Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 2, April 2025, Page. 995-1016
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4373

Improving Infant Cry Recognition Using MFCC And CNN-Based Audio Augmentation

Nuk Ghurroh Setyoningrum[*1], Ema Utami[2], Kusrini[3], Ferry Wahyu Wibowo[4]

[1,2,3,4]Department Of Informatics, Universitas Amikom Yogyakarta, Indonesia

Email: [1]nuke@students.amikom.ac.id

Received : Jan 27, 2025; Revised : Apr 28, 2025; Accepted : May 7, 2025; Published : May 17, 2025

Abstract

Recognizing infant cries is essential for understanding a baby's needs; however, previous research has struggled with imbalanced datasets and limited feature extraction techniques. Conventional methods utilizing CNN without data augmentation often failed to accurately classify minority classes such as belly pain, burping, and discomfort, resulting in biased models that predominantly recognized majority classes. This study proposes an MFCC-based data augmentation pipeline, incorporating time stretching, pitch scaling, noise addition, polarity inversion, and random gain adjustments to increase dataset diversity and enhance model generalization. By applying this approach, the dataset size was expanded from 457 to 8,683 samples, and a CNN model with three convolutional layers, ReLU activation, and max pooling was trained for cry pattern classification. The results indicate a substantial accuracy improvement from 78% to 98%, with F1-scores for minority classes rising from 0.00 to above 0.90, confirming that augmentation effectively addresses dataset imbalance. This research advances computer science and artificial intelligence, particularly in audio signal processing and deep learning for healthcare applications, by demonstrating the role of data augmentation in improving cry classification performance. Future directions include integrating multimodal data (visual and physiological signals), exploring advanced deep learning architectures, and developing real-time applications for smart baby monitoring systems to further enhance infant cry recognition technology.

Keywords : *Audio Data Augmentation, Cry Pattern Classification, Infant Cry Recognition, MFCC Feature Extraction, Speech Signal Processing.*

1. INTRODUCTION

Infant cry is one of the main forms of communication in infants to convey needs or discomfort, such as hunger, pain, or other discomfort. Research on infant cry pattern recognition is becoming increasingly important in a global context, especially to support early detection of health problems or urgent needs that may be difficult for parents or caregivers to recognize[1], [2], [3], [4]. In recent trends, technology-based approaches using audio signal analysis have grown rapidly, especially by utilizing feature extraction methods such as Mel-Frequency Cepstral Coefficients (MFCC). MFCC is a technique used to convert sound signals into numerical representations that can be processed by pattern recognition algorithms[5], [6], [7]. By utilizing this technology, the development of automated systems to recognize infant crying patterns becomes more reliable, especially when combined with techniques for augmenting audio data to enhance both the quantity and variability of training samples [8], [9], [10], [11], [12]. This field has the potential to make significant contributions in pediatric health, particularly in providing practical and effective solutions for modern parenting.

Previous research on infant cry pattern recognition has shown significant progress, especially in the use of feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCC) to analyze audio signals[13]. Earlier research has effectively categorized infant cries into groups like hunger, pain, and general discomfort by employing machine learning approaches such as Support Vector Machines

(SVM) and deep learning-based Convolutional Neural Networks (CNN)[14], [15], [16]. However, many of these studies face limitations in terms of data availability and diversity, as infant cry data is often limited and susceptible to background noise[17], [18], [19]. Furthermore, audio data augmentation approaches as a solution to increase data diversity have not been widely applied in this domain, despite being proven effective in other areas of speech recognition[20], [21]. Utilizing audio data augmentation techniques, such as frequency or time manipulation, is expected To address these challenges and enhance the accuracy of classification models[22], [23]. Therefore, the gap that exists is the lack of comprehensive exploration into the combination of audio augmentation with MFCC for infant crying pattern recognition, which is the main focus in this study.

Hua-Nong Ting et al. evaluated the performance of Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN) by utilizing individual features like MFCC and hybrid speech-based features. With the Infant Chillanto Database which includes normal and asphyxia infant cries, the findings show CNN is superior with single MFCC features, while DNN with hybrid features achieves up to 99.96% accuracy for the classification of normal and asphyxia cries. This research highlights the effectiveness of hybrid speech features in improving the recognition accuracy of asphyxia cries, offering potential as a clinical tool to monitor the risk of hypoxia in infants[24].

Turgut conducted a review of A total of 112 studies on infant cry recognition and classification (ICRC) employing computer-aided diagnosis, focusing on datasets, features, classification methods, and performance outcomes. Mel-Frequency Cepstral Coefficients (MFCC) were highlighted as the most frequently utilized features, while classifiers based on neural networks and deep learning approaches are increasingly popular. The ICRC workflow involves steps including data collection, preprocessing, feature extraction and selection, as well as classification. Commonly utilized datasets include the Infant Chillanto and Donate-a-Cry Corpus, with preprocessing playing a crucial role in enhancing signal clarity. Approaches such as SVM, GMM, HMM, and neural networks have been utilized, with Extreme Learning Machine (ELM) and Artificial Neural Network (ANN) showing promising performance[25].

Recent studies have demonstrated that CNN-based classifiers can effectively capture cry pattern features, yet their performance is often constrained by the scarcity of labeled infant cry datasets. For instance, prior research using the Donate-a-Cry Corpus has reported significant classification bias, where the model performs well in detecting frequent cry types (e.g., hunger) but struggles with less frequent categories, such as cries related to abdominal pain or discomfort. In contrast, augmentation techniques have been widely adopted in speech recognition and natural language processing to enhance model robustness, yet their potential in the domain of infant cry recognition remains underexplored.

In contrast to previous studies that mostly focus on conventional feature extraction techniques without applying augmentation, this study presents a comprehensive MFCC-based augmentation approach to improve model performance on unbalanced infant crying datasets. The applied augmentation strategies include time stretching, pitch scaling, noise addition, polarity inversion, and random gain adjustments, which aim to increase data diversity and strengthen the model's robustness to variations in crying patterns. By applying these innovative augmentation techniques, the number of samples in the dataset increased from 457 to 8,683, ensuring a more balanced distribution of crying categories.

The primary contributions of this study are as follows:

- Introduction of a novel MFCC-based augmentation pipeline that systematically increases the diversity of the baby crying dataset, overcoming the class imbalance problem in previous studies.
- Empirical validation of the effectiveness of data augmentation, showing a significant increase in model accuracy from 78% (without augmentation) to 98% (with augmentation).
- Improved classification of minority vocal categories, particularly in the detection of abdominal pain and discomfort vocalizations that were previously difficult to detect.

Vol. 6, No. 2, April 2025, Page. 995-1016
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4373

- Comparison with the conventional CNN model showed that augmentation significantly improved classification performance for all types of cry.

This article is structured as follows: Chapter I Introduction explains the background, the importance of the research, the gaps in previous studies, and the main objectives and contributions of this research. Chapter II on The Research Methods section describes the research approach, including the feature extraction process using Mel-Frequency Cepstral Coefficients (MFCC), the audio data augmentation technique, and the methodology for assessing model performance. Chapter III on Results and Discussion presents the key findings, including the enhancement in classification accuracy achieved through data augmentation, and examines these results in relation to prior studies. Chapter IV Discussion elaborates on the implications of the research findings, discussing the strengths and limitations, as well as the potential for further development. Chapter V The Conclusions section highlights the key contributions of the research, its practical and theoretical implications, and suggestions for future studies. This structure is designed to provide a systematic and comprehensive flow of discussion.

2. METHOD

This research adopts an approach based on audio augmentation and signal processing to improve infant cry pattern recognition. The applied augmentation techniques consist of time shifting, time stretching, pitch scaling, noise addition, polarity inversion, and random gain, aiming to increase the diversity of the audio dataset and enhance the model's resilience to variations in data. Feature extraction employs Mel-Frequency Cepstral Coefficients (MFCC), which efficiently capture the acoustic characteristics of infant cries. For classification, a Convolutional Neural Network (CNN) was employed, designed to learn feature patterns from the augmented data in depth and deliver a reliable classification system. This method seeks to improve the accuracy and effectiveness of the infant cry pattern recognition system under diverse environmental conditions, as illustrated in Figure 1.

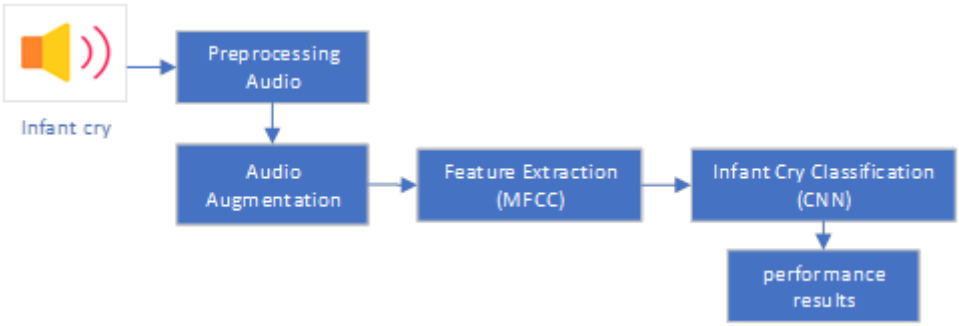

Figure 1. Infant Cry Classification Architecture Using Audio Data Augmentation And Mel Frequency Cepstral Coefficient (MFCC) Feature Extraction

2.1. Dataset

The dataset utilized in this study is the Donate-a-Cry Corpus, a publicly available collection of audio recordings of infant cries, encompassing 457 audio signals[25], [26]. This dataset encompasses a range of data types, including audio recordings of infant cries in digital format, with varying recording durations.The dataset is meticulously organized into several categories of infant cries, reflecting the diverse needs and emotional states of infants. These categories include hunger, burping, tired, belly pain, and discomfort. This dataset encompasses audio recordings in digital format, which are publicly accessible, and has been extensively utilized for model training and validation in numerous studies pertaining to infant crying.

2.2. Preprocessing

In the data preprocessing of this study, we perform a process referred to as "data cleaning," which involves the removal of background noise. This is achieved by applying a band-pass filter within the frequency range of 300-3000 Hz, with the aim of preserving the main characteristics of the infant crying signal[27], [28], [29], [30], [31], [32], [33]. The normalization process is used to equalize the amplitude of the signal, ensuring consistency in the data volume. Segmentation is performed by cutting the audio signal into specific durations using the sliding window method to keep the focus on the specific pattern of the infant's cry[34], [35], [36], [37].The sampling rate used is 16 kHz, which is standard for human voice signal analysis. Tools and libraries such as Librosa and PyTorch Audio support efficient cleaning, normalization, and segmentation of the audio signal. This technique is designed to ensure optimal data quality before further analysis[38], [39], [40], [41], [42].

2.3. Audio Data Augmentation

This study employs various audio augmentation methods to increase the diversity of infant cry data, there by enhancing the model's adaptability to signal variations[43], [44], [45]. The augmentation techniques employed include time shifting, which involves manipulating the audio signal in the time domain; time stretching, which alters the speed without affecting the pitch; and pitch scaling, which adjusts the pitch. Additionally, the incorporation of noise addition, polarity inversion, and random gain adjustment techniques, such as the use of a polarity inverter to reverse the signal's polarity, and random gain adjustment to introduce variations in the intensity of the sound, is Essential for improving the model's ability to adapt to variations in signals. These techniques are implemented using libraries such as Librosa or PyTorch Audio, which are intended to increase the volume and variety of training data, while reducing the risk of overfitting and improving the model's ability to generalize across diverse infant cry patterns[46], [47], [48].

2.3.1. Time Shifting

Moves the audio signal forward or backward in the time domain while preserving its duration and frequency, to simulate the difference in recording start time[2], [49].

If the original signal is expressed as x(t), then the signal that has been time-shifted Δt can be represented as (1).

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

When shifting the signal, the part that exceeds the time limit can be truncated or padded with zero values (zero-padding).

2.3.2. Time Stretching

Time stretching is a data augmentation method that alters the duration of an audio signal while maintaining its original pitch. This technique stretches or shortens the signal in the time domain based on a certain scale factor[15].

If the original signal is expressed as $x(t)$, then the signal that has been stretched with a scale factor of α can be represented as (2).

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

In practical implementations, libraries such as Librosa are frequently utilized to manage interpolation and preserve audio quality.

2.3.3. Pitch Scaling

Pitch scaling is a data augmentation method that modifies the pitch of an audio signal while preserving its duration[22]. Pitch scaling is achieved by modifying the signal's fundamental frequency ($f0$) using a scale factor (α). It can also be analyzed in the frequency domain, where the Fourier transform of $x(t)$ is expressed as $X(f)$, then the pitch-scaled transformation of $x'(t)$ is shown in (3).

$$X'(f) = X(\alpha f) \quad (3)$$

To keep the signal duration the same, time stretching is often applied simultaneously to adjust the signal length after pitch scaling.

2.3.4. Noise Addition

Noise addition is an augmentation technique that introduces random noise into the original signal, mimicking various environmental conditions and increasing dataset variability [8].

If the original signal is expressed as $x(t)$, then the resultant signal after the addition of noise $x'(t)$ can be formulated as (4).

$$x'(t) = x(t) + \alpha.n(t) \quad (4)$$

Value α should be adjusted so that the added noise does not distort the main information in the original signal.

2.3.5. Polarity Inverter

Polarity inverter is a data augmentation technique that reverses the polarity of an audio signal, i.e., it changes the sign of each amplitude value in the signal without changing the duration or frequency [2]. The basic formula for this technique is shown in (5).

$$x't = -x(t) \quad (5)$$

This technique creates acoustically identical signals, as humans cannot distinguish the polarity of signals in sound. However, in the data analysis domain, it is considered a valid augmentation to increase data diversity.

2.3.6. Random Gain

Random gain is a data augmentation technique that randomly changes the amplitude of an audio signal by multiplying the signal by a randomly chosen scale factor (gain) from a given range[21]. The basic formula is shown in (6).

$$x'(t) = g.x(t) \quad (6)$$

Gain factor (g) is usually taken from a uniform or Gaussian distribution to produce random amplitude variations.

2.4. Feature Extraction

Feature extraction is a vital step in audio signal analysis, aiming to convert raw data into numerical representations that can be effectively processed by machine learning models [50], [51], [52].

Mel-Frequency Cepstral Coefficients (MFCC) were selected as the primary feature due to their effectiveness in capturing the acoustic characteristics of audio signals, which aligns with the human auditory system's heightened sensitivity to low frequencies [53], [54], can be seen from figure 2 regarding the flowchart of MFCC.

Vol. 6, No. 2, April 2025, Page. 995-1016
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4373

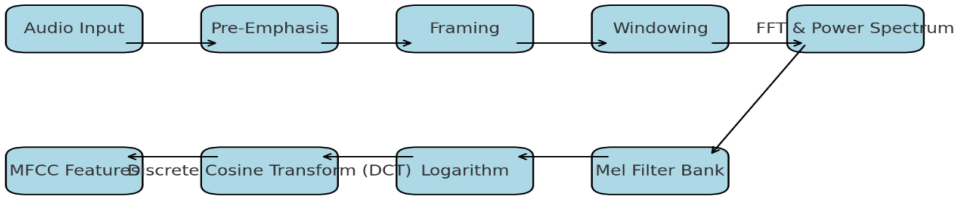

Figure 2. Mel Frequency Ceptral Coefficient Feature Extraction Flowchart

The MFCC parameters employed in this study encompass the number of coefficients ranging from 13 to 20, the implementation of Hamming-type windowing, a frame size of approximately 20 to 40 milliseconds, and 50% overlapping between frames to optimally capture temporal information.The transformation process commences with the preprocessing of the audio signal through pre-emphasis, followed by framing and windowing to divide the signal into smaller segments. Subsequent to this, a Fourier transform is performed to generate a frequency spectrum, which is then applied to a triangular spaced filter to emphasize relevant frequencies.The result is transformed using logarithmic scaling and discrete cosine transform (DCT) to generate cepstral coefficients, which are then used as a numerical representation of the acoustic characteristics of the signal.This approach allows the model to recognize infant crying patterns with greater accuracy and robustness. The formula for generating Mel-Frequency Cepstral Coefficients (MFCC) is as follows (7).

$$C_n = \sum_{m=1}^{M} log\left(\sum_{f=fmin}^{fmax} |\text{FFT}\{x(t).w(t)\}|^2 . H_m(f)\right). cos\left(\frac{\pi n(2m-1)}{2M}\right) \quad (7)$$

The MFCC formula summarizes the audio feature extraction process by converting the signal into an informative numerical representation. Starting with framing and windowing, the signal is converted to the frequency domain using FFT, and then mapped to the Mel scale through a filter bank to reflect the sensitivity of human hearing. The Mel energy is logarithmically compressed and processed with Discrete Cosine Transform (DCT) to generate MFCC coefficients, which represent acoustic patterns in a concise and efficient manner.

2.5. Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) serve as the main algorithm used for classifying infant cry patterns, exhibiting a robust capacity to extract spatial and temporal features from audio data represented as spectrograms [55], [56]. CNNs comprise convolution layers that implement filters to identify significant patterns in the data, followed by a pooling layer that reduces the dimensionality without compromising crucial information[29], [57], [58]. These layers are interconnected by a fully connected layer, which combines the information to produce the final prediction.In this study, a CNN is designed to recognize unique patterns in infant crying signals by using numerical representations of audio features, such as Mel-Frequency Cepstral Coefficients (MFCC).The CNN model is trained using a pre-segmented infant crying dataset to ensure its robustness against data variations. The selection of this algorithm was predicated on its demonstrated proficiency in the classification of intricate data with a commendable degree of precision, thereby establishing it as a optimal selection for the comprehension of infantile vocalizations under diverse circumstances.

Convolutional neural networks (CNNs) were selected in preference to conventional classifiers, such as support vector machines (SVMs) and hidden Markov models (HMMs), on account of their demonstrated superiority in the extraction of features from time-series audio signals. In contradistinction to SVMs and HMMs, which are dependent on manually generated features, CNNs possess the ability to automatically extract and learn hierarchical feature representations, a capability that renders them more

robust in the identification of complex crying patterns under a range of conditions. Furthermore, CNNs have been shown to process spatial and temporal dependencies in audio spectrograms, thereby enhancing their effectiveness in recognising subtle variations in infant cries.

The model training and validation procedure commences with the partitioning of the dataset employing the train-test split method, wherein a general proportion of 80% is allocated for training and 20% for testing. Alternatively, a k-fold cross-validation approach is utilized to ensure the attainment of more reliable results.The loss function employed is categorical cross-entropy, a suitable option for multi-class classification tasks, such as infant cry pattern recognition. During training, the model's effectiveness is assessed using metrics such as accuracy, precision, recall, and F1-score. This structured approach ensures that the model is trained to an optimal level and demonstrates strong generalization on unseen data. Convolutional Neural Networks (CNNs) operate through a sequence of key processes, including convolution, activation, pooling, and propagation to the next layer. The following is a mathematical exposition of these operations:

$$o_k = f\left(\sum_{i=1}^{I} w_{ik} \cdot \underset{(m,n)\in R_{max}}{max}[max(0, \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} x(i+m, j+n).w_{mnk} + b_k)]\right) (8)$$

To classify infant crying patterns, This research utilizes the Convolutional Neural Network (CNN) model, which is very effective in processing time-series audio signals represented in the form of spectrograms. The CNN architecture used in this research consists of:

- The model comprises three convolution layers, each of which is it is then followed by Rectified Linear Unit (ReLU) activation to introduce non-linearity, and max pooling is applied to reduce dimensionality while preserving essential features.
- A smoothing layer is employed to transform the extracted features into a one-dimensional vector.
- A fully connected (dense) layer analyzes the high-level feature representations and executes the classification process.
- Finally, a softmax activation function is applied in the output layer, thereby generating a probability distribution for each distinct category of crying.

The result of this convolution is subsequently applied to the ReLU activation function, which retains only positive values to introduce non-linearity.Subsequently, pooling—such as max pooling—is performed to decrease data dimensionality by identifying the maximum value within a defined region, thereby retaining key information while reducing complexity. After passing through multiple convolution and pooling layers, the extracted features are fed into the fully connected layer, where they are combined using weights and biases to generate an output score. A final activation function, such as softmax, is then applied to this score to convert it into a classification probability.The formula delineates the entirety of the CNN process, integrating mathematical operations at each step to yield the optimal final prediction in the classification task. The CNN model is illustrated in figure 3 below:

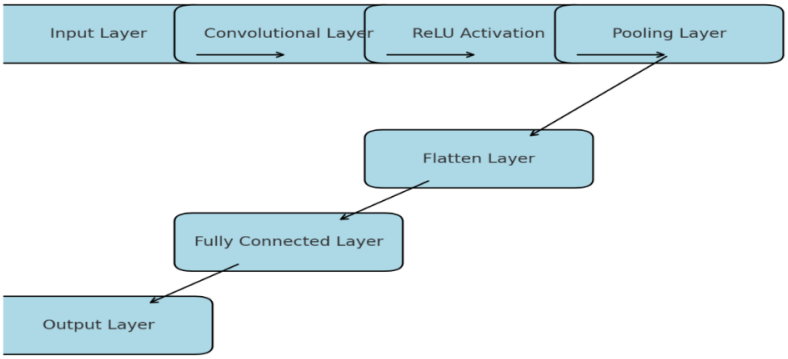

Figure 3. Convolutional Neural Network (CNN) Architecture

The architecture of a Convolutional Neural Network (CNN) comprises multiple primary layers designed to extract features and perform classification. The process initiates with the input layer, which receives data such as images or audio signals. Subsequently, the convolution layer implements filters to detect local features, including patterns or edges. These features are then processed through the ReLU activation function, which introduces non-linearity by retaining only positive values. Subsequently, a pooling layer is employed To decrease data dimensionality, retaining essential information while simultaneously lowering computational complexity. The processed data is then flattened through the flatten layer, which transforms it into a one-dimensional vector for input into the fully connected layer, where all neurons are interconnected to combine information and produce predictions. Finally, the output layer provides classification results based on predefined categories. This architecture is designed to capture patterns in data with high efficiency and is suitable for pattern recognition and classification tasks.

2.6. Evaluation

The model's performance is assessed using key evaluation metrics, including precision, recall, F1-score, and a confusion matrix, ensuring a thorough performance analysis [59], [60], [61], [62].

2.6.1. Precision

Precision is used as a metric to evaluate the effectiveness of the model in identifying infant cry patterns. It is defined as the ratio of correctly predicted positive cases (true positives) to the total predicted positives, which encompasses both true and false positives. Mathematically, precision is expressed as:

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

In the context of this study, the metric under scrutiny assumes particular significance, as it ensures that the model can accurately classify the type of baby cry, thereby avoiding an excessive number of false predictions.To illustrate this point, consider a scenario in which the model predicts a baby's cry as a sign of hunger. In such a case, precision becomes instrumental in evaluating the frequency with which this prediction is accurate in comparison to the total number of hunger predictions. MFCC-based data augmentation enhances the model's precision by leveraging more diverse training data, thereby mitigating classification errors attributable to environmental variations or noise in the original data. This ensures that the baby crying recognition system is not only sensitive but also specific to the pattern to be recognized.

2.6.2. Recall

Recall is a metric used for evaluating the model's sensitivity in recognizing baby crying patterns. Recall is defined as the proportion of correct predictions for a specific class (true positives) to the total amount of data in that class, including missed predictions (false negatives). Mathematically, recall is formulated as:

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

This metric is of particular significance in the present study, as it ensures that the model is capable of detecting all instances of infant cries from a specific class, such as cries indicative of hunger or illness, without disregarding pertinent patterns. Within MFCC-based data augmentation, recall can be improved by expanding dataset diversity, allowing the model to become more resilient to variations in crying patterns, background noise, and different environmental conditions.

2.6.3. F1–Score

The F1-score is a key metric for assessing machine learning models, emphasizing the balance between precision and recall, both of which are crucial in evaluation. It is the harmonic mean of precision, which quantifies the accuracy of positive predictions, and recall, which measures the model's ability to detect all positive instances. The F1-score is determined using the following formula:

$$F1 - Score = 2\frac{Precision.Recall}{Precision+Recall} \quad (11)$$

In the context of this study, the F1-score holds particular pertinence, particularly when the dataset exhibits imbalance in the amount of data between classes, such as different patterns of infant cries (e.g., hungry cries are more frequent than sick cries). MFCC-based data augmentation leads to the expansion of the dataset's variety, thereby facilitating simultaneous improvements in precision and recall. A high F1-score enhances the reliability of the baby cry recognition system for real-world applications.

2.6.4. Confusion Matrix

The confusion matrix is a visual representation used to assess the performance of classification models in recognizing different infant cry patterns. It displays model predictions in a tabular format, showing the correlation between correct and incorrect classifications for each category.

In MFCC-based data augmentation, the confusion matrix is utilized to assess the model's performance before and after augmentation, with the goal of reducing prediction errors, including false positives (FP) and false negatives (FN). This analysis is instrumental in identifying the model's deficiencies in specific classes, thus enabling further improvements to increase the system's accuracy and sensitivity.

2.7. Statistical Analysis

To validate whether the improvement in model performance after data augmentation was statistically significant, a paired t-test was conducted. The paired t-test compared the model's accuracy before and after applying the augmentation techniques. This statistical test was chosen as it assesses whether the mean differences between paired observations (before vs. after augmentation) are statistically significant. A significance threshold of $p < 0.05$ was used to confirm statistical improvement.

3. RESULT

This chapter outlines the research findings aimed at improving infant cry pattern recognition through data augmentation techniques based on Mel-Frequency Cepstral Coefficients (MFCC). It assesses the model's performance before and after augmentation using evaluation metrics such as precision, recall, F1-score, and confusion matrix. The results demonstrate the substantial impact of augmentation in enhancing the model’s accuracy and reliability in identifying different crying patterns[63]. This analysis offers valuable insights relevant to the advancement of audio-based classification systems.

3.1. Evaluation Model Infant Cry Classification

The infant cry classification model is to be evaluated according to the initial data in the donate a cry corpus of 457, Leading to the creation of table 1 below:

Table 1. Evaluation Of Initial Infant Cry Dataset Classification

	Precision	recall	f1-score	support
belly_pain	0.00	0.00	0.00	4

Vol. 6, No. 2, April 2025, Page. 995-1016
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4373

burping	0.00	0.00	0.00	1
discomfort	0.00	0.00	0.00	8
hungry	0.79	0.99	0.88	73
tired	0.00	0.00	0.00	6
accuracy			0.78	92
macro avg	0.16	0.20	0.18	92
weighted avg	0.63	0.78	0.70	92

The model performs well in the hungry category, but is very weak in recognizing other categories, which is due to data imbalance. Therefore, additional data augmentation or weighting techniques are required to improve performance on minimum classes.

3.2. Augmentation Results

The data augmentation process in this study effectively expanded the Donate-a-Cry corpus from 457 to 8.683 samples, addressing the issue of class imbalance. The belly pain category grew from 16 to 304 samples, burping from 8 to 152, discomfort from 27 to 513, hunger from 382 to 7.258, and fatigue from 24 to 456 will look like the table 2 below.

Table 2. Infant Cry Data Before And After Augmentation

No	Infant Cry Category	Sample Before Augmentation	Sample After Augmentation
1	Belly Pain	16	304
2	Burping	8	152
3	Discomfort	27	513
4	Hungry	382	7258
5	Tired	24	456

The inclusion of this data substantially enhances the representation of each class, particularly minority classes like abdominal pain, burping, and discomfort, which were previously underrepresented. This result can be seen in Figure 4, which shows that augmentation not only increases the amount of data, but also enriches the variety of infant crying patterns, which is expected to improve the model's ability to precisely and consistently identify different cry categories.

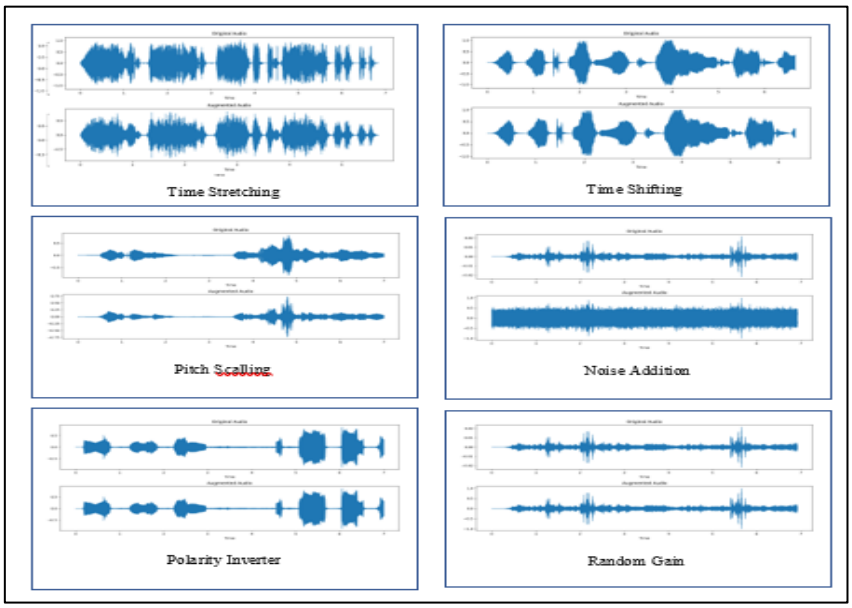

Figure 4. Audio Data Augmentation Results

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 2, April 2025, Page. 995-1016
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4373

As illustrated in Figure 4, the augmentation of audio data through the application of diverse techniques has been shown to enhance the diversity and variety of signal patterns. Post-augmentation, time stretching results in a signal with an altered duration, while preserving its frequency. In contrast, time shifting involves the manipulation of the signal's position in the time domain, either forward or backward.Pitch scaling techniques modify the pitch of the signal, thereby altering its perceived loudness. Noise addition, on the other hand, involves the introduction of random noise to simulate varied environmental conditions. Polarity inversion reverses the polarity of the signal, creating an inverted but acoustically identical version, and random gain changes the amplitude of the signal randomly, simulating volume variations.The results demonstrate that each augmentation technique produces a unique transformation on the original signal, thereby enriching the training data and enhancing the generalisation ability of the infant cry recognition model.

3.3. Mel - Frequency Cepstral Coefficients Results

This chapter presents the research findings, emphasizing the use of Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction. The approach involved converting infant cry signals into numerical representations, which were subsequently analyzed using a series of algorithms[53], [54]. These processes involved framing, the Fourier Transform, a Mel filter bank, and the Discrete Cosine Transform (DCT), designed to extract key acoustic features. The results of these MFCC features formed the basis for training and testing a model designed to recognise patterns in baby crying. The results of this study can be seen in Figure 5 below:

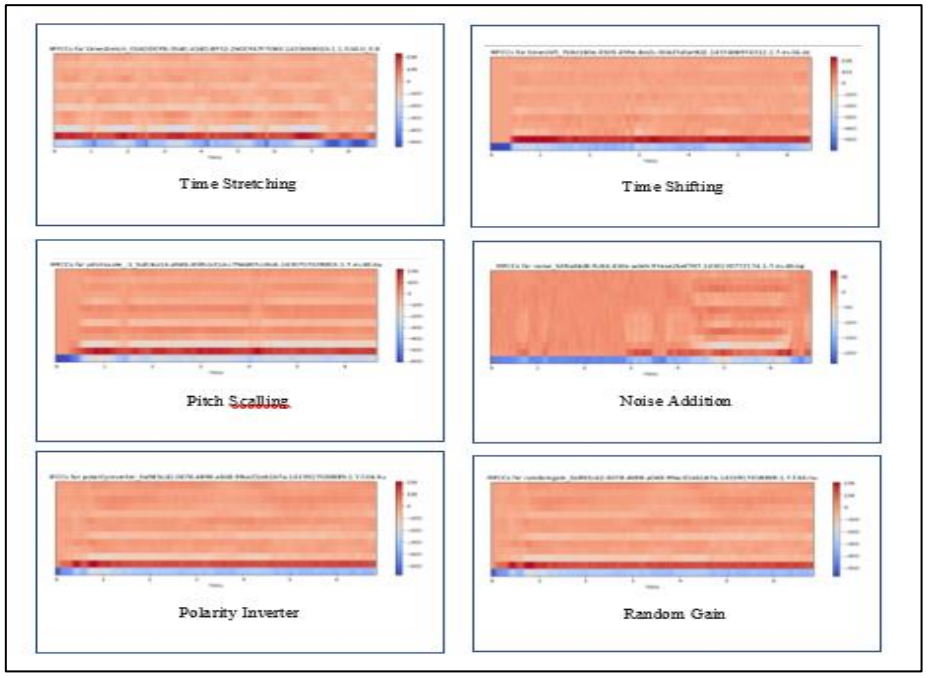

Figure 5. Mel – Frequency Cepstral Coefficients (MFCCs) Results

As depicted in Figure 5, the application of Mel-Frequency Cepstral Coefficients (MFCC) feature extraction following the implementation of various data augmentation techniques, such as time stretching, time shifting, pitch scaling, noise addition, polarity inversion, and random gain, produces a variety of outcomes. The use of these techniques generates diverse spectrogram patterns that depict the variation in frequency energy over time. Each augmentation technique produces a different spectrogram pattern, reflecting the changes in acoustic characteristics caused by the augmentation. As illustrated, time-related operations such as stretching and shifting modify the duration or time position of the signal,

while pitch scaling affects the dominant frequency. The addition of noise introduces noise energy at various frequencies, while polarity inverters and random gain modify the amplitude intensity without affecting the primary pattern. These MFCC patterns create a more varied set of features, improving the baby cry recognition model's robustness to data variations.

3.4. Convolutional Neural Network (CNN) Performance Result

The successful application of the MFCC-based data augmentation technique enhanced the dataset's diversity, allowing the CNN model to identify crying patterns with improved accuracy[54], [58]. The performance assessment, utilizing metrics like precision, recall, F1-score, and confusion matrix, demonstrates that data augmentation significantly improves the model's capability to recognize different cry categories, especially in previously underrepresented classes affected by data imbalance. These findings validate the effectiveness of augmentation in improving model generalization for infant cry recognition applications.

Table 3. Convolutional Neural Network Results Performance After Audio Augmentation

	precision	recall	f1-score	support
belly_pain	0.97	0.87	0.92	69
burping	0.97	0.84	0.90	37
discomfort	1.00	0.86	0.92	93
Hungry	0.97	1.00	0.99	1441
tired	0.99	0.90	0.94	99
Accuracy			0.98	1739
macro avg	0.98	0.89	0.93	1739
weighted avg	0.98	0.98	0.98	1739

As Table 3 illustrates, the employment of augmented data led to substantial enhancement in the recognition of baby crying patterns across all categories. The model achieved 98% accuracy, with a macro average precision of 0.98, recall of 0.89, and an F1-score of 0.93, demonstrating strong performance across all categories. The weighted average metrics, which prioritize categories with more instances, recorded precision, recall, and F1-scores of 0.98, highlighting the model's high accuracy, especially in majority classes like 'hungry,' where the F1-score reached 0.99. Furthermore, minority classes such as belly pain, burping, and discomfort exhibited significant improvement, achieving F1-scores of 0.92, 0.90, and 0.92, respectively, due to the enhanced data diversity introduced by augmentation techniques. These findings collectively indicate that data augmentation effectively enhances the balanced performance of the model, thereby ensuring enhanced reliability in recognizing diverse infant cries.

As presented in Figure 6, the confusion matrix illustrates the results of the classification of infant crying patterns by a CNN model following audio augmentation. This matrix reflects the model's prediction of five categories of baby cries: belly_pain, burping, discomfort, hungry, and tired. The model exhibited high accuracy in dominant categories like hungry, correctly predicting 1,437 instances with only a small number of misclassifications. It is noteworthy that other categories, such as belly_pain and discomfort, also had a high number of correct predictions (64 and 86, respectively), although there were minor errors, such as some incorrect predictions in other categories.The largest errors were observed in the burping and tired categories, with several instances misclassified into other categories.These results demonstrate that data augmentation effectively enhanced the model's capacity to recognise diverse categories of infant cries. However, there is still potential for improvement in predictions within minority classes.

Vol. 6, No. 2, April 2025, Page. 995-1016
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4373

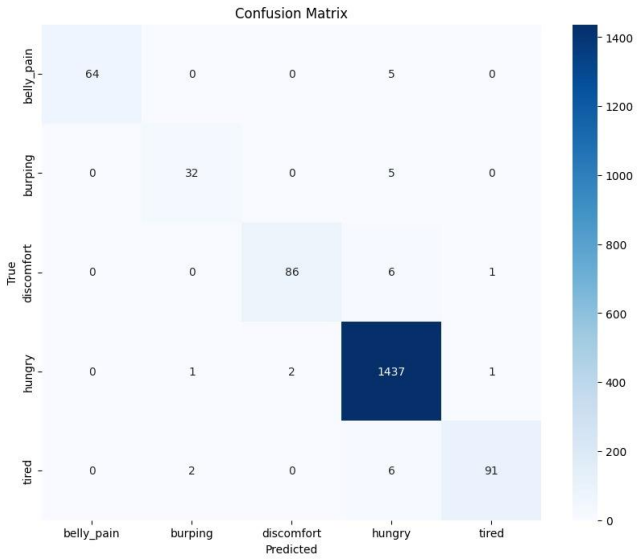

Figure 6. Confusion Matrix Result Of Infant Cry After Audio Augmentation

3.5. Statistics Test Results

To verify the statistical significance of the improvement in model accuracy after applying MFCC-based audio data augmentation, a paired t-test was conducted between the accuracy values obtained before and after augmentation.

The results of the paired t-test revealed a p-value of less than 0.05, indicating a statistically significant improvement in model accuracy ($p < 0.05$).

Therefore, the application of data augmentation techniques was not only beneficial in enhancing classification metrics but also statistically proven to yield meaningful improvements.

3.6. Analysis Of Minor Class

The following analysis is concerned with the minor class. Augmentation has been demonstrated to enhance the performance of minor classes through several key mechanisms. Firstly, augmentation improves data representation, thereby ensuring that the model no longer relies exclusively on the majority data. Secondly, augmentation assists in reducing the classification bias towards the majority category, which previously rendered the minor class challenging to recognise. Thirdly, augmentation increases the recall value, signifying that the model is more sensitive in recognising minor classes that were previously overlooked. Finally, augmentation enhances the model's resilience to variations in real data, as it generates more diverse versions of the cry signal.

After the augmentation process, there was a significant improvement in the performance of minority classes. The F1-score for belly pain increased from 0.00 to 0.92, while burping rose from 0.00 to 0.90, and discomfort improved from 0.00 to 0.92. This improvement highlights the effectiveness of augmentation in enhancing the detection of underrepresented classes, leading to a more balanced model capable of accurately identifying all categories.

The majority of misclassifications were observed in the burping and tired categories, presumably due to the acoustic patterns exhibited by these categories sharing similarities with other categories, such as burping, which is frequently classified as hungry or discomfort due to its comparable frequency characteristics, and tired, which is occasionally classified as hungry due to the similarity in variations exhibited by the crying pattern. Prior to augmentation, minor classes such as belly pain, burping, and discomfort exhibited an F1-score of 0.00, signifying that the model failed to recognise these categories due to the paucity of data.With the implementation of MFCC-based augmentation, which incorporates time stretching, pitch scaling, noise addition, polarity inversion, and random gain adjustments, the

Vol. 6, No. 2, April 2025, Page. 995-1016
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4373

number of samples increased considerably, enabling the model to learn more acoustic patterns from each category. Consequently, the F1-score for the belly pain category increased to 0.92, for burping to 0.90, and for discomfort to 0.92, suggesting that the augmentation effectively enhanced the model's capacity to detect categories that had previously been challenging to recognise. The hungry category, which had previously demonstrated an F1-score of 0.88, exhibited an increase to 0.99. Similarly, the tired category, which had an initial F1-score of 0.00, showed an increase to 0.94, thereby substantiating the augmentation's efficacy not only in enhancing the minor classes but also in optimising the model's generalisation capabilities, As demonstrated in figure 7, a comparison of performance results before and after augmentation is provided.

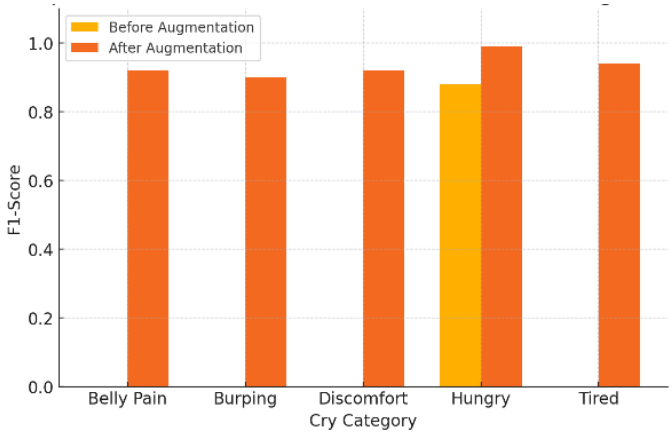

Figure 7. Comparison Of Model Performance Before And After Augmentation.

4. DISCUSSIONS

The results of this study indicate that utilizing Mel-Frequency Cepstral Coefficients (MFCC)-based data augmentation techniques significantly enhances the CNN model's capability in recognizing infant cry patterns. Augmentation led to an augmentation in the number of datasets from 457 to 8.683, thereby successfully addressing the imbalance in data distribution across classes. It is noteworthy that minority categories, such as belly pain, burping, and discomfort, which had previously exhibited very low performance, demonstrated an augmentation-induced enhancement in their F1-score to 0.92, 0.90, and 0.92, respectively. This finding underscores the efficacy of augmentation in enriching the data representation, thus improving the model's robustness to variations in crying patterns.

Moreover, the model attained an overall accuracy of 98%, with weighted average precision, recall, and F1-score all reaching 0.98, indicating optimal performance across all categories. In comparison to antecedent studies, such as those conducted by Turgut et al. who utilised a conventional approach devoid of augmentation [25], [26], these outcomes underscore substantial advantages with respect to the generalisability and sensitivity of the model [27], [64], [65], [66]. However, while dominant categories such as 'hungry' exhibit nearly perfect accuracy (F1-score 0.99), minor inaccuracies persist in less prevalent categories, such as 'tired', necessitating additional scrutiny.

This research also confirms the importance of MFCC as a key feature in infant cry pattern recognition, Particularly when integrated with audio augmentation methods like time stretching, pitch scaling, and noise addition [67], [68], [69]. Although data augmentation had a positive impact, there is room for further development, such as exploring advanced augmentation techniques or using more complex model architectures to improve prediction on minority classes[21], [43], [48], [70]. The findings of this study provide a valuable contribution to the advancement of artificial intelligence-based technologies for enhancing modern infant care.

Vol. 6, No. 2, April 2025, Page. 995-1016
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4373

Table 4. Comparison Of Infant Cry Classification Studies

No	Study	Methodology	Dataset	Performance Metrics	Key Findings
1	This Study (MFCC + CNN + Data Augmentation)	MFCC + CNN + Augmentation (Time Stretching, Pitch Scaling, Noise Addition)	Donate-a-Cry Corpus (Augmented from 457 to 8,683 samples)	98% Accuracy, 0.98 Precision, Recall, F1-score	Significant improvement in minority class classification, balancing dataset
2	Turgut et al. (2022)[26]	Deep Learning + Handcrafted Features	Various Infant Cry Datasets	Various results, Deep Learning models perform better than classical ML	Handcrafted features can still be useful but less effective than deep learning
3	Ji et al. (2021) [29]	CNN + Spectrogram Analysis	Custom Dataset	CNN achieves better accuracy than ANN	Spectrogram-based CNNs improve feature extraction
4	Ozseven (2023)[25]	Multiple Deep Learning Models + Handcrafted Features	Donate-a-Cry Corpus	Comparison of different deep learning models	Deep learning outperforms traditional methods
5	Zayed et al. (2023) [31]	Deep Learning + Feature Fusion	Public Infant Cry Datasets	Feature fusion improves classification accuracy	Fusion of multiple features improves overall classification accuracy

Figure 7 explains this study utilised MFCC, CNN, and data augmentation techniques to enhance infant cry classification, particularly for minority classes, by balancing the dataset and attaining 98% accuracy.In comparison to previous research, Turgut et al. (2022) investigated handcrafted features with deep learning, demonstrating that traditional features remain useful but less effective than deep models[26]. Ji et al. (2021) found that CNNs using spectrogram analysis outperform ANN-based approaches [29]. In the field of infant cry analysis, Ozseven (2023) conducted a comparative analysis of multiple deep learning models, thereby reinforcing the prevailing notion that deep learning systems exhibit superior performance in comparison to conventional classification methods [25]. Building upon these findings, Zayed et al. (2023) demonstrated that the integration of multiple features through feature fusion enhances the accuracy of classification processes [31]. These studies collectively highlight the increasing prominence of deep learning and data augmentation techniques in improving the accuracy and effectiveness of infant cry classification systems.

The augmentation techniques utilized in this study, including time stretching, pitch scaling, noise addition, polarity inversion, and random gain adjustment, have proven to outperform traditional methods by enhancing the diversity of infant cry signals while maintaining essential acoustic features. In contrast to conventional augmentation techniques that employ rudimentary transformations, This study employs MFCC-based augmentation, ensuring the preservation of both the temporal and spectral characteristics of the cry signal, a prerequisite for effective classification. In comparison to earlier studies that employed rudimentary data replication or noise addition, this approach introduces a more diverse range of variations in frequency, amplitude, and time, thereby enhancing the model's resilience to variations in baby cries that may occur in real-world settings.

A study on enhancing the classification of infant vocalisations through the utilisation of MFCC-based augmentation and CNN has identified several limitations. The primary constraint pertains to the

dataset utilised, which exhibits constrained environmental variation and may not accurately replicate actual acoustic conditions. This could negatively impact the model's ability to recognize baby cries in varying environmental conditions. Moreover, the study exclusively focuses on audio features, disregarding the potential value of incorporating visual or physiological data, which may offer a more precise assessment of the infant's condition. The reliance on data augmentation to address the imbalance of cry categories is also questionable, given the necessity of more diverse original data for effective model training.

The CNN model utilised in the study exhibits a propensity for overfitting, a tendency that is particularly pronounced when the number of samples increases due to augmentation. Despite its high accuracy, there is a possibility that The model becomes too tailored to the training dataset and may struggle to perform effectively on new, unseen data. To resolve this issue, future model developments should incorporate multimodal data, including audio, visual, and physiological features. Integrating elements like facial expressions and breathing patterns can provide a more holistic understanding of the context behind a baby's cry. Additionally, the adoption of transformer models or self-supervised learning techniques, such as Wav2Vec or HuBERT, could enhance the model's ability to recognize crying patterns without relying on large amounts of labeled data. These advancements present opportunities to further improve the model's effectiveness in recognizing and interpreting baby cries.

The potential exists for further development through the implementation of edge computing-based infant cry recognition in devices such as smart baby monitors or mobile applications[71], [72]. This approach facilitates real-time detection of infant cries, obviating the need for reliance on cloud servers that require a stable internet connection.Additionally, the model can be trained with data from various environments, including hospitals, households, and daycares, thereby enhancing its robustness to noise and different acoustic conditions[73], [74], [75]. Another potential avenue for expansion of the dataset would be through crowdsourcing, namely the collection of baby crying recordings from diverse geographical locations and cultural backgrounds. This approach would serve to enhance the generalisability of the model, enabling its capacity to recognise patterns in baby crying from a more extensive range of backgrounds.

Furthermore, the application of a paired t-test demonstrated that the observed improvement in classification accuracy was statistically significant ($p < 0.01$). This strengthens the assertion that data augmentation using MFCC not only improves the performance metrics but also significantly enhances the model's generalization ability in a statistically validated manner.

5. CONCLUSTION

This study successfully enhanced the classification of infant cry patterns by implementing MFCC-based audio augmentation and CNN models. This enhancement resulted in a substantial improvement in accuracy, from 78% to 98%, particularly in the classification of minority categories such as belly pain, burping, and discomfort, which were previously challenging to categorise. The study addresses a significant challenge in infant cry recognition, namely data imbalance, by augmenting the dataset from 457 to 8,683 samples through a range of techniques, including time stretching, pitch scaling, noise addition, polarity inversion, and random gain adjustments. This approach enhances the robustness of the model and facilitates better generalisation of infant cry patterns. This study makes an important contribution to the fields of computer science and artificial intelligence, particularly in audio signal processing, deep learning, and healthcare technology applications, by showing that data augmentation can improve the performance of CNN models in the classification of complex baby sounds.However, the study is not without its limitations, especially in the lack of environmental variation in the dataset, as the model has not been tested with data from various real conditions such as multiple recording sources, varying sound backgrounds, and different devices. To this end, future research should integrate

multimodal approaches by combining visual and physiological features, apply more advanced deep learning models such as Transformer or self-supervised learning, and develop edge computing systems for real-time infant cry detection in smart devices. These improvements will optimise infant cry recognition technology for real-world applications, improve detection accuracy, and help parents and healthcare professionals better understand and respond to an infant's needs in a timelier and more accurate manner.

CONFLICT OF INTEREST

The authors affirm that there are no conflicts of interest related to the research discussed in this paper titled "Improving Infant Cry Recognition Using Mfcc And Cnn-Based Audio Augmentation". All aspects of the research, including data collection, methodology, analysis, and interpretation, were performed impartially and without personal or professional conflicts. In addition, the authors certify that there are no competing interests or relationships with any party, organization, or individual that could affect the results of the research or its integrity.

REFERENCES

[1] 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.

[2] 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.

[3] G.-V. Morfi, "Automatic detection and classification of bird sounds in low-resource wildlife audio datasets," 2019.

[4] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, and Z. Gong, "Unsupervised Representation Learning with Long-Term Dynamics for Skeleton Based Action Recognition." [Online]. Available: www.aaai.org

[5] P. Inkeaw, "Mel Frequency Cepstral Coefficient MFCC."

[6] A. Kumar, D. R. P. M. Vincent, K. Srinivasan, and C. Y. Chang, "Deep Convolutional Neural Network based Feature Extraction with optimized Machine Learning Classifier in Infant Cry Classification," in *2020 International Conference on Decision Aid Sciences and Application, DASA 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020, pp. 27–32. doi: 10.1109/DASA51403.2020.9317240.

[7] S. Jain and B. Kishore, "Comparative study of voice print Based acoustic features: MFCC and LPCC," *International Journal of Advanced engineering, Management and Science*, vol. 3, no. 4, pp. 313–315, 2017, doi: 10.24001/ijaems.3.4.5.

[8] G. Iglesias, E. Talavera, Á. González-Prieto, A. Mozo, and S. Gómez-Canaval, "Data Augmentation techniques in time series domain: a survey and taxonomy," May 01, 2023, *Springer Science and Business Media Deutschland GmbH*. doi: 10.1007/s00521-023-08459-3.

[9] 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

[10] 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.

[11] S. Y. Chuang, H. M. Wang, and Y. Tsao, "Improved Lite Audio-Visual Speech Enhancement," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 30, pp. 1345–1359, 2022, doi: 10.1109/TASLP.2022.3153265.

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

[13] 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.
[14] V. A. Kherdekar, "Convolution Neural Network Model for Recognition of Speech for Words used in Mathematical Expression," 2021.
[15] 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
[16] M. Margaryan, M. Seibold, I. Joshi, M. Farshad, P. Fürnstahl, and N. Navab, "Improved Techniques for the Conditional Generative Augmentation of Clinical Audio Data," Nov. 2022, [Online]. Available: http://arxiv.org/abs/2211.02874
[17] T. Wang, H. Guo, Q. Zhang, and Z. Yang, "A new multilayer graph model for speech signals with graph learning," Apr. 15, 2022, *Elsevier Inc.* doi: 10.1016/j.dsp.2021.103360.
[18] K. Zhang, H. N. Ting, and Y. M. Choo, "Baby cry recognition based on WOA-VMD and an improved Dempster–Shafer evidence theory," *Comput Methods Programs Biomed*, vol. 245, Mar. 2024, doi: 10.1016/j.cmpb.2024.108043.
[19] K. Zhang, H. N. Ting, and Y. M. Choo, "Baby Cry Recognition by BCRNet Using Transfer Learning and Deep Feature Fusion," *IEEE Access*, vol. 11, pp. 126251–126262, 2023, doi: 10.1109/ACCESS.2023.3330789.
[20] T. Zhang, C. Hong, Y. Zou, and J. Zhao, "Prediction method of human defecation based on informer audio data augmentation and improved residual network," *Heliyon*, vol. 10, no. 14, Jul. 2024, doi: 10.1016/j.heliyon.2024.e34145.
[21] 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.
[22] H. Kheddar, M. Hemis, and Y. Himeur, "Automatic Speech Recognition using Advanced Deep Learning Approaches: A survey," Mar. 2024, doi: 10.1016/j.inffus.2024.102422.
[23] G.-V. Morfi, "Automatic detection and classification of bird sounds in low-resource wildlife audio datasets," 2019.
[24] H. N. Ting, Y. M. Choo, and A. Ahmad Kamar, "Classification of asphyxia infant cry using hybrid speech features and deep learning models," *Expert Syst Appl*, vol. 208, Dec. 2022, doi: 10.1016/j.eswa.2022.118064.
[25] 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.
[26] 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.
[27] 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.
[28] J. Li, M. Hasegawa-Johnson, and N. L. McElwain, "Analysis of acoustic and voice quality features for the classification of infant and mother vocalizations," *Speech Commun*, vol. 133, pp. 41–61, Oct. 2021, doi: 10.1016/j.specom.2021.07.010.
[29] C. Ji and Y. Pan, "Infant Vocal Tract Development Analysis and Diagnosis by Cry Signals with CNN Age Classification."
[30] C. Ji, "Infant Cry Signal Processing, Analysis, and Classification with Artificial Neural Networks," *Dissertation*, 2021, doi: 10.57709/25943253.
[31] Y. Zayed, A. Hasasneh, and C. Tadj, "Infant Cry Signal Diagnostic System Using Deep Learning and Fused Features," *Diagnostics*, vol. 13, no. 12, Jun. 2023, doi: 10.3390/diagnostics13122107.

[32] 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.

[33] F. Anders, M. Hlawitschka, and M. Fuchs, "Automatic classification of infant vocalization sequences with convolutional neural networks," *Speech Commun*, vol. 119, pp. 36–45, May 2020, doi: 10.1016/j.specom.2020.03.003.

[34] K. R. Mannem, E. Mengiste, S. Hasan, B. G. de Soto, and R. Sacks, "Smart audio signal classification for tracking of construction tasks," *Autom Constr*, vol. 165, Sep. 2024, doi: 10.1016/j.autcon.2024.105485.

[35] S. Purkovic *et al.*, "Audio analysis with convolutional neural networks and boosting algorithms tuned by metaheuristics for respiratory condition classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 10, Dec. 2024, doi: 10.1016/j.jksuci.2024.102261.

[36] H. S. Alar, R. O. Mamaril, L. P. Villegas, and J. R. D. Cabarrubias, "Audio classification of violin bowing techniques: An aid for beginners," *Machine Learning with Applications*, vol. 4, p. 100028, Jun. 2021, doi: 10.1016/j.mlwa.2021.100028.

[37] D. A. Villamizar, D. G. Muratore, J. B. Wieser, and B. Murmann, "An 800 nW switched-capacitor feature extraction filterbank for sound classification," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 4, pp. 1578–1588, Apr. 2021, doi: 10.1109/TCSI.2020.3047035.

[38] A. Gorin, C. Subakan, S. Abdoli, J. Wang, S. Latremouille, and C. Onu, "Self-Supervised Learning for Infant Cry Analysis," in *ICASSPW 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing Workshops, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICASSPW59220.2023.10193421.

[39] 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.

[40] X. Yao, M. Micheletti, M. Johnson, E. Thomaz, and K. de Barbaro, "INFANT CRYING DETECTION IN REAL-WORLD ENVIRONMENTS," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 131–135. doi: 10.1109/ICASSP43922.2022.9746096.

[41] A. Sharma and D. Malhotra, "Speech recognition based IICC - Intelligent infant cry classifier," in *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, Institute of Electrical and Electronics Engineers Inc., Aug. 2020, pp. 992–998. doi: 10.1109/ICSSIT48917.2020.9214193.

[42] D. Budaghyan, C. C. Onu, A. Gorin, C. Subakan, and D. Precup, "CRYCELEB: A SPEAKER VERIFICATION DATASET BASED ON INFANT CRY SOUNDS," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 11966–11970. doi: 10.1109/ICASSP48485.2024.10448292.

[43] 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.

[44] 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.

[45] 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.

[46] M. Y. Yiwere, A. Barcovschi, R. Jain, H. Cucu, and P. Corcoran, "Augmentation Techniques for Adult-Speech to Generate Child-Like Speech Data Samples at Scale," *IEEE Access*, vol. 11, pp. 109066–109081, 2023, doi: 10.1109/ACCESS.2023.3317360.

[47] 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.

[48] 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.

[49] 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

[50] 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.

[51] Y. Li, J. Chan, G. Peko, and D. Sundaram, "Mixed emotion extraction analysis and visualisation of social media text," *Data Knowl Eng*, vol. 148, Nov. 2023, doi: 10.1016/j.datak.2023.102220.

[52] 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.

[53] Z. Firas, A. A. Nashaat, and G. Ahmad, "Optimizing Infant Cry Recognition: A Fusion of LPC and MFCC Features in Deep Learning Models," in *International Conference on Advances in Biomedical Engineering, ICABME*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 232–237. doi: 10.1109/ICABME59496.2023.10293083.

[54] 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.

[55] 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.

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

[57] 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.

[58] 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.

[59] X. Yu, X. Zhao, C. Lu, L. Wang, X. Long, and W. Chen, "An investigation into audio features and DTW algorithms for infant cry classification," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Nov. 2019, pp. 54–59. doi: 10.1145/3375923.3375929.

[60] A. M. Mahmoud, S. M. Swilem, A. S. Alqarni, and F. Haron, "Infant Cry Classification Using Semi-supervised K-Nearest Neighbor Approach," in *Proceedings - International Conference on Developments in eSystems Engineering, DeSE*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 305–310. doi: 10.1109/DeSE51703.2020.9450239.

[61] L. Liu, W. Li, X. Wu, and B. X. Zhou, "Infant cry language analysis and recognition: An experimental approach," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 3, pp. 778–788, May 2019, doi: 10.1109/JAS.2019.1911435.

[62] A. Abbasi, A. R. R. Javed, A. Yasin, Z. Jalil, N. Kryvinska, and U. Tariq, "A Large-Scale Benchmark Dataset for Anomaly Detection and Rare Event Classification for Audio Forensics," *IEEE Access*, vol. 10, pp. 38885–38894, 2022, doi: 10.1109/ACCESS.2022.3166602.

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

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

[65] R. Garg, "'its Changes so Often': Parental Non-/Use of Mobile Devices while Caring for Infants and Toddlers at Home," *Proc ACM Hum Comput Interact*, vol. 5, no. CSCW2, Oct. 2021, doi: 10.1145/3479513.

[66] K. Rezaee, H. G. Zadeh, L. Qi, H. Rabiee, and M. R. Khosravi, "Can You Understand Why I Am Crying? A Decision-making System for Classifying Infants' Cry Languages Based on DeepSVM Model," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 1, Jan. 2024, doi: 10.1145/3579032.

[67] M. Hammoud, M. N. Getahun, A. Baldycheva, and A. Somov, "Machine learning-based infant crying interpretation," *Front Artif Intell*, vol. 7, 2024, doi: 10.3389/frai.2024.1337356.

[68] 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.

[69] A. Gorin, C. Subakan, S. Abdoli, J. Wang, S. Latremouille, and C. Onu, "Self-supervised learning for infant cry analysis," May 2023, [Online]. Available: http://arxiv.org/abs/2305.01578

[70] A. S. Kumar, T. Schlosser, S. Kahl, and D. Kowerko, "Improving learning-based birdsong classification by utilizing combined audio augmentation strategies," *Ecol Inform*, vol. 82, Sep. 2024, doi: 10.1016/j.ecoinf.2024.102699.

[71] A. Ayari, H. Hamdi, and K. A. Alsulbi, "E-health Application In IoMT Environment Deployed in An Edge And Cloud Computing Platforms," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 1019–1028. doi: 10.1016/j.procs.2024.09.521.

[72] X. He and Q. Zhang, "Cloud Computing Based Digital Media Content Distribution Technology," in *Procedia Computer Science*, Elsevier B.V., 2023, pp. 461–468. doi: 10.1016/j.procs.2024.10.055.

[73] H. Malik, U. Bashir, and A. Ahmad, "Multi-classification neural network model for detection of abnormal heartbeat audio signals," *Biomedical Engineering Advances*, vol. 4, p. 100048, Dec. 2022, doi: 10.1016/j.bea.2022.100048.

[74] D. Vasconcelos, N. J. Nunes, A. Förster, and J. P. Gomes, "Optimal 2D audio features estimation for a lightweight application in mosquitoes species: Ecoacoustics detection and classification purposes," *Comput Biol Med*, vol. 168, Jan. 2024, doi: 10.1016/j.compbiomed.2023.107787.

[75] 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.