12-13-2021

Infant Cry Signal Processing, Analysis, and Classification with Artificial Neural Networks

Chunyan Ji

Recommended Citation

Ji, Chunyan, "Infant Cry Signal Processing, Analysis, and Classification with Artificial Neural Networks." Dissertation, Georgia State University, 2021.

doi: https://doi.org/10.57709/25943253

INFANT CRY SIGNAL PROCESSING, ANALYSIS, AND CLASSIFICATION WITH ARTIFICIAL NEURAL NETWORKS

by

CHUNYAN JI

Under the Direction of Yi Pan, Ph.D.

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2021

ABSTRACT

As a special type of speech and environmental sound, infant cry has been a growing research area covering infant cry reason classification, pathological infant cry identification, and infant cry detection in the past two decades. In this dissertation, we build a new dataset, explore new feature extraction methods, and propose novel classification approaches, to improve the infant cry classification accuracy and identify diseases by learning infant cry signals.

We propose a method through generating weighted prosodic features combined with acoustic features for a deep learning model to improve the performance of asphyxiated infant cry identification. The combined feature matrix captures the diversity of variations within infant cries and the result outperforms all other related studies on asphyxiated baby crying classification. We propose a non-invasive fast method of using infant cry signals with convolutional neural network (CNN) based age classification to diagnose the abnormality of infant vocal tract development as early as 4-month age. Experiments discover the pattern and tendency of the vocal tract changes and predict the abnormality of infant vocal tract by classifying the cry signals into younger age category. We propose an approach of generating hybrid feature set and using prior knowledge in a multi-stage CNNs model for robust infant sound classification. The dominant and auxiliary features within the set are beneficial to enlarge the coverage as well as keeping a good resolution for modeling the diversity of variations within infant sound and the experimental results give encouraging improvements on two relative databases. We propose an approach of graph convolutional network (GCN) with transfer learning for robust infant cry reason classification. Non-fully connected graphs based on the similarities among the relevant nodes are built to consider the short-term and long-term effects of infant cry signals related to inner-class and inter-class messages. With as limited as 20% of labeled training data, our model outperforms that of the CNN model with 80% labeled training data in both supervised and semi-supervised settings. Lastly, we apply mel-spectrogram decomposition to infant cry classification and propose a fusion method to further improve the infant cry classification performance.

INDEX WORDS: Infant cry classification, Neural networks, Machine learning, Graph convolutional networks, Mel-spectrogram decomposition, Spectrograms

INFANT CRY SIGNAL PROCESSING, ANALYSIS, AND CLASSIFICATION WITH ARTIFICIAL NEURAL NETWORKS

by

CHUNYAN JI

Committee Chair: Yi Pan

Committee: Rajshekhar Sunderraman

Jonathan Shihao Ji

Mark Keil

Electronic Version Approved:

Office of Graduate Services
College of Arts and Sciences
Georgia State University
December 2021

DEDICATION

I would love to dedicate this to my parents, who are both teachers and always believe that education is the most important factor in life.

ACKNOWLEDGEMENTS

It has been a wonderful journey with expectations, disappointments, frustrations, excitements, and tear of success. It is true that it is never too late to pursue your dreams. I deeply thank my advisor, Dr. Yi Pan, who has believed in me and led me to this path all these years. His state-of-the-art expertise and persistent spirit of professionalism guide me to be a lifelong learner and perform beyond. His encouragement and guidance have always accelerated my eagerness to learn more and go further.

I share my gratitude to all my committee members, Dr. Raj Sunderraman, Dr. Jonathan Shihao Ji, and Dr. Mark Keil, for their constructive comments, suggestions, and support. I thank all professors, who taught me computer science topics and other foundation knowledge at GSU, for their dedication of educating students. I thank all staff working in the computer science department for being helpful for students. Special thanks to Tammie Dudley who is always so patient and warm helping everyone. I thank UIC, where I worked for 10 years, for motivating me to look further and act. Special thanks to Dr. Lilian Kwan, former associate president of UIC, who encouraged me to pursue my study regardless of any difficulties.

I thank Dr. Orion Reyes and Dr. Carlos A. Reyes for providing the access to the Baby Chillanto database, Dr. Eduard Franti for providing Dunstan Baby Language database, and all parents, doctors and nurses who support recording the infant cry databases. This research is supported by the department of computer science and Molecular Basis of Disease (MBD) at GSU. We also acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU.

Lastly, I thank my loved ones, family, and friends, who have always been there for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Artificial neural networks, the fundamental of deep learning, have been widely used in image recognition, speech recognition, and robotics, etc. Neural networks are computing systems, containing interconnected neurons, inspired by biological brain system. Input vectors, neurons, weights, activation functions, and output are the main elements in a neural network. Each neuron has a value computed in the forward propagation process based on the weights of each connection and bias of each layer. Activation functions are used to achieve nonlinearity in the network. The back propagation is the key algorithm to train the model and minimize the loss function, which evaluates how well the model fits the dataset. The tremendous success of ImageNet leads to enormous amount of research on image processing in the past decade. In recent years, more research effort is putting on audio and video processing. Besides speech recognition, which has always been an important research direction, music and environmental sound processing attract more researchers. Audio classification can be applied to many real-life applications such as health care monitoring systems, disease identification, and public safety systems, etc. Infant crying is one special type of environmental sounds that happen in our lives and attracts more research effort.

About 130 million babies are born globally each year. Taking good care of newborns is a big challenge, especially for first time parents. Following the suggestions from other parents and books is not enough to solve the problems in practice. The main reason is because it is difficult to understand the meaning of the infant cries, which is the only way that infants communicate with the world. Experienced parents, caregivers, doctors, and nurses understand the cries based on their experiences. Young parents get frustrated and have trouble calming down their babies because all cry signals sound the same to them. Accurately interpreting infant cry sound and detecting infant cry signals automatically can help parents and care givers provide better care to their babies. Early diagnosis of diseases and disorders using cry signals is non-invasive and can be performed without professionals around, hence, it can save more lives especially in underdeveloped areas. Research on infant cry started as early as the 1960s when Wasz–Hockert research group identified the four types of the cries (pain, hunger, birth, and pleasure) auditorily by trained nurses [1]. In the early years, researchers have determined that different types of cries can be differentiated auditorily by trained adult listeners. But training human perception for infant cry is much harder than training machine learning models. In Mukhopadhyay's study, the highest classification accuracy by training a group of people to recognize some cry sounds is 33.09% while machine learning algorithm based on spectral and prosodic features can recognize the same set of data and reach 80.56% accuracy [2]. Research show that machine learning algorithms achieve promising results on detecting and classifying infant cry signals. Building smart machines to understand infant cry signals leads the way to build intelligent robot caregivers in the future.

As shown in Figure 1.1, automatic infant cry research generally involves five stages: data acquisition, pre-processing, feature extraction, feature selection, and classification. Discovering novel methods in any of the stages can help improve the performance of the final classification accuracy. Due to the sensitivity of cry data, it has been difficult for researchers to acquire data needed. To date, researchers either record cry clips by themselves or ask permission for datasets from other authors. Most databases are recorded in hospital, Neonatal Intensive Care Unit (NICU), home, and clinics by recording in real time or by setting up electronic recording devices close to infants' cribs for a long period of time. Most recordings contain noises because usually infants will not be left alone in a quiet room. Signal processing techniques remove background noises and perform signal segmentation to build cry databases. Segmentation is the method to cut the cry recordings into shorter length cries that don't have breaks and unrelated sounds. Once the database is available, feature extraction is the step to extract features from different domains of the cry signals. Features extracted from time domain, cepstral domain, or prosodic domain represent different aspects of cry signals. Selecting the most appropriate features and reducing the feature dimensions are another task to save computational time and build effective classification models. Applying appropriate machine learning models for specific cry features is vital for classification or detection performance. This research explores and creates data acquisition, pre-processing, feature extraction, and classification methods to improve the accuracy of infant pathological cry identification and cry reason classification.

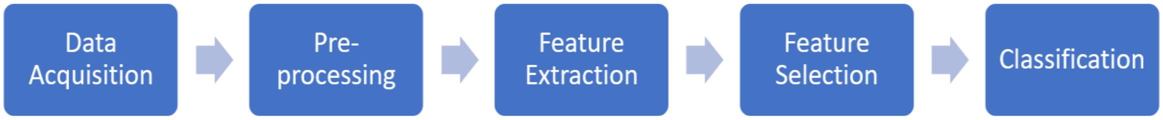

Figure 1.1:Five stages of infant cry research.

1.1 Problem Definition

This dissertation reviews infant cry research literatures in the past decade, designs and develops novel approaches to solve some problems in infant cry research by identifying pathological cries and classifying reasons behind the cries.

Working on pathological cries, we focus on designing efficient neural network models for early disease diagnosis. Firstly, we explore better algorithms to improve the accuracy of asphyxiated infant cry identification. Asphyxia is one of the main causes of infant death. Traditional methods such as blood tests and other physical procedures are time consuming, which can easily miss the best treatment time [3]. Asphyxiated cry identification by machine is very meaningful because the non-invasive method can be easily performed without professionals and high-quality equipment. Secondly, we use CNN model for infant monthly age classification to diagnose the abnormality of infant vocal tract development. It is shown that the postnatal development of vocal tract is associated with cry signals [4]. Diseases can lead to vocal tract development retardation and some healthy infants may also suffer from vocal tract development delay. Guiding infants, especially infants with tardy vocal tract development, to practice certain sounds and syllables as early as possible promotes their speech development. Finding out whether infants' vocal tract is developing normally as expected is vital for parents to take timely measures against the problems found. Compared to MRI with image processing, early diagnosis of infant vocal tract abnormality by cry signals is a non-invasive fast method that can help alert parents the possibility of any diseases or disorders that need medical assistance as early as 4-month-old.

We focus on classifying the reasons behind the crying for healthy infants. Having newborns bring happiness to the family but taking care of babies is tiring and frustrating due to the lack of understanding of infants' needs. Infants only communicate with the world by crying and normal people do not understand the meaning of the crying sound. There are many reasons behind the baby crying such as pain, discomfort, and hunger, etc. In recent years, exploring the meaning of the cries attract more research interests, but it remains in the challenging stage with relatively low accuracy due to the lack of standard public datasets and less effort on the domain. Infant sound is associated with infant speech and infant crying. Pediatricians and professionals can distinguish different types of infant sounds. It is shown that infant sound is made of four types of sound: one coming from the expiration phase, a brief pause, and a sound coming from the inspiration phase followed by another pause. An infant sound is a short-term stationary signal, which is assumed to be more stationary than a speech signal because of infants' lack of full control of the vocal tract. Figure 1.2 gives a comparison of spectrograms between infant sound and adult speech. We can see that the variations within waveform and spectrum are quite different, especially in the areas of energy, intensity, and formants. Variations in intensity, fundamental frequency (F0), formants, and duration are typical acoustic cues for infant sound and speech [5]. Adult speech's F0 ranges from 85Hz to 200Hz while infant crying signal is characterized by its high F0 within 250-700Hz. F0 is commonly computed using an auto correlation-based method provided by Praat [6]. From Figure

1.2, we can also see that the corresponding clear harmonics in the lower frequency region below 2KHz in adult speech, whereas the harmonic structure becomes drastically weaker as the frequency increases. In other words, the lower frequency region covers more energy and the transitional pattern of speech manifold in that region. This is the reason why mel-scale frequency warping is promising for speech recognition. Figure 1.2 (left) shows that the envelop of the intensity of normal baby cry signal is rhythmic and has cyclic changes due to the natural breath. It has a high pitch of about 500Hz. Further, infant sound is characterized by its periodic nature alternating crying and inspirations.

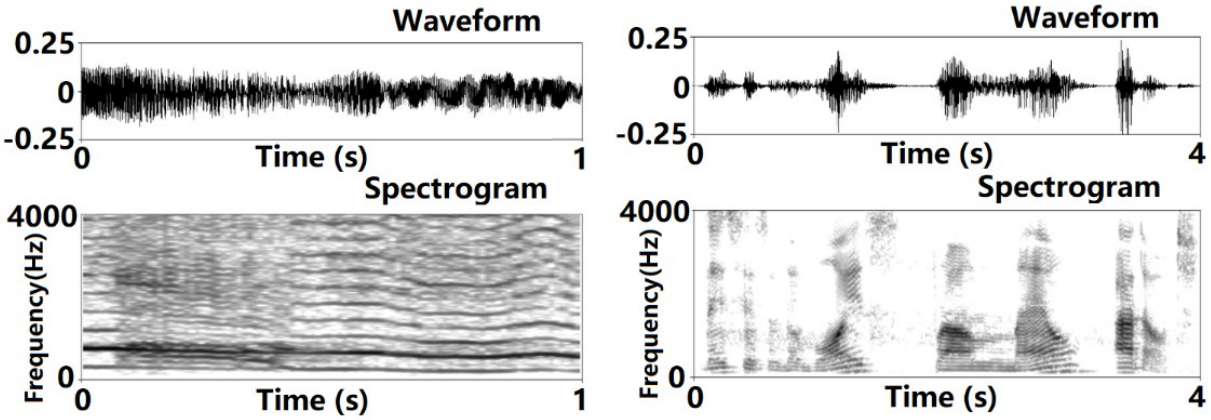

Figure 1.2: Infant cry (left) vs. adult speech (right) signal in time and frequency domain. The top images are the waveforms, and the bottom images are the spectrograms generated by Praat.

1.2 Existing Work

As the second Artificial Intelligence (AI) winter ends in the 1990s [7], neural networks emerge as a popular method in infant cry research. During the 2000s, most methods adopted in infant research are related to neural networks including scaled conjugate gradient neural network, multi-layer perceptron, general regression neural network, evolutionary neural network, probabilistic neural network, neuro-fuzzy network, and time delay neural network, etc. Hidden Markov model and Support Vector Machine (SVM) were also adopted in the 2000s. In the recent decade, many traditional machine learning methods, such as SVM, K-Nearest Neighbor (KNN), Gaussian Mixture Model (GMM), fuzzy classifier, logistic regression, K-means clustering, and Random Forest, are applied to pathological cry classification, cry reason classification, and cry sound detection. Recently, novel neural networks architectures are used pervasively in industry and research. Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), CNN-RNN, Capsule Net, Reservoir Network, and neuro-fuzzy networks start a new chapter in infant cry research. Most research works continue focusing on pathological infant cry identification, infant cry sound detection, and infant cry reason classification.

1.2.1 Databases

Finding suitable datasets is vital in machine learning. Due to the lack of public infant cry datasets, most researchers need to record their own datasets on their own. The data acquisition stage includes recording the infant cry sounds and labeling. Most databases are recorded in hospitals or homes, labeled by doctor, nurses, or parents. Digital recorders are placed close to infants and are either operated on the spot to capture the cry signals one by one or left on to record the sound events around the infants for a long period of time. Due to the limitation of resources and the sensitivity of infant cry data collection process, the number of infant cry databases are very limited.

Table 1.1 shows the main databases used in literatures in the past two decades. We can see that most datasets are in small sizes. The average size is 2983 samples, and the biggest dataset is less than 20,000 samples. So far, the most often used database in infant cry research is Baby Chillanto database. Baby Chillanto database was collected by the National Institute of Astrophysics and Optical Electronics, CONACYT Mexico [8]. It contains five types of cry signals including deaf, asphyxia, normal, hungry, and pain. Each cry is equally segmented into one second

Table 1.1 Main databases used in literatures

Database	**Creator**	**Data**	
Baby Chillanto (2004)	National Institute of Astrophysics and Optical Electronics, CONACYT Mexico	Deaf	879
		Asphyxia	340
		Normal	507
		Hunger	350
		Pain	192
		Total	**2268**
Dustan Baby Language (2018)	Extracted from Dunstan Baby Language DVD by authors	Hungry	56
		Sleepy	106
		Burping	55
		BellyPain	37
		Discomfort	61
		Total	**315**
Donate A Cry (2015)	https://github.com/gveres/donateacry-corpus	Hungry	382
		Tired	24
		Burping	8
		BellyPain	16
		Discomfort	27
		Total	**457**
CRIED (2018)	INTERSPEECH 2018 computational paralinguistic challenge	Neutral	2292
		Fussing	368
		Crying	178
		Total	**2838**
iCOPE (2019)	**https://www.infantcope.org/**	Pain	42
		NoPain	71
		Total	**113**
ChatterBaby (2020)	**https://chatterbaby.org/**	Fuzzy	171
		Hungry	167
		Pain	353
		Colic	380
		Total	**1071**
SPLANN (2015)	“Sf. Pantelimon” Emergency Clinical Hospital within the SPLANN research project (project coordinator SOFTWIN SRL). Bucharest, Romania	Colic	225
		Eructation	505
		Discomfort	2210
		Hungry	5536
		Pain	4404
		Pathology	459
		Tired	34
		Total	**13373**
self-recorded database (2019)	Recorded by first-time parents	Pain	5445
		Hungry	6263
		Sleepy	4927
		WetDiaper	3056
		Total	**19691**
self-recorded database (2019)	Self-recorded in NICU of a hospital	Hungry	422
		Diaper	137
		Attention	151
		Sleepy	79
		Discomfort	182

		Total	**971**
self-recorded database (2016)	Collected from National Taiwan University Hospital	Hungry	586
		Pain	723
		Sleep	860
		Total	**2169**
DA-IICT Cry (2018)	Self-recorded	Normal	793
		Asphyxia	215
		Asthma	182
		Total	**1190**
Autism database (2019)	Self-recorded	Normal	64
		Autism	20
		Total	**84**
Hypothyroid database (2009)	Self-recorded	Normal	45
		hypothyroid	43
		Total	**88**
ADEL database (2010)	Self-recorded	Normal	22
		ADEL	17
		Total	**39**

long and the total number of cries is 2268. Another database used in multiple literatures is named Dunstan Baby Language database [9], which is extracted from the Dunstan baby video tutorial presented by Priscilla Dunstan who invented the Dunstan Baby Language theory. There are several versions of Dunstan Baby Language database since authors extracted the audio clips in their own ways. The version described in [9] consists of 315 wave files, sampled at 16 kHz, with a variable length between 0.3 and 1.6s. Each utterance is a word of infant speech corresponding to one of the five “Dunstan words,” which were translated as “Neh” = hungry, “Eh” = need to burp, “Oah” = tired, “Eairh” = low belly pain, and “Heh” = physical discomfort. Many databases are self-recorded for research. One database named “Donate A Cry” [10] is available online, but it is not well labeled and only one literature is found using this database. Some databases are recorded in the NICUs, pediatric clinics, or baby-sitting environments. Some cry audio signals online are also collected, and some synthetic databases are created by the authors to compare the performances of the proposed methods on real databases and synthetic databases. In Ferretti’s work [11], the CNN detects the cry signal better on the synthetic database than the real database. It shows that the automatic detection and classification of real-time infant cry is still challenging because the real-time environment may exist many types of complications that can affect the quality of the cry signals. Synthetic databases can be generated by adding noises to clean cry recordings or combining different cries together. Training models on synthetic databases can avoid requiring a large amount of data to be acquired in sensible environments such as NICUs. To solve limited data problem, researchers use data augmentation techniques. Zhang et al. created new waveform images from training datasets by transforming these waveform images into slightly faster or slightly slower waveforms for the purpose of increasing training datasets to overcome overfitting problem [12]. Several other data augmentation techniques, such as noise variation, signal intensity variation, tonality variation, and spectrogram's size alteration, were used to artificially increase either the number of audio signals or the number of spectrograms [13]. The experimental results show that these data augmentation methods cannot lead to accuracy improvement. The reasons lie in the fact that the limited data cannot capture the diversity of variations within infant cry signals. Recently, Yao et al. uses Specaugment [14], which flips the crying windows horizontally as well as used time masking deformation to randomly mask blocks of time steps for each window [15], and it helps with unbalanced dataset problem. Other data augmentation techniques need to be explored and recording larger size datasets is still urgently needed.

1.2.2 Signal Processing

The main tasks in pre-processing stage are denoising and audio segmentation. The complication of the recording environment leads to noisy infant cry recordings. In a neonatal care unit, besides infant cry signals, there could be many kinds of sounds such as footsteps, adult's speech, air-conditioner sound, and alarm sound, etc. To detect or classify cry signals accurately, cleaning up the recorded data at the pre-processing stage is a crucial step. To clean up a signal, the first task is denoising, which removes the background sounds such as speech, fan, and footstep, etc. Turan and Erzin applied high-pass FIR filter to remove the speech sound and low frequency noise in the recording [16]. Ferretti et al. reduced coherent noise source by a filter-and-sum beamformer and uses OMLSA post-filter to reduce the residual diffuse noise [11]. Gu et al. used optimized Blackman window to handle each frame signal, which is the result after the endpoint detection [17]. The signal noise is significantly reduced after filtering. Recorded audio files are usually large files containing many crying segments due to the characteristics of cry signals that contains respiration. In early years, equally segmenting the audio recordings into equal length of segments to build the dataset is used such as Baby Chillanto database. It may cause some complete cries broken into several pieces and affect the machine learning algorithm to recognize the overall characteristics of a cry signal. Audio segmentation task is now commonly performed using Voice Activity Detection (VAD). VAD technique is widely used in speech recognition to detect the human speech in audio signals. Researchers also use it to detect the infant cry and remove the silent duration in a sample recording. It is used to detect the presence or absence of a baby cry in a noisy environment to improve the overall baby cry recognition rate and it is used to detect the sections of the audio with sufficient audio activity, but it still faces the challenge of separating the cry and noise. Implementing a basic VAD algorithm, which uses short time features of audio frames and a decision strategy for determining sound and silence frames can be helpful to segment a large dataset, but the result may not be effective due to the limitation of the VAD algorithm. Sometimes researchers also manually cut the samples to remove the silent part and the voice interference part, and only the continuous crying part of the sound was retained. We also choose to manually segment recordings for Baby2020 database to try to achieve the best result due to the complication of different infants and recording environments.

1.2.3 Feature Extraction

Feature extraction is the stage to extract the discriminative features from the audio signals and later feed into the machine learning algorithms. It is one of the most vital parts of a machine learning process [18]. Performing feature extraction task either in time or frequency domain addresses the fundamental work of baby cry analysis and processing. Time domain features, such as zero-crossing rate, amplitude, and energy-based features, etc., are simple and straightforward to compute. While time domain features are not robust enough to cover the variations within infant cry signals and the features are sensitive to background noises, the frequency domain features have strong ability to model the characteristics within infant cry signals. Most research use frequency domain features and time domains are sometimes combined with frequency domain features to obtain better performance.

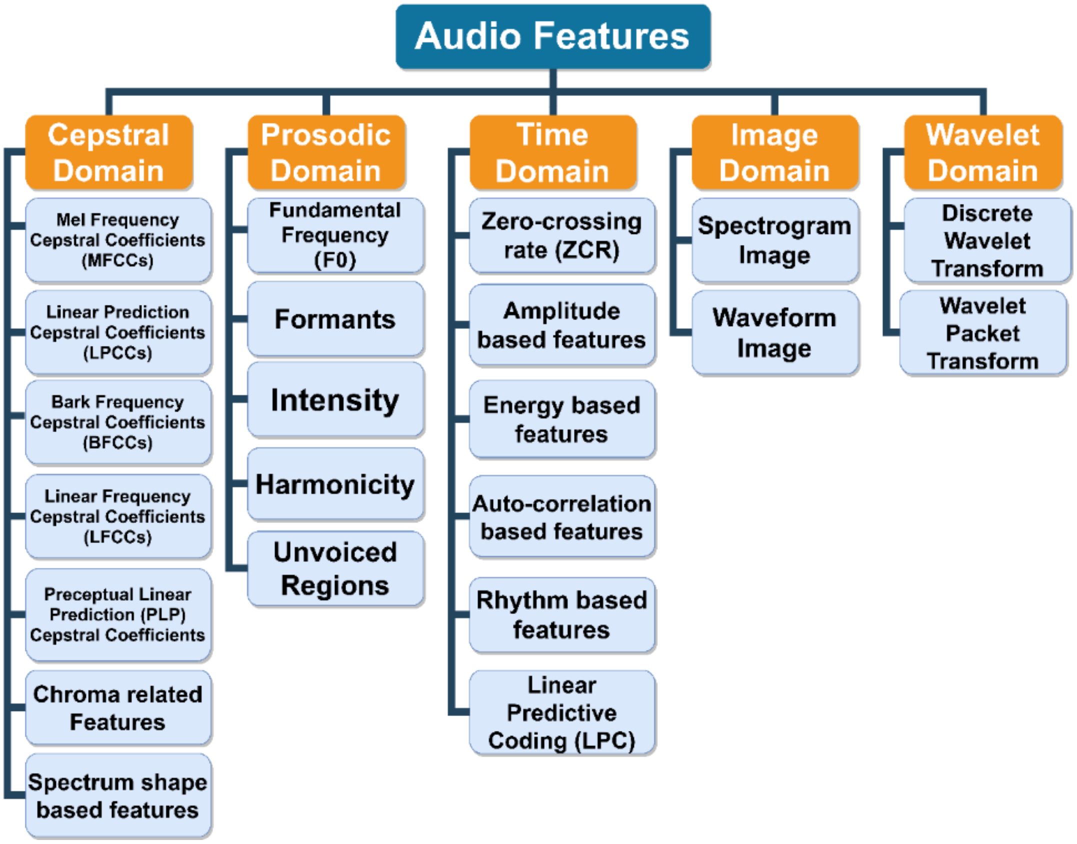
Audio Features
Cepstral Domain
Mel Frequency Cepstral Coefficients (MFCCs)
Linear Prediction Cepstral Coefficients (LPCCs)
Bark Frequency Cepstral Coefficients (BFCCs)
Linear Frequency Cepstral Coefficients (LFCCs)
Preceptual Linear Prediction (PLP) Cepstral Coefficients
Chroma related Features
Spectrum shape based features
Prosodic Domain
Fundamental Frequency (F0)
Formants
Intensity
Harmonicity
Unvoiced Regions
Time Domain
Zero-crossing rate (ZCR)
Amplitude based features
Energy based features
Auto-correlation based features
Rhythm based features
Linear Predictive Coding (LPC)
Image Domain
Spectrogram Image
Waveform Image
Wavelet Domain
Discrete Wavelet Transform
Wavelet Packet Transform

Figure 1.3:Main audio feature categories.

The commonly used Mel-frequency cepstral coefficient (MFCC), Linear Prediction Cepstral Coefficients (LPCC), and Linear Frequency Cepstral Coefficients (LFCC) have been proven better performance than using time domain features. On the other hand, it is shown that infant cry signal is rhythmic and has cyclic changes due to the natural interruption and breath. As infant cry signals are different from speech signals, not only the words or meaning in the signal is important, the tone or the style of the sound also represent the meaning of the cry. The high-level information, such as prosodic features, is also important to improve the discriminative ability within signals. Therefore, attaching prosodic domain features together with time or frequency domain is capable for capturing both physical and physiological information. In addition, images such as spectrogram are time-frequency representation of an audio clip. It is known that spectrogram has a strong ability to present the signal and include both acoustic and prosodic information, so image domain features are also widely used in infant cry research. Figure 1.3 depicts the main categories of the audio features that are applied to research related to speech, music, and environmental sounds. Acoustic and prosodic features are commonly used for infant cry detection and classification. Cepstral domain features, prosodic features, and image-based features are widely used in speech processing and infant cry processing with a proportion over 70% research articles.

Cepstral domain features. MFCC is widely used in speech recognition. It is a cepstral representation of the audio signals. Researchers use it to test proposed approaches [19] and often use it for baseline experiments [20]. According to auditory perception models, MFCC coefficients are more robust than other coefficients such as LPC coefficients. Liu et al. used MFCC along with two other cepstral features LPCC and Bark Frequency Cepstral Coefficients (BFCC) for infant cry reason classification. The result shows that the BFCC with a neural network model produces the best recognition rate of 76.47% [21]. The main idea of LPCC is to remove the redundancy from a signal and tries to predict next values by linearly combining the previous known coefficients. LFCC extraction process is similar to MFCC extraction. The difference is that it uses a linear filter-bank instead of the Mel filter-bank. Some works show that LFCC performs better than MFCC in discriminating high frequency audio signals such as female voice and baby cry signals. Researchers have also used other cepstral features such as Fast Fourier Transform (FFT), Log-Mel feature, Mel Scale, Constant-Q Chromagram, Log-mel spectrum, and delta spectrum, etc.

Prosodic domain features. It is shown that infant cry is made of four types of sound: one coming from the expiration phase, a brief pause, and a sound coming from the inspiration phase followed by another pause. Variations in intensity, fundamental frequency (F0), formants, and duration are typical acoustic cues that carry prosodic information about infant cry and speech. It is shown that the above prosodic features are efficient to identify the types of infant cry. Adult F0 ranges between 85 and 200Hz while infant crying F0 is characterized by its high F0 250–700Hz. F0 is commonly computed using an autocorrelation-based method provided by Praat [6]. According to auditory perception models, MFCC coefficients are more robust than other coefficients such as LPC coefficients. Our work [22] shows that combining weighted prosodic features with MFCC features helps improve the classification accuracy in a deep learning model. Features such as mean, median, standard deviation, and minimum and maximum of F0 and F123 voiced/unvoiced counter, consecutive F0, and harmonic ratio accumulation are also useful features to explore in infant cry research.

Image domain features. Currently, there are four types of images related to audio signal in literatures. They are spectrogram, mel-spectrograms, waveforms, and prosodic feature images. A spectrogram is an image that is a time-frequency representation of an audio clip. It is known that spectrogram has a strong ability to present the signal and include both acoustic and prosodic information. Spectrograms can be extracted through framing and windowing, FFT, and calculating the log of the filtered spectrum steps, which are illustrated in Figure 1.4. Feeding spectrograms into classifiers can solve the problem of different cry signals having different durations. Instead of using zero padding to achieve same length of feature vectors, normalization is applied in the process of spectrogram generation, which produces the same size images without changing the original signal. A mel-spectrogram is another visual representation of a signal indicating how people hear sound by converting the Y-axis to mel-scale. Mel-spectrogram can relatively represent human sound perception characteristics, which presents the linear distribution under the 1000Hz and the logarithm growth above the 1000Hz on a logarithmic scale rather than a linear scale. People are more sensitive to lower frequency sound and the difference between high frequency sound is not as easy to distinguish as the ones between lower frequency sound. We hear frequencies on a logarithmic scale rather than a linear scale. In mel-spectrogram, each unit is judged by listeners to be equal in pitch distance from the next. Waveforms, representing the pattern of sound pressure amplitude in the time domain, are not as effective as spectrograms in deep learning models for infant cry classification as shown in [23]. Prosodic feature images created in [23] contain the prosodic feature lines including F0, intensity, and formants and can be used as auxiliary features because the information contained is only certain aspects of the original audio signals.

Other relevant domain features. Other domain features used in infant cry research include time domain features such as zero-crossing rate, short time energy, and voiced-unvoiced regions. Zero crossing rate is the rate at which the signal passes zeros and changes signs. It can be used in conjunction with short-time energy to detect endpoints of speech utterances, hence, to detect the existence of the cry sound from other sounds happening in the environment. Since the amplitude of an audio signal varies with time, the short-time energy can serve to differentiate voiced and unvoiced segments, which can be used for infant cry detection and classification.

Voiced-unvoiced counter, Linear Predictive Coding (LPC), wavelet transform, and waveform packet transform are also used by research. Researchers also calculate the statistical natural parameters of the data such as mean frequency, standard deviation, and third quartile range, to help infant cry detection and classification. Feature extraction is a critical step in audio processing. Besides aforementioned Praat software, feature extraction tools such as librosa library [24] and OpenSMILE toolkit [25] have made audio feature extraction easier.

In the methods we proposed for this dissertation, we extract MFCC, spectrograms, waveforms, prosodic lines images, and mel-spectrograms from infant cry signals and feed them into our neural network models for disease identification and cry type classification.

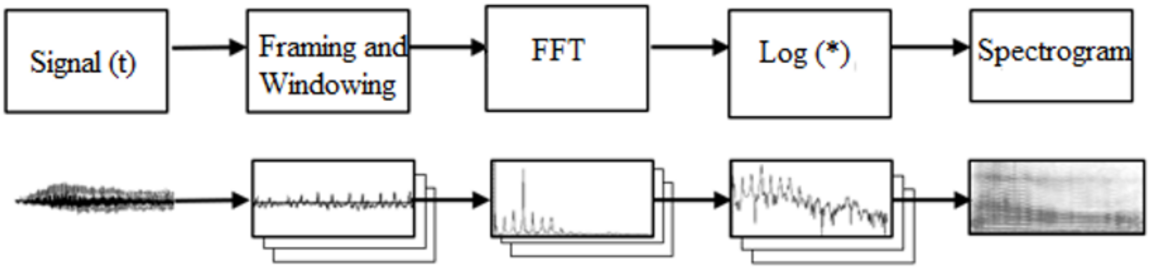

Figure 1.4:The flowchart of spectrogram generation.

1.2.4 Classification Methods

With data cleaned and segmented, features extracted, selected, and normalized, finding the appropriate classifier is the most important stage in the machine learning process. In the past decade, traditional machine learning methods and new methods are both applied to infant cry research. The main methods are listed as follows.

Support vector machine (SVM). The goal of SVM algorithm is to find decision boundaries in a multiple dimensional space that can separate the data points. It is the most popular probabilistic classifier used in the infant cry classification including multi-class SVMS, linear and RBF kernels binary SVM, and incremental SVM learning model, which keeps adding new data into the dataset in each training step. Features being fed into SVM models include temporal features, prosodic features, and cepstral features. SVM is often used to compare with other non-linear classifiers like neural networks on infant cry classification because SVMs can work effectively with limited examples and high-dimensional data [26].

K-nearest neighbor (KNN). KNN is a well-known pattern recognition method used in classification. With k nearest neighbors in the feature space, the goal is to assign the test sample to the class that its nearest neighbor belongs to. If k is greater than 1, the nearest neighbor is selected based on the number of nearest neighbors. In the case of infant cry classification, researchers use Euclidean distance, Minkowski distance, and other methods to measure the distance between two sample feature vectors and the feature vectors selected are usually MFCC and LFCC.

Gaussian mixture models (GMM). GMM is a probabilistic model that assumes the datapoints are in gaussian distribution of some mean and variance. The idea is to learn the parameters to model the provided training data as mixture of several gaussian distributions, and then the test data can be classified by the trained model. Expectation maximization (EM) algorithm is used for finding the maximum likelihood estimates of the parameters under GMM based structures. GMM clustering method is compared to hierarchical clustering and k-means clustering on cry features and shows the GMM model produces the best result with least amount of overlapping datapoints with a certain database [27]. Research also shows that GMM based classifiers are sensitive to environments and cannot lead to satisfied results especially with limited training and development data.

Fuzzy classifier. Fuzzy logic systems have been used in many applications such as transmission systems, power systems, and wireless network routing. In infant cry classification, it is used to detect infant cry signals from laughter signals [28]. Fuzzy decision tree, fuzzy decision forest, fuzzy KNN, and fuzzy relational neural network classifier are used for pathological cry classification [29]. Type-2 fuzzy pattern matching algorithm is used to classify asphyxia, normal, and hyperbilirubinemia [30], and it outperforms the SVM and the logistic regression classifier on classifying hunger and pain.

Logistic regression classifier. Logistic regression classifier is a low-complexity supervised algorithm that uses the Sigmoid function to predict observations to a discrete set of classes, and it is usually used as a referencing experiment for infant cry research. Lavner et al. use it to show that CNN performs better on cry detection [31] and Orlandi et al. use it to compare with many other classifiers, in which random forest performs the best on classifying full-term and preterm infant cries [32].

K-means clustering. K-mean clustering represents an unsupervised algorithm mainly used for clustering. Unlabeled data points can be gradually separated into groups based on the mean value and centroid moving. Sharma et al. use K-means clustering to show that the GMM model has better performance differentiating different types of cry [27]. In [33], K-means clustering is used to build a speaker database for speaker recognition.

Ensemble decision trees. Bagging, boosted trees, and random forest are techniques that perform ensemble decision trees. They all combine multiple decision trees to produce better performance. Experiments have shown that they are powerful on infant cry classification. Tree classifiers are used to compare with over 40 classifiers and showed the best overall performance comparing to bayes classifiers, lazy classifiers, function classifiers, and rule classifiers, etc. [34].

Neural networks (NN). Artificial neural network is a machine learning computing system, containing interconnected neurons, inspired by biological brain system. In 1995, Petroni et al. made the first attempt of ANN in infant cry classification [35]. Many types of NN are used in infant cry research recently including feed forward neural network (FFNN), multi-layer perceptron (MLP), probabilistic neural network (PNN), general regression neural network (GRNN), time-delay neural network (TDNN), convolutional neural network (CNN), long short-term memory (LSTM), the combination of CNN and RNN (CNN-RNN), the combination of fuzzy logic and neural network (neuro-fuzzy networks), capsule network (CAPSNET), Reservoir networks (RN). FFNN is the simplest neural network that passes information in one direction and MLP is a type of FFNN that contains at least three layers. CNN is a deep learning algorithm that is commonly applied to images and uses shared-weight architecture of convolutional layers, and it has been successfully used in computer vision, language processing, and other domains achieving unprecedented high accuracy. LSTM is a type of RNN that has internal states to make accepting sequence of data possible and is known as best neural network for time series data such as language translation and speech recognition. Neuro-fuzzy Network combines fuzzy logic with neural networks, and it has been used successfully in infant classification. Capsule Network is a deep learning topology that adds a structure called capsules into the CNN model. As maxpooling in CNN only picks the maximum value within a region and throws away information in certain positions, higher-level capsules cover larger regions of the image and performs routing by agreement instead. RN is a neural network model derived from RNN. Its input nodes connect to a nontrainable reservoir, which contains connected nonlinear units with randomly generated fixed weights. These neural network approaches perform relatively better in certain tasks with certain datasets. Each of them has advantages and disadvantages and no algorithm is perfect for every dataset and task. Selecting a suitable model to achieve high performance is still challenging. There are not very deep models with big data involved in the infant cry research yet due to the difficulty of data acquisition. At present, the largest dataset has less than 20000 samples. The small and imbalanced datasets lead to high classification accuracy but low confidence for some of the tasks. To achieve high performance with high confidence, real big data with real deep learning models are to be explored. To determine the classification ability of the different models, Fuhr et al.

experimented differentiating healthy infant cries and cries of infants suffering from several diseases using 12 classifiers including SVM, decision tree, KNN, MLP, etc. The result shows only C5 decision tree and KNN achieved greater than 90% accuracy [36]. Applying many algorithms on a task before selecting the algorithm to use is impractical. How to select a suitable algorithm for a specific task should be explored and an appropriate algorithm should be chosen accordingly for specific datasets and tasks. Comparing the machine learning algorithms used in infant research, we analyze them from the following aspects.

Time complexity. It includes training time and classification time relying on the data size, searching space, and the complexity of coefficients. In general, traditional methods such as SVM, K-means clustering, and GMM-based approaches are relatively simple and straightforward. Smaller sample size is acceptable, which differs from neural network methods. Hence, training time, searching time, and classification time are much less than those of neural network methods. Also, fine tuning in neural network models also requires more development time.

Sample complexity. It indicates whether the model requires large size of data or not to learn. It depends on the complexity of the data and the complexity of the algorithms. To reach better performance, neural network methods generally require larger sample size for complex searching space than other traditional algorithms. Larger size infant cry databases are needed for deep neural networks.

Parametricity. It indicates if the number of the parameters used in the model is fixed or it varies along when new data is brought in. Linear regression, GMM, and neural networks are parametric methods while KNN and SVM are nonparametric models.

Feature complexity. Features extracted from either time domain or frequency domain have the same abilities to represent the different characteristics of the cry signals in different models. There is no feature complexity difference involved for traditional models or neural network-based models. But using too many features to represent one sample may cause overfitting issue, therefore, selecting the most appropriate features and using appropriate feature reduction method specific models are critical.

Parallelizability. Parallelizability is a pivotal feature for saving the training time of machine learning methods. Large amount of data in neural networks is associated with high computation cost in both time and space. Parallelizability with Graphics Processing Unit (GPU) computing greatly reduces the training time and made deep learning possible. Other method such as KNN is easy to parallel, but parallelism is tricky if the next step is based on the previous step result such as decision trees.

1.2.5 Infant Cry Applications

Researchers use different classifiers to perform infant cry tasks. In the past decade, most research works continue to pay effort to improve the classification accuracy of infant cry signals including differentiating the pathological cries from the normal cries and understand the meaning behind the cry signals. Many significant works on infant cry classification and detection are done in the past decade.

Infant cry reason classification. In the early years of infant cry research, more work is done on automatically differentiating the cries of healthy infants from pathological cries and most work are on Baby Chillanto database. In recent years, exploring the meaning of the crying attract more research interests. As shown in Table 1.2, some significant works are done on classifying the reason behind the infant cries. Most methods on infant cry reason classification have accuracy around 80% while a few methods can reach over 90% accuracy on Dunstan Baby database and a binary classification on Baby Chillanto database reaches 97%. It’s noticeable that researchers are using different datasets, most of which are self-recorded. It is difficult to evaluate the performance of the different approaches because they are using different databases. With different datasets in similar research, even the classification types are the same, it is unfair to make direct comparison on the performances of the proposed methods. The infant classification remains in challenging stage due to lack of standard public datasets and the classification accuracy is still relatively low.

Table 1.2:Significant works on infant cry reason classification

First Author	**Dataset**	**Features**	**Classifiers**	**Best Performance**
Felipe (2019)	**iCOPE** **https://www.infantcope.org/** pain vs. no pain	Mel Scale (MS), MFCC, Constant-Q Chromagram (CQC); Local Binary Pattern (LBP), Local Phase Quantization (LPQ), Robust Local Binary Pattern (RLBP) extracted from spectrogram	**SVM**	**71.68%**
Sharma (2019)	Donate A Cry (Hungry, Burp needed, Belly pain, Discomfort, Tired, Lonely, Feeling cold/hot, Scared, Unidentified)	Mean frequency Standard deviation Median frequency Third quartile range Spectral entropy Kurtosis, skewness Spectral flatness, etc.	K-Means Clustering Hierarchical Clustering Gaussian Mixture Models clustering machine learning technique	**81.27%**
Maghfira (2019)	**Dunstan Baby database** (Pain, Hunger, Discomfort, Need to burp, BellyPain)	**Spectrogram**	CNN-RNN	**94.97%**
Franti (2018)	**Dunstan Baby database** (Pain, Hunger, Discomfort, Need to burp, BellyPain)	**Spectrogram**	**CNN**	**89%**
Liu (2018)	**NICU recorded** **Draw attention cry, Diaper change needed cry, and Hungry**	**LPC, LPCC, MFCC, BFCC**	**Nearest Neighbor** **Artificial Neural Network**	**76.4%**
Turan (2018)	**CRIED**	**Spectrogram**	**Capsule Network**	**86.1%**
Osmani (2017)	**Dunstan Baby database (Hunger, Pain, Tiredness, BellyPain, NeedBurp)**	Spectrum, Pitch, zero-crossing rate, root mean square, intensity, energy along with their calculated statistics (mean, variance, skewness, etc)	SVM Bagging and Boosted Trees Decision Tree classification	**N/A**
Chang (2016)	**collected from National Taiwan University Hospital** (hungry, pain, and sleep)	**Spectrogram**	**CNN**	**78.5%**
Bano (2015)	**Self-recorded in hospital** Neh (Hungry), Eh(Pain/burp-me: Pinching/Drawing blood), Owh(Sleepy),	Pitch Short-time energy MFCC Statistical properties of MFCC	**KNN**	**86%**

	Eairh(Pain), Heh(Discomfort).			
Orlandi (2015)	Self-recorded (Full term vs. Preterm)	22 acoustical parameters CU length, F0 median, F0 mean, F0 standard deviation (F0 std), F0 minimum (F0 min), F0 maximum (F0 max), number of estimated F0 values, F123 median, F123 mean, F123 standard deviation, F123 minimum, F123 maximum	Logistic Regression, Multilayer Perceptron NN, Support Vector Machine, and **Random Forest**	**87%**
Bhagatpatil (2015)	**Self-recorded** (Pain, Hunger, Discomfort, Need to burp, BellyPain)	**LFCC** MFCC	**K-mean clustering** **KNN**	**91.58%**
Rosales-Péreza (2014)	**Baby Chillanto** Hungry vs. Pain	**MFCC, LPC**	**Fuzzy Model**	**97.96%**
Yamamoto (2013)	**Self-recorded** Discomfortable, Hungry, Sleepy	**FFT**	**Nearest Neighbor**	**62.1%**

Infant pathological cry classification. Infant cry signals have been used to identify many diseases such as asphyxia, hypo-acoustic (hearing disorder), hypothyroidism, Hyperbilirubinemia, Cleft Palate, Respiratory Distress Syndrome, Ankyloglossia with deviation of the epiglottis and larynx, etc. In the past decade, researchers continue to apply novel methods to classify normal cry and pathological cry. Asphyxiated cry is the most popular disease in infant cry research. Table 1.3 shows the latest works on classifying normal cry from asphyxiated cry. Researchers have been using Baby Chillanto database to perform the binary classification. In 2012, Probabilistic Neural Network (PNN) and General regression neural network (GRNN) reached 99% accuracy [37][38], the latest SVM model can reach 97.7% accuracy [39], and the deep learning model reaches 96.74% accuracy [22].

Table 1.3:Classification of asphyxiated cry from other cries

First Author	**Dataset**	**Features**	**Classifiers**	**Performance**
Ji (2019)	Baby Chillanto	MFCC + Weighted Prosodic	DNN	96.74%
Badreldine (2018)	Baby Chillanto 340Normal,340Asphyxia	DWT-MFCC	RBR Kernel SVM	98.5% (40% test) 97.7% (10 fold)

Zabidi (2017)	Baby Chillanto database with database from University of Milano-Bicocca	MFCC Image	CNN	92.8%
Onu [(2017)	Baby Chillanto	MFCC	SVM	85%
Sachin (2017)	Baby Chillanto	Waveforms	AlexNet	92%
Moharir (2017)	Baby Chillanto	Waveforms	GoogleNet,AlexNet	94%
Rosales-Péreza (2014)	Baby Chillanto	MFCC, LPC	Fuzzy Model	90.68%
Saraswathy (2012)	Baby Chillanto	Wavelet Packet Transform	PNN, GRNN	99.04%
Hariharan (2012)	Baby Chillanto	STFT	PNN, GRNN, TDNN, MLP	99%
Hariharan (2012)	Baby Chillanto	Weighted LPCC	PNN	99%

Besides identifying asphyxia, other types of diseases have also been studied. According to Gianluca Esposito's review [40], it is shown that the infant cry signals are useful for early diagnosis of Autism Spectrum Disorder (ASD). In 2012, Orlandi et al analyzed the cry signals of the high-risk infants whose siblings have already been diagnosed to be ASD. It is noticed that less cry episodes occur, F0 is lower, and Formants reach high values for high-risk infants than healthy infants [41]. Although some babies are born with ASD, it is usually diagnosed when they are two to three years old since the diagnosis involves observing the behavior of children. This leads to the difficulty of the cry signal acquisition for autism babies. In 2019, Wu et al recorded twenty audio samples of autistic children whose ages are between 2 to 3 years old. They reached 96% accuracy by using SVM classifier with MFCC features [42], but the small sample size makes the experimental result not very convincing. Identifying hypo acoustic cry signal has been successful in the early years. In 2011, Hariharan's General Regression Neural Network reached 99% on Baby Chillanto database [43] and in 2009, O.F. Reyes-Galaviz et al. used evolutionary neural network system to reach almost 100% on Mexican-Cuba database [44]. Then in 2014, Rosales-Pérez et al used fuzzy model and genetic algorithm to reach 99.42% on Baby Chillanto database [45]. Other types of diseases such as Hypothyroidism, Respiratory Distress Syndrome, Cleft Palate, and Ankyloglossia with Deviation of the Epiglottis and Larynx were studied in the early years [46]. In 2014, Feier et al. studied newborns' cries within minutes after birth. Random tree and random forest methods were able to classify cries of healthy newborns from premature newborns, newborns with umbilical cord strangulation during birth, and newborns with other pathologies with accuracy above 95% [47].

Infant cry detection. Infant cry detection is considered as a binary classification with cry and not-cry categories. It is another attractive research topic in the latest decade. The goal is to detect the infant cry signals efficiently and accurately in various environments, such as car, home, and hospital etc., while other sounds happening at the same time. Since the data is recorded during a long period of time in a certain environment such as home or hospital, the detection algorithm needs to be able to detect the cry sound despite the background sounds happening in the environment. Table 1.4 shows some recent significant works on infant cry detection.

Table 1.4:Significant works on infant cry detection

First Author	Dataset	Features	Classifiers	Best Performance
Chang (2019)	Self-recorded (Crying with TV, Speech, etc.)	Spectrogram	CNN	99.83%
Manikanta (2019)	Recorded in homes (Crying with AC, Fan, etc.)	MFCC	1D-CNN FFNN SVM	86%
Dewi (2019)	Self-recorded samples Cry and Not Cry	LFCC	KNN	90%
Gu (2018)	Self-recorded (Crying with laughter, barking, etc.)	LPC	Dynamic time warping algorithm	97.1%
Ferretti (2018)	Real Dataset: recorded in the NICU of a hospital. Synthetic DB: Crying with speech, "beep" sounds, etc.)	Log-Mel Coefficients	DNN	Real dataset 86.58% Synthetic DB 92.92%
Feier (2017)	TUT Rare Sound Events 2017 (Crying with "glass breaking", "gunshot", etc.)	log-amplitude mel-spectrogram	CRNN	85% for baby crying detection

				(87% for all three targets, First place in the competition)
Torres (2017)	Online resources (Crying with adult cry, vacuum cleaning, etc.)	Voiced unvoiced counter, consecutive F0 and harmonic ratio accumulation, MFCC	support vector data description (SVDD) CNN	AUC 92%
Lavner (2016)	Recorded in domestic environment (Crying with speech, door opening etc.)	MFCC, Pitch, Formants, etc.	CNN	95%

It is seen from Table 1.4 that CNN based approaches reach good performance under clean and constrained conditions, e.g., over 90% accuracy. On the other hand, with noisy environment and limited training data, classifiers are sensitive at the boundary and easy to be confused and overlapped with noise signals.

1.3 Challenges

With the improvement of computational ability and the use of neural network-based algorithms, infant cry research has attracted much effort in recent years, but the following challenges remain in the field.

Lack of data and scalability of research. Research is based on different datasets recorded by authors. Therefore, it is difficult to compare the performances of methods experimented on different datasets. The only database shared by some research is Baby Chillanto database, which has been around for two decades. The total amount of Baby Chillanto database is 2268 and the largest private database has less than 20,000 samples, which is insufficient for deep learning NN models. Data is the key elements of machine learning, especially deep learning. We notice that although some deep learning methods such as CNN and CNN-RNN are used in infant cry research, the datasets are relatively small, and the architectures of models contains less than four layers. The main reason is that the deep models underfit the small training dataset and lead to poor performance. To take advantage of deep learning, large-scale databases with sufficient samples covering diverse changes within acoustic and prosodic features of different babies are in need.

Collecting data and labeling is a time-consuming process and requires skilled labors. Most databases used so far are self-recorded by authors and private to certain people or organizations. Although some online resources are available such as videos on Youtube, which is what Google AudioSet [48] links to. Most cry clips have no relevant labels, and many recordings are full of background sound containing both infant cries, adult speech, and other noises. To accelerate the progress of building automatic infant cry classifiers and smart cradle systems, and further to build robotic babysitter caregivers, effort to make public comprehensive well-structured and labeled databases are urgently in need. In addition, databases that contain samples from specific babies that can track their cries at different ages are needed. This type of database is essential to study the characteristic of infant cry along with their body development. Setting up recording devices on infants' cradles and recording real-time cry signals using cell phones by caregivers are the main methods used by data collectors. Baby cry translator mobile applications such as ChatterBaby [49] help predict infant cry reasons and made data collection easier. It will be more beneficial to the development of infant cry research if some newly collected datasets can be made public.

Poor connection between medical professionals and researchers. Cooperation between medical professionals and computer science researchers are vital to reach higher level of achievements in this area. Poor connection among them diminishes the ability of interdisciplinary mutual promotion. Research have proven that classifying infant cry signals is a non-invasive method and can be very helpful in some early disease diagnosis such as asphyxia, autism, cleft palate, and hypothyroidism, etc. But most of the pathological disease research with infant cry were performed before 2010, and the sizes of the datasets were very small. The difficulty of data acquisition may be the biggest obstacles in this research area. The ethical and legal issues involved in data collection process hinder the development of infant cry research. Cooperation between medical professionals and computer scientists may trigger some opportunities in this life saving research topic.

1.4 Contributions

Dealing with the challenges described above in infant cry research area, this dissertation aims to make effort and design novel methods to solve some of the problems and improve the performance of machine learning algorithms in this area, hence, to contribute to the development of infant cry research. Our contributions are as follows.

Data acquisition. We create Baby2020 database, which is a new dataset that contains more than 40000 cry samples. The cry samples were collected in natural real-world home or hospital environments. Parents, doctors, and nurses were instructed to place their smartphones close to the crying infants and record the crying signals using the recording applications on the smartphones. The length of each recording is less than 3 minutes. More than 100 healthy infants from newborn to 9-month-old participated in this project. The parents, doctors, and nurses capture the cries and perform the annotation at the ending of recording by including the label in the name of the recording file. Each recording is labeled by the monthly age of the infant, gender of the infant, and the reason of the cry. The cry reason annotation is based on caregivers' experiences and the real time situation when the cry occurs. For example, hungry cries most happen multiple hours after previous feeding time and feeding stops the cries. The wakeup cries are recorded when infants just wake up and the sleepy cries are identified by infants' face expressions and whether they fell asleep afterwards. The uncomfortable cries include some situations that cause discomfort such as wet diapers, mouth washing, bathing, and passive exercising, etc. Most raw audio recordings collected are in m4a format, some are in mp3 format, and a few are in amr format. We use FFmpeg [50] to convert all recordings into wav files. Each wav file is manually segmented into multiple cry samples using the Transcriber tool [51]. Manual segmentation ensures that each cry sample is a complete cry starting with a sound from the expiration phase and ending around a sound in an inspiration phase. If the one complete cry is less than 1 second, multiple complete cries will be included in a sample, otherwise there is no breath break in the sample. The durations of the segmented cry samples are between 1 second to 7 seconds.

Feature extraction. We thoroughly analyze the audio features in time domain, frequency domain, prosodic domain, and image domain, and propose to combine prosodic features with MFCC features. Our method generates weighted prosodic features and then combine them with acoustic features to classify asphyxiated infant cry effectively. We create a novel prosodic feature image as input for neural network models for infant cry reason classification. We propose a method extracting high level features by transfer learning with ResNet50 as the base model and using the features for graphical neural network (GCN) classification. We also propose to fuse the spectrogram features and mel-spectrogram features extracted from transfer learning model and decomposition model, respectively.

Classification models. Designing novel classification architectures is the most vital part of this research. We propose new approaches to improve the infant cry classification accuracy and disease identification: (1) we propose a non-invasive fast method of using infant cry signals with convolutional neural network based age classification to diagnose the abnormality of vocal tract development as early as 4-month age; (2) we propose an approach of generating hybrid feature set and using prior knowledge in a multi-stage CNNs for robust infant sound classification; (3) we propose an approach of using GCN to improve the infant cry reason classification; (4) we propose to use mel-spectrogram decomposition on infant cry classification and together with feature fusion to further improve the classification accuracy.

Some other ideas are also experimented in this research, but the results are not ideal. The experiments we perform include LSTM model, Graph Learning Convolutional Networks (GLCN) model, Graph Attention Network (GAT) model, signed GCN model, and some data augmentation techniques. More research will be explored on these methods in the future.

1.5 Organizations

Chapter 2 introduces the approach of asphyxiated infant cry identification with deep learning. Chapter 3 presents the age classification method to discover the association of the infant vocal tract development and cry signals and diagnose the abnormality of the infant vocal tract development. Chapter 4 introduces the multi-stage CNN method with hybrid feature set and prior knowledge approach for infant sound reason classification. Chapter 5 presents the infant cry classification method using graph convolutional neural network. Chapter 6 introduces the method of improving infant cry classification accuracy by mel-spectrogram decomposition and feature fusion, with conclusion and future work in Chapter 7.

2 ASPHYXIATED INFANT CRY IDENTIFICATION WITH DEEP LEARNING

Asphyxia is a respiratory injury that leads to a serious damage for infants. Early detection of asphyxia using artificially intelligent technology helps in reducing infant mortality rate when compared to traditional medical diagnosis, which is time consuming. In this study, we propose a novel method through generating weighted prosodic features combined with acoustic features to form a merged feature matrix to classify asphyxiated baby crying effectively. The weights of the prosodic features are trained at the frame level with labeled data and can be optimized using deep learning approach with neural networks. The novel merged feature matrix is established with both acoustic and weighted prosodic features. The matrix has good ability to capture the diversity of variations within infant cries, especially for asphyxiated samples. Our method has the benefits of keeping the robustness and resolution of the classification model simultaneously. The effectiveness of this approach is evaluated on Baby Chillanto database. Our method yields a significant reduction of 3.11%, 3.23%, and 1.43% absolute classification error rate compared with the results using single acoustic features, single prosodic features, and both acoustic and prosodic features, respectively. The testing accuracy in our method reaches 96.74%, which outperforms all other related studies on asphyxiated baby crying classification.

2.1 Asphyxiated Infant Cry Identification and Related Work

Asphyxia is one of the main causes of infant death. Traditional methods such as blood tests and other physical procedures are time consuming, so it can easily miss the best treatment time. It is very important to diagnose the asphyxia as early as possible. Baby crying is the only way of communication for infants. A lot of acoustic and prosodic information in baby crying expresses many cues for babies' basic needs and status. In recent years, many researchers focus on using machine learning approaches for baby crying analysis and diagnosing asphyxiated baby crying.

Commonly used methods for studying baby crying relate to speech processing either in time domain or in frequency domain. An infant cry recognition system was developed that classifies three types of cries from normal, deaf, and asphyxiated infants, and the classification accuracy was able to reach 86.06% [52]. SVM and its specialized versions were used in [53][54], and the classification accuracy reached 93.8% on a specific database. A system using MFCC and SVM was deployed via smart phones and wearable technology to diagnose asphyxia through automated analysis of infant crying [26]. A direct method of converting time domain asphyxiated audio samples to images as the input for neural networks achieved 94% accuracy [3]. A convolutional neural network was used to classify the asphyxiated infant crying and achieved 92.8% accuracy in a specific testing dataset [55]. The effects of both Linear Predictive Coding (LPC) and MFCC of infant cry signals were investigated with nearest neighbor approach and neural networks [21]. It is shown that fundamental frequency F0 is an essential feature for baby crying analysis and classification [56][57]. Recently, frame level features including MFCCs, pitch, and short-time energy were used for infant cry analysis and detection [58][59]. In addition, k-nearest neighbor algorithm was developed to classify the signal and added some constraints for pitch frequency and short-time energy to improve the efficiency. An approach with weighted linear prediction cepstral coefficients was proposed to minimize the sensitivities of both low-order and high-order coefficients for spectral slope and noise [60]. More recently, machine learning methods together with different level features for infant cry processing have been proposed. A multi-stage approach using deep neural networks for infant crying detection improved the robustness and performance compared with voice activity detection [11].

Using different audio features both in time and frequency domain together with deep learning approach addresses the fundamental work of baby crying analysis and processing. Challenges remain in these approaches, especially for asphyxiated baby crying classification tasks.

The only use of either converted time signals or MFCC/LPC coefficients without high level prosodic information leads to low discriminative ability of models for the diversity of variations in asphyxiated baby samples. The commonly used frame-level multiple features are capable for modeling acoustic or prosodic characteristics, but features are extracted separately and the valuable information about the physical and physiological condition is not related. Machine learning approaches with two level audio features improved the performance for baby crying identification on a specific database [56]. Whereas, all prosodic features are set with single equal weights, resulting in the degradation of discriminate ability to model the rich variations in baby crying samples.

In this work, we combine acoustic features with weighted prosodic features to generate a merged feature matrix for deep learning neural networks. The weights are trained at frame level with labeled data and can be optimized using deep learning approach with neural networks. The merged feature matrix is established with both acoustic and weighted prosodic features. Compared to using acoustic features or prosodic features solely, the merged feature matrix integrates both segmental and supra-segmental features to capture the diversity of variations within baby crying, especially for asphyxiated samples. Compared to traditional multiple feature combination method, the merged feature matrix differentiates contributions of prosodic information with different weights. Furthermore, exponential smoothing method can be used to indicate the overall change of the audio signal. Combining smoothed prosodic features with the weighted prosodic features before integrating with acoustic information achieves better result and improves the robustness ability because different contribution of each feature and variation of all features are represented. To achieve the same amount of accuracy increase, a single acoustic neural network may require huge amount of data. Our method overcomes the data size requirement of deep learning for baby crying classification by using weighted prosodic features and merged feature matrix. The novelty of this work includes the following: (1) we obtain the weights of the prosodic features by training the prosodic features on deep learning neural network; (2) we combine weighted prosodic features with acoustic features for asphyxiated baby crying classification and achieve better accuracy; (3) we use exponential smoothing method to catch the overall variation of the audio signal and combine it with the weighted prosodic features and acoustic features for asphyxiated baby crying classification. This method achieves the best accuracy; (4) we use a merged feature matrix together with neural network for classification at the frame level reserving the identity information of different prosodic features and achieves a good tradeoff between resolution and robustness of the model.

2.2 Acoustic Features and Prosodic Features

Acoustic features. MFCC is widely used in speech processing to present acoustic level information. The mel frequency cepstrum can be obtained through five steps: framing and windowing, fast Fourier transform, triangularly shaped band pass filtering, calculating the log of the filtered spectrum, and Discrete Cosine Transform (DCT). Hamming or other type of windowing is commonly used to avoid spectral leakage. The short-time Fourier transform of the signal converts the signal from the time domain to the frequency domain. It obtains the quasi-stationary short-time power spectrum F(f)=F{f(t)}. Then the frequency portion of the spectrum is mapped to the mel scale perceptual filter bank using M triangularly shaped ideal band pass filters equally spaced on the mel range of frequency $F(m)$. The central frequencies and widths are arranged according to a mel frequency scale as shown in (1):

$$\mathrm{M(f)} = 2595 * \log\,(1 + \frac{f}{700}) \qquad (1)$$

The total spectral energy $E[i]$ contained in each filter is computed and finally a DCT is performed to obtain the MFCC coefficients:

$$\mathrm{MFCC(l)} = \frac{1}{M}\sum_{i=0}^{M-1}\log\big(E(i)\big)\cdot\cos\left(\frac{2\pi}{M}\left(i+\frac{1}{2}\right)\cdot l\right) \quad (2)$$

for $\mathrm{l} = 0, \cdots, \mathrm{M} - 1$, where l is the order of MFCC coefficients. According to auditory perception model, MFCC coefficients are more robust than other coefficients such as LPC coefficients. In our study, we observe that the acoustic features of normal baby crying signals are quite different from the asphyxiated ones as shown in Figure 2.1. It indicates that the value range and tendency of acoustic features of normal and asphyxiated baby crying are different. The phenomenon can also be found in other frames, whose MFCC values are shown in different color lines in Figure 2.1, which illustrates a 12-order MFCC values in each frame with each different color representing one frame. The normal cry values are cyclic and relatively stable while the asphyxiated cry MFCC values have very high and very low values and it tends to go flat due to infant's lack of breath.

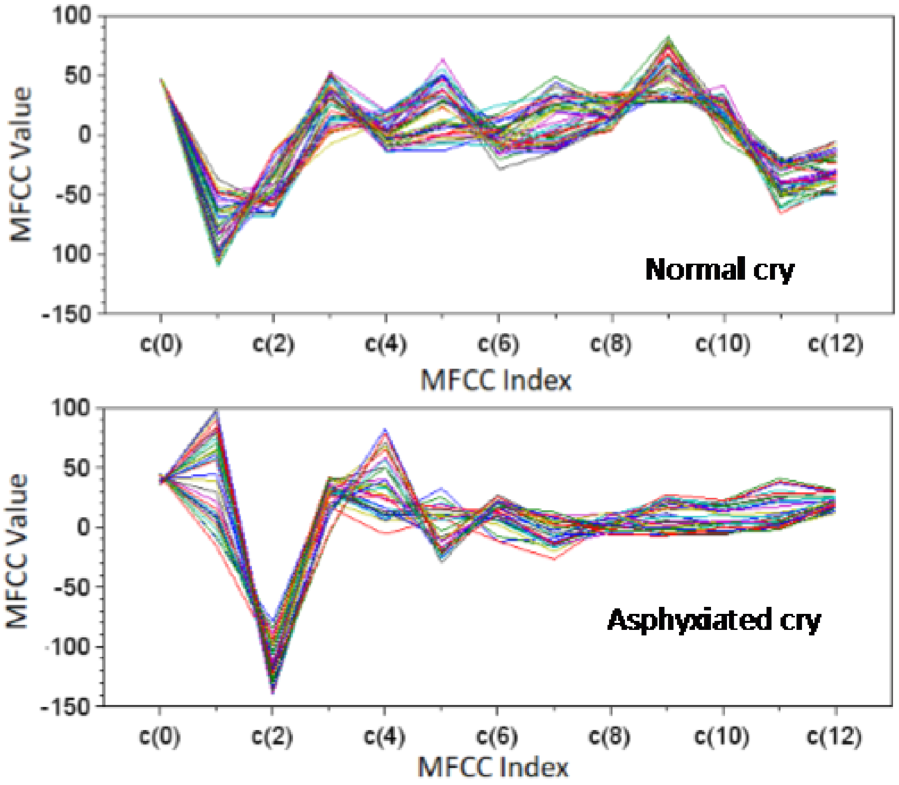

Figure 2.1:12-order MFCC features of normal and asphyxiated infant cry. Each colored line represents the values in each frame and there are total 96 frames.

Prosodic features. In speech recognition, the content of the audio is the key element that the machine algorithm needs to recognize. In human emotional recognition, audio signals are classified to different categories such as happy, sad, and angry, based on both the content of the speech and the style of the vocalization including the pitch, intensity, etc. Variations in intensity, fundamental frequency, formants, and duration are typical acoustic cues that carry prosodic information about infant crying and speech [5][56] . Infant cry is a type of sound in between the speech and other acoustic scene sounds such as animal sounds, and both the content and style of the voice represent important information of infants' needs. It is shown that the above prosodic features are efficient to identify the types of infant crying. For example, the F0 is a critical feature to represent different types of sound. Adult's F0 ranges between 85Hz and 200Hz while infant crying F0 is characterized by its high F0 ranging between 250Hz and 700Hz. We believe the prosodic features play an important role in different types of infant cry samples, especially when pathological cries are in the dataset. The pathological cries contain different acoustic features than normal cries due to infants' short of breath and other symptoms involved like coughing. In the following figures, we present the comparison of prosodic features between normal and asphyxiated crying signals. The left pictures on Figure 2.2 shows the waveform and spectrogram of a normal baby cry signal. The left pictures Figure 2.3, Figure 2.4, and Figure 2.5 show the pitch, intensity, formants graphs, respectively, from the same normal baby cry signal. The right picture on Figure 2.2 shows the waveform and spectrogram of a asphyxiated baby cry signal and the right pictures of Figure 2.3, Figure 2.4, and Figure 2.5 show the pitch, intensity, formants graphs, from the same asphyxiated baby cry signal.

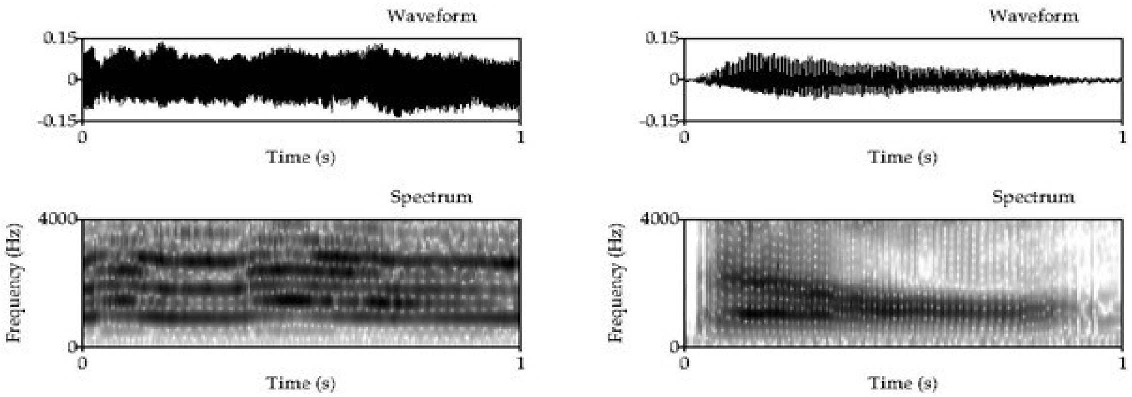

Figure 2.2:Waveform and spectrogram of a normal cry and asphyxiated cry signal.

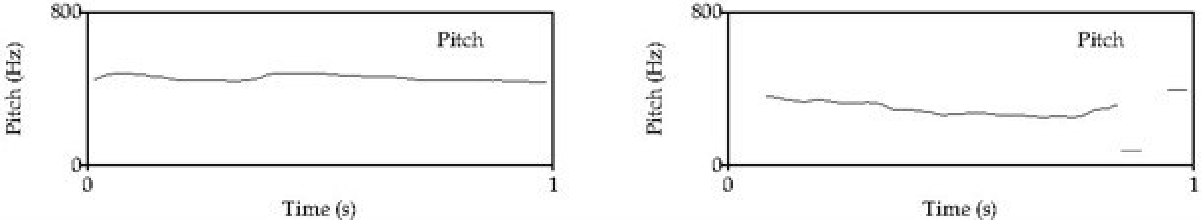

Figure 2.3:Pitch of the normal baby cry signal and asphyxiated cry signal.

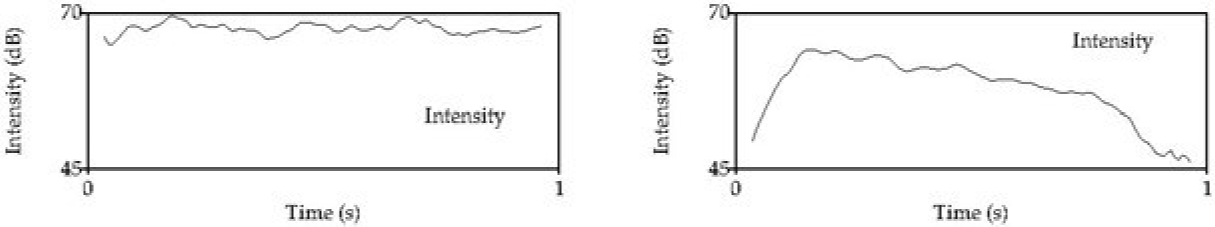

Figure 2.4:Intensity of a normal baby cry signal and asphyxiated cry signal.

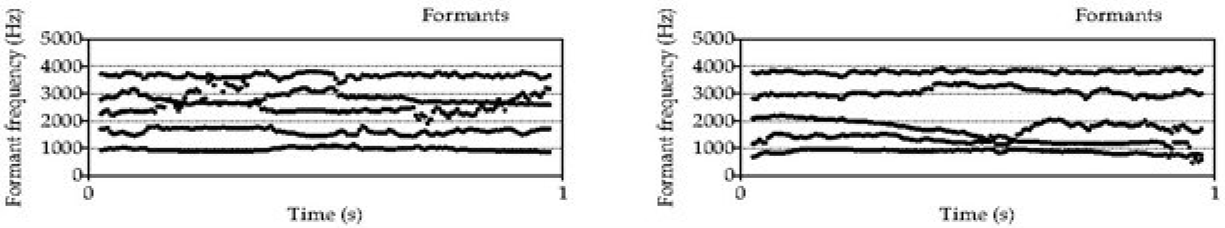

Figure 2.5:Formants of a normal baby cry signal and asphyxiated cry signal.

We investigate the above figures and find out the following: (1) the envelop of the intensity of normal baby cry signal is rhythmic and has cyclic changes due to the natural interruption, breath, and the energy is much higher than that of asphyxiated cry signal. The tendency of asphyxiated cry signal is gradient descent caused by the special crying movement of asphyxia; (2) the intensity is a good feature to capture the variant between normal and asphyxiate cry signals; (3) the pitch contour as well as the envelops of F123 formants from normal baby cry are relatively flat compared to asphyxiated cry signals. It means the tone of normal baby crying is consistent and the fluctuations are relatively shorter. Moreover, the value of F0 in normal infant spectrum is higher than that in asphyxiated infant as shown on Figure 2.2, which is accordant with the description in [57].

2.3 Merged Acoustic Features and Weighted Prosodic Features

Weighting prosodic features leads to improved resolution to model the difference for classification. To calculate the weight for each prosodic feature at frame level, we use a neural network. The architecture of the NN we use to generate the weights is presented in Figure 2.6. It is a fully connected deep learning neural network. Most researchers agree that a deep learning neural network often refers to a complex neural network whose Credit Assignment Path (CAP) is more than 2, where the depth of the CAPs is the number of the hidden layers plus one [61]. Deep learning architecture contains interconnected neurons in layers. The forward propagation process computes the values of each neuron based on the weights of each connection and bias for each layer. Activation functions are used in the hidden layers and in the output layer to perform classification or prediction. The back propagation is the key method to train the model to minimize the loss of the computation which makes the prediction as close to the expected results as possible. Deep learning outperforms in many applications in text, image processing, and audio classification. Our work uses a widely used deep learning neural network to classify the asphyxiated baby crying audio signals. The NN consists of two hidden layers containing different number of neurons in accordance with the size of features. The output layer is composed of two neurons identifying normal or asphyxiated crying. Training of the network is performed by minimizing the binary cross-entropy loss. The initial weights are defined with Gaussian distribution and the weights are trained with labeled transcriptions using data driven method. The activation function used in the hidden layer is tanh, which is a widely used function in deep learning. The softmax activation function is used for classification layer and the softmax cross entropy loss function is used in back propagation shown as follows:

$$Loss \ = -\frac{1}{\mathrm{N}}\sum_{i=1}^{2} y_i * logS_i \qquad (3)$$

where N is the total number of the training set, y is the one hot label, and the S_i is the output from the softmax function in the last layer.

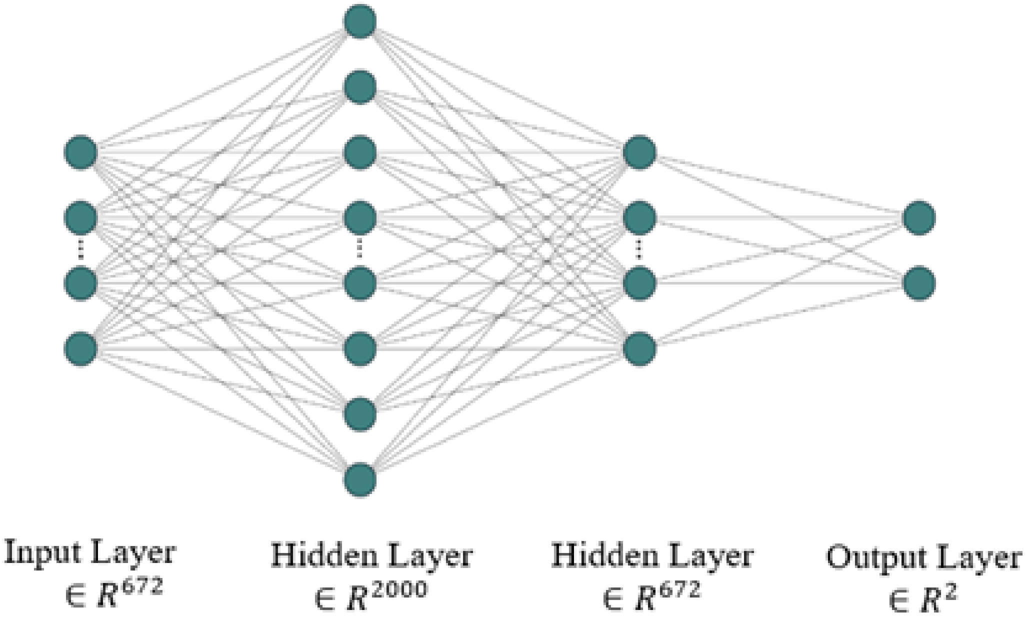

Figure 2.6:The neural network architecture for generating prosodic feature weights.

We select seven typical prosodic features in our study. Some of them, especially the F0, are used in other research, but they are not used together in any other literatures. To extract the MFCC features, framing and windowing are done to each audio files. Each audio signal contains 96 frames where each frame contains its MFCC data and prosodic features. The merged feature matrix is a combination of acoustic features and prosodic features. Acoustic features represent the segmental information of the audio signals. Prosodic features represent high level information of infant crying. C_0 represents the energy of the sound, which is transferred through the substance in a wave. The pitch frequency shows the highness or lowness of the sound. The intensity represents the strength, which is the power carried by the audio wave per unit area. The F0 is the lowest frequency and each formant representing the tone color of the sound corresponds to a resonance in the vocal tract. All these prosodic features represent the style of the cry sound. In our study, we concatenate the MFCC data and prosodic features, and feed them into the deep learning model for asphyxiated cry identification. It is believed that the merged feature matrix can also be fed into other models such as SVM. We use NN because it outperforms other models in recent asphyxiated baby crying detection research [3]. Figure 2.7 shows the architecture of the NN for asphyxiated baby crying classification where we use two thousand neurons in the two hidden layers. The input layer accepts the concatenated acoustic feature and prosodic feature with $m + p$ dimensions, where m represents the number of acoustic features and p represents the number of prosodic features. The output layer has two neurons representing the normal and asphyxiated cry.

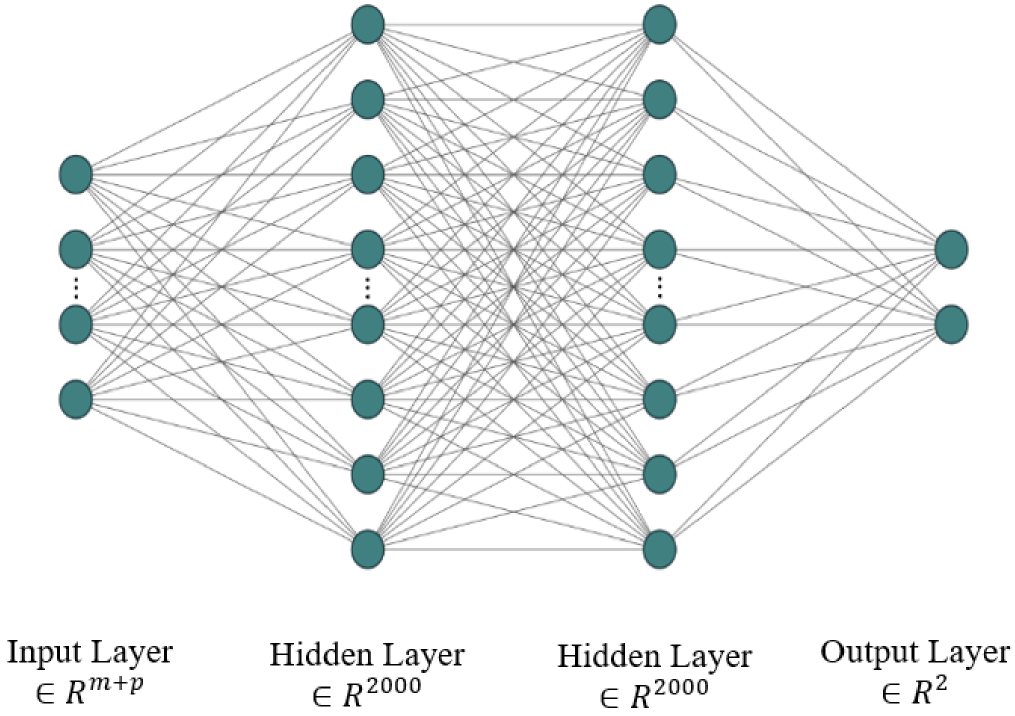

Figure 2.7:The neural network architecture for classification.

The merged feature matrix is fed into the classification NN. It is a combination of MFCCs and all weighted prosodic features. In each frame, the relevant prosodic coefficients are appended to MFCCs. The structure is shown in Figure 2.8, which illustrates that prosodic features of each frame are appended to each MFCC feature for the corresponding frame. Experiments show that merging features in each frame level performs better than appending prosodic features at the end of all MFCC data. The flowchart for generating merged feature matrix for deep learning network is shown in Figure 2.9. Each audio signal is used to generate the prosodic features and MFCC features. The prosodic features are fed into the weighted training NN, and the output is the weighted prosodic features. The MFCC features are then merged with weighted prosodic features and fed into the classification NN for classification.

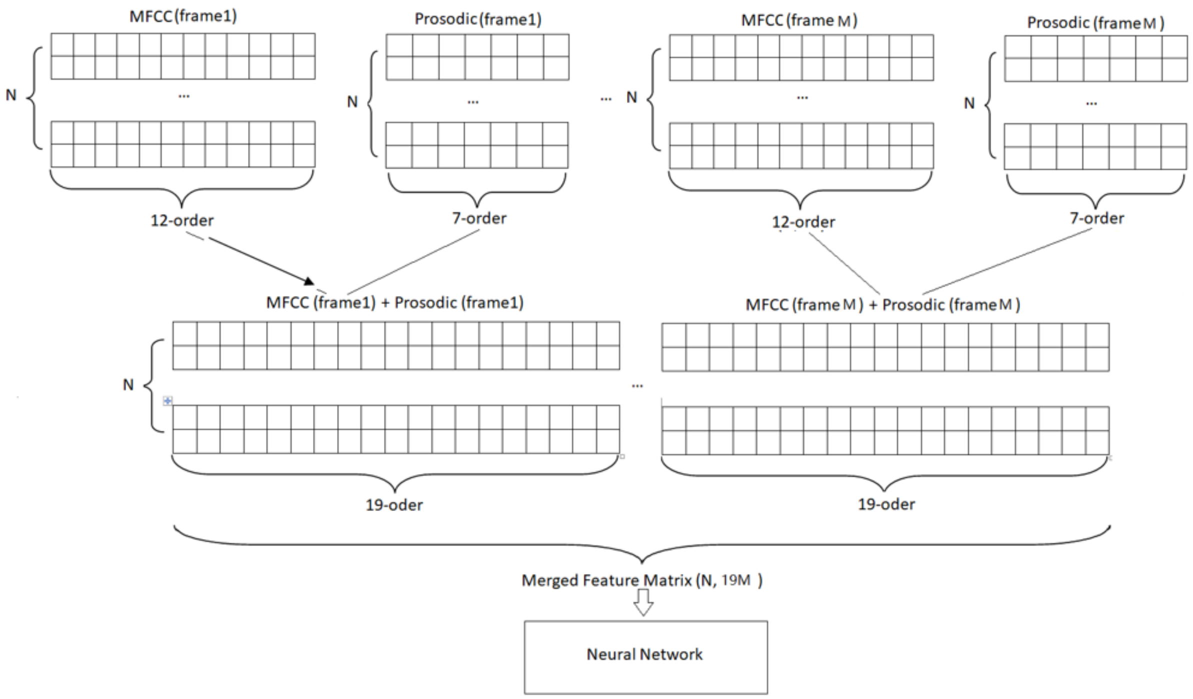

Figure 2.8:The architecture of merged feature matrix.

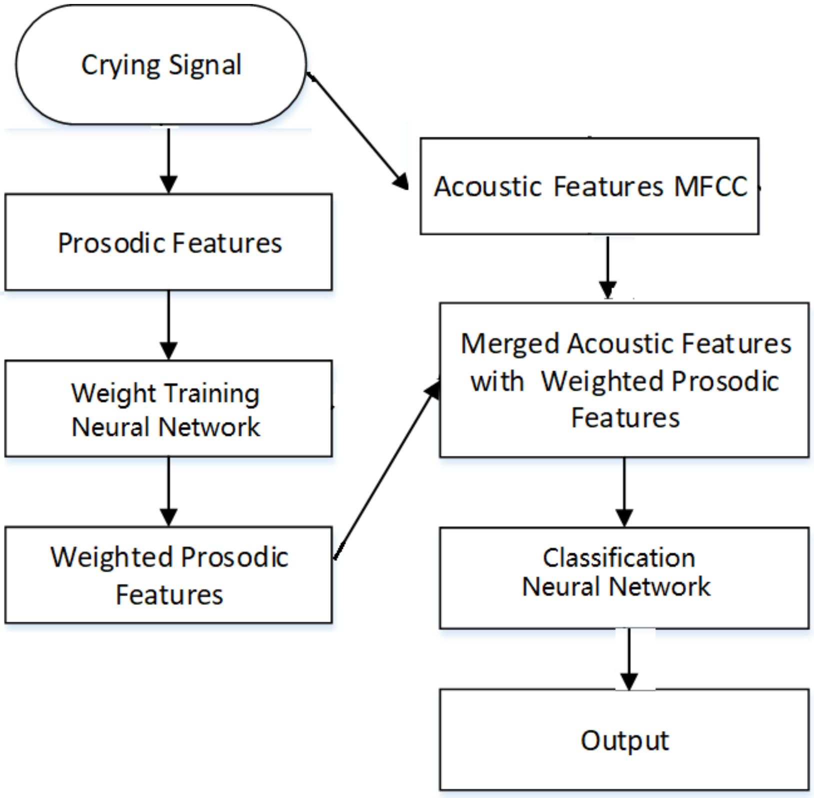

Figure 2.9:Flowchart of generating merged feature matrix.

The architectures of the neural networks for prosodic feature weights training and asphyxia identification are similar. The only difference is the composition of neurons. With the merged feature matrix, acoustic and prosodic features are included synchronously for neural network, and frame level prosodic features are attached with different weights to achieve a better classification.

We also use the first order exponential smoothing method to smooth variations in the crying signal which is time series data. The formula for smoothing is as follows:

$$S_0 = x_0$$

$$S_t = \propto x_{t-1} + (1{-}\propto) S_{t-1} \qquad (4)$$

where $\propto$ is the smoothing factor, x_{t-1} is the original data at the $t-1$ time point, and the smoothing output is written as S_t, which can be considered as the best estimate of the value of the next x.

2.4 Experimental Results and Analysis

2.4.1 Dataset and Experimental Setup

The effectiveness of proposed method is evaluated on the Baby Chillanto database. Baby Chillanto database is the most widely used infant cry database by researchers and was collected by National Institute of Astrophysics and Optical Electronics, CONACYT Mexico [44]. The babies being recorded are ranging from newborn to nine months of age and each sample is a one-second-long audio wav file. The database consists of 2268 baby cry samples in five categories as shown in Table 2.1. The recordings are segmented into 1-second-long wav files programmatically.

Table 2.1:Baby Chillanto database data samples

Category	**Asphyxia**	**Deaf**	**Hunger**	**Normal**	**Pain**
No. of Samples	340	879	350	507	192
Total	2268				

The architecture of neural networks used in the classification is described in Figure 2.7. Two thousand neurons are set for two hidden layers due to the complexity of NN structure. The learning rate used is 0.01 and the number of training epochs is 1000. The data structure of the merged feature matrix is presented in Figure 2.8. The hyperparameter values used in these experiments are the good ones obtained from a series of experiments that use different set of the hyperparameters. In the weight training neural network is shown in Figure 2.6, the number of neurons of the second hidden layer must match the number of features being trained because the weighted features from the second hidden layer signify the transformed weighted features. At the last epoch, the weighted feature matrix values are pulled out from the second hidden layer before it goes to the next layers for classification. The pulled-out matrix represents the weighted prosodic features. In the classification NN, the loss and accuracy with different epochs are investigated. After 1000 epochs of training, the training accuracy reaches 99.3%, which means the model fits the dataset well. We also use regularization and drop out to prevent overfitting.

Table 2.2:Results of experiments using acoustic features and prosodic features.

	Testing Accuracy	**Improvement over MFCC**
Acoustic (MFCC)	**93.63%**	---
Prosodic only	93.51%	-0.12%
MFCC + C0	93.60%	0.03%
MFCC + Pitch	93.26%	-0.37%
MFCC + Intensity	93.77%	0.14%
MFCC + F0	93.80%	0.13%
MFCC + F123	94.93%	1.30%
MFCC + F0123	95.13%	1.50%
MFCC + Prosodic	**95.31%**	**1.68%**

The results of using acoustic features, prosodic features, acoustic feature together with one type of prosodic feature, as well as the combination of acoustic and prosodic features are shown in Table 2.2. These different analyses show that the use of both acoustic and all seven prosodic features achieves the best performance of 95.31% compared with solely using acoustic and prosodic features of 93.63% and 93.51%, respectively. It gives 1.68% and 1.80% absolute classification error reduction. The results prove that prosodic features are efficient to capture the diversity of asphyxia variations compared with normal sounds. In addition, we can conclude that a single prosodic feature used independently together with an acoustic feature cannot improve the classification performance effectively. The reason may be because that the prosodic features are correlated and should be used together.

2.4.2 Results and Analysis

In Table 2.3, we first investigate the results of using the smoothing method compared with the combination with prosodic and the weighted prosodic features. It is shown that smoothing the prosodic features only gives slight improvement less than 1%. Using the merged feature matrix (weighted prosodic features included) yields additional 1% classification error rate reduction, which means the weights attached to prosodic features represent the different contribution of these features, hence improving the ability to differentiate asphyxiated crying from other signals. The weights assigned to each feature were obtained by training the original prosodic features separately in a deep learning network. At the last epoch of the training, the weighted features were pulled-up from the last hidden layer of the network. Moreover, the last result of combing merged feature matrix together with smoothing method achieved the best performance of 96.74% with respect to 93.63% of the baseline and 95.31% of using both acoustic and prosodic features. The results suggest that reserving the identity information of different prosodic features using weights together with variable information between adjacent sequences is beneficial for keeping the robustness and resolution of the model.

Table 2.3:Results of experiments using weighted prosodic features

Features	Testing Accuracy	Improvement over MFCC	Improvement over Prosodic	Improvement over both MFCC and Prosodic
Acoustic (MFCC) only	**93.63%**	---	0.12%	-1.68%
Prosodic only	**93.51%**	-0.12%	---	-1.80%
MFCC + Prosodic	**95.31%**	1.68%	1.80%	---
MFCC+ Smoothing MFCC	94.61%	0.98%	1.10%	-0.70%
MFCC + Smoothing Prosodic	96.10%	2.47%	2.59%	0.79%
WFM (MFCC+Weighted Prosodic)	96.31%	2.68%	2.80%	1.00%
WFM + Smoothing	**96.74%**	**3.11%**	**3.23%**	**1.43%**

Table 2.4:Accuracy comparison with other methods

Models	**Accuracy**
MFCC to SVM [26]	85.0%
MFCC to CNN [55]	92.8%
WaveForm Images to GoogleNet [62]	94.0%
WFM + Smoothing to DL	**96.74%**

Table 2.4 presents the performance comparisons between our method and other relevant experimental results with the same database. It is shown that the methods used in other studies only focused on acoustic features and reached 94% accuracy as the best result. The results prove that prosodic features are efficient to capture the diversity of asphyxia variations compared with normal sounds. The proposed weighted feature matrix together with smoothing on deep learning model, with 96.74% accuracy, outperforms all other related studies on asphyxiated baby crying classification.

2.5 Summary

The weights attached with different prosodic features at frame level are efficiently extracted and optimized using deep learning neural networks. The generated merged feature matrix keeps the good robustness ability of the model. In addition, the use of weights associated with prosodic features increases the distinction degree of the features. Hence, the matrix has a good ability to capture the diversity of variations within infant crying signals. It has been shown that this new merged feature matrix together with weighted prosodic features method provides a good balance of robustness and resolution. It gives an encouraging reduction of 3.11%, 3.23%, 1.43% absolute classification error rate compared with the results from using single acoustic features, single prosodic features, and both acoustic and prosodic features, respectively.

3 INFANT VOCAL TRACT DEVELOPMENT ANALYSIS AND DIAGNOSIS BY CRY SIGNALS WITH CNN AGE CLASSIFICATION

From crying to babbling and then to speech, infants' vocal tract goes through anatomic restructuring. In this study, we propose a non-invasive fast method of using infant cry signals with convolutional neural network (CNN) based age classification to diagnose the abnormality of vocal tract development as early as 4-month age. We study F0, F1, F2, spectrograms of the audio signals and relate them to the postnatal development of infant vocalization. We perform two age classification experiments: vocal tract development experiment and vocal tract development diagnosis experiment. The vocal tract development experiment trained on Baby2020 database discovers the pattern and tendency of the vocal tract changes, and the result matches the anatomical development of the vocal tract. The vocal tract development diagnosis experiment predicts the abnormality of infant vocal tract by classifying the cry signals into younger age category. The diagnosis model is trained on healthy infant cries from Baby2020 database. Cries from other infants in Baby2020 and Baby Chillanto database are used as testing sets. The diagnosis experiment yields 79.20% accuracy on healthy infants, 84.80% asphyxiated infant cries and 91.20% deaf cries are diagnosed as cries younger than 4-month although they are from infants up to 9-month-old. The results indicate the delayed developed cries are associated with abnormal vocal tract development.

3.1 Infant Vocal Tract Development and Related Work

Novice parents are excited to hear their newborn's first cry and care about their health the most. Infants express their needs, such as pain, discomfort, and hunger, by crying. It is shown that the postnatal development of vocal tract is associated with cry signals [4]. Diseases can lead to vocal tract development retardation and some healthy infants may also suffer from vocal tract development delay, which will affect infants or children speech development. Many studies explore the anatomical and acoustic features of adult vocal tract, only a few for children, and even less studies are for infants. Medical methods such as computed tomography (CT) and magnetic resonance imaging (MRI) techniques discover anatomical vocal tract development. It is known that an infant's vocal tract is not simply a miniature version of an adult's vocal tract [4]. Previous research has shown that infant vocal tract increases more than twofold in length and its geometric proportions also change [63]. The shape of the infant vocal tract changes from infancy to adulthood. In the process of the vocal tract development, the bend in the oropharyngeal region gradually forms a right angle, both larynx and the posterior part of the tongue descend and the distance between the soft palate and epiglottis is enlarged. Fitch and Giedd study MRI images of subjects from 2 to 25 years old and point out the first phase larynx descend occurs early in life and the second large descend, which is restricted to males, occurs at puberty [64]. Researchers discover that the acoustic features of infant vocalization reflect the changes in the vocal tract. Kent and Murray discover the ranges of both F1 and F2 frequencies increase as infant grow from 3-month to 6-month age [65]. Machine learning methods have been used in studying infant growth. Pruett et al. perform age classification on 6 versus 12-month old infants by functional connectivity magnetic resonance imaging (fcMRI) data to study brain and behavioral development using support vector machine [66].

Speech emergence and development is presumed to be dependent, at least in part, on the physical changes that the vocal tract structures undergo during development [63]. Speech development starts as early as infant crying. Robb et al. confirm the production of laryngeal constriction during infants' 3-5 months supports the notion that infants start to test and practice their phonetic production skills in the first several months of life [67]. Guiding infants, especially infants with tardy vocal tract development, to practice certain sounds and syllables as early as possible promotes their speech development. Finding out whether infants' vocal tract is developing normally as expected is vital for parents to take timely measures against the problems found. Compared to MRI with image processing, infant cry analysis and classification is non-invasive. The combination of signal processing and machine learning technologies on portable devices leads to simple and easy procedures, which can be performed without professionals.

In this study, we analyze diverse cry signals from different monthly age of infants and extract typical features such as F0, F1, F2, and spectrograms to investigate the relationship between anatomic changes of vocal tract and the characteristics of cries. Through this study, we discover that 4-month age is a key turning point of infant vocal tract growth. Moreover, we apply efficient neural networks to discover the pattern of the changes and diagnose the abnormality of the infant vocal tract by age classification. To our best knowledge, this is the first work of age classification via classifying infant cry signals. In this work, our major contributions include: (1) we propose using the characteristics of infant cry signals to evaluate the development of vocal tract development; (2) fundamental frequency (F0), formants (F1 and F2), and spectrograms of infant cries are investigated and related to the postnatal development of infant vocalization, and we show that 4-month age is a key turning point of infant vocal tract development; (3) an efficient convolutional neural network (CNN) approach is applied to infant cry binary monthly age classification to discover the trend of infants' vocal tract development and diagnose abnormal vocal tract development as early as 4-month-old.

3.2 Infant Vocal Tract Analysis

3.2.1 General Description of Infant Vocalizations

The shape of a newborn's vocal tract is more like a chimpanzee than a human adult [68]. The high position of the larynx and Hyoid bone causes the difficulty of controlling the tongue making infants unconducive to pronunciation. In the postnatal development phase, infants' vocal tract restructures gradually and develops mature speech ability accordingly. From infancy to adulthood, the length of the vocal tract develops from about 8cm to 17cm. It is shown that the vocal tract is nonlinear gradual and the growth curve of it can be fitted with fourth degree polynomial model [4]. Figure 3.1 shows the changes of the vocal tract from infants to adults. In the process of the vocal tract development, the bend in the oropharyngeal region gradually forms a right angle. Both infants' larynx and the posterior part of the tongue descend and the distance between the soft palate and epiglottis is enlarged. Hence, the infant vocalization is without resonant effect and the vowels sound within cries is nasalized resulting in quite different distribution of F1 and F2. Infants' speech development starts from crying. Previous research shows that healthy infants cry for around 1.75 hours per day by the second week of life, reach the peak of 2.75 hours by 6 weeks, and decrease gradually to 0.75 hours by 12 weeks [15]. Figure 3.2 presents a comparison of time and frequency domain between infant cry and adult speech. The areas included in the green rectangles are basic cries and the ones in the blue rectangles are cries ending with creaks. The spectrogram of basic cries is with clear bars, which is similar to that of vowels (green rectangles) shown in the right image of Figure 3.2. The creak at the end of a cry is like choking or interruptions with no vibration of vocal cord. Comparing to infants' cry signals, adults' speech signals are more complex and richer with energy, intensity, and formants changes representing a variety of the expressions. Because of infants' lack of full control of the vocal tract, they can only control the breath force from the lung to generate different types of cries for diverse purposes. The effect of movement of vocal cords is based on Bernoulli's effects [69]. Bernoulli effect determines the movement of the vocal cords to present such characteristics as the higher the flow rate, the lower the pressure. The flow rate increases when air comes from the lungs and passes through the narrow glottis. According to the Bernoulli's principle, the pressure at the vocal cords is reduced and the vocal cords are closed, and then the subsequent air opens the vocal cords again. Consequently, the sound is produced because the vocal cords keep moving up and down repeatedly. The harder the infant breathes, the faster the frequency of the opening and closing of the vocal cords and the greater the pitch and loudness of the sound.

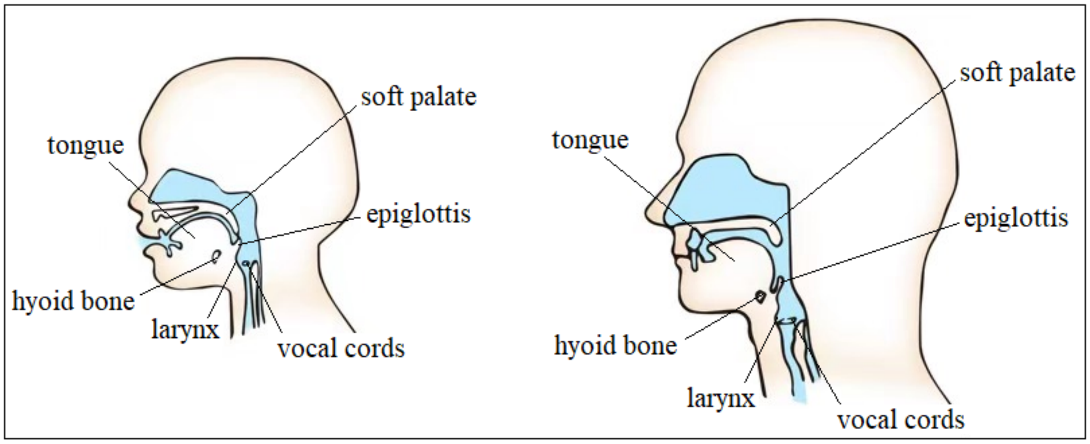

Figure 3.1:Comparison of vocal tract structure of newborn and adult.

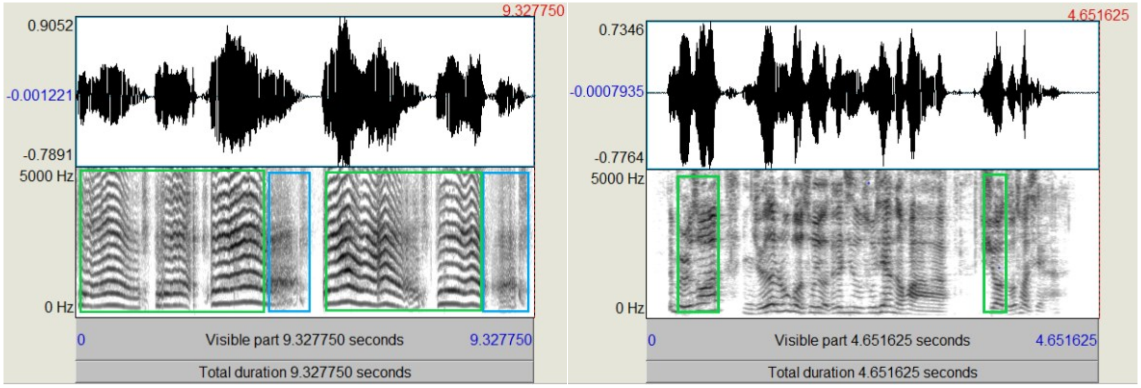

Figure 3.2:Waveform and spectrogram of infant cry (left) and an adult speech (right).

3.2.2 Analysis of Different Monthly Age Infant Cries

A spectrogram is a visual representation of an audio signals showing the amplitude of a particular frequency at a particular time. The spectrograms shown in Figure 3.3 illustrate the difference between a common 1-month cry and 4-month cry. In the earliest three months, an infant cry is characterized by its periodic nature, which alternates crying and inspirations. We can see that the clear harmonics are both in the lower frequency region below 3KHz, which covers more energy represented by lighter colors in the spectrograms. The harmonic structure becomes drastically weaker as the frequency increases for both spectrograms, but the 4-month spectrogram contains stronger energy in the low frequency than the 1-month spectrogram. Figure 3.3 right illustrates a gap, at around 0.85 second, which is the effect of glottis closure. When we listen to this audio signal, it contains an unclear “mama” sound. It means the 4-month age infant acquires the ability to close the glottis to form a certain level discontinuous vocalization during crying. It means the infant can generate discontinuous speech with different blocks in a whole articulation and is ready for the first word pronunciation.

The shape of vocal tract decides the resonant characteristics of its vocalizations. Infant cry signal is characterized by its high F0 within 250-700Hz compared to 85Hz to 200Hz of adult. The first two formants (F1 and F2) determine the vowel sounds, relating to the length and place of narrowing of the vocal tract and the F0 is corresponding to the increases in the length and volume of a vocal cord. F1 corresponds to the vertical height (high or low) of the tongue, F2 relates to the horizontal position (forward and backward) of the tongue [4]. In Figure 3.4, we plot F0, F1, and F2 using Praat [6] tool for a typical male infant cry for hunger at 1-month age and 4-month age. It is shown that the coefficients of F0 relating to vocal cord vary slightly between 1-month age and 4-month age baby. It indicates that the length and the volume of the vocal cord may not change much during the postnatal development of the first 3 months. On the other hand, values of F1 and F2 (F1=1921Hz vs 1470Hz, F2=4423Hz vs 2339 Hz) for 4-month vocalization increase significantly, which is in accordance with our previous analysis. Since F1 and F2 are strongly related to resonant cavity and the tongue, the increase of F1 and F2 indicates a great change of tongue location, oral, and nasal cavity extension for word pronunciation. To show the distribution of F1 and F2 along with the vocal tract development, we plot the scatter graphs for 0-month and 4-month cry samples from the same boy infant in Figure 3.5. Each graph contains the same number of values extracted from 100 cry samples. The horizontal axis is the F1 value, and the vertical axis is the F2 value. Figure 5 indicates that with the development of vocal tract, the average values of F1 and F2 increase. For example, F1 over 800Hz and F2 over 2400Hz are covered by more samples from 4-month compared to those of 0-month. In addition, the distribution of samples from 4-month is quite different with respect to 0- month samples. The standard deviation of F1 and F2 are increased with age. It indicates that with the improved ability of controlling the vocal tract, infants start to generate different formants changes representing different expressions at 4-month age, which is a turning point of the development of vocal tract.

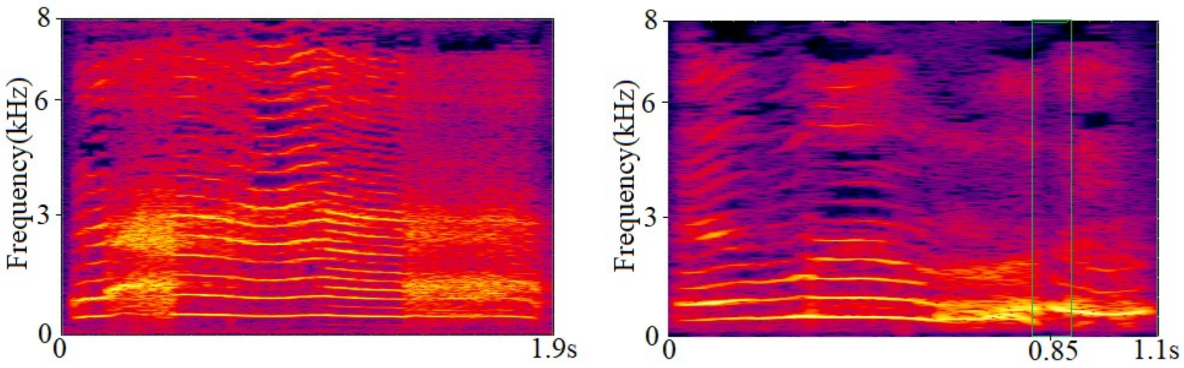

Figure 3.3:Spectrograms of a 1-month sleepy cry (left) and a 4-month sleepy cry (right).

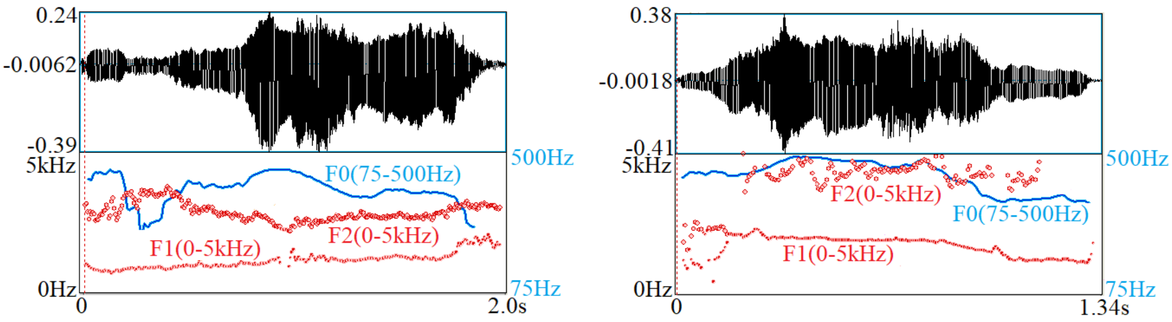

Figure 3.4:F0, F1, F2 comparisons of the same infant at 1-month (left) and 4-month (right). 1-month hungry cry has F0=442Hz, F1=1470Hz, F2=2339Hz; 4-month hungry cry has F0=457Hz, F1=1921Hz, F2=4423Hz.

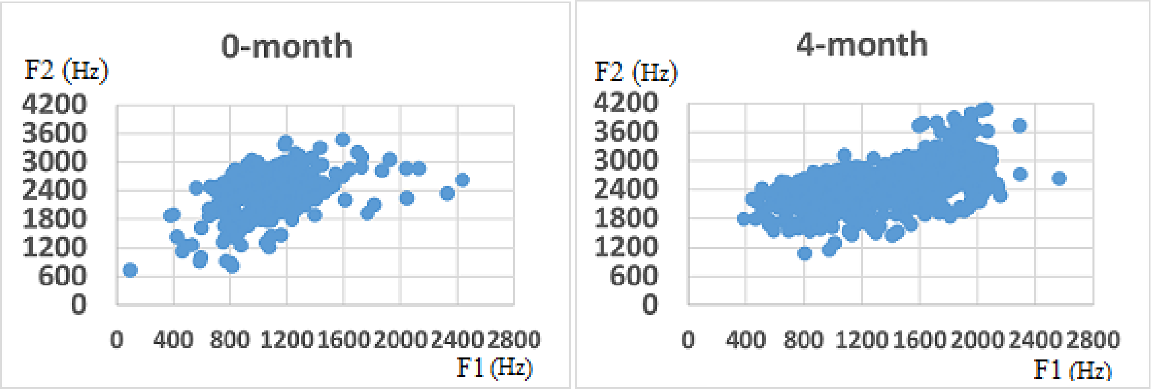

Figure 3.5:F1 and F2 distribution from a certain infant of 0-month (left) and 4-month (right).

3.3 Age Classification with Convolutional Neural Networks

Convolutional neural network is one of the deep learning models that is widely used in many research domains such as image classification, object detection, and signal processing, etc. Comparing to fully connected neural network, CNN is better at extracting the high-level features from the images with less parameters to train, which leads to better performance and less training time. In the training phase, each labeled image passes through a certain number of share-weights convolutional layers with selected filters, selected pooling that reduces the dimensionality, fully connected layers, and at last a softmax activation function that is applied to the last dense layer to generate a probabilistic value between 0 and 1 for classification. While training, the filter weights get updated by backpropagation algorithm to ensure the result matches the label of the image. In the testing phase, testing images pass through the trained CNN model to get the classification labels.

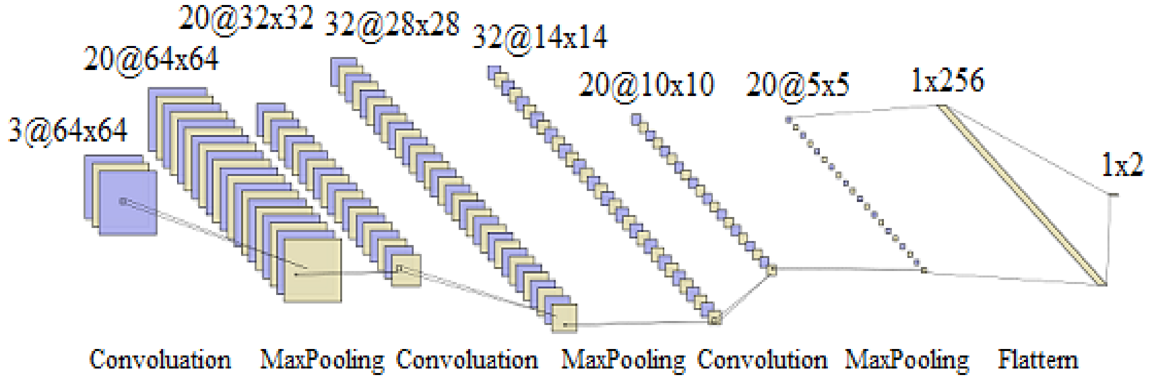

Figure 3.6:CNN architecture of our approach for age classification.

In recent years, CNN is used in infant cry reason classification and infant cry detection. The input images used include waveforms, spectrograms, and prosodic feature images. Research shows that the spectrograms perform the best on classifying the cry signals comparing to waveforms and prosodic line images [23]. In this study, we extract the spectrograms from the cry signals and feed them into the CNN model for age classification. The implementation of the CNN uses Keras framework with Tensorflow backend [70]. The architecture of our CNN model is illustrated in Figure 3.6. Spectrograms with size 64 × 64 are fed into the CNN model that contains three convolutional layers, each of which is followed by a max pooling. The network is then flattened into a 256-neuron fully connected layer, and the softmax is used in the last layer for classification.

3.4 Experimental Setup and Results

3.4.1 Datasets

We use subsets of our developing Baby2020 database and Baby Chillanto database in this work. The Baby Chillanto database is the same database as described in 2.4.1. We use the asphyxiated baby cries and deaf baby cries as the pathological testing sets. Baby2020 database is as described in 1.4. we use a subset of the Baby2020 database, and the label used in this study is infants' monthly ages. The reason we use two datasets is that Baby2020 database contains sufficient labeled samples from healthy infants and Baby Chillanto database contains pathological cries. We use samples in Baby2020 database for vocal tract development experiment and train the vocal tract development diagnosis model. Samples in Baby Chillanto database are used in the vocal tract development diagnosis experiment as the testing sets.

3.4.2 Experimental Setup and Results

A CNN binary classifier is used for infant cry age classification. The input images for all the experiments are spectrograms generated by Sound eXchange (Sox) software [71]. The original spectrograms generated from Sox are in 513×800 size and we resize them into 64×64. There are three convolutional layers, each of which is followed by a max-pooling layer in the model. 5×5 filter is used in each convolution layer and max pooling uses 2 ×2 filter with stride 2. The first convolutional layer uses "same" padding, and the default "valid" padding is used in other two convolutional layers. The first and third convolutional layers use 20 filters, and the second convolutional layer uses 32 filters. After the third max-pooling, the network is flattened into a 256-neuron dense layer. The ReLu activation function is used in each convolution layer and Adam is used as the optimizer. In the dense layer, the softmax activation function is applied for final classification. All models are trained with 100 epochs and all testing accuracies are the average of 10 runs.

Table 3.1:Vocal tract development experiments with binary monthly age classification

No. of Cries	Category1	Category2	Accuracy
	0-month	1-month	88.69%
		2-month	93.47%
		3-month	95.91%
		4-month	96.27%
		5-month	96.07%
		6-month	95.70%

3000	1-month	2-month	88.16%
		3-month	89.98%
		4-month	92.33%
		5-month	93.68%
		6-month	92.58%
	2-month	3-month	85.16%
		4-month	90.14%
		5-month	90.01%
		6-month	90.22%

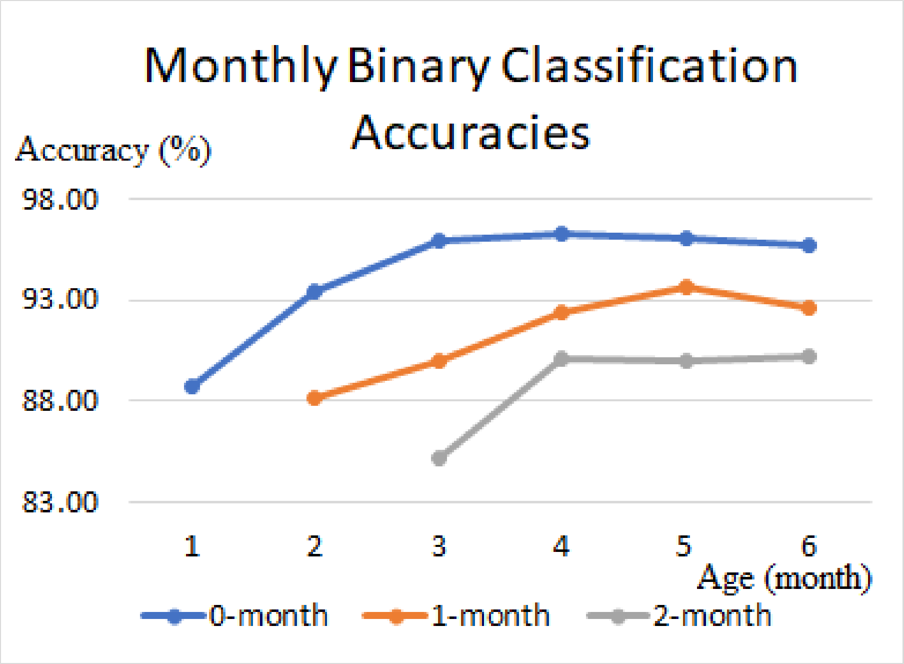

Figure 3.7:Line chart of monthly binary age classification.

Vocal tract development experiment. We design a binary age classification to identify the association between the infants' monthly growth and cry signal changes. In section 3.2 we analyze that 4-month age is a vital turning point of vocal tract development. This experiment is to confirm the ability of the model to recognize the vocal tract development changes. We use cry samples from infants between newborn (0-month) to 6-month. As shown in Table 3.1, each binary classification contains 3000 samples (1500 for each month) selected from multiple infants. The training and testing samples are from the same group of babies with 5-fold cross validation performed in this experiment. The labels of the samples are 0, 1, 2, 3, 4, 5 and 6 representing the monthly age of the infants. We generate different binary pairs such as 01, 02, 03, 04, 05, 06, 12, 13, 14, 15, 16, 23, 24, 25, and 26. For example, 01 means classifying 0-month samples from 1-month samples. Table 3.1 and Figure 3.7 give the classification results of all pairs. CNN achieves over 85% for all pairs indicating its strong ability to differentiate the monthly cries. As shown in Figure 3.7, when 0-month, 1-month, and 2-month cries are compared to cries in other months, the accuracies consistently increase as infants grow. The turning points arrive at 4-month where the accuracies stop increasing or frustrates through 6-month. The classifiers cannot differentiate the cries of 5-month and 6-month cries from 0-2-month cries better than the 4-month cries indicating the change of the infant vocal tract reaches a certain stable stage after the infants reach 4-month-old. In our multi-category classification experiment with 7 labels (0 month to 6 month), the accuracy is 66.75%, which also indicates that some signals in different months are too close to classify. In another experiment, we separate the male cries and female cries into two datasets and perform the binary age classification separately, we discover that both datasets show the same trend as the combined dataset shows, which is the accuracies stop increasing when infants reach 4-month-old.

Table 3.2:Results of vocal tract development diagnosis experiments

No. of Cries	Category1	Category2	Accuracy
3000 (training)	4-month cry	Younger than 4-month cry	---
1100 healthy cries (testing)			79.20%
879 asphyxiated cries (testing)			84.80%
340 deaf cries (testing)			91.20%

Vocal tract development diagnosis experiment. We perform age classification to diagnose abnormality of the infant vocal tract development. When an infant reaches 4-month age and his cry signals are classified as younger month cries by the model, it indicates that his vocal tract development is in growth retardation or related diseases may be involved. Our age classification model is to identify if a 4-month or older infant's vocal tract has developed normally as other healthy infants. In this experiment, the training set are from healthy infants containing 1500 samples in each category (0, 1, 2, 3 cries as one category and month 4 as the other category). The testing sets are from different groups of infants than the training set because the trained model cannot contain new patients' cries in real testing environment and new patients could be healthy or unhealthy infants. The three testing sets are 1100 healthy samples, 879 asphyxiated samples, and 340 deaf samples. We use the same CNN binary classification model described above. Table 3.2 shows the accuracies of the infant vocal tract development diagnosis age classification tested on healthy infants, asphyxiated infants, and deaf infants. For healthy infants, 79.20% testing samples can be classified into the correct age. 84.80% asphyxiated infant cries and 91.20% deaf cries are diagnosed as cries younger than 4-month cries. This means the model identifies pathological cries as younger age cries although the infants are up to 9-month-old. When infants' vocal tracts are affected by abnormality or diseases, their cry signals don't contain certain features that the corresponding age healthy cries do. The loss of hearing of deaf infants has great impact of the development of infants' vocal tract and asphyxiated infants show the delay of development of vocal tract. The experimental results show that pathological cries of older age infants are classified as younger age category due to the delayed developed vocal tract. Hence, our approach can be used to diagnose certain abnormality of the vocal tract development as early as 4-month age.

3.5 Summary

In this paper, we demonstrated that infant cry age classification with CNN is an efficient non-invasive method to diagnose abnormality of the vocal tract development of early age infants. The analysis of acoustic features from different monthly age infants shows that 4-month age is a key turning point of infant vocal tract growth. We have shown that the length and volume of vocal cord may not change much during the postnatal development of the first 3 months. F1 and F2 increase significantly within the early 4 months indicating a great change of tongue location, oral, and nasal cavity extension for word pronunciation. We applied CNN age classification on vocal tract development experiment and vocal tract development diagnosis experiment. The vocal tract development experiment discovered the pattern and tendency of the vocal tract changes, which matches the anatomical changes of the infant vocal tract discovered in research. The vocal tract development diagnosis experiment showed that infant cries can be used to identify if a 4-month or older infant's vocal tract has developed normally as other healthy infants. Our method achieved 79.20% accuracy for healthy infant cries on Baby2020 database, 84.80% asphyxiated infant cries and 91.20% of the deaf cries are diagnosed as cries younger than 4-month-old although they are up to 9-month-old in Baby Chillanto database.

4 INFANT SOUND CLASSIFICATION ON MULTI-STAGE CNNS WITH HYBRID FEATURES AND PRIOR KNOWLEDGE

In this chapter, we introduce our proposed approach of generating hybrid feature set and using prior knowledge in a multi-stage CNNs for robust infant sound classification. The dominant and auxiliary features within the set are beneficial to enlarge the coverage as well as keeping a good resolution for modeling the diversity of variations within infant sound. The novel multi-stage CNNs method work together with prior knowledge constraints in decision making to overcome the data sparse problem in infant sound classification. Prior knowledge either from rules or from statistical results provides a good guidance for searching and classification. The effectiveness of proposed method is evaluated on commonly used Dustan Baby Language Database and Baby Chillanto database. It gives an encouraging reduction of 4.14% absolute classification error rate compared with the results from the best model using one-stage CNN. In addition, on Baby Chillanto database, a significant absolute error reduction of 5.33% is achieved compared to one-stage CNN and it outperforms all other existing related studies.

4.1 Infant Sound Classification and Related Work

There are many reasons behind the baby crying such as pain, discomfort, and hunger, etc. Previous work shows that baby crying is a short-term stationary signal and only contains non-speech information [72]. In recent years, Priscilla Dustan showed that baby crying is a complicated procedure consisting of baby language and baby crying parts [9]. The Infant sound concept was proposed to cover both baby language and crying. In addition, Dustan's theory points out that baby language consists of five words associated with infants' five basic needs. Many researchers focus on using Dustan theory for baby sound analysis and processing, especially in testing the universal baby language hypothesis using speech recognition methods such as GMM, HMM, and CNN for classification.

Priscilla Dunstan discovered that babies use a proto-language with five “words” to express their needs [9]. It is shown that the proto-language is universal. Dustan translated the words as “Neh”=hungry; “Eh”=need to burp; “Oah”=tired; “Eairh”=low belly pain; and “Heh”=physical discomfort. Infants first express a certain need with one of these phonemes. If the need is not taken care of, they will soon start to cry. Mel Frequency Cepstral Coefficients (MFCC) together with KNN was used to achieve 79% for Dunstan five-word classification. Linear Frequency Cepstrum Coefficient (LFCC) was proven to be effective and the classification accuracy reached around 90% on limited testing data [9]. Other researchers collected the raw data using Dustan definition and used MFCC with K-nearest neighbor classifier to obtain around 70% accuracy [73]. An automatic method for infant cry classification proposed in [72] used GMM-UBM as well as i-vectors modeling methods to achieve average accuracy around 70%. A method of converting infant crying audio samples to spectrogram images as the input for neural networks achieved 89% accuracy [9]. In this method, a Convolutional Neural Network (CNN) was used to classify the five “words” with a fixed specific testing data. More recently, machine learning methods together with prosodic features for infant cry processing have been proposed. It is shown that fundamental frequency F0 is an essential feature for baby crying classification [56][57]. Recently, frame level features including MFCCs, pitch, and short-time energy were used for infant cry analysis and detection [59].

Using different features or spectrogram images together with machine learning approaches addresses the fundamental work of infant crying classification. Challenges remain in these approaches, especially for infant speech classification tasks. Infant speech is different from infant crying. Baby crying is considered more stationary than the speech since infant cannot fully control the vocal tract. Applying speech recognition approaches leads to inferior performance due to the difference between speech and non-speech signals. Infant sound is a time sequence with four steps, including infant speech and crying [72]. The only use of either speech coefficients such as MFCC/LPC/LFCC or converted spectrogram images as input for the machine learning models is not able to capture the diversity of variations within the sound produced by different age infants. In addition, data sparse is a severe problem in the task. The size of Dustan infant speech database is small. The total amount of transcribed samples is very limited for robust neural network classification structure. Automatic infant speech classifier with CNN approach improved the performance for infant speech classification on the Dustan database [9]. Whereas the testing environment is set to be very specific and both test set, and configurations are fixed strictly. It is essential to generate an efficient approach for processing both infant speech and infant crying under limited data samples.

In this study, we propose generating hybrid feature set and using prior knowledge to guide the training of a multi-stage CNNs model for robust infant sound classification. We investigate the detailed difference between infant speech and crying both in time domain and frequency domain. We compare infant speech and crying to traditional normal speech to discover the hidden characteristics in the sound. We establish dominant and auxiliary features to form a hybrid feature set to take advantage of different discrimination ability of each CNN. Compared to using traditional features solely, the hybrid feature set uses the auxiliary features as supplement to capture the diversity of variations within infant speech and crying. Furthermore, the prior knowledge either from rules or from statistical results is used to guide the multi-stage CNNs classification. With the use of prior knowledge and hybrid feature set, the searching space of CNN classification is constrained so the system is robust under limited data samples. The effectiveness of proposed method is evaluated on commonly used both Dustan Baby Language database and

Baby Chillanto database. Our method overcomes the data sparse challenge for both infant speech and infant crying classification. In this paper, our major contributions include the following: (1) a novel approach of generating hybrid feature including prosodic feature images; (2) we introduce a method to use different feature images to feed into multiple CNN models for robust classification; (3) we propose a multi-stage CNNs model that can take advantage of discriminative ability of each individual model; (4) we use prior knowledge in decision making to guide the training process in the multi-stage model.

4.2 Feature Extraction

Dustan translated the words as "Neh" means hungry; "Eh" means need to burp; "Oah" means tired; "Eairh" means low belly pain; and "Heh" means physical discomfort. To see the differences of five words of infant speech, we plot both spectrograms and prosodic feature lines including F0, intensity and F12345.

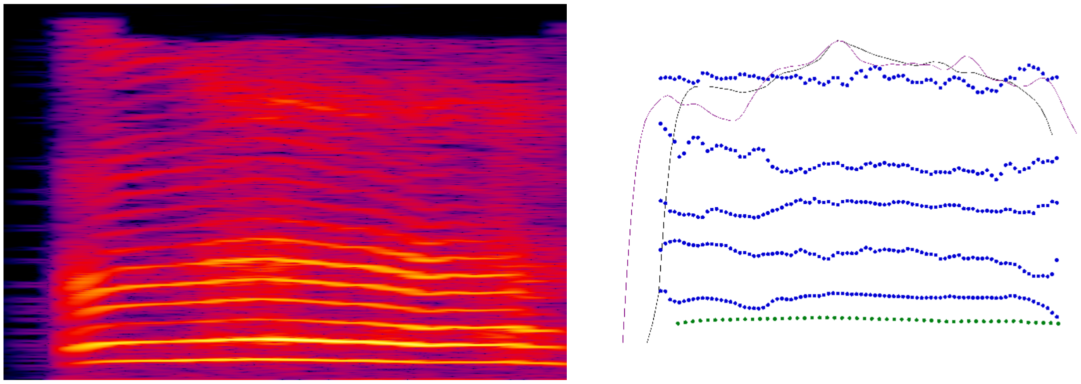

Figure 4.1:The spectrogram and prosodic lines for infant word of "Neh".

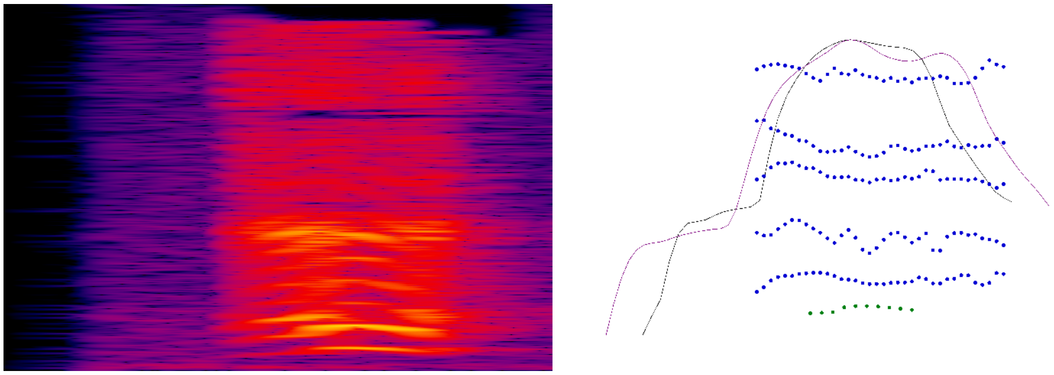

Figure 4.2:The spectrogram and prosodic lines for infant word of "Eh".

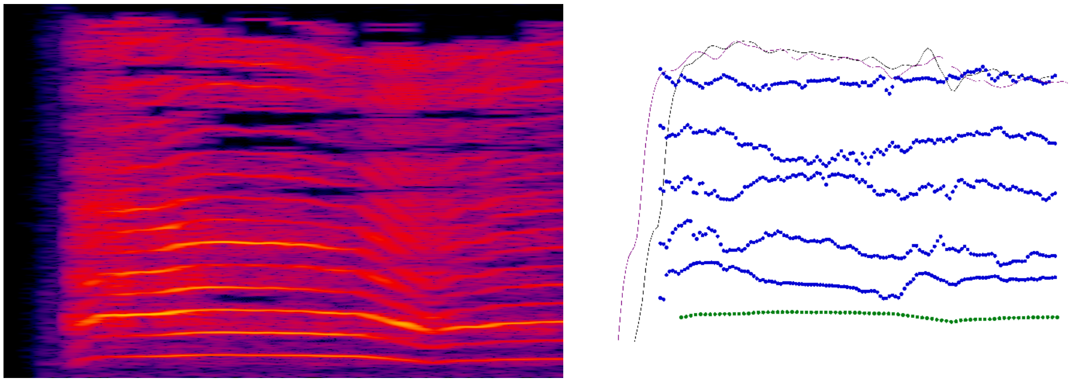

Figure 4.3:The spectrogram and prosodic lines for infant word of "Oah".

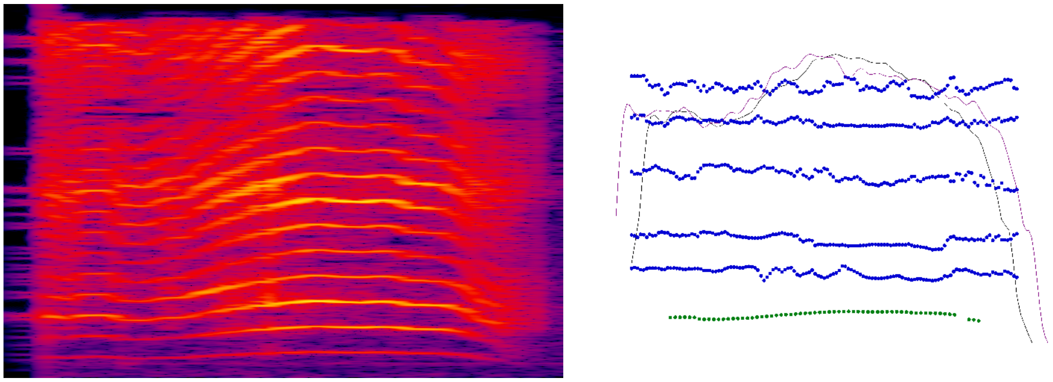

Figure 4.4:The spectrogram and prosodic lines for infant word of "Eairh".

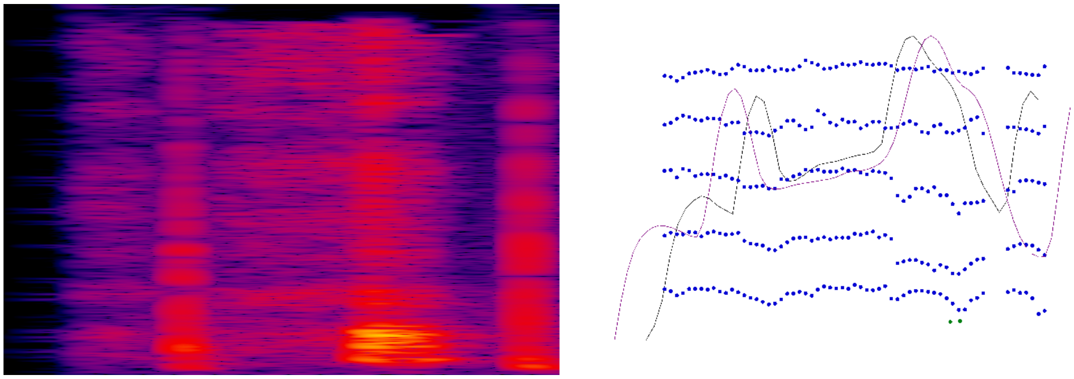

Figure 4.5:The spectrogram and prosodic lines for infant word of "Heh".

We investigate the above five figures and find out the following: (1) the energy shown in the spectrograms at different frequency of five infant words is quite different. For example, the word "heh" has the lowest energy in all frequency band while "eairh" has the highest, which is in accordance with the infant status of physical discomfort and stomach cramp. In addition, the figures also present that infant can pronounce vowels; (2) prosodic features have good resolution to characterize the difference within infant sound. For instance, the envelop of the intensity of "eh" is approximate rhythmic and has cyclic changes due to the reason of the need of burp. The tendency of "oah" is gradient descent caused by tiredness of the infant. The F0 as well as the envelop of formants have good discriminative ability to classify five infant words; (3) the spectrogram is a good feature to describe the characteristics of infant sound signals. It is assumed that both acoustic and prosodic information are included in spectrograms. The combined prosodic features are good auxiliary features with fine resolution to describe the variations hidden in the infant sound.

4.3 Multi-stage CNNs Model with Hybrid Features

Data limitation is always a challenge for neural network classification tasks. The search space constraint approaches are effective for better performance. Our multi-stage CNNs model uses the hybrid features set and applies the rule-based or statistic-based prior knowledge during the decision-making process. The searching space of CNN classification is narrowed, and hence the performance of classification is improved.

4.3.1 Hybrid Feature Multi-stage CNNs Model

For speech recognition tasks, phoneme units are commonly used. Acoustic coefficients are concatenated and trained at frame level by CNN based classification structure [74]. On the other hand, infant speech and crying are different regarding as non-speech signals. It is not confirmed that phoneme-based structure is suitable for classifying such non-speech signals. Inadequate hand labeled transcriptions cannot support robust model training under CNN framework. Usually, different feature sets have different discrimination ability for different target (e.g., Dunstan infant speech or different type of crying). So, we analyze the confidence measure of each feature set with all test samples and the corresponding model. We calculate the confidence measure of each feature i to identify that feature i has higher accuracy on target k, but not strong in other targets. Here $i = 1 \cdots N,\ N$ is the number of feature sets. $k = 1 \cdots M, M$ is the total number of the categories. Based on the order of the classification accuracy on each target k using each feature set i, we can consider using a N-stage classifier to combine the ability of all N feature sets. In the N-stage classifier, each feature set is only used in its corresponding model to classify the categories that has higher confidence.

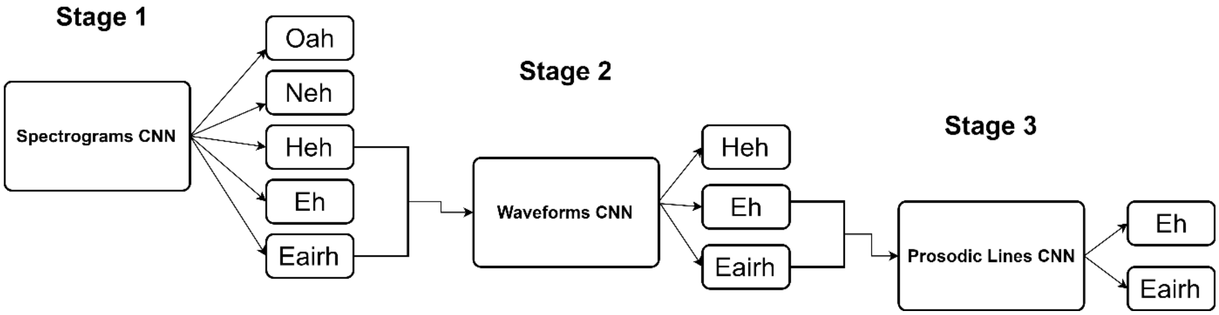

Figure 4.6:Hybrid-feature multi-stage method.

Figure 4.6 shows the hybrid feature multi-stage method used in Dustan baby language classification. We use the spectrogram CNN model to perform 5-category classification. The confidence measure, the accuracy for each category, is calculated. The best two categories will not involve in the second stage. In the case of Figure 4.6, the Oah and Neh can be classified well in the first stage. Then, Heh, Eh, and Eairh's waveform images will be fed into the second stage CNN for 3-category classification. In the third stage, only the Eh and Eairh's prosodic line images will be used in the third CNN for binary classification since the Heh sound has been classified relatively well in the second stage. The confidence measure can be the classification accuracy of each category. We use a multi-stage classifier to find such comparative advantages among different feature sets.

4.3.2 Prior Knowledge Generation

Prior knowledge can be defined either from statistic method or from rules. The statistic-based knowledge is used to decide which model should be used to classify the relevant categories. Other rule-based knowledge such as a vowel sound should be easier to differentiate from a consonant sound is also used in the decision-making process. Due to limited data in our task, an efficient multi-stage classifier is performed to see if we can find such comparative advantage among different feature sets.

Different feature sets have different discrimination ability for different target. Hence, we train and validate the individual spectrogram CNN, waveform CNN as well as prosodic lines CNN separately to obtain different classification accuracy. It indicates that spectrograms predict certain signals more accurately while waveforms can predict another type of signal better. Similarly, this applies to other input images as well. The accuracy can be regarded as confidence measure for prior knowledge. We use the calculated statistic-based prior knowledge to decide which model should be used to classify which categories. In addition, rule-based knowledge from linguistic information were added as another prior knowledge. For example, high energy sound should be easier to differentiate from low energy sound, a vowel sound should be easier to differentiate from a consonant sound. They are used to decide which category should be classified together with relevant categories. Prior knowledge is integrated into multi-stage CNN classification task as follows: (1) first stage: use the best network, the spectrogram model, to perform five-category classification. The classification accuracy of each category is calculated. The weakest three categories will be classified in the following stages; (2) second stage: use waveform model to perform three-categories based on the confidence measure calculated in the first stage. In the case of Figure 4.6, the waveform model is selected because it can recognize "eairh" sound better than other two models. In the case of Baby Chillanto database, the binary classification is decided based on the prior knowledge by analyzing the differences among images. The high energy sounds pain and hunger should not be classified together but they can be classified very well separately with another low energy sound such as asphyxia; (3) third stage: use the last model to classify the last two types of sound.

With the mixed feature set and the use of prior knowledge during the decision-making process, the searching space of CNN classification is constrained, hence the system is more robust under limited data samples. Meanwhile, with the guidance of prior knowledge, the following steps of the classification can be divided with different discrimination.

4.4 Experimental Results and Analysis

4.4.1 Datasets

The effectiveness of proposed method is evaluated on both Dunstan Baby Language database and the Baby Chillanto database for infant sound classification. Baby Chillanto database is described in section 2.4.1. As shown in Table 4.1, Dunstan Baby Language database consists of 315 wave files, sampled at 16KHz, with a variable length between 0.3 to 1.6 second. Each utterance is a word of infant speech corresponding to one of the five “Dunstan words” transcribed by Dunstan herself or other Dunstan certified experts [9].

Table 4.1:Dunstan Baby Language database data samples

Category	Neh(Hungry)	Oah(Sleepy)	Eh(Need burping)	Eairh(Belly Pain)	Heh(Discomfort)
No. of Samples	56	106	55	37	61
Total	315				

4.4.2 Results and Analysis

The implementation of CNNs uses Keras framework with Tensorflow backend [70]. The architecture of spectrogram CNN is shown in Figure 4.7. The convolution layer uses twenty 5 × 5 filters, 2 × 2 pooling size with stride 2 are used in the max pooling layer. In the waveform model, we use five 5 × 5 filters instead to reach relatively higher accuracy. Five 3 × 3 filters are used instead in the prosodic line model for higher accuracy. Other configurations remain the same in the waveform model and prosodic line model. 100 epochs were performed in the training process.

For both datasets, we perform five-fold cross validation classification due to the limitation of available samples. We use 80% samples for training and 20% samples for testing in all the experiments. Spectrograms are generated by the Sound eXchange (Sox) software [71]. The waveforms and prosodic features images are extracted using the Praat tool [6]. The default parameters are used when extracting the waveforms and the prosodic feature lines including C0, pitch, intensity, and formants. All images extracted are then resized into 60 pixels in height and 90 pixels in width for unisize input. The performance of each type of image on each sound is shown on Table 4.2 and the results of using dominant feature and auxiliary feature set separately, as well as merging the three models as a CNN late fusion model are shown in Table 4.3.

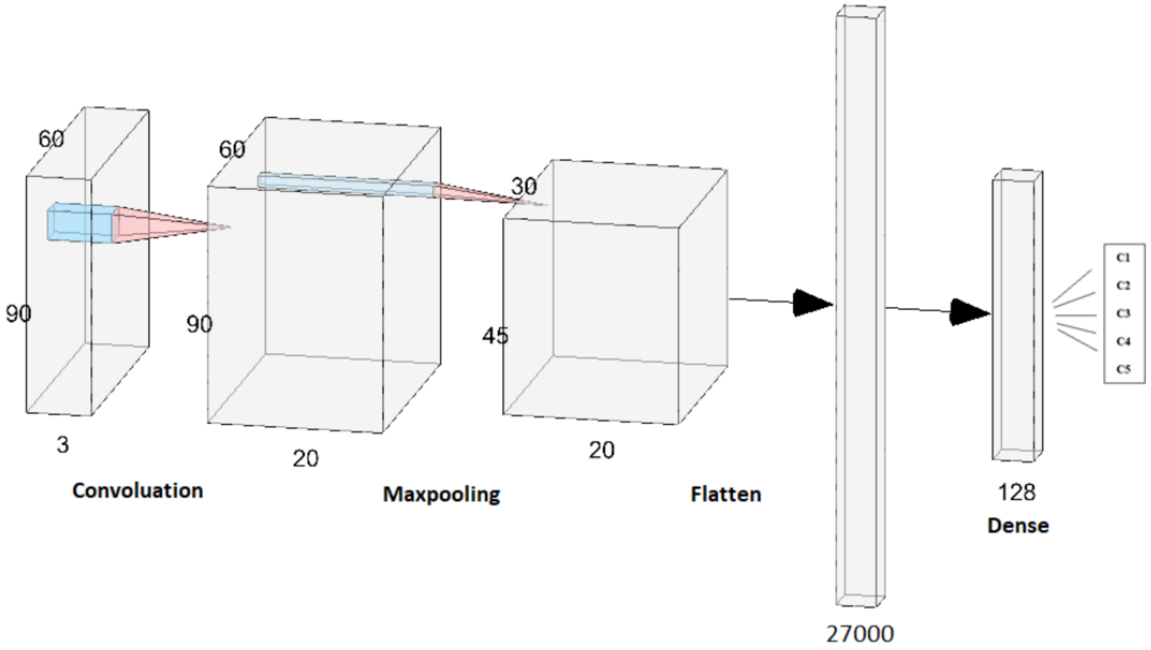

Figure 4.7:CNN architecture of the baseline spectrogram model.

The analysis shows that the use of spectrogram as a sole feature for five infant words classification achieves the best performance of 84.08% compared with solely using waveform and prosodic feature set of 53.84% and 69.33%, respectively. It proves that spectrogram includes both acoustic and prosodic information and is suitable to be the dominant feature for infant speech classification under small data size. In addition, it is seen that merging spectrogram, waveform, and prosodic features models cannot improve the performance. The reason lies in the fact that the short stationary of infant speech is different from normal speech, resulting in not obvious changes in acoustic and prosodic features.

Waveform and prosodic lines have distinguished ability to model some certain infant speech as illustrated in Table 4.2. We observe that different types of images are good at classifying different types of sounds. For example, the “Eairh” sound is the worst to identify in the spectrograms model, but the waveform feature can do it better; the spectrogram and waveform models both cannot recognize the “Heh” sound well, but it has the 2nd best classification accuracy in the prosodic lines model. Therefore, we can use the knowledge obtained above to generate hybrid-feature set and use the multi-stage approach to achieve a better performance. The experiments yield promising improvements. The five-fold cross validation classification accuracy is 88.22%, which improved 4.14% compared to using the spectrogram model. Compared to traditional CNN classification method, our multi-stage approach makes a pre-separation of searching space with prior knowledge at each step, resulting in better performance with limited data trained model. In addition, hybrid feature set consisting of dominant feature as well as auxiliary features with different discriminative ability provides different level of resolution for better classification.

Table 4.2:Accuracy of each category in descending order (Dunstan Baby Language)

Input	Best Accuracy	2nd	3rd	4th	5th
Spectrograms	Oah	Neh	Eh	Heh	Eairh
Waveforms	Oah	Neh	Eairh	Heh	Eh
Prosodic Lines	Oah	Heh	Neh	Eh	Eairh

Table 4.3:Results of using different feature combinations (Dunstan Baby Language)

Feature Combination	Accuracy	Relative changes to Spectrogram
Spectrograms to CNN	84.08%	0%
Waveform to CNN	53.84%	- 30.24%
Prosodic features to CNN	69.33%	- 14.75%
Three CNNs late fusion model	83.48%	- 0.6%
Hybrid-feature multi-stage model	**88.22%**	4.14%

Table 4.4:Accuracy of each category in descending order (Baby Chillanto)

Input	Best Accuracy	2nd	3rd	4th	5th
Spectrograms	Deaf	Asphyxia	Normal	Hunger	Pain
Waveforms	Deaf	Hungry	Normal	Pain	Asphyxia
Prosodic Lines	Deaf	Asphyxia	Hungry	Normal	Pain

Table 4.5:Accuracy comparison with other models on Baby Chillanto database

Input Features	**Method**	**Accuracy**
Spectrograms	CNN	89.77%
Spectrograms	Transfer Learning CNN, SVM Ensemble	90.80%
MFCC and LPCC	MLP and Radial Basis Function Network	93.43%
Spectrograms, Waveforms, Prosodic	**Hybrid-feature Multi-stage method**	**95.10%**

We further evaluate our proposed approach on Baby Chillanto database, which is an infant cry database. The reason to evaluate our approach on two databases is to show that our method is effective on both infant speech and cry signal classification tasks. The size of the database is 6 times larger than the Dunstan database, but it is more unbalanced. Table 4.5 illustrates the results of using multi-stage classification on Baby Chillanto database comparing to other methods. As shown in Table 4.4, we observe that the three networks are good at recognizing different types of crying signals. For example, Spectrograms are good at differentiating the normal crying, the waveform does better job recognizing the hungry crying, and the prosodic feature lines images can classify asphyxia crying well while the waveforms cannot. To take advantages of all features and all models, we apply our hybrid multi-stage approach to classify the baby crying signals. Five-category spectrogram model is used to classify deaf and normal sounds because they perform the best in the spectrogram model. Binary waveform model is used to classify hungry and asphyxia, and prosodic lines are used to classify pain and asphyxia sound in the binary classification model. As shown in Table 4.5, the methods used in other studies reached 93.43% accuracy as the best result. The proposed hybrid feature with multi-stage CNNs achieves 95.10% accuracy and it outperforms all other related studies on the Baby Chillanto database. The results suggest that reserving the specific identity ability from different features to form hybrid feature set is also efficient in infant crying classification tasks. Even more, the hybrid feature set can be used to get prior knowledge information as confidence measure, which is beneficial for providing constraints in searching and classification. Our method is effective in balanced database and unbalanced database and can be extended to other acoustic event databases.

4.5 Summary

Our approach of using hybrid feature set with multi-stage CNNs performs well in infant sound classification. Infant sound is a complicated procedure including infant speech and crying, which have different acoustic and prosodic characteristics. Our experiments show that different features have different discrimination ability to model the diversity of variations within infant sound. To take advantage of the best ability of each feature, the use of dominant and auxiliary features is beneficial to enlarge the coverage as well as keeping a good resolution. We used multi-stage CNNs method together with prior knowledge constraints in decision making to improve the classification accuracy in infant sound classification. Prior knowledge information either from rules or from statistical result provides a good guidance for searching and decision making. The effectiveness of our method was evaluated on commonly used both Dustan Baby Language database and Baby Chillanto database for infant speech and crying classification tasks, respectively. On Dunstan Baby Language database, the experiment gives an encouraging reduction of 4.14% absolute classification error rate compared with the results from using one-stage CNNs with spectrogram feature. In addition, on infant crying Baby Chillanto database, our approach outperforms all other studies on this classification task. A significant absolute reduction of 5.23%, 4.30%, and 1.67% are achieved compared to one-stage CNN, the transfer learning CNN and SVM ensemble model, and the MLP with Radial Basis Function Network, respectively.

5 INFANT CRY CLASSIFICATION WITH GRAPH CONVOLUTIONAL NETWORKS

In this chapter, we present our proposed approach of graph convolutional networks for robust infant cry classification. We construct non-fully connected graphs with weighted edges based on the similarities among the relevant nodes and feed them into convolutional neural networks to consider the short-term and long-term effects of infant cry signals related to inner-class and inter-class effects. The approach captures the diversity of variations within infant cries and gives encouraging results in both supervised and semi-supervised node classification. The effectiveness of this approach is evaluated on Baby Chillanto database and Baby2020 database. With limited 20% of labeled training data, our model outperforms the CNN model with 80% of labeled training data and the accuracy stably improves as the number of labeled training samples increases. The best results give significant improvements of 7.36% and 3.59% compared with the results of the CNN models on Baby Chillanto database and Baby2020 database, respectively.

5.1 Graph Convolutional Network and Related Work

Neural Networks (NNs) have been applied for infant cry feature extraction and classification. As a type of NN, the cutting-edge method of Graph Neural Networks (GNNs) demonstrated ground-breaking performance on many tasks. The connectionist models capture the dependence of graphs via message passing between the nodes of graphs [75]. Graph Convolutional Networks (GCNs) is a branch of GNNs that have a good balance of spectral and spatial representation [76]. To date, GNN is rarely used in audio research domain yet and a thorough search yields two relevant studies. GCN with a fully connected graph was used to classify music genre and achieves high accuracy [77]. The attentional GNN based few-shot learning was proposed in environmental sound classification achieving satisfying improvements [78]. Previous

NN based methods address the fundamental work of infant cry classification. Challenges remain in this domain, especially under the limited labeled data with inconsistent transcriptions.

The available labeled infant cry data is very limited because infant cry data acquisition is sensitive and transcribing such raw data is time consuming and requires pediatricians or experienced care givers. Moreover, infant cry is more stationary and includes complicated acoustic and prosodic phenomena related to short-term and long-term influence. The uncertainty and inconsistency of the labels are higher than that of speech because the meaning behind infant cries is complex and possibly mixed. CNN approaches improve the performance of infant cry classification, whereas the use of fixed filter sizes in CNN cannot consider the long-term effects within the stationary infant cry signals. Transfer learning is powerful on image processing, whereas spectrograms are quite different from images. It is essential to discover novel efficient approaches to learn the short-term and long-term effects among multiple cry signals within the same category and across different categories for robust infant cry classification under limited samples recorded in real environments.

In this study, we propose applying GCN for infant cry classification. We establish the graphs connected by weighted edges based on the similarities of the relevant nodes. The non-fully connected graphs feeding to the GCN is constructed considering the short-term and long-term effects of infant cry signals related to inner-class and inter-class effects. Our GCN approach captures the long-distance variations within infant cries, especially under the limited training samples. The effectiveness of proposed method is evaluated on the Baby Chillanto database that contains pathological cries and Baby2020 database, which has larger size and was recorded from healthy infants in real life environment such as homes and hospitals. Our major contributions include: (1) we apply GCN on robust infant cry classification and achieve significant improvements; (2) our approach with 20% of labeled data outperforms the CNN model with 80%

of labeled data; (3) the novel non-fully connected graph with GCN is efficient to capture the long-distance effects in infant cry for better discriminative ability; (4) the challenge of limited training data with uncertain labels can be solved using weighted class-crossing edges in the graph of GCNs.

5.2 Infant Cry Classification with GCN

5.2.1 The Structure of Proposed Approach

The architecture of our model for infant cry classification is shown in Figure 5.1. The whole procedure includes three parts: feature extraction, transfer learning CNN, and GCN classification. As illustrated in Figure 5.1, spectrograms are generated from the audio wav samples and are then fed into the transfer learning CNN feature extractor to extract multi-dimensional feature vectors, which is the input of the GCN classifier. Supervised or semi-supervised audio classification with GCN is performed to get the outputs.

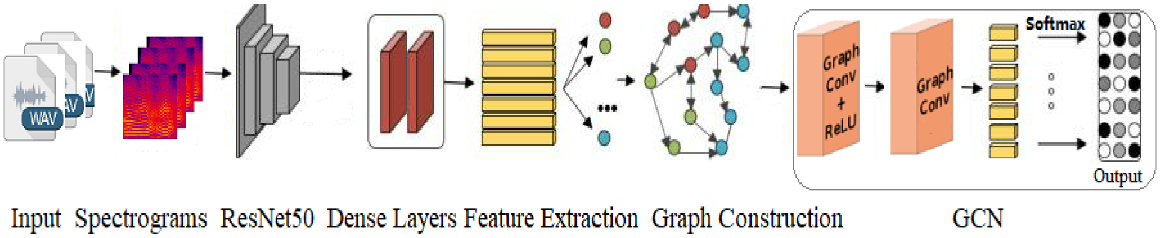

Figure 5.1:Our proposed model for infant cry classification with GCN.

Since spectrograms have strong relationship with images and ResNet50 is proven good performance for image classification, we feed the spectrograms into the feature extractor, which is built using a ResNet50 based transfer learning CNN model. The layers before the fully connected layer in ResNet50 are saved as the base model of our feature extractor. Appending some custom layers after the base model, we train the model using the training set and extract the features from the last fully connected layer and get the multi-dimensional feature vectors when predicting the test set.

5.2.2 Graph Convolutional Network Node Classification

As CNN being widely used for images classification, it learns the features within the images. By passing messages among neighbors and aggregation of features, GCN is a promising choice to learn hidden relationship among images. GCN node classification takes a graph as input and determine labels of nodes by learning the features of their own and associated neighbors. The input graph is defined as $G = (V, \mathbf{A})$, where V represents the vertex set consisting of nodes $\{v_1, . . . , v_n \}$ and $\mathbf{A} \in \mathrm{R}^{n\times n}$ is the adjacency matrix where $a_{ij} \geq 0$ denotes the edge weight between nodes v_i and v_j . The GCN we consider in our experiments have the similarity of each pair of images as the edge weight and each node v_i has a corresponding *1024*-dimensional feature vector $x_i \in R^{1024}$. Each entry on the diagonal degree matrix $D = Diag(d_i, \ldots, d_n)$ is equal to the row-sum of the adjacency matrix $d_i = \sum_0^j a_{ij}$. GCN learns a new feature representation for the feature x_i of each node over multiple layers and subsequently use it as input into a linear classifier. In each graph convolution layer, the features x_i of each node v_i are updated via averaging within its local neighborhood by feature propagation:

$$x_i^{(k)} \leftarrow \frac{1}{d_i+1} x_i^{(k-1)} + \sum_{j=1}^{n} \frac{1}{\sqrt{d_i+1}\sqrt{d_j+1}} x_i^{(k-1)} \qquad (5)$$

Intuitively, this step smooths the hidden representations locally along the edges of the graph and ultimately encourages similar predictions among locally connected nodes. After the local smoothing, the operation goes to a standard multi-layer perceptron with a weight matrix $\mathbf{\Theta}^{(k)}$ and a nonlinear activation function, such as ReLU in our experiment, is applied pointwise before outputting the new feature representation $\mathbf{X}^{(k)}$. The last layer of a GCN predicts the labels using the softmax activation function, by which each node belongs to one out of *C* classes and can be labeled with a *C*-dimensional one-hot vector $y_i \in \{0, 1\}^C$.

5.2.3 Graph Construction for GCN

Semi-supervised graph construction for GCN. The graph fed to the GCN is built based on the similarities calculated using all samples including labeled training set and unlabeled testing set. We calculate the similarity of each pair of nodes by their Euclidean distance because it performs the best comparing to Cosine similarity and Gaussian similarity. Then for each node, the similarities to all other nodes are ordered and a hyperparameter k is tuned to decide the degree of the edges. Once two nodes are connected, the similarity between them is used as the weight on the corresponding edge. The weight matrix (2267×2267), which shows the similarity values if two nodes are connected and 0 otherwise.

Supervised graph construction for GCN. What's different for supervised method is that it builds the training graph and testing graph separately. 80% of data is used to construct the graph for training and 20% data for testing. Same as semi-supervised method, the graph is constructed by connecting k number of closest neighbors from each node. After training, the testing graph is fed into the trained GCN model for node classification.

Figure 5.2 illustrates a graph constructed before feeding to the GCN. It has 15 nodes in five classes with k value equals to 2 and each node contains a feature vector extracted from CNN. Double arrowed edges represent both nodes are the k closest neighbors of each other. Single arrow means only the arrow-side node is the k closest neighbors of the node without the arrow, not in the opposite way. For example, node 0 and node 5 are close to each other since they have double arrows. A single arrow is in between node 1 and node 2 because node 2 is one of the closest to node 1 but node 1 is not one of the closest two neighbors of node 2.

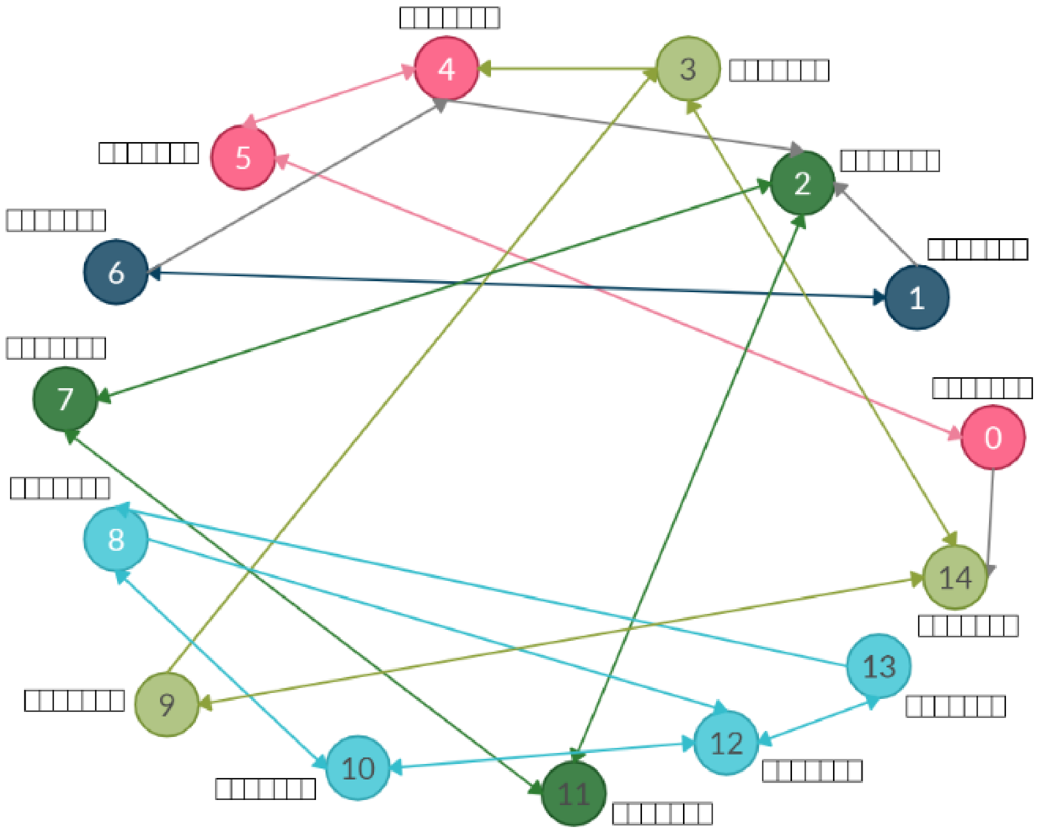

Figure 5.2: Graph constructed for GCN.

GCN with supervised graphs and semi-supervised graph are both valuable in infant cry classification task due to the lack of available labeled crying data. Discovering effective semi-supervised method can help make good use of large amount of unlabeled cry samples. Graph-based methods have been demonstrated as one of the most effective approaches for semi-supervised learning, as they can exploit the connectivity patterns between labeled and unlabeled data samples to improve learning performance. In GCN setting, semi-supervised classification has advantage because semi-supervised graph is constructed by the whole dataset, which means the distances between the testing samples and training samples are also calculated. Connecting unlabeled testing nodes to labeled training nodes before training brings some level of prior knowledge to the unlabeled data samples, which helps to achieve to better performance.

5.3 Experimental Results and Analysis

5.3.1 Datasets

We evaluate our approach on Baby Chillanto database and Baby2020 dataset. Baby Chillanto database is described in section 2.4.1. Baby2020 database is described in section 1.4. In this study, we use the subset of the Bab2020 database, which contains 5540 cry samples of three types (2880 hungry, 1700 Sleepy, 960 Wakeup) from healthy infants of 0 to 3 months old.

Sound eXchange (Sox) software [71] is used to generate spectrograms from the wav file samples. Each spectrogram image is in size 256×256. From the trained ResNet50, we take the convolutional layers before the fully connect layer as the base model. We then append custom layers, a GlobalMaxPooling layer, two dense layers with 1024 neurons, a dropout layer with a rate of 0.25, and another 1024-neuron dense layer, to the base ResNet50 model. The TL CNN model is trained on 80% of samples and used to extract the features for the rest of the 20% testing samples. For example, in Baby Chillanto database, the total 2267 samples are divided into five-fold training and testing sets. After five times training and extracting, the features for all 2267 samples are extracted, and then are combined to be the feature vectors stored in a csv file, which is a 2267×1024 matrix. We do the same training and extracting process for Baby2020 dataset and get a 5540×1024 feature matrix.

The feature matrix is then used to calculate the similarity of each pair of nodes using the Euclidean distance and the graphs are constructed for the GCN, which is introduced in [76]. The hyperparameters are tuned and the same values are used for all experiments on the same dataset. The values of the hyperparameters on Baby Chillanto database are: two-layer GCN, 32 number of hidden neurons in GCN, 2000 epochs, 0.001 learning rate, 0.1 dropout, and the best k value is 3. The hyperparameter values for Baby2020 database are the same except the k value is equal to 1, which produces the best performance. All experiments are performed with five-fold cross validation to ensure the reliability of the accuracies.

5.3.2 Results and Analysis

Table 5.1 shows our method outperforms other methods on Baby Chillanto database. Comparing to the standard CNN model, our approach in supervised and semi-supervised settings improves testing accuracy by 4.98% and 7.36% respectively. Our GCN approach aggregates similar features with short and long distance in the time series. Aggregating features of nodes in different classes can also help soothe the mislabeling effect, which is unavoidable because cry reasons may be mixed in one cry sample. Semi-supervised graph provides some prior knowledge of the relationship between testing data and the labeled training data, hence, produces better accuracy than the supervised classification with GCN.

Table 5.1: Results comparisons on Baby Chillanto database

Features	**Model**	**Accuracy**	**Improvement**
Spectrogram	CNN [79]	**87.03%**	baseline
Spectrogram	TLCNN & SVM [79]	90.80%	+3.77%
Spectrogram → TL CNN	GCN (supervised)	92.01%	+4.98%
Spectrogram → TL CNN	GCN (semi-supervised)	**94.39%**	**+7.36%**

Table 5.2: Results comparisons on Baby2020 database

Features	**Model**	**Accuracy**	**Improvement**
Spectrogram	CNN	**70.78%**	baseline
Spectrogram → TL CNN	GCN (supervised)	74.24%	+3.46%
Spectrogram → TL CNN	GCN (semi-supervised)	**74.37%**	**+3.59%**

Table 5.2 shows the performances of our approach on Baby2020 dataset. It outperforms standard CNN model in both supervised and semi-supervised settings by improving 3.46% and

3.59% accuracy, respectively. The classification accuracy on Baby2020 dataset is not as high as the Baby Chillanto database. We believe it is because data samples recorded by recording APP installed on mobile devices are in mp3 and m4a formats, which use lossy compression technique causing loss of some high frequency information of the audios. The recording environments may also have some background noises and the three types of crying, hungry, sleepy, and wakeup, are from healthy infants and have high level of similarity. The effectiveness of GCN on such dataset indicates that the relationship among same class and different class data samples are considered in learning and help improve the learning outcomes. The Baby2020 dataset will be our focus in the future because it has the same setting as real time infant cry classification system, by which the user can catch infant cry signal using mobile devices and immediately tell infants' needs.

Table 5.3: Results of experiments with different train test ratios

Dataset	**Model (Train:Test)**	**Accuracy**	**Improvement**
Chillanto	**CNN (80:20)**	**87.03%**	---
Chillanto	GCN (supervised 20:80)	89.89%	+2.86%
Chillanto	GCN (supervised 40:60)	91.00%	+3.97%
Chillanto	GCN (supervised 50:50)	91.73%	+4.70%
Chillanto	GCN (supervised 60:40)	91.79%	+4.76%
Chillanto	GCN (supervised 80:20)	**92.01%**	**+4.98%**
Chillanto	GCN (semi-supervised 20:80)	91.68%	+4.65%
Chillanto	GCN (semi-supervised 40:60)	92.33%	+5.30%
Chillanto	GCN (semi-supervised 50:50)	92.80%	+5.77%
Chillanto	GCN (semi-supervised 60:40)	92.80%	+5.77%
Chillanto	GCN (semi-supervised 80:20)	**94.39%**	**+7.36%**
Baby2020	**CNN (80:20)**	**70.78%**	---
Baby2020	GCN (supervised 20:80)	71.58%	+0.8%
Baby2020	GCN (supervised 40:60)	71.99%	+1.21%
Baby2020	GCN (supervised 50:50)	72.81%	+2.03%
Baby2020	GCN (supervised 60:40)	73.42%	+2.64%
Baby2020	GCN (supervised 80:20)	**74.24%**	**+3.46%**
Baby2020	GCN (semi-supervised 20:80)	73.45%	+2.67%
Baby2020	GCN (semi-supervised 40:60)	73.53%	+2.75%
Baby2020	GCN (semi-supervised 50:50)	73.79%	+3.01%
Baby2020	GCN (semi-supervised 60:40)	73.95%	+3.17%
Baby2020	GCN (semi-supervised 80:20)	**74.37%**	**+3.59%**

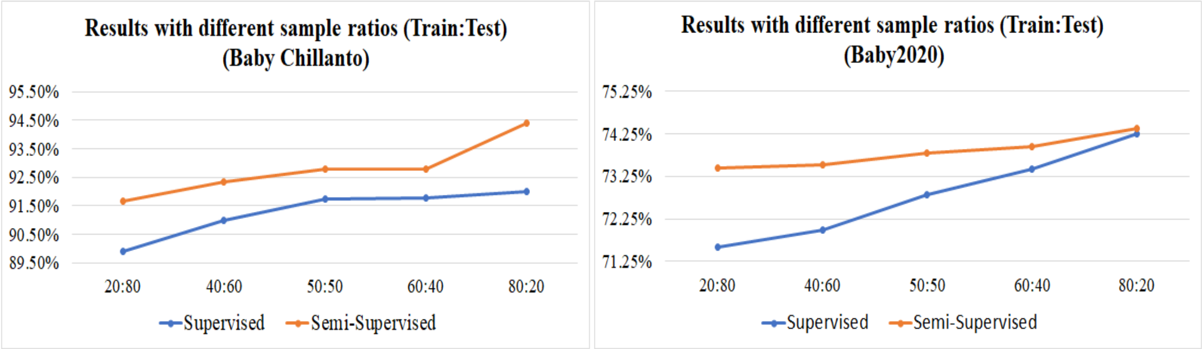

Figure 5.3: Results of experiments with different sample ratios. The semi-supervised method outperforms the supervised method in both databases. The testing accuracy stably improves as the number of labeled training samples increases in both semi-supervised and supervised settings.

Table 5.3 and Figure 5.3 show that with limited 20% of labeled training data, our model outperforms the CNN model with 80% of labeled training data. The testing accuracy stably improves as the number of labeled training samples increases in both semi-supervised and supervised settings. The semi-supervised method performs better because the graph with all samples included contains more knowledge among labeled data and unlabeled testing samples.

5.4 Summary

GCN with transfer learning CNN extracted features measured with similarity is efficient for robust infant cry classification. The non-fully connected graphs of GCN are constructed to consider the short-term and long-term effects of infant cry signals related to inner-class and inter-class messages. The improved discriminative ability of local nodes has the benefits of capturing the long-distance variations within infant cries, especially under the limited training data. The experimental results show that even with as limited as 20% of labeled data, our model outperforms that of the CNN model with 80% labeled training data and the accuracy increases as more training samples are used. The best accuracy improves 7.36% and 3.59% on Baby Chillanto database and Baby2020 database, respectively. Our future work includes generating reliable graph edges for signed GCNs via the use of prior knowledge of class, investigating the effects of different similarity kernels for edges, as well as designing a deeper network architecture for more infant cry samples. Our approach outperforms other methods in infant cry classification and can be easily extended to other acoustic event classification tasks.

6 INFANT CRY CLASSIFICATION BASED ON FEATURE FUSION AND MEL-SPECTROGRAM DECOMPOSITION WITH CNNS

In this chapter, we introduce the proposed novel method of using feature fusion and model fusion to improve infant cry classification performance. Spectrogram features extracted from transfer learning convolutional neural network (CNN) model and mel-spectrogram features are extracted from mel-spectrogram decomposition model. Fused spectrogram features and mel-spectrogram features are fed into an MLP for better classification accuracy. The mel-spectrogram decomposition method feeds band-wise crops of the mel-spectrograms into multiple CNNs followed by a merged global classifier to capture more enhanced discriminative features. Feature fusion brings higher dimensional detailed information and features more in line with human hearing perception together to achieve performance on CNN. The evaluation of the approach is conducted on Baby Chillanto database and Baby2020 database. Our approach yields a significant reduction of 4.72% absolute classification error rate compared with the result using single mel-spectrogram images with CNN model on Baby Chillanto database and our testing accuracy reaches 99.26%, which outperforms all other methods with this five-category classification task. The gender classification experiment on Baby2020 database also shows 3.87% accuracy improvement compared with the CNN model using single spectrograms.

6.1 Feature Fusion and Mel-spectrogram Decomposition

Infant cry research usually starts with data acquisition and preprocessing, followed by feature extraction and feature selection, and finally the classification model. To classify or detect certain types of cry signals, researchers have been using many types of features in the past decades for infant cry detection and classification. The commonly used features include cepstral domain features such as Mel-frequency cepstral coefficient (MFCC) [31], Linear Prediction Cepstral

Coefficients (LPCC) and Bark Frequency Cepstral Coefficients (BFCC) [5], prosodic features such as F0, formants, intensity [22][45], and image-based features such as spectrograms [16][80][81], waveforms [3][82], and mel-spectrograms [83].

Spectrograms are widely used in neural network models such as transformer model [84], Siamese neural network for animal audio classification [85], Convolutional neural network (CNN) [86] for environmental sound classification, CNN models for infant cry classification [23] [81][87]. Many successful works are performed using mel-spectrograms on audio research. Augmentation technique on mel-spectrogram is effective on sequence-to-sequence voice conversion [88]. Mel-spectrograms are used to produce high quality speech synthesis [89][90].

SubspectralNet, which is a mel-spectrogram decomposition method, was proposed in [91] and it feeds band-wise crops of the mel-spectrograms into multiple CNN models followed by a merged global classifier to capture more enhanced discriminative features on acoustic scene classification. Zhang et al. further segmented the mel-spectrogram in both time and frequency dimension and proposed the mel-spectrogram decomposition model to further improve the performance of acoustic scene classification [92]. Feature fusion, the combination of features from different layers or branches, is an omnipresent part of modern network architectures [93]. It is used in image segmentation [94], image classification [93], audio-visual speech enhancement [95], and audio event classification [96], etc. In infant cry research, MFCC and prosodic feature are fused for a FFNN to asphyxiated cry identification [22]. Cry signal spectrogram features and environmental features are fused to improve infant cry reason classification in an infant care center environment [97]. In this study, we propose an approach to fuse the features extracted from spectrograms and mel-spectrograms. We use transfer learning model to extract spectrogram features and use mel-spectrogram decomposition multi-CNN merging model to extract mel-spectrogram features. From the Linear discriminant analysis (LDA) plots, we visualize the fused features and see they are more discriminative than either of the two feature representations individually, hence, the multiple layer perception (MLP) classifier produces better overall classification accuracy.

In this study, our contributions include: (1) we apply mel-spectrogram decomposition with multiple CNNs and model merging method onto infant cry classification to improve the classification accuracy; (2) we propose to fuse spectrogram features and mel-spectrogram features to obtain more discriminative features of infant cry signals, as well as keeping the coverage of crying characteristics; (3) transfer learning CNN captures the high-level features of the spectrograms and mel-spectrogram decomposition makes local features more prominent. Fusing features from these two CNN feature extractors generates more thorough features of the original signals.

6.2 Spectrograms vs. Mel-spectrograms

A spectrogram is a visual representation of a signal at different frequencies as it varies with time. It is a two-dimensional heat map, in which the X-axis is the time, the Y-axis is the frequency, and the brightness of the color indicates the strength of the signal at a certain time and frequency. Spectrograms extensively used in the field of speech, linguistic, music, animal sound, and others. The spectrogram can be obtained through four steps: framing, windowing, fast Fourier transform (FFT), and stacking the results of each frame. Framing is to divide the original signal into certain number of frames with a length, an overlap, and a framing hop. Hamming or other windowing is commonly used to avoid spectral leakage. The FFT converts the signal from the time domain to the frequency domain. Stacking up the result of FFT of each frame produces the spectrogram of a certain audio signal. A mel-spectrogram is another visual representation of a signal indicating how people hear sound by converting the Y-axis from frequency to mel-scale. Mel-spectrogram can relatively represent human sound perception characteristics, which presents the linear distribution under the 1000Hz and the logarithm growth above the 1000Hz on a logarithmic scale rather than a linear scale. People are more sensitive to lower frequency sound and the difference between high frequency sound is not as easy to distinguish as the ones between lower frequency sound. After framing, windowing, and FFT, frequency portion of the spectrum is mapped to the mel scale perceptual filter bank using M triangular-shaped filter bank equally spaced on the mel range of frequency. The relationship between frequencies and mel frequency scale is as shown in (1).

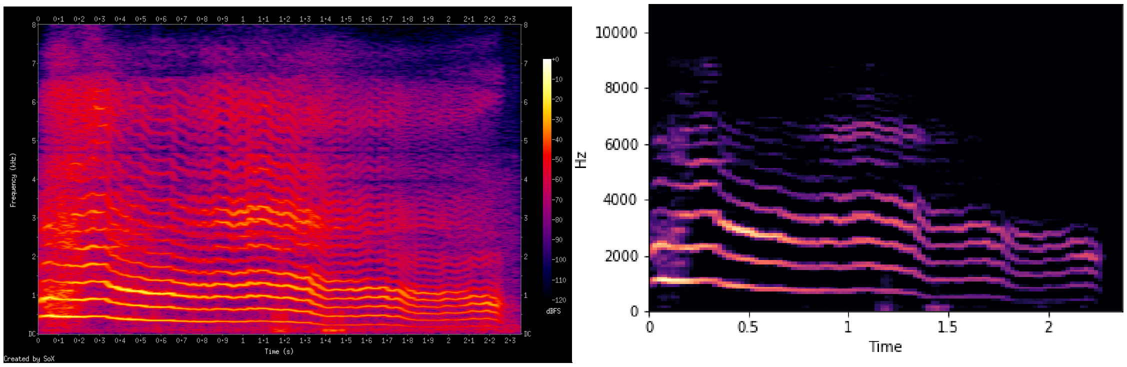

Figure 6.1:Spectrogram (left) and mel-spectrogram (right) of a hungry cry.

Figure 6.1 shows a hungry cry in spectrogram and mel-spectrogram. Comparing to mel-spectrograms, spectrograms contain higher dimensional features extracted from each frame of the signal, hence, the image contains more details, but some detailed information is not necessarily useful for certain classification tasks. Mel-spectrogram is generated by adding filter banks on the mapping step and the commonly used filter banks are the same area triangular filter banks, which reduces higher frequency information more and the results is more in line with human hearing perception. As Figure 6.1 illustrates, the higher frequency information on the mel-spectrogram is weakened and the mel-spectrogram has a lower dimension of features and carries less information, but certain frequency resolution is enhanced. The harmonic structure becomes drastically weaker as the frequency increases. In other words, the lower frequency region covers more energy and the transitional pattern of speech manifold in that region. This is the reason why mel-scale frequency warping is promising for speech recognition. Figure 6.2 illustrates the spectrograms and mel-spectrograms of three types of sound: an adult speech, an infant hungry cry, and a cat meow sound. We consider cat meow sound a live acoustic scene sound. We can see the differences among these three types of sound as follows: (1) adult speech contains obvious changes in vocal tract and the vibration of the vocal cords, producing unvoiced and voiced sounds to form meaningful speech; (2) live acoustic scene event such as cat meow sound can be clearly seen that there is no articulation ability and there is no control in breathing, which is shown as a flat bar in the spectrogram without pause or changes of inspiration and expiration phases; (3) the spectrogram of an infant cry is likely in between the speech signal and cat sound signal. It is more stationary compared to human speech while includes more variations of the signal and more changes of inspiration and expiration with respect to live animal sound. As mel-spectrograms are more suitable for speech recognition since it is more in line with human hearing perception, the spectrograms contain more information for other acoustic scene sounds including animal sounds. As a signal in between the adult speech and acoustic scene, infant cry signal should benefit from both spectrograms and mel-spectrograms. Combining detailed spectrogram features and human perception style mel-spectrogram features can cover a wider range of diversity of changes within infant cry signals leading to robust classification performance for CNN models.

6.3 Fusing Spectrogram Features and Mel-spectrogram Decomposition Features

In this section, we’ll describe our proposed method, which is fusing the extracted spectrogram features from transfer learning model and extracted mel-spectrogram features from mel-spectrogram decomposition model.

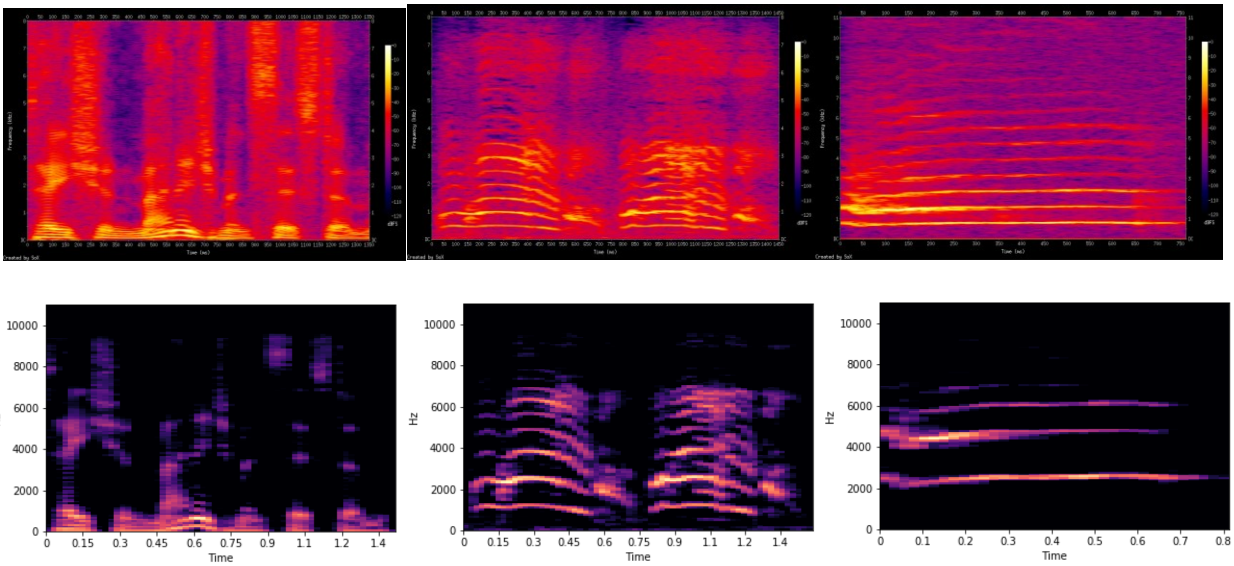

Figure 6.2:Spectrograms (the top row) and mel-spectrograms (the bottom row) of adult speech (left), infant cry (middle), and cat sound (right).

The Figure 6.3 illustrates our fusion method. The original wav files are used to generate spectrograms and mel-spectrograms using Sox [71] and librosa [24] software, respectively. The generated images are resized into the appropriate size as model input for the following two feature extractors, which solves the problem of variable length of the original wav files. We feed the spectrograms into the spectrogram feature extractor, which is built using a ResNet50 based transfer learning CNN model. The layers before the fully connected layer in ResNet50 are saved as the base model and custom layers are appended after the base model. We train the model using the training set and then extract the feature vectors of the testing set from the last fully connected layer to be the extracted spectrogram features. In the mel-spectrogram feature extractor, each mel-spectrogram is decomposed equally into 4 sub mel-spectrograms horizontally, which contains 32 mel-bins in each slice. We choose to decompose the mel-spectrogram into 4 slices based on the experimental results, which show that 4 slices or 5 slices outperform the 2, 3, 6, and 7 slices. We choose to equally decompose the images without any overlapping because the performance of the decomposition model is not affected much by overlapping the slices in our experiments. There are four CNNs used in this decomposition model, in which each sub mel-spectrogram is input for one CNN model containing a convolutional layer, a max pooling, a flattened layer, and a dense layer. The outputs of all CNN model are concatenated followed by a dense layer and a softmax layer for classification. The four CNN models and merged layers are trained together on 80% samples and the features from the last dense layer are extracted as the mel-spectrogram decomposition features. Spectrogram features and mel-spectrogram features are then concatenated to be the new feature vector for the final classification model, which is a Multiple layer perception (MLP) model in this approach. The output of the whole model is a softmax layer classifying different types of cry samples.

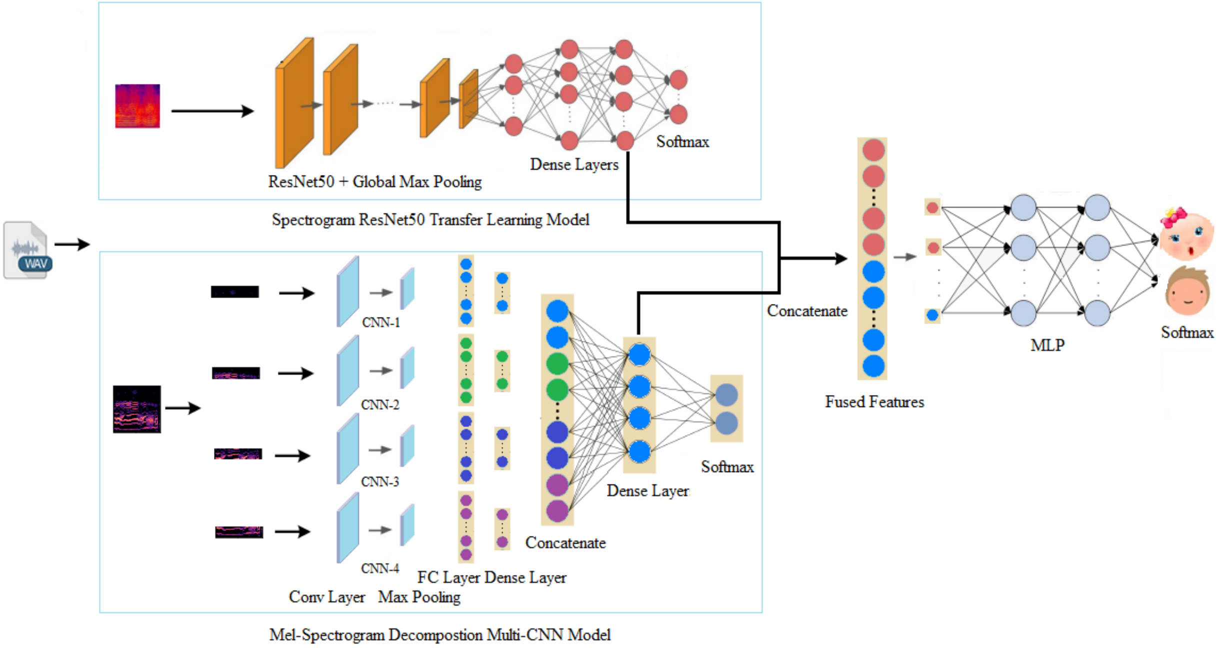

Figure 6.3:Model architecture of our fusion approach. Audio wave files are used to generate spectrograms and mel-spectrograms and they are fed into ResNet50 transfer learning model and decomposition model respectively to extract spectrogram features and mel-spectrogram features. The fused the features are then fed into an MLP for final classification.

6.4 Experimental Results and Analysis

6.4.1 Datasets

We evaluate the effectiveness of the approach on two datasets. They are a subset of our developing Baby2020 database and Baby Chillanto database. In this study, we use 4000 cry samples from Baby2020, including 2000 cries from baby boy younger than 3 months and 2000 cries from baby girl younger than 3 months. From Baby Chillanto database, we use 340 asphyxia samples, 879 deaf samples, 350 hunger samples, 506 normal cries, and 192 pain cries.

6.4.2 Experimental Setup

Spectrogram transfer learning. Sound eXchange (Sox) [71] is used to generate spectrograms from the wav files. In the feature extractor of ResNet50 transfer learning model, the spectrograms are in size 256×256 for Chillanto dabase and 128×128 for Baby2020 database due to the size of the datasets. From the trained ResNet50, we take the convolutional layers before the fully connect 1000 layer as the base model. The custom layers appended to the base model contains a GlobalMaxPooling layer, one 1024-neuron dense layer, a 512-neuron dense layer, a dropout layer with a rate of 0.25, and another 32-neuron dense layer. The transfer learning CNN model is trained on 80% of samples and used to extract the features for the rest of the 20% testing samples. After five times training and extracting, the features for all samples are extracted, and then are combined to be the feature vectors stored in a csv file, which contains a n × 32 matrix where n is the number of samples.

Mel-spectrogram decomposition. The mel-spectrogram is extracted using librosa [24] with 2048 FFT points, 128 mel-bins, and a hop-length of 256. The sampling rate used is the original one from the cry samples, which is 16K for Baby2020 cries, 11250, 8000, 22050 for Baby Chillanto asphyxia, deaf and hunger, normal and pain, respectively. The amplitude of the mel-spectrogram is scaled logarithmically. The scaled mel-spectrogram is decomposed into 4 slices equally and they are resized to 35 × 50 (height × width) to fit the CNN model input size. Each CNN for each slice has the same architecture, which contains a convolutional layer with twenty 5 × 5 filters followed by a max-pooling layer, which uses 2 × 2 filter with stride 2. Then there is a flatten layer and a dense layer with 128 neurons. The output of each CNN is concatenated appended by one 64 neuron dense layer and a 32-neuron dense layer with ReLu activation function, at last a softmax activation function is applied to the last dense layer for classification. We use this mel-spectrogram decomposition and merged model to extract mel-spectrogram features. The 32 features in the last dense layer are extracted and concatenated to the 32 features from the transfer learning spectrogram model. The fused features are then input into an MLP, which is written in Pytorch containing four hidden layers with 1056, 512, 256, 64 neurons, respectively. The optimized used is Adam and learning rate is set to 0.001 and the activation function for the output layer is softmax.

6.4.3 Results and Analysis

The two experiments designed to evaluate the approach are infant cry reason classification and gender classification. Many previous research has classified the infant cry reason and reached good accuracy and we hope to further improve the performance. Gender classification is specifically designed for this approach to show that our proposed method is also effective on all healthy cry samples with precise labels.

Spectrograms are used in other related studies for infant cry classification on Baby Chillanto database. As shown in Table 6.1, the five-category classification reaches 94.39% using spectrograms in graph convolutional network method [98]. Using mel-spectrogram as feature input, the accuracy of a simple CNN model reaches 94.54%, which outperforms the best model using spectrograms. Mel-spectrogram decomposition method improves the accuracy up to 98.70%

and our proposed fusion method reaches 99.26%. In [91], authors show that depending on the scene class, there is a specific frequency band showing most activity, hence providing discriminative features for that class. We believe the reason of mel-spectrogram decomposition performing very well in Baby Chillanto database is because the five types of cries include both healthy cries and pathological cries, which contain discriminative features in different bands. In the gender classification experiment, simple spectrogram CNN outperforms the vanilla mel-spectrograms CNN. This result shows that spectrogram and mel-spectrogram have their own advantages on different audio classification tasks. As shown on Table 6.2, the mel-spectrogram decomposition method is powerful and our method with MLP produces 96.64% accuracy that outperforms all other types of methods on Baby2020 database.

Table 6.1:Experimental results on Baby Chillanto database

Features	**Model**	**Accuracy**
Spectrograms	CNN	87.03%
Spectrograms	Resnet50 Transfer Learning (TLCNN)	90.08%
Spectrograms	TLCNN + GCN	94.39%
Mel-spectrograms	CNN	94.54%
Mel-spectrograms	Decomposition 4-CNN	98.70%
Spectrograms + Mel-spectrograms	TLCNN + Decomposition 4-CNN + MLP	99.26%

Table 6.2:Experimental results on Baby2020 database

Features	**Model**	**Accuracy**
Spectrograms	CNN	92.77%
Spectrograms	Resnet50 Transfer Learning (TLCNN)	94.84%
Mel-spectrograms	CNN	90.26%
Mel-spectrograms	Decomposition 4-CNN	93.15%
Spectrograms+Mel-spectrograms	TLCNN + Decomposition 4-CNN + MLP	96.64%

6.5 Summary

In this study, we demonstrated that mel-spectrogram decomposition is effective on infant cry classification and fusing spectrogram features and mel-spectrogram feature can further improve classification performance. Transfer learning CNN captures the high-level features of the spectrograms and mel-spectrogram decomposition extracts the detailed local features, and more thorough features of the original signals are generated by fusing these two types of features, which improves the final classification accuracy. Cry reason classification on Baby Chillanto database and gender classification on Baby2020 database were used to evaluate the proposed method. Our approach yields 4.72% accuracy improvement than the result using single mel-spectrograms with CNN model on Baby Chillanto database and our testing accuracy reaches 99.26%, which outperforms all other methods. The gender classification experiment on Baby2020 database also shows 3.87% accuracy improvement compared with the CNN model using single spectrograms. Our method focuses on infant cry classification but can be extended to other acoustic scene classification tasks.

7 CONCLUSION AND FUTURE WORK

7.1 Conclusion

In this dissertation, we built Baby2020 infant cry database, reviewed infant cry research in the past decade, and proposed five methods to improve the accuracies of pathological infant cry identification and infant cry reason classification.

In the previous chapters, we have described an approach of generating merged feature matrix with the combination of weighted prosodic features and acoustic features for asphyxiated baby crying classification. The weights attached with different prosodic features at frame level are efficiently extracted and optimized using deep learning neural networks. The generated merged feature matrix keeps the good robustness ability of the model. In addition, the use of weights associated with prosodic features increases the distinction degree of the features. Hence, the matrix has a good ability to capture the diversity of variations within infant crying signals. It has been shown that this new merged feature matrix together with weighted prosodic features method provides a good balance of robustness and resolution. It gives an encouraging reduction of 3.11%, 3.23%, 1.43% absolute classification error rate compared with the results from using single acoustic features, single prosodic features, and both acoustic and prosodic features, respectively. Our method focuses on classification of normal and asphyxiated infant crying but can be easily extended to other infant crying identification tasks.

We have presented that infant cry age classification with CNN is an efficient non-invasive method to diagnose the abnormality of the vocal tract development of early age infants. The analysis of acoustic features from different monthly age infants shows that 4-month age is a key turning point of infant vocal tract growth. We have shown that the length and volume of vocal cord may not change much during the postnatal development of the first 3 months. F1 and F2 increase significantly within the early 4 months indicating a great change of tongue location, oral, and nasal cavity extension for word pronunciation. We applied CNN to discover the pattern of the changes and diagnose the abnormality of the infant vocal tract by age classification. Our method achieved 79.20% accuracy for healthy infant cries on Baby2020 database, obtained 84.80% of asphyxiated cries, and 91.20% of the deaf cries for abnormal vocal tract development diagnosis on Baby Chillanto database. In the future, we will apply other machine learning methods, such as SVM and graph neural networks, to improve the age classification accuracy. We plan to expand the dataset to include older infant cries and young children's early speech to study the vocal tract development in children's first several years of growth.

We have described an approach of using hybrid feature set with multi-stage CNNs for robust infant sound classification. We have shown infant sound is a complicated procedure including infant speech and crying which have different acoustic and prosodic characteristics. Different features have different discrimination ability to model the diversity of variations within infant sound. The use of dominant and auxiliary features is beneficial to enlarge the coverage as well as keeping a good resolution. We used multi-stage CNNs method together with prior knowledge constraints in decision making to overcome the data sparse problem in infant sound classification. Prior knowledge information either from rules or from statistical result provides a good guidance for searching and decision making. The effectiveness of our method was evaluated on commonly used both Dustan Baby Language database and Baby Chillanto database for infant speech and crying classification tasks. It gives an encouraging reduction of 4.14% absolute classification error rate compared with the results from using one-stage CNNs with spectrogram feature. In addition, on infant crying Baby Chillanto Database, our approach outperforms all other studies on this classification task. A significant absolute reduction of 5.23%， 4.30%, and 1.67%

is achieved compared to one-stage CNN, the transfer learning CNN and SVM ensemble model, and the MLP with Radial Basis Function Network, respectively. Our method generates a mechanism of hybrid features and multi-stage CNNs and can be extended to other infant signal processing tasks as well as other classification tasks.

We also demonstrated that using GCN with transfer learning CNN extracted features measured with Euclidean similarity is efficient for robust infant cry classification. The non-fully connected graphs of GCN are constructed to consider the short-term and long-term effects of infant cry signals related to inner-class and inter-class effects. The improved discriminative ability of local nodes has the benefits of capturing the long-distance variations within infant cries, especially under the limited training data. The experimental results show that even with limited 20% of labeled data, our model outperforms that of the CNN model with 80% of labeled training data and the accuracy increases as more training samples are used. The best accuracy improves 7.36% and 3.59% on Baby Chillanto database and Baby2020 database, respectively. Our future work includes generating reliable graph edges for signed GCNs via the use of prior knowledge of class, investigating the effects of different similarity kernels for edges, as well as designing a deeper network architecture for more infant cry samples. Our approach outperforms other methods in infant cry classification and can be easily extended to other acoustic event classification tasks.

We also proposed a novel method of using feature fusion and model fusion to improve infant cry classification performance. Spectrogram features extracted from transfer learning CNN model and mel-spectrogram features are extracted from mel-spectrogram decomposition and merge model. Fused spectrogram features and mel-spectrogram features are fed into a MLP for better classification accuracy. The evaluation of the approach is conducted on Baby Chillanto database and Baby2020 database. Our approach reaches 99.26% classification accuracy on Baby Chillanto database, which outperforms all other classification methods with this five -category classification task. The experimental results on Baby2020 database also show 3.49% and 3.87% improvement than CNN with mel-spectrogram and CNN with spectrogram, respectively.

7.2 Future Work

In the future, we will continue working on infant cry research including infant cry reason classification, pathological infant cry identification, and infant cry detection. Since data is precious in this field, we will make effort to complete and enhance the Baby2020 database and try to record more datasets with different infants in different ages, which will involve children's speech in the future. Because some diseases can only be diagnosed once certain age is reached, we will also try to record some children's speech, especially children's speech with diseases or disorders. The current Baby2020 has more than 40000 samples, but it contains unbalanced data samples. Some categories have very less samples such as pain type of cries while some category has many samples such as hungry type. We hope to record more samples and try to make it more balanced. Also, the Baby2020 database only contains audio files. We believe video files including infants' face expression and information surrounding environment will be beneficial for future research. Infant cry reason classification is still challenging due to lack of datasets, and it is very important for developing future automatic infant care-giving systems. To improve the classification performance, not only audio files are needed, but also the environmental details are necessary to provide clues to understand infants' needs. Chang et al. combine the infant cry signals and environmental context to reach better performance on infant cry reason classification in day centers [97]. It is promising to study infant cry with multi-view data and models. Video files including infants' face expressions will also be helpful to be combined with the audio information to improve the performance of infant cry classification. To reach our goal, we need to work with more parents and infant care givers to record more related videos for further research.

Another important focus will be disease prediction by infant and children's sound. We will put more effort on identifying diseases from sound including cry samples and speech. Due to the sensitivity of the data, it needs more collaboration with medical professionals to collect disease related cries or speech samples such as cry and speech samples from autistic children. Disease prediction by sound can help early diagnosis of certain diseases and we hope it can be part of the multimodal research in disease prediction by many factors such as genome sequencing, medical images, symptoms, and environmental factors, etc. The age classification performance is relatively low as described in chapter 3. In the future, we will apply other machine learning methods to improve age classification accuracy and hope to test the model with more pathological data and improve the performance. We will analyze cry signals using other methods such as the complexity-based approach that uses normalized compression distance as a similarity measure [99]. We plan to expand the dataset to include older infant cries and young children's early speech to study the vocal tract development in children's first several years of growth. With large dataset and the combination of different audio features, we plan to apply rough set theory [100][101] into our audio classification for better performance in time and accuracy.

Deep learning architectures embedded with prior knowledge can also be explored. In this dissertation, we have used prior knowledge to guide the neural network design. The future goal is to embed the prior knowledge into neural network training, which should be more powerful, especially with medical image classification. More recent works utilize the domain knowledge from medical doctors, to create networks that resemble how medical doctors are trained, mimic their diagnostic patterns, or focus on the features or areas they pay particular attention to [102]. We will study the current progress on integrating medical domain knowledge into deep learning and design novel methods to contribute to this research area. In 5.2, we presented that the GCN is effective to further improve the infant cry classification and semi-supervised method is a promising direction for future research because there are abundant recordings available without labels. To construct the graph, we use features extracted from the CNN transfer learning model to calculate the audio signal similarity. We will further explore adding more features, such as infants' age, gender, cry reasons, to provide more information representing each individual cry signal. In audio classification, GCN is also used in music genre classification and Siamese neural networks are used as an analogous to GNN for learning edge similarity weights [77]. We will further explore different methods for edge similarity calculation including the Siamese neural networks. Our proposed method using GCN in infant cry classification can also be extended to music classification and other audio classification research.

Disease prediction from speech will also be explored in the future. While we analyze the difference between infant cries and adult speech, speech related topics also attract our attention. We will explore the disease prediction by speech including for children and adults. For example, speech recording from Alzheimer's patients can be explored and be used in machine learning model for early Alzheimer's diagnosis. It will help many families and take care of their elder people as early as possible.

In addition, some experiments in this research didn't reach ideal performance that we expect, but it is still promising to continue. The methods include LSTM model, GLCN model, GAT model, and signed GCN model, etc. We have experimented some data augmentation methods such as SpecAugment [14] and merging feature images. Although the experimental results are not ideal, we believe data augmentation is still an area that will improve the performance of infant cry research. The mel-spectrogram decomposition method introduced in 6.3 is close to the masking method of data augmentation. In the future, we will continue exploring data augmentation methods. With bigger datasets and multi-view architectures, more related research is worth to explore in the future. We believe the work in this dissertation has built the solid foundation and has led us to the right direction of our future research. We will continue our effort in this area and extend it to other related research areas. With more databases available in the future, we believe that more machine learning methods can be explored in this area. Combining new audio signal processing methods and novel machine learning methods will lead this research to a remarkable future, which will change people’s lives by providing affordable infant automatic caregiving.

REFERENCES

[1] O. Wasz-Höckert, T. J. Partanen, V. Vuorenkoski, K. Michelsson, and E. Valanne, “The identification of some specific meanings in infant vocalization,” *Experientia*, 1964.

[2] J. Mukhopadhyay, B. Saha, B. Majumdar, A. K. Majumdar, S. Gorain, B. K. Arya, S. Das Bhattacharya, and A. Singh, “An evaluation of human perception for neonatal cry using a database of cry and underlying cause,” in *2013 Indian Conference on Medical Informatics and Telemedicine, ICMIT 2013*, 2013.

[3] M. U. Sachin, R. Nagaraj, M. Samiksha, S. Rao, and M. Moharir, “GPU based Deep Learning to Detect Asphyxia in Neonates,” *Indian J. Sci. Technol.*, vol. 10, no. 3, 2017.

[4] R. Mugitani and S. Hiroya, “Development of vocal tract and acoustic features in children,” *Acoust. Sci. Technol.*, vol. 33, no. 4, pp. 215–220, 2012.

[5] L. Liu, Y. Li, and K. Kuo, “Infant cry signal detection, pattern extraction and recognition,” in *2018 International Conference on Information and Computer Technologies, ICICT 2018*, 2018.

[6] “Praat: doing Phonetics by Computer.” [Online]. Available: https://www.fon.hum.uva.nl/praat/. [Accessed: 07-Aug-2020].

[7] L. Floridi, “AI and Its New Winter: from Myths to Realities,” *Philosophy and Technology*. 2020.

[8] O. F. Reyes-Galaviz, E. A. Tirado, and C. A. Reyes-Garcia, “Classification of infant crying to identify pathologies in recently born babies with ANFIS,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3118, pp. 408–415, 2004.

[9] E. Franti, I. Ispas, and M. Dascalu, “Testing the Universal Baby Language Hypothesis -

Automatic Infant Speech Recognition with CNNs," *2018 41st Int. Conf. Telecommun. Signal Process. TSP 2018*, 2018.

[10] "GitHub - gveres/donateacry-corpus: An infant cry audio corpus that's being built through the Donate-a-cry campaign - see http://donateacry.com." [Online]. Available: https://github.com/gveres/donateacry-corpus. [Accessed: 07-Aug-2020].

[11] D. Ferretti, M. Severini, E. Principi, A. Cenci, and S. Squartini, "Infant cry detection in adverse acoustic environments by using deep neural networks," *Eur. Signal Process. Conf.*, vol. 2018-Septe, pp. 992–996, 2018.

[12] X. Zhang, Y. Zou, and Y. Liu, "AICDS: An infant crying detection system based on lightweight convolutional neural network," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018.

[13] G. Z. Felipe, R. L. Aguiat, Y. M. G. Costa, C. N. Silla, S. Brahnam, L. Nanni, and S. McMurtrey, "Identification of Infants' Cry Motivation Using Spectrograms," *Int. Conf. Syst. Signals, Image Process.*, vol. 2019-June, pp. 181–186, 2019.

[14] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019.

[15] X. Yao, M. Micheletti, M. Johnson, and K. de Barbaro, "Detection of Infant Crying in Real-World Home Environments Using Deep Learning," 2020. [Online]. Available: http://arxiv.org/abs/2005.07036.

[16] M. A. Tugtekin Turan and E. Erzin, "Monitoring infant's emotional cry in domestic environments using the capsule network architecture," in *Proceedings of the Annual*

Conference of the International Speech Communication Association, INTERSPEECH, 2018.

[17] G. Gu, X. Shen, and P. Xu, “A Set of DSP System to Detect Baby Crying,” *Proc. 2018 2nd IEEE Adv. Inf. Manag. Commun. Electron. Autom. Control Conf. IMCEC 2018*, no. Imcec, pp. 411–415, 2018.

[18] G. Sharma, K. Umapathy, and S. Krishnan, “Trends in audio signal feature extraction methods,” *Appl. Acoust.*, vol. 158, p. 107020, 2020.

[19] R. Cohen and Y. Lavner, “Infant cry analysis and detection,” *2012 IEEE 27th Conv. Electr. Electron. Eng. Isr. IEEEI 2012*, pp. 1–5, 2012.

[20] A. Zabidi, L. Y. Khuan, W. Mansor, I. M. Yassin, and R. Sahak, “Classification of infant cries with asphyxia using multilayer perceptron neural network,” *2010 2nd Int. Conf. Comput. Eng. Appl. ICCEA 2010*, vol. 1, pp. 204–208, 2010.

[21] L. Liu, Y. Li, and K. Kuo, “Infant cry signal detection, pattern extraction and recognition,” *2018 Int. Conf. Inf. Comput. Technol. ICICT 2018*, no. 2, pp. 159–163, 2018.

[22] C. Ji, X. Xiao, S. Basodi, and Y. Pan, “Deep learning for asphyxiated infant cry classification based on acoustic features and weighted prosodic features,” *Proc. - 2019 IEEE Int. Congr. Cybermatics 12th IEEE Int. Conf. Internet Things, 15th IEEE Int. Conf. Green Comput. Commun. 12th IEEE Int. Conf. Cyber, Phys. So*, 2019.

[23] C. Ji, S. Basodi, X. Xiao, and Y. Pan, “Infant Sound Classification on Multi-stage CNNs with Hybrid Features and Prior Knowledge,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020.

[24] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and Music Signal Analysis in Python,” in *Proceedings of the 14th Python in Science*

Conference, 2015.

[25] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE - The Munich versatile and fast open-source audio feature extractor," in *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*, 2010.

[26] C. C. Onu, I. Udeogu, E. Ndiomu, U. Kengni, D. Precup, G. M. Sant'anna, E. Alikor, and P. Opara, "Ubenwa: Cry-based Diagnosis of Birth Asphyxia," 2017. [Online]. Available: http://arxiv.org/abs/1711.06405.

[27] K. Sharma, C. Gupta, and S. Gupta, "Infant Weeping Calls Decoder using Statistical Feature Extraction and Gaussian Mixture Models," *2019 10th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2019*, pp. 1–6, 2019.

[28] M. Kia, S. Kia, N. Davoudi, and R. Biniazan, "A detection system of infant cry using fuzzy classification including dialing alarm calls function," *2nd Int. Conf. Innov. Comput. Technol. INTECH 2012*, pp. 224–229, 2012.

[29] A. Rosales-Pérez, C. A. Reyes-García, J. A. Gonzalez, O. F. Reyes-Galaviz, H. J. Escalante, and S. Orlandi, "Classifying infant cry patterns by the Genetic Selection of a Fuzzy Model," *Biomed. Signal Process. Control*, 2015.

[30] K. Santiago-Sánchez, C. A. Reyes-García, and P. Gómez-Gil, "Type-2 fuzzy sets applied to pattern matching for the classification of cries of infants under neurological risk," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5754 LNCS, no. 1, pp. 201–210, 2009.

[31] Y. Lavner, R. Cohen, D. Ruinskiy, and H. Ijzerman, "Baby cry detection in domestic environment using deep learning," *2016 IEEE Int. Conf. Sci. Electr. Eng. ICSEE 2016*, 2017.

[32] S. Orlandi, C. A. Reyes Garcia, A. Bandini, G. Donzelli, and C. Manfredi, "Application of

Pattern Recognition Techniques to the Classification of Full-Term and Preterm Infant Cry," *J. Voice*, vol. 30, no. 6, pp. 656–663, 2016.

[33] M. V Varsharani Bhagatpatil, "An Automatic Infant's Cry Detection Using Linear Frequency Cepstrum Coefficients (LFCC)," *Int. J. Sci. Eng. Res.*, vol. 5, no. 12, pp. 1379–1383, 2014.

[34] R. I. Tuduce, H. Cucu, and C. Burileanu, "Why is My Baby Crying? An In-Depth Analysis of Paralinguistic Features and Classical Machine Learning Algorithms for Baby Cry Classification," *2018 41st Int. Conf. Telecommun. Signal Process. TSP 2018*, pp. 1–4, 2018.

[35] M. Petroni, A. S. Malowany, C. C. Johnston, and B. J. Stevens, "Classification of infant cry vocalizations using artificial neural networks (ANNs)," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 5, no. Cim, pp. 3475–3478, 1995.

[36] T. Fuhr, H. Reetz, and C. Wegener, "Comparison of Supervised-learning Models for Infant Cry Classification / Vergleich von Klassifikationsmodellen zur Säuglingsschreianalyse," *Int. J. Heal. Prof.*, vol. 2, no. 1, pp. 4–15, 2015.

[37] M. Hariharan, L. S. Chee, and S. Yaacob, "Analysis of infant cry through weighted linear prediction cepstral coefficients and probabilistic neural network," *J. Med. Syst.*, vol. 36, no. 3, pp. 1309–1315, 2012.

[38] M. Hariharan, J. Saraswathy, R. Sindhu, W. Khairunizam, and S. Yaacob, "Infant cry classification to identify asphyxia using time-frequency analysis and radial basis neural networks," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9515–9523, 2012.

[39] O. M. Badreldine, N. A. Elbeheiry, A. N. M. Haroon, S. Elshehaby, and E. M. Marzook, "Automatic diagnosis of asphyxia infant cry signals using wavelet based mel frequency cepstrum features," in *ICENCO 2018 - 14th International Computer Engineering Conference: Secure Smart Societies*, 2019.

[40] G. Esposito, N. Hiroi, and M. L. Scattoni, “Cry, baby, cry: Expression of distress as a biomarker and modulator in autism spectrum disorder,” *Int. J. Neuropsychopharmacol.*, vol. 20, no. 6, pp. 498–503, 2017.

[41] S. Orlandi, C. Manfredi, L. Bocchi, and M. L. Scattoni, “Automatic newborn cry analysis: A Non-invasive tool to help autism early diagnosis,” *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 2953–2956, 2012.

[42] K. Wu, C. Zhang, X. Wu, D. Wu, and X. Niu, “Research on acoustic feature extraction of crying for early screening of children with autism,” *Proc. - 2019 34rd Youth Acad. Annu. Conf. Chinese Assoc. Autom. YAC 2019*, pp. 290–295, 2019.

[43] M. Hariharan, R. Sindhu, and S. Yaacob, “Normal and hypoacoustic infant cry signal classification using time-frequency analysis and general regression neural network,” *Comput. Methods Programs Biomed.*, vol. 108, no. 2, pp. 559–569, 2012.

[44] O. F. Reyes-Galaviz, S. D. Cano-Ortiz, and C. A. Reyes-García, “Evolutionary-neural system to classify infant cry units for pathologies identification in recently born babies,” in *7th Mexican International Conference on Artificial Intelligence - Proceedings of the Special Session, MICAI 2008*, 2008.

[45] A. Rosales-Pérez, C. A. Reyes-García, J. A. Gonzalez, O. F. Reyes-Galaviz, H. J. Escalante, and S. Orlandi, “Classifying infant cry patterns by the Genetic Selection of a Fuzzy Model,” *Biomed. Signal Process. Control*, vol. 17, pp. 38–46, 2015.

[46] J. Saraswathy, M. Hariharan, S. Yaacob, and W. Khairunizam, “Automatic classification of infant cry: A review,” *2012 Int. Conf. Biomed. Eng. ICoBE 2012*, no. February, pp. 543–548, 2012.

[47] F. Feier, I. Enatescu, C. Ilie, and I. Silea, “Newborns’ cry analysis classification using signal processing and data mining,” *2014 Int. Conf. Optim. Electr. Electron. Equipment, OPTIM*

2014, pp. 880–885, 2014.

[48] "Google Audioset." [Online]. Available: https://research.google.com/audioset/.

[49] J. J. Parga, S. Lewin, J. Lewis, D. Montoya-Williams, A. Alwan, B. Shaul, C. Han, S. Y. Bookheimer, S. Eyer, M. Dapretto, L. Zeltzer, L. Dunlap, U. Nookala, D. Sun, B. H. Dang, and A. E. Anderson, "Defining and distinguishing infant behavioral states using acoustic cry analysis: is colic painful?," *Pediatr. Res.*, vol. 87, no. 3, pp. 576–580, 2020.

[50] "ffmpeg website." [Online]. Available: https://www.ffmpeg.org/.

[51] "Transcriber homepage." [Online]. Available: https://catalog.ldc.upenn.edu/LDC2005S15.

[52] O. Reyes-Galaviz and C. Reyes-Garcia, "A system for the processing of infant cry to recognize pathologies in recently born babies with neural networks," in *9th Conference Speech and Computer*, 2004.

[53] R. Sahak, W. Mansor, Y. K. Lee, A. I. M. Yassin, and A. Zabidi, "Performance of Combined Support Vector Machine and Principal Component Analysis in recognizing infant cry with asphyxia," *2010 Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBC'10*, pp. 6292–6295, 2010.

[54] R. Sahak, Y. K. Lee, W. Mansor, A. I. M. Yassin, and A. Zabidi, "Detection of asphyxiated infant cry using support vector machine integrated with principal component analysis," *Proc. 2010 IEEE EMBS Conf. Biomed. Eng. Sci. IECBES 2010*, no. December, pp. 485–488, 2010.

[55] A. Zabidi, I. M. Yassin, H. A. Hassan, N. Ismail, M. M. A. M. Hamzah, Z. I. Rizman, and H. Z. Abidin, "Detection of asphyxia in infants using deep learning Convolutional Neural Network (CNN) trained on Mel Frequency Cepstrum Coefficient (MFCC) features extracted from cry sounds," *J. Fundam. Appl. Sci.*, 2018.

[56] A. Osmani, M. Hamidi, and A. Chibani, "Machine learning approach for infant cry

interpretation," in *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, 2018.

[57] Z. W. YS. Lei, "The Characteristic of Infant cries," *Natl. Conf. Man-Machine Speech Commun.*, 2011.

[58] R. Cohen and Y. Lavner, "Infant cry analysis and detection," in *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel, IEEEI 2012*, 2012.

[59] S. Asthana, N. Varma, and V. K. Mittal, "An investigation into classification of infant cries using modified signal processing methods," in *2nd International Conference on Signal Processing and Integrated Networks, SPIN 2015*, 2015.

[60] M. Hariharan, L. S. Chee, and S. Yaacob, "Analysis of infant cry through weighted linear prediction cepstral coefficients and probabilistic neural network," *J. Med. Syst.*, 2012.

[61] "Deep Learning." [Online]. Available: https://en.wikipedia.org/wiki/Deep_learning.

[62] M. Moharir, M. U. Sachin, R. Nagaraj, M. Samiksha, and S. Rao, "Identification of asphyxia in newborns using GPU for deep learning," in *2017 2nd International Conference for Convergence in Technology, I2CT 2017*, 2017.

[63] H. K. Vorperian, R. D. Kent, L. R. Gentry, and B. S. Yandell, "Magnetic resonance imaging procedures to study the concurrent anatomic development of vocal tract structures: Preliminary results," *Int. J. Pediatr. Otorhinolaryngol.*, vol. 49, no. 3, pp. 197–206, 1999.

[64] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *J. Acoust. Soc. Am.*, 1999.

[65] R. D. Kent and A. D. Murray, "Acoustic features of infant vocalic utterances at 3, 6, and 9 months," *J. Acoust. Soc. Am.*, vol. 72, no. 2, pp. 353–365, 1982.

[66] J. R. Pruett, S. Kandala, S. Hoertel, A. Z. Snyder, J. T. Elison, T. Nishino, E. Feczko, N. U. F. Dosenbach, B. Nardos, J. D. Power, B. Adeyemo, K. N. Botteron, R. C. McKinstry, A.

C. Evans, H. C. Hazlett, S. R. Dager, S. Paterson, R. T. Schultz, D. L. Collins, V. S. Fonov, M. Styner, G. Gerig, S. Das, P. Kostopoulos, J. N. Constantino, A. M. Estes, S. E. Petersen, B. L. Schlaggar, and J. Piven, "Accurate age classification of 6 and 12 month-old infants based on resting-state functional connectivity magnetic resonance imaging data," *Dev. Cogn. Neurosci.*, vol. 12, pp. 123–133, 2015.

[67] M. P. Robb, F. Yavarzadeh, P. J. Schluter, V. Voit, W. Shehata-Dieler, and K. Wermke, "Laryngeal constriction phenomena in infant vocalizations," *J. Speech, Lang. Hear. Res.*, 2020.

[68] T. Fitch, "Production of Vocalizations in Mammals," in *Encyclopedia of Language & Linguistics*, 2006.

[69] I. R. Titze, "The physics of small-amplitude oscillation of the vocal folds," *J. Acoust. Soc. Am.*, vol. 83, no. 4, pp. 1536–1552, 1988.

[70] "Tensorflow homepage." [Online]. Available: https://www.tensorflow.org/.

[71] "Sox Homepage." [Online]. Available: https://en.wikipedia.org/wiki/SoX.

[72] I. A. Banica, H. Cucu, A. Buzo, D. Burileanu, and C. Burileanu, "Automatic methods for infant cry classification," *IEEE Int. Conf. Commun.*, vol. 2016-Augus, pp. 51–54, 2016.

[73] S. Bano and K. M. Ravikumar, "Decoding baby talk: A novel approach for normal infant cry signal classification," *Proc. IEEE Int. Conf. Soft-Computing Netw. Secur. ICSNS 2015*, pp. 24–26, 2015.

[74] N. Takahashi, M. Gygli, B. Pfister, and L. Van Gool, "Deep convolutional neural networks and data augmentation for acoustic event recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2016.

[75] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph Neural

Networks: A Review of Methods and Applications," 2018. [Online]. Available: http://arxiv.org/abs/1812.08434.

[76] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc.*, pp. 1–14, 2017.

[77] S. Dokania and V. Singh, "Graph Representation learning for Audio & Music genre Classification," 2019. [Online]. Available: http://arxiv.org/abs/1910.11117.

[78] S. Zhang, Y. Qin, K. Sun, and Y. Lin, "Few-shot audio classification with attentional graph neural networks," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2019-Septe, pp. 3649–3653, 2019.

[79] L. Le, A. N. M. H. Kabir, C. Ji, S. Basodi, and Y. Pan, "Using Transfer Learning, SVM, and Ensemble Classification to Classify Baby Cries Based on Their Spectrogram Images," in *Proceedings - 2019 IEEE 16th International Conference on Mobile Ad Hoc and Smart Systems Workshops, MASSW 2019*, 2019.

[80] E. Franti, I. Ispas, and M. Dascalu, "Testing the Universal Baby Language Hypothesis - Automatic Infant Speech Recognition with CNNs," *2018 41st Int. Conf. Telecommun. Signal Process. TSP 2018*, pp. 1–4, 2018.

[81] C. Y. Chang and J. J. Li, "Application of deep learning for recognizing infant cries," *2016 IEEE Int. Conf. Consum. Electron. ICCE-TW 2016*, 2016.

[82] M. Moharir, M. U. Sachin, R. Nagaraj, M. Samiksha, and S. Rao, "Identification of asphyxia in newborns using GPU for deep learning," *2017 2nd Int. Conf. Converg. Technol. I2CT 2017*, vol. 2017-Janua, pp. 236–239, 2017.

[83] H. Lim, J. Park, K. Lee, and Y. Han, "Rare Sound Event Detection Using 1D Convolutional Recurrent Neural Networks," *Dcase 2017 Proc.*, no. November, pp. 2–6, 2017.

[84] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," 2021.

[Online]. Available: http://arxiv.org/abs/2104.01778.

[85] L. Nanni, A. Rigo, A. Lumini, and S. Brahnam, “Spectrogram classification using dissimilarity space,” *Appl. Sci.*, vol. 10, no. 12, pp. 1–17, 2020.

[86] K. Palanisamy, D. Singhania, and A. Yao, “Rethinking CNN Models for Audio Classification,” 2020.

[87] C. Y. Chang and L. Y. Tsai, “A CNN-Based Method for Infant Cry Detection and Recognition,” in *Advances in Intelligent Systems and Computing*, 2019.

[88] Y. Hwang, H. Cho, H. Yang, D.-O. Won, I. Oh, and S.-W. Lee, “Mel-spectrogram augmentation for sequence to sequence voice conversion,” 2020.

[89] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, “Gelp: GAN-excited linear prediction for speech synthesis from mel-spectrogram,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2019-Septe, pp. 694–698, 2019.

[90] L. Sheng, D.-Y. Huang, and E. N. Pavlovskiy, “High-quality Speech Synthesis Using Super-resolution Mel-Spectrogram,” 2019. [Online]. Available: http://arxiv.org/abs/1912.01167.

[91] S. S. R. Phaye, E. Benetos, and Y. Wang, “SubSpectralNet - Using Sub-spectrogram Based Convolutional Neural Networks for Acoustic Scene Classification,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2019.

[92] T. Zhang, G. Feng, J. Liang, and T. An, “Acoustic scene classification based on Mel spectrogram decomposition and model merging,” *Appl. Acoust.*, vol. 182, p. 108258, 2021.

[93] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, “Attentional Feature Fusion,” 2021. .

[94] Y. Chen, J. Tao, L. Liu, J. Xiong, R. Xia, J. Xie, Q. Zhang, and K. Yang, “Research of improving semantic image segmentation based on a feature fusion model,” *J. Ambient Intell.*

Humaniz. Comput., no. 3, 2020.

[95] X. Xu, Y. Wang, D. Xu, Y. Peng, C. Zhang, J. Jia, and B. Chen, “AMFFCN: Attentional Multi-layer Feature Fusion Convolution Network for Audio-visual Speech Enhancement,” 2021. [Online]. Available: http://arxiv.org/abs/2101.06268.

[96] I. McLoughlin, Z. Xie, Y. Song, H. Phan, and R. Palaniappan, “Time–Frequency Feature Fusion for Noise Robust Audio Event Classification,” *Circuits, Syst. Signal Process.*, vol. 39, no. 3, pp. 1672–1687, 2020.

[97] C. M. Chang, H. Y. Chen, H. C. Chen, and C. C. Lee, “Sensing with Contexts: Crying Reason Classification for Infant Care Center with Environmental Fusion,” *2020 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA ASC 2020 - Proc.*, pp. 314–318, 2020.

[98] C. Ji, M. Chen, B. Li, and Y. Pan, “Infant Cry Classification with Graph Convolutional Networks,” in *2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS)*, 2021, pp. 322–327.

[99] A. Radoi and C. Burileanu, “Infant Cry Classification Using Compression-Based Similarity Metric,” in *2018 International Conference on Communications (COMM)*, 2018, pp. 67–70.

[100] B. K. Baniya and J. Lee, “Rough set-based approach for automatic emotion classification of music,” *J. Inf. Process. Syst.*, 2017.

[101] J. Zhang, J. S. Wong, T. Li, and Y. Pan, “A comparison of parallel large-scale knowledge acquisition using rough set theory on different MapReduce runtime systems,” *Int. J. Approx. Reason.*, 2014.

[102] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, and S. Yu, “A survey on incorporating domain knowledge into deep learning for medical image analysis,” *Med. Image Anal.*, vol. 69, pp. 1–27, 2021.