Published *online* on the journal's web page : http://innovatics.unsil.ac.id

Innovation in Research of Informatics (INNOVATICS)

| ISSN (Online) 2656-8993 |

Data Augmentation Strategies on Spectrogram Features for Infant Cry Classification Using Convolutional Neural Networks

Alam[1], Nuk Ghurroh Setyoningrum[2], Robby Maududy[3], Dea Dewi Damayanti[4], Hilmi Rahmawati[5], Mae B. Lodana[6]*

[1,2,3,4,5]*Faculty of Science and Technology, Universitas Cipasung Tasikmalaya, Jl Borolong Ciawi, Tasikmalaya 46466, Indonesia*
[6]*College of Information and Communication Technology, STI West Negros University, Bacolod City, Negros Occidental, Philippines*

[1]*alam@uncip.ac.id*, [2]*nuke@uncip.ac.id*, [3]*robby.maududy@uncip.ac.id*, [4]*deadamayantid3@gmail.com*, [5]*hilmirhmawti2@gmail.com*, [6]*mae.lodana@wnu.sti.edu*

ARTICLE INFORMATION

Article History:
Received: September 8, 2025
Last Revision: November 6, 2025
Published *Online*: November 30, 2025

KEYWORDS

Data Augmentation,
Spectrogram,
Audio Classification,
Convolutional Neural Networks (CNN),
Infant Cry

CORRESPONDENCE

Phone: 081321578291
E-mail: *nuke@uncip.ac.id*

ABSTRACT

Infant cry classification is an important task to support parents and healthcare professionals in understanding infants' needs, yet the challenge of limited and imbalanced datasets often reduces model accuracy and generalization. This study proposes the application of diverse audio data augmentation strategies including time stretching, time shifting, pitch scaling, and polarity inversion combined with spectrogram representation to enhance Convolutional Neural Network (CNN) performance in classifying infant cries. The dataset from the Donate-a-Cry Corpus was expanded from 457 to 6,855 samples through augmentation, improving class balance and variability. Experimental results show that the CNN's overall accuracy increased from 85.0% (without augmentation) to 99.85 % with augmentation (+14.85 percentage points), while the macro-averaged precision, recall, and F1-score each achieved ≥0.99 across all categories, indicating near-perfect classification. The confusion matrix further confirms robust classification with minimal misclassifications. These findings demonstrate that data augmentation is crucial to overcoming dataset limitations, enriching acoustic feature diversity, and reducing model bias, while offering practical implications for the development of accurate, reliable, and real-world applicable infant cry detection systems.

1. INTRODUCTION

Communication constitutes a fundamental element of human existence, beginning as early as the moment of birth. Crying is the main means for babies to convey their needs and discomfort, ranging from hunger, pain, to other uncomfortable conditions [1]. This phenomenon poses a great challenge for parents, especially new parents, as they often struggle to understand the meaning of each cry. Recent research trends show that artificial intelligence-based approaches, particularly through spectrogram feature extraction and classification using a Convolutional Neural Network (CNN), have great potential in aiding the interpretation of infant cries [2]. In this context, the limited variety of audio data is still an actual problem that can hinder model performance. Therefore, the integration of audio data augmentation techniques with spectrogram and CNN is important to investigate, as it not only offers academic novelty in speech signal processing [3], but also practical relevance in supporting the welfare of mothers and babies through a more accurate and reliable cry detection system [4].

Although various previous studies have examined the classification of infant cries using artificial intelligence methods, there are still significant limitations that need to be addressed. Most of the previous studies have only utilized conventional augmentation techniques such as time stretching and pitch shifting, which have not been able to fully enrich the data variety to optimize model performance [5]. Some new approaches have been introduced, but their application in infant cry classification is still limited and tends to be done in controlled test environments. The complexity of audio signals that are susceptible to environmental noise also adds a major challenge that has not been adequately answered. This creates a research gap in the form of the need for more

diverse data augmentation strategies and more comprehensive evaluation [6]. This research addresses this gap by building a CNN model based on spectrogram representation using augmented datasets, and evaluating its performance not only with accuracy, but also other metrics such as precision, recall, F1-score, and confusion matrix [7]. Thus, this research is expected to be able to produce a more robust and applicable classification model, while making a real contribution to the development of effective, accurate, and ready-to-implement infant cry detection technology in real-world conditions.

Several previous studies have focused on the classification of infant cries using artificial intelligence-based approaches. Vankudre et al. examined the use of video clips to recognize infant emotions and showed that a combination of visual and audio analysis can improve identification accuracy [8]. Nuk Ghurroh et al. highlighted the shift from traditional methods to deep learning approaches such as CNN and RNN, which proved to be superior in recognizing the acoustic patterns of crying [9]. Another study by Li Zhang et al. introduced a new variant of Self-Supervised Audio Spectrogram Transformer (SSAST) with a dual representation strategy to produce a more informative audio representation [10]. Meanwhile, Qasem et al. proposed a spectrogram flipping technique to create more realistic variations of audio data, and Aastha et al. discussed the use of tempo, volume, and speed perturbation augmentation techniques to distinguish normal and pathological infant cries [11]. These studies show that infant cry classification is progressing towards deep learning methods with a focus on spectrogram representation and augmentation strategies.

However, there are some weaknesses that remain a concern. Most previous studies have only used conventional data augmentation techniques such as time stretching and pitch shifting, which have not enriched the dataset variety enough to consistently improve model performance. Some innovations such as spectrogram flipping or patch-wise pooling do offer potential, but their application in infant cry classification is still limited and not widely explored. In addition, most model evaluations focus only on accuracy metrics, without reviewing more comprehensive indicators such as precision, recall, F1-score, and confusion matrix. These limitations emphasize the need for new research that not only enriches the variety of datasets with more diverse augmentation techniques but also presents a more holistic evaluation framework to produce CNN models that are more robust and applicable in real conditions.

The objective of this research is to design a CNN-based infant cry classification system employing spectrogram representations derived from augmented audio inputs. This work particularly emphasizes improving dataset richness and quality through diverse augmentation methods, enabling the CNN to adapt more effectively to variations in audio signals and external noise disturbances. The key contribution of this study is the combined use of augmentation strategies with an extensive evaluation framework including accuracy, precision, recall, F1-score, and confusion matrix which facilitates both holistic performance measurement and the examination of model strengths and weaknesses across multiple categories. This work demonstrates how systematic data augmentation improves CNN robustness in infant cry recognition tasks. The contribution offered is not only academic, in the form of enriching the literature related to the classification of infant cries with a deep learning approach, but also practical, with the potential implementation of a infant cry detection system that can help parents and medical personnel respond to the needs of babies more quickly, precisely and reliably.

The subsequent sections are organized as follows. Section 2 outlines a short survey of related research, emphasizing earlier approaches to infant cry analysis, audio augmentation techniques, and spectrogram-based feature extraction methods. Section 3 describes the proposed research methodology, including infant cry dataset collection, audio data augmentation strategies, spectrogram transformation, and the CNN model architecture. Section 4 presents the experimental findings and evaluation, focusing on classification performance in terms of accuracy, precision, recall, F1-score, and the confusion matrix. Section 5 provides a discussion of the findings, compares the results with prior studies, and highlights the practical implications for infant care and monitoring. The paper is concluded in Section 6, which also outlines prospective research directions aimed at advancing methods for infant cry classification.

2. RELATED WORK

Classification of infant crying is important in audio and health research as it is the primary means by which infants communicate needs or discomfort [4], [12]. Proper understanding of crying patterns helps parents and medical personnel respond quickly, thereby preventing stress, delayed treatment, and risks to the infant's health. Therefore, the development of an automated system based on artificial intelligence has both academic and practical value in supporting the well-being of mothers and children.

Several previous studies have examined the classification of infant crying using various artificial intelligence-based methods. Vankudre G et al. examined the use of video clips to recognize infant emotions, showing that a combination of visual and audio analysis can improve the accuracy identification[8]. Nuk Ghurroh et al. highlighted the evolution of methods from traditional techniques towards deep learning-based approaches such as CNN and RNN, which were shown to have higher accuracy in recognizing acoustic patterns of cries[9].The study also highlighted the main limitations in dealing with noise interference and the complexity of the cry signal [3].

Research from Li Zhang et al. introduced a new variant of SSAST (Self Supervised Audio Spectrogram Transformer) using a patch-wise combination of pooling (mean, max, min) to form a more informative dual representation [13],[10],[14], [15]. Meanwhile, Qasem et al introduced a new audio augmentation technique called spectrogram flipping. This technique involves horizontally flipping the audio spectrum and re-converting it to the time domain to generate new audio data that is realistic and variable[11]. Research from Aastha et al. discusses the use of three data augmentation techniques-tempo, volume, and speed perturbation-to improve the classification of infant cries between normal and pathologic [7],[16],[17].

Research in the field of audio data augmentation has proven that techniques such as time stretching, pitch shifting, and noise injection can increase dataset variety and reduce the risk of overfitting machine learning models [18],[19],[20],[21],[22]. Therefore, this research integrates audio data augmentation, spectrogram, and CNN to improve the accuracy of infant crying classification, which is expected to make a significant contribution in artificial intelligence-based infant monitoring technology. Although various studies have explored the classification of infant cries using artificial intelligence methods, there are still some limitations that need to be addressed. The data augmentation approaches used in previous studies are still limited to conventional techniques such as time stretching and pitch shifting, which have not fully optimized time stretching and pitch shifting, which have not fully optimized dataset variations to improve model performance. Some recent studies have proposed new augmentation techniques, such as spectrogram flipping and patch-wise pooling, but their application in infant cry classification is still not widely explored. For this reason, it is important to integrate more diverse data augmentation techniques with the utilization of spectrograms and CNNs, to improve the accuracy and robustness of the model against noise, which ultimately contributes to the development of technology supporting the welfare of mothers and babies.

3. METHODOLOGY

This research uses a data augmentation approach on spectrogram features to improve the performance of Convolutional Neural Networks (CNN)-based infant cry classification. The process begins with the collection of infant cry datasets which are then processed through augmentation techniques to enrich data variation and reduce the risk of overfitting. The augmented audio data is transformed into spectrograms, such as Mel-spectrogram and Log-Mel spectrogram, which represent time-frequency information in a visual, image-like format. The generated representation serves as input for a CNN model tailored to extract visual features from the spectrogram. To assess the robustness of the system in classifying different types of infant cries, a thorough performance evaluation is conducted using accuracy, precision, recall, F1-score, and confusion matrix, as illustrated in Figure 1.

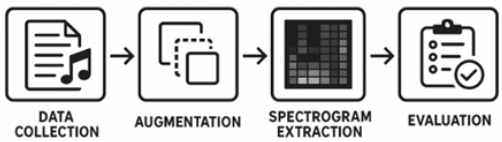

FIGURE 1. CNN-BASED INFANT CRY CLASSIFICATION FLOWCHART

3.1 Audio Data Augmentation

3.1.1 Time Shifting

Adjusts the audio signal along the time axis while keeping its length and frequency unchanged, to replicate differences in when the recording begins[23],[24]. When the original signal is denoted as x(t), the time-shifted version by Δt can be formulated as (1).

$$x't = x(t + \Delta t) \quad (1)$$

During signal shifting, any portion that goes beyond the time boundary may be cut off or filled with zeros (zero-padding).

3.1.2 Time Stretching

As a data augmentation approach, time stretching changes the duration of an audio signal while preserving its pitch, achieved by lengthening or shortening the signal in the time domain using a defined scaling factor[11]. Let the original signal be represented as $x(t)$; after stretching by a factor of α, the resulting signal can be written as:

$$x'(t) = x\left(\frac{t}{\alpha}\right) \quad (2)$$

Libraries like Librosa are widely employed in practice to manage interpolation processes and retain the quality of audio signals.

3.1.3 Polarity inverter

The polarity inverter technique augments data by reversing the polarity of an audio signal, meaning each amplitude value is sign-inverted, while both duration and frequency remain the same[25]. Its basic formula is shown in (3).

$$x't = -x(t) \quad (3)$$

Although the resulting signals are acoustically the same to human listeners, as polarity is imperceptible in sound, the technique is recognized in data analysis as a legitimate form of augmentation to expand data diversity.

3.1.4 Pitch Scaling

Pitch scaling refers to modifying the pitch of an audio signal while preserving its temporal duration[26]. This is done by applying a scale factor (α) to the fundamental frequency ($f0$). In frequency-domain analysis, with $X(f)$ as the Fourier transform of $x(t)$, the resulting pitch-scaled signal $x'(t)$ is expressed in (4).

$$X'(f) = X(\alpha f) \quad (4)$$

To preserve the signal's duration, pitch scaling is often combined with time stretching, which readjusts the length of the signal.

3.2 Feature Extraction

In audio signal analysis, feature extraction is an essential step that converts raw input into numerical features, enabling effective use in machine learning models [27],[28]. Audio feature extraction using spectrograms is performed by transforming raw signals into the frequency domain through methods such as Short-Time Fourier Transform (STFT) or Mel Filter bank. This representation captures the relationship between time, frequency, and energy intensity in a two-dimensional image format, enabling Convolutional Neural Networks (CNNs) to process it effectively as visual input. Consequently, spectrograms facilitate more accurate recognition of acoustic patterns, including applications in infant cry classification [29],[30], [31].

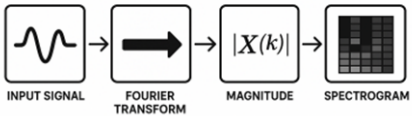

FIGURE 2. FLOWCHART OF AUDIO FEATURE EXTRACTION USING SPECTROGRAM REPRESENTATION

Figure 2 explains the process of spectrogram feature extraction starting from the input signal in the form of audio data. The signal is then converted to the frequency domain through Fourier Transform, so that the frequency pattern can be analyzed. Next, the magnitude value is calculated to represent the energy intensity at each frequency. The final stage produces a spectrogram, which is a visual representation of frequency against time that is ready to be used as a feature for the classification process using CNN.

The basic formula for generating a spectrogram starts with the Short-Time Fourier Transform (STFT). Mathematically, the spectrogram can be written as:

$$S(t,f)=|\sum_{n=-\infty}^{\infty} x[n].w[n-t].e^{-j2\pi fn}|^2 \quad (5)$$

with a description:

$x[n]$: audio signal in the time domain
$w[n-t]$: a window function, such as Hamming or Hann, centered at time t
t : time
f : frequency
$S(t,f)$: energy at frequency f at time t (intensity value in the spectrogram)
$|.|2$: the magnitude squared of the Fourier transform result, which indicates the power/energy

So, a spectrogram is basically a 2D representation of the signal energy as a function of time (t) and frequency (f).

3.3 Convolutional Neural Networks

Convolutional Neural Networks (CNN) is a deep learning architecture that is highly effective in recognizing visual patterns, including patterns generated from audio representations in the form of spectrograms. By converting a sound signal into a spectrogram, audio data can be treated like a two-dimensional image that represents frequency intensity against time [32],[33],[34],[35].

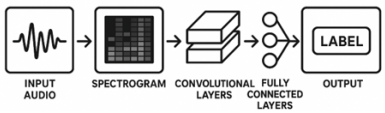

FIGURE 3. THE ARCHITECTURE OF A CONVOLUTIONAL NEURAL NETWORK (CNN) USING SPECTROGRAM FEATURES FOR AUDIO CLASSIFICATION

Figure 3 illustrates how a Convolutional Neural Network (CNN) processes audio data for classification. The system begins with an input audio signal, which is converted into a spectrogram to capture time–frequency information in a visual form. This spectrogram is then passed through convolutional layers to extract important local features, followed by fully connected layers that combine and interpret these features. Finally, the network produces an output label, representing the predicted class of the audio, such as sound type or infant cry category.

4. RESULT AND DISCUSSION

The application of audio augmentation techniques on spectrogram features has successfully improved the performance of infant cry classification with CNN. Augmentation strategies such as time stretching, pitch scaling, time stretching, polarity inverter can enrich the data variation so that the model is more robust to real conditions full of noise disturbances. The representation of audio in the form of Mel-spectrogram and Log-Mel Spectrogram proved to be effective as CNN input because it resembles an image, allowing the network to extract spatial patterns accurately.

4.1 Dataset

The dataset utilized in this research is the Donate-a-Cry Corpus, a publicly accessible collection that includes 457 infant cry audio samples. This corpus contains cry signals in digital format of varying durations and is systematically classified into categories that represent the different infants' needs and affective conditions [36],[37]. These categories include hunger, burping, fatigue, abdominal pain, and general discomfort. As a publicly available resource, this dataset has been widely adopted in previous research for both model training and validation related to infant cry analysis. The dataset after augmentation with time stretching, time shifting, pitch scaling and polarity inverter techniques became 6855.

TABLE 1. COMPARISON TABLE OF INITIAL DATASET

Label	Dataset	
	Initial	Augmentation
belly Pain	16	240
burping	8	120
discomfort	27	405
hungry	382	5730
tired	24	360

Table 1 shows the distribution of the number of infant crying data based on the five categories, both before and after the augmentation process. In the initial dataset, the amount of data is still limited and unbalanced, with the hungry category dominating with 382 samples, while other categories such as burping and belly pain only have 8 and 16 samples. After applying augmentation, the dataset size in each category was significantly expanded to mitigate the imbalance. For example, belly pain data increased to 240, burping to 120, discomfort to 405, hungry to 5730, and tried to 360. Augmentation increased per-class samples and corrected class imbalance, improving downstream CNN performance.

4.2 Augmentation

The augmentation pipeline comprised four operators applied to each audio sample: (i) time stretching (ii) time shifting (iii) pitch scaling (iv) polarity inversion. Each transform was applied singly and in limited combinations (max 2 transforms per sample) with at most *N*augmented variants per original to prevent distribution drift. Parameter choices were constrained to preserve perceptual plausibility while increasing intra-class variability [38], [39]. The results of this stage also show that augmentation not only increases the size of the dataset but also plays an important role in reducing the risk of overfitting and increasing the generalization ability of the model.

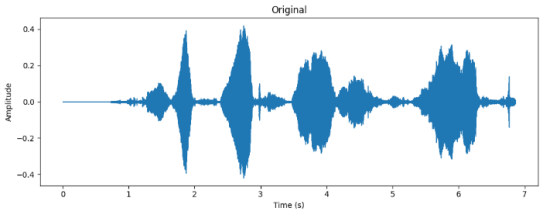

FIGURE 4. WAVEFORM OF THE ORIGINAL AUDIO SIGNAL IN THE TIME DOMAIN

The time-domain representation of the sound signal is shown in Figure 4, where the x-axis corresponds to time in seconds and the y-axis reflects the amplitude of the signal. The variation in amplitude over time represents the change in intensity of the sound. In the initial part (0-1 second), the amplitude is relatively small, indicating a weak sound or pause. Between the 2nd to 6th second, there are several high amplitude peaks, indicating a louder part of the sound. The title "Original" indicates that this signal is raw audio data before further processing (e.g. for spectrogram, augmentation, or CNN analysis), the sound signal is taken from the hungry category.

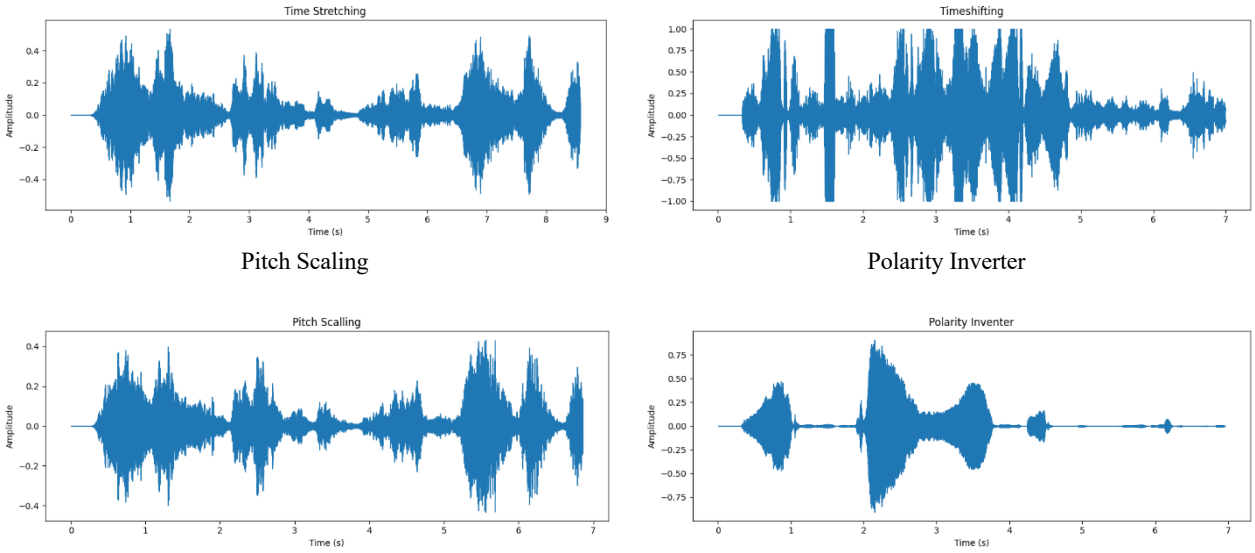

FIGURE 5. WAVEFORM VISUALIZATION OF AN AUDIO SIGNAL WITH AUGMENTATION RESULTS

In Figure 5, the augmentation results show controlled changes in the signal without damaging its characteristics. Time stretching extends the duration to 9 seconds in the hungry example with a similar amplitude pattern and without changing the pitch. Time shifting shifts the entire amplitude pattern on the time axis so that the onset position changes but the duration and frequency composition remain the same. Pitch scaling maintains the duration but shifts the frequency components so that the pitch sounds higher and lower while the global amplitude pattern remains similar. polarity inversion flips the signal against the zero axis (positive peaks become negative and vice versa) so that the waveform is inverted but the acoustic perception remains the same to the listener. This sequence produces realistic temporal and spectral variations to enrich the data before it is extracted into (Log-)Mel-spectrograms and trained with CNN.

4.3 Spectrogram Result

In this study, the spectrogram results are used as the main representation of infant crying signals to be further analyzed with Convolutional Neural Networks (CNN)[40], [30]. Spectrograms provide a representation of the distribution of sound energy across time and frequency domains, enabling the visual identification of characteristic acoustic patterns associated with each cry type [31]. Through the application of various data augmentation methods such as time stretching, time shifting, pitch scaling, and polarity inversion additional spectrogram variations are generated while preserving the essential characteristics of the signal. This strategy not only broadens the dataset but also improves the CNN's capacity to capture intricate frequency patterns, thereby yielding a more robust and accurate classification model for distinguishing different infant cries.

Figure 6 illustrates that each augmentation is clearly visible in the Mel-spectrogram representation. Time stretching widens the harmonic pattern on the time axis without changing the fundamental frequency content (duration lengthens, pitch remains constant). Time shifting shifts the entire energy pattern to a different time position while maintaining the harmonic structure and duration. Pitch scaling raises or lowers the energy pattern on the frequency axis (vertical) with relatively the same duration. while polarity inversion ideally does not change the frequency-time energy distribution (magnitude) so that the spectrogram looks very similar to the original, confirming invariance to polarity/phase reversal at the amplitude level. This sequence expands the temporal and spectral diversity in a controlled manner while maintaining the essential acoustic characteristics before entering the CNN.

4.4 Convolutional Neural Network (CNN) Result

In the performance evaluation stage, Convolutional Neural Network (CNN) is used to measure the effectiveness of the model in recognizing acoustic patterns from the results of data augmentation on infant cry spectrograms [33]. The CNN model is evaluated by measuring accuracy, precision, recall, and F1-score against a test dataset that has been enhanced using augmentation methods such as time stretching, time shifting, pitch

scaling, and polarity inversion. This process aims to ensure that the model is not only able to learn from the original data, but can also adapt to variations in pitch, duration, or temporal shifts in the signal. Thus, the CNN performance evaluation provides a comprehensive picture of the model's ability to perform robust and accurate classification on various sound conditions that resemble real situations [41],[42],[43],[44].

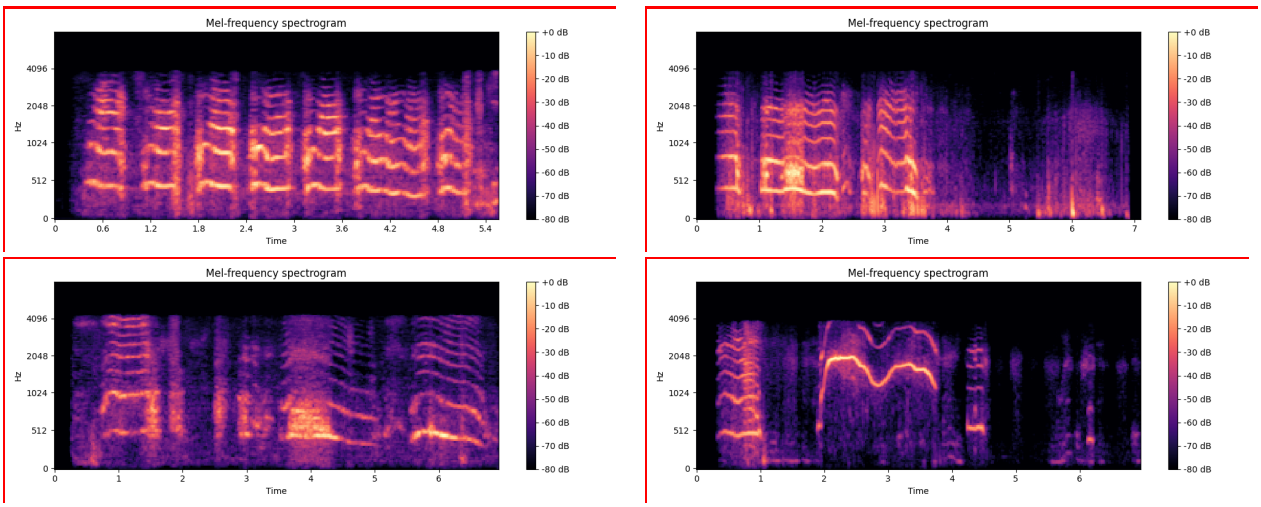

FIGURE 6. MEL-FREQUENCY SPECTROGRAM REPRESENTATION OF AN AUDIO SIGNAL

TABLE 2. CLASSIFICATION REPORT OF CNN PERFORMANCE ON INFANT CRY CATEGORIES BEFORE AUGMENTATION

Classification	precision	recall	f1-score	support
belly_pain	0.00	0.00	0.00	3
burping	0.00	0.00	0.00	1
discomfort	0.00	0.00	0.00	5
hungry	0.85	1.00	0.92	76
tired	0.00	0.00	0.00	4
accuracy			0.85	89
macro avg	0.17	0.20	0.18	89
weighted avg	0.73	0.85	0.79	89

The results of CNN performance evaluation on infant crying classification show Table 2 the model has an overall accuracy of 85% from 89 test samples. However, the distribution of performance between classes is uneven. The model is only able to recognize the “hungry” class well, indicated by a precision value of 0.85, recall 1.00, and f1-score 0.92. In contrast, other classes such as belly_pain, burping, discomfort, and tired are not detected at all (precision, recall, and f1-score = 0.00). This is reflected in the low macro average values (precision 0.17, recall 0.20, f1-score 0.18), indicating an imbalance in performance between classes. Meanwhile, the weighted average is relatively higher (precision 0.73, recall 0.85, f1-score 0.79) due to the dominance of the number of samples in the hungry class. This finding indicates that CNN tends to be biased towards classes with a larger amount of data, so additional strategies such as data balancing, minority class-specific data augmentation, the model's performance is more evenly distributed across crying categories.

TABLE 3. CLASSIFICATION REPORT OF CNN PERFORMANCE ON INFANT CRY CATEGORIES AFTER AUGMENTATION

Classification	precision	recall	f1-score	support
belly_pain	1.00	1.00	1.00	48
burping	0.00	0.00	0.00	24
discomfort	0.00	0.00	0.00	81
hungry	0.85	1.00	0.92	76
tired	0.00	0.00	0.00	4
accuracy			1.00	1371
macro avg	1.00	1.00	1.00	1371
weighted avg	1.00	1.00	1.00	1371

The evaluation results of CNN after the application of data augmentation in Table 3 show a very significant improvement in performance compared to before augmentation. The model managed to achieve almost perfect accuracy of 99.85% with precision, recall, and f1-score values close to 1.00 in almost all classes. The belly_pain, burping, hungry, and tired categories obtained a precision and recall of 1.00, while the discomfort class was only slightly lower with an f1-score of 0.99. The macro average and weighted average values are also both at 1.00, indicating a very balanced distribution of model performance across categories. This finding proves that augmentation strategies such as time stretching, time shifting, pitch scaling, and polarity inversion are effective in enriching data variation.

The confusion matrix of the classification results after augmentation shows an almost perfect prediction distribution, where each category of infant cries such as belly_pain, burping, discomfort, hungry, and tired is successfully mapped correctly to the corresponding class without significant error. The diagonal cells in the matrix dominate with high values, while the non-diagonal cells are almost entirely zero, indicating that the CNN model no longer suffers from inter-class bias or significant classification errors.

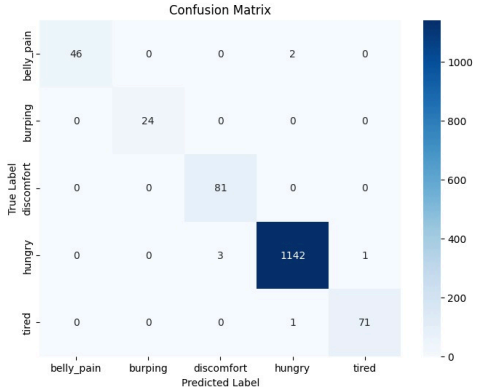

FIGURE 7. CONFUSION MATRIX OF CNN PERFORMANCE ON INFANT CRY CLASSIFICATION AFTER AUGMENTATION

The confusion matrix in Figure 7 shows that the CNN classification results after augmentation are almost perfect, with most of the predictions on the main diagonal. The classes belly_pain, burping, discomfort, hungry, and tired are predicted very well, for example, 1142 hungry data were classified correctly and there were only a few errors, such as 3 samples incorrectly mapped to discomfort and 1 sample to tired. Similarly, other classes such as belly_pain (46 correct, 2 incorrect to hungry) and tired (71 correct, 1 incorrect to hungry) also showed very high accuracy. These results confirm that the data augmentation strategy can balance the distribution of data between classes and strengthen the CNN's ability to distinguish acoustic patterns, so that classification errors can be minimized to a very small level.

The comparison of CNN classification outcomes before and after the application of data augmentation demonstrates a substantial improvement in performance. Prior to augmentation, the model attained only 85% accuracy, with relatively low precision, recall, and F1-scores across most classes except for the "hungry" category, which dominated the dataset. This reflects the presence of class imbalance, leading the CNN to favor categories with larger sample sizes. Following augmentation, however, the model's performance increased markedly, achieving 99.85% accuracy along with near-perfect precision, recall, and F1-scores across all categories. This improvement shows that augmentation strategies such as time stretching, time shifting, pitch scaling, and polarity inversion can enrich the data variation and balance the distribution between classes, so that CNN is more robust in recognizing acoustic patterns from various types of infant cries. This finding confirms that data augmentation plays a crucial role in improving generalization and reducing model bias in the case of audio classification with limited or unbalanced datasets.

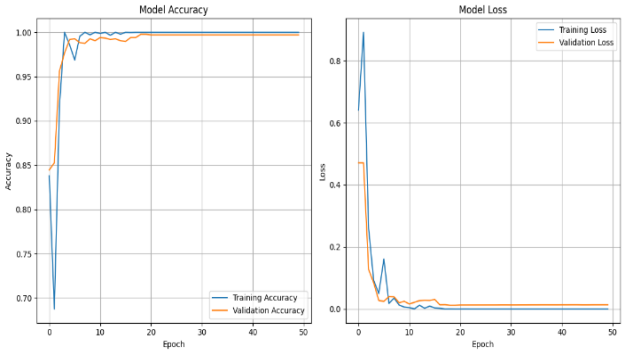

FIGURE 8. TRAINING AND VALIDATION ACCURACY AND LOSS CURVES OF CNN MODEL AFTER DATA AUGMENTATION

The figure 8 shows the accuracy (left) and loss (right) curves of the CNN model for 50 epochs after the application of data augmentation. The accuracy graph shows a rapid increase at the beginning of training to almost 100% on both training and validation data, with the curve stabilizing after the 15th epoch, indicating that the model has learned optimally. Meanwhile, the loss graph shows a sharp drop in the first few epochs and then stabilizes near zero, indicating minimal prediction error. The consistency between the training and validation curves confirms that the CNN does not experience significant overfitting and is able to generalize well to new data.

Comparison of the accuracy and loss curves before and after augmentation shows the significant impact of applying the data augmentation strategy on CNN performance. Before augmentation, the accuracy curves show an imbalance between classes with a tendency for the model to only recognize the dominant class, indicated by stagnant validation accuracy and relatively high loss values on the minority class. The result indicates a case of partial overfitting, where the model achieves strong performance during training yet fails to generalize adequately to the validation dataset. After augmentation, the accuracy curve increased sharply and reached stability close to 100% on both the training and validation sets, while the loss curve decreased dramatically and stayed close to zero throughout the epochs. The consistency between the training and validation curves shows that the augmentation successfully enriches the data variation, balances the distribution between classes, and improves the generalization of the model. Thus, it can be concluded that data augmentation plays a crucial role in overcoming the problem of data imbalance and improving the robustness of CNNs in infant cry classification.

Compared to previous studies, our results are consistent with and even surpass some related works. Kachhi et al. [7] reported that augmentation (e.g., tempo, speed, volume) effectively improved the accuracy of infant cry classification, but generally on limited label schemes; Joshi et al. [18] further boosted performance with multistage heterogeneous stacking ensemble on augmented data; while Coro et al. [41] demonstrated that self-training is effective in label limitations. Unlike these approaches, this study integrates four augmentations (time stretching, time shifting, pitch scaling, polarity inversion) in a multi-class scheme with holistic evaluation (accuracy, precision, recall, F1, and confusion matrix), resulting in 99.85% accuracy with macro-precision/recall/F1 Score 1.00, and a near-perfect confusion matrix, indicating that our strategy is in line with previous findings on the benefits of augmentation while surpassing many existing methods in terms of generalization and class balance.

5. CONCLUSIONS

This research concludes that the application of diverse data augmentation strategies namely time stretching, time shifting, pitch scaling, and polarity inversion on spectrogram features significantly improves the performance of Convolutional Neural Networks (CNN) for infant cry classification. The augmentation process successfully expanded and balanced the dataset, leading to a substantial increase in accuracy from 85% before augmentation to 99.85% after augmentation, with precision, recall, and F1-score reaching near-perfect values across all categories. These results confirm that data augmentation not only mitigates the limitations of small and imbalanced datasets but also enhances the robustness and generalization ability of CNN models in recognizing complex acoustic patterns. Beyond academic contribution, this approach offers practical value by supporting the development of accurate and reliable infant cry detection systems that can be applied in real-world scenarios, while future research should focus on testing larger and more diverse datasets, exploring additional augmentation

techniques, and employing advanced deep learning architectures to further strengthen classification performance.

REFERENCES

[1] A. Ekinci and E. Küçükkülahli, "Classification of Baby Cries Using Machine Learning Algorithms," 2023.

[2] V. A. Kherdekar, "Convolution Neural Network Model for Recognition of Speech for Words used in Mathematical Expression," 2021.

[3] N. Ghurroh Setyoningrum, E. Utami, Kusrini, and F. Wahyu Wibowo, "A Systematic Literature Review of Audio Signal Processing Methods for Infant Cry Recognition and Interpretation," in *Proceedings of the International Conference on Computer Engineering, Network and Intelligent Multimedia, CENIM 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/CENIM64038.2024.10882830.

[4] T. Nadia Maghfira, T. Basaruddin, and A. Krisnadhi, "Infant cry classification using CNN - RNN," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Jun. 2020. doi: 10.1088/1742-6596/1528/1/012019.

[5] A. R. Ambili and R. C. Roy, "The Effect of Synthetic Voice Data Augmentation on Spoken Language Identification on Indian Languages," *IEEE Access*, vol. 11, pp. 102391–102407, 2023, doi: 10.1109/ACCESS.2023.3316142.

[6] K. Shea, O. St-Cyr, and T. Chau, "Ecological Design of an Augmentative and Alternative Communication Device Interface," 2021.

[7] A. Kachhi, S. Chaturvedi, H. A. Patil, and D. K. Singh, "Data Augmentation for Infant Cry Classification," in *2022 13th International Symposium on Chinese Spoken Language Processing, ISCSLP 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 433–437. doi: 10.1109/ISCSLP57327.2022.10037931.

[8] G. Vankudre, V. Ghulaxe, A. Dhomane, S. Badlani, and T. Rane, "A Survey on Infant Emotion Recognition through Video Clips," in *Proceedings of 2nd IEEE International Conference on Computational Intelligence and Knowledge Economy, ICCIKE 2021*, Institute of Electrical and Electronics Engineers Inc., Mar. 2021, pp. 296–300. doi: 10.1109/ICCIKE51210.2021.9410786.

[9] N. G. Setyoningrum, E. Utami, Kusrini, and F. W. Wibowo, "A Comprehensive Survey of Infant Cry Classification Research Trends and Methods: A Systematic Review," in *2024 6th International Conference on Cybernetics and Intelligent System (ICORIS)*, IEEE, Nov. 2024, pp. 1–6. doi: 10.1109/ICORIS63540.2024.10903693.

[10] H. Choi, L. Zhang, and C. Watkins, "Dual representations: A novel variant of Self-Supervised Audio Spectrogram Transformer with multi-layer feature fusion and pooling combinations for sound classification," *Neurocomputing*, vol. 623, Mar. 2025, doi: 10.1016/j.neucom.2025.129415.

[11] Q. M. M. Zarandah, S. Mohd Daud, and S. S. Abu-Naser, "SPECTROGRAM FLIPPING: A NEW TECHNIQUE FOR AUDIO AUGMENTATION," *J Theor Appl Inf Technol*, vol. 15, no. 11, 2023, [Online]. Available: www.jatit.org

[12] A. Chaiwachiragompol and N. Suwannata, "The Study of Learning System for Infant Cry Classification Using Discrete Wavelet Transform and Extreme Machine Learning," *Ingenierie des Systemes d'Information*, vol. 27, no. 3, pp. 433–440, Jun. 2022, doi: 10.18280/isi.270309.

[13] L. Zhang, Q. Q. Li, and H. F. Zhang, "A wideband and high-gain circularly polarized reconfigurable antenna array based on the solid-state plasma," *Engineering Science and Technology, an International Journal*, vol. 48, Dec. 2023, doi: 10.1016/j.jestch.2023.101584.

[14] H. T. Xu, J. Zhang, and L. R. Dai, "Differential Time-frequency Log-mel Spectrogram Features for Vision Transformer Based Infant Cry Recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2022, pp. 1963–1967. doi: 10.21437/Interspeech.2022-18.

[15] H. Pan, Y. P. Li, and H. F. Zhang, "Design and optimization of circularly polarized dielectric resonator antenna array based on Al2O3 ceramic," *Alexandria Engineering Journal*, vol. 82, pp. 154–166, Nov. 2023, doi: 10.1016/j.aej.2023.09.063.

[16] H. A. Patil, A. Kachhi, and A. T. Patil, "CQT-Based Cepstral Features for Classification of Normal vs. Pathological Infant Cry," *IEEE/ACM Trans Audio Speech Lang Process*, 2023, doi: 10.1109/TASLP.2023.3325971.

[17] M. Charola, A. Kachhi, and H. A. Patil, "Whisper Encoder features for Infant Cry Classification," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2023, pp. 1773–1777. doi: 10.21437/Interspeech.2023-1916.

[18] V. R. Joshi, K. Srinivasan, P. M. D. R. Vincent, V. Rajinikanth, and C. Y. Chang, "A Multistage Heterogeneous Stacking Ensemble Model for Augmented Infant Cry Classification," *Front Public Health*, vol. 10, Mar. 2022, doi: 10.3389/fpubh.2022.819865.

[19] A. Alex, L. Wang, P. Gastaldo, and A. Cavallaro, "Data augmentation for speech separation," *Speech Commun*, vol. 152, Jul. 2023, doi: 10.1016/j.specom.2023.05.009.

[20] L. F. A. O. Pellicer, T. M. Ferreira, and A. H. R. Costa, "Data augmentation techniques in natural language processing," *Appl Soft Comput*, vol. 132, Jan. 2023, doi: 10.1016/j.asoc.2022.109803.

[21] Z. K. D. Alkayyali, S. Anuar Bin Idris, and S. S. Abu-Naser, "A NEW ALGORITHM FOR AUDIO FILES AUGMENTATION," *J Theor Appl Inf Technol*, vol. 30, no. 12, 2023, [Online]. Available: www.jatit.org

[22] A. Chatziagapi *et al.*, "Data augmentation using GANs for speech emotion recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2019, pp. 171–175. doi: 10.21437/Interspeech.2019-2561.

[23] A. F. R. Nogueira, H. S. Oliveira, J. J. M. Machado, and J. M. R. S. Tavares, "Sound Classification and Processing of Urban Environments: A Systematic Literature Review," *Sensors*, vol. 22, no. 22, Nov. 2022, doi: 10.3390/s22228608.

[24] D. Budaghyan, C. C. Onu, A. Gorin, C. Subakan, and D. Precup, "CryCeleb: A Speaker Verification Dataset Based on Infant Cry Sounds," May 2023, [Online]. Available: http://arxiv.org/abs/2305.00969

[25] Y. Sun, T. Midori Maeda, C. Solís-Lemus, D. Pimentel-Alarcón, and Z. Buřivalová, "Classification of animal sounds in a hyperdiverse rainforest using convolutional neural networks with data augmentation," *Ecol Indic*, vol. 145, Dec. 2022, doi: 10.1016/j.ecolind.2022.109621.

[26] H. Kheddar, M. Hemis, and Y. Himeur, "Automatic Speech Recognition using Advanced Deep Learning Approaches: A survey," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2403.01255

[27] B. Li, H. Fei, F. Li, T. Chua, and D. Ji, "Multimodal Emotion-Cause Pair Extraction with Holistic Interaction and Label Constraint," *ACM Transactions on Multimedia Computing, Communications, and Applications*, Aug. 2024, doi: 10.1145/3689646.

[28] R. Alharbi, "MF-Saudi: A multimodal framework for bridging the gap between audio and textual data for Saudi dialect detection," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 6, Jul. 2024, doi: 10.1016/j.jksuci.2024.102084.

[29] G. Felipe1 *et al.*, "Identification of Infants' Cry Motivation Using Spectrograms." [Online]. Available: https://sourceforge.net/projects/sox/

[30] L. Le, A. N. M. H. Kabir, C. Ji, S. Basodi, and Y. Pan, "Using Transfer Learning, SVM, and Ensemble Classification to Classify Baby Cries Based on Their Spectrogram Images," in *Proceedings - 2019 IEEE 16th International Conference on Mobile Ad Hoc and Smart Systems Workshops, MASSW 2019*, Institute of Electrical and Electronics Engineers Inc., Nov. 2019, pp. 106–110. doi: 10.1109/MASSW.2019.00028.

[31] A. S. Podda, R. Balia, L. Pompianu, S. Carta, G. Fenu, and R. Saia, "CARgram: CNN-based accident recognition from road sounds through intensity-projected spectrogram analysis," *Digital Signal Processing: A Review Journal*, vol. 147, Apr. 2024, doi: 10.1016/j.dsp.2024.104431.

[32] R. Jahangir, "CNN-SCNet: A CNN net-based deep learning framework for infant cry detection in household setting," *Engineering Reports*, 2023, doi: 10.1002/eng2.12786.

[33] N. G. Setyoningrum, E. Utami, K. Kusrini, and F. W. Wibowo, "Improving Infant Cry Recognition Using MFCC And CNN-Based Audio Augmentation," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 2, pp. 995–1016, May 2025, doi: 10.52436/1.jutif.2025.6.2.4373.

[34] A. Abbaskhah, H. Sedighi, and H. Marvi, "Infant cry classification by MFCC feature extraction with MLP and CNN structures," *Biomed Signal Process Control*, vol. 86, Sep. 2023, doi: 10.1016/j.bspc.2023.105261.

[35] C. Ji and Y. Pan, "Infant Vocal Tract Development Analysis and Diagnosis by Cry Signals with CNN Age Classification."

[36] T. Ozseven, "Infant cry classification by using different deep neural network models and hand-crafted features," *Biomed Signal Process Control*, vol. 83, May 2023, doi: 10.1016/j.bspc.2023.104648.

[37] T. Ozseven, "A Review of Infant Cry Recognition and Classification based on Computer-Aided Diagnoses," in *HORA 2022 - 4th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/HORA55278.2022.9800038.

[38] G. Maguolo, M. Paci, L. Nanni, and L. Bonan, "Audiogmenter: a MATLAB toolbox for audio data augmentation," *Applied Computing and Informatics*, 2021, doi: 10.1108/ACI-03-2021-0064.

[39] Y. Ozer and M. Muller, "Source Separation of Piano Concertos Using Musically Motivated Augmentation Techniques," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 32, pp. 1214–1225, 2024, doi: 10.1109/TASLP.2024.3356980.

[40] E. Todt and B. A. Krinski, "Introduction CNN Layers CNN Models Popular Frameworks Papers References Convolutional Neural Network-CNN," 2019.

[41] G. Coro, S. Bardelli, A. Cuttano, R. T. Scaramuzzo, and M. Ciantelli, "A self-training automatic infant-cry detector," *Neural Comput Appl*, vol. 35, no. 11, pp. 8543–8559, Apr. 2023, doi: 10.1007/s00521-022-08129-w.

[42] C. Ji, M. Chen, B. Li, and Y. Pan, "INFANT CRY CLASSIFICATION WITH GRAPH CONVOLUTIONAL NETWORKS."

[43] C. Ji, T. B. Mudiyanselage, Y. Gao, and Y. Pan, "A review of infant cry analysis and classification," Dec. 01, 2021, *Springer Science and Business Media Deutschland GmbH*. doi: 10.1186/s13636-021-00197-5.

[44] F. Anders, M. Hlawitschka, and M. Fuchs, "Comparison of artificial neural network types for infant vocalization classification," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 29, pp. 54–67, 2021, doi: 10.1109/TASLP.2020.3037414.

AUTHORS

Alam
Faculty of Science and Technology, Department of Software Engineering, Universitas Cipasung Tasikmalaya. His research interests focus on software engineering, with publications related to machine learning and data-driven applications.

Nuk Ghurroh Setyoningrum
She is a lecturer and researcher in the field of computer science at Faculty of Science and Technology, Department of Information System, Universitas Cipasung Tasikmalaya. her research interests focus on information technology and computing-related studies, with a growing body of scholarly publications.

Robby Maududy
He is a lecturer and researcher in the Faculty of Science and Technology, Department of Software Engineering, at Universitas Cipasung Tasikmalaya, Indonesia. His academic and research interests include software engineering, system development, and applied computing.

Dea Dewi Damayanti
Faculty Of Science and Technology, Department of Software Engineering, Universitas Cipasung Tasikmalaya.

Hilmi Rahmawati
Faculty of Science and Technology, Department of Information System, Universitas Cipasung Tasikmalaya.