Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# A Lightweight Audio-LLM Framework for Infant Cry Behavior Recognition Using Wav2Vec2 and TinyLlama

**Muhammad Sheraz Khan**
Kookmin University
**Falak Niaz**
Kookmin University
**Irshad Ahmad**
Islamia College University
**Jehad Ali**
Ajou University
**Ijaz Ahmad Khan**
Kookmin University
**yeong Min Jang**
yjang@kookmin.ac.kr

Kookmin University

**Article**

**Keywords:**

**Posted Date:** November 19th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11338456/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# A Lightweight Audio-LLM Framework for Infant Cry Behavior Recognition Using Wav2Vec2 and TinyLlama

**Muhammad Sheraz Khan**[1], **Falak Niaz**[1], **Irshad Ahmad**[3], **Jehad Ali**[4], **Ijaz Ahmad Khan**[1], **and Yeong Min Jang**[1,*]

[1]Department of Electronics Engineering, Kookmin University, Jeongneung-ro 77, Seoul, 02707, Seongbuk-gu, South Korea
[2]Department of Computer Science, Islamia College, Grand Trunk Rd Peshawar, Pakistan
[3]Department of AI Convergence Network, Ajou University, Suwon, South Korea
[*]Corresponding Author: Yeong Min Jang; yjang@kookmin.ac.kr

## ABSTRACT

Infant crying is a primary form of non-verbal communication through which infants express physiological and behavioral needs. However, accurately interpreting infant cries remains challenging because cry signals exhibit substantial acoustic and temporal variability, particularly under noisy conditions. To address this challenge, we propose a lightweight Audio-Large Language Model (Audio-LLM) framework for infant cry behavior recognition by integrating a pre-trained Wav2Vec2-base audio encoder with the TinyLlama-1.1B language model. Wav2Vec2-base extracts high-level acoustic representations directly from the infant cry waveform, which are temporally aggregated and transformed through a linear projection and LayerNorm into an LLM-compatible audio token. The resulting audio representation is combined with a task-specific text prompt and processed by TinyLlama, while Low-Rank Adaptation (LoRA) enables parameter-efficient adaptation of the language model. The framework performs five-class infant cry classification, including belly pain, burping, discomfort, hunger, and tiredness, using the Donate-a-Cry Corpus dataset. The model achieved 99.34% validation accuracy under the clean training condition. To further evaluate robustness, controlled noise experiments were conducted at 0.1%, 0.2%, 0.5%, and 2% noise levels, achieving validation accuracies of 95.33%, 86.00%, 71.92%, and 67.42%, respectively. These results demonstrate the effectiveness of pre-trained acoustic representations combined with a compact LLM for infant cry behavior recognition and provide a quantitative assessment of performance degradation under increasing noise conditions. The proposed framework offers a lightweight alternative to conventional approaches and establishes an Audio-LLM-based direction for infant cry analysis.

## 1 Introduction

Infant crying is one of the primary forms of communication available to newborns and can convey information associated with physiological, behavioral, and emotional states. Differences in cry characteristics may provide useful information about conditions such as hunger, discomfort, pain, and tiredness, but reliable interpretation from acoustic signals remains challenging because infant cries exhibit substantial variability across infants and recording conditions. Recent machine-learning (ML) and deep-learning (DL) studies have demonstrated the feasibility of automatically classifying infant cry signals into meaningful categories[1–3]. Automated infant-cry analysis has traditionally been described as a closed-set classification problem in which an acoustic representation is first extracted and subsequently supplied to a task-specific classifier. Recent systems have employed convolutional, recurrent, residual, and Transformer-based architectures to model the spectral and temporal characteristics of infant cries. For example, SE-ResNet-Transformer models have been investigated for infant-cry classification using engineered Mel-Frequency Cepstral Coefficient (MFCC) representations, while hybrid MFCC, Mel-spectrogram, and Tonnetz representations have been combined with residual Long-Short Term Memory (LSTM) architectures to capture complementary acoustic information[2,4]. The rapid development of LLMs has introduced another paradigm for representation learning and task adaptation. Rather than designing an independent model for every downstream task, large pre-trained language models can learn general-purpose representations and be adapted through task-specific instructions or parameter-efficient mechanisms. The scaling of pre-trained language and multimodal language models has enabled broad task adaptation through large Transformer-based representations and efficient transfer mechanisms[5]. Although the learning objective differs from acoustic pre-training, this work provides an additional example of exploiting structure in unlabeled observations. The capabilities of language models have subsequently been extended beyond text through multimodal architectures. Recent multimodal-LLM research has emphasized the integration of heterogeneous inputs within a shared language-model representation space, providing a

conceptual basis for connecting non-textual information with language-model processing[6]. For audio, this direction has led to Audio-Language Models (ALMs), in which acoustic representations are connected to language-model components to enable audio-conditioned understanding and generation. Recent ALM studies have demonstrated that pre-trained language models can be conditioned on acoustic information. Acoustic Prompt Tuning investigates learned acoustic representations for conditioning LLMs, while broader auditory-intelligence work considers unified processing of machine listening and speaking[7,8]. These studies support the use of continuous acoustic representations as inputs to language-model processing rather than restricting the system to a conventional classification head.

The Audio-LLM paradigm has continued to develop toward instruction following and general-purpose auditory systems. Recent work has examined general auditory intelligence, audio-language alignment, and biomedical audio adaptation, illustrating how acoustic representations can be combined with language-model processing for tasks beyond conventional speech recognition[8–10]. Audio-caption datasets and multimodal-language-model surveys further highlight the importance of audio-language data construction and cross-modal representation learning[5,11].

A key requirement for such architectures is a strong acoustic representation that can be transferred to a downstream task. Self-supervised speech representation learning provides an attractive solution because large amounts of unlabeled audio can be used during pretraining, reducing dependence on task-specific labeled data. Wav2Vec 2.0 introduced a self-supervised framework that learns contextualized representations directly from raw speech audio through masked latent representations and contrastive learning[12]. The resulting representations have demonstrated strong transfer to downstream speech tasks, including scenarios with limited labeled data. Computational efficiency and generalization under resource constraints are also active research concerns in other areas of ML. Niaz et al.[13] reviewed quantum and hybrid quantum-classical approaches to object detection and segmentation, identifying computational requirements, scalability, and deployment limitations as important challenges. These considerations provide broader motivation for investigating resource-efficient learning frameworks, although the underlying methods differ from the Audio-LLM architecture considered here. Large-scale pre-trained speech models have also demonstrated improved robustness across diverse acoustic conditions. Whisper uses large-scale weakly supervised training on diverse audio data and employs encoder-decoder Transformer architecture for speech recognition and related tasks[14]. Such pre-trained acoustic models are therefore relevant to infant-cry analysis, where labeled datasets are typically much smaller than the datasets used to train general-purpose speech and audio models. Adapting a language model to a specialized task also requires consideration of computational cost and the amount of available task-specific data. Full fine-tuning updates all model parameters and can become inefficient as model size increases. LoRA addresses this issue by freezing the pre-trained model weights and introducing trainable low-rank matrices into selected Transformer layers, thereby substantially reducing the number of trainable parameters during adaptation[15]. This property makes parameter- efficient adaptation particularly suitable for specialized low-resource applications.

Despite these advances, the integration of pre-trained acoustic representations with a compact language model for infant-cry behavior classification remains insufficiently explored. Existing infant-cry studies have mainly employed handcrafted acoustic representations, deep neural classifiers, attention-based architectures, or pre-trained audio encoders followed by dedicated classification models[1,2,4,16]. In parallel, Audio-LLM research has demonstrated the feasibility of connecting continuous acoustic representations with language-model processing, but has primarily focused on speech, general environmental audio, audio question answering, captioning, and broader auditory reasoning rather than five-class infant-cry behavior recognition[7,9]. This leaves a gap between pre-trained acoustic representation learning for infant cries and audio-conditioned language-model classification. To address this gap, this study proposes an Audio-LLM framework for five-class infant cry behavior classification. The proposed architecture uses Wav2Vec2-base as a pre-trained acoustic encoder, followed by temporal mean aggregation to obtain a fixed-dimensional acoustic representation. A learnable projection maps this representation into the embedding space of the TinyLlama-1.1B language model, where it is introduced as an audio token together with a task-specific textual prompt. LoRA is employed to adapt the language-model component in a parameter-efficient manner. The complete framework is evaluated on the Donate-a-Cry corpus using five infant-cry categories and is further examined under controlled acoustic-noise conditions.

The novel contributions of this study are summarized as follows:

1. A lightweight Audio-LLM framework is developed by integrating Wav2Vec2-base with the TinyLlama-1.1B language model for five-class infant cry behavior classification.

2. A learnable Audio-to-LLM projection is introduced to convert Wav2Vec2 acoustic features into an LLM-compatible audio token.

3. LoRA-based parameter-efficient fine-tuning is applied to adapt TinyLlama to infant cry classification with a reduced number of trainable parameters.

4. A preprocessing and augmentation pipeline is developed to increase training-data diversity and improve the class distribution of the Donate-a-Cry-based dataset.

5. Robustness is analyzed under controlled levels of Root-Mean-Square (RMS) scaled additive Gaussian noise (0.1%, 0.2%, 0.5%, and 2.0%), with a 0% clean condition used as the baseline.

The remaining work is organized as follows: Section 2 presents a review of recent research on infant cry classification, pre-trained acoustic representations, and audio-language models. Section 3 describes the proposed methodology, including dataset description and preparation, audio data augmentation, Wav2Vec2-based acoustic representation extraction, temporal aggregation, Audio-to-LLM projection, and the integration of TinyLlama with LoRA for five-class infant cry classification. Section 4 presents and analyzes the experimental results using classification performance metrics, confusion matrices, learning curves, and evaluation under different noise conditions, followed by a comparison with representative existing methods. Finally, Section 5 summarizes the main findings of the study and discusses potential directions for future research.

## 2 Related Work

Automatic infant cry classification has been investigated using handcrafted acoustic descriptors, conventional machine-learning classifiers, and increasingly deep neural networks. Earlier systems frequently relied on acoustic features such as MFCC, spectral descriptors, pitch, and related time-frequency characteristics, followed by task-specific classifiers. Recent systematic reviews show that infant-cry research covers both needs-related cry classification and pathological-cry detection, with substantial variation in datasets, features, classifiers, and evaluation protocols[17,18].

DL approaches have subsequently reduced the dependence on manually designed decision rules by learning discriminative patterns directly from acoustic or time-frequency representations. Hammoud et al.[1] investigated multiple acoustic representations, including MFCC, spectrogram, and scalogram representations, together with ML and DL classifiers for infant crying interpretation. Li et al.[2] proposed an SE-ResNet-Transformer architecture that combines residual learning, squeeze-and-excitation mechanisms, and Transformer-based sequence modelling for infant cry classification. Qiu et al.[4] further investigated a hybrid representation consisting of MFCC, Mel-spectrogram, and Tonnetz features and combined these representations with a residual LSTM architecture. Additional recent studies have examined lightweight and DL-based infant-cry systems, including MFCC-based Multi-Layer Perceptron/Convolutional Neural Network (MLP/CNN) classification, spectrogram-based deep models, and recent baby-cry classification architectures[19–21]. CNN/Transformer approaches have also been evaluated for baby-cry sound analysis[3]. Attention-based architectures have also been explored to capture complementary acoustic information across different feature domains. An efficient graph and attention-based infant-cry classification framework combines spectrogram, Mel-spectrogram, and MFCC information before classification and evaluates its performance on several infant cry datasets, including Donate-a-Cry[22]. These developments demonstrate the effectiveness of increasingly sophisticated feature-learning and sequence-modelling architectures. Recent studies have also investigated time-frequency and augmented representations, adaptive classification, TinyML deployment, and biomedical acoustic features for infant-cry analysis [23–27]. Nevertheless, the majority of these systems remain task-specific pipelines in which the acoustic representation is ultimately mapped to a predefined class label.

A more recent direction is to replace manually engineered acoustic features with representations obtained from pre-trained audio and speech encoders. Self-supervised representation learning is attractive for infant-cry analysis because the amount of labeled infant-cry data is relatively limited compared with the large audio corpora used to pretrain general-purpose acoustic models. Wav2Vec 2.0 learns contextualized representations from unlabeled speech and has demonstrated transferability to downstream speech tasks [12]. Whisper provides another large-scale pre-trained speech representation, and recent work has examined its fine-tuning strategies for low-resource conditions[14].

The application of these pre-trained representations to infant cry classification has recently been investigated directly. Muralikrishna et al.[16] evaluated Whisper and Wav2Vec2 representations for infant-cry classification, demonstrating the potential of pre-trained speech representations for this domain. A separate neonatal-cry study has also used wav2vec 2.0, WavLM, and HuBERT for newborn cry-signal analysis[28]. Muralikrishna et al.[16] subsequently conducted a more detailed investigation of Wav2Vec 2.0 and Whisper representations under low-resource infant-cry classification conditions. Their study examined information encoded at different model layers and evaluated statistical pooling, attentive statistical pooling, linear merging, and attention-based merging strategies. The authors reported that attention-based aggregation of Wav2Vec 2.0 representations combined with an x-vector classifier provided their strongest configuration[16].

Some studies have continued to improve infant-cry classification through specialized feature representations, adaptive architectures, and robustness-oriented learning. Jayasree and Blessy[29] introduced a Discrete Stockwell Transform-based representation and IncryNet variants for infant cry classification. Their experiments included both the Donate-a-Cry and Neo-Cry datasets and demonstrated that time-frequency representations can provide highly discriminative information for

infant cry recognition. However, the approach relies on a dedicated signal transformation followed by specialized image-based DL architectures.

Owino et al.[25] proposed an adaptive infant cry classification framework based on multi-armed-bandit modality selection and an attentive convolutional recurrent neural network(CRNN). The approach combines acoustic, spectral, and temporal information and was evaluated using infant-cry recordings from the Baby Chillanto and Donate Cry databases. The study specifically considers variability in cry patterns and background noise, highlighting the importance of robust acoustic representations for practical infant-cry recognition.

The development of LLMs has created a transition from task-specific audio classifiers toward multimodal systems in which acoustic information can be interpreted jointly with language. ALMs extend this paradigm by connecting continuous acoustic representations with language-model representations, allowing audio information to participate in language-based understanding and reasoning. A recent survey of ALMs identifies audio representation learning, audio-language pre-training, task-specific adaptation, and evaluation as major components of this rapidly developing research area[8].

Several recent studies demonstrate different strategies for connecting audio encoders with language models. Acoustic Prompt Tuning provides a learned interface between acoustic representations and LLMs, while general auditory-intelligence and DeSTA2.5-Audio studies consider broader audio-language processing and cross-modal alignment[7–9]. Biomedical audio experiments further illustrate the potential of adapting audio-language models to domain-specific acoustic features[10].

Recent research has also investigated mechanisms for aligning acoustic representations with language-model embedding spaces. Acoustic Prompt Tuning explores the use of learned acoustic representations to condition LLMs, while newer general-purpose audio-language architectures investigate cross-modal alignment and self-generated audio-text representations[7,9,30]. These approaches demonstrate that a continuous audio representation can serve as an input to a language-model processing pipeline rather than being restricted to a conventional classification head.

Despite this progress, the application of an audio-conditioned LLM to multi-class infant cry behavior recognition remains comparatively underexplored. Existing infant-cry studies have demonstrated the value of DL and pre-trained acoustic representations[1,2,16], whereas Audio-LLM studies have mainly targeted general speech, environmental audio, captioning, question answering, and auditory reasoning[7,9]. More recently, McCofie et al.[31] investigated few-shot prompting with Generative Pre-trained Transformer (GPT-4o) for infant pain classification using mel-spectrograms and Wav2Vec2 embeddings. Their results demonstrate the potential of pre-trained acoustic representations combined with Vision-Language Model (VLM) for infant-cry analysis. However, their approach relies on few-shot prompting of a large pre-trained VLM and focuses on binary pain classification. In contrast, the proposed framework introduces a learnable Audio-to-LLM projection and parameter-efficient LoRA adaptation of TinyLlama for five-class infant cry behavior recognition. The present work addresses this intersection by using Wav2Vec2-base as the acoustic encoder and TinyLlama as the language-model backbone, with a learnable projection between the two representation spaces. The reviewed literature reveals three main limitations that motivate the present study. First, conventional infant-cry classification continues to rely substantially on manually designed acoustic or time-frequency representations and task-specific classifiers[1,2,4]. Second, pre-trained acoustic encoders such as Wav2Vec 2.0 and Whisper have recently demonstrated their usefulness for infant-cry representation learning, but existing approaches primarily connect these representations to dedicated classification architectures[16]. Third, ALMs have demonstrated the ability to connect continuous acoustic information with language-model representations, yet their application to five-class infant cry behavior classification has received limited attention. Representative studies related to infant cry classification are summarized in Table 1. Accordingly, the proposed framework investigates an intermediate architecture between pre-trained acoustic representation learning and Audio-LLM processing. Wav2Vec2-base is used to extract contextualized frame-level acoustic representations, temporal mean pooling produces a fixed-dimensional representation, and a learnable projection maps the acoustic representation into the embedding space of TinyLlama. The resulting audio representation is introduced together with a textual task prompt, while LoRA provides parameter-efficient adaptation of the language-model component. This design directly connects pre-trained acoustic representation learning with language-model-based infant cry classification. Lightweight DL architectures have also shown potential for automated health-related classification. Alnafisah et al.[32] investigated a MobileNetV3-based framework for nail disease detection, highlighting the practical value of compact models for specialized classification tasks.

## 3 Methodology

The proposed framework is designed as an audio-conditioned language model for five-class infant cry behavior classification. Unlike conventional infant-cry classification pipelines that explicitly construct acoustic features and subsequently apply a dedicated classifier, the proposed approach uses a pre-trained Wav2Vec2-base encoder to obtain contextualized acoustic representations and connects these representations to the embedding space of a pre-trained TinyLlama language model. A learnable audio projection is used to perform cross-modal dimensional alignment, while LoRA is employed for parameter-efficient adaptation of the language-model component. The complete processing pipeline consists of main stages: dataset preparation, waveform preprocessing and noise augmentation, pre-trained acoustic representation extraction, temporal aggregation and

**Table 1**
Summary of representative studies related to infant cry classification, pre-trained acoustic representations, and ALM.

| Ref. | Dataset | Feature Extraction | Model | Main contribution |
|---|---|---|---|---|
| [1] | Infant cry | MFCC, spectrogram, and scalogram | ML + DL classifiers | Multiple acoustic representations for infant cry interpretation |
| [2] | Donate-a-Cry + Baby Chillanto | MFCC | SE-ResNet-Transformer | Residual, squeeze-excitation, and Transformer modelling |
| [3] | Donate-a-Cry | Mel-spectrogram | CNN + Vision Transformer | Comparison of CNN and ViT approaches |
| [4] | Donate-a-Cry | MFCC + Mel-spectrogram + Tonnetz | ResLSTM | Hybrid acoustic representation and temporal modelling |
| [7] | General audio | Learned acoustic embeddings | Audio-language model | Acoustic prompt conditioning of language models |
| [8] | General auditory tasks | Audio-language representations | Auditory intelligence framework | Unified machine listening and speaking |
| [9] | General audio | Cross-modal audio representations | Audio-language model | General-purpose audio-language alignment |
| [10] | Biomedical audio | Heart-murmur acoustic features | Audio-LLM | Domain-specific biomedical audio adaptation |
| [11] | General audio | Audio-caption representations | Audio-language data pipeline | Automated audio-caption data construction |
| [16] | Infant cry | Wav2Vec2 + Whisper | X-vector + LSTM | Layer-wise pre-trained acoustic representations and aggregation |
| [18] | Infant cry | Hand-crafted acoustic features | Deep neural networks | Comparison of deep models and hand-crafted features |
| [19] | Infant cry | MFCC | MLP + CNN | MFCC-based infant cry classification |
| [20] | Infant cry | Spectrogram + MFCC | Deep-learning models | Audio classification for baby monitoring |
| [22] | Infant cry | Spectrogram, Mel-spectrogram, MFCC | Graph + attention model | Efficient graph and attention-based classification |
| [25] | Donate Cry | Acoustic, spectral, and temporal | Attentive CRNN | Adaptive modality selection for cry classification |
| [26] | Infant cry | Compact acoustic features | TinyML system | Lightweight infant cry classification prototype |
| [28] | Newborn cry | Self-supervised representations | Wav2Vec2 + WavLM + HuBERT | Self-supervised features for neonatal cry analysis |
| [29] | Donate-a-Cry + Neo-Cry | Discrete Stockwell Transform | IncryNet | Time–frequency representation for infant cry classification |
| [30] | Multimodal tasks | Cross-modal representations | Alignment + fusion methods | Survey of multimodal alignment and fusion |
| [31] | USF-MNPAD-II | Wav2Vec2 embeddings + Mel-spectrogram | GPT-4o + VLM prompting | Few-shot infant pain classification using audio representations and a vision-language model |
| **Proposed** | **Donate-a-Cry** | **Wav2Vec2-base** | **TinyLlama-1.1B + LoRA** | **Audio-conditioned language-model framework for five-class infant cry behavior recognition** |

Audio-to-LLM projection, audio-language integration with TinyLlama, and parameter-efficient classification. The overall architecture is illustrated in Fig. 1, while the detailed feature-extraction and Audio-to-LLM processing sequence is shown in Fig. 2.

### 3.1 Dataset Collection

This study used the publicly available Donate-a-Cry corpus dataset maintained by G. Veres[33]. The dataset was originally developed through the Donate-a-Cry campaign, in which infant-cry audio recordings were uploaded through the associated mobile applications. The authors of the present study did not recruit participants or collect new infant-cry recordings; the work constitutes a secondary analysis of the existing dataset. The resulting dataset contained 457 original audio recordings. The

**Table 2**
Class-wise dataset distribution and audio augmentation configuration.

| Dataset Distribution | | | | Augmentation Configuration Applied to Each Class | | | |
|---|---|---|---|---|---|---|---|
| **Class** | **Train** | **Val** | **Aug. Train** | **Technique** | **Min Range** | **Max Range** | **Probability** |
| Belly Pain | 13 | 3 | 1187 | Time Stretch | 0.1 ms | 10.0 ms | 0.0–1.0 |
| Burping | 6 | 2 | 1194 | Time Masking | 0.0 ms | 1.0 ms | 0.0–1.0 |
| Discomfort | 22 | 5 | 1178 | Time Shifting | -1 ms | 1 ms | 0.0–1.0 |
| Hungry | 308 | 74 | 892 | Resampling | 8000 Hz | 44100 Hz | 0.0–1.0 |
| Tired | 19 | 5 | 1181 | Pitch Shifting | -12 st | 12 st | 0.0–1.0 |
| | | | | Low-Pass Filter | 150 Hz | 7500 Hz | 0.0–1.0 |
| **Total** | **368** | **89** | **5632** | *All augmentation techniques were applied to each training class.* | | | |

class distribution was 16 belly-pain, 8 burping, 27 discomfort, 382 hungry, and 24 tired recordings. This five-class distribution is also reported independently in previous peer-reviewed studies using the Donate-a-Cry corpus[1,2]. The cleaned dataset was subsequently divided into training and validation subsets before the augmentation procedure described in the next section.

### 3.2 Audio Preprocessing and Augmentation

As shown in Table 2 the training subset were augmented using time stretching, time shifting, pitch shifting, resampling, and low-pass filtering to increase acoustic variability and reduce class imbalance structure[34]. Each sample was processed at a sampling rate of 16 kHz and a maximum duration of 5 s. To construct a balanced training dataset, 1,200 samples were prepared for each class by combining the original training samples with augmented samples. For example, the Belly Pain class comprised 13 original training samples and 1,187 augmented samples, resulting in 1,200 samples. The same procedure was applied to all five classes, yielding 1,200 training samples per class and a balanced training dataset of 6,000 samples. Fig. 3 presents the class-wise distribution of the augmented training samples. Recent infant-cry recognition research has also demonstrated the value of combining lightweight architectures with hybrid augmentation strategies, including temporal, pitch, and noise transformations, for improving robustness under limited-data conditions[35]. The resulting waveform is normalized according to the preprocessing requirements of the Wav2Vec2 processor. The processed one-dimensional waveform is then directly supplied to the pre-trained acoustic encoder. To improve robustness to acoustic perturbations, controlled additive noise was introduced during the training experiments. To improve robustness to acoustic perturbations, controlled additive noise was introduced during the training experiments. The use of training-only audio augmentation, including noise injection and temporal or pitch perturbations, is also consistent with recent audio-classification studies that apply augmentation exclusively to training samples to improve robustness while preserving independent evaluation data[36]. To evaluate robustness to acoustic perturbations, controlled additive Gaussian noise was introduced during the noise experiments. The noise was generated as a zero-mean Gaussian process and scaled relative to the RMS amplitude of each input waveform. This type of controlled noise perturbation is relevant to infant-cry recognition because acoustic noise and environmental variability can substantially affect classification robustness[25]. The importance of noise-aware evaluation has also been demonstrated in other signal-classification applications.

For an input waveform $x[n]$ containing $N$ samples, the RMS amplitude is defined as in Equation (1).

$$x_{\mathrm{RMS}} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x^2[n]}. \tag{1}$$

Where $N$ is the number of waveform samples. A zero-mean Gaussian noise signal is then generated according to Equation (2).

$$v[n] \sim \mathcal{N}\left(0, x_{\mathrm{RMS}}^2\right), \tag{2}$$

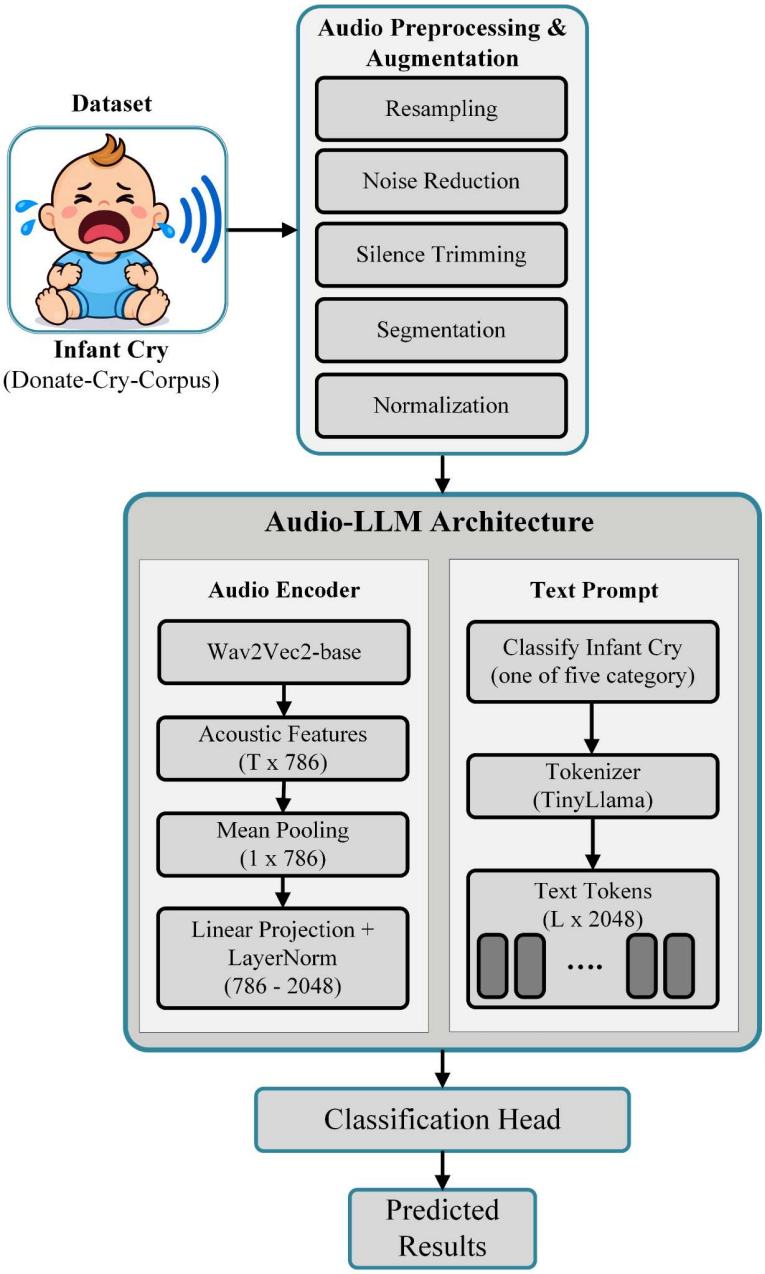



**Figure 1.** Proposed infant-cry classification framework, audio pre-processing, and audio-LLM architecture using Wav2Vec2-base audio encoder, LoRA, and TinyLlama.

The $x_{\mathrm{RMS}}$ determines standard deviation of the generated noise. The noisy waveform is subsequently obtained by scaling the Gaussian noise according to the prescribed noise percentage $p$ in Equation (3).

$$x_{\mathrm{noisy}}[n] = x[n] + \frac{p}{100} v[n], \tag{3}$$

Where $p$ represents the selected noise level. In this study, the non-zero noise conditions were 0.1%, 0.2%, 0.5%, and 2.0%, while the 0% condition was retained as the clean dataset baseline. The same noise level was specified for the corresponding training and validation experiments. The noise conditions were evaluated independently so that the effect of increasing waveform perturbation on classification performance could be quantified. The corresponding confusion matrices are presented in Section 4.

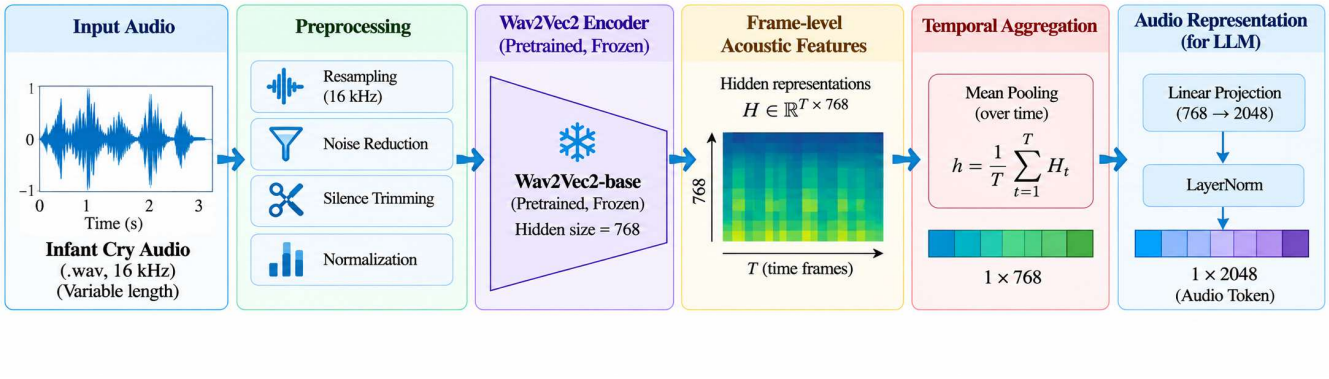



**Figure 2.** Proposed waveform frozen Wav2Vec2 feature extraction, temporal mean pooling, and projection of the acoustic representation into the 2048-dimensional LLM embedding space.

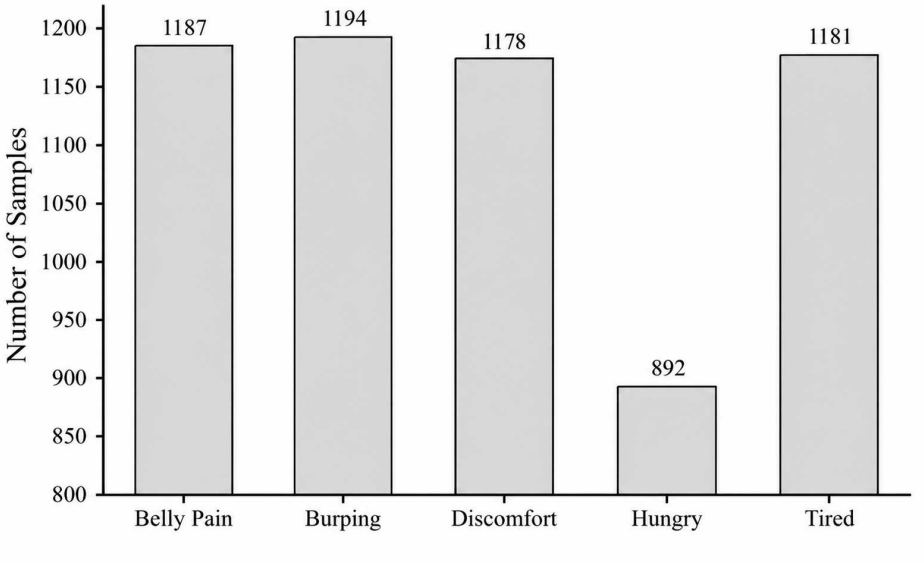



**Figure 3.** Class-wise distribution of augmented training samples for the five infant cry categories.

### 3.3 Wav2Vec2 Acoustic Representation

The proposed framework uses the pre-trained wav2vec2-base model as the acoustic representation encoder, shown in Fig. 2. The waveform is then directly supplied to the pretrained acoustic encoder, where Wav2Vec2-base is used to extract frame-level acoustic representations. Similar pretrained Wav2Vec2-based audio encoders have recently been used as modality-specific acoustic representation extractors in multimodal learning frameworks[37]. Wav2Vec2 is a self-supervised speech representation model that learns contextualized representations directly from raw waveforms[12]. In the present framework, the pre-trained encoder is used to transform each processed infant cry waveform into a sequence of contextual acoustic representations. Let the preprocessed waveform be denoted by $\mathbf{x} \in \mathbb{R}^N$. The Wav2Vec2 encoder transforms the waveform into a sequence, Equation (4).

$$\mathbf{H} = f_{\mathrm{W2V2}}(\mathbf{x}), \qquad \mathbf{H} \in \mathbb{R}^{T \times 768} \tag{4}$$

$T$ denotes the number of temporal representation frames and 768 is the dimensionality of the contextual representation produced by the base configuration.

The encoder parameters are kept frozen during task-specific training. Thus, Wav2Vec2 provides a pre-trained acoustic representation without requiring the complete acoustic encoder to be fine-tuned on the relatively small infant-cry dataset. This

design also separates general acoustic representation learning from the subsequent audio-language adaptation stage.

### 3.4 Temporal Aggregation and Audio-to-LLM Projection

The resulting acoustic representation is projected into the embedding space expected by the language model. This design follows the broader multimodal learning paradigm in which acoustic representations are aligned with pretrained language-model representations to bridge the audio–language modality gap[38]. The Wav2Vec2-base encoder produces a sequence of contextualized acoustic representations $\mathbf{H} \in \mathbb{R}^{T\times 768}$, where $T$ denotes the number of temporal frames. Since the subsequent language-model component requires a fixed-dimensional representation, temporal aggregation is performed across the frame dimension. Mean pooling is used to obtain a single acoustic representation from the complete infant cry sequence.

$$\mathbf{z} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t, \qquad \mathbf{z} \in \mathbb{R}^{768} \tag{5}$$

Equation (5) $\mathbf{h}_t \in \mathbb{R}^{768}$ denotes the contextualized acoustic representation at the $t$-th temporal frame. This operation converts the variable-length frame representation $T \times 768$ into a fixed-dimensional $1 \times 768$ acoustic vector. The resulting acoustic vector is transformed into the embedding space of the TinyLlama language model through a learnable linear projection. Projecting continuous acoustic representations into the latent space of a language model provides a practical mechanism for integrating audio with pre-trained language-model representations[39].

$$\mathbf{p} = \mathbf{W}_p\mathbf{z} + \mathbf{b}_p \qquad \mathbf{p} \in \mathbb{R}^{2048} \tag{6}$$

Equation (6) $\mathbf{W}_p \in \mathbb{R}^{2048\times 768}$ is the learnable projection matrix and $\mathbf{b}_p \in \mathbb{R}^{2048}$ is the corresponding bias vector. Layer normalization is subsequently applied:

$$\mathbf{a} = \mathrm{LayerNorm}(\mathbf{p}), \qquad \mathbf{a} \in \mathbb{R}^{2048} \tag{7}$$

The resulting $\mathbf{a}$ in Equation (7) is treated as a single LLM-compatible audio token. It is subsequently combined with the tokenized textual prompt before being processed by the TinyLlama language-model backbone.

### 3.5 Audio-Language Integration with TinyLlama

The projected acoustic representation is integrated with a pre-trained TinyLlama language model to form the audio-language processing stage. The model used in the implementation is TinyLlama-1.1B. Recent work has also demonstrated adaptation of TinyLlama with LoRA for a specialized domain, providing a relevant example of compact-LLM adaptation [40]. The textual input consists of a fixed task-specific prompt describing the required infant cry classification task.

Let the tokenized textual prompt be represented by Equation (8).

$$\mathbf{E}_{\text{text}} = \mathrm{Embed}(\mathbf{s}) \tag{8}$$

Where $\mathbf{s}$ denotes the input token sequence and $\mathbf{E}_{\text{text}}$ represents the corresponding language-model embeddings. The projected acoustic token $\mathbf{a}$ is incorporated into the input embedding sequence together with the textual representation. The resulting multimodal sequence can be expressed conceptually as, Equation (9).

$$\mathbf{E} = [\mathbf{a}; \mathbf{E}_{\text{text}}] \tag{9}$$

The combined representation is then processed by TinyLlama to obtain a contextual representation for the classification task. This design allows the acoustic signal to influence language-model processing without converting the audio representation into a conventional handcrafted feature vector.

### 3.6 Parameter-Efficient Adaptation and Classification

To adapt the pre-trained language model to infant cry classification, LoRA is employed. LoRA freezes the original pre-trained parameters and introduces trainable low-rank matrices into selected Transformer layers, reducing the number of parameters that must be updated during task-specific training[15]. Recent work has further explored reducing the memory requirements of LoRA-based adaptation, reinforcing the suitability of low-rank adaptation for resource-constrained language-model fine-tuning[41]. For a pre-trained weight matrix $\mathbf{W}_0$, the adapted weight can be represented as Equation (10).

$$\mathbf{W} = \mathbf{W}_0 + \mathbf{AB} \tag{10}$$

**A** and **B** are trainable low-rank matrices. The pre-trained matrix $\mathbf{W}_0$ remains frozen while the low-rank update is optimized for the infant-cry classification task.

The implementation uses LoRA on the TinyLlama language-model component. The training configuration contains approximately 1.126 million trainable LoRA parameters out of approximately 1.101 billion language-model parameters. A separate classification head is used to map the resulting model representation to the five infant cry classes.

$$\hat{\mathbf{y}} = \text{softmax}\left(\mathbf{W}_c\mathbf{h} + \mathbf{b}_c\right) \tag{11}$$

Where in Equation (11) $\mathbf{h}$ denotes the final task representation, $\mathbf{W}_c$ and $\mathbf{b}_c$ are the classifier parameters, and $\hat{\mathbf{y}} \in \mathbb{R}^5$ contains the class probabilities.

### 3.7 Training and Evaluation Configuration

The model was trained using a fixed random seed of 42 to improve experimental reproducibility. The optimization procedure used separate learning rates for the LoRA parameters and the classification head. As shown in Table 3, the reported tuned configuration used a learning rate of $2 \times 10^{-5}$ for LoRA parameters and $2 \times 10^{-4}$ for the classification head, together with a weight decrease of $10^{-2}$ and label smoothing of 0.05. Gradient accumulation was set in two steps. The training procedure monitored validation performance independently from the training objective. The validation set was not used for gradient updates and was used for model-selection decisions. For the robustness experiments, the same balanced validation protocol was used under the specified noise conditions. Classification performance was evaluated using overall accuracy and class-wise confusion matrices. For a validation set containing $N$ samples, accuracy is defined as, Equation (12).

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N} \times 100 \tag{12}$$

$N_{\text{correct}}$ denotes the number of correctly classified validation samples. The diagonal entries of the confusion matrix directly represent the number of correctly classified samples for each class. Additional precision, recall, and F1-score measures are reported where appropriate in the experimental evaluation.

## 4 Results and Discussion

This section presents the experimental evaluation of the proposed Wav2Vec2–TinyLlama audio-language framework for five-class infant cry classification. Recent infant-cry recognition systems have also emphasized lightweight deployment in resource-constrained environments, including real-time operation under background-noise conditions[42]. The analysis first reports the overall validation and classification performance, followed by class-wise evaluation using confusion matrices and training learning curves. The robustness of the proposed framework is then examined under controlled acoustic noise conditions of different intensities. Finally, the obtained results are compared with representative existing infant-cry classification approaches to provide a quantitative context for the proposed method. All reported results were obtained using the experimental configuration described in Section 3, with the validation.

### 4.1 Experimental Evaluation

The proposed Wav2Vec2–TinyLlama framework was evaluated using the balanced infant-cry dataset described in Section 3. The dataset contains 6,000 samples distributed equally across five cry-behavior categories: Belly Pain, Burping, Discomfort, Hungry, and Tired. The data were divided into 1200 samples per class as training subset.

The evaluation was conducted using the trained Audio-LLM framework, where Wav2Vec2-base generates the acoustic representation and TinyLlama performs language-model-based processing after audio-to-LLM projection. The language-model component was adapted using LoRA, while the Wav2Vec2 acoustic encoder remained pre-trained for acoustic feature extraction. Overall classification performance was measured using accuracy, precision, recall, and F1-score. In addition to aggregate metrics, confusion matrices were analyzed to identify class-specific prediction patterns and the principal sources of classification errors.

**Table 3**
Main training and model configuration of the proposed Audio-LLM framework.

| Category | Parameter | Configuration |
|---|---|---|
| *Data* | Dataset | Donate-a-Cry-based balanced dataset |
| | Classes | 5 |
| | Total samples | 6,000 |
| *Audio* | Sampling rate | 16 kHz |
| | Maximum duration | 5 s |
| *Encoder* | Acoustic encoder | Wav2Vec2-base |
| | Feature dimension | 768 |
| | Temporal aggregation | Mean pooling |
| | Projection dimension | 2,048 |
| *LLM* | Language model | TinyLlama-1.1B-Chat-v1.0 |
| | Adaptation | LoRA |
| *Training* | LoRA learning rate | $2 \times 10^{-5}$ |
| | Classifier-head learning rate | $2 \times 10^{-4}$ |
| | Weight decay | $1 \times 10^{-2}$ |
| | Label smoothing | 0.05 |
| | Gradient accumulation | 2 |
| | Random seed | 42 |

### 4.2 Confusion-Matrix Analysis

The class-wise behavior of the proposed framework was further examined using confusion matrices. The confusion matrix provides the number of samples assigned to each predicted class for every true class, thereby revealing both correct classifications and inter-class confusion. The diagonal elements therefore represent correctly classified samples, whereas the off-diagonal elements indicate misclassification between cry categories. The clean (0% noise) confusion matrix is presented in Fig. 6. The concentration of predictions along the main diagonal demonstrates the class-discriminative capability of the proposed audio-language representation. The off-diagonal elements are particularly useful for identifying acoustically similar cry categories. Rather than evaluating the model only through aggregate accuracy, this analysis provides class-level evidence of how the acoustic representation behaves for different infant-cry behaviors.

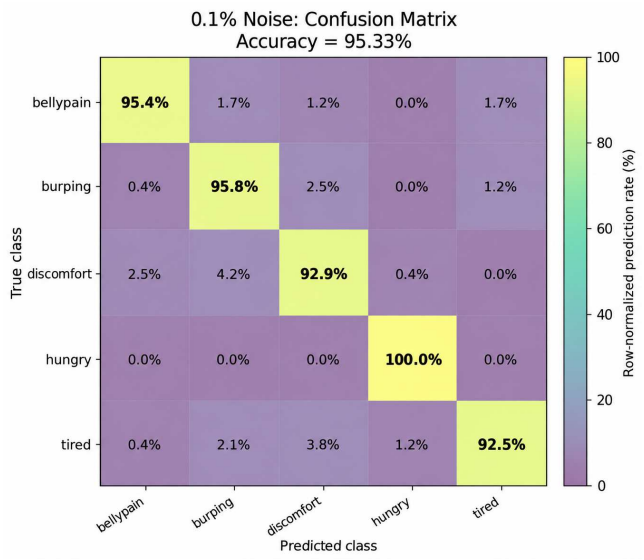
### 4.3 Learning-Curve Analysis

The training dynamics of the proposed framework were analyzed using the learning curves obtained during model optimization. Fig. 5 shows the evolution of the training and validation performance across the training epochs. During the initial training stage, the model gradually learns the relationship between the acoustic representations and the target cry categories. As optimization proceeds, the classification performance increases, indicating progressive adaptation of the trainable projection, classification head, and LoRA parameters.

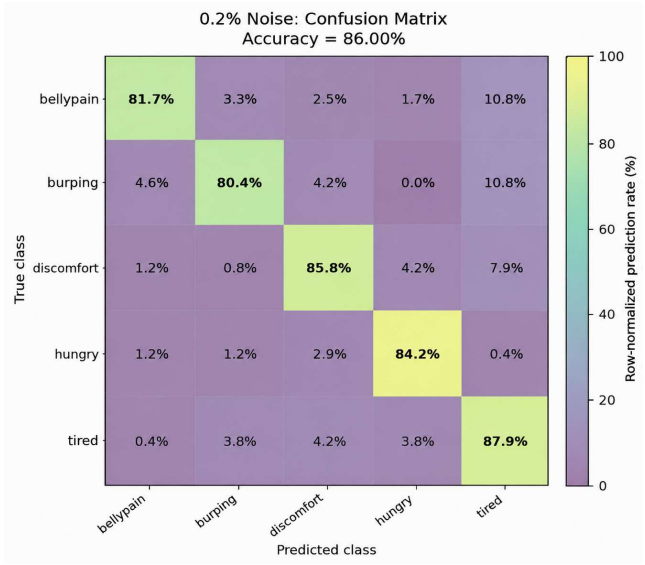
The comparison between training and validation curves is also useful for assessing overfitting. A small and stable difference between the two curves indicates that the model maintains consistent performance on unseen validation samples. Conversely, a persistent divergence between training and validation performance would indicate increasing specialization to the training data. Therefore, the learning curves are considered together with the independent evaluation rather than being used as a substitute for final generalization assessment.

### 4.4 Robustness Under Acoustic Noise

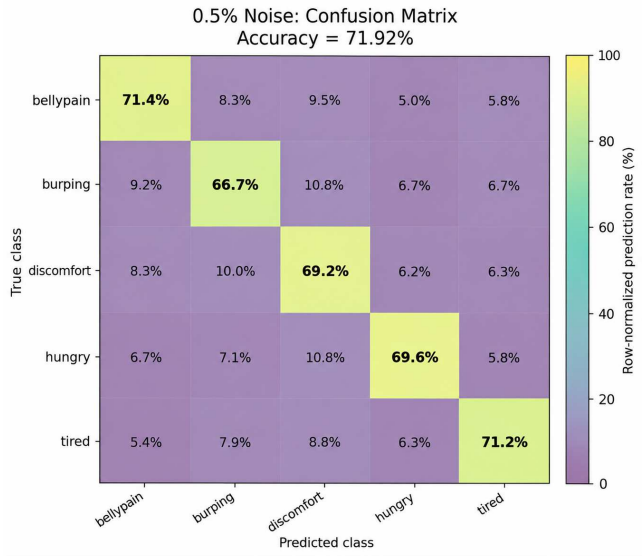
To investigate robustness against acoustic perturbation, the proposed framework was evaluated under controlled additive-noise conditions. Noise amplitude was defined relative to the RMS level of the input waveform. Four noise conditions were considered: 0.1%, 0.2%, 0.5%, and 2.0% of the waveform RMS level. The resulting confusion matrices are shown in Figs. 4a–4d. These matrices provide a direct view of how increasing acoustic perturbation changes the class-wise prediction behavior. At lower noise levels, the prediction distribution remains concentrated around the diagonal, indicating that the learned



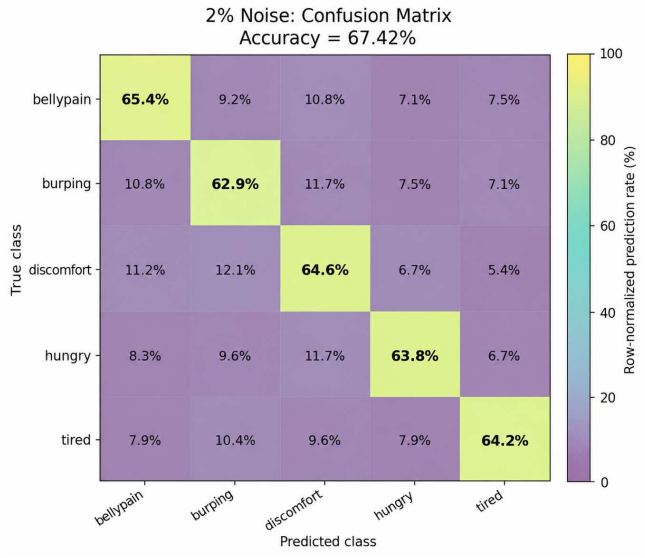

**(a)** 0.1% noise




**(b)** 0.2% noise




**(c)** 0.5% noise




**(d)** 2.0% noise

**Figure 4.** Confusion matrices of the proposed Audio-LLM framework under different waveform RMS noise conditions.

acoustic representation retains substantial discriminative information despite the perturbation. As the noise level increases, the off-diagonal entries become more pronounced, indicating increasing confusion among acoustically similar cry categories.

The quantitative robustness results should be interpreted together with the confusion matrices. In particular, the 0.5% noise experiment available in the training records produced an accuracy of 71.92%, with macro precision, recall, and F1-score of 73.17%, 71.92%, and 71.90%, respectively.

### 4.5 Combined Analysis Across Noise Conditions

The individual noise-level confusion matrices provide class-wise information for each perturbation condition, while Table 4 summarizes the overall effect of increasing acoustic noise. The results are compared across 0.1%, 0.2%, 0.5%, and 2.0% RMS noise conditions using the same independent data.

The combined analysis is intended to characterize the sensitivity of the learned acoustic-language representation to progressively stronger waveform perturbations. Performance degradation is expected to become more evident as the noise amplitude increases because the perturbation modifies the acoustic characteristics available to the Wav2Vec2 encoder. The

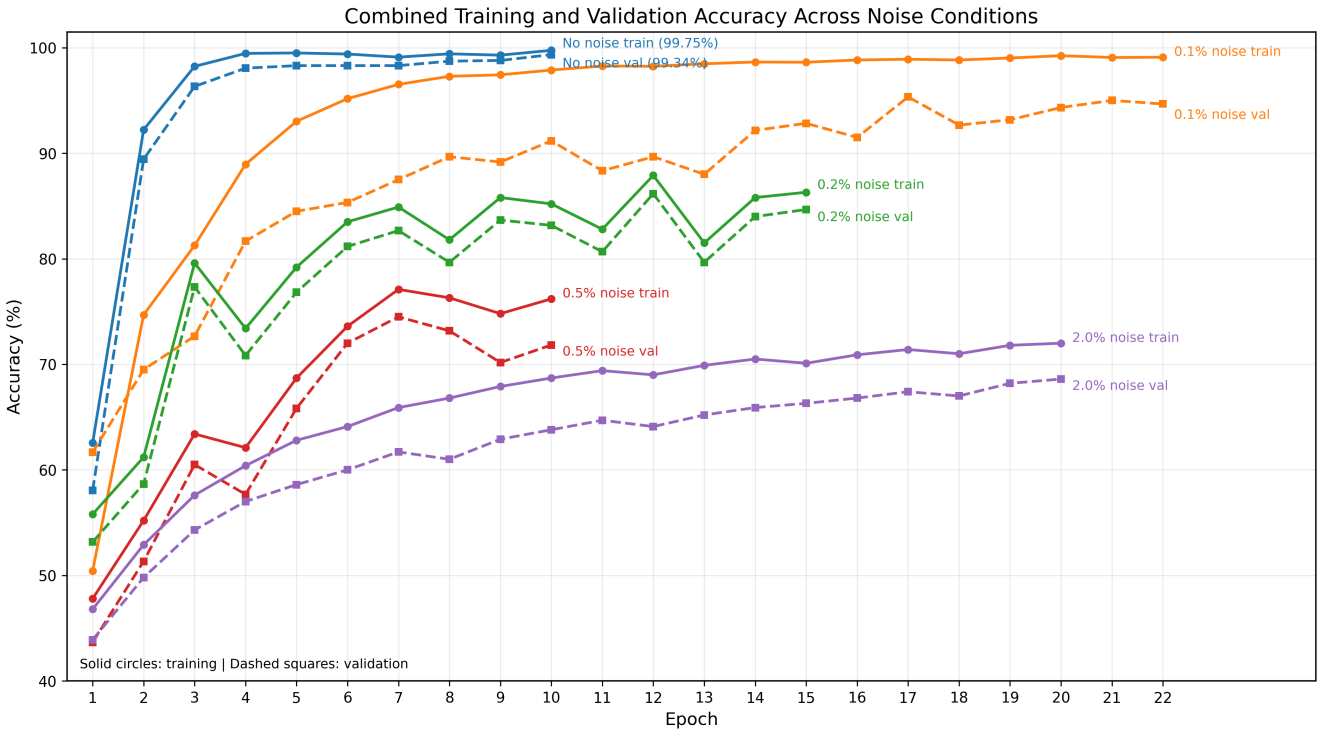



**Figure 5.** Training and validation learning curves of the proposed Audio-LLM framework, showing classification epochs.

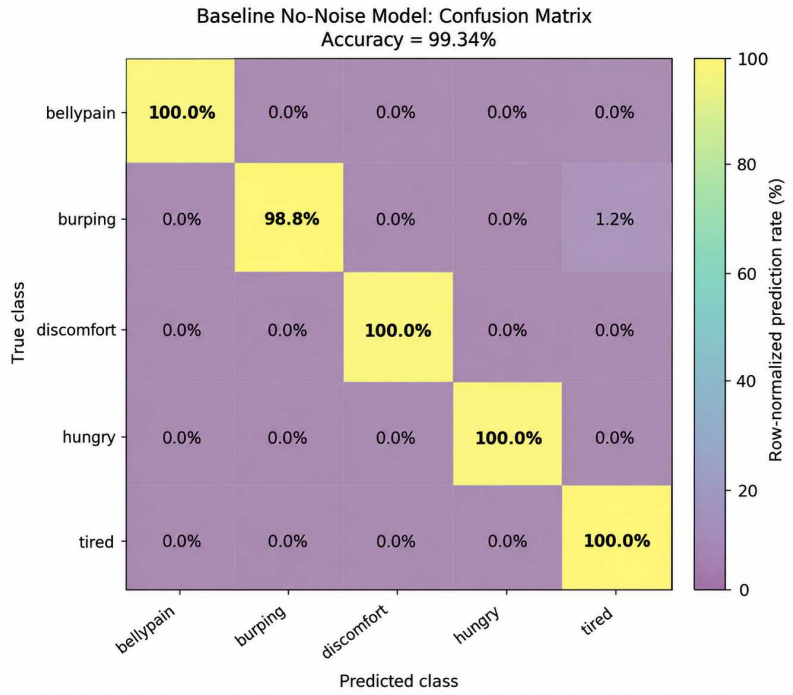



**Figure 6.** Confusion matrix without additional noise (0%), showing an overall classification accuracy.

corresponding accuracy trend and class-wise behavior are shown in Fig. 4.

This analysis complements the individual confusion matrices by providing a single view of robustness across the complete noise range. It also separates the effect of acoustic perturbation from the clean-data validation performance reported earlier.

The validation set performance was first evaluated without additional waveform noise (0%). Fig. 6 shows the corresponding

**Table 4**
Overall classification accuracy under different waveform noise conditions.

| Noise level | Correct/Total | Accuracy (%) |
|---|---|---|
| 0% | 1192/1200 | 99.34% |
| 0.1% | 1144/1200 | 95.33% |
| 0.2% | 1032/1200 | 86.00% |
| 0.5% | 863/1200 | 71.92% |
| 2.0% | 809/1200 | 67.42% |

**Table 5**
Performance comparison of the proposed framework with representative infant-cry classification methods.

| Method | Dataset | Feature | Accuracy (%) |
|---|---|---|---|
| Machine learning-based infant crying interpretation[1] | Donate-a-Cry | MFCC + ZCR + RMS | 96.39% |
| SE-ResNet-Transformer[2] | Donate-a-Cry | MFCC | 93.00% |
| CNN + Vision Transformer[3] | Donate-a-Cry | Mel-spectrogram | 98.33% |
| Hybrid features + ResLSTM[4] | Donate-a-Cry | MFCC + Mel-spectrogram + Tonnetz | 95.98% |
| Wav2Vec2 + Whisper[16] | Baby Chillanto | Wav2Vec2 + Whisper | 93.27% |
| | **Proposed Work** | | |
| **Wav2Vec2-base + TinyLlama-1.1B + LoRA** | **Donate-a-Cry** | **Wav2Vec2-base + LoRA** | **99.34%** |

confusion matrix. The model correctly classified 1192 of the 1200 training subset samples, achieving an overall accuracy of 99.34%. Belly Pain, Discomfort, Hungry, and Tired were classified without errors, while three Burping samples were incorrectly predicted as Tired. This clean-condition result is used as the baseline for the subsequent noise-robustness analysis.

### 4.6 Comparison with Existing Infant-Cry Classification Methods

The proposed framework was compared with representative infant-cry classification approaches reported in the recent literature. The comparison focuses on studies that provide a directly reported classification accuracy and use infant-cry data or pre-trained acoustic representations relevant to the present problem.

The comparison is intended to provide contextual information about the performance of different representation and modelling strategies. Differences in datasets, class definitions, data partitions, augmentation procedures, and evaluation protocols must be considered when interpreting reported accuracy values. Therefore, the values in Table 5 are reported as published and are not treated as directly equivalent measurements.

For the proposed framework, the 99.34% accuracy is reported separately from the independent results.

## 5 Conclusion and Future Work

This study presented a lightweight Audio-LLM framework for five-class infant cry behavior classification using a pre-trained Wav2Vec2-base acoustic encoder and the TinyLlama-1.1B language model. The proposed architecture transforms infant cry waveforms into contextualized acoustic representations, temporally aggregates the encoded features, and projects them into the language-model embedding space. TinyLlama then processes the projected acoustic representation together with a task-specific prompt, while LoRA provides parameter-efficient adaptation of the language-model component.

The experimental evaluation demonstrated strong classification performance on the balanced Donate-a-Cry dataset. The proposed framework achieved a validation accuracy of 99.34% on the held-out validation set. Class-wise confusion-matrix analysis further showed that the learned representation can distinguish the five target cry categories, while the noise experiments demonstrated the effect of progressively increasing waveform perturbation on classification performance. The accuracy decreased from 95.33% under 0.1% RMS noise to 86.00%, 71.92%, and 67.42% under 0.2%, 0.5%, and 2.0% noise, respectively. These results indicate that acoustic perturbation remains an important factor affecting infant-cry classification and should be considered when developing models for practical recording environments.

The main contribution of this work is the investigation of an audio-conditioned language-model architecture for infant cry behavior recognition. Unlike conventional approaches that directly connect handcrafted or pre-trained acoustic features to task-

specific classifiers, the proposed framework establishes an explicit audio-to-language-model interface between Wav2Vec2-base and TinyLlama. This provides a foundation for exploring language-model based processing of infant acoustic signals while maintaining parameter-efficient adaptation through LoRA. Despite the promising validation performance, several limitations remain. The experiments were conducted using a single infant-cry dataset and five predefined cry categories. The noise evaluation was performed using controlled waveform perturbations and therefore does not fully represent the variability of real household, clinical, or mobile recording environments. In addition, the present study focuses on classification rather than clinical diagnosis or comprehensive infant-state interpretation. Consequently, the reported results should be interpreted as evidence of classification capability rather than as a clinical decision-making system.

Future work will investigate cross-dataset and cross-subject generalization using more diverse infant-cry recordings. Broader evaluation using heterogeneous datasets is also important for understanding model robustness. For instance, Niaz et al.[43] introduced UAV-PanoSeg, a dual-modality benchmark for evaluating panoptic segmentation under varying acquisition conditions. Although focused on aerial vision, the study illustrates the role of diverse benchmark data in assessing perception systems across changing environments. Additional real-world acoustic conditions, including environmental and device-dependent distortions, can be incorporated to evaluate robustness more comprehensively. Further research can also explore larger or more capable ALMs, improved Audio-to-LLM alignment mechanisms, multimodal inputs, and explainable outputs for caregiver-oriented infant monitoring applications. Future extensions may also investigate explainable and reliability-aware decision mechanisms. These directions could extend the proposed framework from five-class acoustic classification toward broader and more context-aware infant audio understanding.

## 6 Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No.2022R1A2C1

## 7 Authorship contribution statement

M.S.K. writes reviews & edits and writes original drafts, visualization, data curation, and conceptualization. F.N. and I.A. contributed to writing and organizing the manuscript content. J.A. contributed to methodology, formal analysis, and validation. I.A.K contributed to writing, Y.M.J. advised, acquired funding, and supervised.

## 8 Data availability

The dataset is publicly available here:
https://github.com/gveres/donateacry-corpus/tree/master/donateacry_corpus_cleaned_and_updated_data

## 9 Code Availability Statement

The code supporting the findings of this study is available at:
https://github.com/msherazkhan/A-Lightweight-Audio-LLM-Framework-for-Infant-Cry-Behavior-Recognition-Using-Wav2Vec2-and-TinyLlama.git
Python 3.11.2:
https://www.python.org/downloads/release/python-3112/
Librosa 0.10.1:
https://github.com/librosa/librosa/compare/0.10.0...0.10.1
OpenCV 4.8.0
https://github.com/opencv/opencv/tree/4.8.0
TensorFlow:
https://www.tensorflow.org/
NumPy:
https://numpy.org/
Scikit-learn:
https://scikit-learn.org/stable/
Matplotlib:
https://matplotlib.org/3.9.3/index.html

## 10 Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References


1. Hammoud, M., Getahun, M. N., Baldycheva, A. & Somov, A. Machine learning-based infant crying interpretation. *Front. Artif. Intell.* **7**, 1337356 (2024).
2. Li, F., Cui, C. & Hu, Y. Classification of infant crying sounds using se-resnet-transformer. *Sensors* **24**, 6575 (2024).
3. Younis, S. A., Sobhy, D. & Tawfik, N. S. Evaluating convolutional neural networks and vision transformers for baby cry sound analysis. *Futur. Internet* **16**, 242 (2024).
4. Qiu, Y. *et al.* Classification of infant cry based on hybrid audio features and reslstm. *J. Voice* (2024).
5. Jin, Y. *et al.* Efficient multimodal large language models: A survey. *Vis. Intell.* **3**, 27 (2025).
6. Xiao, H. *et al.* A comprehensive survey of large language models and multimodal large language models in medicine. *Inf. Fusion* **117**, 102888 (2025).
7. Liang, J. *et al.* Acoustic prompt tuning: Empowering large language models with audition capabilities. *IEEE Transactions on Audio, Speech Lang. Process.* **33**, 949–961 (2025).
8. Wang, S. *et al.* Towards general auditory intelligence for machine listening and speaking. *Nat. Mach. Intell.* **8**, 1201–1220 (2026).
9. Lu, K.-H. *et al.* Desta2.5-audio: Toward general-purpose large audio language model with self-generated cross-modal alignment. *IEEE Transactions on Audio, Speech Lang. Process.* **34**, 2062–2076 (2026).
10. Florea, A., Jiang, X., Mesgarani, N. & Jiang, X. Exploring finetuned audio-llm on heart murmur features. *Smart Heal.* **36**, 100557 (2025).
11. Bai, J. *et al.* Audiosetcaps: An enriched audio-caption dataset using automated generation pipeline with large audio and language models. *IEEE Transactions on Audio, Speech Lang. Process.* **33**, 2817–2829 (2025).
12. Kunešová, M., Zajíc, Z., Šmídl, L. & Karafiát, M. Comparison of wav2vec 2.0 models on three speech processing tasks. *Int. J. Speech Technol.* **27**, 847–859 (2024).
13. Niaz, F., Utama, I. B. K. Y. & Jang, Y. M. Quantum machine learning for object detection and segmentation: A comprehensive survey and future research directions. *ICT Express* (2026).
14. Liu, Y., Yang, X. & Qu, D. Exploration of whisper fine-tuning strategies for low-resource asr. *EURASIP J. on Audio, Speech, Music. Process.* **2024**, 29 (2024).
15. Mao, Y. *et al.* A survey on lora of large language models. *Front. Comput. Sci.* **19**, 197605 (2025).
16. Muralikrishna, H., Angra, A., Dileep, D., Madhushankara, M. & Reddy, C. S. Exploring layer-wise features of whisper and wav2vec 2.0 models and their aggregation for infant cry classification in low-resource conditions. *Int. J. Inf. Technol.* 1–14 (2025).
17. Owino, G. & Shibwabo, B. Advances in infant cry paralinguistic classification—methods, implementation, and applications: Systematic review. *JMIR Rehabil. Assist. Technol.* **12**, e69457 (2025).
18. Ozseven, T. Infant cry classification by using different deep neural network models and hand-crafted features. *Biomed. Signal Process. Control.* **83**, 104648 (2023).
19. Abbaskhah, A., Sedighi, H. & Marvi, H. Infant cry classification by mfcc feature extraction with mlp and cnn structures. *Biomed. Signal Process. Control.* **86**, 105261 (2023).
20. Herlea, D. M., Iancu, B. & Ardelean, E.-R. A study of deep learning models for audio classification of infant crying in a baby monitoring system. *Informatics* **12**, 50 (2025).
21. Dudhal, J., Shinde, S. & Kharade, K. Infant crying detection and classification using deep learning. *Int. J. on Adv. Comput. Eng. Commun. Technol.* **14**, 97–101 (2025).
22. Qiao, X. *et al.* Infant cry classification using an efficient graph structure and attention-based model. *Kuwait J. Sci.* **51**, 100221 (2024).
23. Jayasree, T. & Blessy, S. Msdp-gapcnn: Modwt based statistical distribution patterns and global average pooling convolutional neural network for improved infant cry classification. *Traitement du Signal* **42**, 1085 (2025).
24. Ozcan, T. & Gungor, H. Baby cry classification using structure-tuned artificial neural networks with data augmentation and mfcc features. *Appl. Sci.* **15**, 2648 (2025).
25. Owino, G., Kamanu, T., Ndiritu, J. & Kikechi, C. B. Adaptive infant cry classification using multi-armed bandit modality selection in an attentive convolutional recurrent neural network model. *Complex & Intell. Syst.* **11**, 375 (2025).

**26.** Eldine, O. B., Selim, M., El-Makky, N. & Ghanem, N. M. Tinyml prototype of infant cry classification system. *IEEE Access* (2025).

**27.** Aggarwal, G., Jhajharia, K., Izhar, J., Kumar, M. & Abualigah, L. A machine learning approach to classify biomedical acoustic features for baby cries. *J. Voice* **39**, 1446–1455 (2025).

**28.** Shayegh, S. V. & Tadj, C. Deep audio features and self-supervised learning for early diagnosis of neonatal diseases: Sepsis and respiratory distress syndrome classification from infant cry signals. *Electronics* **14**, 248 (2025).

**29.** Jayasree, T. & Blessy, S. Infant cry classification via deep learning based infant cry networks using discrete stockwell transform. *Eng. Appl. Artif. Intell.* **160**, 112008 (2025).

**30.** Li, S. & Tang, H. Multimodal alignment and fusion: A survey. *Int. J. Comput. Vis.* **134**, 103 (2026).

**31.** McCofie, A., Kandiyana, A., Mouton, P. R., Sun, Y. & Goldgof, D. B. Few-shot prompting with vision language model for pain classification in infant cry sounds. In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, 857–862 (2025).

**32.** Alnafisah, K. H., Niaz, F., Ahmad, I., Faheem, Z. B. & Ali, J. Intelligent detection of nail diseases using an ai-powered approach. *Results Eng.* 110084 (2026).

**33.** Veres, G. Donateacry corpus: An infant cry audio corpus. https://github.com/gveres/donateacry-corpus. GitHub repository, accessed October 1, 2026.

**34.** Setyonogrum, N. G., Utami, E., Kusrini & Wibowo, F. W. Improving infant cry recognition using mfcc and cnn-based audio augmentation. *J. Tek. Informatika* **6**, 995–1016 (2025).

**35.** Zhang, M. *et al.* Enhancing infant cry recognition using lightweight cnn with hybrid feature augmentation. *Biomed. Signal Process. Control.* **121**, 110367 (2026).

**36.** Chourasia, N., Lamba, C. S. & Gupta, A. K. A 1d-cnn with advanced data augmentation for robust speech emotion recognition. *Sci. Reports* **16**, 27744 (2026).

**37.** Alrayzah, A. Tri-mca fusion: cross-modal attention and dynamic gating for multimodal sentiment analysis. *Sci. Reports* **16**, 25201 (2026).

**38.** Yu, J. Acoustic feature decoupling and pre-trained language model integration for music-to-text cross-modal generation. *Sci. Reports* (2026).

**39.** Wang, W., Hou, R., Chang, H., Shan, S. & Chen, X. MATS: An audio language model under text-only supervision. In *Proceedings of the 42nd International Conference on Machine Learning*, vol. 267 of *Proceedings of Machine Learning Research*, 63016–63035 (PMLR, 2025).

**40.** Bassani, A. *et al.* Adapting a pretrained llm to rapidly recover earthquake magnitude and location. *J. Geophys. Res. Mach. Learn. Comput.* **3**, e2026JH001578 (2026).

**41.** Zhou, Z., Zhang, Q., Kumbong, H. & Olukotun, K. LowRA: Accurate and efficient LoRA fine-tuning of LLMs under 2 bits. In *Proceedings of the 42nd International Conference on Machine Learning*, vol. 267 of *Proceedings of Machine Learning Research*, 79570–79594 (PMLR, 2025).

**42.** Mpia, H. N. *et al.* Lightweight deep learning system for infant-cry recognition with real-time notification in resource-constrained environments. *Mach. Learn. with Appl.* **22**, 100791 (2025).

**43.** Niaz, F., Utama, I. B. K. Y., Ko, S. M. & Jang, Y. M. Uav-panoseg: A large-scale dual-modality benchmark for uav-based panoptic segmentation. *Alex. Eng. J.* **152**, 77–93 (2026).