Improving Infant Cry Recognition with CNNs and Imbalance Mitigation

Michael Indrawan*, Ardytha Luthfiarta, Muhammad Daffa Al Fahrzea, Muhammad Rafid

Abstract—The classification of baby cries using machine learning is essential for developing automated systems that can assist caregivers in identifying and responding to the needs of infants promptly and accurately. This study aims to improve upon previous research relating to the Cry Baby Dataset, which has highly imbalanced data. We combine oversampling and undersampling techniques using SMOTE and ENN, along with data augmentation through pitch shifting and noise addition to address the data imbalance issue. The processed data was then modeled using Convolutional Neural Networks (CNN). The study yielded an overall accuracy of 88%, with balanced accuracy observed across all classes, effectively mitigating data imbalance. This represents a notable advancement compared to previous research, which often encountered challenges with unbalanced accuracies across classes. The classes identified include recordings of baby cries attributed to belly pain caused by colic, recordings related to burping, recordings associated with discomfort or other symptoms, recordings of hungry baby cries, and recordings indicating fatigue or the need for sleep. This shows a significant improvement from previous studies, which had very unbalanced accuracy for each class.

Keywords: Baby Cry Classification; Neural Network; Handling Data Imbalance; Audio Analysis

1. INTRODUCTION

The intricate dynamics of infant communication, particularly through crying, have long been recognized as a foundational element in establishing a secure and nurturing bond between newborns and their caregivers[1]. This primal form of expression becomes a crucial avenue through which infants convey their needs, discomfort, or distress, necessitating a nuanced understanding from caregivers to appropriately respond. Given the multifaceted nature of infant cries and the necessity for prompt and accurate response, the classification of baby cries using machine learning is essential for developing automated systems that can assist caregivers in identifying and responding to the needs of infants promptly and accurately. This comprehensive approach to classification and interpretation is vital, prompting various studies to explore diverse methodologies.

An exemplary study by Dewi et al. delved into the application of Linear Predictive Coding (LPC) and Euclidean Distance methods for identifying and distinguishing between different types of baby cries[2]. While their results were promising, it is imperative to acknowledge the inherent limitations within these methodologies. Building on this foundation, Prayogi et al. undertook an investigation that incorporated prosody features, Moments of Distribution, and K-Nearest Neighbors methods to discern and categorize infant cries[3]. Despite their endeavors, the attained accuracy revealed ongoing challenges in achieving a satisfactory level of precision.

Concurrently, Yusdiantoro et al. contributed valuable insights by achieving commendable accuracy in identifying baby cries; however, a lingering imbalance in accuracy metrics across distinct classes surfaced, raising concerns regarding the practical utility of the findings[4]. These collective observations and challenges set the stage for the current research, which endeavors to transcend the limitations of previous methodologies in identifying and categorizing infant cries.

Central to this research approach is the integration of Convolutional Neural Networks (CNN)[5], [6], [7], recognized for their proficiency in discerning intricate patterns within diverse datasets. The adaptability of CNNs to both image and audio data positions them as a promising solution for the nuanced task of identifying and categorizing the nuanced acoustic features of infant cries. However, acknowledging the pervasive issue of data imbalance, this study employs a comprehensive strategy that encompasses oversampling and undersampling techniques.

Synthetic Minority Over Sampling Technique (SMOTE)[8], [9] is employed to address underrepresented classes, ensuring a more robust representation of diverse cry types. Simultaneously, Edited Nearest Neighbors (ENN)[10], [11] is utilized to mitigate the impact of overrepresented classes, fostering a more balanced and equitable dataset. This meticulous data balancing strategy is poised to enhance the efficacy of the CNN model in accurately classifying infant cries.

In the pursuit of addressing data imbalances within the research framework, a paramount emphasis is placed on the integration of advanced data augmentation techniques, representing a formidable stride towards methodological refinement. These sophisticated strategies include, but are not limited to, the strategic implementation of pitch shifting and noise addition, which function as pivotal mechanisms strategically deployed to introduce a heightened degree of variability into the dataset.

The deliberate injection of variability stands as a linchpin in fortifying the model's robustness, cultivating an environment where the model is not only adept at capturing diverse patterns within the data but also exhibits a
heightened capacity for generalized performance. The augmentation strategies, therefore, transcend the mere quest for overall accuracy improvement; they are crafted with the explicit intent of achieving a more equitable and balanced accuracy distribution across the distinct classes of infant cries.

The overarching objective of this research extends beyond the conventional bounds of algorithmic enhancements, representing an ambitious endeavor to elevate the precision in the nuanced identification and categorization of baby cries. Simultaneously, the research is unwaveringly committed to ensuring a more nuanced and balanced distribution of accuracy, acknowledging the inherent complexity and diversity within the spectrum of infant vocalizations. This multifaceted approach holds immense promise in furnishing caregivers and medical professionals with an advanced and nuanced toolkit, empowering them to comprehend and respond to the intricate array of needs expressed through infant cries.

In summation, this research epitomizes a comprehensive and intricate undertaking, dedicated not only to advancing the accuracy of methodologies for identifying and categorizing baby cries but also to pushing the boundaries of practical applicability. By seamlessly integrating cutting-edge technologies, notably Convolutional Neural Networks (CNNs), with state-of-the-art data balancing and augmentation techniques, the study not only contributes to the academic knowledge base but also seeks to exert a tangible and transformative influence on the practical landscape of infant care. The potential implications of this research extend far beyond the confines of academia, permeating into the tangible and consequential realm of caregiving practices and infant well-being, thereby ushering in a new epoch characterized by precision and sophistication in early childcare methodologies.

2. RESEARCH METHODOLOGY

In this chapter, we will describe in detail the methods used in this research to overcome the challenges of identifying baby cries. The method is divided into three main sections, namely Section 2.1 describes dataset, Section 2.2 related to data preprocessing, Section 2.3 describes the convolution model used, and Section 2.4 discusses the model training method.

Figure 1. Research Stage Flowchart

The research stage is a comprehensive process that initiates with the meticulous preparation of the dataset. Subsequently, the focus shifts towards addressing imbalances within the dataset through a strategic combination of ENN (Edited Nearest Neighbors) and SMOTE (Synthetic Minority Over-sampling Technique). This involves an initial application of ENN, followed seamlessly by the integration of SMOTE techniques. A pivotal component of this stage involves dataset augmentation, enriching the dataset with additional instances.

Following the pre-processing steps, the dataset undergoes a crucial phase of division into distinct training and validation subsets. The training subset is then channeled into an intensive modeling session, where the algorithm learns and adapts to the intricacies of the data. This phase is marked by iterative refinement and optimization to enhance the model's predictive capabilities.

The culminating phase of the research encompasses the evaluation of the model's performance, a critical assessment conducted using the validation dataset. This evaluation session scrutinizes the model's accuracy, precision, recall, and other pertinent metrics, ensuring a comprehensive understanding of its efficacy and robustness. The seamless flow from dataset preparation to model evaluation underscores the systematic and
rigorous nature of the research process. The entire research process, from dataset preparation, addressing imbalances, and model evaluation, as described, is depicted in Figure 1.

2.1 Dataset

In this experiment, we used the Baby Cry Detection dataset curated by Priscilla Dunstan[12]. This dataset serves as the main data source for developing the baby cry identification model and consists of a diverse collection of audio recordings of baby cries categorized into five classification categories based on their causes. These five categories include:

a. Belly Pain: This category contains 16 audio recordings of baby crying caused by colic.
b. Burping: There are 8 audio recordings of baby cries related to the burping process.
c. Discomfort: This category includes 27 audio recordings of baby cries caused by discomfort or other symptoms.
d. Hungry: This is the largest category, with 382 audio recordings of hungry baby cries.
e. Tired: There are 24 audio recordings of baby cries related to fatigue or the desire to sleep.

It is imperative to underscore the salient observation that the dataset we are working with is characterized by a marked incongruity in the distribution of classes, manifesting as a significant imbalance in terms of sample representation across the diverse categories. The preponderance of instances is conspicuously concentrated within the confines of the “Hungry” category, while a notable dearth of samples is observed in several other categories. This pronounced class imbalance introduces a formidable challenge that demands nuanced and strategic interventions in order to cultivate a balanced and unbiased environment for the successful identification of infant cries.

Navigating through this intricate landscape of imbalanced class distribution, our forthcoming chapters will meticulously elucidate the multifaceted strategies employed to rectify and harmonize the dataset. This undertaking involves a meticulous exploration of methodologies such as oversampling minority classes, undersampling the majority class, or resorting to advanced techniques like synthetic data generation to rectify the skewed class distribution.

Furthermore, we will expound upon the intricate interplay between the rectified dataset and its pivotal role in both the training and testing phases of our baby cry identification model. By delving into the specifics of how these balanced datasets are meticulously curated and employed, we aim to provide a comprehensive understanding of the foundational elements that contribute to the robustness and generalizability of the developed model. Through these detailed insights, our objective is to showcase not only the challenges posed by class imbalance but also the innovative strategies and methodologies devised to surmount them in the pursuit of a sophisticated and effective infant cry identification system.

2.2 Preprocessing

First, we load the baby crying audio data from a file. We set the sample frequency to 44,100 Hz to ensure high resolution in the audio data. This step is crucial to ensure that the audio data is of good enough quality for further analysis.

![Figure 2. Preprocessing workflow](image-url)

The next step in data preprocessing is to generate Mel-Frequency Cepstral Coefficients (MFCC) features from the baby cry audio data. MFCC is a commonly used feature representation in speech analysis, particularly in speech pattern recognition. In this context, we used 20 MFCC coefficients to describe the characteristics of baby crying sounds. To obtain the MFCC coefficients, we also consider the skip length, which is equivalent to 1/5 of the sample frequency, which affects how often audio frames are taken for feature extraction. In addition, we specified other parameters such as the number of frames to be 8192 and the number of filter banks to be 24. The number of frames represents the frames used in the analysis, while the number of filter banks affects how the frequency spectrum of the sound is decomposed into MFCC features[13]. Proper selection of these parameters is essential to produce an accurate representation of the baby crying audio data, which will be used in the next step of the baby crying identification process. With this MFCC feature extraction, the audio data has been transformed...
into a more representative form, making it easier to model and recognize patterns using Convolutional Neural Networks (CNN). Calculate the FFT value and square it to get the power spectrum:

\[ P(f) = |FFT(frame)|^2 \]  

We can see in (1). Fast Fourier Transform (FFT) is a technique used to transform a signal from the time domain to the frequency domain, which allows analysis of the frequency components of the signal[14]. The FFT results in a frequency spectrum that shows the contribution of the frequency components to the signal. \[ |FFT(frame)| \] refers to the magnitude (size) of the FFT result, measured as an absolute value, removes phase information, and highlights the strength of the frequency components in the signal. Squaring the magnitude \[ |FFT(frame)|^2 \] used to calculate the power spectrum, removes negative signs and gives greater emphasis to the strong frequency components in the signal. So \( P(f) \) is the power spectrum of the signal frame.

Apply the Mel filter bank to the power spectrum to obtain the energy of the filter bank: Create a set of Mel filters (usually in the form of triangular filters) on the Mel frequency scale[15]. For each filter, calculate the weighted sum of the power spectrum in the frequency range covered by that filter. The signal energy in the Mel m filter bank is given by:

\[ E_m = \sum_{f_{\text{min}}}^{f_{\text{max}}} P(f) \cdot H_m(f) \]  

We can see in (2). \( E_m \) is the energy in the Mel m filter bank, and \( H_m(f) \) is the value of the Mel m filter bank at frequency f. This involves summing the power spectrum components of the signal from \( f_{\text{min}} \) to \( f_{\text{max}} \) which represents how various frequencies in the signal contribute to a particular Mel filter. Apply the natural logarithm to each filter bank energy value to reduce the dynamic range:

\[ L_m = \ln(E_m) \]  

We can see in (3). \( L_m \) is the natural logarithm of the energy measured in the Mel m filter bank. Discrete Cosine Transform (DCT) is a signal processing technique used to transform data from the spatial domain to the frequency domain, which is often used in image compression and audio compression[16]. Apply the Discrete Cosine Transform (DCT) to the energy values of the log filter bank to correlate the coefficients and obtain a series of MFCC coefficients:

\[ c_m = \sum_{m=1}^{M} L_m \cdot \cos\left(\frac{\pi i (m-0.5)}{M}\right) \]  

We can see in (4). \( c_m \) is the mth MFCC coefficient you want to calculate. M is the number of Mel filters or the number of MFCC coefficients you want to generate. is the natural logarithm (base e logarithm) of the filter bank energy. \( L_m \) obtained by applying the natural logarithm to the energy measured in the m Mel filter, as you have mentioned earlier. i is the index that represents the frequency component in the calculation. This usually ranges from 1 to the number of relevant frequencies.

By performing this preprocessing with the help of the Librosa library[17] ensures that the baby cry audio data is primed for use in convolutional models, enhancing pattern recognition. Refer to Figure 2 for essential preprocessing steps in baby cry identification, vital for accuracy and data balance.

2.3 Convolution Model

The Convolutional Neural Network (CNN) model is a key component in the baby crying identification process. Convolutional Neural Network (CNN) is a type of artificial neural network architecture specifically designed for pattern recognition tasks on data, mainly used in image processing and visual object recognition[5]. You can see an illustration of the model in Figure 3.
First, there is an initial convolution layer with 8 filters of size (5,5) and a ReLU activation function. ReLU (Rectified Linear Unit) is an activation function in artificial neural networks that replaces negative values in the input with zero and leaves positive values unchanged, which is often used to speed up and improve machine learning[18]. The input is the result of MFCC feature extraction with the form (20, 28, 1). The settings for this layer indicate that it uses no bias. Next, there is a first max-pooling layer with window size (2,2) and stride (1,1) to reduce the dimensionality of the data. The model continues with a second convolution layer with the same configuration. The second max-pooling layer has similar settings. The third convolution layer allows for deeper feature extraction. After the convolution operation, the output is flattened to be used as input to the fully connected layer.

The architectural design of the model is intricately crafted, incorporating multiple layers to optimize its performance in identifying and classifying baby cries based on Mel-frequency cepstral coefficients (MFCC) features. Commencing the architecture is the inclusion of a first dropout layer, strategically configured with a dropout rate of 0.2. This initial dropout layer plays a pivotal role in the model's training process by addressing the issue of overfitting, thereby enhancing its generalization capabilities.

Moving forward, the fully connected layer takes center stage, exhibiting a comprehensive connection with 48 neurons. This layer is meticulously configured to leverage the Rectified Linear Unit (ReLU) as its chosen activation function, contributing to the model's ability to capture and propagate relevant features through the neural network. The utilization of ReLU is instrumental in introducing non-linearity to the model, enhancing its capacity to learn complex patterns inherent in the MFCC features.

To further fortify the model against overfitting, a second dropout layer is seamlessly integrated into the architecture, mirroring the design of the initial dropout layer with a dropout rate of 0.2. This dual-dropout strategy serves as a robust mechanism to reduce the risk of overfitting during the training phase, promoting the model's adaptability to diverse datasets.

As the architecture unfolds, the final layer emerges in the form of the softmax output layer. This layer is adept at generating probability distributions for a set of 5 classes, each representing distinct categories of baby cries. The softmax function, a cornerstone of this layer, plays a crucial role in transforming raw scores or values into a coherent probability distribution. In the context of this model, the softmax function is particularly instrumental in facilitating multiclass classification, providing a nuanced understanding of the likelihoods associated with each potential class[19]

The culmination of this intricate configuration lies in the determination of the final prediction. The model discerns the class with the highest probability from the softmax output layer, effectively making the ultimate prediction regarding the category of the baby cry. This meticulously designed Convolutional Neural Network (CNN) architecture, with its layered approach and thoughtful integration of dropout mechanisms and activation functions, stands as a testament to its prowess in accurately identifying and classifying baby cries with a notable degree of accuracy based on the underlying MFCC features.

2.4 Training Method

The intricate process of training the Convolutional Neural Network (CNN) model to optimize its proficiency in identifying baby cries unfolds as a multifaceted orchestration, entailing a nuanced exploration of numerous sophisticated components. This pivotal stage in model development traverses through a labyrinth of decisions and methodologies, each contributing to the intricate dance of parameters and algorithms that define the model's ultimate predictive prowess.

Initiating our journey at the genesis of model construction, we find ourselves immersed in a realm of meticulous decisions, each carrying profound implications for the trajectory of the model's learning. At the forefront of these decisions is the selection of the optimizer, a pivotal aspect that sets the tone for the model's adaptability and responsiveness. In this instance, the Nesterov Accelerated Adaptive Moment assumes a central role, wielding a learning rate of 0.00175. This optimizer, an amalgamation of Nesterov Accelerated Gradient and Adaptive Moment Estimation (Adam) methodologies, unfolds as a sophisticated algorithm, dynamically adjusting the learning rate throughout the intricate process of training deep learning models[20], [21]. In concert with this, the choice of the loss function adds another layer of complexity to the model's architecture, with the Cross Entropy Loss method standing as the metric of choice, meticulously scrutinizing the model's prediction errors with granular precision.

Transitioning seamlessly into the subsequent phase, the incorporation of checkpoints emerges as a strategic fortification for the model. The meticulous establishment of a checkpoint protocol serves as a sentinel, safeguarding only the paragon of excellence, the model boasting the highest validation accuracy. This strategic approach ensures not only the proficiency of the retained model but also validates its prowess through stringent validation scrutiny, establishing it as the consummate performer among its algorithmic peers.

As we traverse the climactic phase of this intricate process, the model undergoes a rigorous training regimen spanning 50 epochs. Each epoch unfolds as a comprehensive iteration, wherein the model processes the entire expanse of the training dataset, engaging in a meticulous dance of parameter adjustments and fine-tuning. Amidst this training trajectory, the infusion of validation data takes center stage, assuming the role of a vigilant litmus test to continually monitor the model's performance and adeptly address any burgeoning signs of overfitting.
Simultaneously, the strategically positioned predefined checkpoints operate dynamically during the training process, astutely saving the model that ascends to the zenith of validation accuracy. In summation, this holistic training stage unveils itself as an intricate tapestry woven with strategic decisions and meticulous methodologies. From the deliberate configuration of the optimizer and the construction of the model adorned with specifically chosen loss functions and metrics to the implementation of checkpoints as guardians of model excellence, the process is orchestrated with an exquisite precision akin to a symphony of algorithms. As the training unfolds over the stipulated number of epochs, the model not only refines its capabilities but undergoes a dynamic evolution towards heightened proficiency, culminating in an augmentation of its predictive acumen that reverberates across the algorithmic landscape.

3. RESULTS AND DISCUSSION

The baby crying dataset was collected and preprocessed, undergoing undersampling using ENN to address class imbalance, coupled with SMOTE for oversampling minority classes with limited samples. Furthermore, data augmentation via pitch shifting and noise addition was conducted to enrich dataset diversity. The subsequent division of the dataset into training and validation sets enabled independent model training and evaluation. Experimental results, depicted in Figure 4, demonstrated a notable enhancement in the model's ability to accurately identify infant cries, despite significant data imbalance.

Figure 4. Workflow for handling data imbalance, adding data, and converting it into training and validation data

3.1 Dataset Imbalance Handling

To address the issue of significant imbalance in the baby crying dataset, we implemented several strategies using the imbalanced-learn library[22]. We recognize that the imbalance in the number of samples between classes may impact the quality of the baby crying identification results. Therefore, we took the following actions:

3.1.1 Undersampling with ENN (edited Nearest Neighbor)

Edited Nearest Neighbors (ENN) is an undersampling method used to reduce the number of irrelevant or disturbing samples in the majority class for classification modeling. [1]. ENN works by removing samples from the majority class that are categorized as noise based on comparison with their nearest neighbors from the same class or minority class. In this way, ENN helps simplify the data by removing samples that may cause confusion in modeling, thereby improving classification accuracy on imbalanced data.

Algorithm 1. ENN

\[
\begin{align*}
D & \leftarrow \text{data} \\
K_{\text{neighbors}} & \leftarrow \text{number of nearest neighbors} \\
\text{clean sample} & \leftarrow \text{list} \\
\text{for} \; \text{sample} \; \epsilon \; D & \; \text{do} \\
\text{neighbors} & \leftarrow \text{find K nearest neighbors of sample base of } K_{\text{neighbors}} \\
\text{same class count} & \leftarrow 0 \\
\text{for} \; \text{neighbor} \; \epsilon \; \text{neighbors} & \; \text{do} \\
\text{if} \; \text{neighbor.class} = \; \text{sample.class} & \; \text{then} \\
\text{same class count} & \leftarrow \text{same class count} + 1 \\
\text{end} \\
\text{end} \\
\text{if} \; \text{same class count} > \; \frac{\text{number of neighbors}}{2} & \; \text{then} \\
\text{clean sample} & \; \text{append sample} \\
\text{end} \\
\text{end} \\
\text{return} \; \text{clean sample}
\end{align*}
\]

We took the approach of reducing the sample size in the majority class, 'hungry' which initially consisted of 382 samples. With the help of ENN, we reduced the sample size of 'hungry' to 25 samples, which balanced the class distribution with the other classes. In addition to reducing the sample size of the 'hungry' class, we also
undersampled the 'discomfort' class, which originally consisted of 27 samples, and reduced it to 15 samples. This was done to maintain the relative balance between the classes.

3.2.2 Over-Sampling with SMOTE (Synthetic Minority Over-Sampling Technique)

SMOTE (Synthetic Minority Over-sampling Technique) is a resampling technique used in imbalanced data processing. SMOTE works by creating additional synthetic samples of the minority class by combining data from existing instances of the minority class [2]. This helps to increase the number of samples in the minority class, thus reducing the class imbalance in the data, and in turn, can improve the performance of the prediction model especially in the case of underrepresented minority classes.

SMOTE is used on the highly imbalanced dataset of baby crying detection after being preprocessed with the edited nearest neighbour algorithm to generate more synthetic samples from the minority class. This is crucial because imbalanced datasets can hinder the model from learning patterns from the minority class effectively. By adding synthetic samples from the minority class using SMOTE, we can improve class balance in the dataset, aiding the model to learn better and make more accurate predictions, especially for previously underrepresented minority classes.

Algorithm 2. SMOTE

\[
\begin{align*}
D & \leftarrow \text{data} \\
\text{minority class} & \leftarrow \text{underrepresented class.} \\
K_{\text{neighbors}} & \leftarrow \text{number of nearest neighbors.} \\
\text{oversampling ratio} & \leftarrow \text{ratio of synthetic samples.} \\
\text{minority samples} & \leftarrow \text{retrieve samples with minority class} \\
\text{synthetic samples} & \leftarrow \text{list} \\
\text{for} \text{sample} \in \text{minority samples} \text{ do} \\
& \text{for} \ i \in \text{number of oversampling ratio do} \\
& \quad \text{nearest neighbors} \leftarrow \text{find k nearest neighbors for sample.} \\
& \quad \text{selected neighbors} \leftarrow \text{randomly select nearest neighbor.} \\
& \quad \text{synthetic sample} \leftarrow \text{sample} + \text{random weight from 0 to 1} \ast (\text{selected neighbors} - \text{sample}) \\
& \quad \text{synthetic samples appned} \text{ synthetic sample} \\
& \text{end} \\
& \text{end} \\
D & \text{append} \text{ synthetic samples} \\
\text{return} \ D
\end{align*}
\]

For classes that initially had a very limited number of samples, such as ‘burping’, ‘belly pain’, and ‘tired’, we applied oversampling with the help of SMOTE. The result of this strategy was an increase in the sample size for each class, with ‘burping’ becoming 33 samples, ‘belly pain’ becoming 30 samples, and ‘tired’ becoming 30 samples.

By strategically amalgamating the sophisticated techniques of undersampling for the majority class and oversampling for the minority class, an intricately balanced dataset has been meticulously curated to serve as the cornerstone for the training and testing phases of the baby crying identification model. This nuanced approach to dataset preparation aims to redress the inherent imbalance by addressing the disproportionate representation of the majority and minority classes. The concerted effort to rectify this imbalance is not merely a technical necessity but a crucial step towards bolstering the accuracy and overall efficacy of the baby cry identification model.

Undoubtedly, the presence of an imbalanced dataset poses challenges to machine learning models, particularly in scenarios where the classes are unevenly distributed. In the realm of identifying baby cries, where the instances of the majority class (non-crying instances) significantly outweigh the minority class (crying instances), traditional training methods may result in biased models that tend to prioritize the majority class. This not only hampers the model's ability to accurately discern and classify instances of infant distress but also compromises its generalizability.

In response to this challenge, the amalgamation of undersampling and oversampling techniques stands out as a sophisticated strategy. Undersampling selectively trims the abundance of instances from the majority class, ensuring a more equitable representation of both classes, while oversampling diligently augments the instances of the minority class, effectively elevating its influence during model training. This intricate dance of data manipulation crafts a harmonious dataset that not only rectifies the initial imbalance but also empowers the model to learn and generalize effectively across both classes.

The ramifications of this balanced dataset extend beyond mere technicalities. By fortifying the model's training data with a more equitable distribution of instances, the resultant baby cry identification model becomes inherently more resilient and adept at handling real-world scenarios. The refined accuracy achieved through this approach contributes to a model that is not only precise in distinguishing between crying and non-crying instances but is also robust in the face of diverse and dynamic environmental factors.
In essence, the integration of undersampling and oversampling techniques represents a sophisticated stride towards enhancing the performance and reliability of the baby cry identification model. It reflects a commitment to not only rectify dataset imbalances but also to cultivate a model that is attuned to the subtleties of infant distress, ensuring its applicability and efficacy in real-world scenarios where the accurate identification of crying instances holds paramount importance.

3.3 Data Augmentation

In an effort to increase the diversity of the dataset and strengthen the ability of the baby crying identification model, we applied data augmentation techniques. Data augmentation is a commonly used technique in audio data processing to introduce variability into an existing dataset. In this context, we applied two different data augmentation techniques:

a. Pitch Shifting: We apply pitch shifting to the existing audio samples. Pitch shifting is the process of changing the pitch of a sound without changing its duration or speed. By applying pitch shifting, we created a variety of sounds with different pitches. This technique is particularly useful because baby cries can vary in pitch, and we wanted to train our model to recognize these variations. The result of pitch shifting is additional audio samples with different pitches.

b. Noise Addition: We also added noise to the original audio samples. The addition of noise aims to simulate different environmental conditions and allow the model to be more resilient to potential audio interference in real situations. By adding noise, we created more realistic audio variations. This technique allows the model to identify baby cries in noisier environments. The result is additional audio samples with varying noise levels.

By applying pitch shifting and noise addition, we managed to increase the number of samples in the dataset. As a result, we generated a total of 399 audio samples in our dataset. The addition of data is an important step in increasing the diversity of the dataset and improving the ability of the model to identify baby cries in different situations and variations in the real world.

3.4 Data Segregation for Training and Validation

In managing the dataset, we split the data with an allocation of 70% for training and 30% for validation. With this ratio, 70% of the dataset is used as training data to train the model in recognizing baby cries. The training data allows the model to understand various patterns and features in the baby crying audio. Meanwhile, the remaining 30% is used as validation data. The validation data is not used during the training process but is set aside as an independent sample to test the performance of the model. The results from validation testing help evaluate how accurately the model can recognize baby cries and identify overfitting or generalization issues in the model. By splitting the data, we ensure that our model has sufficient opportunity to learn from significant data and be tested with independent data, thus supporting the development of a reliable and accurate model.

By splitting the data after the augmentation process and handling imbalanced data using a combination of Edited Nearest Neighbor (ENN) and SMOTE, we want to emphasize that these steps are designed to minimize the impact on the quality of the original data. The ENN process at the beginning of the data flow aims to filter out significant data that may have imbalances with other data, ensuring that the addition of data through SMOTE occurs with minimal changes, thus preserving data consistency.

By dividing the data after these steps, we can ensure that the representation of each class remains balanced during the evaluation stage. This helps prevent anomalies or imbalances that may occur during the model evaluation on augmented data. In other words, separating the data after augmentation and imbalanced data handling can enhance the representation and consistency of the data, ultimately leading to a more accurate evaluation of the model’s ability to recognize baby cries.

3.5 Experiment Results

<table>
<thead>
<tr>
<th>Class</th>
<th>Addressed Imbalance (Ours)</th>
<th>Previous research by Yusdiantoro et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Remember</td>
</tr>
<tr>
<td>Belly pain</td>
<td>0.92</td>
<td>1.00</td>
</tr>
<tr>
<td>Burning</td>
<td>0.89</td>
<td>0.97</td>
</tr>
<tr>
<td>Discmofort</td>
<td>1.00</td>
<td>0.83</td>
</tr>
<tr>
<td>Hungry</td>
<td>0.68</td>
<td>0.81</td>
</tr>
<tr>
<td>Tired</td>
<td>0.96</td>
<td>0.74</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Macro Averages</td>
<td>0.89</td>
<td>0.87</td>
</tr>
<tr>
<td>Weighted Average</td>
<td>0.89</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Table 1. Comparison of Results
Our experimental results show a significant improvement in the identification of infant cries compared to previous studies. In our study, for the 'belly pain' category, we achieved a precision of 0.92, recall of 1.00, and F1-score of 0.96, which was supported by 23 samples. These results show a much higher level of accuracy in recognizing baby cries caused by belly pain. In contrast, a previous study achieved a precision of 0.75, recall of 0.43, and F1-score of 0.55 for the same category, with a much smaller sample size.

For the 'burping' category, we achieved a precision of 0.89, recall of 0.97, and F1 score of 0.93, supported by 33 samples. These results show a significant improvement in our model's ability to recognize baby cries associated with burping. Meanwhile, previous studies achieved a precision of 0.5, recall of 0.25, and F1-score of only 0.33 for the same category.

In the 'discomfort' category, our study achieved a precision of 1.00, recall of 0.83, and F1 score of 0.91, which was supported by 12 samples. These results indicate a much better ability in identifying infant cries caused by discomfort. In contrast, the previous study achieved a precision of 1.00 but a recall of only 0.40 and an F1-score of 0.57 for the same category.

In the 'hungry' category, we achieved a precision of 0.68, recall of 0.81, and F1 score of 0.74, which was supported by 21 samples. These results show a significant improvement in recognizing hungry baby cries. In a previous study, the precision reached 0.87, but the recall was 0.99, with an F1-score of 0.92 for the same category.

Finally, for the 'tired' category, our study achieved a precision of 0.96, recall of 0.74, and F1 score of 0.84, supported by 31 samples. These results demonstrate the ability of our model to identify infant cries associated with sleepiness. In contrast, a previous study only achieved a precision of 0.12, recall of 0.33, and F1-score of 0.18 for the same category, with a very limited number of samples.

Overall, our experimental results yielded an overall accuracy of 0.88, with an average precision, recall, and F1-score of 0.89, 0.87, and 0.87, respectively. In contrast, the previous study achieved an overall accuracy of 0.86, with an average precision of 0.62, recall of 0.41, and F1-score of 0.47.

The results, as conveyed in Table 1, highlight that the model developed in our study substantially enhances the identification of infant cries with elevated accuracy and improved evaluation metrics, even amidst considerable data imbalance. Our research offers a robust solution for infant cry identification across diverse conditions, with positive implications for infant health monitoring and the alleviation of parental stress.

4. CONCLUSIONS

This research has skillfully navigated the intricate challenges associated with identifying baby crying, employing a comprehensive methodology that amalgamates oversampling, undersampling, and data augmentation techniques, coupled with the implementation of Convolutional Neural Networks (CNN). By strategically addressing the issue of data imbalance through the strategic application of undersampling and oversampling methods, the initially skewed baby cry dataset underwent a metamorphosis, emerging as a more equally balanced dataset. Noteworthy is the model's achievement of an impressive 88% accuracy, representing a discernible improvement from the 86% reported in the previous study. Furthermore, commendable evaluation metrics were observed for each specific crying category, encompassing 'belly pain,' 'burping,' 'discomfort,' 'hunger,' and 'tired.' The incorporation of sophisticated data augmentation techniques, including but not limited to pitch shifting and noise addition, played a pivotal role in not just expanding the dataset but also in successfully enhancing its variability. The outcomes of the experiments not only showcase significant strides in terms of accuracy but also underscore a substantial enhancement in the overall balance of the baby crying identification model. The findings from this study not only establish a robust foundation for the evolution of superior infant cry identification models but also set the stage for the development of applications that could revolutionize infant health monitoring. While the current model has demonstrated significant strides, it is imperative for future research endeavors to delve deeper into exploring alternative models and conducting rigorous real-world testing. This would not only fortify the reliability and applicability of the model but also facilitate its seamless integration into a broader spectrum of applications, thereby maximizing its impact on infant health monitoring. The ongoing pursuit of advancements in this field holds the promise of yielding even more refined and versatile solutions for infant health monitoring. Infant cry identification, paving the way for a new era of innovation in infant care and health monitoring applications.

REFERENCES


