| Volume 02 | Issue 03 | August 2026 | ISSN: 3139-1737

International Journal of Advances and Applications in Convergent Studies & Transdisciplinary (IJ-AACST)

Bridging Disciplines, Bridging Knowledge

Aligned with the Vision of NEP 2020 and UGC Guidelines

Website: https://aacstij.org || Email: journaleditor@aacstij.org

Audio Feature Extraction Methods for Machine Learning and Deep Learning-Driven Sound Classification: A Scientific Literature-Based Analysis

DOI: https://doi.org/10.5281/zenodo.21817718

Sandhya Rani K.M[1]*, Dr. S D Khamitkar[2], Dr. Parag U Balachandra[3]

*Corresponding Author: [1] Research Scholar, School of Computational Sciences, SRTM University, Nanded
sandhya.kiran.rao@gmail.com

ORCID: https://orcid.org/0009-0009-2027-8513

[2]School of Computational Sciences, SRTM University, Nanded
s.khamitkar@gmail.com

[3]School of Computational Sciences, SRTM University, Nanded
srtmun.parag@gmail.com

ORCID: https://orcid.org/0000-0001-5422-9423

ABSTRACT

Sound classification has become an important task in computational audio analysis because audio signals can provide meaningful information regarding the environment, speech, music, machines, animals, health conditions, and forensic evidence. Although recent studies often emphasize machine learning and deep learning architectures, the representation of audio signals before classification remains a central methodological consideration. This study presents a literature-based analysis of audio feature extraction techniques employed in learning-driven sound classification. By synthesizing recent research, it identifies key feature categories, examines the relationships between features and model architectures, explores domain-specific feature preferences, and addresses reproducibility issues. The findings highlight that commonly used features include MFCCs, Mel spectrograms, STFT/spectrograms, chroma, RMS) energy, zero-crossing rate, spectral descriptors, pitch and rhythm features, and multi-feature fusion. Spectrogram-like inputs are predominantly utilized with CNN models, whereas classical machine learning approaches typically rely on concise handcrafted features. This study emphasizes that feature extraction should be considered a domain- and model-dependent design decision rather than a routine preprocessing step.

KEYWORDS: Audio feature extraction; sound classification; machine learning; deep learning; MFCC; Mel spectrogram; feature fusion; Librosa.

1. INTRODUCTION

Sound has become an important source of information for the automated classification of animal behavior. Audio signals can reveal patterns associated with the environment, speech, emotions, music, machine operation, animal activity, biomedical conditions, and forensic evidence. Accordingly, sound classification has been applied in environmental sound recognition, speech and voice analysis, music information retrieval, respiratory and cardiac

sound assessment, bioacoustic monitoring, industrial sound analysis, emergency sound detection, and digital audio forensics [1], [2], [3], [4], [5], [6], [7]. These applications show that audio is not merely a recorded signal but a source from which meaningful computational patterns can be derived.

Despite its practical value, audio is difficult to classify directly using raw waveforms. The waveform is continuous, time-dependent, and sensitive to the recording conditions, background noise, device characteristics, and signal duration. Therefore, most sound classification systems transform raw audio data into more informative representations before classification. Features such as Mel-frequency cepstral coefficients (MFCCs), spectrograms, Mel spectrograms, chroma, zero-crossing rate, root mean square (RMS) energy, spectral descriptors, pitch-related measures, rhythm descriptors, and statistical acoustic features are commonly used to represent the temporal, spectral, cepstral, harmonic, and energy-related properties of sound [8], [9], [10].

Earlier studies generally treated feature extraction as a fixed vector preparation stage for classical machine learning models. In such systems, descriptors such as MFCCs, temporal features, spectral features, and statistical summaries are extracted and classified using support vector machines, random forests, k-nearest neighbors, decision trees, Naïve Bayes, logistic regression, or gradient boosting methods. These approaches remain relevant because they are computationally efficient and interpretable in nature. However, their success strongly depends on whether the selected features capture the discriminative acoustic properties of the target class.

In contrast, recent deep learning studies have increasingly used CNNs, RNNs, LSTMs, GRUs, CRNNs, attention mechanisms, and related representation-learning approaches for sound classification [11], [12], [13], [14], [15]. These models can learn complex patterns from high-dimensional inputs; however, they do not eliminate the need for audio representations. Many deep learning systems still use Mel spectrograms, STFT spectrograms, MFCC maps, harmonic-contrast representations, or fused feature inputs. This suggests that deep learning has transformed the role of feature extraction rather than replacing it. In classical machine learning, features usually act as compact numerical vectors, whereas in deep learning, they often serve as structured time–frequency or cepstral representations that support hierarchical learning.

A remaining gap in the literature is that sound classification studies frequently emphasize model architecture while giving less analytical attention to feature choice, feature–model compatibility, and domain-wise feature suitability. However, the selected feature representation determines the acoustic information available to the classifier. MFCCs provide compact information about the spectral envelope and timbre; spectrogram and Mel spectrogram representations preserve the time–frequency structure; chroma, CQT, and Tonnetz features capture tonal and harmonic information; and RMS, ZCR, and spectral descriptors represent signal energy and frequency-shape characteristics [16], [17], [18]. Because speech, music, environmental, medical, industrial, and forensic sounds differ in their acoustic structures, feature selection should be treated as a scientific design decision rather than a routine pre-processing step. Unlike model-centered reviews that mainly compare classification architectures, this study places audio feature extraction at the center of the analysis. It synthesizes feature families, model compatibility, domain-wise feature preferences and Librosa-compatible implementation mapping. This contribution lies in connecting signal representation choices with learning model design and reproducibility

requirements. Therefore, this study analyzed audio feature extraction methods, feature–model pairing, and domain-wise feature preferences across 70 selected studies.

2. LITERATURE REVIEW / RELATED WORK / BACKGROUND STUDY

This study adopted a literature-based analytical design rather than a strict PRISMA review or quantitative meta-analysis. The selected publications were used as an evidence base to examine how audio feature extraction methods are employed in machine learning- and deep learning–driven sound classification.

The analysis is based on a curated sample of studies on audio and sound classification. For each study, the following information was extracted, where available: publication year, application domain, classification task, feature extraction methods, model family, dataset or corpus details, clearly reported performance values, and relevant methodological observations. The missing details were neither inferred nor reconstructed. When key information, such as dataset size, feature extraction parameters, or validation protocols, was not reported, the corresponding study was treated cautiously and excluded from unsupported quantitative comparisons.

Feature extraction methods were grouped into major families: cepstral features, time–frequency representations, temporal descriptors, spectral descriptors, harmonic and pitch-based features, rhythm-related features, statistical acoustic descriptors, and multi-feature fusion approaches. The model families were organized into classical machine learning and deep learning approaches. Classical machine learning includes models such as Support Vector Machines, Random Forests, k-nearest neighbours, decision trees, Naïve Bayes, logistic regression, gradient boosting, and XGBoost. Deep learning encompasses architectures such as Convolutional Neural Networks, Deep Neural Networks/Multilayer Perceptron’s, Recurrent Neural Networks, Long Short-Term Memory networks, Gated Recurrent Units, Convolutional Recurrent Neural Networks, Transformer-based and attention-based models, and hybrid architectures.

The analysis was descriptive and interpretative. Descriptive counts were used to summarize publication trends, whereas qualitative comparisons were used to interpret feature families, model compatibility, and domain-wise suitability. Because the reviewed studies differed in datasets, class structures, validation methods, preprocessing procedures, and performance metrics, direct ranking of models or features based only on reported accuracy was avoided. Therefore, emphasis is placed on the scientific interpretation of audio representations, rather than performance-based comparisons.

3. OVERVIEW OF THE SELECTED STUDIES

The 70 selected studies were published between 2019 and 2026. The annual distribution indicates that learning-based sound classification using audio feature extraction has received increasing attention in recent years. Only one study was identified in 2019, followed by an increase in 2020. The highest number of studies appeared in 2024, accounting for 17 studies (24.3% of corpus). The years 2023 and 2025 also showed strong activity, with 15 and 12 studies published in those years, respectively. The smaller number of studies in 2026 may reflect incomplete indexing or the partial nature of the year of data collection.

The selected studies covered a broad range of application areas. Environmental and urban sound classification has appeared frequently, reflecting interest in acoustic scene analysis, smart city monitoring, and real-world sound event recognition [11], [12], [19], [20], [21]. Speech- and voice-related studies include speech classification, speaker- or gender-related analysis, speech emotion recognition, and voice classification [16], [22], [23]. Music and music information retrieval studies have addressed musical instrument classification, music classification, raga identification, and music sentiment-related tasks [2], [24], [25], [26]. Medical sound studies have focused on respiratory, lung, cardiac, and biomedical acoustic classifications [1], [4], [27], [28], [29], [30]. Other domains include animal and bioacoustic sound classification, industrial or appliance sound recognition, emergency vehicle detection, and audio forensics [5], [6], [7], [14], [31]

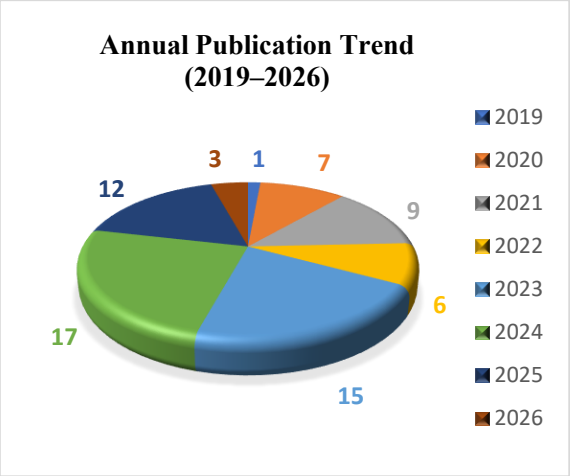

Figure 1. Annual publication trends of selected studies from 2019 to 2026.

Both classical machine learning and deep learning approaches were represented in the study. Classical models are mainly associated with fixed-length handcrafted features, whereas deep learning models are more commonly paired with spectrogram-like inputs, MFCC maps, fused features, and learned representations. This distribution supports the central observation that feature extraction and model architecture are closely linked in sound classification.

4. Audio Feature Extraction Methods in Sound Classification

The core analytical focus of this study is audio feature extraction. The selected studies show that feature extraction is not a uniform operation; rather, it involves choosing an acoustic representation that matches the signal characteristics, classification objectives, and the learning model. The major feature families observed in the literature include cepstral, time–frequency, temporal, spectral, harmonic, pitch-based, rhythm-based, statistical, and fused features.

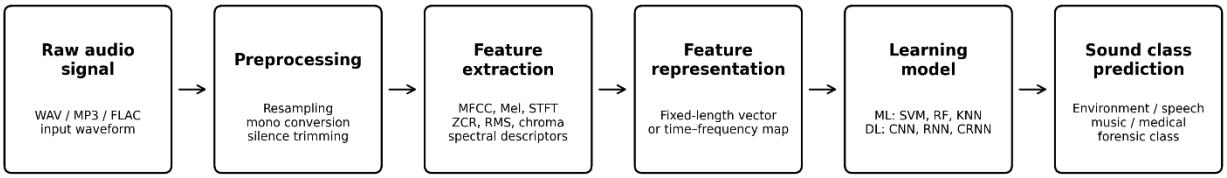

Figure 2. General pipeline of audio feature extraction for learning-based sound classification.

4.1 Cepstral Features

Cepstral features, particularly MFCCs, are among the most widely used representations in sound classification. MFCCs provide a compact description of the spectral envelope of an audio signal and are closely associated with timbre-related information. Their popularity across speech, music, respiratory, cardiac, environmental, and

bioacoustic classification tasks is partly due to their ability to summarize perceptually meaningful frequency characteristics using a relatively small number of coefficients [16], [32].

Compared with raw spectral descriptors, MFCCs are compact and suitable for conventional machine learning models. They can be summarized statistically using the mean, standard deviation, or other measures to form fixed-length input vectors for SVM, Random Forest, KNN, or Gradient Boosting models. MFCCs have also been used with CNNs and RNNs when represented as frame-level matrices or sequential feature maps [9]. This flexibility explains their continued use in both traditional and deep learning-based pipelines.

The compactness of the MFCCs can also be a limitation. Because MFCCs compress spectral information, they may discard fine time–frequency details that are useful for complex acoustic scenes such as speech. For instance, environmental and biomedical sounds may contain short-duration events, subtle frequency variations, or localized acoustic patterns that are not fully captured by summary-level MFCC features. Delta and delta-delta MFCCs can partly address this limitation by representing temporal changes in MFCC trajectories, although their usefulness depends on the temporal structure of the task and the model used. This explains why MFCCs remain useful in compact machine learning pipelines, particularly when interpretability and lower dimensionality are important. However, when the classification task depends on localized temporal–spectral patterns, MFCCs may need to be complemented with spectrogram-based or fused representations.

4.2 Time–Frequency Representations

The time–frequency representation describes how the frequency content of a signal changes over time. STFT, Mel, and log-Mel spectrograms are particularly important in deep learning-based sound classification because they can be treated as two-dimensional inputs. In these representations, local acoustic structures such as bursts, harmonics, frequency sweeps, and energy changes can be learned using CNN filters.

The STFT spectrogram preserves detailed frequency-time information and is useful when temporal changes in the spectral structure are significant. Mel spectrograms transform frequency information using the Mel scale, providing a perceptually motivated representation that is often more compact than a full STFT spectrogram. Log-Mel spectrograms further compress the energy values and can reduce the dominance of high-magnitude spectral components. These properties make Mel-based representations common in CNN-based environmental, appliance, speech, and general audio classification studies [17].

Compared with MFCC vectors, spectrogram-based features preserve a richer time–frequency structure. This is advantageous for deep learning models but may require more data, computations, and careful parameter selections. Choices such as the sampling rate, window length, hop length, FFT size, and number of Mel bands can influence the preserved acoustic patterns. Therefore, spectrogram-based classification should not be considered automatically superior; its effectiveness depends on both the extraction settings and the model design

4.3 Temporal and Spectral Descriptors

Temporal and spectral descriptors provide simpler yet interpretable information about audio signals. The zero-crossing rate measures how often the waveform crosses the zero-amplitude axis and can reflect noisiness or high-frequency activity. The RMS and short-time energies describe the signal strength or loudness patterns. These

features are useful in tasks where intensity variation, silence, bursts, or waveform activity help to distinguish sound classes.

Spectral descriptors describe the distribution and shape of the frequency content. The spectral centroid is often associated with brightness or the center of the spectral mass. Spectral roll-off indicates the frequency below which most of the spectral energy is concentrated, whereas the spectral bandwidth reflects the spread of the frequency content. Spectral contrast evaluates the variation between the highest and lowest points within spectral bands, while spectral flatness aids in differentiating between tonal and noise-like sounds.

These features are lightweight and interpretable, making them attractive for real-time and resource-limited applications. Emergency vehicle classification, speech/music classification, and general audio classification studies have demonstrated the use of temporal and spectral descriptors with classical machine learning models. However, when used alone, these descriptors may not capture the full complexity of sound scenes. They are often more effective when combined with cepstral or time–frequency representation.

4.4 Harmonic, Pitch, and Rhythm Features

Harmonic, pitch, and rhythm features are more domain-specific than MFCCs and general spectrograms. In music-related tasks, chroma features play a crucial role as they illustrate the distribution of energy across various pitch classes. The constant-Q transform provides a logarithmic frequency representation that is well aligned with the musical pitch structure. Tonnetz features describe tonal relationships, whereas tempo, beat, and onset descriptors capture rhythmic organization [18].

Music and music information retrieval tasks often require these features because music signals contain harmonic progressions, pitch structures, rhythms, and timbral variations. For example, musical instrument classification, raga identification, and music classification tasks may benefit from features that preserve harmonic and rhythmic characteristics rather than relying only on MFCCs or general spectrograms [25], [26].

Pitch- and energy-related features are also relevant in speech, voice, emotion, and bioacoustic tasks. Speech emotion recognition may depend on pitch contour, loudness, and prosodic variation, whereas animal sound classification may involve species-specific vocal frequency patterns [7], [23]. These examples show that feature selection should be guided by the physical and perceptual structures of the sound domain.

4.5 Statistical Acoustic Features

Many audio features are originally extracted on a frame-by-frame basis. To use them with classical machine learning models, they are often summarized into fixed-length vectors using statistical measures such as the mean, standard deviation, variance, minimum, maximum, skewness, kurtosis, or entropy. This approach converts variable-length audio signals into a format suitable for models such as SVM, Random Forest, KNN, and Gradient Boosting.

Statistical summaries are useful because they reduce the dimensionality and simplify the classification. They are particularly suitable when the datasets are small or when computational simplicity is important. However,

statistical summarization may remove temporal ordering and local acoustic structures. Therefore, although statistical features can support efficient classification, they may be less suitable for tasks in which the sequence or timing of acoustic events is important.

4.6 Multi-Feature Fusion

Multi-feature fusion is used because different feature families capture different acoustic properties of the sound. MFCCs represent cepstral and timbral information, Mel spectrograms capture time–frequency energy distribution, chroma reflects harmonic structure, RMS and ZCR describe temporal energy and waveform variation, and spectral descriptors capture frequency shape characteristics. The combination of features can produce a more complete acoustic representation.

Feature fusion has been reported in environmental sound classification, lung and heart sound classification, cardiac sound classification, and general audio classification studies. Some studies combine handcrafted features, whereas others combine handcrafted and deep features. Fusion may improve representation by capturing complementary information; however, it also increases dimensionality, computational cost, and potential overfitting risk. Therefore, the benefits of fusion should ideally be evaluated using ablation studies that compare individual features, fused features, and model outputs under the same conditions.

Because many sound classification studies report standard acoustic descriptors without always naming the extraction library, it is useful to distinguish between explicitly reported Librosa use and Librosa-compatible feature extraction. The feature families discussed in this study can be implemented using standard Librosa functions, as summarized in Table 2. This mapping supports transparency and clarifies how commonly reported descriptors, such as MFCC, Mel spectrogram, chroma, RMS energy, zero-crossing rate, spectral centroid, spectral contrast, CQT, Tonnetz, and tempo, can be operationalized in Python-based audio analysis. The table should not be interpreted as evidence that all the reviewed studies explicitly used Librosa. The functions listed in Table 2 are implementation equivalents for reproducibility and are not used as inclusion criteria.

Table 2. Audio feature families and corresponding Librosa functions for reproducible sound classification workflows

Feature family	*Example features*	*Librosa function / method*	*Acoustic information captured*
Audio loading / preprocessing	Load audio, resampling, trimming silence	librosa.load(), librosa.resample(), librosa.effects.trim()	Prepares waveform for feature extraction
Cepstral features	MFCC	librosa.feature.mfcc()	Spectral envelope, timbre, perceptual frequency structure
Dynamic cepstral features	Delta MFCC, delta-delta MFCC	librosa.feature.delta()	Temporal change in MFCC patterns
Time–frequency features	STFT, spectrogram	librosa.stft(), librosa.amplitude_to_db()	Frequency variation over time
Mel-scale features	Mel spectrogram, log-Mel spectrogram	librosa.feature.melspectrogram(), librosa.power_to_db()	Perceptual time–frequency energy distribution
Temporal features	Zero-crossing rate	librosa.feature.zero_crossing_rate()	Waveform sign changes, noisiness, signal variation

Energy features	RMS energy	librosa.feature.rms()	Frame-level signal energy or loudness
Spectral-shape features	Spectral centroid	librosa.feature.spectral_centroid()	Brightness or centre of spectral mass
Spectral-spread features	Spectral bandwidth	librosa.feature.spectral_bandwidth()	Spread of frequency content
High-frequency distribution	Spectral roll-off	librosa.feature.spectral_rolloff()	Frequency below which most spectral energy is concentrated
Spectral texture	Spectral contrast	librosa.feature.spectral_contrast()	Difference between spectral peaks and valleys
Tonal/noise-like quality	Spectral flatness	librosa.feature.spectral_flatness()	Tonal versus noise-like character
Harmonic features	Chroma	librosa.feature.chroma_stft(), librosa.feature.chroma_cqt(), librosa.feature.chroma_cens()	Pitch-class and harmonic structure
Constant-Q features	CQT	librosa.cqt()	Log-frequency representation useful for music and pitch-related tasks
Tonal centroid features	Tonnetz	librosa.feature.tonnetz()	Tonal and harmonic relationships
Rhythm features	Tempo, beat	librosa.beat.beat_track()	Beat and rhythmic structure
Onset features	Onset strength	librosa.onset.onset_strength()	Sudden energy changes and event beginnings
Feature fusion	MFCC + Mel + chroma + spectral features	Combination of multiple Librosa outputs	Complementary acoustic representation

5. FEATURE–MODEL PAIRING AND DOMAIN-WISE INTERPRETATION

5.1 Feature–Model Compatibility

The feature representation and model architecture should be selected together. Classical machine learning models generally require fixed-length feature vectors. Therefore, MFCC statistics, spectral descriptors, RMS, ZCR, and other statistical acoustic features are suitable for SVM, Random Forest, KNN, Gradient Boosting, and related models. These pairings are useful when interpretability, lower computational cost, or limited training data are required.

CNN-based models are more compatible with spectrogram-like representations because CNN filters can learn the local time–frequency structures. Studies on environmental sound classification, Mel-filter bank features, appliance sound classification, and general audio classification have demonstrated the use of CNNs with spectrogram or Mel spectrogram inputs[21]. In such systems, the feature map behaves similarly to an image, allowing the convolutional layers to detect patterns across time and frequency.

Sequential architectures, such as LSTM, GRU, and CRNN, are more appropriate when the temporal progression of acoustic events is important. Speech emotion, respiratory sound, heart sound, and audio event classification tasks may require models to capture changes across frames rather than static summaries [18], [45] [51], [67]. Attention-based and representation-learning approaches may be useful when the classifier needs to focus on

selected time–frequency regions or a broader acoustic context [46], [69]. However, these models require carefully designed inputs and sufficient data to prevent overfitting.

Table 3. Feature–model compatibility in learning-based sound classification

Feature representation	Suitable model family	Reason for compatibility	Example application domain
MFCC/statistical features	SVM, RF, KNN, Gradient Boosting	Fixed-length compact representation	Speech/music, emergency sound
Spectrogram/Mel spectrogram	CNN, ResNet, transfer learning CNN	Image-like time–frequency structure	Environmental sound, appliance sound
Sequential MFCC/spectrogram frames	LSTM, GRU, CRNN	Preserves temporal progression	Speech emotion, medical sound
Harmonic/rhythm features	ML models, CNN, hybrid models	Captures pitch, harmony, rhythm	Music/MIR, instrument classification
Fused acoustic features	Hybrid ML/DL models	Combines complementary sound properties	Environmental, medical, general audio
Embedding or feature maps	Attention/representation-learning models	Supports selective focus or learned audio embeddings	Speech, audio event, complex sound scenes

5.2 Domain-Wise Feature Preferences

The selected studies indicate that the feature choice is strongly domain-sensitive. Environmental and urban sound classification often uses MFCCs, Mel spectrograms, and STFT/spectrograms because these features capture broad and non-stationary time–frequency patterns. Speech and voice tasks commonly use MFCCs, pitch-related features, Mel spectrograms, and statistical descriptors because speech contains spectral envelope, prosodic cues, and temporal cues].

Music and music information retrieval require features that capture the timbre, harmony, pitch class, and rhythm. Chroma, CQT, Tonnetz, tempo, rhythm descriptors, MFCCs, and Mel spectrograms are therefore more relevant in music-related classification. In contrast, respiratory and cardiac sound classification often relies on MFCCs, spectrograms, Mel spectrograms, or fused representations because medically relevant sound patterns may be subtle and localized.

Bioacoustic classification may benefit from MFCCs, spectrograms, Mel spectrograms, and pitch-related descriptors because animal vocalizations often contain frequency patterns and call structures. Industrial, appliance, emergency, and forensic sound tasks may require spectral, energy, statistical, MFCC, and Mel spectrogram features, depending on whether the target signal is tonal, mechanical, impulsive, or artifact-related.

Table 4. Domain-wise feature preferences in sound classification

Application domain	Common feature types	Reason for suitability	Representative studies
Environmental/urban sound	MFCC, Mel spectrogram, STFT/spectrogram	Captures diverse non-stationary events	[11], [12], [19], [20]
Speech/voice	MFCC, pitch/F0, Mel spectrogram, statistics	Represents spectral envelope and prosody	[16], [22], [33]
Speech emotion	MFCC, pitch, RMS/energy, Mel spectrogram	Captures emotional voice variation	[16], [23]

Music/MIR	Chroma, CQT, Tonnetz, tempo, MFCC	Represents harmony, rhythm, and timbre	[2], [18], [25], [26]
Medical/respiratory/cardiac	MFCC, spectrogram, Mel spectrogram, fusion	Captures subtle biomedical acoustic patterns	[1], [4], [29], [30]
Bioacoustics	MFCC, spectrogram, pitch-related features	Represents vocal frequency and call patterns	[7], [34]
Industrial/forensic sound	Spectral, energy, MFCC, Mel spectrogram, statistics	Captures mechanical, artifact, and source cues	[5], [6], [14], [31]

6. DISCUSSION, IMPLICATIONS, LIMITATIONS

6.1 Scientific Discussion

The analysis indicates that audio feature extraction remains relevant, even in the deep learning era. Although deep learning models can learn complex patterns, many studies still depend on transformed inputs, such as MFCCs, Mel spectrograms, STFT spectrograms, or fused acoustic features. This suggests that the debate should not be framed as handcrafted features versus deep learning, but as a question of which representation best supports the model and target domain.

No single feature family is suitable for all sound classification tasks. MFCCs are compact and widely applicable; however, they may not preserve all fine-grained time–frequency information. Spectrogram and Mel spectrogram representations provide a richer structure for CNN models, but they are more sensitive to extraction parameters and computational costs. Harmonic and rhythm features are valuable in music-related tasks, whereas pitch and energy features are important in voice- and emotion-related applications. Therefore, feature selection should be guided by the acoustic characteristics of the target sound.

Feature fusion offers a way to combine complementary information, but it should be used with caution. When fusion increases dimensionality without adding meaningful information, it may increase the computational burden and overfitting risk. Studies using fusion would be strengthened by ablation analysis, which identifies the contribution of each feature group. Similarly, feature–model compatibility should be considered a design principle. A complex neural network may not compensate for a poorly chosen representation, whereas a simpler machine learning model may perform effectively when the feature set is well aligned with the task.

6.2 Practical Implications

For future sound classification studies, feature extraction should be reported in sufficient detail to support reproducibility. At minimum, researchers should specify the feature extraction library or software (for example, Librosa, pyAudioAnalysis, custom code, or another tool), the target sampling rate, frame length and hop length, FFT size, and the number of Mel bands when applicable. It is also important to report the MFCC configuration (number of coefficients and whether delta and delta–delta coefficients are used), any normalization or preprocessing applied (such as trimming, scaling, denoising, or augmentation), and the segmentation strategy used to convert continuous audio into analyzable units.

In addition, studies should clearly describe the dataset and data split, including the dataset name, class distribution, and train/validation/test partitioning, so that validity and generalization can be interpreted appropriately. Where

fused feature sets are proposed, ablation analyses comparing single-feature and fused-feature configurations under the same conditions are recommended to demonstrate whether fusion offers a genuine performance gain. Finally, fair model comparison requires evaluating classical machine learning and deep learning models under comparable preprocessing and training settings, and sharing code, configuration files, or explicit extraction parameters can substantially improve the reproducibility of published work.

6.3 Limitations

However, this analysis was limited by the reporting quality of the selected studies. Not all studies clearly reported the feature extraction parameters, dataset details, validation strategies, or performance values. Some studies also reported standard audio descriptors without explicitly naming the feature extraction library. Because the datasets, class structures, preprocessing choices, validation methods, and metrics varied across studies, a direct performance comparison was not attempted. Therefore, this study focuses on feature trends, feature–model relationships, and domain-wise interpretation rather than quantitative meta-analysis.

7. CONCLUSION

This study presents a scientific literature-based analysis of audio feature extraction methods used in machine learning and deep learning-driven sound classification. The analysis showed that MFCCs, Mel spectrograms, STFT/spectrogram representations, chroma, temporal features, spectral descriptors, pitch/rhythm features, statistical summaries, and fused representations each captured different acoustic properties. CNN-based models align well with spectrogram-like inputs, whereas classical machine learning models remain useful with compact handcrafted features. These findings suggest that feature selection should be considered both domain-sensitive and model-dependent. Future sound classification studies

8. CONFLICT OF INTEREST

The authors declare that there are no financial, commercial, institutional, or personal relationships that could have influenced the work reported in this research.

9. RESEARCH INTEGRITY STATEMENT

The authors declare that the manuscript is original, complies with the journal's plagiarism and AI usage requirements, has not been published or submitted elsewhere, and that all applicable ethical guidelines have been followed in conducting, reporting, and publishing the research.

10. REFERENCES

[1] A. Srivastava, S. Jain, R. Miranda, S. Patil, S. Pandya, and K. Kotecha, “Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease,” *PeerJ Comput. Sci.*, vol. 7, p. e369, Feb. 2021, doi: 10.7717/peerj-cs.369.

[2] Gst. A. V. M. Giri and M. L. Radhitya, “Musical Instrument Classification using Audio Features and Convolutional Neural Network,” *J. Appl. Inform. Comput.*, vol. 8, no. 1, pp. 226–234, Jul. 2024, doi: 10.30871/jaic.v8i1.8058.

[3] S. Gupta, V. Srivastava, and D. Kumar, "Environment Sound Classification using stacked features and convolutional neural network," in *Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing*, Noida India: ACM, Aug. 2024, pp. 42–50. doi: 10.1145/3675888.3676028.

[4] Z. Tariq, S. K. Shah, and Y. Lee, "Feature-Based Fusion Using CNN for Lung and Heart Sound Classification," *Sensors*, vol. 22, no. 4, p. 1521, Feb. 2022, doi: 10.3390/s22041521.

[5] D. Jayakumar, M. Krishnaiah, S. Kollem, S. Peddakrishna, N. Chandrasekhar, and M. Thirupathi, "Emergency Vehicle Classification Using Combined Temporal and Spectral Audio Features with Machine Learning Algorithms," *Electronics*, vol. 13, no. 19, p. 3873, Sep. 2024, doi: 10.3390/electronics13193873.

[6] M. A. Qamhan, H. Altaheri, A. H. Meftah, G. Muhammad, and Y. A. Alotaibi, "Digital Audio Forensics: Microphone and Environment Classification Using Deep Learning," *IEEE Access*, vol. 9, pp. 62719–62733, 2021, doi: 10.1109/ACCESS.2021.3073786.

[7] S. Carvalho and E. F. Gomes, "Automatic Classification of Bird Sounds: Using MFCC and Mel Spectrogram Features with Deep Learning," *Vietnam J. Comput. Sci.*, vol. 10, no. 01, pp. 39–54, Feb. 2023, doi: 10.1142/S2196888822500300.

[8] M. K. Gourisaria, R. Agrawal, M. Sahni, and P. K. Singh, "Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques," *Discov. Internet Things*, vol. 4, no. 1, p. 1, Jan. 2024, doi: 10.1007/s43926-023-00049-y.

[9] K. M. Rezaul *et al.*, "Enhancing Audio Classification Through MFCC Feature Extraction and Data Augmentation with CNN and RNN Models," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 7, 2024, doi: 10.14569/IJACSA.2024.0150704.

[10] M. Bhattacharjee, S. R. M. Prasanna, and P. Guha, "Speech/Music Classification Using Features From Spectral Peaks," *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1549–1559, 2020, doi: 10.1109/TASLP.2020.2993152.

[11] R. Jahangir, M. Asif Nauman, R. Alroobaea, J. Almotiri, M. Mohsin Malik, and S. M. Alzahrani, "Deep Learning-based Environmental Sound Classification Using Feature Fusion and Data Enhancement," *Comput. Mater. Contin.*, vol. 74, no. 1, pp. 1069–1091, 2023, doi: 10.32604/cmc.2023.032719.

[12] F. Demir, D. A. Abdullah, and A. Sengur, "A New Deep CNN Model for Environmental Sound Classification," *IEEE Access*, vol. 8, pp. 66529–66537, 2020, doi: 10.1109/ACCESS.2020.2984903.

[13] C. Yang, X. Gan, A. Peng, and X. Yuan, "ResNet Based on Multi-Feature Attention Mechanism for Sound Classification in Noisy Environments," *Sustainability*, vol. 15, no. 14, p. 10762, Jul. 2023, doi: 10.3390/su151410762.

[14] M. Dimbiniaina, D. P. Pau, and T. A. Naramo, "Mel Power Spectrogram Approximation By Tiny Neural Networks for Home Appliances Classification," in *2023 IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4.0&IoT)*, Brescia, Italy: IEEE, Jun. 2023, pp. 60–65. doi: 10.1109/MetroInd4.0IoT57462.2023.10180197.

[15] M. A. Nessiem, M. M. Amin, and B. W. Schuller, "SMILENets: Audio Representation Learning via Neural Knowledge Distillation of Traditional Audio-Feature Extractors," in *2023 8th International Conference on*

Frontiers of Signal Processing (ICFSP), Corfu, Greece: IEEE, Oct. 2023, pp. 32–37. doi: 10.1109/ICFSP59764.2023.10372936.

[16] P. A. Babu, V. Siva Nagaraju, and R. R. Vallabhuni, "Speech Emotion Recognition System With Librosa," in *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, Bhopal, India: IEEE, Jun. 2021, pp. 421–424. doi: 10.1109/CSNT51715.2021.9509714.

[17] R. Mushi and Y.-P. Huang, "Assessment of Mel-Filter Bank Features on Sound Classifications Using Deep Convolutional Neural Network," in *2021 International Conference on System Science and Engineering (ICSSE)*, Ho Chi Minh City, Vietnam: IEEE, Aug. 2021, pp. 334–339. doi: 10.1109/ICSSE52999.2021.9538433.

[18] G. K. Birajdar and M. D. Patil, "Speech/music classification using visual and spectral chromagram features," *J. Ambient Intell. Humaniz. Comput.*, vol. 11, no. 1, pp. 329–347, Jan. 2020, doi: 10.1007/s12652-019-01303-4.

[19] H.-C. Chu, Y.-L. Zhang, and H.-C. Chiang, "A CNN Sound Classification Mechanism Using Data Augmentation," *Sensors*, vol. 23, no. 15, p. 6972, Aug. 2023, doi: 10.3390/s23156972.

[20] Y. Wang *et al.*, "Sound as a bell: a deep learning approach for health status classification through speech acoustic biomarkers," *Chin. Med.*, vol. 19, no. 1, p. 101, Jul. 2024, doi: 10.1186/s13020-024-00973-3.

[21] L. Vujosevic and S. Dukanovic, "Deep learning-based classification of environmental sounds," in *2021 25th International Conference on Information Technology (IT)*, Zabljak, Montenegro: IEEE, Feb. 2021, pp. 1–4. doi: 10.1109/IT51528.2021.9390124.

[22] A. Yousuf and D. S. George, "Feature extraction of audio data for speaker's gender classification," *J. Phys. Conf. Ser.*, vol. 2998, no. 1, p. 012003, Apr. 2025, doi: 10.1088/1742-6596/2998/1/012003.

[23] A. Aggarwal *et al.*, "Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning," *Sensors*, vol. 22, no. 6, p. 2378, Mar. 2022, doi: 10.3390/s22062378.

[24] C. Zhen and L. Changhui, "Music Audio Sentiment Classification Based on CNN-BiLSTM and Attention Model," in *2021 4th International Conference on Robotics, Control and Automation Engineering (RCAE)*, Wuhan, China: IEEE, Nov. 2021, pp. 156–160. doi: 10.1109/RCAE53607.2021.9638811.

[25] A. R. Chhetri, K. Kumar, M. P. Muthyala, S. M R, and R. A. Bangalore, "Carnatic Music Identification of Melakarta Ragas through Machine and Deep Learning using Audio Signal Processing," in *2023 4th International Conference for Emerging Technology (INCET)*, Belgaum, India: IEEE, May 2023, pp. 1–5. doi: 10.1109/INCET57972.2023.10170568.

[26] S. Telsang *et al.*, "A Novel Voice-Processing Method Leveraging Polyphonic Audio for Instrument Recognition Using Machine Learning," in *2025 9th International Conference on Computing, Communication, Control and Automation (ICCCBEA)*, Pune, India: IEEE, Aug. 2025, pp. 1–6. doi: 10.1109/ICCUBEA65967.2025.11284237.

[27] M. R. Koravanavar and C. B. Kolanur, "Lung Sound Based Pulmonary Disease Classification Using Deep Learning," in *2023 2nd International Conference on Futuristic Technologies (INCOFT)*, Belagavi, Karnataka, India: IEEE, Nov. 2023, pp. 1–4. doi: 10.1109/INCOFT60753.2023.10425681.

[28] S. Yadav, P. Agarwal, and S. A. M. Rizvi, "Automated Multiclass Classification of Lung Diseases with Deep Learning: A Hybrid CNN-LSTM Approach," in *2024 4th International Conference on Technological*

Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan: IEEE, Nov. 2024, pp. 1622–1627. doi: 10.1109/ICTACS62700.2024.10841124.

[29] M. Deng, T. Meng, J. Cao, S. Wang, J. Zhang, and H. Fan, "Heart sound classification based on improved MFCC features and convolutional recurrent neural networks," *Neural Netw.*, vol. 130, pp. 22–32, Oct. 2020, doi: 10.1016/j.neunet.2020.06.015.

[30] M. Bahreini, R. Barati, and A. Kamali, "Cardiac sound classification using a hybrid approach: MFCC-based feature fusion and CNN deep features," *EURASIP J. Adv. Signal Process.*, vol. 2025, no. 1, p. 2, Jan. 2025, doi: 10.1186/s13634-025-01203-0.

[31] M. Olowe *et al.*, "Spectral Features Analysis for Print Quality Prediction in Additive Manufacturing: An Acoustics-Based Approach," *Sensors*, vol. 24, no. 15, p. 4864, Jul. 2024, doi: 10.3390/s24154864.

[32] V. M.R.M, A. Ingavale, V. Khullar, S. S. Tirth, B. Pandey, and H. P. Singh, "Identification of Sounds Using Deep Learning with MFCC Features Extraction," in *2024 IEEE AITU: Digital Generation*, Astana, Kazakhstan: IEEE, Apr. 2024, pp. 60–64. doi: 10.1109/IEEECONF61558.2024.10585335.

[33] N. S M and S. A, "Speech Emotion Recognition Using ML," in *2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, Bangalore, India: IEEE, Nov. 2023, pp. 1–6. doi: 10.1109/CSITSS60515.2023.10334088.

[34] M. Cheema and V. R. Basker, "Animal Sound Classification for Wildlife Surveillance," in *2025 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, Bhubaneswar, India: IEEE, Feb. 2025, pp. 497–501. doi: 10.1109/ESIC64052.2025.10962711.