An Analytical Study of Data Augmentation Across Audio Representations for Infant Cry Classification

Meriyem Ghanjaoui[1], Abdelaziz Daaif[2], Abdelmajid Bousselham[3],
Sajid Rahim[4], Ahmed Bouatmane[5], Mohamed Elyoussfi[6]
L2IAC ENSET Hassan 2 University, Mohammedia, Morocco[1, 2, 3, 5, 6]
Computing and Software, McMaster University, 1280 Main Street West, Hamilton, L8S 4L7, ON, Canada[4]

***Abstract*—Several multidisciplinary studies consider an infant's cry as a valuable source of information, particularly for parents, caregivers, and medical professionals. From a signal processing viewpoint, infant cries can be represented either in the time domain (one-dimensional or 1D raw waveform) or in the time–frequency domain (two-dimensional or 2D spectrogram-based representations). However, the impact of these representations on classification performance, particularly under constrained and imbalanced dataset conditions, remains insufficiently explored. This study presents a comparative analysis of 1D and 2D convolutional neural networks applied to waveform and spectrogram representations of infant cries. Due to the significant class imbalance of the dataset, we employed data augmentation techniques. Experimental results show that the 1D CNN achieved 95% training accuracy and 91% validation accuracy, indicating a relatively small generalization gap. In contrast, 2D CNN reached 98% training accuracy but remained below 91% on the validation set, revealing a larger gap and suggesting potential overfitting to the augmented data.**

***Keywords*—CNN; waveform; spectrogram; deep learning; baby cries**

I. Introduction

Numerous motivations drive the analysis of infant cries, particularly distinguishing normal cries, caused by hunger, thirst, sleep, or other needs, from abnormal cry patterns that may indicate illness or neurological disorders. Insights derived from cry analysis can assist parents and caregivers in responding appropriately to a baby's needs, thereby supporting healthy emotional and physical development.

In recent years, deep learning techniques have significantly advanced the field of audio analysis. Convolutional Neural Networks (CNNS), for example, have demonstrated strong performance in various audio-related tasks, including speech recognition, emotion detection, and pathological voice classification. These approaches enable automatic feature learning directly from raw or transformed audio signals, reducing dependence on handcrafted features.

However, audio signals inherently contain multiple feature representations, and deciding the most suitable representation for a specific application remains an open research question. In the context of infant cry classification, it is still unclear whether time-domain representations (raw waveform) or time–frequency representations (such as spectrograms) provide more robust and generalizable performance, especially under constrained and imbalanced dataset conditions.

The main contributions of this study are summarized as follows:

- A systematic comparative analysis between time-domain (1D waveform) and time–frequency domain (2D spectrogram) representations for infant cry classification using a fixed deep learning model.
- An experimental evaluation of data augmentation strategies applied to a highly imbalanced dataset, examining their influence on classification performance.

This study is organized as follows: Section II provides the theoretical foundations for audio processing; Section III outlines the datasets, processing, and CNN architectures used. Sections IV and V present and discuss the results, including comparisons with existing studies. Finally, Section VI concludes the study and outlines future research directions.

II. Audio Representations: Theoretical Foundations

Our ears can naturally distinguish between different sounds, voices, and music. Behind this ability lies the perception of key attributes, also called features. Human ears receive an audio signal and implicitly decode it into meaningful characteristics. Similarly, when processing audio, a machine gets a numerical waveform and attempts to extract its underlying features. Because raw audio is complex and high-dimensional, features allow machine learning models to focus on the most informative patterns rather than the entire waveform.

In general, audio features are categorized into prosodic and acoustic features, and each classification task depends on which features are most relevant.

Prosodic features focus more on how something is spoken, but not what is spoken. They include aspects such as intensity or loudness, pitch (F0), rhythm, duration, and voice quality. Acoustic features, in contrast, focus on the physical properties of the signal itself. They represent low-level measurable attributes, such as: Root Mean Square (RMS) energy, Zero Crossing Rate, Spectral centroid, Spectral roll-off, Mel-Frequency Cepstral Coefficients (MFCC), Formants, and Harmonic structure. In summary, the prosodic features are high-level abstractions derived from low-level acoustic features, which themselves originate from either a time-domain, frequency domain, or time-frequency domain representation.

A. Audio Domain Representations

Basically, audio signals can be represented in different domains, each capturing complementary information:

1) Time-domain representations: The main idea about the time domain revolves around analyzing waveforms (Fig. 1), a waveform is a graphical representation of a signal that illustrates its variation over time [1]. Accordingly, the x-axis represents time, while the y-axis represents the amplitude. giving a direct visualization of the signal's comportment by tracing the amplitude in every time unit.

Mathematically, the real-world audio waveform can be expressed as a sum of sinusoidal components (1) [1].

$$x(t) = \sum_{k=1}^{K} Ak\ sin\ sin\ (2\pi ft + \emptyset K) \quad (1)$$

It means that the time-domain signal $x(t)$is shaped by the superposition of multiple sinusoids, which is characterized by an amplitude A_k, a frequency f_k, and a phase ϕ_k.

2) Frequency-domain: The second way to represent an audio is by transforming it into the frequency domain. Frequency domain representation describes the signal from the frequency components it contains. Instead of showing how the amplitude varies over time. It reveals the pitch. fundamental frequency. Harmonic, spectral harmonic, amongst others. In this domain, signals are represented by amplitude squared or power on the Y-axis and frequency on the X-axis. This is carried out using the means of the Discrete Fourier Transform or Fast Fourier Transform (FFT) as shown in Fig. 2. Mathematically, digital signals are finite and processed with the Discrete Fourier Transform (DFT) (2) [2],[3]:

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty}\left(a_n\, cos\, \frac{n\pi x}{L} + b_n\, sin\, \frac{n\pi x}{L}\right) \quad (2)$$

3) Time-frequency domain: Time-frequency domain enables the description of how the spectral content of a signal varies over time as it merges the amplitude vs time and the amplitude vs frequency representations. The result is a two-dimensional representation with time in the X-axis, frequency in the Y-axes and the amplitude as color. This transformation is done by means of Short-Time Fourier Transform (STFT) (3), producing a spectrogram as shown in Fig. 3[4].

$$X(t,f)=\int x(\tau)w(\tau-t)e^{-j2\pi f\tau}\, d\tau \quad (3)$$

where w is a time-localizing window. The spectrogram, given by $|\ X(t,f)\ |^2$, displays energy as a function of time and frequency.

When coming to the time-frequency domain, we find three known graphical representations available in the literature. These are Spectrogram, Mel spectrogram and Constant Q transform (CQT). Each has its use case as summarized in Table I.

To summarize, audio representations depend on time and frequency domains, thus the study's hypothesis is as follows. In order to maximize the relevance of the feature extracted, it is recommended to use the time-frequency domain.

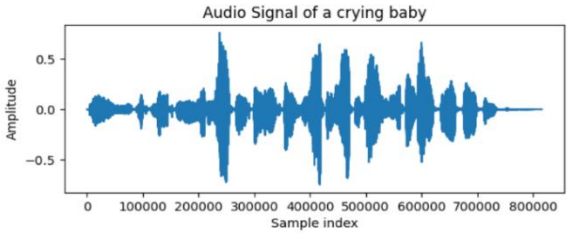

Fig. 1. An example of a crying baby waveform.

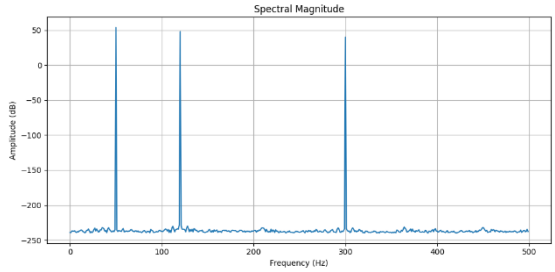

Fig. 2. Frequency domain: Spectral magnitude.

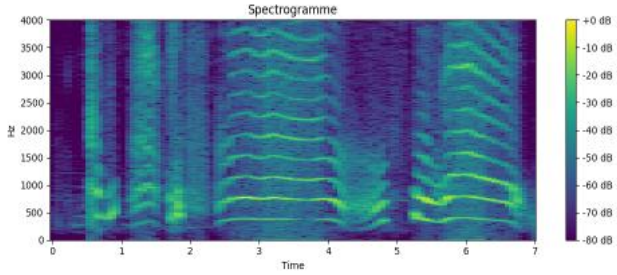

Fig. 3. A Spectrogram example of a pain cry.

TABLE I. COMPARISON OF MOST USED TIME-FREQUENCY DOMAIN REPRESENTATIONS

Representation	Frequency scale	Advantages	Disadvantages
Raw waveform (1D CNN)[5]	Time domain	Direct learning from original signal, no handcrafted transformation, reduced preprocessing	Sensitive to noise, less interpretable.
Spectrogram[6], [7]	Linear	True physical frequencies, Baseline representation.	Low frequencies are not taken into consideration, large dimensionality, and non-perceptual.
Mel [8], [9], [10]Spectrogram	Mel scale	Perceptual meaningful, Compact, Robust in feature extraction, widely used.	Low frequency precision, Possibility of losing information
CQT[11], [12]	Logarithmic	High resolution at low frequencies,	Computationally heavy

III. Related works

Several studies have mentioned the importance of increasing data size in infant cry analysis and audio classification tasks. Due to dataset imbalance and limited sample availability, many reported results fail to achieve robust generalization. For this reason, several works apply various data augmentation techniques, particularly classical methods such as time stretching, pitch shifting, and noise addition. For example, Abayomi-Alli et al. [13] conclude that noise addition is among the most commonly used augmentation techniques across existing audio classification studies.

However, despite the known use of data augmentation, selecting a suitable audio representation remains a significant challenge Table II. Some studies operate directly on one-dimensional raw waveforms, while others underline the importance of time–frequency representations.

For instance, Abdoli et al. [6] conducted experiments on the UrbanSound8K dataset (8,732 samples), using raw 1D waveform inputs with a 1D CNN architecture, achieving 89% accuracy without applying data augmentation. In contrast, Doungpaisan and Khunarsa [7] employed Constant-Q Transform (CQT) representations with a ResNet-based 2D CNN on a gunshot dataset of 2,148 samples, reporting 95% accuracy.

More recently, transformer-based approaches have gained attention. Jeong et al. [8] applied the Audio Spectrogram Transformer (AST) on spectrogram representations of Parkinson's speech data (5,000 samples), achieving 92% accuracy without augmentation. Similarly, Li et al. [9] explored speech emotion recognition using Mel-spectrograms combined with CNN/Transformer architectures on a 5,000-sample dataset, incorporating augmentation techniques and reporting 92% accuracy.

These studies prove that both representation choices significantly impact model performance. However, most works rely on relatively large datasets compared to infant cry datasets, which are often limited in size and highly imbalanced.

TABLE II. Summary of Related Works

Studies	Dataset	Samples	Representation	Model	Augmentation	Reported Accuracy
Abdoli et al. [6]	UrbanSound8K	8 732	1D waveform	1D CNN	No	89%
Doungpaisan & Khunarsa [7]	GunShot	2148	CQT	ResNet based on 2D CNN	No	95%
Jeong *et al.* [8]	Parkinson Speech	5000	Spectrogram	Audio spectrogram transformer (AST)	No	92%
Li *et al.* [9]	Speech Emotion	5 000	Mel-Spectrogram	CNN/Transformer	Yes	92%
Proposed Work	Infant Cry	475	1D & 2D	CNN	Yes	91%

IV. Methodology

Founded on the representations of audio cited earlier in this research, the objective of this experimentation is to evaluate the different representations, in other word to compare the time domain representation, which is a one-dimensional input vector, to the time-frequency domain, which is a two-dimensional vector. We consider our classification experiment using the Donate a Cry dataset [14].a public dataset of baby cries containing labeled samples for five different types of cries: 6 audio clips labeled as belly pain, 8 as burping, 16 as discomfort, 382 as hunger, and 27 as tiredness. In our case, we merged this data into two classes: hungry and not-hungry, in order to proceed with five class classification. We aim by this operation to overcome the limitation of the dataset, which is the imbalance of classes. We also applied augmentation directly on the waveform in the evaluation data: pitch shifting, time stretching, random noise injection, and small temporal shifts.

The first part consists of feeding a waveform, which is in the time domain, into a 1D CNN model in order to classify baby cries into hungry or not. In the second part, we transformed the waveform into a mel frequency, which is the time-frequency representation and fed it into a 2D CNN model, in order to do the same classification.

This setup allows us to directly test how the representations described above perform in real-world cry classification tasks.

1) Pipeline description: For our experimentation, we evaluate the time domain and the time-frequency domain by conducting 2 sets of experiments, the first set aims to apply a 1D CNN on the data waveform without augmentation, then applying 1D CNN on the waveform data after augmentation. The augmentation is done directly on the waveform by applying pitch shifting, time stretching, random noise injection, and changing speed [15].

The second setup concerns applying a 2D CNN on the three types of time-frequency domain, which are Spectrograms, Melspectrograms, and CQT. First, without augmentation, then on augmented data, similarly to the first experiment, we applied augmentation directly to the audio waveform.

In details Fig. 4:

Fig. 4. Pipeline description.

For the augmentation, we applied a conditional augmentation to make the classes balanced, and since we have different numbers in each class, we target a number of 800 per class, which is considered enough for a CNN model to train it, which means we apply (4):

$$\text{needed} = \text{target} - \text{current\_count} \quad (4)$$

and then the augmentation files are (5)

aug_per_file = ceil(needed / current_count) (5)

So, the proportions would be as follows, Table III:

TABLE III. NUMBER OF AUGMENTATION AUDIO CLIP PER CLASS

Class	Number of Audio Clips	Augmentation Per File
tired	24	needed = 800 - 24 = 776 aug_per_file = ceil(776 / 24) = 33
hungry	382	needed = 800 - 382 = 418 aug_per_file = ceil(418 / 382) = 2
discomfort	16	needed = 800 - 16 = 784 aug_per_file = ceil(784 / 16) = 49
burping	8	needed = 800 - 8 = 792 aug_per_file = ceil(792 / 8) = 99
belly pain	6	needed = 800 - 6 = 794 aug_per_file = ceil(794 / 6) = 133

The 1D CNN processes the raw audio waveform directly. The architecture consists of an input layer, a raw waveform of 7 seconds at 16 kHz, with one channel. Convolutional layers have three Conv1D layers with increasing filter sizes 16, 32, and 64 with a kernel size of 9. Each is followed by Batch Normalization and MaxPooling1D to reduce dimensionality and capture local patterns. The fully connected layer comprises a Dense layer with 128 neurons and ReLU activation. This is followed by Dropout (0.3) to reduce overfitting. Finally, an output layer is a single neuron with sigmoid activation for five classes, as shown in Fig. 5.

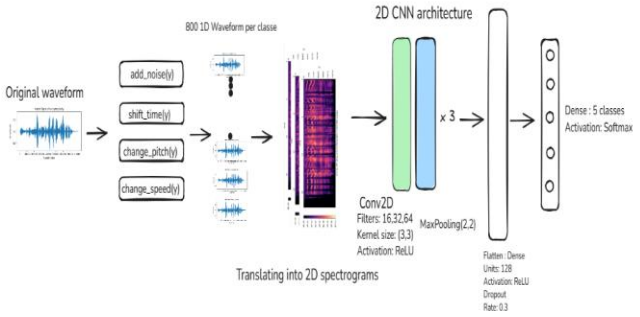

Fig. 5. The Pipeline used in the first set of experiments.

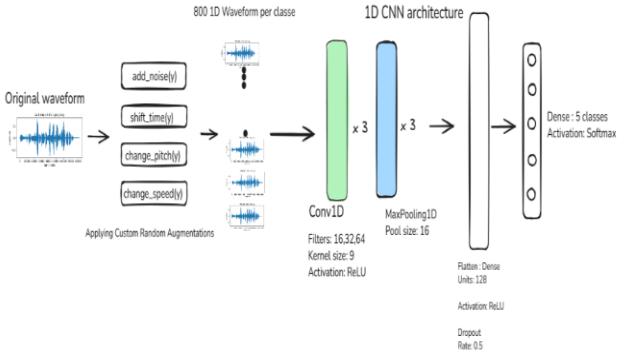

Fig. 6. Time-frequency domain 2D architecture.

The 2D CNN Time-Frequency Domain Spectrogram's architecture using 2D CNN is shown in Fig. 6. The 2D CNN architecture consists of an input layer, namely a spectrogram. The Convolutional layers comprise three Conv2D layers with increasing filters 32, 64, 128, kernel size (3×3), padding='same'. Every convolution is followed by Batch Normalization and MaxPooling2D to reduce spatial dimensions. Last layers are fully connected using a Dense layer with 128 neurons, ReLU activation, and Dropout (0.3), plus an output layer, which is a SoftMax of 5 classes.

Both architectures were trained using the Adam optimizer and categorical cross-entropy.

V. RESULTS

Before applying augmentation on the baby cries open-source dataset "Donate a Cry", the results are shown in Fig. 7. For the 2D CNN applied on spectrograms generated from real imbalanced data, it gave a good accuracy in training, reaching 97%. However, validation did not exceed 82%. 1D CNN did not reach 85% in both, with a drop at the 18th epoch. This highlights the problem of overfitting.

After augmentation, the performance of the CNN model using the time domain 1D raw waveform and time-frequency domain 2D spectrogram-based is noted in Table IV.

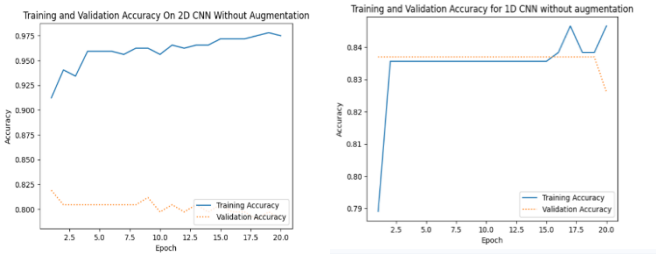

Fig. 7. Training and validation accuracy on 1D vs 2D dataset before augmentation.

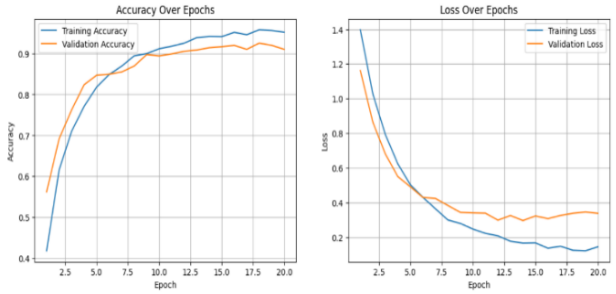

Fig. 8. Accuracy and loss over epochs for 1D CNN model on the time domain.

TABLE IV. CLASSIFICATION METRICS

	Precision	recall	F1 score	Support
Tired	0.89	0.84	0.80	163
Hungry	0.81	0.84	0.82	230
Discomfort	0.95	0.87	0.91	270
Burping	0.98	0.98	0.98	160
Belly Pain	0.92	0.99	0.95	429

The accuracy of the 1D CNN based on augmented data achieved 95% for training and 91% for validation, while the loss is 13% and stays under 20% in the validation, as shown in Fig. 8.

For the second experiment, by applying a 2D CNN on spectrograms, training accuracy is 3% better than that of a 1D CNN. The results of validation and loss curve are not favorable as shown in Fig. 9.

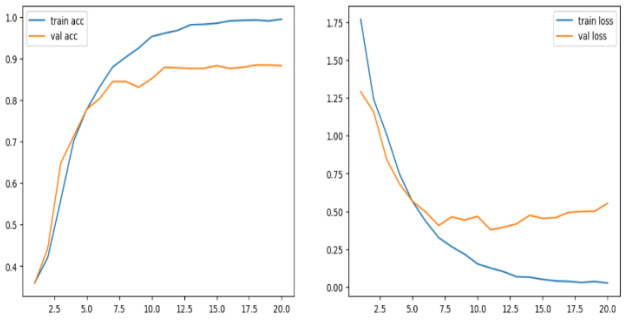

Fig. 9. Accuracy and loss over epochs for 2D CNN model on the time-frequency domain.

VI. Discussion

In this study, two types of representations were evaluated, namely, the time domain waveform and time frequency spectrogram. For this dataset, an augmentation technique was necessary. It was applied directly to the waveform, which helped both models to learn effectively. The first model, a 1D CNN applied on the raw audio waveform, achieved 95% training data and approximately 91% in validation. The reduced gap between training and validation indicates a good level of generalization. In contrast, the second model, 2D CNN applied to the time-frequency domain, particularly the spectrogram, reached 98% on the training set but remained below 91% on the validation set. This shows that the model learns well the training data, but does not generalize to unseen data, which is characteristic of overfitting behavior.

Although the time-frequency representation provides a richer source of features that can be learned, the model used has high capacity and therefore needs more data or more regularization to better validate. Meanwhile, the Time domain features extracted are likely to better classify in real scarce datasets.

VII. Limitations

Despite the promising performance of both 1D waveform and time-frequency spectrogram representations, several limitations should be acknowledged. First, the dataset used in this study is relatively small and highly imbalanced, which may affect the generalization capability of models, particularly the high-capacity 2D CNN trained on spectrogram representations.

Second, although data augmentation was employed to alleviate class imbalance and overcome data scarcity, synthetic samples cannot fully reproduce the natural variability and acoustic diversity of real-world infant cries. The heavy reliance on augmented data for minority classes may therefore influence performance stability.

Third, the choice of time–frequency transformations (e.g., Mel spectrogram or Constant-Q Transform) require careful parameter tuning, which may impact feature quality and overall model performance. Although experiments were repeated multiple times to ensure consistency, a formal cross-validation protocol and statistical significance testing were not conducted, which may limit the statistical reliability of the reported comparisons.

Finally, this study focuses exclusively on convolutional neural network architectures. Other deep learning paradigms, such as transformer-based models or hybrid architectures, were not explored and may offer additional improvements in future work.

VIII. Conclusion

In our work, we aimed to evaluate two principal representations of infant cry signals: The time domain waveform and the time frequency spectrogram, to understand how this choice influences model behavior and generalization. By comparing a 1D CNN on raw waveform to a 2D CNN on spectrograms, we verified that although time-frequency features provide richer information, they also increase the probability of overfitting when real data is limited. On the other hand, time domain features, although being simpler, enabled the model to generalize effectively.

This analysis stands as a building block for our future work, especially as we are leveraging transformer-based architecture, which relies on the quality of input representations. Moreover, the limitations of CNNs on spectrograms lead to models that are better in capturing temporal dependencies as well as experimenting with other types of representations, Mel spectrogram and CQT.

References

[1] T. F. Quatieri, Discrete-time speech signal processing: principles & practice. in Prentice Hall signal processing series. Upper Saddle River, N.J. London: Prentice Hall PTR, 2002.

[2] H. H. Donaldson, The rat; reference tables and data for the albino rat (Mus norvegicus albinus) and the Norway rat (Mus norvegicus). Philadelphia: [s.n.], 1915. doi: 10.5962/bhl.title.22441.

[3] C.-I. Páez-Rueda, A. Fajardo, M. Pérez, G. Yamhure, and G. Perilla, "Exploring the Potential of Mixed Fourier Series in Signal Processing Applications Using One-Dimensional Smooth Closed-Form Functions with Compact Support: A Comprehensive Tutorial," MCA, vol. 28, no. 5, p. 93, Sep. 2023, doi: 10.3390/mca28050093.

[4] E. Sejdić, I. Djurović, and J. Jiang, "Time–frequency feature representation using energy concentration: An overview of recent advances," Digital Signal Processing, vol. 19, no. 1, pp. 153–183, Jan. 2009, doi: 10.1016/j.dsp.2007.12.004.

[5] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-End Environmental Sound Classification using a 1D Convolutional Neural Network," 2019, arXiv. doi: 10.48550/ARXIV.1904.08990.

[6] P. Doungpaisan and P. Khunarsa, "Deep Spectrogram Learning for Gunshot Classification: A Comparative Study of CNN Architectures and Time-Frequency Representations," J. Imaging, vol. 11, no. 8, p. 281, Aug. 2025, doi: 10.3390/jimaging11080281.

[7] S.-M. Jeong, S. Kim, E. C. Lee, and H. J. Kim, "Exploring Spectrogram-Based Audio Classification for Parkinson's Disease: A Study on Speech Classification and Qualitative Reliability Verification," Sensors, vol. 24, no. 14, p. 4625, Jul. 2024, doi: 10.3390/s24144625.

[8] H. Li, J. Li, H. Liu, T. Liu, Q. Chen, and X. You, "MelTrans: Mel-Spectrogram Relationship-Learning for Speech Emotion Recognition via

Transformers," Sensors, vol. 24, no. 17, p. 5506, Aug. 2024, doi: 10.3390/s24175506.

[9] R. N. Bashir et al., "Voice pathology identification using mel spectrogram features and deep learning," SIViP, vol. 19, no. 11, p. 909, Nov. 2025, doi: 10.1007/s11760-025-04527-4.

[10] V. Sareen and S. K.R., "Speech Emotion Recognition using Mel Spectrogram and Convolutional Neural Networks (CNN)," Procedia Computer Science, vol. 258, pp. 3693–3702, 2025, doi: 10.1016/j.procs.2025.04.624.

[11] F. Wolf-Monheim, "Spectral and Rhythm Feature Performance Evaluation for Category and Class Level Audio Classification with Deep Convolutional Neural Networks," Sep. 12, 2025, arXiv: arXiv:2509.07756. doi: 10.48550/arXiv.2509.07756.

[12] J. Yan et al., "Graph Convolutional Network Based on CQT Spectrogram for Bearing Fault Diagnosis," Machines, vol. 12, no. 3, p. 179, Mar. 2024, doi: 10.3390/machines12030179.

[13] O. O. Abayomi-Alli, R. Damaševičius, A. Qazi, M. Adedoyin-Olowe, and S. Misra, "Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review," Electronics, vol. 11, no. 22, p. 3795, Nov. 2022, doi: 10.3390/electronics11223795.

[14] Bhoomika Valani, "donate-a-cry-corpus-features-dataset." Kaggle. doi: 10.34740/KAGGLE/DSV/5215181.

[15] K. Nugroho, I. H. Al Amin, N. A. Noviasari, and D. R. I. M. Setiadi, "Prosodic Spatio-Temporal Feature Fusion with Attention Mechanisms for Speech Emotion Recognition," Computers, vol. 14, no. 9, p. 361, Aug. 2025, doi: 10.3390/computers14090361.