

Rachna_Thesis_Intro

by Rachna Patil

Submission date: 21-May-2026 07:46AM (UTC+0530)

Submission ID: 2966052576

File name: Thesis_11_-pages-2.pdf (2.19M)

Word count: 6268

Character count: 35580

Chapter 1

INTRODUCTION

1.1 Motivation

Infant crying is ¹⁶ one of the earliest forms of non-verbal communication used by infants to communicate ⁷ their physical and emotional conditions such as hunger, distress, pain, and stress. Caregivers and healthcare providers often use cry sounds to make judgments about the physiological state of the infant and provide necessary attention. Hence, proper analysis of cries could help in diagnosing the conditions at an early stage especially in neonatal intensive care units (NICU) and home care settings.

Traditional infant cry signal classification methodologies mainly utilized handcrafted acoustic features such as MFCC and LPC along with machine learning (ML) classifiers like SVM and RF. While these approaches yielded some success, they were not robust enough to cope with noise and variations in signals, and handcrafted acoustic features may not be able to capture all complexities of infant cry sounds.

Recently, there have been significant improvements in audio classification using DL techniques. The use of CNNs, RNNs, LSTMs, and GRUs indicates that these models are capable of efficiently learning the temporal and spectral features present in audio signals. In addition, oversampling techniques such as SMOTE, ADASYN, Borderline-SMOTE, and RMO-DA have ³² proven to be an effective means of addressing the problem of class imbalance in infant cry datasets. This has prompted further investigation into the impact of the combination of DL techniques and class balancing on infant cry classification accuracy.

1.2 Problem Statement

While there have been significant advancements in the area of infant cry classification, some issues still persist. For instance, the current databases for infants' cries are generally small, highly imbalanced, and recorded in an uncontrolled environment where noise is inevitable. Minor classes such as burping and belly pain are also neglected, causing bias in the models towards majority classes, such as hungry or crying. Previous researches have mainly focused on single DL algorithms or used different preprocessing pipelines, augmentation strategies,

and evaluation setup, making fair comparison difficult. Standard ML algorithms have shown poor results in terms of generalizability under noisy environments, mostly because of their dependence on hand-crafted features. While oversampling methods like SMOTE and ADASYN have been recommended to address the issue of imbalanced classes, there has been little work done to compare different oversampling techniques in the context of DL models within the same experimental setup.

There is a clear necessity for developing a standardized process for assessing CNN, RNN, LSTM, and GRU models based on consistent preprocessing, data augmentation, feature extraction, and oversampling techniques to determine the best way forward for cry classification.

1.3 Overview of the Proposed Work

The research implements a comparative analysis of four deep learning approaches that perform the task of detecting and classifying the sounds made by babies. The proposed methodology consists of preprocessing, feature extraction, data augmentation, oversampling, and model evaluation in an equal experimental setting to ensure a fair comparison among the models. The dataset used in this study includes different kinds of baby cries, namely hunger cry, burping cry, abdominal pain cry, distress cry, crying, laughing, noise, and silence. The preprocessing stage includes normalization, silence trimming, mono conversion, and resampling of raw audio. Further, several salient audio features are extracted from spectrogram representations of the input data. These include MFCC (Mel-frequency cepstral coefficients), ZCR (Zero Crossing Rate), RMS (Root Mean Square), and chroma features.

Data augmentation techniques used in this study include adding noise, pitch shifting, time stretching, time shifting, and SpecAugment. Five oversampling techniques, which include Random Oversampling, SMOTE, Borderline-SMOTE, ADASYN, and RMO-DA, are implemented to solve class imbalance problem. Finally, CNN, RNN, LSTM, and GRU models are developed using five-fold cross-validation technique. Various performance metrics like accuracy, precision, recall, F1-score, and confusion matrix are used to evaluate the performance of the models.

1.4 Objectives of the Study

The following are the main purposes of this research work:

1. To improve automatic classification of infant cries using deep learning (DL).

2. To apply different DL algorithms such as CNN, RNN, LSTM, and GRU to develop infant cry detection models.
3. The study aims to evaluate the effects of using various oversampling approaches ³⁸ such as Random Oversampling, SMOTE, Borderline-SMOTE, ADASYN, and RMO-DA on balancing classes.
4. Performance evaluation will be done ² based on accuracy, precision, recall, and F1-score.
5. Improve the minority class effectiveness and increase overall accuracy rates in infant cry classifiers.
6. To give insights into the relationship between DL architectures and imbalance control plan for robust infant cry detection

Chapter 2

RELATED WORK

From the perspective of algorithms used in the classification of baby crying sounds, the development is from classical machine learning algorithms to modern deep learning algorithms. While the former focused mainly on handcrafted acoustic feature engineering along with classical classification, the latter focuses on deep learning and its variants such as attention-based, oversampling, and transformer methods. Although substantial advancements have been made in this field, there are still open issues such as class imbalance, noisy signals, small datasets, and lack of a unified pipeline for evaluation. This chapter discusses the existing literature on infant cry classification, deep learning models, and data augmentation techniques.

2.1 Traditional Machine Learning Approaches

Studies conducted in an attempt to classify baby cries using traditional machine learning methods were primarily focused on approaches such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF). Traditional machine learning was mainly dependent on the manually extracted features of acoustic parameters like Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC). Manual feature extraction was intended to study the underlying spectral and temporal properties of the cry signal.

While these approaches achieved reasonable accuracy, they had some limitations. The handcrafted features often lacked the ability to incorporate the temporal-spectral information of infant cry signals. Additionally, the conventional classifiers showed limited robustness in situations where there is noise and variability in the signals.

2.2 Deep Learning Based Infant Cry Classification

In recent times, there have been several developments in the field of deep learning approaches, which have led to marked improvement in the area of classification of infant cries. Convolutional neural networks, recurrent neural networks, LSTM (Long Short-Term Memory), and GRU (Gated Recurrent Unit) networks can be considered powerful when it comes to discerning audio-based patterns. CNN based architectures have been shown excellent performance in

extracting local spectral interpretations of spectrograms, while sequential models such as RNN, LSTM, and GRU efficiently model temporal dependencies in cry signals. Comparative studies indicate that CNN generally achieves higher total classification accuracy, whereas sequential architectures enhance recall for minority classes due to their temporal learning capability. A number of researchers have proposed hybrid DL frameworks for improving classification performance. Maghfira et al. [1] presented CNN-RNN hybrid architectures for joint spectral temporal modelling. Özseven [2] joint handcrafted acoustic features with DNN (Deep Neural Networks) and demonstrable progress in infant cry recognition tasks. Likewise, Alam and Mamun [3] proposed CNN+DNN hybrid frameworks with improved augmentation and feature extraction strategies, achieving state-of-the-art achievement.

Moreover, self supervised and transfer learning approaches have also been explored. Gorin et al. [4] used self supervised learning to leverage unlabeled infant cry datasets, while Anjali et al. [5] examined the transfer learning approach using pre-trained models for infant cry classification.

2.3 Transformer-Based and Attention-Based Models

The current research trend has focused on using transformer-like models for studying infant cries, which take into consideration larger time-dependent factors when compared to classical recurrent neural networks.

In addition, transformer models similar to CryFormer and spectrogram fusion attention network have shown great robustness even under noisy recording. Transformer models outperform classical sequential models in multiple audio classification tasks.

2.4 Data Augmentation Techniques

The application of ¹¹ data augmentation has become an essential method in improving the robustness and generalization ability of infant cry classifiers. The fact that infant cry databases usually have limited samples means that data augmentation is used to add variability to the data. Commonly used augmentation techniques include:

- Noise injection
- Pitch shifting
- Time stretching

- Time shifting
- Gain variation
- SpecAugment

Ko et al. [8] showed the effectiveness of audio augmentation techniques in speech recognition tasks, whereas Park et al. [9] introduced SpecAugment for improving robustness through time frequency masking. Generative augmentation techniques additionally had just been explored to create more realistic synthetic cry samples. Meephiw and Leesutthipornchai [10] working MFCC based feature selection strategies in order to improve class separation, whereas Firas et al. [11] integrated LPC and MFCC features about DL models to enhance recognition performance.

2.5 Oversampling and Class Imbalance Handling

Class imbalance remains among the most significant challenges in infant cry classification. Underrepresented groups such as burping and belly pain often contain very few samples, causing models to transform into biased toward majority classes such as hungry or crying. In the interest of this issue, several oversampling techniques have been proposed. SMOTE [2] creates synthetic minority samples using interpolation between neighboring instances. Borderline SMOTE concentrates on generating samples near decision boundaries, whilst ADASYN adaptively creates synthetic data for harder to learn minority samples. Random Oversampling furthermore its extension RMO DA have also shown promising results in enhancing minority class recall. Lee et al. [13] demonstrated that SMOTE based balancing greatly increases DL performance in speech related audio tasks. A recent survey suggests that combining imbalance aware oversampling with attention based architectures boosts robustness and minority class recognition performance [14, 15].

2.6 Research Gaps Identified from Literature

Despite considerable advancements in the field of baby crying classifications, some issues still need to be studied further. Current literature focuses more on architecture-based research without comparing CNN, RNN, LSTM, and GRU under equal pre-processing conditions. Furthermore, many studies utilize different datasets, feature extraction techniques, augmentation methods, and test scenarios, making fair comparison difficult. A more big limitation is the lack of

systematic evaluation of multiple oversampling techniques with DL architectures. Even though augmentation and balancing methods have improved minority class recognition, one can still find limited understanding of how different balancing strategies affect various DL models. Besides, most datasets remain small, noisy, and highly imbalanced, reducing the generalizability and clinical applicability of existing rules. Transformer based and attention driven approaches have proved promising results, but their integration with imbalance aware frameworks remains inadequately discussed. Thus, this research aims to address these gaps by systematically evaluating CNN, RNN, LSTM, and GRU models comprising several oversampling techniques under a standardized preprocessing, augmentation, and evaluation pipeline in Table 2.1.

Table 2.1: Comparison of Existing Infant Cry Classification Techniques

Author	Method	Features Used	Model	Balancing Technique	Key Findings
Maghfira et al.	Hybrid DL	MFCC	CNN-RNN	None	Improved temporal-spectral learning
Özseven	DNN + Hand-crafted Features	LPC, MFCC	DNN	None	Better feature representation
Alam and Mamun	CNN+DNN Hybrid	MFCC + Augmentation	CNN+DNN	Data Augmentation	High classification accuracy
Gorin et al.	Self-supervised Learning	Spectrogram	SSL Model	None	Utilized unlabeled data effectively
Lee et al.	Audio Classification	MFCC	Deep Learning Models	SMOTE	Improved minority-class recall
Hasasneh et al.	Vision Transformer	Spectrogram	ViT	Attention Mechanism	Improved explainability and robustness
Proposed Work	Comparative Study	MFCC, ZCR, RMS, Chroma	CNN/RNN/LSTM/GRU	SMOTE, ADASYN, Borderline-SMOTE, RMO-DA	Fair comparison of deep learning and balancing techniques

Chapter 3

METHODOLOGY

The proposed methodology is the purpose of developing a robust infant cry classification framework using DL architectures and imbalance handling techniques. The process flow incorporates dataset collection, preprocessing, feature extraction, data augmentation, oversampling, DL model training, and performance measurement. Initially, raw infant cry recordings are filtered using normalization, silence trimming, mono conversion, and resampling techniques. Spectral and temporal auditory properties such as MFCC, ZCR, RMS, and Chroma are then extracted from the enhanced signals.

For making the system more robust and to balance out the classes, various methods for augmentation and oversampling are used, such as noise addition, pitch shift, time stretch, time shift, SpecAugment, SMOTE, ADASYN, Borderline-SMOTE, random oversampling, and RMO-DA. The final step includes training and evaluation of the CNN, RNN, LSTM, and GRU models using five-fold cross-validation. Figure 3.1 shows the entire flowchart of the proposed infant cry classification system.

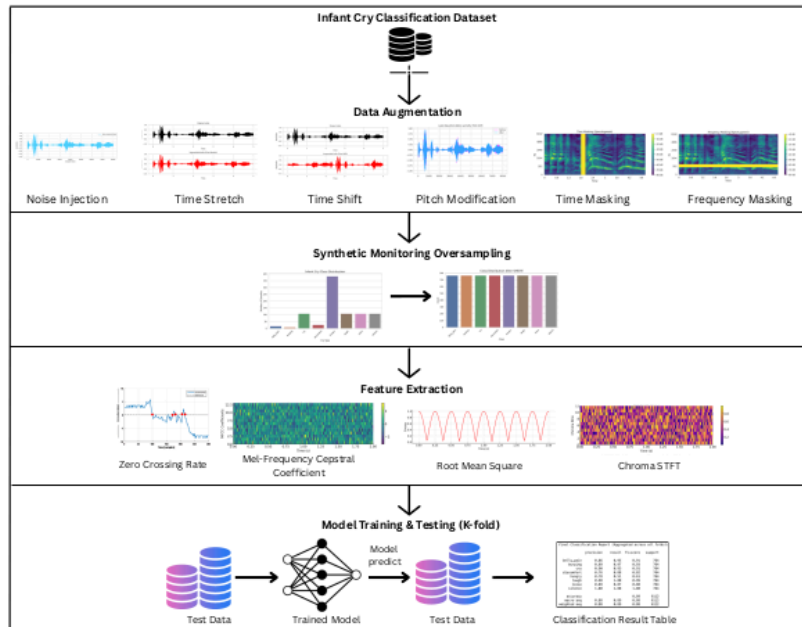


Figure 3.1: Proposed method for identifying baby cries

3.1 Dataset

The data set used in this study has been collected from two open Kaggle data sets: Donate a Cry Corpus [16] and Infant Cry Sounds data set [17]. In both of these data sets, infant cry sounds are recorded which are affected by various physical and environmental parameters.

The consolidated dataset consists of 873 audio samples distributed across eight different classes, specifically: *hungry* (383), *burping* (9), *belly_pain* (17), *discomfort* (28), *cry* (109), *laugh* (109), *noise* (109), and *silence* (109). There is significant class imbalance in this dataset, where some classes, for instance burping and *belly_pain*, have far fewer entries compared to the majority classes like hungry.

The sound recordings were all recorded in .wav format and vary in terms of recording environment, noise, and intensity of crying. This makes the problem of classifying infant cries difficult in practice.

Figure 3.2 shows the distribution of data classes before applying augmentation or over-sampling techniques. From the figure, we can see the imbalance among the classes, hence the need to balance the classes is imperative.

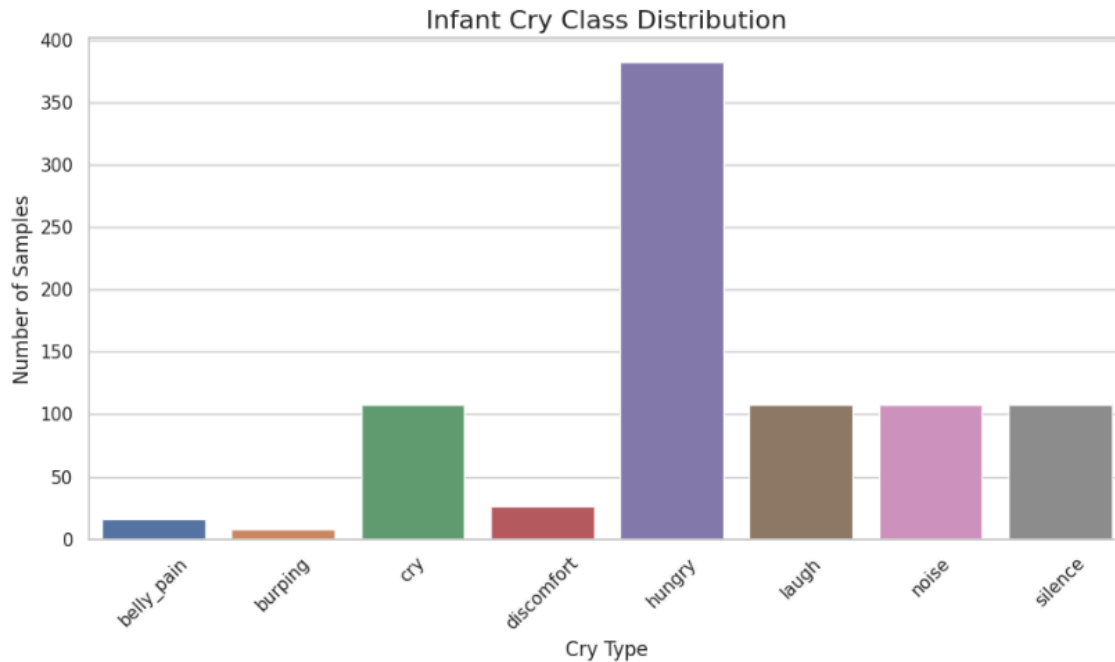


Figure 3.2: Class distribution of the infant cry dataset prior to SMOTE

3.2 Audio Preprocessing

Firstly, the audio dataset was preprocessed by standardization and mono channelization of the audio samples. Then, the audio samples were resampled at a rate of 16kHz. The silence trimming technique was applied to eliminate any silent parts in the audio samples.

After this, the audio samples were split into frames of 25ms, and the spectrogram of each frame was generated using STFT method and a window size of 512 FFT points.

31

3.3 Feature Extraction

Acoustic feature extraction is vital in order to represent the unique acoustic properties of baby crying sound signals. These acoustic features have been extracted from the spectrogram images which have been computed using the Python Librosa package.

The audio features that were extracted include: MFCC, ZCR, RMS, and Chroma Features. These are some of the key features that are needed to differentiate between different types of baby cries.

3.4 Data Augmentation

10

Data augmentation methods were used in order to improve the accuracy of the infant cry classification model. As there was a limited amount of balanced data available for infant cries, data augmentation became essential in order to avoid overfitting.

Librosa is a Python library used in this experiment due to its ability to perform numerous transformations on audio signals. The list of augmentation procedures used in this experiment is provided below:

1. **Noise Injection:** Random Gaussian noise was introduced into the audio signal to mimic environmental effects and improve noise resistance.
2. **Pitch Shifting:** The pitch of the synthesized audio baby cry was changed without changing its duration.
3. **Time Stretching:** The tempo of the audio signal was varied without altering its pitch to achieve greater temporal variability.
4. **Time Shifting:** The time shift operation was applied to the waveform to simulate different timing in recordings.

5. **Gain Variation:** The amplitude of the audio signal was raised or lowered to simulate intensity variations in recordings.

6. **SpecAugment:** Time masking and frequency masking were imposed directly to spectrogram representations in order to be more robust against spectral variations.

The aforementioned augmentation strategies increase the ability of the model to generalize under different recording environments and also benefit the minority classes. Figures 3.3, 3.4, 3.5, 3.6, and 3.7 provide example data using the above-mentioned augmentation

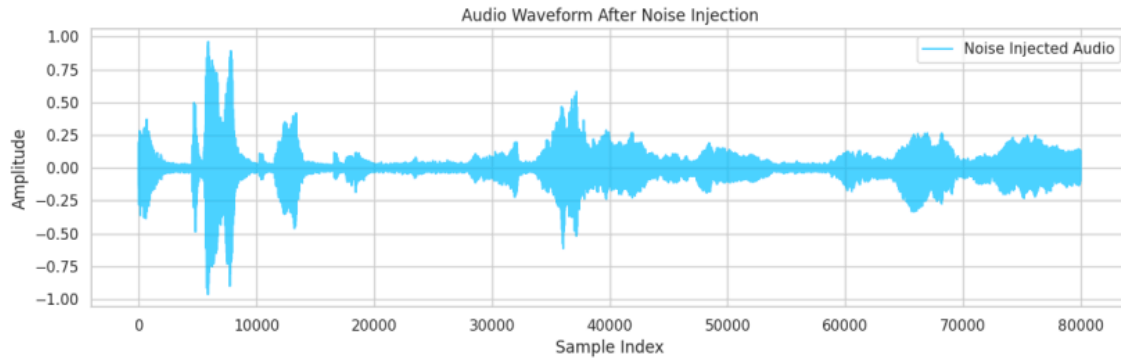


Figure 3.3: noise injection to audio of infant cries

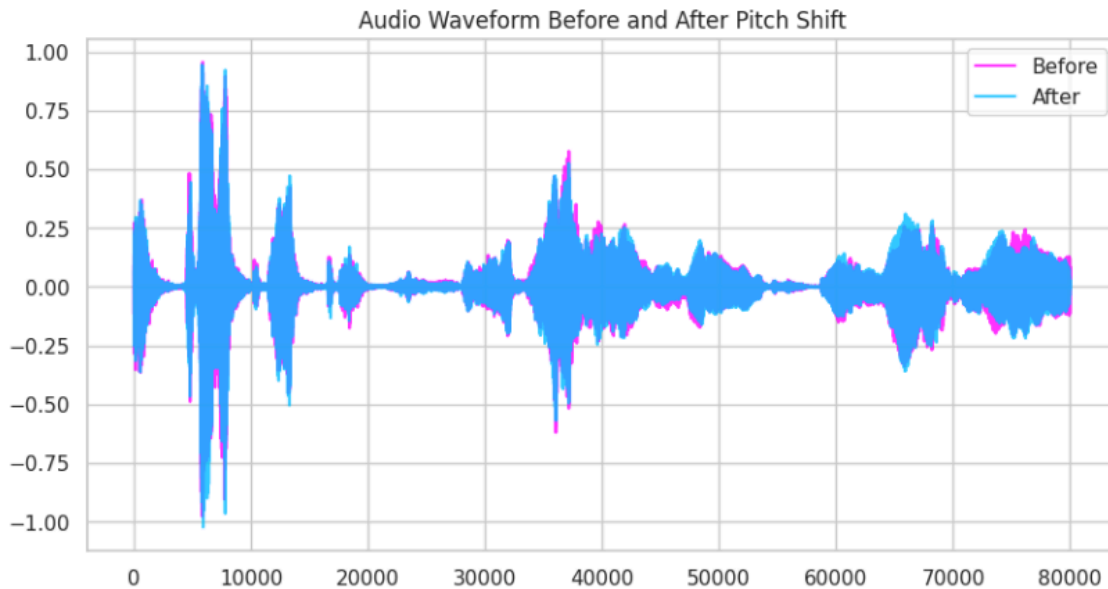


Figure 3.4: Pitch shift augmentation

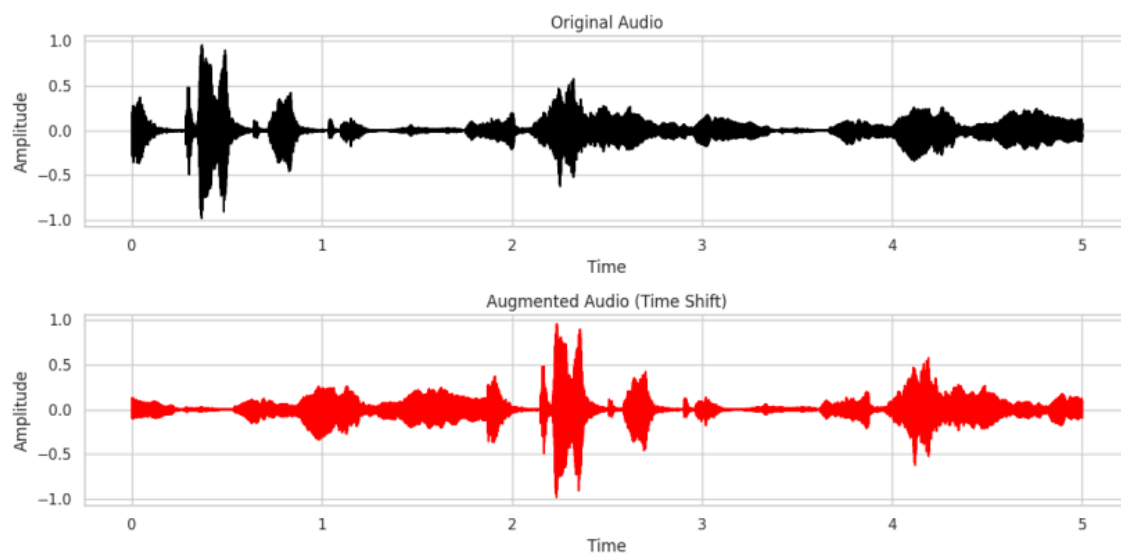


Figure 3.5: Time shift augmentation

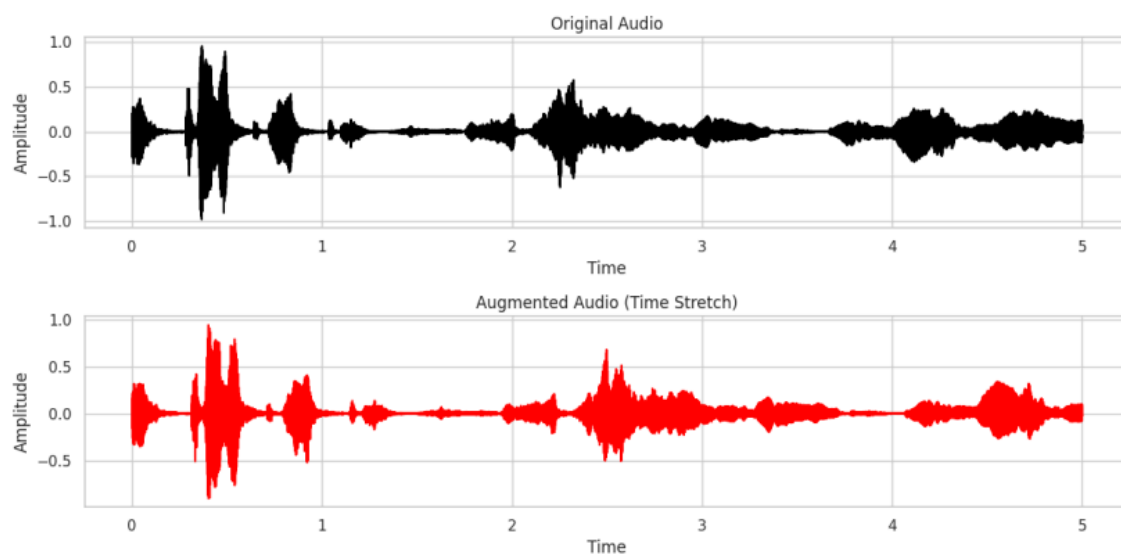


Figure 3.6: Augmentation through time stretching

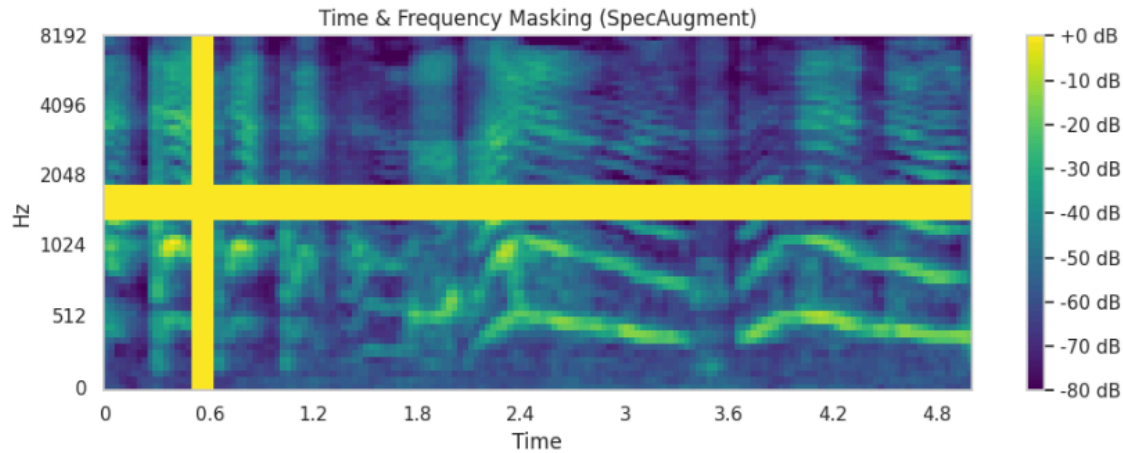


Figure 3.7: SpecAugment with time and frequency masking

3.5 Class Balancing Techniques

Class imbalance is among major challenges in infant cry classification, as certain cry categories hold significantly fewer samples compared to majority classes. The imbalance leads to deep learning being biased towards the majority class, hence reducing the classification of other minor cry types like burping and *belly_pain*. To overcome this challenge, several resampling approaches have been applied in this research paper.

These oversampling approaches have been applied through the `scikit-learn` and `imbalanced-learn` packages due to their accurate resampling algorithms. In the five-fold cross-validation process, all methods of balancing were performed on the training data only, without touching the test data at all, in order to avoid any data leakage.

Class balancing was achieved by using the following techniques:

1. **Random Oversampling:** The strategy ensures balanced data by replicating the data samples of minority classes randomly until all classes have an equal number of samples. Although easy to implement, too much replication can make the model susceptible to overfitting.
2. **SMOTE (Synthetic Minority Oversampling Technique):** The SMOTE technique creates artificial samples of minority classes using interpolation of minority class data samples (please refer to reference [2]). The method ensures class balancing without making the model prone to overfitting.
3. **Borderline SMOTE:** The method involves generating synthetic samples in the vicinity

of class boundaries. Since errors are more probable in such areas, the technique helps classify minority classes effectively.

4. **ADASYN (Adaptive Synthetic Sampling):** ADASYN adaptively constructs synthetic samples for difficult to learn minority instances. greater quantity synthetic data is created for samples that are harder to classify, thereby progressing minority class learning performance.
5. **RMO-DA (Random Oversampling with Data Augmentation):** The RMO-DA method involves Random Oversampling combined with audio-based augmentations such as pitch shifting, noise injection, and time stretching. The method achieves two objectives: increasing the sample size of the minority class and improving data diversity.

These oversampling techniques contribute to improving minority class recognition, reduce classification bias, and strengthen the overall robustness of the infant cry classification system. After addressing balancing methods, the dataset turn significantly more uniform across all classes.

Figure 3.8 illustrates the dataset distribution after applying balancing techniques.

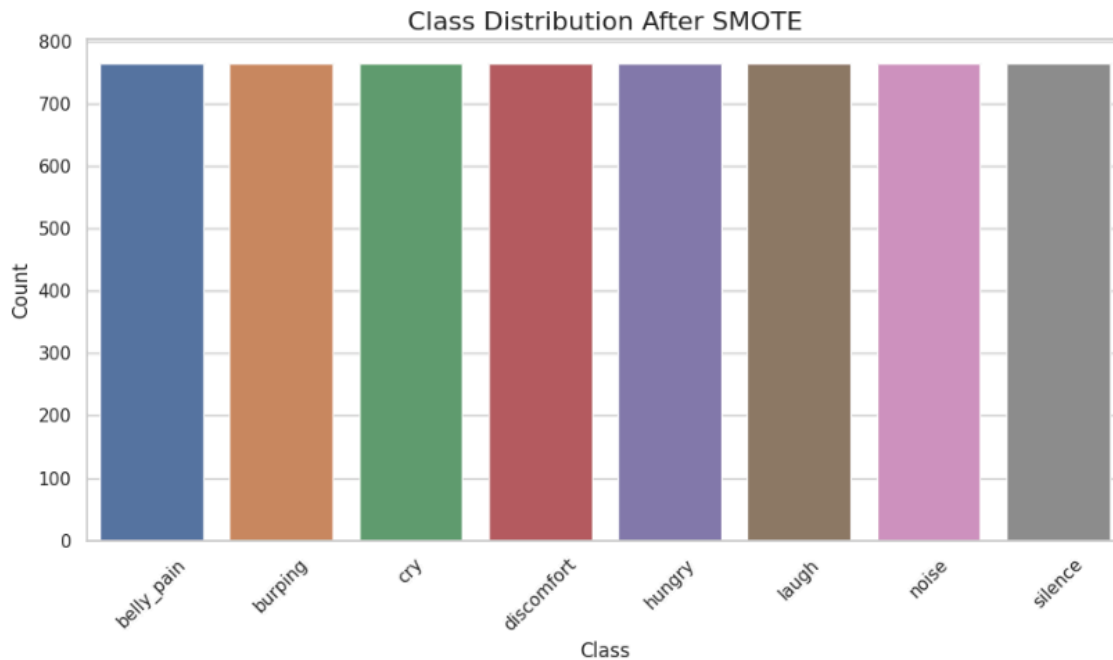


Figure 3.8: Distribution of balanced dataset post-SMOTE

3.6 Deep Learning Models

Four distinct DL architectures were implemented and evaluated in this study for infant cry classification. These models were chosen because of their strong capability in learning temporal and spectral representations with audio data. All models were trained under identical preprocessing, augmentation, balancing, and evaluation setups to ensure fair comparison.

After feature extraction, the class labels were converted into integer values using `LabelEncoder` and additionally transformed into one hot encoded vector for multi class category. The extracted feature vectors were normalized and reshaped into three dimensional tensors of magnitude $(n_{\text{samples}}, n_{\text{features}}, 1)$ to make them compatible through DL architectures.

3.6.1 Convolutional Neural Network

CNN have been extensively used in the application of audio and image classification owing to their ability to learn spatial and spectral features independently from the representations such as spectrogram. For the current experiment, the architecture of the CNN involved three convolutional layers consisting of 32, 64, and 128 filters. After each convolution process, max pooling layers were added to minimize the dimensions and reduce complexity.

The final outputs from the convolutions were flattened before being connected to dense layers. The final dense layer was activated by the softmax layer for the purpose of multi-classification.

3.6.2 Recurrent Neural Network

RNN models are designed specially to work with sequential data and are able to discover temporal dependencies in time series data. As cry signal samples have some kind of temporal acoustic variation, we used RNN to get sequential data from our feature set.

In the current project, we applied the following configuration for RNN: two recurrent layers with 128 hidden units and a dropout rate of 0.3. Finally, we use the SoftMax activation function for the output layer.

3.6.3 Long Short-Term Memory

LSTM is an advanced variant of RNN intended to overcome the vanishing gradient problem faced in traditional recurrent networks. LSTM networks utilize memory cells and gating mechanisms in order to stay in long-term temporal dependencies within sequential data.

In this experiment, an LSTM model was utilized with two LSTM layers, where both layers had 128 hidden units. Furthermore, a dropout ratio of 0.3 was implemented in the model to enhance its learning capabilities. This model was capable of recognizing temporal patterns within infant cry sounds with improved recall ability for minor classes.

3.6.4 ³ Gated Recurrent Unit

GRU is yet another type of recurrent neural network model that offers a simplified structure of the LSTM model without losing the ability to learn temporal relationships. GRU uses fewer parameters compared to LSTM, making it possible to compute quickly and consume less memory.

In this research, the proposed model included two GRU layers with 128 units and a dropout of 0.3. Even though the model had effective computations and converged faster than LSTM, its classification accuracy was relatively poorer compared to CNN and RNN models for the task of infant cry detection.

In this experiment, models were evaluated using five-fold cross-validation. In every fold, there were four folds used for training, and one fold used ²² for testing. Accuracy, Precision, Recall, F1 score, and confusion matrices were used for evaluating the proposed models.

3.7 Model Training and Validation

All models were trained using five-fold cross-validation in order to provide an unbiased evaluation of their performance. Each iteration included four datasets for training and one dataset for validation. All experiments were run on Google Colab with Tensorflow and Keras using GPU.

¹⁹ The learning rate was set to 0.001 and Adam optimizer was used as an optimizer while ¹⁷ categorical cross-entropy was applied as a loss function. Batch size was set to 32 and each experiment consisted of 50 epochs.

3.8 ⁷ Evaluation Metrics

The performance of the proposed models ⁸ estimated using Accuracy, Precision, Recall, and F1-score. Confusion matrices and classification reports too generated to analyze class wise performance, particularly for minority classes.

Chapter 4

EXPERIMENTAL SETUP AND RESULTS

All experimental tasks were executed in Google Colab with GPU support enabled. The proposed deep learning approaches were implemented using Python 3.8, TensorFlow 2.19.0, and Keras 3.10.0. A consistent and unbiased comparison between the considered models was provided by applying the same data preparation and transformation processes in all experimental tasks.

To ensure the reproducibility of the experiments, the random number generators were seeded with the number 42 using Python's random library, NumPy, TensorFlow, and imbalanced learn libraries. All experimental tasks were based on fivefold cross-validation, in which 80% of the data were used for training and 20% for validation in each fold.

The hyperparameter optimization was performed by Keras Tuner. All models were trained with the Adam optimizer and categorical cross-entropy loss, using a learning rate of 0.001. The chosen batch size was 32, while 50 training epochs were defined for achieving a good balance between training speed and overfitting risk. Data augmentation and oversampling were applied only to the training data in each fold.

4.1 Evaluation Metrics

The proposed models were evaluated based on the Accuracy, Precision, Recall, and F1-score metrics. The choice of such metrics was made with the aim of evaluating both minority and majority classes in an adequate way.

- **Accuracy:** measures the overall accuracy of predictions.
- **Precision:** indicates the proportion of correct positive predictions from all positive predictions made.
- **Recall:** Measures the capability of the model to properly identify actual positive samples.
- **F1-score:** The harmonic mean of precision and recall; included because it addresses class imbalance.

Confusion matrices and classification reports were created in order to measure the performance of the classifier for each class separately.

4.2 Performance Comparison of Deep Learning Models

A comparison of the performances of the CNN, RNN, LSTM, and GRU after applying the class balancing techniques is shown in Table 4.1.

Table 4.1: Comparison of performance among CNN, RNN, LSTM, and GRU

Model	Accuracy	Precision	Recall	F1-score
CNN	0.93	0.93	0.93	0.93
RNN	0.89	0.88	0.89	0.88
LSTM	0.81	0.80	0.81	0.79
GRU	0.71	0.69	0.71	0.68

From the results, the CNN model performed best with 93% accuracy. The superior performance of the CNN can be attributed to its effectiveness in feature extraction from the cry spectrograms of infants.

RNN accomplished the second highest performance while demonstrating balanced recall for minority classes because of its sequential learning capability. LSTM improved minority class sensitivity but demonstrated slightly reduced precision because of increased false positives. GRU displayed comparatively lower classification accuracy, indicating limited capability in handling complex temporal variations existing in infant cry signals.

4.3 Oversampling Analysis

As to measure the influence of imbalance handling strategies, multiple oversampling techniques as well as Random Oversampling, SMOTE, Borderline-SMOTE, ADASYN, and RMO-DA were compared with all DL models. The comparative performance achieved by the use of various oversampling methods is presented in Table 4.2.

The results prove that oversampling significantly improved minority class recognition performance. RMO DA and Borderline SMOTE constantly achieved better performance in comparison with other balancing techniques. Sequential architectures such as RNN and LSTM specifically benefit from imbalance aware augmentation methods.

Table 4.2: Performance of models under different oversampling techniques

Model	ADASYN				Borderline-SMOTE				RMO-DA				Random Oversampling				SMOTE			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
RNN	0.93	0.93	0.93	0.92	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.95	0.89	0.88	0.89	0.88	0.93	0.93	0.93	0.93
LSTM	0.74	0.71	0.74	0.71	0.74	0.72	0.74	0.71	0.87	0.87	0.88	0.87	0.78	0.77	0.78	0.76	0.87	0.86	0.87	0.86
GRU	0.82	0.81	0.82	0.81	0.89	0.89	0.89	0.89	0.92	0.92	0.92	0.91	0.95	0.95	0.95	0.94	0.89	0.88	0.89	0.88
CNN	0.92	0.92	0.92	0.91	0.86	0.87	0.87	0.87	0.86	0.87	0.87	0.87	0.78	0.76	0.78	0.76	0.81	0.80	0.81	0.79

4.4 Confusion Matrix Analysis

Confusion matrices had been generated for all implemented models to analyze class wise prediction behavior and incorrect classification trends as illustrated in Figure 4.1.

CNN and RNN displayed stronger diagonal dominance, indicating accurate prediction for most infant cry classifications. LSTM improved minority class recall with fewer false negatives for classes for instance *belly pain* and *burping*. GRU showed comparatively weaker class distinctions and higher confusion among acoustically similar classes.

4.5 Statistical Validation

In order to examine the efficiency of oversampling techniques, Wilcoxon signed-rank test has been conducted on the oversampled data versus non-oversampled data. The results of statistical validation are shown in Table 4.3.

Table 4.3: Wilcoxon signed-rank test results

Metric	Wilcoxon Statistic	p-value
Accuracy	0.0	0.125
Precision	0.0	0.125
Recall	0.0	0.125
F1-score	0.0	0.125

In spite of the fact that oversampling improved average model performance, the obtained p values surpass 0.05, indicating that the improvements were not statistically significant. This indicates that larger and more diverse datasets are required for stronger statistical validation.

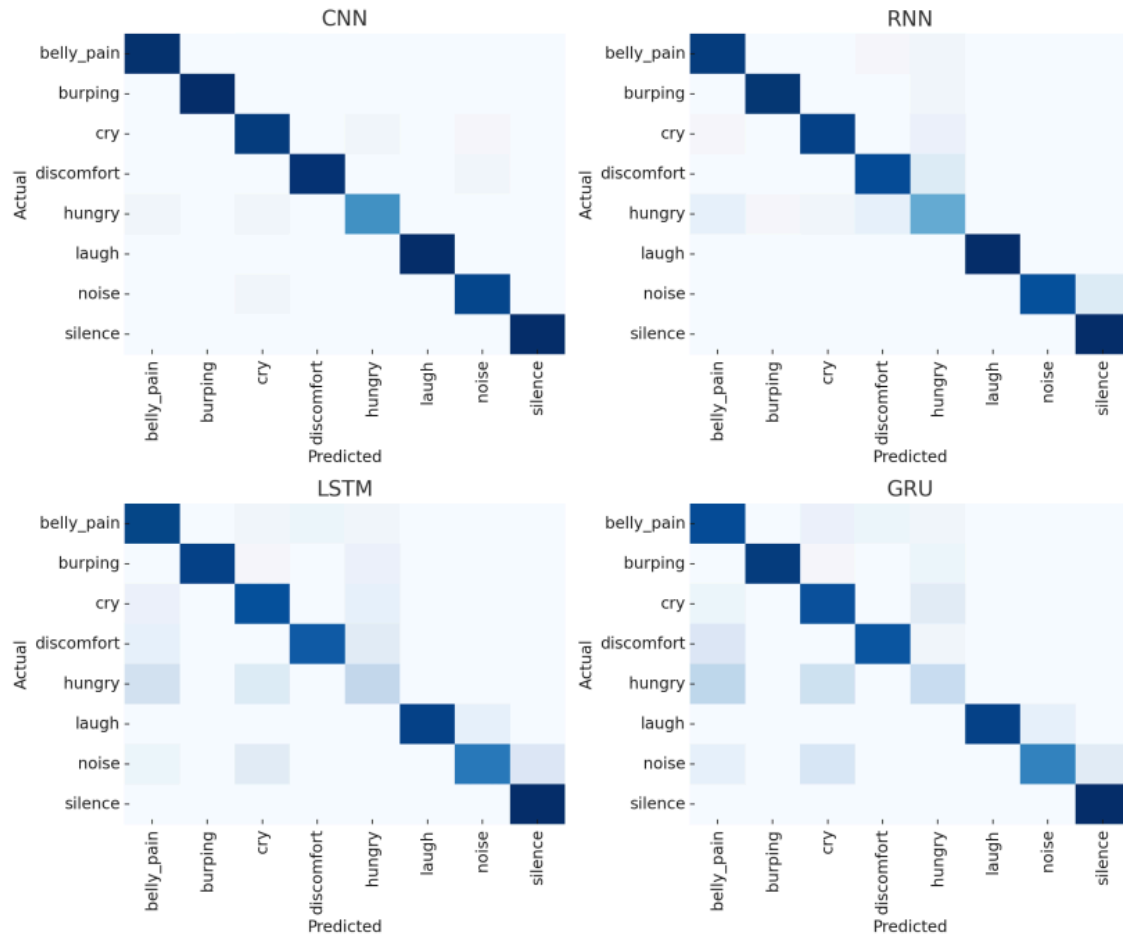


Figure 4.1: Confusion matrix for CNN, RNN, LSTM, and GRU models

4.6 Discussion

The comparative analysis exhibits that DL architectures behave differently under class imbalance conditions. CNN accomplished the highest overall classification accuracy because of its functions to efficiently capture local spectral patterns from infant cry spectrograms.

Sequential architectures for instance RNN and LSTM improved minority class recall due to their expertise in modeling temporal dependencies. However, such an improvement is accompanied by higher computation complexity as well as a slight decrease in accuracy.

Between oversampling methods, RMO-DA and Borderline-SMOTE provided better results as both helped achieve class balancing and maintain real data variability. Random Oversampling posed an increased danger of overfitting because of the replication of the minority class data.

The study reveals that the use of DL algorithms in combination with methods for addressing data imbalance is expected to improve the results of classification of infants' cries. To study the effect of imbalance handling, a comparative analysis was carried out to test the performance of the model with and without oversampling. The aim of such a comparison was to investigate if oversampling affects the accuracy of the model and its ability to detect the minority classes. The comparative analysis results are summarized in Table 4.4.

Table 4.4: Model performance with and without oversampling

Model	With Oversampling				Without Oversampling			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
RNN	0.93	0.93	0.93	0.93	0.81	0.76	0.81	0.78
LSTM	0.80	0.79	0.80	0.78	0.73	0.68	0.73	0.69
GRU	0.89	0.89	0.89	0.89	0.81	0.77	0.81	0.78
CNN	0.85	0.84	0.85	0.84	0.83	0.78	0.83	0.80
Average	0.87	0.86	0.87	0.86	0.80	0.75	0.80	0.76

Based on the results, it is evident that the application of oversampling methods contributed to an increase in performance for all three neural network models. The biggest increase was observed in the case of the RNN model, whose accuracy rate increased by 20% after oversampling from 81% to 93%. Similar results were observed in the GRU and LSTM models.

From the accomplished goals, it is clear that imbalance-aware learning significantly enhances the ability of the classifier to identify the minority classes and minimize any bias towards the dominant cry types. The CNNs were not as effective compared to other deep learning

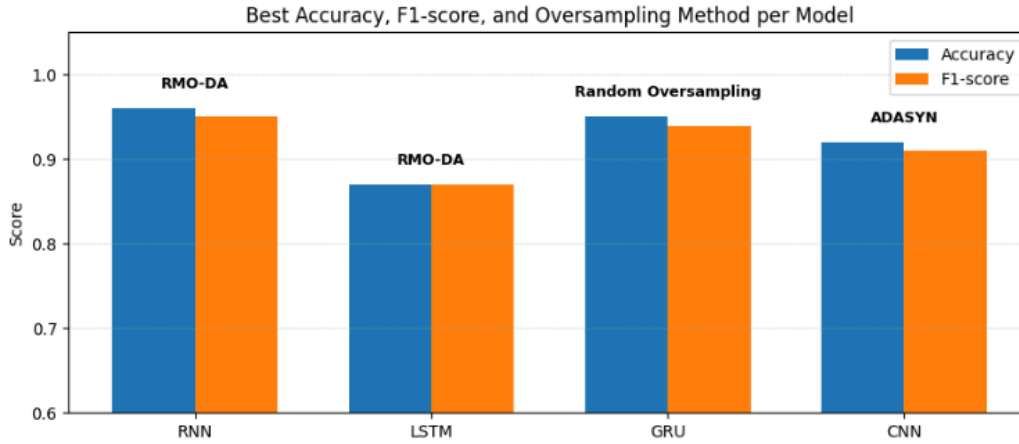


Figure 4.2: Best Accuracy and F1-score per model with corresponding optimal oversampling method.

approaches because they had a relatively good capability of extracting spectral features even under imbalanced scenarios.

4.7 Limitations

However, despite producing encouraging results, the suggested model suffers from various limitations. For instance, the data set is quite small and unbalanced. Moreover, the signals used for analysis have been acquired in an uncontrolled environment, complete with ambient noise.

Even though the use of oversampling methods improved the classification of the minority classes, the synthetic examples might not fully represent the actual cry variations. In addition, since the model lacks clinical validation, its implementation in practical applications is limited.

4.8 Societal Impact

This work has the potential to early detect and monitor infant distress, which supports clinicians and caregivers to timely alert and respond so as to improve their outcomes. Practical deployment leads to ethical and societal considerations, such as automated cry classifiers can produce false positives/negatives that can cause unnecessary alarm or misclassifications, and the models trained on limited datasets, which lead to risk-biased performance across populations and recording conditions. Ethical use of the dataset, such as privacy and consent, is most essential when collecting or deploying audio, where recordings must be stored and transmitted securely and parental consent is required for clinical use of the audio. To mitigate risks, we

recommend (i) continued validation on large, diverse, clinically valid datasets, (ii) ‘human-in-the-loop’ deployments, where automated outputs augment replacing caregiver judgment, and (iii) strict privacy and data security practices in real-world systems.

Chapter 5

CONCLUSION AND FUTURE WORK

In this research, the classification performance of four types of deep neural networks, namely, CNN, RNN, LSTM, and GRU, for infant cries is compared based on different class balancing techniques. The key problem tackled by this research in infant-cry analysis is class imbalance, where different oversampling methods like random oversampling, SMOTE, Borderline-SMOTE, ADASYN, and RMO-DA are used for enhancing the classification accuracy of minority classes such as burping and belly pain in infant cries.

From experimentation, it is evident that oversampling methods improve the classification ability of minority classes. Among all the deep networks tested in this experiment, the CNN network produced the best classification accuracy because of its superior feature extraction capacity from infant-cry spectra. The sequential models like the RNN and LSTM performed better when it came to minority class recall because of their ability to detect temporal relationships in the cry signal. The RNN was the best-performing model when paired with the rhythmic model optimization data augmentation (RMO-DA) approach, while the GRU was less accurate in classification but consumed fewer resources.

The experiment further highlighted the superiority of imbalance-aware data augmentation approaches such as RMO DA and Borderline SMOTE compared to conventional oversampling methods in terms of broadening the data scope and improving boundary learning. While there were clear performance improvements from oversampling through the Wilcoxon signed-rank test, the p-value results did not achieve significance because of the small sample size used. This highlights the importance of combining appropriate deep learning algorithms with proper balancing techniques for cry recognition system development.

The above framework has several possible applications in various fields including health care, newborn monitoring system, and intelligent caregiving environment. Automatic analysis of infant crying can help both the caregiver and clinician in the identification of infant distress and in continuous monitoring of infants.

5.1 Future Work

Further analysis can work towards collecting bigger and more diverse infant cry datasets, which are also clinically validated, thereby reducing the dependency on artificial data samples for training. Complex architectures, such as transformer and attention, can be helpful in learning time-dependent features. Pretrained audio models like Wav2Vec, YAMNet, and Whisper using transfer learning techniques might be more beneficial.

Moreover, multimodal approaches that combine audio, physiological, and visual modalities could help improve the accuracy of diagnoses and create more precise monitoring systems for infants. The use of lightweight solutions, designed to be implemented on mobile and IoT platforms, would make it possible to develop infant cry monitoring solutions in real-time. Lastly, explainable AI technologies combined with clinical validation will be essential for developing such applications.

REFERENCES

- [1] N. Maghfira, T. Basaruddin and A. Krisnadhi, “Infant cry classification using CNN-RNN,” *Journal of Physics: Conference Series*, vol. 1528, pp. 012019, 2020.
- [2] T. Özseven, “Infant cry classification by using different deep neural network models and hand-crafted features,” *Biomedical Signal Processing and Control*, vol. 83, 2023.
- [3] K. Alam and K. A. Mamun, “From cries to answers: A comprehensive CNN+DNN hybrid model for infant cry classification with enhanced data augmentation and feature extraction,” in *Proc. IEEE Int. Conf. Advanced Computing, Communication and Systems (iCACCESS)*, 2024.
- [4] A. Gorin, C. Subakan, S. Abdoli, J. Wang, S. Latremouille and C. Onu, “Self-supervised learning for infant cry analysis,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing Workshops (ICASSPW)*, 2023, pp. 1–5.
- [5] G. Anjali, S. Sanjeev, A. Mounika, G. Suhas, G. P. Reddy and Y. Kshiraja, “Infant cry classification using transfer learning,” in *Proc. IEEE Region 10 Conf. (TENCON)*, 2022, pp. 1–7.
- [6] Y. Zhang, M. Liu and H. Zheng, “Deep Learning-Based Infant Cry Emotion Recognition Using Spectrogram Fusion and Attention Mechanisms,” *IEEE Access*, vol. 12, pp. 14567–14579, 2024.
- [7] S. Kamble and R. Thakur, “CryFormer: A Transformer-Based Framework for Robust Infant Cry Classification Under Noisy Conditions,” *IEEE Transactions on Affective Computing*, early access, 2025.
- [8] T. Ko, V. Peddinti, D. Povey and S. Khudanpur, “Audio augmentation for speech recognition,” in *Proc. Interspeech*, 2015, pp. 3586–3589.
- [9] D. S. Park et al., “SpecAugment: A simple data augmentation method for automatic speech recognition,” in *Proc. Interspeech*, 2019, pp. 2613–2617.
- [10] N. Meephiw and P. Leesutthipornchai, “MFCC feature selection for infant cry classification,” in *Proc. Int. Conf. Science, Engineering and Communications (ICSEC)*, 2022, pp. 123–127.
- [11] Z. Firas, A. A. Nashaat and G. Ahmad, “Optimizing infant cry recognition: A fusion of LPC and MFCC features in deep learning models,” in *Proc. Int. Conf. Advances in Biomedical Engineering (ICABME)*, 2023, pp. 1–6.
- [12] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Oversampling Technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

- [13] J. N. Lee, H. Kim and S. Park, “An Efficient SMOTE-Based Deep Learning Model for Voice Pathology Detection,” *Applied Sciences*, vol. 13, no. 6, pp. 3571, 2023. doi:10.3390/app13063571
- [14] H. Lin and K. Chen, “Boundary-Aware Synthetic Oversampling for Imbalanced Deep Learning,” *Pattern Recognition*, vol. 144, 109842, 2023.
- [15] A. Romero, L. Cuadra and M. Nieto, “Imbalance-Aware Deep Learning for Audio Classification Using Adaptive Synthetic Sampling and Attention Networks,” *Neural Computing and Applications*, vol. 36, pp. 10231–10245, 2024.
- [16] B. Valani, “Donate-a-cry-corpus-features-dataset,” Kaggle, 2020. [Online]. Available: <https://www.kaggle.com/datasets/balraj98/donateacry>
- [17] R. Jahangir, “Infant cry sounds,” Kaggle, 2020. [Online]. Available: <https://www.kaggle.com/datasets/raufjahangir/infant-cry-sounds>

ORIGINALITY REPORT

8%

SIMILARITY INDEX

6%

INTERNET SOURCES

6%

PUBLICATIONS

2%

STUDENT PAPERS

PRIMARY SOURCES

- 1 Arvind Dagur, Sohit Agarwal, Dharendra Kumar Shukla, Shabir Ali, Sandhya Sharma. "Artificial Intelligence and Sustainable Innovation - Volume 1", CRC Press, 2026
Publication
- 2 www.corpuspublishers.com
Internet Source
- 3 schneppat.com
Internet Source
- 4 export.arxiv.org
Internet Source
- 5 nepjol.info
Internet Source
- 6 spj.science.org
Internet Source
- 7 B. Satheesh, Kumaresan Angappan. "TriSense-AttentionNet: A CNN-LSTM With Multi-Head Attention Fusion Model for Infant Cry Classification", IEEE Access, 2026
Publication
- 8 www.ijirset.com
Internet Source
- 9 ia601407.us.archive.org
Internet Source
- 10 jutif.if.unsoed.ac.id
Internet Source
- 11 Mohit Dua, Nidhi Chakravarty, Shelza Dua. "Audio Spoof Detection from Theory to Practical Application", CRC Press, 2026
Publication
- 12 Submitted to Sheffield Hallam University
Student Paper
- 13 bthek1.github.io
Internet Source
- 14 Submitted to Nottingham Trent University
Student Paper

15 jamh.uomosul.edu.iq
Internet Source

16 Submitted to Anna University
Student Paper

17 Michele Rossi, Giovanni Iacca, Luca Turchet. "Advancing guitar emotion recognition through audio data augmentation to enhance smart musical instruments", EURASIP Journal on Audio, Speech, and Music Processing, 2025
Publication

18 docplayer.net
Internet Source

19 github.com
Internet Source

20 csis.pace.edu
Internet Source

21 ir-library.ku.ac.ke
Internet Source

22 pci.cyber.az
Internet Source

23 www.vniitti.ru
Internet Source

24 assets-eu.researchsquare.com
Internet Source

25 downloads.hindawi.com
Internet Source

26 saucis.sakarya.edu.tr
Internet Source

27 www.themoonlight.io
Internet Source

28
Publication

29 Georgios Douzas, Fernando Bacao. "Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning", Expert Systems with Applications, 2017
Publication

30 Muhammad Hassan Jamal, Naila Naz, Muazzam A. Khan Khattak, Faisal Saeed, Saad Nasser Altamimi, Sultan Noman Qasem. "A Comparison of Re-Sampling Techniques for Detection of Multi-Step Attacks on Deep Learning Models", IEEE Access, 2023

- 31 R. Sabitha, P. Poonkodi, M. S. Kavitha, S. Karthik. "Premature Infant Cry Classification via Deep Convolutional Recurrent Neural Network Based on Multi-class Features", Circuits Systems, and Signal Processing, 2023

Publication

- 32 Taghi M. Khoshgoftaar, Naeem Seliya, Dennis J. Drown. "Evolutionary data analysis for 1 class imbalance problem", Intelligent Data Analysis, 2010

Publication

- 33 Wu, Xingfang. "Representation Learning for Enhanced Software Analytics.", Ecole Polytechnique, Montreal (Canada)

Publication

- 34 Wurzel, Karen K.. "Graph Neural Networks Predict Anti-Epileptic Treatment Response in Non-Lesional Infantile Epileptic Spasm Syndrome.", Marquette University

Publication

- 35 dx.doi.org

Internet Source

- 36 new.esp.org

Internet Source

- 37 openaccess.altinbas.edu.tr

Internet Source

- 38 scholar.archive.org

Internet Source

- 39 www.ijraset.com

Internet Source

- 40 Sarah Jane Kho, Brian Loh Chung Shi, Vong Wan-Tze, Law Kian Boon et al. "Malaysian cough sound analysis and COVID-19 classification with deep learning", Intelligence-Based Medicine, 2023

Publication

Exclude quotes

On

Exclude matches

Off

Exclude bibliography

On