do**
    - **for** $m \leftarrow 1$ **to** M **do**
      - $L^{(t)}[m] \leftarrow \ln(S^{(t)}[m] + \epsilon)$
    - **for** $c \leftarrow 1$ **to** C **do**
      - $C^{(t)}[c] \leftarrow \sum_{m=1}^{M} L^{(t)}[m] \cos\left(\frac{\pi c}{M}(m - \frac{1}{2})\right)$
    - $\mathbf{F}^{(t)} \leftarrow \mathbf{C}^{(t)}$

```
// Step 3.2: Temporal Derivative Computation (Optional)
```
if *Dynamic features required* **then**
- **for** $t \leftarrow 1$ **to** $T-2$ **do**
  - $\Delta\mathbf{F}^{(t)} \leftarrow \mathbf{F}^{(t+1)} - \mathbf{F}^{(t-1)}$
- **for** $t \leftarrow 2$ **to** $T-3$ **do**
  - $\Delta^2\mathbf{F}^{(t)} \leftarrow \Delta\mathbf{F}^{(t+1)} - \Delta\mathbf{F}^{(t-1)}$
- $\mathbf{F}_{combined} \leftarrow [\mathbf{F}; \Delta\mathbf{F}; \Delta^2\mathbf{F}]$

else
- $\mathbf{F}_{combined} \leftarrow \mathbf{F}$

```
// Step 3.3: Feature Normalisation (Optional)
```
if *Normalization required* **then**
- **for** *each feature dimension* d **do**
  - $\mu_d \leftarrow \frac{1}{\mathbf{T}}\sum_t \mathbf{F}_{combined}[d,t]$
  - $\sigma_d \leftarrow \sqrt{\frac{1}{\mathbf{T-1}}\sum_t (\mathbf{F}_{combined}[d,t] - \mu_d)^2}$
  - $\mathbf{F}_{final}[d,:] \leftarrow \frac{\mathbf{F}_{combined}[d,:] - \mu_d}{\sigma_d}$

else
- $\mathbf{F}_{final} \leftarrow \mathbf{F}_{combined}$

return $\mathbf{F}_{final}$

3. AEDC methods based on Mel-based features

This section analyses the 373 AEDC studies identified through the systematic search, organised by the type of Mel feature used. For each feature, the four key parameters (NF, frame length, OP, and window type) are examined through frequency distributions and temporal trends, and the classifier architectures are catalogued. The detailed study level tables are provided in Appendix B. AEDC methods that use Mel-based features can be divided into categories based on the type of Mel feature. Fig. 6 shows the usual Mel-based features researchers examined in their AEDC research. The Cochleagram, Mean magnitude spectrogram, Mel-fgram, Polygram Mel-scaled, SALSA, and Weighted Mel spectrogram have also been used, but rarely, and discarded in the current review.

As shown in Fig. 6, the LMS feature stands out significantly from the rest. It was used in 237 research studies, much more than any other

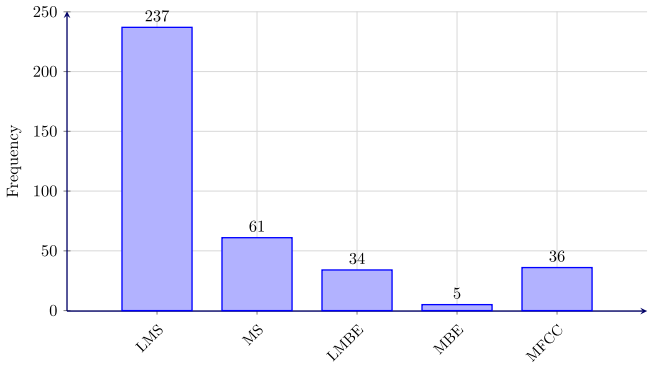


Fig. 6. Frequency of each feature in AEDC methods.

feature. Conversely, MBE had the lowest usage with only five studies. LMS's popularity in AEDC research is derived from its comprehensive spectral characterisation. It gives complete details of sound characteristics, and most of the critical information needed to identify acoustic events can be extracted from this feature. The two main advantages of LMS are its logarithmic processing, which smooths out the variations in output, and its two-dimensional (2D) structure, making it compatible with most deep learning frameworks that researchers have commonly used in recent years. Since LMS, MS, LMBE, and MFCC features were frequently used in the studies, the current review analyzes them in different time intervals. This makes it easier to look at the results and see what patterns emerge. For each study, four key parameters that matter for all Mel-based features:

- NF;
- Window Type;
- OP; and
- Frame Length;

were considered. The type of classifier used in each AEDC study is also considered. The NF directly affects the number of features the system processes and the computational cost. When NF is increased, more detailed frequency information is extracted from the input sound. The second parameter is frame length. The frame length affects how quickly the system can detect events and how much noise it introduces. The frame length and sampling frequency determine the number of samples in feature extraction. The frame length and the OP also influence how much training and testing data you can get from each sound file.

The frame length and OP need some explanation, as they were reported differently. Frame length can be reported as the number of samples or time duration. In the current study, when Frame length was reported as the number of samples, it was converted to time in "ms" using the known sampling frequency. The OP has even more variety. Studies use two different approaches: Hop length and OP. Hop length refers to the step size between frames, while OP, as an inverse parameter, shows how much consecutive frames overlap. Both can be reported as a percentage of frame length, as several samples, or as a time duration. All these parameters were reported as percentages of the frame length to keep everything consistent. The following equation is used to convert the Hop length to OP:

$$\mathrm{OP} = \left(1 - \frac{\text{Hop length}}{\text{Frame length}}\right) \times 100. \tag{7}$$

For classifiers, the acronyms of each classifier that are fully explained in the acronym list at the beginning of this article were used.

3.1. Log Mel spectrogram

LMS has been the most widely used Mel-based feature in AEDC research. Because so many studies have used this feature, its parameters are separated into yearly periods. The full list of all 237 LMS studies with their reported parameters is provided in Appendix B. (Tables A.1 through A.7). The figures and the accompanying analysis below illustrate how each parameter is distributed across the selected studies and how these patterns have evolved over time. The values of NF, as the first analysed parameters, were divided into five intervals, shown in Fig. 7 (part A). For each interval, the frequency of AEDC methods is calculated. When a study did not report the NF value, it was categorised as "Unknown". The results in Fig. 7 show that most studies using LMS fell into two [128–255] and [64–127] intervals.

The least selected was NF below 32, which appeared in only a single study. The dominance of the [64-127] and [128-255] NF intervals in LMS-based AEDC systems is closely related to the spectral resolution requirements of typical acoustic events. Increasing NF effectively increases the number of triangular filters applied along the Mel scale, resulting in finer partitioning of the frequency axis. This is particularly important when distinguishing events with overlapping harmonic structures, such as machinery noise, urban sound mixtures, or animal vocalisations with closely spaced spectral components. When NF is below 32, the Mel filterbank becomes too coarse to resolve harmonics that are often separated by only tens of Hertz. In such cases, distinct spectral signatures may collapse into the same filter response, reducing discriminative power. The near absence of NF $<$ 32 in LMS-based studies, therefore, reflects a practical constraint rather than a stylistic choice. To provide a clearer comparison, Fig. 8 presents the relative frequency of NF across different years.

Fig. 8 (parts A and B) shows that the most common NF value in 2020 fell in the [64–127] interval. In later years, NF increased as deep learning algorithms became more powerful at handling higher-dimensional inputs. Research focused on the [64–127] and [128–255] ranges. The preference for the [128-255] range in more recent studies also aligns with the increasing adoption of CNNs, which benefit from richer 2D spectral representations. A higher NF provides more detailed spectral maps that convolutional filters can exploit spatially. However, this improvement comes at the cost of increased computational complexity and memory usage, which explains why moderate NF values [64-127] remain widely used in resource-constrained or real-time applications. These findings suggest that NF selection is implicitly shaped by a balance between spectral discrimination capability, classifier architecture compatibility, and computational constraints, rather than in an arbitrary manner. The gradual shift towards higher NF values over the 2020-2025 period further indicates a methodological transition driven by deep learning architectures rather than changes in acoustic event characteristics. About 10% of the found studies in this field do not specify this parameter, making it nearly impossible to reproduce the corresponding study and verify its results, which is a real weakness since NF is considered the most important parameter for Mel-based features.

The second analysed parameter is frame length, Fig. 9. Considering the information reported in Tables A.1 through A.7 (Appendix B.), six intervals were selected to analyse frame length. Also, the "unknown" category was defined for studies that did not report Frame length. The results show that the [32–64) ms was the most common choice for frame length. [32–64) shows as a closed-open interval, meaning 32 is included in the interval, but 64 is not. The lowest frequency for Frame length is in intervals $\geq 256\ ms$ and $< 20\ ms$. These two extremes highlight the trade-off between response speed and the number of samples available for processing. In Frame length $\geq 256\ ms$, too many samples require more computational power in the feature extraction and classification steps. The system responds very quickly in intervals $< 20\ ms$, but there are not enough samples to properly process and recognise event patterns, so the results will not be reliable. The concentration in the [32–64) ms range suits most acoustic events studied with LMS. This range captures enough cycles for stable spectral estimation of sounds like car horns or alarms, whilst remaining short enough to track temporal changes. Impulsive events lasting under 150 ms would be averaged out by frames longer than 128 ms, explaining the low usage of that range. For a more de-

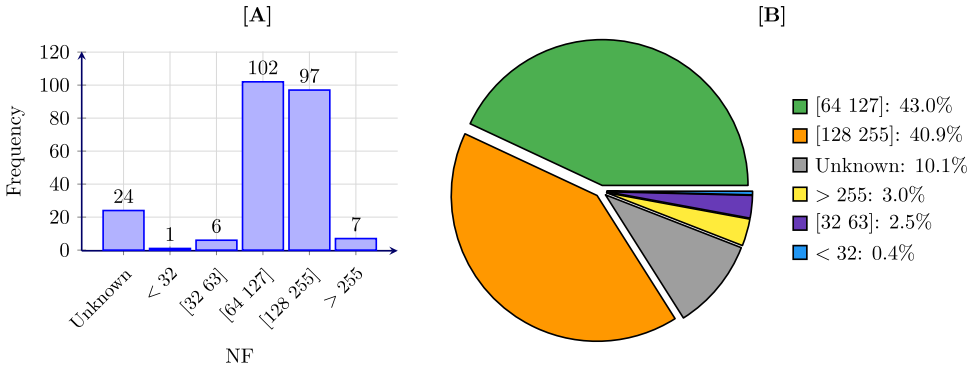
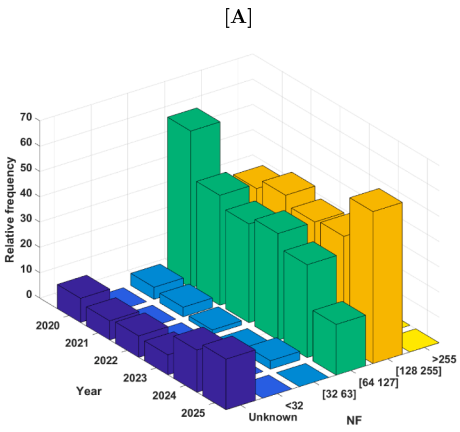


Fig. 7. NF in LMS-based AEDC methods: (A) Frequency and (B) Relative frequency.



Year	NF						Total
	Unknown	<32	[32–63]	[64–127]	[128–255]	>255	
2020	4	0	2	26	10	0	42
2021	4	0	2	23	22	2	53
2022	5	0	1	23	27	3	59
2023	3	1	0	16	16	2	38
2024	5	0	1	11	13	0	30
2025	3	0	0	3	9	0	15
Total	24	1	6	102	97	7	**237**

Fig. 8. Changes in NF frequency over time in LMS-based AEDC methods: [A] 3D view of yearly distribution patterns and [B] Frequency table of NF over time.

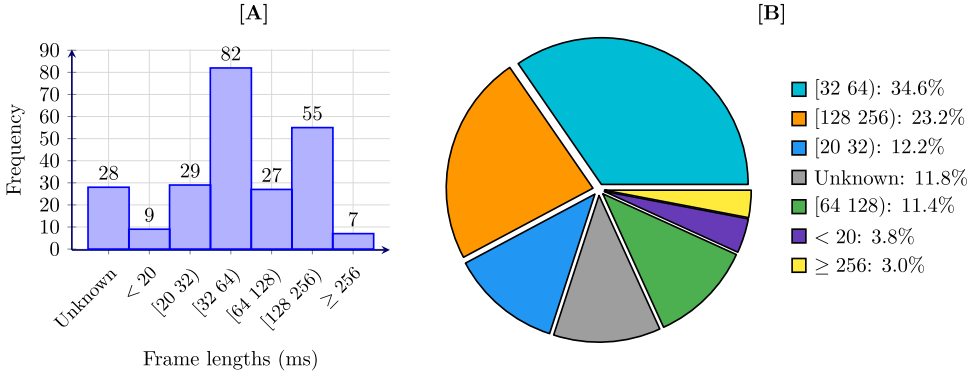


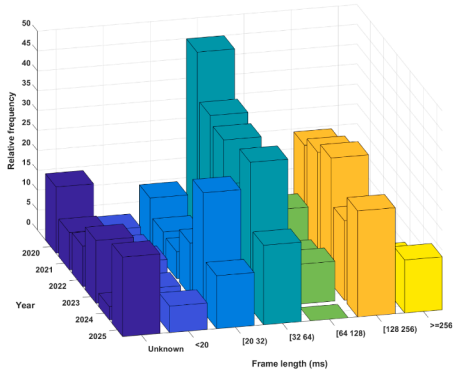
Fig. 9. Frame lengths (ms) in LMS-based AEDC methods: [A] Frequency and [B] Relative frequency.

tailed comparison between the different Frame length intervals, Fig. 10 shows the relative frequency of this parameter in other years.

Based on the statistics reported in Fig. 10 (parts A and B), the [128–256) ms interval was also important, although it appeared less frequently than the [32–64) ms range. Another significant finding is that about 11.8% of studies did not report Frame length (Fig. 9). This makes it practically impossible to reimplement the study and verify its results. As a summary, the dominance of the [32-64) ms interval reflects the fundamental trade-off between temporal and spectral resolution inherent in short-time spectral analysis. Shorter frames (< 32 *ms*) provide improved temporal localisation and may benefit the detection of very brief impulsive events such as gunshots or mechanical clicks. However, they reduce frequency resolution, making it more difficult to separate closely spaced harmonics or tonal components. Conversely, longer frames (> 64 *ms*) improve frequency discrimination but introduce temporal smearing, potentially blurring the onset and offset boundaries of short-duration events. The widespread adoption of the [32-64) ms range suggests, therefore, that researchers implicitly converge on a compromise configuration that captures at least one full pitch cycle for most acoustic events above approximately 30 Hz while preserving sufficient temporal precision for practical AEDC applications. This indicates that the observed consensus is not arbitrary but grounded in signal-processing constraints rather than mere convention.

Moreover, this pattern is consistent across both spectrogram-based and cepstral features, indicating a shared methodological assumption in the field. The third parameter, which affects the number of training samples and the system's inherent accuracy in detecting start and end times of events, is OP. Reducing OP improves the accuracy of detecting when events start and end. However, reducing this parameter too much, especially at the beginning and end of events, results in the acoustic

[A] [B]



Year	Frame length (ms)							Total
	Unknown	<20	[20–32)	[32–64)	[64–128)	[128–256)	≥256	
2020	7	2	5	20	6	2	0	42
2021	5	2	4	19	6	14	3	53
2022	6	3	4	20	9	17	0	59
2023	6	1	5	9	3	12	2	38
2024	1	0	9	11	3	6	0	30
2025	3	1	2	3	0	4	2	15
Total	28	9	29	82	27	55	7	**237**

Fig. 10. Frame length distribution over time in LMS-based AEDC methods: [A] 3D visualisation of yearly frame length patterns and [B] Frequency table of Frame length over time.

[A] [B]

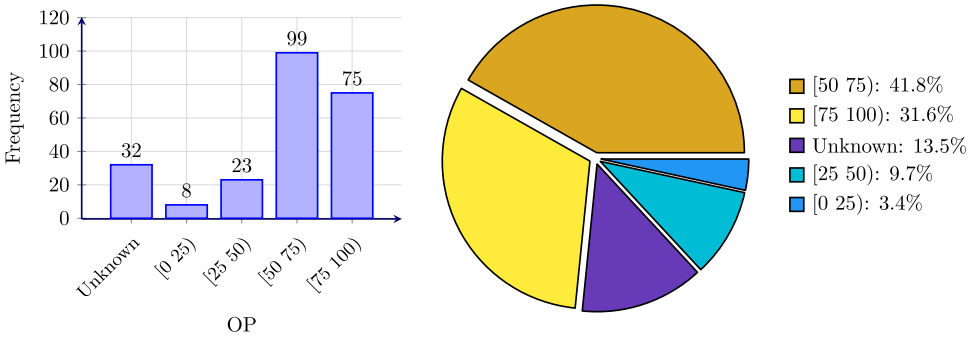
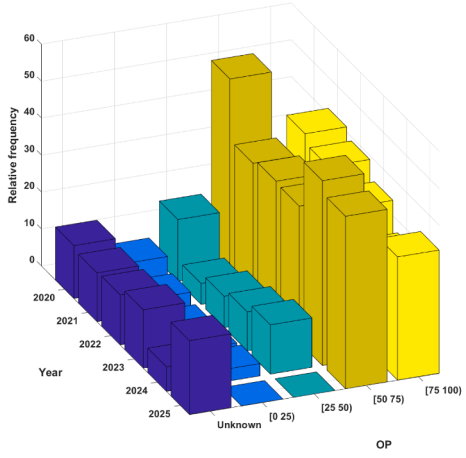


Fig. 11. OP in LMS-based AEDC methods: [A] Frequency and [B] Relative frequency.

[A] [B]



Year	OP					Total
	Unknown	[0–25)	[25–50)	[50–75)	[75–100)	
2020	6	3	7	22	4	42
2021	7	2	3	19	22	53
2022	8	1	5	22	23	59
2023	6	1	4	14	13	38
2024	2	1	4	15	8	30
2025	3	0	0	7	5	15
Total	32	8	23	99	75	**237**

Fig. 12. OP distribution over time in LMS-based AEDC methods: [A] 3D visualisation of yearly OP patterns and [B] Frequency table of OP over time.

event occupying a smaller share of the total frame time; consequently, increasing the probability of the AEDC error. Based on Fig. 11, this parameter was selected to be greater than 50% of the frame length in most studies.

The preference for OP above 50% makes practical sense because higher overlap creates smoother transitions in the spectrogram, helping convolutional layers learn temporal patterns. The [75–100) range was particularly common in studies dealing with longer duration events like machinery sounds, where dense temporal sampling captures gradual spectral changes. To analyse how this variable has changed over time, its frequency by year is shown in Fig. 12 (parts A and B). Fig. 12 shows that the relative frequency of OP has not changed much over time. The most common choices have always been in the [50–75) and [75–100) ranges. What is interesting is the relatively high number of studies that

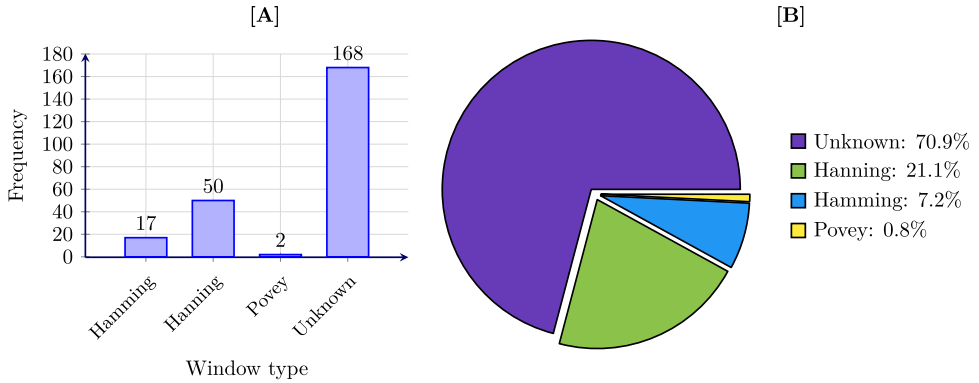


Fig. 13. Window types in LMS-based AEDC methods: [A] Frequency and [B] Relative frequency.

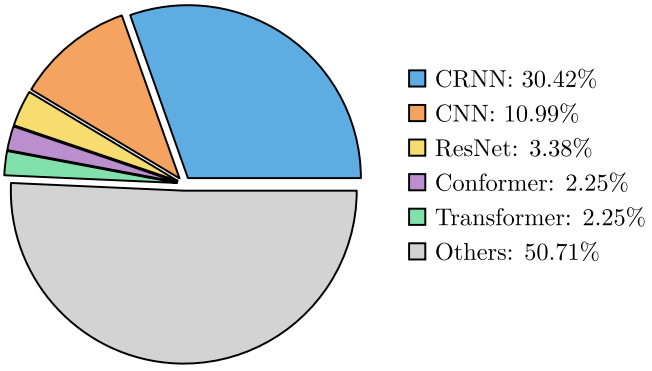


Fig. 14. Percentage of the found classification methods with relative frequency >1% in LMS-based AEDC methods.

Table 1
Predominant classifiers in the found LMS-based AEDC methods over time.

Model	2020	2021	2022	2023	2024	2025
CRNN	*	*	*	*	*	*
CNN	*	*	*	*	*	
DNN	*					
Transformer	*		*		*	
Conformer		*				
CapsNet			*			
ResNet			*	*	*	
FDY				*		

did not report this parameter (13.5% according to Fig. 11). This parameter directly affects the number of training and testing samples, so knowing it is crucial for verifying research results.

Based on Fig. 13, the window types used in LMS-based AEDC methods were mainly Hamming or Hanning. Hanning has been used about three times as often as Hamming. The higher preference for Hanning in LMS is likely because its smoother spectral transition produces cleaner spectrogram representations for convolutional classifiers. Hamming gives slightly better frequency resolution but higher sidelobe levels, which can introduce spectral artefacts into the spectrogram. Another finding from Fig. 13 is the high number of studies (71%) that did not specify the window type. This high rate of non-reporting suggests that many researchers simply accepted software library defaults without considering how window response affects feature quality. Checking the frequency is not straightforward for classifiers because most are built by combining multiple layers or by modifying existing networks.

In this study, the classifier architecture reported in each article was manually extracted and recorded in a structured database. String matching was then applied to separate the names of constituent layers from composite architectures (e.g., extracting “CNN” and “RNN” from “CRNN”) and to standardise naming across studies. The dominant approaches were identified by counting the most frequently repeated structures, and the final results were manually verified against the original articles to correct any extraction errors. Fig. 14 shows the percentage of different structures that appeared with a relative frequency higher than 1%.

As shown in Table 1, CRNN was the most frequently used classifier in the LMS-based AEDC methods. After CRNN, three structures, ResNet, Transformer, and CNN, were among the most repeated classifiers used over different periods.

3.2. Mel spectrogram

Mel spectrograms were the second-most frequently used feature in AEDC research. The full list of all 61 MS studies with their reported parameters is provided in Appendix B. (Table A.8 for studies from 2020 to 2022, and Table A.9 for studies from 2023 to 2025). Fig. 15 shows that in MS, the highest frequency for the NF parameter was similar to LMS in the [128–255] and [64–127] ranges, although the frequency difference between these two intervals is much greater than LMS. In MS, about 10% of the total studies did not specify the NF. The other intervals show a significant difference in frequency compared to these two. For a more detailed comparison between different intervals, Fig. 16 shows the relative frequency of NF for MS over time.

Fig. 16 (parts A and B) demonstrates that the highest frequency of NF falls in the [64–127] interval only in 2020. The [128–255] interval had a higher frequency in subsequent years. The faster shift towards higher NF in MS compared to LMS is worth noting. Because MS keeps the linear amplitude scale without logarithmic compression, smaller differences in spectral energy between overlapping events are harder to distinguish. Using more filters compensates for this by providing finer frequency resolution, which helps the classifier detect subtle spectral differences that the logarithmic scaling in LMS would naturally amplify. This explains why MS researchers moved to the [128–255] range earlier than their LMS counterparts.

According to Fig. 17, the frequency pattern of MS is very similar to LMS; for example, the highest frequency for Frame length occurred in the middle intervals of the table. Also, some differences in relative frequency can be observed: in MS, about 18% of studies did not report the frame length value, whereas in LMS this value was only 12%. Fig. 18 (parts A and B) shows the relative frequency and frequency of frame length over time.

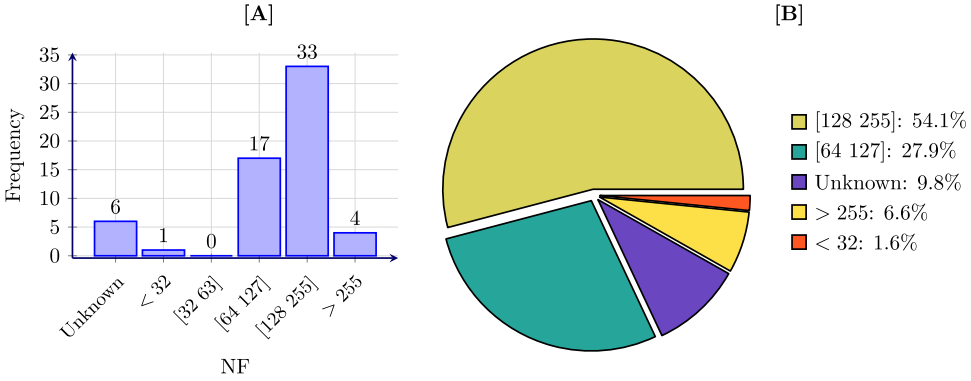
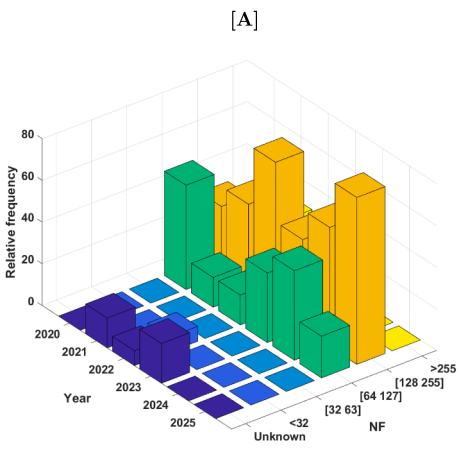


Fig. 15. NF in MS-based AEDC methods: [A] Frequency and [B] Relative frequency.



Year	NF						Total
	Unknown	<32	[32–63]	[64–127]	[128–255]	>255	
2020	0	0	0	3	2	1	6
2021	1	0	0	1	3	2	7
2022	1	1	0	2	10	0	14
2023	4	0	0	7	9	1	21
2024	0	0	0	3	5	0	8
2025	0	0	0	1	4	0	5
Total	**6**	**1**	**0**	**17**	**33**	**4**	**61**

Fig. 16. Changes in NF frequency over time in MS-based AEDC methods: [A] 3D view of yearly distribution patterns and [B] Frequency table of NF over time.

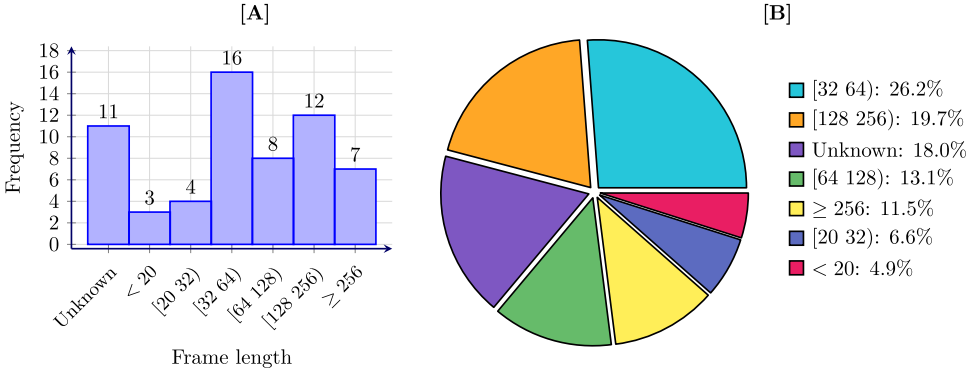
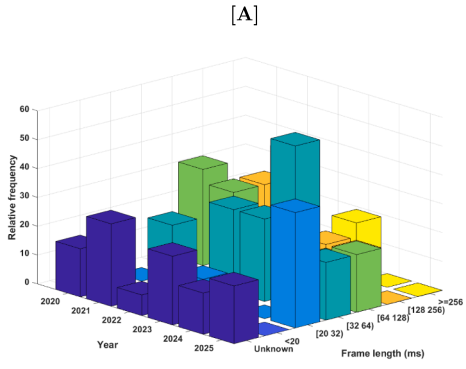


Fig. 17. Frame lengths (ms) in MS-based AEDC methods: [A] Frequency and [B] Relative frequency.



Year	Frame length (ms)							Total
	Unknown	<20	[20–32)	[32–64)	[64–128)	[128–256)	≥256	
2020	1	1	0	1	2	0	1	6
2021	2	0	0	0	2	2	1	7
2022	1	1	1	4	0	6	1	14
2023	5	1	1	6	2	2	4	21
2024	1	0	0	4	1	2	0	8
2025	1	0	2	1	1	0	0	5
Total	**11**	**3**	**4**	**16**	**8**	**12**	**7**	**61**

Fig. 18. Frame length distribution over time in MS-based AEDC methods: [A] 3D visualisation of yearly frame length patterns and [B] Frequency table of Frame length over time.

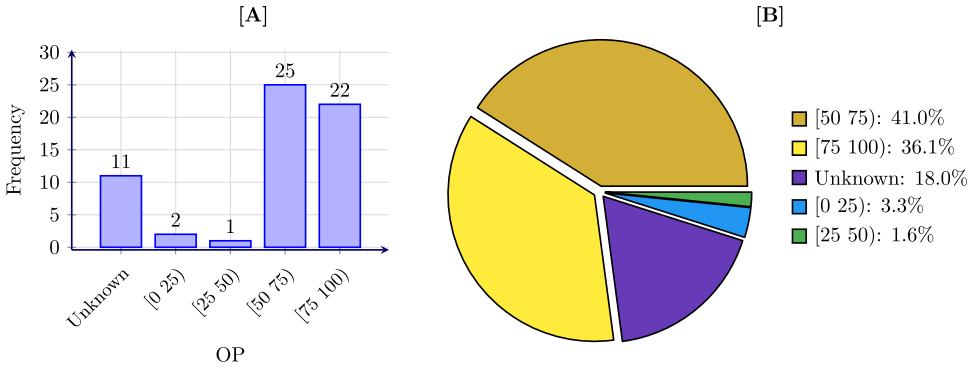
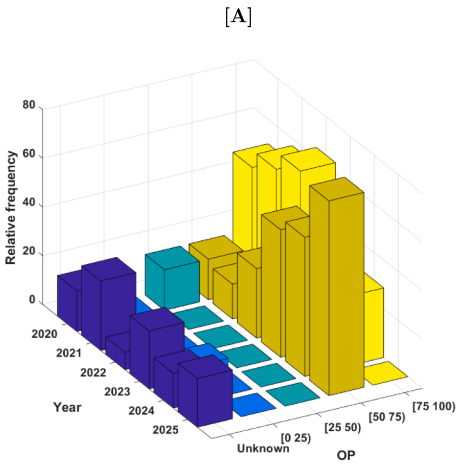


Fig. 19. OP in MS-based AEDC methods: [A] Frequency and [B] Relative frequency.



Year	OP					Total
	Unknown	[0–25)	[25–50)	[50–75)	[75–100)	
2020	1	0	1	1	3	6
2021	2	0	0	1	4	7
2022	1	0	0	4	9	14
2023	5	2	0	11	3	21
2024	1	0	0	4	3	8
2025	1	0	0	4	0	5
Total	**11**	**2**	**1**	**25**	**22**	**61**

Fig. 20. OP distribution over time in MS-based AEDC methods: [A] 3D visualisation of yearly OP patterns and [B] Frequency table of OP over time.

According to the results of Fig. 18, it can be seen that in 2020 and 2021, the interval [64 128) was more frequent, and in recent years, shorter Frame lengths, such as [32 64), were considered; thus, the time trend was for the frame length to become shorter. This shift towards shorter frames in MS likely reflects a growing focus on detecting events with rapid onset characteristics, such as impacts or alarms. Without the logarithmic compression that LMS applies, the raw spectral energy values in MS are more sensitive to sudden amplitude changes. Shorter frames let the system capture these sharp transitions more precisely, rather than averaging them out over a longer window. The higher proportion of studies not reporting frame length in MS (18%) compared to LMS (12%) also suggests that MS is more often used with pretrained models or standard pipelines where the frame length is set internally and not always documented by the researchers.

The OP in MS-based AEDC methods is shown in Fig. 19. Like LMS, OP was selected to be greater than 50% of the frame length in most studies. The similarity in OP patterns between MS and LMS is expected because both features produce two-dimensional spectral representations that are processed in much the same way by convolutional classifiers. The temporal resolution requirements for tracking how sounds evolve do not change just because the amplitude scaling is different. What is slightly different in MS is the higher proportion of studies in the [75–100) range compared to LMS (36.1% vs 31.6%, Figs. 19 and 11). Since MS preserves the linear amplitude scale, neighbouring frames with high overlap provide the classifier with a more gradual view of how the raw energy distribution shifts over time, which can be helpful when the system needs to track slow spectral changes in events like continuous machinery noise.

Fig. 20 indicates that the temporal trend of OP in MS also has a very high similarity to LMS. Fig. 21 shows that the window types used in MS-based AEDC methods were mainly Hamming or Hanning. In the MS feature, Hanning also indicates a significant difference in frequency compared to Hamming. The preference for Hanning in MS follows the same logic as in LMS: its smoother spectral transitions produce cleaner representations for convolutional classifiers.

However, the slightly higher proportion of Hamming usage in MS compared to LMS (11.5% vs 7.2%, Figs. 21 and 13) may reflect that some MS studies prioritise preserving finer spectral detail, since MS already lacks the smoothing effect that logarithmic compression provides in LMS. In those cases, Hamming's narrower main lobe gives a small advantage in frequency resolution. The unreported rate for window type in MS (66%) is slightly lower than in LMS (71%), but still very high.

Table 2 and Fig. 22 show the temporal pattern of different classifiers used and the percentage of repetition for different classifiers with relative frequency higher than 1% in MS-based AEDC systems. Like LMS, CRNN, and CNN, Conformer was among the most frequently used classifiers. The dominance of CRNN across both LMS and MS is not surpris-

Table 2
Predominant classifiers in MS-based AEDC methods found over time.

Model	2020	2021	2022	2023	2024	2025
CRNN	*	*	*	*		
CNN		*				
Conformer				*		
EINV2				*		
STRFNet				*		

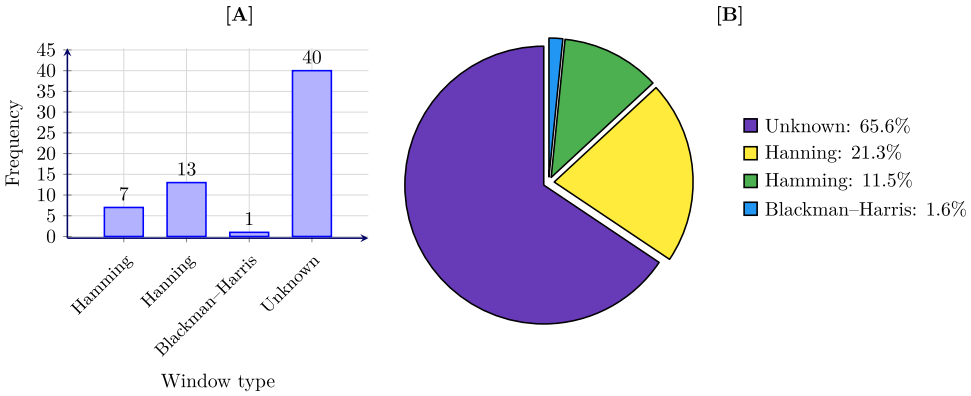


Fig. 21. Window types in MS-based AEDC methods: [A] Frequency and [B] Relative frequency.

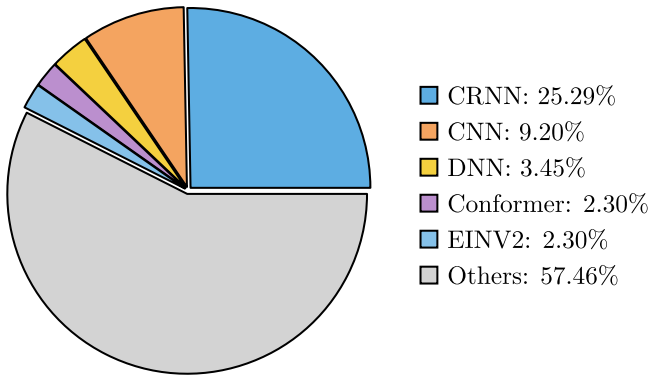


Fig. 22. Percentage of classification methods with relative frequency >1% in MS-based AEDC methods.

Table 3
The frequency of NF in LMBE-based AEDC methods found over time.

Year	NF						Total
	Unknown	< 32	[32–63]	[64–127]	[128–255]	> 255	
2020	1	0	4	5	1	1	12
2021	0	0	2	3	1	0	6
2022	0	0	2	6	0	0	8
2023	0	0	1	3	1	0	5
2024	0	1	0	0	2	0	3
2025	0	0	0	0	0	0	0
Total	1	1	9	17	5	1	**34**

ing, given that both features produce spectral representations where the convolutional layers handle the frequency patterns, and the recurrent layers track how those patterns change over time. The appearance of EINV2 in the MS classifier list (Table 2) but not in LMS is notable, as EINV2 was specifically designed for sound event detection tasks where preserving raw amplitude relationships, which MS provides, can benefit the detection of overlapping events.

3.3. Log Mel band energies

Due to the smaller number of AEDC studies using LMBE compared to LMS and MS, the full list of all 34 LMBE studies with their reported parameters is provided in Appendix B. (Table A.10). Fig. 23 shows that the highest frequency for the NF parameter was in the [64–127] range, and the two intervals [128–255] and [32–63] also had significant frequencies. The main difference between spectrogram-based (LMS and MS) features, and this feature is the increased frequency of the [32–63] range. The reason LMBE works acceptably with fewer filters is that this feature serves a different purpose. Unlike LMS or MS, which keep the full spectral detail within each frame, LMBE averages the energy across each Mel band. This averaging already smooths out fine spectral variations, so having lots of filters does not add as much useful information as it would for a spectrogram. For tasks like binary event detection, where the system only needs to tell whether a particular sound is present or absent rather than separate multiple overlapping events, 32 to 63 filters can capture enough of the broad frequency shape to get reliable results.

The number of methods with unspecified NF was also lower in this feature than in the previous two features, likely reflecting the smaller, more focused research community working with LMBE, where parameter choices tend to be more deliberate and better documented. Table 3 shows the frequency of LMBE-based AEDC studies found by year for each interval. The Frame length pattern in LMBE is also very different from spectrogram-based features.

Fig. 24 shows this parameter's frequency and relative frequency in LMBE-based AEDC systems. The highest frequency for Frame length was in the [32–64) range, with a much stronger concentration than in LMS or MS (50% in LMBE vs 34.6% in LMS, Figs. 24 and 9). This makes sense given what LMBE captures. Since this feature averages spectral energy within each band, it does not need the long frames that spectrogram features require for stable frequency estimation. Instead, shorter frames give LMBE better temporal resolution, which is more useful for this feature because it enables the system to track rapid changes in the energy distribution across the Mel bands. This is particularly valuable for detecting events with sharp onsets, such as alarms or impacts, where the energy profile shifts rapidly between consecutive frames. Additionally, the number of studies that did not report frame length for LMBE was much lower than in the previous two cases (5.9% vs 11.8% in LMS, Figs. 24 and 9).

Table 4 shows the frequency of Frame length in LMBE-based AEDC methods over time. The results show that in the years 2020 to 2022, frame length was usually selected at or below 64 ms in LMBE, but after 2022, the trend changed. The appearance of longer frames (128 ms and above) in more recent LMBE studies may reflect the growing use of this feature in combination with attention-based classifiers, which can handle longer input sequences and benefit from the broader temporal context that longer frames provide. In the OP parameter, the pattern also shows very significant differences from the LMS and MS, Fig. 25.

In LMBE, OP has a very high concentration in the [50–75) range (61.8% compared to 41.8% in LMS and 41.0% in MS, Figs. 25, 11, and 19). The much lower presence of the [75–100) range in LMBE (14.7% vs 31.6% in LMS) is a meaningful difference.

Because LMBE already smooths spectral content through band energy averaging, very high overlap would create heavily redundant frames with almost identical energy values.

This redundancy would not provide the classifier with much new information between consecutive frames and could even harm generali-

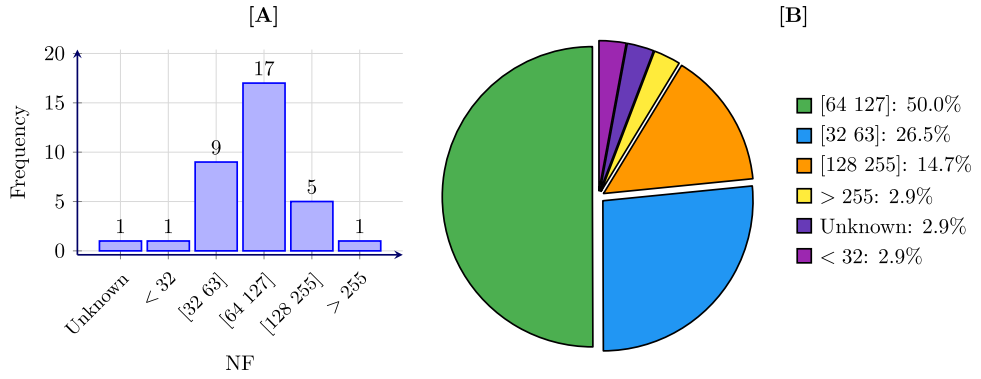


Fig. 23. NF in LMBE-based AEDC methods: [A] Frequency and [B] Relative frequency.

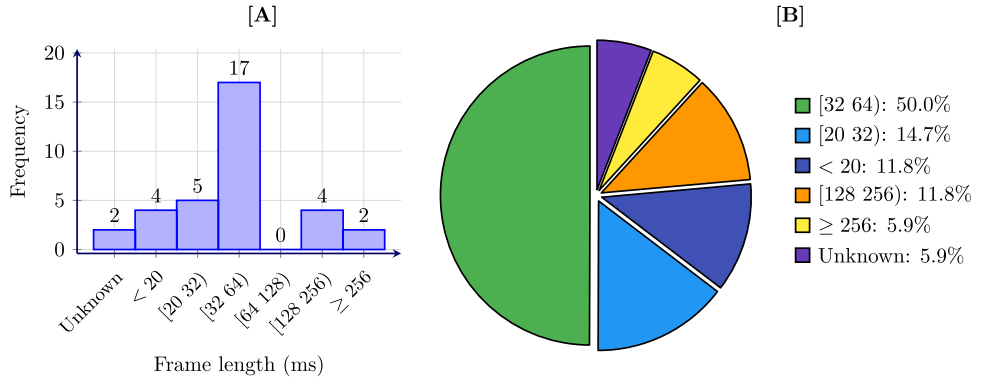


Fig. 24. Frame length (ms) distribution in LMBE-based AEDC methods: [A] Frequency and [B] Relative frequency.

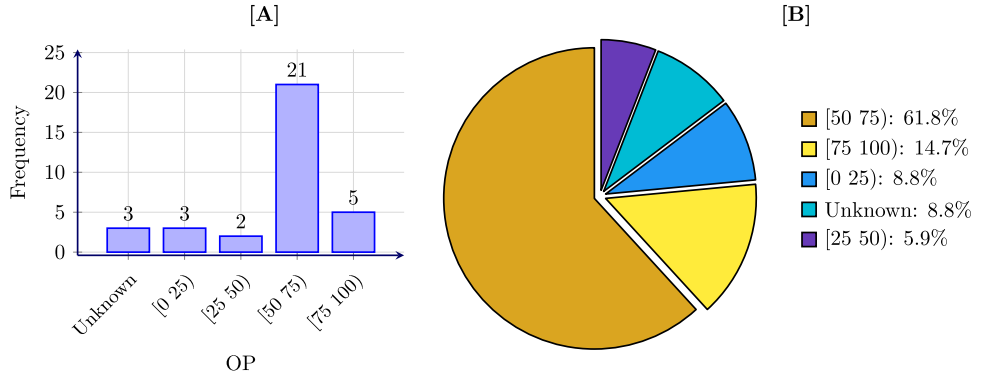


Fig. 25. OP in LMBE-based AEDC methods: [A] Frequency and [B] Relative frequency.

Table 4
The frequency of Frame length (ms) in LMBE-based AEDC methods found over time.

Year	Frame length (ms)							Total
	Unknown	< 20	[20–32)	[32–64)	[64–128)	[128–256)	≥ 256	
2020	1	3	3	5	0	0	0	12
2021	0	1	0	4	0	1	0	6
2022	1	0	1	6	0	0	0	8
2023	0	0	0	2	0	1	2	5
2024	0	0	1	0	0	2	0	3
2025	0	0	0	0	0	0	0	0
Total	2	4	5	17	0	4	2	**34**

Table 5
The Frequency of OP in LMBE-based AEDC methods found over time.

Year	OP					Total
	Unknown	[0–25)	[25–50)	[50–75)	[75–100)	
2020	1	2	1	8	0	12
2021	0	0	0	5	1	6
2022	1	1	0	5	1	8
2023	1	0	0	3	1	5
2024	0	0	1	0	2	3
2025	0	0	0	0	0	0
Total	3	3	2	21	5	**34**

sation by making the training data too repetitive. The moderate [50–75) range provides enough temporal density to track energy changes without excessive redundancy. The number of studies that did not report OP was much lower than the previous two features. Table 5 shows the yearly frequency of OP for LMBE-based studies.

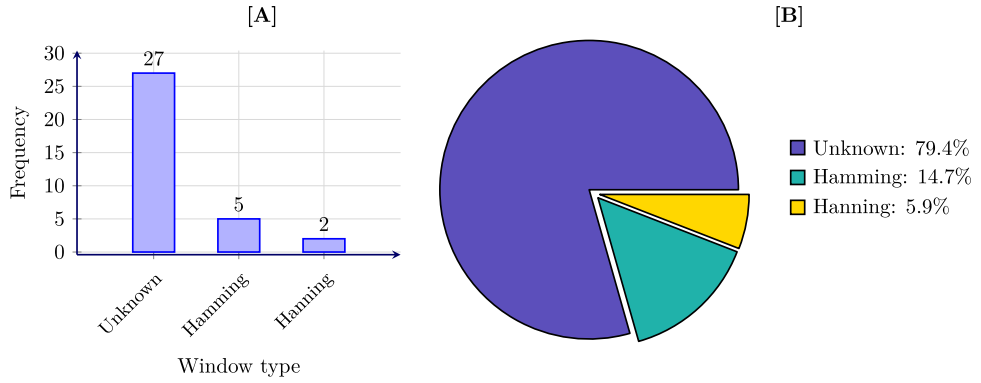


Fig. 26. Window types in LMBE-based AEDC methods: [A] Frequency and [B] Relative frequency.

Table 6
Predominant classifiers in LMBE-based AEDC methods found over time (shaded* = used).

Model	2020	2021	2022	2023	2024	2025
CRNN	*	*	*	*		
CNN	*					
FCN		*				
RNN	*			*		
DNN	*					

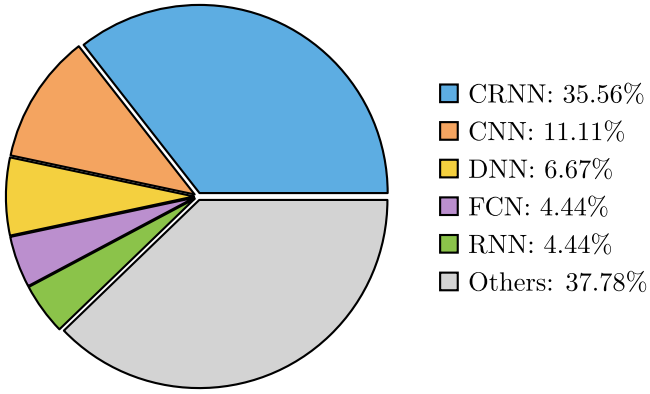


Fig. 27. Percentage of classification methods with relative frequency >1% in LMBE-based AEDC methods.

Fig. 26 indicates that the window types used in LMBE-based AEDC methods were also mainly Hamming or Hanning, with the difference that, in this feature, Hamming has a higher frequency than Hanning (14.7% vs 5.9%). This is the opposite pattern from LMS and MS, where Hanning was clearly preferred (Figs. 13 and 21). This reversal can be explained by the underlying mechanism of LMBE. Since LMBE sums up all the energy within each Mel band, sharp spectral peaks matter more than smooth sidelobe suppression. Hamming's narrower main lobe better preserves those peaks. In spectrogram features, the full spectral shape is kept, so Hanning's smoother transitions help the classifier. But in LMBE, where averaging already removes fine spectral detail, Hamming's sharper peak response helps the system locate the dominant energy. Also, 79% of studies did not specify the window type, which is the highest unreported rate among all features examined.

Table 6 shows predominant classifiers in LMBE-based AEDC methods over time. Fig. 27 shows the percentage of repetition for different classifiers with relative frequency higher than 1%. As to LMBE, Table 6 and Fig. 27 indicate that CRNN and CNN were the most frequently used classifiers. The higher proportion of simpler architectures, such as CNNs and DNNs, in LMBE compared to LMS and MS (Figs. 27, 14, and 22) aligns with the lower dimensionality of this feature. Since LMBE produces a compact representation with typically 40 to 64 values per frame, rather than the hundreds or thousands in a spectrogram, less complex classifiers can handle the data effectively, making LMBE a practical choice for systems with limited processing power.

3.4. Mel band energies

Mel-band energies are the least frequently investigated feature, with only five studies identified (Table A.11 in Appendix B.). In terms of time span, one study was conducted each year with this feature, and in 2021, no studies with this feature were found. For NF, the number 64 was used most frequently, in three studies. The frame length was mainly in the range [32–64) ms, and the Hamming window was the most common. In the OP, the numbers show high dispersion, but all but one study were 50% or higher. The classifier used in all five studies was CRNN. The Hamming function was used twice in these five studies, whereas the window type was not reported in the other three. Although the number of MBE studies is too small to draw strong conclusions, the patterns observed are consistent with what would be expected from this feature's characteristics. MBE represent the raw spectral energy in each Mel band without logarithmic compression, making it similar to MS but with reduced dimensionality (similar to the relationship between LMBE and LMS). The preference for NF of 64 and frame lengths in the [32–64) ms range mirrors the LMBE patterns (Figs. 23 and 24), which is logical since both features perform band energy averaging. The exclusive use of CRNN across all five studies suggests that the recurrent component is particularly useful for tracking how raw energy values evolve over time, compensating for the lack of spectral detail in this compact representation.

3.5. MFCC

The full list of all 36 MFCC studies with their reported parameters is provided in Appendix B. (Table A.12). For the NF parameter, the highest frequency was in the [64–127] range, and the 2nd widely used is the interval below 32 (Fig. 28). The NF frequency pattern in MFCC differs clearly from previous features, as it shows a high frequency below 32. This is the most distinctive NF pattern among the five features, and the reason lies in how MFCC works.

After applying the Mel filter bank, MFCC takes the logarithm and then applies the discrete cosine transform (DCT), which compresses the spectral information into a small set of cepstral coefficients. In practice, researchers typically keep only the first 12 to 20 coefficients regardless of how many Mel filters were used, because most of the useful spectral

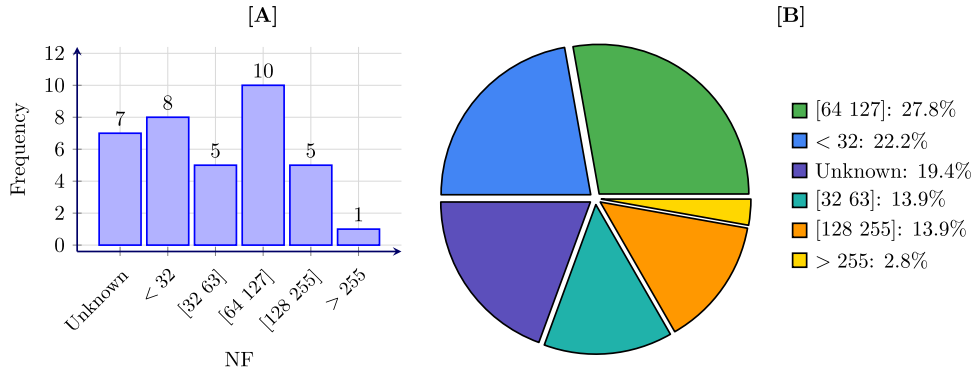


Fig. 28. NF in MFCC-based AEDC methods: [A] Frequency and [B] Relative frequency.

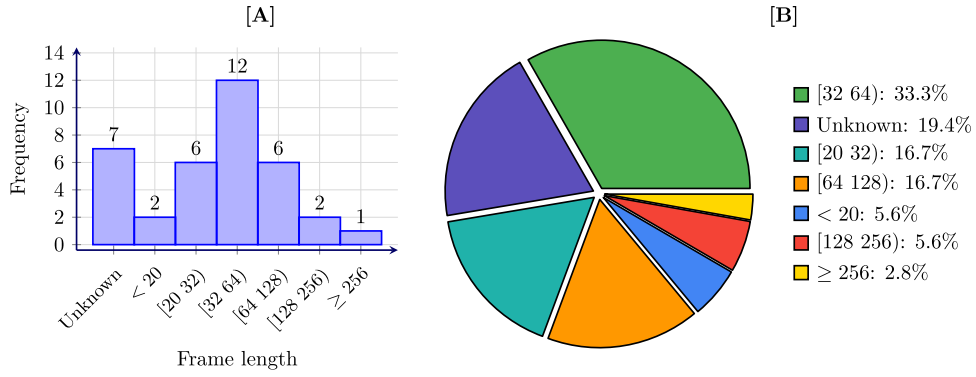


Fig. 29. Frequency of frame lengths (ms) in MFCC-based AEDC methods: [A] Frequency and [B] Relative frequency.

Table 7
The frequency of NF in MFCC-based AEDC methods found over time.

Year	NF						Total
	Unknown	< 32	[32–63]	[64–127]	[128–255]	> 255	
2020	2	2	1	2	0	0	7
2021	3	2	2	0	2	0	9
2022	1	1	0	4	2	1	9
2023	0	2	1	3	0	0	6
2024	1	0	1	1	1	0	4
2025	0	1	0	0	0	0	1
Total	7	8	5	10	5	1	**36**

Table 8
The frequency of Frame length (ms) in MFCC-based AEDC methods found over time.

Year	Frame length (ms)							Total
	Unknown	< 20	[20–32)	[32–64)	[64–128)	[128–256)	≥ 256	
2020	1	2	1	2	1	0	0	7
2021	2	0	1	3	2	0	1	9
2022	2	0	1	4	1	1	0	9
2023	2	0	1	2	0	1	0	6
2024	0	0	2	0	2	0	0	4
2025	0	0	0	1	0	0	0	1
Total	7	2	6	12	6	2	1	**36**

envelope information is concentrated in these lower-order coefficients. This means that using a very high number of Mel filters does not translate into a proportionally richer final feature for the classifier, unlike in LMS or MS, where the full spectral detail passes through. Studies using NF below 32 for MFCC essentially recognise that, for the types of acoustic events they study, 20 to 30 filters provide sufficient spectral information for the DCT to produce meaningful cepstral coefficients. The number of methods with unspecified NF in this feature is also higher than in all previous features (19.4%, Fig. 28), which may partly reflect that some researchers consider NF less critical for MFCC precisely because the DCT step reduces the output dimensionality regardless of the input filter count.

Table 7 shows the frequency of MFCC-based studies broken down by year for each range. For Frame length, the highest frequency was in the [32–64) ms range, similar to previous features. However, the distribution across intervals is noticeably more spread out in MFCC than in the other features. The [20–32) ms interval has a higher relative frequency in MFCC (16.7%) than in LMS (12.2%) or MS (6.6%), as shown in Figs. 29, 9, and 17.

This broader spread can be explained by the historical use of MFCCs with a wider range of classifiers, including traditional machine learning approaches such as SVM, GMM, and HMM (Fig. 32), each of which imposes different input requirements compared to deep learning models. Traditional classifiers that operate on individual frames often benefit from shorter frame lengths that capture rapid spectral variations, whereas deep learning methods that process sequences of frames can effectively leverage longer frame durations. The number of studies that did not report frame length for MFCC is relatively high (19.4%, Fig. 29), the highest among all features examined.

Table 8 shows the Frame length frequency by year. Fig. 30 demonstrates that although in the OP parameter the two intervals above 50 have higher frequency, similar to spectrogram and energy-based features, the difference in frequency is noticeable. Specifically, the difference between intervals with OP above 50 and other intervals has decreased, and additionally, the percentage of studies that did not report this parameter has also increased (30.6%, far above the 13.5% in LMS, Figs. 30 and 11). The more even spread of OP in MFCC, and particularly the presence of studies using very low overlap in the [0–25) range

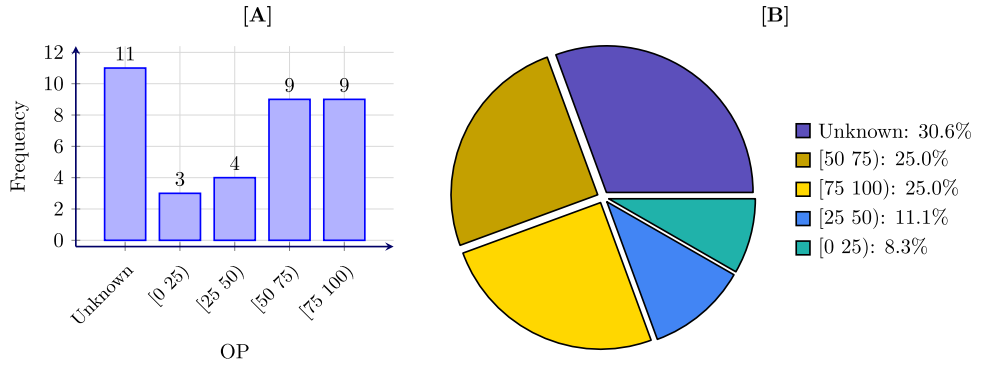


Fig. 30. OP in MFCC-based AEDC methods: [A] Frequency and [B] Relative frequency.

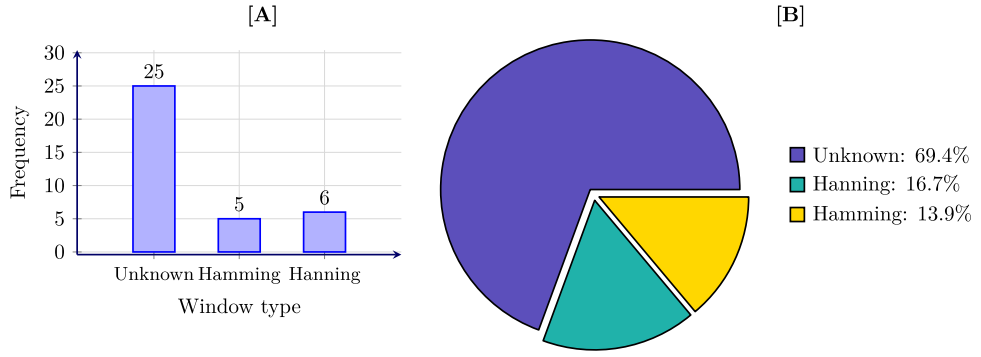


Fig. 31. Window types in MFCC-based AEDC methods: [A] Frequency and [B] Relative frequency.

Table 9
The frequency of OP in MFCC-based AEDC methods found over time.

Year	OP bins					Total
	Unknown	[0–25)	[25–50)	[50–75)	[75–100)	
2020	2	2	0	2	1	7
2021	3	0	2	1	3	9
2022	3	0	1	2	3	9
2023	2	1	0	2	1	6
2024	0	0	1	2	1	4
2025	1	0	0	0	0	1
Total	11	3	4	9	9	**36**

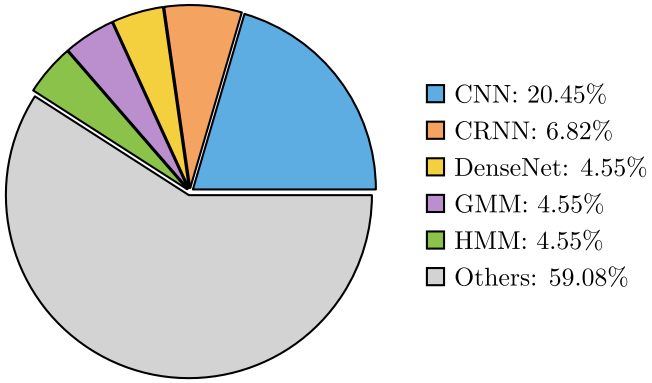


Fig. 32. Percentage of classification methods with relative frequency >1% in MFCC-based AEDC methods.

(8.3% vs 3.4% in LMS), connects to how MFCC features are used by classifiers. Because the DCT decorrelates the spectral information, consecutive MFCC frames are already less correlated with each other than consecutive spectrogram frames. This means the system can get away with lower overlap without losing important temporal patterns between frames.

For traditional classifiers like GMM or HMM that model each frame somewhat independently, low overlap simply provides more diverse training samples rather than the smooth temporal continuity that deep learning classifiers need from spectrogram inputs.

Table 9 shows the OP frequency by year for MFCC-based AEDC studies. The window type used in MFCC-based AEDC methods was mainly Hanning (16.7%) followed by Hamming (13.9%), as shown in Fig. 31.

The near-equal split is uncommon among the features examined. In LMS and MS, Hanning was clearly preferred for its smoother spectral transitions, while in LMBE, Hamming was more common for its sharper peak response. The fact that MFCC shows only a slight preference is consistent with the DCT processing step. Since the DCT decorrelates and compresses the log-Mel spectral representation into cepstral coefficients that characterize the spectral envelope, local spectral variations introduced by different window functions (e.g., Hamming vs. Hann) exert only a limited influence on the resulting MFCC features compared to their impact on raw spectrogram or Mel-band energy representations. Like all previous features, the window type was the most unreported parameter in AEDC studies (69.4%, Fig. 31).

Fig. 32 shows the percentage of repetition for different classifiers with relative frequency higher than 1% in MFCC-based AEDC studies. Unlike all previous features where CRNN was the top classifier, in MFCC the order is reversed: CNN leads at 20.45% whilst CRNN drops to 6.82%. This shift reflects a real difference in how MFCC gets used compared to spectrogram features. Because the DCT already captures temporal spectral patterns within each set of coefficients, the recurrent component of CRNN adds less value than it does for spectrograms, where temporal tracking across frames is essential. Simple CNNs can classify MFCC features effectively by treating the cepstral coefficients as a compact set of descriptors for each frame. The presence of traditional classi-

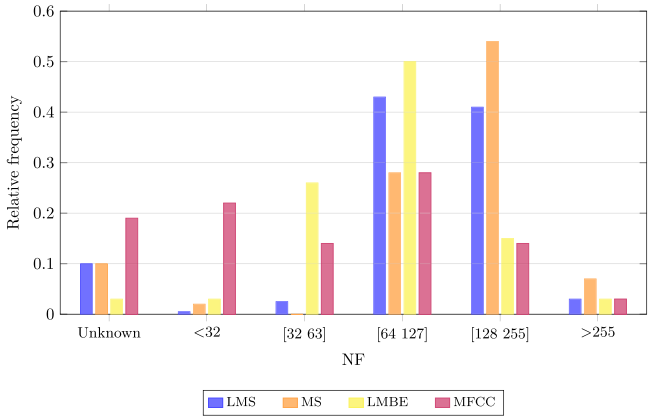


Fig. 33. Relative frequency of NF across all examined features.

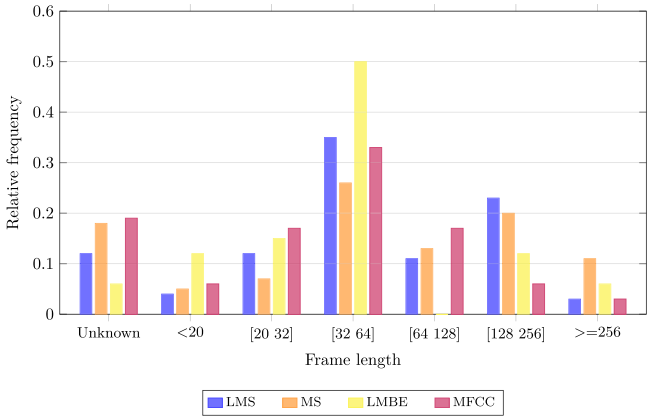


Fig. 34. Relative frequency of Frame length across all examined features.

fiers like GMM (4.55%), HMM (4.55%), and DenseNet (4.55%) in the MFCC results, which are almost absent from the LMS and MS classifiers (Figs. 14 and 22), confirms that MFCC remains the preferred Mel feature for researchers using non-deep learning approaches. MFCC's compact, decorrelated output works well with statistical models that assume the features are independent. Spectrograms, on the other hand, are high-dimensional and correlated, so they require the pattern recognition capabilities of deep learning to be used effectively.

4. Discussion

This section examines the findings from Section 3 through three complementary approaches: overall parameter comparison across all features, temporal pattern analysis using Jensen-Shannon divergence to quantify how parameter choices have changed over time, and detailed explanations of why specific parameters suit different acoustic conditions. The section also considers how application domains influence parameter selection and identifies current research limitations and future directions. In this section, a comparative analysis of the parameters and classifiers used in Mel-based AEDC research was conducted. The comparison was conducted in three subsections: Overall Analysis, Temporal Analysis, and Analytical Description. Due to differences in the number of studies conducted across various features, relative frequency was used for the comparisons. The fourth subsection identified challenges and future approaches in this field.

4.1. Overall analysis

Fig. 33 shows the relative frequency of NF for different Mel-based features. As shown, the two spectrogram-based features had high relative frequency in the [64–127] and [128–255] intervals, and their distributions at the two ends of the range were almost identical. The LMBE and MFCC were similar at > 255, but their frequency differed in the other intervals. The MFCC-based AEDC systems have the highest relative frequency in intervals with fewer NF. Fig. 34 illustrates the distribution of frame lengths across the examined features, highlighting systematic differences among them.

The only similar pattern across the four features is that the [32–64) interval had the highest relative frequency in all of them. In OP, all four features show higher relative frequency in intervals above 50% (Fig. 35). This parameter's percentage of unreported values was higher than that of previous parameters. The MFCC feature, with over 30% of values unreported, had the highest ambiguity among the features examined so far. Hamming and Hanning window types were the common choice for all features. The Hanning window had a higher relative frequency than the Hamming window for the two spectrogram-based features.

In the MFCC feature, Hanning had a slightly higher relative distribution than Hamming. In LMBE, the Hamming window had a higher

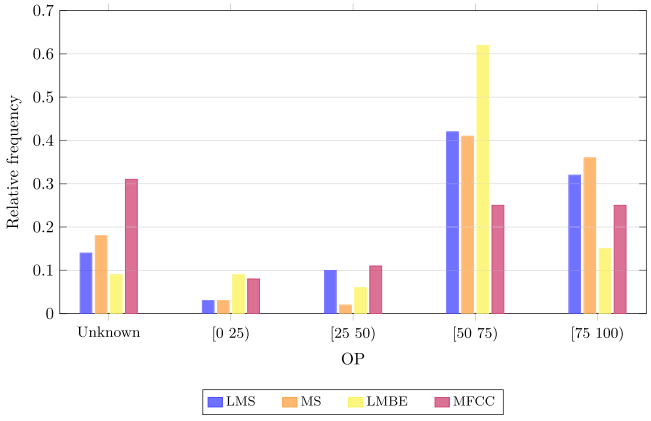


Fig. 35. Relative frequency of OP across all examined features.

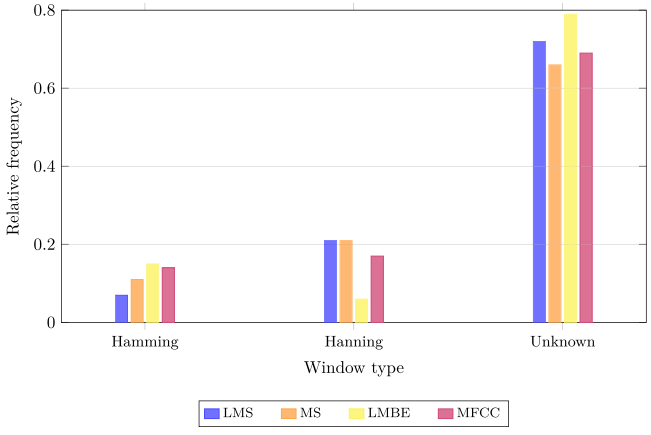


Fig. 36. Relative frequency of the window type across all examined features.

frequency (Fig. 36). This parameter has the highest percentage of being unspecified across all AEDC studies, suggesting that researchers considered it less important than other Mel feature parameters.

The overall comparison reveals that the five Mel features fall into two broad groups with distinct parameter patterns. The spectrogram group (LMS and MS) tends towards higher NF values, longer frames, higher overlap, and Hanning windows, reflecting the needs of deep learning classifiers processing detailed two-dimensional spectral representations. The energy and cepstral groups (LMBE, MBE, and MFCC) show greater variety in NF choices (including lower values), stronger preference for shorter frames, more moderate overlap, and either a Hamming preference or no clear window preference. These differences are not arbitrary;

Table 10
JSD values between temporal patterns of NF in different Mel-based features.

Features	LMS	MS	LMBE	MFCC
LMS	**0.000**	0.064	0.075	0.061
MS	0.064	**0.000**	0.078	0.040
LMBE	0.075	0.078	**0.000**	0.035
MFCC	0.061	0.040	0.035	**0.000**

they follow directly from how each feature processes spectral information and from the type of classifier that typically receives the output. The detailed reasoning behind these group-level differences is discussed in Section 4.3.

4.2. Temporal analysis

Examining the temporal patterns of Mel parameters is important because it shows whether their use has followed a specific pattern over time or remained constant. To examine the similarity of temporal trends for each parameter in different Mel-based features, considering that the frequency of different features varies significantly, in the first step, the frequency of each parameter at each time step was divided by the total AEDC research related to that feature to convert the temporal pattern of each parameter in each feature into a 2D probability density function. Various measures exist to compare discrete 2D probability density functions. The KLD measure, one of the most commonly used divergence measures, is defined as:

$$D_{KL}(P \parallel Q) = \sum_{i,j} P(i,j) Ln \frac{P(i,j)}{Q(i,j)}, \tag{8}$$

where P and Q are discrete probability density functions, D_{KL} indicates the distance between two distributions, and Ln is the natural logarithm (Basseville, 2013; Joyce, 2011). KLD measures the degree of deviation between two probability distributions. When two distributions are completely similar, this measure equals 0 (zero). This measure is not symmetric and is widely used in information theory (Kullback, 1997). KLD does not satisfy the triangle inequality and lacks symmetry, so it is not a valid distance metric. In applications such as machine learning, JSD offers a symmetric and bounded alternative to KLD for comparing 2D probability distributions (Menéndez et al., 1997):

$$JSD(P||Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M), \tag{9}$$

where D_{KL} is KLD and $M = \frac{1}{2}(P+Q)$. In this study, JSD was calculated for each parameter between every pair of features, and the final result was reported. The JSD similarity measure approaches 0 (zero). Table 10 shows JSD values in temporal patterns of NF between different features.

Given that the JSD ranges from 0 (zero) to $Ln(2) = 0.693$, interpreting similarity and difference results is possible based on defining a threshold level (0.05 in this study). Table 10 demonstrates that the most similar temporal pattern to LMS and MS in NF belongs to MFCC. The MS feature is identical to the MFCC, and, due to the assumed threshold, they can be interpreted similarly. LMBE and MFCC show the highest similarity in NF; moreover, the worst cases are LMBE with MS and LMS. The zero values on the main diagonal are due to the absolute similarity of each feature's temporal pattern to itself.

Table 11 shows JSD for the frame length temporal pattern. The LMS and MS features are similar to MFCC features, with the JSD value in LMS higher and that in MS lower than the corresponding NF value. The LMBE and MFCC features, similar to NF, showed the highest similarity, and MFCC showed the highest similarity with all other features.

Overall, the similarities in temporal patterns of frame length are higher than those in NF. The JSD values reported in Table 12 show that only MFCC and LMS are similar due to the selected threshold, and

Table 11
JSD values between temporal patterns of frame length in different Mel-based features.

Features	LMS	MS	LMBE	MFCC
LMS	**0.000**	0.058	0.068	0.045
MS	0.058	**0.000**	0.069	0.049
LMBE	0.068	0.069	**0.000**	0.029
MFCC	0.045	0.049	0.029	**0.000**

Table 12
JSD values between temporal patterns of OP in different Mel-based features.

Features	LMS	MS	LMBE	MFCC
LMS	**0.000**	0.070	0.062	0.050
MS	0.070	**0.000**	0.143	0.062
LMBE	0.062	0.143	**0.000**	0.092
MFCC	0.050	0.062	0.092	**0.000**

Table 13
JSD values between temporal patterns of window type in different Mel-based features.

Features	LMS	MS	LMBE	MFCC
LMS	**0.000**	0.048	0.028	0.033
MS	0.048	**0.000**	0.056	0.066
LMBE	0.028	0.056	**0.000**	0.040
MFCC	0.033	0.066	0.040	**0.000**

no similarity is observed in other temporal patterns of OP between Mel-based features.

In the Window Type parameter (Table 13), the JSD values showed similarity in most cases. The worst cases belonged to MS, two MFCC, and LMBE. The JSD values of this parameter are the lowest among all the analysed parameters. As a final summary, the following conclusions can be drawn:

- Among the different features, MFCC patterns exhibit the highest similarity, whereas MS shows the lowest similarity relative to the other features.
- No two features exhibit similar temporal patterns across all parameters.

4.3. Analytical description

4.3.1. NF Selection

The selection of NF depends on the spectral characteristics of target acoustic events. NF determines the frequency resolution in the Mel band, which must be sufficient to separate the fundamental frequencies of different events. For instance, in urban acoustic events (such as vehicle sounds, horns, and mechanical noise), fundamental frequencies and their harmonics are typically spaced by 100 Hz or more. At a 16 kHz sampling rate, 64 filters provide adequate resolution to separate these spectral components. Wildlife bioacoustics presents different requirements. Many species produce vocalisations with fundamental frequencies differing by 50 Hz or less. This necessitates a higher spectral resolution, which explains the shift toward 128–255 filters. Medical respiratory sounds fall between these extremes. Normal breathing shows harmonic spacing around 100 to 150 Hz, while pathological sounds

(wheezes, crackles) contain components spaced 150 to 300 Hz apart. For separating normal breathing from wheezes, NF = 64 is acceptable, but for separating crackles from wheezes, the AEDC system needs higher NF values.

In LMBE and MFCC, choosing NF differs from MS and LMS because these features include additional parameters that mitigate the frequency-resolution effect on system efficiency. In LMBE, the band energy averaging step means that the output for each filter is a single energy value rather than a detailed spectral shape. Adding more filters beyond 64 provides diminishing returns, since averaging already removes the fine spectral details that additional filters would capture. The data from this review support this observation: 50% of LMBE studies employed NF values in the [64-127] range, and 26.5% used the [32-63] range (Fig. 23). In contrast, only 2.5% of LMS studies selected the [32-63] range (Fig. 7). For MFCC, this trend is even more pronounced. The DCT compresses the log Mel spectrum into a handful of cepstral coefficients, typically 12 to 20. Since these coefficients represent the broad spectral envelope rather than individual frequency details, the number of input filters has a limited effect on the final output quality. This explains why 22.2% of MFCC studies used NF below 32 (Fig. 28), a range that is almost nonexistent in spectrogram studies.

The practical consequence for system designers is clear. When using LMS or MS, the choice of NF directly affects feature quality and should be carefully selected based on the spectral spacing of the target events. When using LMBE or MFCC, NF matters less, and other factors such as the number of cepstral coefficients or the temporal resolution become more important for system performance. The acoustic environment also affects NF selection, but this interaction has received very little attention in the reviewed studies.

4.3.2. Frame length

The selection of Frame length is a trade-off between the minimum event duration required for reliable spectral estimation and the need to avoid excessive temporal averaging. Impulsive acoustic events (gunshot, impacts, breaking sounds, firearm discharges) last 50–150 ms. Frame lengths exceeding 64 ms average the sharp transient with the surrounding background, reducing efficiency. Sustained harmonic sounds require longer frames for stable frequency estimation. A 100 Hz vibration completes one cycle every 10 ms. Reliable spectral analysis requires multiple complete cycles within each frame. Frames shorter than 32 ms capture insufficient cycles, producing unstable frequency estimates. The LMBE feature shows distinct behaviour with the highest prevalence in 32 to 64 ms frames (50%, Fig. 24), shorter than spectrogram features. Since LMBE averages spectral energy across each filter band, it sacrifices fine spectral detail. Shorter frames compensate by providing better temporal resolution for tracking rapid energy changes, making LMBE suitable for onset detection in events with abrupt energy transitions.

MFCC shows a different pattern again. Whilst the [32–64) ms range is still the most common (33.3%, Fig. 29), the distribution is notably wider than in other features. The [20–32) ms interval reaches 16.7% in MFCC compared to 12.2% in LMS (Fig. 9). This wider spread reflects the diversity of classifiers used with MFCC. Traditional machine learning methods like GMM and HMM (Fig. 32), which process individual frames rather than sequences, can work effectively with shorter frames because they do not rely on smooth temporal continuity between consecutive frames. Deep learning classifiers used with LMS and MS, on the other hand, process sequences of frames and benefit from consistent frame lengths that produce regular temporal patterns. Across all features, the frame length choice is most critical for spectrogram features (LMS and MS), as these representations preserve full spectral detail, which requires enough samples per frame for stable frequency estimation. For LMBE and MFCC, the downstream processing steps (band averaging and DCT, respectively) reduce the sensitivity to frame length, giving researchers more flexibility in this parameter. The interaction between frame length and noise level has also not been adequately investigated in the AEDC literature.

4.3.3. OP

OP changes the number of samples and influences model generalisation. The observed prevalence of 50 to 75% overlap (Figs. 11, 19) represents a practical compromise. Higher OP increases the number of samples. For instance, a 40 ms frame with 75% overlap advances 10 ms between consecutive frames, yielding 100 frames per second. At 50% overlap, this reduces to 50 frames per second. However, excessive OP creates temporal dependency between adjacent frames. At 75% overlap, consecutive frames share three-quarters of their samples, resulting in highly correlated spectral content. This can lead models to learn temporal smoothness specific to training data rather than generalising to event signatures. Application-specific patterns emerge from these considerations. Healthcare respiratory monitoring exhibits higher OP values (often 75 to 87.5%, Tables A.1 and A.10 in Appendix B.) because gradual transitions between normal and failure conditions require dense temporal sampling for precise onset detection.

The cross-feature comparison of OP reveals meaningful differences that connect to how each feature processes spectral information. LMS and MS show very similar OP patterns (41.8% and 41.0% in the [50–75) range, Figs. 11 and 19), which is expected since both produce two-dimensional spectral representations processed by similar classifier architectures. LMBE, however, concentrates much more heavily in the [50–75) range (61.8%, Fig. 25) with relatively little usage of very high overlap. Because LMBE averages energy within each Mel band, consecutive frames with very high overlap would contain nearly identical energy values, adding redundancy without useful information. MFCC shows the most spread out OP distribution with the highest unreported rate (30.6%, Fig. 30). The DCT decorrelation in MFCC means that consecutive frames are already less similar to each other than in spectrogram features, so the system can tolerate lower overlap without losing important temporal information. The effect of noise on the best OP choice has likewise not been studied.

4.3.4. Window function

For impulsive events confined to single frames, boundary effects usually occur at the leading and trailing edges; the choice of window function matters less because most frames contain either pure events or pure backgrounds. The main concern with this parameter arises in musical events or other sounds with strong harmonic content, where spectral leakage from the window's sidelobes can blur the boundaries between harmonics, making it harder for the system to identify the correct spectral pattern. The data from this review shows that window preferences vary across features in ways that connect to their processing characteristics. In LMS and MS, Hanning is preferred (21% and 21.3%, respectively; Figs. 13 and 21) because its smoother sidelobe suppression produces cleaner spectrogram representations for convolutional classifiers. In LMBE, Hamming has a higher prevalence (14.7% vs 5.9%, Fig. 26). Since LMBE performs frequency band averaging, preserving sharp spectral peaks through Hamming's narrower main lobe matters more than minimizing sidelobes, because the averaging step already smooths out sidelobe effects.

In MFCC, Hanning is slightly more common (16.7%) than Hamming (13.9%), as shown in Fig. 31, which reflects the DCT step reducing the impact of window choice on the final cepstral coefficients. How different window types handle background noise is another question that remains open. However, the high non-reporting rate (66% to 79% across features) indicates many researchers accept default settings without considering acoustic implications. This is the most underexplored parameter in Mel-based AEDC research, and systematic studies comparing window types across different acoustic conditions could yield practical improvements in system performance. One gap across all four parameters is the lack of research on how noise conditions affect parameter selection. Higher NF values may reveal event features hidden between noise components, but they also let more noise energy through. Longer frames average out short noise bursts but can also blur brief events. Higher overlap yields denser sampling, helping separate events from noise, but

Table 14
Practical guidance for Mel feature selection based on acoustic conditions.

Condition	Recommended feature	Reasoning
Clean, overlapping events	LMS	Full spectral detail aids event separation
Clean, single events	LMS or LMBE	Both perform well; LMBE is more efficient
Moderate noise	LMBE	Band averaging provides noise smoothing
High noise, deep learning	LMS	Classifiers can learn to ignore noise
High noise, traditional ML	MFCC	Coefficient truncation reduces noise
Limited computing power	LMBE or MFCC	Compact representations suit embedded systems

risks repeating noise patterns across frames. And different window types handle noise leakage across frequency bands differently. None of these interactions has been adequately studied in the AEDC literature, and controlled experiments comparing parameter choices across different noise levels would provide practical guidance that the field currently lacks.

4.3.5. Feature type selection

Different Mel features suit different acoustic conditions based on their information content and computational requirements. LMS provides a complete time-frequency representation suitable for complex overlapping events. Multiple simultaneous sounds produce additive patterns in spectrograms that neural architectures can learn to separate. Urban soundscapes with overlapping traffic, voices, and construction sounds benefit from this richness, which explains LMS's prevalence (64%, Fig. 6). The logarithmic amplitude scaling provides robustness against recording-level variations. MS without logarithmic scaling preserves linear amplitude relationships, useful when relative loudness carries diagnostic information.

However, this sensitivity to absolute levels makes MS susceptible to microphone distance variations, explaining its lower prevalence (16%). LMBE discards fine spectral structure by averaging energy within each Mel band, reducing dimensionality substantially (typically 40 to 64 values per frame versus 1024 to 2048 for spectrograms). This suits applications where broad frequency trends matter more than exact spectral shape. Binary event detection (present versus absent) often requires only coarse frequency distribution, not fine harmonic structure. The computational efficiency of LMBE (9% prevalence) enables it to run on embedded systems with limited processing capability. MFCC produces decorrelated coefficients that efficiently represent the spectral envelope. Traditional machine learning methods that struggle with high-dimensional inputs perform well with MFCC's compact representation. However, MFCC discards phase information and detailed spectral structure, limiting performance on overlapping events compared to spectrograms. The temporal shift toward LMS and away from MFCC (Fig. 3) reflects deep learning adoption. CNNs handle spectrogram dimensionality effectively, removing the efficiency advantage previously enjoyed by MFCC with traditional classifiers.

The choice between feature types also depends on the noise environment. LMS provides the richest representation in clean conditions, and its logarithmic compression helps in moderate noise by reducing dynamic range differences. However, in very noisy conditions, LMS preserves noise patterns alongside event patterns, which can confuse classifiers. MFCC offers a natural advantage over noise because the DCT concentrates useful spectral information in lower-order coefficients, and discarding higher-order coefficients effectively removes noise. This built-in noise reduction is one reason MFCC remained popular in speech recognition long before deep learning. LMBE sits between these extremes. Its band energy averaging provides some noise smoothing whilst retaining more information than MFCC. For applications where computational cost matters and noise is moderate, LMBE can be a practical middle ground between the richness of LMS and the compactness of MFCC. Table 14 summarises how different noise conditions and event characteristics influence the choice of Mel feature type based on the patterns observed in this review.

4.3.6. Mel-based features vs end-to-end learning

An ongoing question concerns whether Mel preprocessing remains necessary given advances in end-to-end learning from raw waveforms. For acoustic events with clear spectral structure, Mel features accelerate learning and improve efficiency, particularly when training data is limited. End-to-end waveform models must learn to discover patterns from scratch, requiring substantially more parameters and data to match the performance of Mel-based AEDC systems. However, for complex acoustic scenes with overlapping events or sounds lacking clear spectral structure, end-to-end approaches sometimes surpass Mel features by learning representations directly rather than being constrained by perceptual frequency scaling.

The continued prevalence of Mel-based methods (237 LMS studies, 61 MS studies, as detailed in Tables A.1 through A.9 in Appendix B.) shows that Mel-based features remain practically advantageous for most AEDC applications, especially when computational resources or training data were constrained. The findings of this review add specificity to this discussion. The shift from MFCC towards LMS over the 2020 to 2025 period (Fig. 3) does not indicate that Mel features are becoming less relevant. Rather, it shows that the type of Mel feature is changing alongside classifier evolution. As deep learning models replaced traditional classifiers, researchers moved from the compact MFCC representation (which suited statistical models) to the richer LMS representation (which suits convolutional architectures). This suggests that Mel features are adapting to new classifier paradigms rather than being replaced by them. The persistence of MFCC in recent studies (Table 7) shows that for specific applications where computational efficiency or noise robustness matters, the older feature type still has a role. End-to-end learning is most likely to replace Mel features in applications where very large training datasets are available, and the acoustic events lack well-defined spectral characteristics. For the majority of AEDC tasks reviewed here, where training data is limited, and events exhibit clear frequency patterns, Mel-based features remain a strong and practical starting point.

4.3.7. Domain specific parameter patterns

The analysis presented so far treats AEDC as a single category, but the 373 reviewed studies span several distinct application domains that impose different acoustic constraints on parameter selection. Based on the study descriptions and datasets reported in Tables A.1 through A.7 (Appendix B.), the most common domains in the reviewed literature are urban and domestic sound event detection, bioacoustics (bird, whale, and other animal vocalisations), industrial machinery monitoring, and surveillance or security related audio analysis. These domains differ in ways that directly affect which Mel parameters work best. Urban and domestic sound detection, which accounts for the largest share of the reviewed studies, typically involves polyphonic scenes in which several everyday sounds overlap. This explains the strong preference for LMS with high NF (128 or above) and moderate frame lengths (32 to 64 ms) observed in the overall statistics (Figs. 7 and 9), since these settings provide the spectral and temporal resolution needed to separate overlapping events in complex acoustic environments. Bioacoustic studies tend to have different requirements.

Animal vocalisations often occupy narrow frequency bands with distinctive harmonic patterns. Higher NF values are useful here too, but for

a different reason: they help resolve the closely spaced harmonics that distinguish one species from another. Frame lengths in bioacoustics are often shorter than in urban monitoring because many animal calls are brief and would be lost in longer analysis windows. MFCC appears more frequently in bioacoustic studies than in urban sound detection (Fig. 32), partly because the compact cepstral representation captures the spectral envelope of vocalisations efficiently and works well with the smaller datasets typical of wildlife research. Industrial monitoring studies focus on detecting faults or anomalies in machinery, where the sounds of interest are often continuous or slowly changing. Longer frame lengths and higher overlap percentages are well-suited to these applications because they yield stable spectral estimates of steady-state machinery noise.

The relevant acoustic events, such as bearing wear or valve leaks, produce gradual spectral changes rather than sudden onsets, which is why dense temporal sampling through high OP values is common in this domain. Surveillance and security applications share some characteristics with urban monitoring but place greater emphasis on detecting specific events, such as glass breaking, screams, or gunshots, in noisy environments. These events tend to be impulsive and brief, favouring shorter frame lengths and moderate overlap. Table 14 in Section 4.3.5 provides practical guidance on feature selection for different noise conditions, which is directly relevant to surveillance applications where background noise levels vary considerably. Although the reviewed studies do not always report their application domain explicitly, and a full statistical breakdown by domain would require individual classification of all 373 studies, the patterns described above are consistent with the parameter distributions observed throughout Section 3. Future reviews could benefit from requiring authors to report their target application domain alongside their parameter choices, enabling more precise domain-stratified analysis.

4.4. Future trends and open research questions

The statistical and analytical findings presented in Sections 4.2 and 4.3 naturally lead to several open research questions.

4.4.1. Spectral resolution optimisation beyond empirical convention

The temporal analysis presented in Section 4.3 shows a gradual shift from the [64-127] NF interval to the [128-255] interval in recent years (See Figs. 16, 23, and Table 3). While this shift coincides with the adoption of convolutional neural networks, it remains unclear whether higher NF values consistently improve classification performance or merely increase representational richness without measurable gains. So an important open question is:

Does increasing NF beyond 128 yield statistically significant improvements across diverse AEDC domains, or does it introduce diminishing returns relative to computational cost? Future research should investigate adaptive or data-driven NF selection strategies that dynamically adjust spectral resolution based on event bandwidth or harmonic density, rather than assuming a fixed value.

4.4.2. Re-evaluating the [32 64) ms frame-length consensus

More than 30% of the reviewed studies used frame lengths of 32–64 ms. While this interval reflects a reasonable compromise between time and frequency resolution, it is unclear whether this consensus represents optimal design or inherited default settings from earlier speech-processing conventions. A key research direction is:

Are alternative frame-length configurations (e.g., multi-scale or adaptive framing) better suited to highly transient or overlapping acoustic scenes? Multi-resolution feature extraction strategies may allow systems to preserve fine temporal detail for impulsive events while retaining spectral stability for sustained sounds.

4.4.3. Domain-specific parameter adaptation

Section 4.3.7 demonstrated clear domain-dependent parameter patterns. For example, bioacoustic applications tend to prefer higher NF and shorter frames, whereas industrial monitoring emphasises stability and sustained frequency tracking. However, current AEDC pipelines typically use fixed parameters across datasets. Future work should explore: Domain-aware or task-conditioned feature extraction modules that adjust Mel parameters automatically based on acoustic scene statistics. Such systems may bridge the gap between handcrafted parameter selection and fully end-to-end learned representations.

4.4.4. Parameter reporting and reproducibility standards

A significant proportion of the reviewed studies did not report window type or overlap percentage (Figs. 35 and 36). This lack of transparency limits reproducibility and prevents systematic meta-analysis.

An important methodological direction is the development of: Standardised reporting guidelines for Mel parameter configuration in AEDC research. Improving reporting consistency would enable more rigorous cross-study evaluation in the future.

4.4.5. Hybrid architectures

The continued dominance of Mel-based features suggests that they remain computationally efficient and structurally compatible with deep learning architectures. However, end-to-end waveform models are increasingly explored. A research opportunity lies in:

Exploring hybrid systems where initial Mel filterbanks are partially learnable, allowing fine adjustment of spectral spacing while retaining perceptual grounding.

This may combine interpretability and adaptability without abandoning Mel-based priors entirely. In summary, integrating AEDC systems into embedded expert systems requires continued development of efficient architectures that balance accuracy with computational constraints, produce reliable estimates for reasoning engines, and enable continual learning capabilities that enable current systems to adapt to different acoustic environments.

5. Conclusion

This systematic review examined 373 AEDC studies published between 2020 and 2025, providing the first focused analysis of Mel-based feature extraction in AEDC. Five primary Mel-based features were investigated: LMS appeared in 237 studies, MS in 61 studies, LMBE in 34 studies, MFCC in 36 studies, and MBE in 5 studies. Four critical Mel-related parameters were extracted from each study: NF, frame length, OP, and window type. The statistical analysis revealed clear differences in parameter selection patterns and temporal trends across feature types. LMS is the dominant Mel-based feature in recent AEDC research, reflecting its compatibility with deep learning architectures and its ability to provide comprehensive spectral characterisation. Mel parameter patterns differ substantially across feature types. AEDC methods employing LMS and MS predominantly use 64–255 Mel filters, whereas MFCC-based approaches typically use NF values below 32. Frame lengths of 32–64 ms are most commonly used across all features, providing a balance between temporal and spectral resolution. OP values exceeding 50% are preferred in most AEDC systems to ensure sufficient temporal sampling density. Window type specification was not reported in 66-79% of the reviewed studies, indicating limited attention to this parameter. Temporal pattern analysis using Jensen-Shannon divergence revealed that MFCC shows the greatest consistency in parameter selection across studies, whereas MS shows the highest variability. The classification method analysis identified CRNN as the dominant architecture across all feature types, followed by CNNs and transformer-based architectures. Current limitations include the absence of systematic parameter selection strategies, insufficient consideration of how acoustic event characteristics should guide feature design, and incomplete reporting of parameter values across studies. Future work should focus on adaptive parameter selection methods tailored to specific acoustic conditions, self-supervised learning strategies to address limited training data, and multimodal systems that integrate

acoustic features with other sensor modalities. Overall, the findings and analyses presented here provide practical guidance for researchers developing Mel-based AEDC systems across diverse application domains.

Data availability

No new data were created in this study.

CRediT authorship contribution statement

Vahid Hajihashemi: Investigation, Data curation, Software, Formal analysis, Writing – original draft; **Abdorreza Alavi Gharahbagh:** Investigation, Data curation, Software, Formal analysis, Writing – original draft; **José J.M. Machado:** Writing – review & editing; **João Manuel R.S. Tavares:** Writing – review & editing, Conceptualisation, Funding acquisition, Supervision.

Data availability

No data was used for the research described in the article.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgements

This article is partially a result of the project “Sensitive Industry”, co-funded by the European Regional Development Fund (ERDF), through the Operational Programme for Competitiveness and Internationalisation (COMPETE 2020), under the PORTUGAL 2020 Partnership Agreement. The first author would like to thank “Fundação para a Ciência e a Tecnologia (FCT)”, in Portugal, for his PhD grant with reference 2021.08660.BD.

Appendix A. Complete list of acronyms

Acronym	Definition
A-CRNN	Adapted CRNN
A2-FPN	Attention and Aggregation-based Feature Pyramid Network
ADD-YOLO	Angular-Distance YOLO
AED	Acoustic/Audio Event Detection
AEDC	Acoustic/Audio Event Detection and Classification
AFL	Asymmetric Focal Loss
AM	Affinity Mixup
APNet	Attentional Pyramid Network
APrNet	Audio Prototype Network
ASiT	Audio Spectrogram Vision Transformer
ASPP	Atrous Spatial Pyramid Pooling
AST	Audio Spectrogram Transformer
ATST	Audio Teacher Student Transformer
AttCapsNet	Attention Capsule Network
AWAM	Axis-Wise Attention Module
BEATS	Bidirectional Encoder representation from Audio Transformers
Bi-GRU	Bidirectional-GRU
Bi-LSTM	Bi-Directional LSTM
BNN	Binary Neural Network
BNMVDR	DNN-supported minimum Variance Distortionless Response
CAM	Class Activation Mapping
CapsNet	Capsule Neural Networks
CBAM	Convolutional Block Attention Module
CCAN	Causal Convolution Alternating Network
CDC	Central Difference Convolution
CDur	Duration robust CRNN
CGI	Cross-modal Graph Interaction
CMA	Cross-Modal Attention
CMAF	Cross-Modal Attentive Fusion
CNN	Convolutional neural network
CNM	Convolutional Neural Module
Conformer	Convolution-Augmented Transformer
CRNN	Convolutional recurrent neural network
CRST	Cross-Referencing Self-Training
CTNN	Convolutional Transformer Neural Network
CVAE	Conditional Variational Autoencoder
DA	Domain Adaptation
DBAM	Dual-branch attention module
DBN	Deep Belief Network
DCMA	Dual Cross-Modal Attention
DCNN	Deep Convolutional Neural Network
DCRNN	Dilated CRNN
DDC	Dilated Depthwise separable Convolution
DenseNet	Densely Connected Convolution Net
DKCNN	Dynamic Kernel Convolutional-based Neural Network
dLKA	downsized Large-Kernel Attention
DMRN	Dual Multimodal Residual Network
DNN	Deep Neural Network
DRNN	Dense Recurrent Neural Network
DSSDB	Deep Space Separable Distillation Block
DWS-CNN	Depthwise Separable CNN
E-ATP	Embedding-level Attention Pooling
ECA-Net	Efficient Channel Attention Network
EINV2	Event-Independent Network V2
EnFC-DNN	Autoencoder-based Fully Connected Deep Neural Network
FA3-Net	Feature Aggregation and Augmentation with Attention Network
FAP	Frequency-dependent Auto-Pooling Function
FCL	Fully Connected Layer

Acronym	Definition
FCN	Fully Convolutional Network
FDY	Frequency Dynamic Convolution
FDY-CRNN	Frequency Dynamic-CRNN
FL	Federated Learning / Focal Loss
FPN	Feature Pyramid Network
FTENet	Fine-Tune Embedded Network
GAN	Generative Adversarial Network
GCBIndRNN	Gated Convolution Bidirectional Independent Recurrent Neural Network
GL	Guided Learning
GLU	Gated Linear Unit
GMM	Gaussian Mixture Model
GNNet	Group Convolution Network
GRU	Gated Recurrent Unit
GTAP	Global Threshold Average Pooling
GWPR	Global Weighted Pooling Layer
HADNet	Highway Audio Detection Network
HiSSNet	Hierarchical SED and Speaker Identification Network
HMM	Hidden Markov Model
HMSED	human-machine collaboration-based SED
HTA	Hierarchical Temporal Aggregation
iCE	iConformer Encoding
I-GAP	Instance-level Global Average Pooling
I-GMP	Instance-level Global Max Pooling
IRN	Inverted Residual Network
IRTAM	IRN + Triplet Attention
JSD	Jensen Shannon divergence
KLD	Kullback-Leibler Divergence
KNN	K-nearest neighbor
LEE	Latent Embedding Estimation
LGAF-TIN	Local-Global Adaptive Fusion and Temporal Importance Network
LGC	Local and Global Consistency
LKA	Large Kernel Attention
LMBE	Log Mel Band Energies
LMS	Log Mel Spectrogram
LSTM	Long Short-Term Memory
LUU-Net	U-Net with Limited Upsampling
MA-JDC CNN	Multi-Scale Attention Based Joint Detection-Classification CNN
MBE	Mel Band Energies
MDSC	Modified Depthwise Separable Convolutions
MDTC	Multi-Dilation Time Convolution
MFCC	Mel-Frequency Cepstral Coefficients
MGA CRNN	Multi-Scale Gated Attention CRNN
MHA	Multi-Head Attention Layer
MHSA	Multi-Head Self-Attention
MIL	Multiple Instance Learning
MIMI	Maximising Inter-class distance and Minimising Intra-class distance
MLFL	Metric Learning And Focal Loss
MLP	Multilayer Perceptron
MLP-CRNN	MLP-Mixer enhanced CRNN
MRNet	Mobile-Regression Net
MS	Mel Spectrogram
MSCA	Multi-Scale Convolution-based Attention Network
MS-FCN	Multi-Scale Fully Convolutional Networks
MSRA	Multi-Scale Residual Attention
MT	Mean teacher
MTF-CRNN	Multiscale Time-Frequency CNN
MTL	Multitask Learning
MVANet	Multi-Stage Video Attention Network
MViT	Multiscale Vision Transformer
NAS-DYMC	Dynamic Multi-Scale Convolutional Neural Network
NF	Number of Mel filters
NMF	Nonnegative Matrix Factorization

Acronym	Definition
OP	Overlap Percentage
PADO	Transformer with Prompt Tuning
PANN	Pretrained Audio Neural Network
PaSST	The Patchout FAST Spectrogram Transformer
PBA	Pixel-Based Attention
PCL	Peer Collaborative Learning
PENet	Polyphonic Evidential Network
PEPE	Plain Efficient Pretrained (Audio) Embeddings
PMAE	Parallel Multi-Attention Enhancement
P-RRNN	Perturbed Residual Recurrent Neural Network
PT-PS model	Professional teacher-Professional student model
Puzzle CAM	Puzzle Class Activation Map
PWFFN	Position-Wise Feed-Forward Networks
RB	ResBlocks
RCnet	Resnet-Conformer Network
RCRNN	Residual Convolutional Recurrent Neural Network
RF	Random Forest
RNN	Recurrent Neural Network
SALSA	Spatial Cue-Augmented Log-Spectrogram
SAM	Spatial Attention Module
SAN	Self-Attention Network
SCT	Self-Consistency Training
SDM	Subband Dependency Model
SE-CRNN	Squeeze-and-Excitation CRNN
SED	Sound Event Detection
SEDC	Sound Event Detection and Classification
SELD	Sound Event Localisation and Detection
SEM	Semantic Embedding Module
SepSELD	Iterative Separation in Embedding Space SELD
SKA	Selective Kernel Attention
SKATTN	SKNet + Self-Attention Blocks
SKNet	Selective Kernel Network
SMN	Sequence Matching Network
SMS	Soft-Median Selection
SMSA	Selective Multi-Scale Attention
SRC-MT	Relation Consistency Mean Teacher
SSA	Sparse Self-Attention
SSAST	Self-Supervised Audio Spectrogram Transformer
SS-FCN	Single-Scale Fully Convolutional Networks
STRF	SpectroTemporal Receptive Field
SVM	Support Vector Machine
TA	Task-Adaptive
TB-STRFNet	Two Branch STRFNet
TCN	Temporal Convolutional Network
TCRNN	Transposed CRNN
TDCAM	Two-Stage Dual-conv Coordinate Attention Model
TFFS	Temporal Frequency attention and Feature Space attention
TFSCCN	Time-Frequency Separable Convolutional Compression Network
TFSIFNN	Time-Frequency Spectral Information Fusion Neural Network
TI	Transductive Inference
TPN	Transductive Propagation Network
T-SELDNet	Transpose Convolution
TSMLF	Task-Specific Meta-Learning Framework
T-UAED	Transformer-based Unified Audio Event Detection
TWEM	Track-Wise Ensemble Model
URNN	U Recurrent Neural Network
VAE	Variational Autoencoder
ViT	Vision Transformer
WPE	Weighted Prediction Error
W-YAMNet	Water YAMNet
YOHO	You Only Hear Once

Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/j.ejor.10.1016/j.eswa.2026.132549.

References

Akbalık, F., Yıldız, A., Ertuğrul, Ö. F., & Zan, H. (2024). Engine fault detection by sound analysis and machine learning. *Applied Sciences*, *14*(15), 6532. https://doi.org/10.3390/app14156532.

Baronetto, A., Fischer, S., Neurath, M. F., & Amft, O. (2025). Automated inflammatory bowel disease detection using wearable bowel sound event spotting. *Frontiers in Digital Health*, *7*, 1514757. https://doi.org/10.3389/fdgth.2025.1514757.

Basseville, M. (2013). Divergence measures for statistical data processing-an annotated bibliography. *Signal Processing*, *93*(4), 621–633. https://doi.org/10.1016/j.sigpro.2012.09.003.

Boztas, G., Tuncer, T., Aydogmus, O., & Yildirim, M. (2024). A DCSLBP based intelligent machine malfunction detection model using sound signals for industrial automation systems. *Computers and Electrical Engineering*, *119*, 109541. https://doi.org/10.1016/j.compeleceng.2024.109541.

Castorena, C., Cobos, M., Lopez-Ballester, J., & Ferri, F. J. (2024). A safety-oriented framework for sound event detection in driving scenarios. *Applied Acoustics*, *215*, 109719. https://doi.org/10.1016/j.apacoust.2023.109719.

Ciaburro, G. (2025). Automated home security system based on sound event detection using deep learning methods. In *Modern advancements in surveillance systems and technologies* (pp. 273–302). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-6996-8.ch012.

Fernando, T., Sridharan, S., Denman, S., Ghaemmaghami, H., & Fookes, C. (2022). Robust and interpretable temporal convolution network for event detection in lung sound recordings. *IEEE Journal of Biomedical and Health Informatics*, *26*(7), 2898–2908. https://doi.org/10.1109/JBHI.2022.3144314.

Hajihashemi, V., Gharahbagh, A. A., Cruz, P. M., Ferreira, M. C., Machado, J. J. M., & Tavares, J. M. R. S. (2022). Binaural acoustic scene classification using wavelet scattering, parallel ensemble classifiers and nonlinear fusion. *Sensors*, *22*(4), 1535. https://doi.org/10.3390/s22041535.

Harimi, A., Majd, Y., Gharahbagh, A. A., Hajihashemi, V., Esmaileyan, Z., Machado, J. J. M., & Tavares, J. M. R. S. (2022). Classification of heart sounds using chaogram transform and deep convolutional neural network transfer learning. *Sensors*, *22*(24), 9569. https://doi.org/10.3390/s22249569.

Jiang, Y., Guo, D., Wang, L., Zhang, H., Dong, H., Qiu, Y., & Zou, H. (2024). Sound event detection in traffic scenes based on graph convolutional network to obtain multi-modal information. *Complex & Intelligent Systems*, *10*(4), 5653–5668. https://doi.org/10.1007/s40747-024-01463-7.

Joyce, J. M. (2011). Kullback-leibler divergence. In *International encyclopedia of statistical science* (pp. 720–722). Springer. https://doi.org/10.1007/978-3-642-04898-2_327.

Juodakis, J., & Marsland, S. (2022). Wind-robust sound event detection and denoising for bioacoustics. *Methods in Ecology and Evolution*, *13*(9), 2005–2017. https://doi.org/10.1111/2041-210X.13928.

Krstulović, S. (2017). Audio event recognition in the smart home. *Computational analysis of sound scenes and events*, (pp. 335–371). https://doi.org/10.1007/978-3-319-63450-0_12.

Kullback, S. (1997). Information theory and statistics. Courier Corporation.

Menéndez, M. L., Pardo, J. A., Pardo, L., & del C., P. M. (1997). The jensen-shannon divergence. *Journal of the Franklin Institute*, *334*(2), 307–318. https://doi.org/10.1016/S0016-0032(96)00063-4.

Nolasco, I., Singh, S., Morfi, V., Lostanlen, V., Strandburg-Peshkin, A., Vidaña-Vila, E., Gill, L., Pamuła, H., Whitehead, H., Kiskin, I. et al. (2023). Learning to detect an animal sound from five examples. *Ecological Informatics*, *77*, 102258. https://doi.org/10.1016/j.ecoinf.2023.102258.

Park, J., Cho, Y., Sim, G., Lee, H., & Choo, J. (2022). Enemy spotted: In-game gun sound dataset for gunshot classification and localization. In *2022 IEEE Conference on games (cog)* (pp. 56–63). IEEE. https://doi.org/10.1109/CoG51982.2022.9893670.

Podda, A. S., Balia, R., Pompianu, L., Carta, S., Fenu, G., & Saia, R. (2024). Cargram: Cnn-based accident recognition from road sounds through intensity-projected spectrogram analysis. *Digital Signal Processing*, *147*, 104431. https://doi.org/10.1016/j.dsp.2024.104431.

Shanmukha, S. V., & Deepak, K. (2024). Sound event detection in constrained real-world environments using fine tuned CNNs. In *2024 International conference on distributed computing and optimization techniques (ICDCOT)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICDCOT61034.2024.10516127.

Tagawa, Y., Maskeliūnas, R., & Damaševičius, R. (2021). Acoustic anomaly detection of mechanical failures in noisy real-life factory environments. *Electronics*, *10*(19), 2329. https://doi.org/10.3390/electronics10192329.

Ueda, M., Tanaka, D., & Tanaka, M. S. (2024). A study on anomalous sound detection in factories for early failure detection using wavelet transform. In *2024 International conference on consumer electronics-Taiwan (ICCE-taiwan)* (pp. 469–470). IEEE. https://doi.org/10.1109/ICCE-Taiwan62264.2024.10674160.

Wißbrock, P., Ren, Z., & Pelkmann, D. (2024). More than spectrograms: Deep representation learning for machinery fault detection. *Applied Acoustics*, *225*, 110178. https://doi.org/10.1016/j.apacoust.2024.110178.