Use of Prosodic Features in Infant Cry Diagnostic System

by

Fatemeh SALEHIANMATIKOLAIE

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE IN PARTIAL FULFILLMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, DECEMBER 17TH 2021

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Chakib Tadj, Thesis supervisor
Department of Electrical Engineering, École de technologie supérieure

Mr. Patrick Cardinal, President of the board of examiners
Department of Software Engineering and IT, École de technologie supérieure

Mr. Christian Gargour, Member of the jury
Department of Electrical Engineering, École de technologie supérieure

Mr. Mohand Said Allili, External examiner
Department of Electrical Engineering, Université du québec en outaouais

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON DECEMBER 7TH 2021

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

This work would not have been possible without the support of many people. First and foremost, I would like to express my special appreciation to my supervisor, Professor Chakib Tadj, for his outstanding support and advice in my doctoral study. Professor Tadj's support is endless. He generously gave wise advice, a positive attitude, professionalism and tremendous experience. Altogether, this made working with Professor Tadj a great pleasure and honour. Thank you for enriching my knowledge, both professionally and personally.

Special thanks to my examiners: Professor Patrick Cardinal, Professor Christian Gargour, and Professor Mohand Said Allili. Thanks for reviewing and evaluating my thesis.

I am also grateful to all former colleagues who worked on the development of the database. I also would like to thank Dr. Yasmina Kheddache for her collaboration with my third paper.

A special acknowledgment to Mr. Prasun Lala from Sara writing group at ETS for his feedback on the papers that are presented in this thesis.

I would like to dedicate this thesis to my parents and my entire family. I would like to extend my most sincere appreciation to my dad for his consistent support during this stage of my life. This was simply not possible without my family's encouragement and unconditional love. Thank you very much for all of your love.

Utilisation des caractéristiques prosodiques dans le système de diagnostic des pleurs des nourrissons

Fatemeh SALEHIANMATIKOLAIE

RÉSUMÉ

Le signal sonore de cri (Cry Audio Signal, CAS) du nouveau-né est constitué d'un son rythmique. Imaginez que les nouveau-nés ne pleurent pas ; dans ce cas, nous n'aurions aucun moyen de les comprendre. Leurs cris expriment la faim, la douleur, la maladie ou simplement le besoin d'un câlin. Lorsqu'un parent entend les pleurs d'un nouveau-né, des hormones de stress sont libérées dans son corps ce qui entraîne une augmentation de la pression artérielle, du rythme cardiaque et de la tension musculaire ; il essaie donc d'arrêter les pleurs en soulageant le nouveau-né. Les pleurs sont expliqués comme un signal graduel qui constitue un stimulus dans le système comportemental. Les nouveau-nés peuvent susciter la réaction de leur entourage en pleurant, et les pleurs des nouveau-nés sont donc considérés comme un comportement précoce de survie dans le système comportemental.

La recherche sur le CAS des nouveau-nés pour étudier le potentiel des caractéristiques discriminantes a commencé dans les années 1960. Elle a commencé par des investigations auditives subjectives et, fait intéressant, plusieurs rapports ont montré que les mères et le personnel hospitalier pouvaient souvent distinguer les besoins des nouveau-nés uniquement en les écoutant. L'étude a ensuite été suivie par des analyses de temps, de fréquence et de domaines spectrographiques. Des chercheurs ont aussi constaté que les CAS des nouveau-nés contenaient des informations sur l'état physique et psychologique du nouveau-né. Les chercheurs ont même décrit comment le cerveau du nouveau-né modifie la traction de ses cordes vocales par l'intermédiaire des nerfs crâniens - établissant un lien entre les pleurs et le cerveau.

Ces examens ont permis de révéler des schémas distinctifs qui déterminent les caractéristiques du groupe. Pour éviter la tâche fastidieuse de l'analyse par des humains d'une grande quantité d'informations contenues dans les CAS des nouveau-nés, une analyse automatisée a été proposée. Un tel système peut considérablement réduire le temps d'investigation et les classer automatiquement. Ainsi des modèles d'apprentissage automatique ont été introduits afin d'établir des statistiques.

Cette thèse vise à développer le système de diagnostic des pleurs du nouveau-né (Newborn Cry Diagnosis System, NCDS) afin d'identifier automatiquement les CAS des nourrissons qu'ils soient malades ou bien portants sans examen physique du nouveau-né. Un NCDS comprend trois étapes principales : Le prétraitement, l'extraction de caractéristiques et l'entraînement du modèle pour la classification. La recherche présentée ici explore les modèles à différents niveaux des CAS des nouveau-nés dans la phase d'extraction des caractéristiques. L'analyse comprend la recherche d'informations à court et à long terme dans les CAS du nouveau-né pour trouver des caractéristiques potentielles à caractère pathologique. Notre principale contribution dans ce travail est l'utilisation des caractéristiques prosodiques pour étudier les modèles statistiques à long terme dans les CAS des nouveau-nés. Nous avons exploré l'efficacité des ensembles de caractéristiques de rythme, d’inclinaison et d’intensité dans le NCDS. Les ensembles de caractéristiques prosodiques d’inclinaison et de rythme n’ont jamais été étudiés dans le NCDS. Il a été constaté que les informations de haut niveau, à savoir les caractéristiques prosodiques, améliorent la capacité de discrimination des signaux audio dans les systèmes de reconnaissance de la parole et du langage.

En ce qui concerne les ensembles de caractéristiques à court terme, l’ensemble de caractéristiques commun examiné avec succès dans le NCDS est celui des coefficients cepstraux de fréquence Mel (Mel Frequency Cepstral Coefficients, MFCC). Une autre innovation dans ce travail est l’utilisation,) pour la première fois dans le NCDS ,de l’ensemble de caractéristiques à court terme de la modulation d’amplitude inspirée par l’auditoire (Auditory-inspired Amplitude Modulation, AAM).

Notre objectif était de comparer la fonctionnalité de l’ensemble de caractéristiques AAM dans le NCDS avec l’ensemble de caractéristiques examiné le plus influent, le MFCC, et d’explorer le potentiel de fusion de cet ensemble de caractéristiques avec le MFCC et l’ensemble de caractéristiques prosodiques.

Les performances de chaque ensemble de caractéristiques ont été évaluées à l’aide d’une série de classificateurs, dont la machine à vecteurs de support, l’arbre de décision, le réseau neuronal à perceptron et l’analyse discriminante. Nous avons également examiné la méthode du vote majoritaire pour améliorer les résultats de la classification, ce qui n’a pas été rapporté auparavant dans la littérature relative au développement d’un NCDS.

Notre étude s’est principalement concentrée sur deux pathologies critiques, la détresse respiratoire et la septicémie, qui sont les 11e et 6e causes de décès au Canada. Au final, nous avons abouti à un modèle complet englobant 34 pathologies courantes chez les nouveau-nés.

Mots-clés: Caractéristiques prosodiques, Rythme, Mélodie, Intensité, Coefficient cepstral de fréquence mélodique, Caractéristiques de modulation d’amplitude inspirées de l’audition, Système de diagnostic des pleurs du nouveau-né, Pleurs du nouveau-né, Pleurs d’expiration et d’inspiration.

Use of Prosodic Features in Infant Cry Diagnostic System

Fatemeh SALEHIANMATIKOLAIE

ABSTRACT

The newborn's Cry Audio Signal (CAS) is made up of a rhythmic sound. Imagine that the newborns would not cry; in this case, we had no way of understanding the newborn's needs. Needs like hunger, pain, illness, or just the need to hug. When a parent hears the sound of a newborn crying, stress hormones are released into the parent's body, which leads to high blood pressure, heart rate, and muscle tension, and thus the parent tries to stop crying by alleviating the newborn. Crying is explained as a graded signal that is a stimulus in the behavioural system. Newborns can elicit the surrounding people's reaction by crying, so newborns' crying is regarded as an early behaviour for survival in the behavioural system.

The cry-researchers found the newborns' CASs having concealed information about the newborn's physical and psychological states. The newborns' brain changes the amount of traction in the vocal cords through the cranial nerves. Because the cranial nerves control crying, the cry-researchers made a connection between crying and the brain. The research on newborns' CAS to investigate the potential of discriminating characteristics started in the 1960s. It started with the subjective auditory investigations, and interestingly, several reports showed that mothers and the hospital staff often could distinguish the needs of newborns only by listening to them. The investigation was then followed by time, frequency, and spectrographic domains analyses. Through these examinations, distinctive patterns were revealed that determine group characteristics. Finally, to avoid the tedious task of analyzing a large amount of information in newborns' CASs by humans, automated machine-based analysis was proposed. Such a system for analyzing newborns' CASs can considerably speed up the investigation time and automatically classify them. This is where machine learning models were introduced to capture the statistics in the newborns' CASs.

This thesis aims to develop the Newborn Cry Diagnostic system (NCDS) to automatically identify sick infants' CASs from healthy ones without any newborn physical examination. An NCDS includes three main stages of preprocessing, feature extraction, and model training for classification. This research presented here explores patterns at different levels of newborns' CASs in the feature extraction phase. The analysis includes investigating the short-term and long-term information in the newborn's CASs for potential pathologically informed features. Our main contribution in this work is the use of the prosodic features to investigate the long-term statistical patterns in newborns' CASs. We explored the effectiveness of rhythm, tilt, and intensity feature sets in NCDS. The prosodic feature sets of tilt and rhythm have never been studied in NCDS. The high-level information, namely prosodic features, was found to improve the discriminative ability within audio signals in speech and language recognition systems.

Regarding the short-term feature sets, the common feature set successfully examined in NCDS is Mel Frequency Cepstral Coefficients (MFCC). Another innovation of this work is that we employed the short-term feature set of Auditory-inspired Amplitude Modulation (AAM) for the first time in the NCDS. Our goal was to compare the functionality of the AAM feature set in NCDS with the most influential examined feature set of MFCC and explore the fusion potential of this feature set with MFCC and the prosodic feature set.

The performance of each feature set was evaluated using a collection of classifiers, including support vector machine, decision tree, perceptron neural network and discriminant analysis. We also examined the majority voting method to upgrade the classification results, which has not previously been reported in the literature relating to developing an NCDS.

Our study primarily focused on two critical pathologies of respiratory distress and sepsis, ranking as the 11th and sixth leading causes of death in Canada. In the end, we came up with a comprehensive model encompassing 34 pathologies common among newborns.

Keywords: Prosodic features, Melody, Rhythm, Intensity, Mel-frequency Cepstral Coefficient, Auditory-inspired Amplitude Modulation features, Newborn cry diagnostic system, Newborn infant cries, Expiration and inspiration cry.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABREVIATIONS

AAM	Auditory-inspired Amplitude Modulation
AE	All Episodes
ALI	Automatic Language Identification
ASR	Automatic Speech Recognition
BPSO	Binary Particle Swarm Optimization
CAS	Cry Audio Signal
CFS	Correlation-based Feature Selection
CNS	Consistency-based Feature Selection
DFT	Discrete Fourier Transform
EXP	Expiration
FCBF	Fast Correlation-Based Filter
FN	False Negative
FP	False Positive
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
IDFT	Inverse Discrete Fourier Transform
INSV	Inspiration Voiced
LFCC	Linear Frequency Cepstrum Coefficients
LPC	Linear Predictive Coefficients

MFCC	Mel-frequency Cepstral Coefficients
MLP	Multilayer Perceptron
NN	Neural Network
nrPVI	(anglais, Pulsed Shaping)
OneR	One Rule
PCA	Principle Component Analysis
PNN	Probabilistic Neural Network
RBF	Radial basis function
RBFN	Radial basis function Network
rPVI	raw Pairwise Variability Index
RDS	Respiratory Distress Syndrome
ROC	Receiver Operating Characteristic
SE	Single Episode
STDFT	Short-Time Discrete Fourier Transform
SVM	Support vector machine
TN	True Negative
TP	True Positive

INTRODUCTION

0.1 Context of research work

Nearly 140 million newborns are born worldwide a year. Newborns are entirely dependent on their parents or the adults around them to provide them with food and comfort them. Crying, like speaking, is a way for newborns to express their needs and request attention for care.

(LaGasse, Neal & Lester (2005)) describes, *"Crying is a biological siren which alerts the caregiving environment about the needs and wants of the infant and motivating the listener to respond"*. The investigation to translate newborns' Cry Audio Signal (CAS) has been of interest in history. Perhaps the very late thought about secrets in crying is the mythological belief that the saints had cried already before birth (Sirviö & Michelsson (1976)). In the documentation of the infants' CAS research, the study begins with the attempts to find the auditory patterns and then is evolved to the spectral analysis (Lynip (1951)). (Sirviö & Michelsson (1976)) anecdotally explains that the initial findings in 1962 revealed that newborns' vocal behaviour during crying is different evidently for four different reasons of birth, pleasure, hunger, and pain CASs. Almost a decade later, the findings of (Michelsson (1971)) revealed that asphyxiated newborns' vocal CAS behaviour have characteristics that differ from the healthy ones, which was the very first report to initiate the idea of the potential of newborns' CAS characteristics as a symptom between healthy and unhealthy newborns.

In the clinical setting, the evaluations show that mothers and also the medical staff who are dealing more often with newborns can understand the needs of the newborns based on their auditory experiences (LaGasse *et al.* (2005); Mukhopadhyay, Saha, Majumdar, Majumdar, Gorain, Arya, Bhattacharya & Singh (2013); Sagi (1981)). Newborns' CAS thus is a multimodal signal that varies the information proportionately to the condition of the newborn.

Through these findings, the infant cry-researchers are now confident that different levels of information are conveyed in infants' CASs. In the domain of newborns' CAS analysis, the most influential attempt has been in developing diagnostically practical machine learning frameworks for the early diagnosis and treatments in newborns. Although there have been some accomplishments to develop a Newborn Cry Diagnostic System (NCDS), the progress in analyzing the newborns' CASs for diagnostic purposes is not yet as developed as other audio processing domains. The complication of finding more pathologically informed patterns in the newborns' CASs remains in an open and undeveloped state.

0.2 Statement of research problem

Over the last few years, there has been increasing attention in promoting the machine learning-based diagnostic system. Automated diagnostic systems have become highly prevalent in today's application domains. Moreover, when it comes to newborns, developing an NCDS is even of more profound concern as they only communicate by crying. Statistically, in recent years, the infant mortality rate in developed countries has decreased. However, this rate is still high in developing countries. Table 0.1 shows the infant death rate in some countries worldwide (CIA).

Table 0.1 Infant mortality rate in some countries. The sign * in the table shows the countries of infants whose CASs were used in this thesis (CIA)

Rank	**Country**	**Deaths/1k live birth**
1	Afghanistan	106.75
2	Somalia	88.03
3	Centra African Republic	84.22
184	Canada *	4.44
161	Lebonon *	7.20

From the clinical procedure point of view, the early diagnosis of diseases, such as hearing disorders, sepsis, and central nervous system disorders, etc. helps the treatment process and probably increases the chance of recovery and reduces the risk of severe health problems later in life (Kheddache & Tadj (2019); Lester & LaGasse (2008); Varallyay, Benyó, Illényi, Farkas & Kovács (2004)). Most notably, in an acute situation, specific characteristics in the newborns' CASs can be explained as sudden death, which can be prognosticated and prevent the newborn's death. The possibility of predictability highlights the importance of regarding newborns' CAS as an acoustic symptom for an early diagnosis. Saving newborns' lives and promoting their health is of particular importance in the health of any nation and for further providing health services. Hence, it would be reasonable and practical to design an NCDS similar to an Automatic Speech Recognition (ASR) system or other audio processing models to detect the newborn's CAS warning cues and thus prevent the damage.

Although various model configurations for NCDSs have been developed, the growth in this domain is still not as advanced as other audio processing domains. In recent years, there have been consistent attempts to promote the proliferation of a more affordable, easy-to-use, reliable, and powerful NCDS. This thesis endeavoured to take the NCDS one step forward by enriching it by introducing more statistically distinctive features using prosodic features.

0.3 Research objectives

This research work aims to develop a non-invasive automatic system that we call NCDS in this write-up. An NCDS uses machine learning techniques to distinguish between healthy and unhealthy newborns based on the their CASs. Compared with other identification and diagnostic systems, there is far less research on developing an NCDS. This is mainly due to the small newborns' CAS dataset in this domain. In the audio processing system, the machine learning researchers vigorously apply deep learning methods due to their very assuring results. However, deep learning configurations require a large amount of dataset for the system to train itself and get statistics within the data.

In the present problem of designing an NCDS, we face the lack of a large CAS dataset from different newborns; thus, deep learning algorithms would not have adequate training samples to have that much insight to learn patterns in the data by itself. Hence, our method is to employ the traditional machine learning approaches to design the steps for the NCDS. For this work, the objective is to introduce some hand-designed feature sets to give helpful knowledge to the model.

Generally, the goal is to design the traditional pipeline framework as the following:

1. Creating the feature extraction phase for the NCDS using features from different levels of information, including short-term and long-term feature sets.
2. Evaluating the performance of these new feature sets for diagnostic purposes.
3. Evaluating the performance of the fusion of short-term and long-term feature sets likewise for diagnostic purposes.
4. Comparing the efficacy of the newly introduced feature extraction phase with previous ones.
5. Use of multiple models for evaluating each feature set and making the final decision by use of majority voting.

0.4 Methodology

As a common practice, pediatricians diagnose infant sickness based on several physical examinations. The idea of identifying sick infants based on their CASs may seem challenging at first glance. However, the fact that mothers and individuals in contact with an infant appear able to predict some of the infant's needs based on their CASs (Moller & Schonweiler (1999); Mukhopadhyay *et al.* (2013)) gives the motivation to develop a system able to determine the infant's illness. Before explaining the steps for developing an NCDS, a brief explanation of the dataset is needed.

0.4.1 Database Arrangement and Specification

The primary stage for developing an automatic recognition system is data acquisition. The CASs of 769 newborns have been recorded from the hospitals of Al-Sahel, Al-Raee of Lebanon, and Ste-Justine of Canada. In the recording procedure, a two-channel sound recorder with a sampling frequency of 44.1 kHz and a resolution of 16 bits was placed at a distance between 10 and 30 cm from the infant. The recordings' length of each record is in the range of two to three minutes. The CASs in the database are either of healthy infants or ones afflicted by one or some diseases. There are 96 types of diseases in the database. For some diseases, the number of infant is limited to one baby with several CASs. Every CAS in the database is arranged based on the following information:

- **Reason of Crying:** includes Abdominal Pain, Aerosol, After Shower, Aspirating of Secretions, Birth CAS, Blood Test, CCVP removal, Circumcision, Cold, Collection of urine, Cranial Ultrasound examination, Diabet Test, Diaper, Discomfort, Eye Pads installation, Fear, Feeding tube, Fever, Glucometer, Hemo test, Hunger, Inserting Nasogastric Tube, Irritable, Irritation and nasal hygiene, IV, IV Installation, Kicked by his twin, Manipulation, Medical Exam, Nasal Cpap, Nebulizer, Neonatal Screening, Ophthalmologist exam, Pain, PH meter, Radiology exam, Reflux, Saturo Installation, Screening Test, Section, Shower, Temperature taking, Treatment, Usual Treatment, Vaccination, vital sign monitoring, Vomiting, Weighting and unknown.

- **Gestational Age:** is in the range of 27 weeks and two days to 41 weeks and four days.

- **Birth weight:** starts from 0.98 kilograms to 5.2 kilograms.

- **Race variety:** includes half-Caucasian and half-Haitian, African, Arabic, Caucasian, Latino, Native Hawaiian and Quebecois.

- **Baby's age during recording:** is in the range of one day to 208 days.

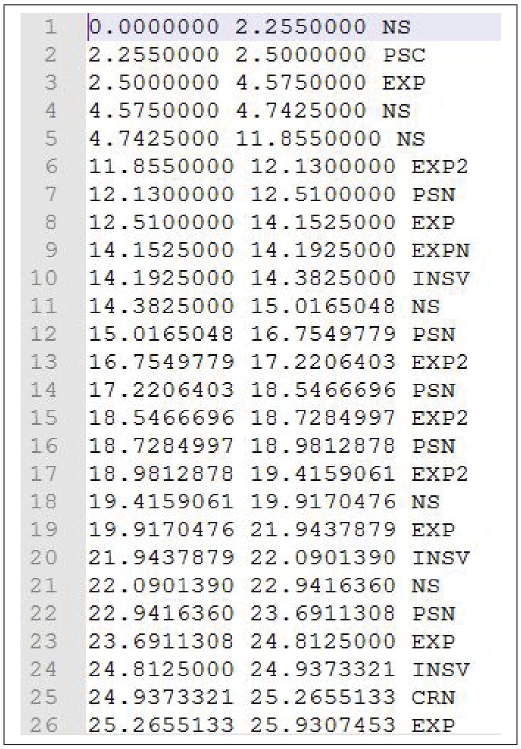
Other details are reporting the gender of the infants, the health condition, and the Apgar test score[1]. The portions of the infants' CASs in the database were labelled in previous work (Abou-Abbas, Tadj, Gargour & Montazeri (2016)). The assigned labels and their descriptions are listed in Table 0.2. The labels were assigned using the WaveSurfer software tool. This program has options for visualizing the waveform and the spectrogram, as well as manual labelling. The manual annotation file is also available for each recording. An example of this file for a portion of a CAS is shown in Figure 0.1.

Table 0.2 The annotated labels to the CAS portions in the database and their descriptions

Labels	**Description**
EXP	Voiced expiration segment during a period of CAS
EXPN	Unvoiced expiration segment during a period of CAS
INS	Unvoiced inspiration segment during a period of CAS
INSV	Voiced inspiration segment during a period of CAS
EXP2	Voiced expiration segment during a period of pseudo-CAS
INS2	Voiced inspiration segment during a period of pseudo-CAS
PSEUDOCRY	Any sound generated by the baby and it is not a CAS
Speech	Sound of the nurse or parents talking around
Background	The kind of noise so low, it is characterized by a very low power-silence affected with little noise
BIP	The sound of the medical instruments next the baby
Noisy CAS	Any sound heard with the CAS: machine's beep sound, water, diaper, etc.
Noisy pseudo-CAS	Any sound heard with the pseudo-CAS
Noise	Like the sound caused by the mic moved by someone, the diaper, a door sound, speech + background, speech +beep.

An example of the labels in a portion of a CAS is portrayed in Figure 0.2. The waveform of the recorded CAS and the corresponding labels in the yellow medium are shown.

[1] The Apgar test is the first test after birth for determining the infant's health condition



```
0.0000000 2.2550000 NS
2.2550000 2.5000000 PSC
2.5000000 4.5750000 EXP
4.5750000 4.7425000 NS
4.7425000 11.8550000 NS
11.8550000 12.1300000 EXP2
12.1300000 12.5100000 PSN
12.5100000 14.1525000 EXP
14.1525000 14.1925000 EXPN
14.1925000 14.3825000 INSV
14.3825000 15.0165048 NS
15.0165048 16.7549779 PSN
16.7549779 17.2206403 EXP2
17.2206403 18.5466696 PSN
18.5466696 18.7284997 EXP2
18.7284997 18.9812878 PSN
18.9812878 19.4159061 EXP2
19.4159061 19.9170476 NS
19.9170476 21.9437879 EXP
21.9437879 22.0901390 INSV
22.0901390 22.9416360 NS
22.9416360 23.6911308 PSN
23.6911308 24.8125000 EXP
24.8125000 24.9373321 INSV
24.9373321 25.2655133 CRN
25.2655133 25.9307453 EXP
```

Figure 0.1 Manual transcription of an example CAS in the database

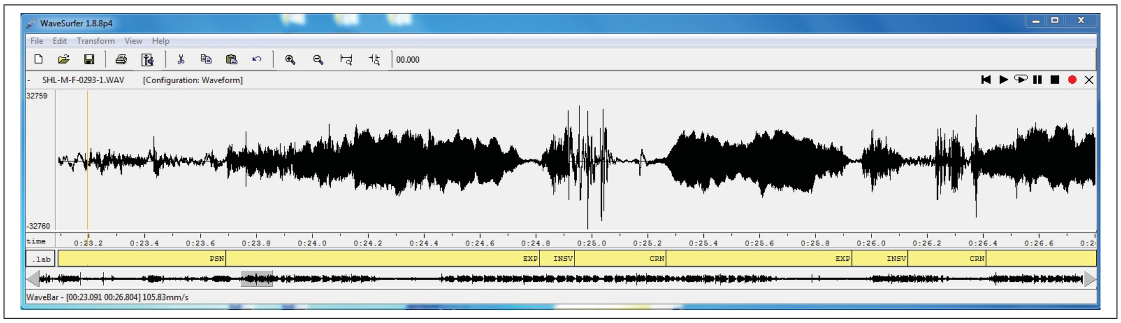


Figure 0.2 An example of a labeled CAS in WaveSurfer software Medium

0.4.2 Proposed Methodologies

The task of the NCDS is to seek an acoustic signature for the CASs of groups of healthy and unhealthy infants. As mentioned earlier, due to the lack of a large dataset, we proposed designing the entire stages of the NCDS, including preprocessing, feature extraction, and classification. An NCDS is a signal-based diagnostic system; its input is the CASs of the newborns, and its output is a predicted label showing that the newborn is healthy or unhealthy. Figure 0.3 shows the stages of our proposed NCDS.

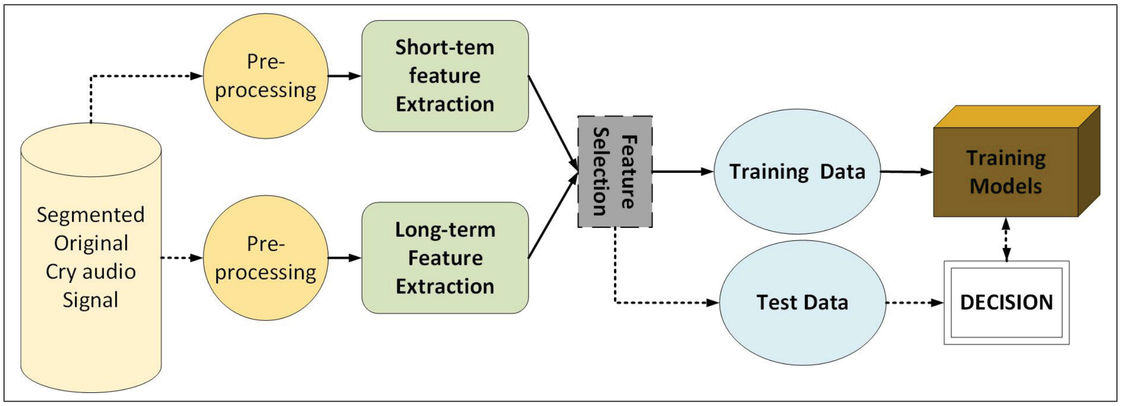

Segmented Original Cry audio Signal
Pre-processing
Pre-processing
Short-tem feature Extraction
Long-term Feature Extraction
Feature Selection
Training Data
Test Data
Training Models
DECISION

Figure 0.3 Block diagram of proposed NCDS

For developing an NCDS, the following methodologies were employed:

1. In the preprocessing step, the infants' CASs are segmented according to the labels in Table 0.2. Then the required segments are fetched for further analysis. In this project, the segments of inspiration and expiration are used. In Table 0.2, these segments are called "EXP" and "INSV". According to several infant cry-researchers, these segments are of importance (Alaie, Abou-Abbas & Tadj (2016)) and hence are more amenable to being used for the analysis of groups of healthy and unhealthy infants.

   The requirement of applying other preprocessing applications varies according to the information level intended to be extracted in the next step (feature extraction). For short-term feature extraction, the infant's CAS is required to be of a smaller size. Hence, extra processing involves pre-emphasizing, windowing, and applying filter banks and other applications proportionate to the short-term feature extraction technique. The preprocessing steps are explained in detail in chapters 2, 3, and 4.

2. Next is the feature extraction phase. Different analyses of the infants' CASs are employed for extracting hidden statistical patterns at different levels. The process involves the widely used Mel-frequency Cepstral Coefficients (MFCC) and the Auditory-inspired Amplitude Modulation (AAM) feature set for extracting short-term patterns. The MFCCs are quite a standard technique in audio processing models such as ASR and Automatic Language Identification (ALI). This technique has been widely used in the NCDS as well. However, the AAM feature set has not been investigated in NCDS. It is the first time to investigate the performance of the AAM feature set. The AAM feature set has been successfully tested in other acoustic recognition systems such as nonverbal human-produced audio events (Bouserhal, Chabot, Sarria-Paja, Cardinal & Voix (2018)), speaker verification (Kinnunen, Lee & Li (2008)), and also was shown to outperform the widely used feature set of MFCC (Sarria-Paja & Falk (2017)).

   We extracted feature sets for investigating long-term cues in newborns' CASs, including melody (tilt), rhythm, and intensity. The melody and rhythm features have not been experimented with in the NCDS. The values used in these techniques were explained in detail in chapters 2 and 3, and in chapter 4. In this thesis, we call the melody feature set tilt. The long-term feature sets were expected to have satisfactory performance in classification; the assumption in this research was that the combination of these features would result in better system performance based on experiments in similar systems (Adami, Mihaescu, Reynolds & Godfrey (2003); Vicsi & Szaszák (2010)).

3. After feature extraction, we used feature selection in NCDS to choose the optimal number of features due to the high dimensionality of feature sets. The use of feature selection techniques was reported to save the computation time, and in some cases, it has increased the system accuracy (Sahak, Mansor, Lee, Yassin & Zabidi (2010a)). Our analysis investigated the usefulness of taking the statistics measures, and we also experimented with the well-known Principal Component Analysis (PCA) to present the feature sets to the model.

4. The last part is for the evaluation of the proposed feature sets and their combinations. In this part, different families of classifiers were used. These classifiers are of the families of decision trees, discriminant analysis, Support Vector Machine (SVM), and Perceptron Neural Network (PNN). For accurate functionality measurement, the F-score and accuracy were computed.

After system development and experimenting with the mentioned techniques in the NCDS, we investigated to obtain the optimal hyperparameters in the classification phase for SVM. We explored the popular method of grid search. Grid search is one of the most accustomed techniques for hyperparameters optimization for the learning algorithms. The advantage of the grid search method is that it is reasonable for problems with a few number of hyperparameters (Aufa, Suyanto & Arifianto (2020)). Thirty point values for hyperparameters were evaluated. Not only does the use of grid search not improve the system performance but also in some cases the suggested values for hyperparameters degraded the system performance. Thus in the experiments reported in chapter 2, 3 and 4, we reported the results with the default values of hyperparameters.

0.5 Organization of the thesis

This manuscript is a thesis by articles. The work carried out during this thesis is presented in the form of two published articles and one article submitted for publication in scientific journals. In the course of the research, developing an NCDS was addressed by the three articles.

The question of using the prosodic feature set to enhance the recognition power of the NCDS to categorize healthy and unhealthy infants' CASs was investigated in the first article. In the first article the CASs of infants with Respiratory distress (RDS) was the point of interest as there is a paucity of literature for infants with RDS.

Due to promising results in the first article with the prosodic feature sets, we were driven to add more prosodic features to the NCDS. Moreover, we found significant improvement in using the majority voting technique in the decision-making for our proposed NCDS. We also found it essential to address a particular pathology to test the system in a more practical way. So the second article is focused on sepsis in infants.

In the third article, it was of interest to investigate other feature sets as a complement or replacement of the conventional MFCCs and assess the system performance using the new proposed short-term feature set with prosodic feature sets. We decided to investigate further the usefulness of a combination of all feature sets with the methodology we have found before.

The following describes the organization of this thesis. The first chapter is devoted to the description of the phenomenon of newborns' CAS and literature reviews. This chapter introduces how the CAS is created, the different methods for analyzing the newborns' CAS, and the studied methods for identifying pathological CASs from healthy ones in the literature. Chapters two to four contains the following articles:

- Salehianmatikolaie, Fatemeh, et Chakib Tadj. 2020. «On the use of long-term features in a newborn cry diagnostic system». Biomedical Signal Processing and Control, vol. 59, p. 101889.

The second chapter contains the article mentioned above. This article was accepted and published in February 2020 in the journal of Biomedical Signal Processing and Control. In this article, for the first time, we introduced the melody feature set of tilt and the feature set of rhythm in NCDS. We also combined the prosodic feature sets of tilt and rhythm with the baseline feature set of MFCC. These sets of features were fed to the linear SVM classifier. In this study, we only focus on healthy infants and infants with respiratory distress pathology. We employed two sets of experiments on voiced expiration and voiced inspiration episodes of newborns' CASs.

- Salehianmatikolaie, Fatemeh, et Chakib Tadj. 2021. «Machine Learning-Based Cry Diagnostic System for Identifying Septic Newborns». Journal of Voice.

The third chapter explains the article mentioned above, submitted to the Journal of Voice in 2021. In this article, we added the intensity feature set to the tilt and rhythm feature sets that we introduced in the first article. Our contribution was twofold. One contribution is how we evaluate short-term and long-term feature sets and how we use these features to make a final decision. The second contribution is to look at the unstudied pathology of sepsis. We investigated the four feature sets of tilt, intensity, durational feature, and MFCC. The learning models of SVM, decision tree, and discriminant analysis were used.

The second article likewise compares the usefulness of two feature selection methods of PCA and statistical measures for beneficially representing the features.

- Salehianmatikolaie, Fatemeh, Kheddache, Yasmina et Chakib Tadj. 2021. « Automated Newborn Cry diagnostic system using Machine Learning Approach». Biomedical Signal Processing and Control.

The fourth chapter explains the third article mentioned above. It was accepted in the Journal of Biomedical Signal Processing and Control in 2021. This article proposed a holistic NCDS that resembles the real-world problem by including 34 clinical states of newborns. Our other innovation was to encompass the short-term feature set of AAM for the first time in NCDS. We applied the methodology of feature combination using the two learning algorithms of SVM and PNN.

Ultimately, the closing chapter is devoted to the summing-up of this thesis report and some additional research recommendations.

CHAPTER 1

LITERATURE REVIEW

1.1 How is the Infants' CAS Generated?

Infants produce CAS by pushing airflow from their lungs to the vocal track (Soltis (2004)), and then airflow vibrates the vocal cords, which generates the sound. Thus, lungs work like power and provide patterns. This explanation is called source-filter theory, which is modelled in Figure 1.1. In general, the infants' CAS results from the altered created sound of the source (vibrating larynx) by the vocal tract. The set from vocal cords to the lips forms the vocal tract (Rutledge (1995)). The form of the vocal tract adjusts the vocalization and works as a filter. It attenuates or amplifies some frequencies. Figure 1.2 illustrates the spectrums of an example of the produced sound and its modifications in each step. Understanding the infants' CAS generation is essential later for determining discriminative features. For instance, fundamental frequency and harmonics are the larynx features, and the formants (or resonance frequencies) correlate to the resonances in the vocal tract.

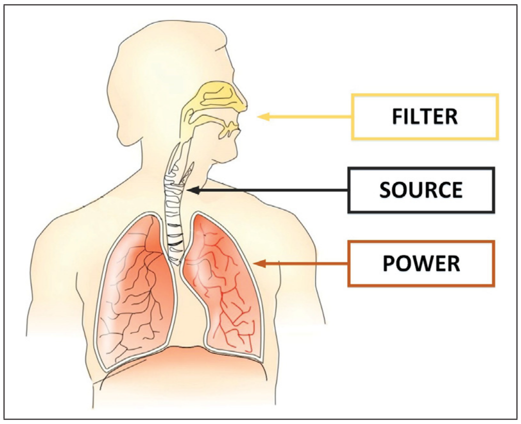


Figure 1.1 The illustration of sound production of an infant based on source filter theory

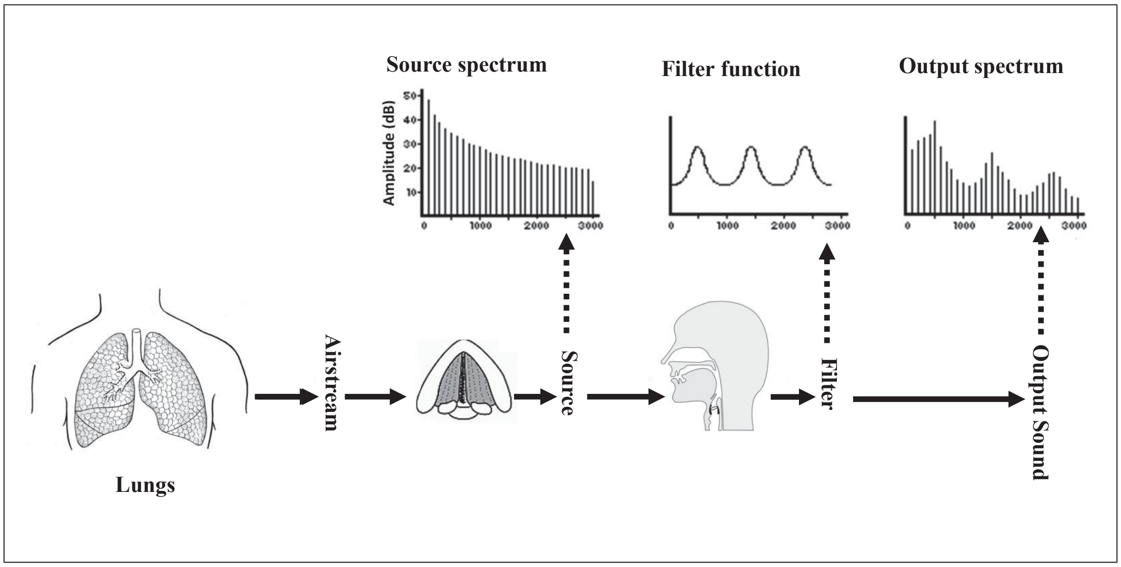


Figure 1.2 The spectrums of an example of the produced sound and its modifications in each step

To better understand the problem, it would be helpful to have a bit of background on how newborns' CAS is controlled physiologically. (Golub (1979)) explains three motors in the newborns' body for controlling the crying procedure. The first controller, which is attributed as "*upper processor*", correlates with the state of the infant, such as fussiness. The second controller, called "*middle processor*", regulates the state of the infant relating to "*vegetative states, such as swallowing, coughing, digestion and crying*". Lastly, the third controller, called "*the lowest processor*", orders the muscles of the face and larynx, which govern the act of crying. Accordingly, crying is an excellent source of information about the state of the infant (Varallyay Jr (2006)).

1.2 Infants' CAS Components

Before proceeding into the preliminary research on newborns' CASs, a brief introduction to the components of the CAS is required. Like the human adults' languages formed of phonemes arranged in a specific order, the newborns' CASs also follow a sequence of expirations, inspirations, and pauses. The infant's CAS respiratory pattern during crying is characterized as read in the following:

- **Expiration** is the acoustical part of the CAS (Robb & Goberman (1997)). The CAS is only produced during the expiratory phase (Chittora & Patil (2013)). (Grau, Robb & Cacace (1995)) reported the expiration before inspiration has the fundamental frequency between 320Hz and 740Hz. In essence, the CAS is the expiration.
- **Inspiration**: is any perceivable sound during inhalation (Grau *et al.* (1995)). Inspiration is breathier and contains a less vocalized sound than the expiration (Aucouturier, Nonaka, Katahira & Okanoya (2011)).Inspiratory CAS was barely seen in healthy infants (Wasz-Hockert (1968a)). The average of inspiration incidents in infants' CASs is in a range of two to 12, and the fundamental frequency is between 367Hz and 1040Hz (Grau *et al.* (1995)). The inspiratory phonation is assumed to be of value, containing information relating to pain and distress (Aucouturier *et al.* (2011)).
- **Silence**: are the soundless gaps between inspiration and expiration which lasts between 50ms to 100ms (Robb & Goberman (1997)).

The diagram in Figure 1.3 shows the distribution of CAS episodes in healthy infants' CASs (Robb & Goberman (1997)).

1.3 Observation of Normal vs Pathological CASs

Infants CAS has been of interest in the last century. Several comparative studies have been done on the CASs of healthy newborns and those with a particular pathology. The analysis of CASs of each healthy individual studied in (Michelsson, Eklund, Leppänen & Lyytinen (2002)) showed that they are characteristically similar both auditory and visually on their spectrogram. However, there are differences between CASs of healthy and sick infants (Mende, Wermke, Schindler, Wilzopolski & Hock (1990); Michelsson *et al.* (2002); Moller & Schonweiler (1999); Wasz-Höckert, Michelsson & Lind (1985)). Differences in the CAS characteristics of several pathologies in infants such as cri du chat syndrome, down's syndrome, hyperbilirubinemia,

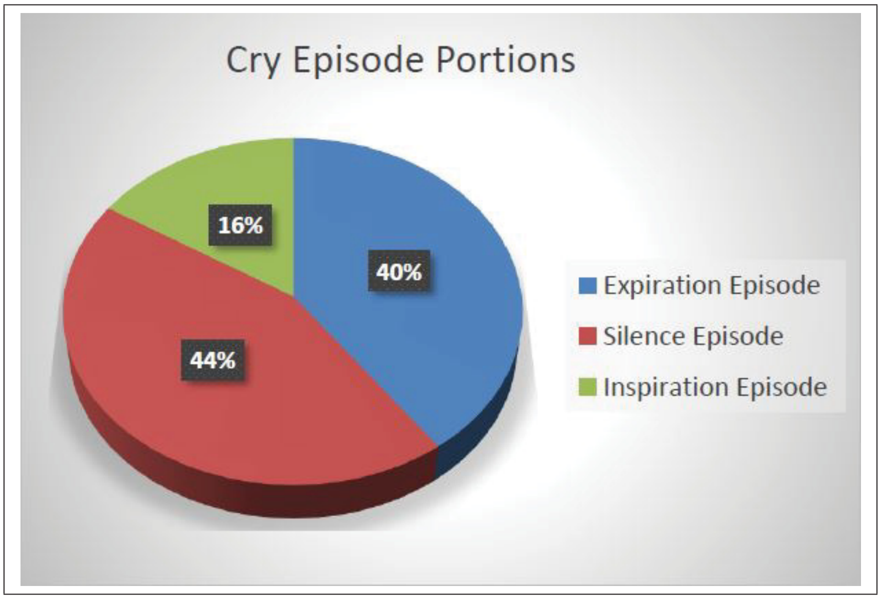


Figure 1.3 Portions of each CAS episode occurrence in healthy infants' CAS
Adapted from Robb & Goberman (1997)

encephalitis, meningitis, asphyxia, and some forms of brain damage with healthy infants were reported in (Michelsson & Wasz-Höckert (1980); Michelsson (1971); Wasz-Hockert (1968b)).

In the following, we look at some patterns identified simply in the CASs of healthy and unhealthy infants groups. The CASs of healthy infants has the fundamental frequency typically between 400 Hz to 600 Hz (Michelsson *et al.* (2002); Ruíz, Altamirano, Reyes & Herrera (2010)). The CASs with high fundamental frequency is referred to as unhealthy infants. In terms of the fundamental frequency contour of the infants' CASs called as melody, it has a rising-falling shape for normal infants, whereas, for sick ones, it is falling, falling-rising, flat (Chittora & Patil (2013); Orozco-García & Reyes-García (2003); Sirviö & Michelsson (1976)). The CASs of healthy infants contain more sounds (Sirviö & Michelsson (1976)) comparing to the sick ones. In the CASs of sick newborns, there are more shifts (Chittora & Patil (2013); Orozco-García & Reyes-García (2003)) and glides than healthy ones (Ruíz *et al.* (2010)). For latency, the mean for the CASs of healthy infants durationally is different from unhealthy ones'. For healthy newborns, it was reported 1.6 seconds; however, it is 2.6 seconds for infants with brain insults (Zeskind & Lester (1978)). The measure furcation does not occur in the CASs of healthy infants. The comparison made in (Kheddache & Tadj (2013b)) revealed that sick newborns' CASs, including full-term and preterm, contain higher hyper-phonic segments and irregularity of fundamental frequency rather than healthy ones. The intensity of the CASs of sick infants is lower than in healthy infants (Orozco-García & Reyes-García (2003)). The hyper-phonation feature is an identifier of the CASs of infants with epilepsy from healthy ones' (Chittora & Patil (2013)). The pathological CASs' spectrogram contains lower intensity than healthy ones' (Orozco-García & Reyes-García (2003)).

1.4 Various domains of investigating the Newborns' CAS

1.4.1 Subjective Auditory investigation

As mentioned previously, the study of the newborns' CAS initially started with the subjective auditory analysis. The cry-researchers showed that various reasons that initiate infants' crying, such as CASs initiated due to birth, pain, pleasure, and hunger, can be verified by the experienced auditory individuals (Wasz-Höckert, Partanen, Vuorenkoski, Michelsson & Valanne (1964)). Accordingly, the most experienced individuals with infants' CASs were scored best in recognizing the newborns' needs, such as mothers and hospital staff (Mukhopadhyay *et al.* (2013); Sagi (1981)).

1.4.2 Time-domain investigation

The impressive conclusions obtained by the research on the subjective auditory investigation led to the time domain investigation. The time organization of newborns' CAS episodes was well studied in this domain. In time-domain research, various duration of infants' CAS episodes was linked to external or internal stimuli (Zeskind, Parker-Price & Barr (1993)). While time-domain research was low-cost computationally and would allow observational investigation, it lacks the spectrum information of the newborns' CAS.

1.4.3 Frequency-domain investigation

The time-domain investigation solely unveils a tiny portion of massive information hidden in the newborns' CASs. Later the frequency domain of newborns' CASs was studied. It was shown that the frequency-domain experiment gives access to the coarse information of the frequency spectrum properties of the newborns' CASs (Lederman & Lederman (2002)). The frequency-domain diagram specifies how much of the signal is placed in a given frequency band in a frequency range; however, it lacks the time domain information.

1.4.4 Spectrographic Investigation

The spectrographic illustration of infants' CASs shows the patterns of CAS energy with time indexes. The horizontal axis shows time, the vertical axis shows frequency, and the grayscale within the diagram shows the intensity of the CAS.

With the use of spectrogram, several CAS characteristics such as the duration of the CAS, frequencies measures, and the melody contour (Chittora & Patil (2013); Sirviö & Michelsson (1976); Varallyay Jr (2006)), the range of fundamental frequency, the presence of harmonic doubling, bi-phonation, shift, and latency (Chittora & Patil (2013); Grau *et al.* (1995)) can be determined, and meanwhile most of the time, information of the signal is maintained (Moller & Schonweiler (1999)). Examples of the spectrogram of the healthy and sick newborns' CASs are illustrated in Figure 1.4, in which visual variations are noticeable.

There are visual cues observable in the spectrogram of infants' CASs that several cry-researchers have described to be characteristically distinct among groups of healthy and unhealthy ones (Boukydis & Lester (2012)). These measures contribute to the interpretation of the infants' CASs. Following some of these patterns are mentioned:

- **Fundamental frequency**: is the lowest measure of frequency in the CAS spectrogram.
- **Latency**: is the interval between once the infant has been stimulated for crying and when the crying starts. The latency time is dependent not only on the infant's disease but also on

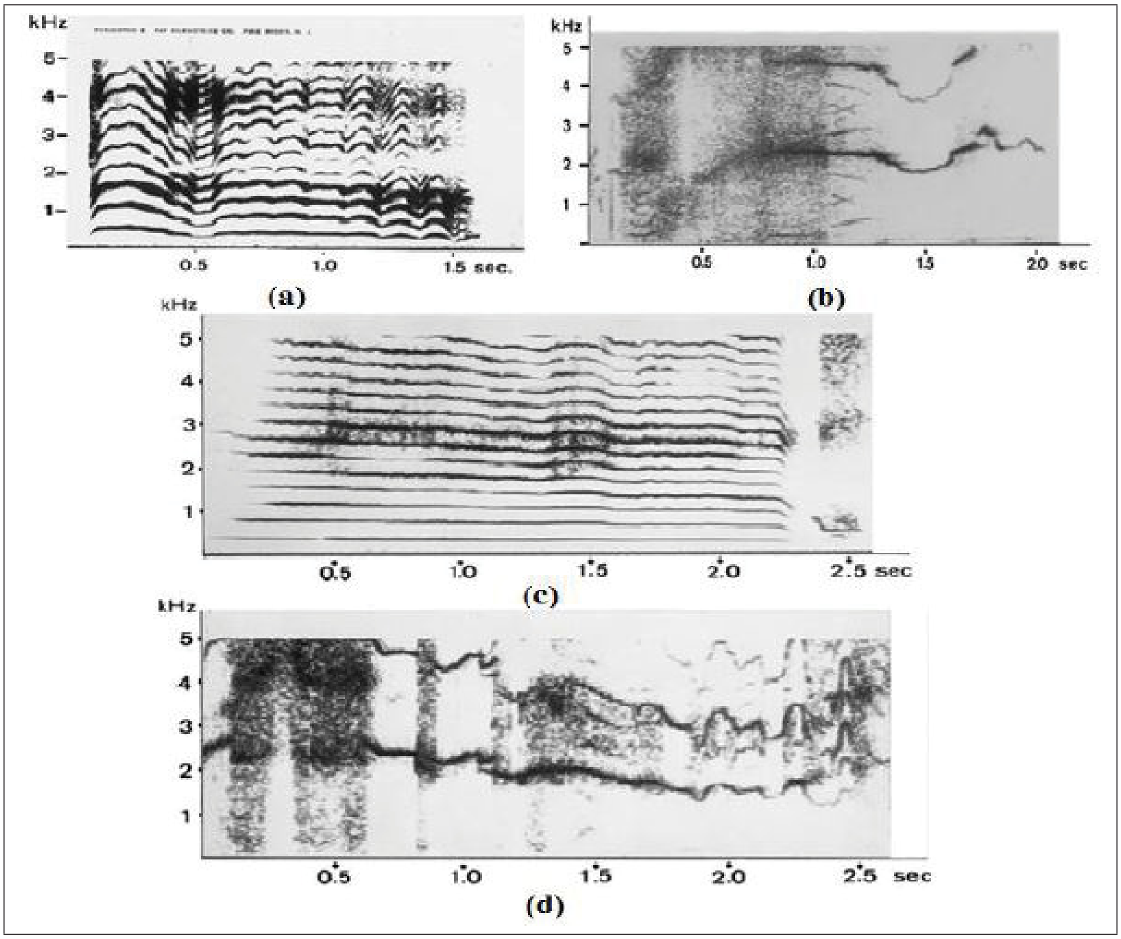


Figure 1.4 The spectrogram of the CAS of a healthy full-term infant, (b), (c) and (d) the spectrograms of the CASs of infants with brain disorder respectively hypothyroidism, asphyxia, and meningitis
Adapted from Michelsson & Michelsson (1999)

time since the last feeding as well as the wakefulness of the infant at the time of the CAS recording (Wasz-Höckert *et al.* (1985)).

- **Utterances**: is the number of vocal sound in the CAS (LaGasse *et al.* (2005)).
- **Shift**: is a sudden jump from one frequency to another (Michelsson *et al.* (2002)). Lack of stability in the neural control of the larynx results in changes in fundamental frequency (LaGasse *et al.* (2005)).

- **The melody type**: means the trend of the CAS melody contour in which six forms were defined for it, comprising of rising, falling, rising-falling, falling-rising, flat, and without melody shape (Ruíz *et al.* (2010)).

- **Duration of the cry**: is the time from an instance that the crying starts to its end, which is dependent on neural control of the respiratory system (LaGasse *et al.* (2005)).

- **Phonation**: The average rate of 25ms blocks having an F_0 in the 350–750 Hz range (Kheddache & Tadj (2013b)).

- **Bi-phonation**: is defined as the presence of two fundamental frequencies extending nonparallel across the spectrogram.

- **Double Harmonic break**: is defined as two nonparallel series of harmonics that have the same melody shape as the fundamental frequency (Sirviö & Michelsson (1976)).

- **Glide**: or "Glottal roll" (Chittora & Patil (2016)) is defined as very sudden change in the fundamental frequency, it is 600Hz or more within about 0.1 sec (Verduzco-Mendoza, Arch-Tirado, García, Ibarra & Bonilla (2009)).

- **Hyper-phonation**: is defined as noisy blocks of 25ms that have the fundamental frequency of higher than 1 kHz (Chittora & Patil (2013); Kheddache & Tadj (2013b)).

- **Dysphonation**: is characterized by the irregular or unregulated distribution of energy, and typically the energy in this region is very high. Heavy turbulence is created in this region (Chittora & Patil (2016)).

- **Furcation**: is an important feature that does not occur in healthy infants' CASs. It is the fundamental frequency branches to several other contours with different fundamental frequencies (Sirviö & Michelsson (1976)).

- **Subharmonic**: "*Subharmonics regime was defined primarily by the abrupt appearance in the narrowband spectrogram of intervening harmonics, doubling, tripling, or even higher integer multiples in relation to the surrounding set.*" (Buder, Chorna, Oller & Robinson (2008)).

1.4.5 Automated Computer-Based Analysis

The domains mentioned above unveil various aspects of valuable information hidden in newborns' CASs that describe groups' behaviours. However, a critical hinder for utilizing this information in these domains to identify the healthy and unhealthy group was the large volume of information. Manually handling the massive quantity of information within the dataset is difficult for humans; thus, an intelligent approach was required to analyze this vast information. Hence the computer-based automatic configuration for analysis of the newborns' CASs was developed (Golub & Corwin (1982)). The following section describes the method of developing an NCDS for analyzing the information in newborns' CASs for diagnosis purposes.

1.5 Newborn Cry Diagnostic System (NCDS)

An NCDS is an infant CAS-based diagnostic system that mainly contains three stages of CAS preparation, feature extraction and selection, and classification. Figure 1.5 is the perspective of an NCDS. The task of this system is to identify the CASs of infants with clinical state from healthy ones based on the patterns that it receives. After preparing the input infants' CASs with preprocessing techniques, the feature extraction block captures the distinctive features and eliminates unnecessary information. Moreover, as the quantity of the input CAS is quite massive for processing, it presents the essential distinguishing CAS features of each category in a manageable form (Orozco-García & Reyes-García (2003); Rao, Reddy & Maity (2015)). The feature extraction phase is for obtaining various distinguished patterns. The features may be in the time domain, frequency domain, and time-frequency domain. The features parametrization is explained in the following section.

In the last stage of the NCDS, learning algorithms such as SVM, nonlinear/linear regression, Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), etc., are employed for training and testing.

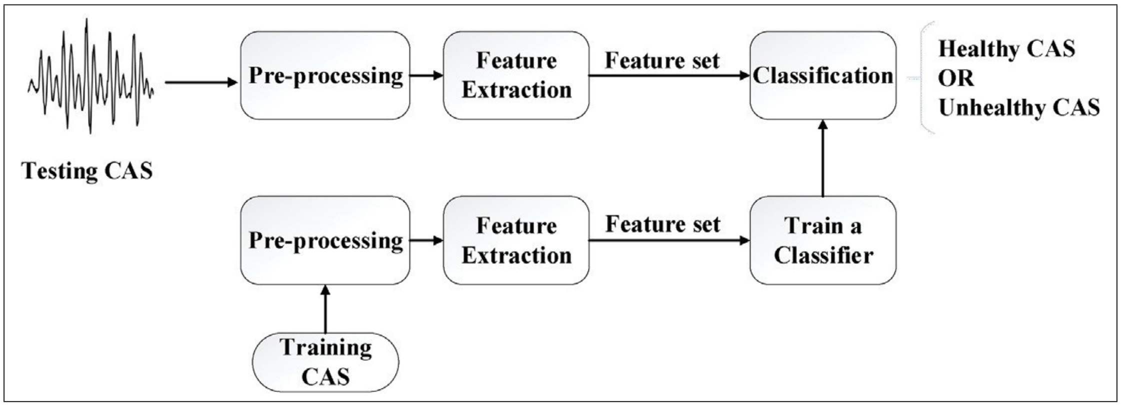


Figure 1.5 The design of the NCDS and its main blocks

1.6 Feature Parametrization

This thesis intends to enrich the feature extraction phase of the NCDS. Hence this section is devoted to the feature parameterization of newborns' CASs using the NCDS. As explained in the previous section, the feature extraction block catches discriminative information for each group and compresses the information into a manageable form for modelling. Hence, with the help of feature extraction techniques, each infant's CAS is reconstructed by a feature vector sequence. The feature extraction techniques use different analyzing domains for representing the audio file; thus, the features may represent the time domain, frequency domain, or the time-frequency domain.

Earlier in this chapter, the filter-source theory was explained. There are various techniques for designing the filter and source features. Besides that, another level of information can experiment with human vocalization based on prosodic features. Prosodic features are obtained by processing the audio file at a global level, unlike the local level mentioned above for filter-source features, which analyze the produced sound in quite a short interval. According to the literature, the prosodic features include the patterns of fundamental frequency variations (or Melody), durational features (or rhythm feature), intonation, stress, intensity, etc.

Accordingly, the study presented here categorizes the CAS features based on feature frame level in two categories: short-term and, long-term features also called prosodic features. The short-term features are extracted on short frames of CAS, while far more extended frame sizes are used for extracting long-term features. The following sections will define feature extraction methods in short and long intervals. First, the long-term features and their presence in infants' CASs are explained, and then, the common short-term features are introduced.

1.6.1 Long-term Features (Prosodic Features)

Prosodic features in the human language processing domain, also known as supra-segmental information (Adami *et al.* (2003); Pattnaik & Dash (2012); Vicsi & Szaszák (2010)) are defined as the long-term information of the voice signal. This level of features is related to the information of the pitch of signal (loudness) (Lieberman (1985)), the duration of utterance (Lieberman (1985); Pattnaik & Dash (2012)), the amplitude of the audio waveform (Lieberman (1985)), and the perceptible breaks that occur during speech. These features are namely intonation, melody, rhythm, and stress. The prosodic features are shown in written sentences by punctuation marks like periods, commas, question marks, etc. How essential these features are for understanding the content becomes apparent if one removes the punctuation in a sentence; so the pauses and the tone of the sentences become ambiguous, and it becomes difficult to tell if the sentence is a question or is conveying news, etc. (Lieberman (1985)).

In the study of the sound interpreting system such as ALI, and ASR, while the main focus is on standard short-term spectral information, several investigations have shown the advancement of systems using prosody features (Adami *et al.* (2003); Dahmani, Selouani, Chetouani & Doghmane (2008); Dahmani, Selouani, Doghmane, O'Shaughnessy & Chetouani (2014); Nisar, Shahzad, Khan & Tariq (2017); Rao *et al.* (2015); Selouani, Dahmani, Amami & Hamam (2012); Shriberg & Stolcke (2004); Vicsi & Szaszák (2010)). In the infants' CAS analyzing domain, (Manfredi, Pieraccini, Viellevoye, Torres-Garcia & Reyes-Garcia (2017); Mende *et al.* (1990); Moller & Schonweiler (1999); Wermke, Birr, Voelter, Shehata-Dieler, Jurkutat, Wermke & Stellzig-Eisenhauer (2011)) also showed successful results for detecting the CASs of sick infants using prosody features in comparative studies. In the following, the definition of the prosodic features is explained.

1.6.1.1 Melody

Melody is the study of the pitch of the audio signal. Melody in the speech processing field is known to convey the emotional state of the speaker (Mampe, Friederici, Christophe & Wermke (2009)). It is the variation of the fundamental frequency of the signal against time (Mende *et al.* (1990); Ruíz *et al.* (2010); Varallyay & Benyó (2007); Wermke, Mende, Manfredi & Bruscaglioni (2002)). The melody shape of the infants' CASs is described in the literature to be continuously connected (Mampe *et al.* (2009)) except when there are sudden jumps from one frequency to another (Moller & Schonweiler (1999)).

It is suggested that the melody form of the CAS correlates with the physical and physiological conditions of the infants (Varallyay & Benyó (2007)). The preliminary study found that the melody of newborns' CASs initiated due to birth, hunger, pain, and pleasure are different. Figure 1.6 shows the melody contour of four reasons of newborns' CASs.

Reason of Crying	**Melody Form(s)**	
Birth Cry	—	╲
Hunger Cry	⌒	
Pain Cry	╲	
Pleasure Cry	⌒	—

Figure 1.6 The most frequent melody form(s) for common infants' CASs and their percentages
Adapted from Wasz-Hockert (1968a)

In literature, cry-researchers proved melody as a distinct pattern between healthy and diseased infants suffering from hearing impairment, brain disorder resulting from severe oxygen deficiency after birth, meningitis, hydrocephalus and central respiratory distress (Boukydis & Lester (2012); Michelsson & Michelsson (1999); Moller & Schonweiler (1999)), and preterm and at-term infants (Manfredi *et al.* (2017)).

Newborns are under the influence of the sound of their mother's speaking, and this early experience impacts their postnatal auditory preferences (DeCasper & Spence (1986)), which later influences their CAS production in terms of melody (Mampe *et al.* (2009); Manfredi *et al.* (2017)). Therefore, this mentioned studies result has to be taken into account in the creation of NCDS. We call this criterion "unbiased by region or language" and observed it in our proposed NCDS. We explain this criterion again in the conclusion section. In this work, we followed the method described in (Mary (2012)) in the ASR domain to use tilt parameterization to describe the melody features of newborns' CASs. The tilt feature set is explained in detail in chapters two, three and four.

1.6.1.2 Rhythm

Rhythm is the repetition of a pattern periodically. Rhythm exists in various sorts in human organs and behavioral systems (Wolff (1967)). The first emergence of rhythmical structure in humans is in neonates' CAS (Wolff (1967)). In the speech domain, processing rhythm is defined as the result of specific phonological phenomena in a given language (Ramus, Nespor & Mehler (1999)). Likewise, in newborns' CASs, particular organized patterns in terms of temporal morphology were documented (Lester, Boukydis, Garcia-Coll, Hole & Peucker (1992); Michelsson, Christensson, Rothgänger & Winberg (1996); Wermke & Mende (2009); Wolff (1967); Zeskind *et al.* (1993)).

Figure 1.7 shows the rhythmicity of hunger and pain CASs of newborns. Rhythm in infants' CASs can be defined as the repetition of a certain element with time order. The rhythmical characteristic of infants' CASs is an order of an expiratory, a pause, an inspiratory within a specific time organization (Wolff (1967)). It is suggested that the duration of CAS segments (expiration, pause and inspiration) is under the impact of external or internal stimuli (Zeskind *et al.* (1993)). Several researchers such as (Michelsson *et al.* (1996); Wolff (1969); Zeskind *et al.* (1993)) reported different temporal sequences of newborns' CASs, such as ones irritated from pain, hunger, and frustration.

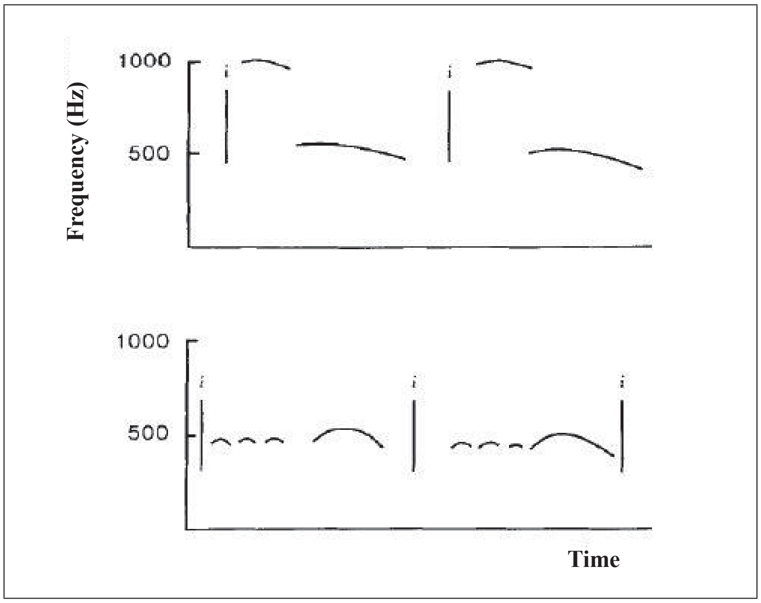


Figure 1.7 The diagram of frequency versus time of pain CAS (top), hunger CAS (bottom) in which the periodicity of specific patterns are evident Adapted from Michelsson *et al.* (1996)

As mentioned earlier the studies showed that the rhythmitial properties of hunger, pain and some other types of CASs are different (Chang & Li (2016); Michelsson *et al.* (1996); Rodriguez & Caluya (2017); Vempada, Kumar & Rao (2012)). Some of CASs initiated by hunger and pain reasons are shown in Figure 1.7; thus, an NCDS should be unaffected by the reason for crying and discards such patterns that are unrelated to infants' clinical state. In our study, we call this criterion "robust to the reason for crying "and discuss it in the conclusion section.

1.6.2 Short-term Features

Short-term features relate to short interval analysis of the audio signal, typically on the order of tens of milliseconds. These features capture information on the voice parameters of the speaker. Some features in this set include Linear Predictive Coefficients (LPC), Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Prediction (PLP), Relative spectra filtering of log domain coefficients (Shrawankar & Thakare (2013)), bark frequency cepstrum coefficient, Linear Frequency Cepstrum Coefficients (LFCC), MFCC and AAM Feature sets. This work focused on the standard method of the MFCC feature set and the AAM feature set for short-term analysis of infants' CASs.

As it is explained in the following section (Review of studies), the MFCC feature set defeated other short-term techniques. Furthermore, we used the AAM feature set, which has never been experimented with in the NCDS. The AAM feature set has been successfully tested in other acoustic recognition systems, such as nonverbal human-produced audio events (Bouserhal *et al.* (2018)), speaker verification (Kinnunen *et al.* (2008)), and specifically was shown to outperform the widely used feature set of MFCC (Sarria-Paja & Falk (2017)). The MFCC and AAM feature sets are explained in detail in chapters two, three and four

1.7 Review of studies

In general, the works made in newborns' CAS analysis include various tasks such as automatic detection of newborns' CASs, among other non-CAS sounds in the environment (Kim, Kim, Hong & Kim (2013)), automatic identification of segments in newborns' CASs including expiration, inspiration and pause (Abou-Abbas *et al.* (2016); Aucouturier *et al.* (2011)), identification of non-pathological reason of crying such as the CASs initiated by hunger, pain, birth etc. (Abdulaziz & Ahmad (2010); Saha, Purkait, Mukherjee, Majumdar, Majumdar & Singh (2013); Wahid, Saad & Hariharan (2016)), and the identification of CASs of sick newborns form healthy ones (Alaie *et al.* (2016); Kheddache & Tadj (2013a); Lahmiri, Tadj, Gargour & Bekiros (2021); Orozco-García & Reyes-García (2003)).

Due to the domain of this research, the following content is focusing on the diagnostic computer-based model for identifying the CASs of sick newborns from healthy ones. As explained earlier, researchers reported that the infants' CASs characteristics could be adopted as symptoms to identify infants' unhealthy state. The following section is grouped by the disease that cry-researchers have worked on.

1.7.1 Hearing Impairment vs. Healthy

Several studies reported using popular feature extraction methods of MFCC, LPC, and the prosodic features to identify the CASs of infants with hearing disorders from healthy ones. In these experiments, it was shown that the different number of coefficients impacts classification results (García & García (2003)), and it is dependent on the dataset that has to be learned through several examinations.

The comparison between MFCC and LPCC feature extraction techniques in the pipeline with Feed-Forward Neural Network (FFNN) model showed better system accuracy using MFCC. The accuracy for LPCC and MFCC were 94.3% and 96.80% respectively (García & García (2003)). However, in another more novel study (Wahid *et al.* (2016)), the appliance of the LPC technique always reported better with neural network comparing with MFCC. Moreover, the use of feature selection increased the system performance and saved the computation time. The higher identification accuracy was reported for the setups of LPCC plus to delta LPCC and delta delta LPCC with Multilayer Perceptron (MLP) and Radial Basis Function Network (RBFN) models with accuracy rates of 99.21% and 99.42%, respectively (Wahid *et al.* (2016)).

With regards to the use of prosodic features, the duration of the CAS, melody contour complexity, as well as energy feature and frequency pattern, were determinative for diagnosing the CASs of newborns with the hearing problem (Moller & Schonweiler (1999)). The length of infants' CASs and the fundamental frequency variation (melody complexity) in infants' CASs are higher in hearing impairment than normal hearing ones. On the other hand, the normal infants' CASs have higher energy in the 2k to 4k Hz band and the 4k to 8k Hz band. In spectral view, basic and fasted frequencies' rhythmicity was found higher in normal infants' CASs. In addition, the second formant is higher and has higher variability in the CASs of normal infants than unhealthy ones.

Using MFCC features, the accuracy was in the range of 62% to 67%, whereas the addition of prosodic features increased the rate up to 75%(Moller & Schonweiler (1999)).

1.7.2 Asphyxia vs Healthy

The disruption of oxygen and carbon dioxide gas exchange in the fetal or infancy period is referred to as asphyxia. This results in neuronal symptoms in the infant can cause mental retardation-cerebral palsy in the future. Among newborns' clinical states, the effect of asphyxia pathology on newborns' CASs has been widely studied.

Initially, the study on the CASs of infants of the healthy group and infants suffering from asphyxia indicated different significant characteristics. These differences were observed for both preterm and full-term infants of groups of healthy and asphyxia (Michelsson (1971)). In full-term infants, the energy of inspiratory segments in infants with respiratory disease, including asphyxia, was reported to be higher than healthy group (Alaie *et al.* (2016)).

The identification of CASs of infants with asphyxia from normal infants and deaf ones was successfully performed by feature extraction technique of MFCC and FFNN (Reyes-Galaviz, Verduzco, Arch-Tirado & Reyes-García (2005)). The classification accuracy rate was up to 97.39%. Later, the configuration of weighted LPCC with PNN increased the recognition accuracy to 99%(Hariharan, Chee & Yaacob (2012a)).

In a more general identification configuration, the identification of asphyxiating CASs from other disease family types of heart problems, neurological disorders, blood abnormalities, and others was designed in (Alaie *et al.* (2016)). The MFCC features techniques, extracted exclusively from two informative episodes of expiration and inspiration was tried with several classification methods, such as MLP, PNN, and SVM, in which GMM showed high ability. It has a maximum classification rate of 74%.

Table 1.1 shows the proposed schemes and the accuracy results for identifying the condition of asphyxia using newborns' CASs. The non-linear SVM with RBF kernel always resulted better than linear kernel with features of MFCC (Sahak *et al.* (2010a); Sahak, Mansor, Lee, Yassin & Zabidi (2010b)). Among the proposed method the configuration of PNN model with energy and entropy of wavelet packet resulted best.

Table 1.1 Classification schemes proposed for identifying asphyxiate CASs from normal ones. * shows the maximum accuracy obtained between the examined methods by Wahid et al.

Reference	Feature extraction techniques	Feature selection technique	classification method	Maximum Accuracy
(Sahak et al., 2010b)	MFCC	Orthogonal Least Square (OLS)	SVM (RBF kernel)	93.16%
(Zabidi, Mansor, Khuan, Yassin, & Sahak, 2010)	MFCC	F-Ratio	MLP	93.38%
(Sahak et al., 2010a)	MFCC	PCA	SVM (RBF kernel)	95.86%
(Zabidi, Mansor, Khuan, Yassin, & Sahak, 2011)	MFCC	BPSO	MLP	96.30%
(Hariharan, Yaacob, & Awang, 2011)	Energy and Entropy of wavelet packet transform	—	PNN	99.49%
(Wahid et al.)	MFCC	OneR	MLP	99.29% *
	MFCC + Δ MFCC + ΔΔ MFCC	ReliefF		
	LPCC	FCBF	RBFN	
	LPCC + Δ LPCC + ΔΔ LPCC	CNS		
		CFS		

Concerning prosodic features, in 30 percent of cases, the melody feature of asphyxiated infants' CASs was observed rising and falling-rising (Wasz-Höckert *et al.* (1985)). Furthermore, in a study adding the prosodic feature to MFCC was shown to decrease the model error rate by more than 3% (Ji, Xiao, Basodi & Pan (2019)).

1.7.3 Cleft Lip and Cleft Palate vs Healthy

Cleft lip and cleft palate are connate malformations of lip and mouth, respectively, one of the most prevalent defects in newborns. The identification of CASs of the cleft palate from normal ones was investigated in (Lederman, Zmora, Hauschildt, Stellzig-Eisenhauer & Wermke (2008)). MFCC features and LPCC features were extracted, and HMM was trained on these feature sets. The MFCC features was found markedly defeating LPCC features.

On the prosodic level, (Michelsson, Sirviö, Koivisto, Sovijärvi & Wasz-Höckert (1975)) examined the difference of melody form in the CASs of healthy infants and infants with cleft palate. No difference was found. In another study presented by (Wermke *et al.* (2011)) reviewed the group differences by use of melodic and rhythmic features. The CASs of healthy infants were observed to have far higher melody complexity than newborns' CASs with cleft palates. In addition, the comparison showed that the CASs of newborns with clefts (both types) contain more segmented multiple-arc melodies than healthy ones.

1.7.4 Other Studied Features for Particular Diseases

Commonly the MFCC features are used as the baseline (Ji, Mudiyanselage, Gao & Pan (2021)). The MFCC features were successfully used to identify the CASs of infants affiliated with hypothyroidism pipelined with MLP (Zabidi, Khuan, Mansor, Yassin & Sahak (2010b)). (Santiago-Sánchez, Reyes-García & Gómez-Gil (2009)) experimented identifying the CASs of normal newborns from those with asphyxia or hyperbilirubinemia. They studied the three features of MFCC, LPC, intensity, and Cochleograms and their combination in configuration with Type 2 fuzzy pattern matching. Among these features, the combination of LPC and Cochleograms defeated others.

Accordingly, the MFCC features have been used repeatedly in NCDS and were found to be a dominant technique for characterizing pathological CASs versus healthy ones. The main advantage of MFCC is its resistance to noise and spectrum estimation errors under different conditions. The disadvantage of MFCC is that it requires far more execution time and mathematical resources than LPCC and mentioned prosodic features in the literature reviews.

The foremost critic in literature for designing the NCDS is that the cry-researchers often fail to observe the independence of folds using cross-validation, while, in recent machine learning applications, this precaution is taken into account that the distribution of the individuals between the folds is managed to allot all samples of a particular individual to one fold. In other words, in developing an NCDS, there has to be no CAS of the same infants in more than one fold. This criterion increases the system's credibility as it is independent of individual characteristics. In this case, the NCDS learns the pathologically discriminant patterns in the newborns' CASs, and its predicted label for a test CAS is more reliable. This criterion is crucial since for instance the MFCC features are distinctive characteristics for recognizing speakers; this poses a problem concerning the distribution of the records of the same infant between the learning part and test parts.

CHAPTER 2

ON THE USE OF LONG-TERM FEATURES IN A NEWBORN CRY DIAGNOSTIC SYSTEM

Fatemeh Salehianmatikolaie[1], Chakib Tadj[1]

[1] Department of Electrical Engineering, École de technologie supérieure,
1100 Notre-Dame West, Montreal, Quebec, Canada H3C1K3

Paper published in the journal of Biomedical Signal Processing and Control, 2020.

Résumé

Cette étude propose d’utiliser une nouvelle combinaison de caractéristiques à court et à long terme provenant de différentes échelles de temps pour développer un système de diagnostic automatique des pleurs des nouveau-nés afin de différencier les signaux audio des pleurs (Cry Audio Signals, CASs) des nourrissons en bonne santé de ceux atteints du syndrome de détresse respiratoire (Respiratory Distress Syndrome, RDS). Les coefficients cepstraux de fréquence mélodique (Mel-frequency cepstral coefficients, MFCC) ont été utilisés comme caractéristiques à court terme, tandis que les caractéristiques mélodiques et rythmiques obtenues sur des échelles de temps plus longues ont été utilisées comme caractéristiques à long terme. Nous avons émis l’hypothèse que les différences entre ces groupes pouvaient se produire sur plusieurs échelles de temps. Enfin, un modèle de machine à vecteur de support a été utilisé pour générer la classification finale. Entre autres résultats, les meilleurs ont été obtenus en combinant les trois ensembles de caractéristiques (les MFCC et les caractéristiques de rythme et de mélodie) dans l’épisode d’expiration ; la combinaison des MFCC et des caractéristiques d’inclinaison a amélioré les performances du classificateur dans l’épisode d’inspiration. En termes de mesure du F-score, dans l’expérience d’inspiration, les caractéristiques d’inclinaison seules étaient les caractéristiques de classification les plus fortes pour différencier les nourrissons atteints de RDS des nourrissons sains. Les résultats indiquent que la combinaison des caractéristiques à court et à long terme fournit une meilleure méthode de classification pour différencier les CAS des nourrissons en bonne santé de ceux atteints de RDS. En outre, les résultats ont confirmé l'importance des caractéristiques à long terme dans les épisodes d'expiration et d'inspiration en tant que marqueurs diagnostiques entre les groupes de nourrissons en bonne santé et les nourrissons atteints du RDS.

Mots-clés: Caractéristiques à long terme; Mélodie; Rythme; Caractéristiques à court terme; Coefficient cepstral de fréquence mélodique; Machine vectorielle de soutien; Cri du nouveau-né; Cri d'expiration et d'inspiration.

2.1 Abstract

This study proposes using a novel combination of short-term and long-term features from different timescales to develop an automatic newborn cry diagnostic system to differentiate the cry audio signals (CASs) of healthy infants from those with respiratory distress syndrome (RDS). Mel-frequency cepstral coefficients (MFCCs) were used as the short-term features, while the melody and rhythm features obtained from longer timescales were used as the long-term features. We hypothesized that the differences between these groups may occur on several timescales. Finally, a support vector machine model was used to generate the final classification. Among other findings, the best results were obtained from the combination of all three feature sets (the MFCCs and the rhythm and melody features) in the expiration episode; the combination of MFCCs and tilt features improved the classifier performance in the inspiration episode. In terms of the F-score measure, in the inspiration experiment, the tilt features alone were the strongest classification features for differentiating infants with RDS from healthy infants. The results indicate that the combination of short-term and long-term features provides a better classification method for differentiating the CASs of healthy infants versus RDS infants. Moreover, the results confirmed the importance of long-term features in the expiration and inspiration episodes as diagnostic markers between groups of healthy infants and RDS infants.

Keywords: : Long-term Features; Melody; Rhythm; Short-term Features; Mel-frequency Cepstral Coefficient; Support Vector Machine; Newborn Infant Cry; Expiration and Inspiration Cry.

2.2 Introduction

2.2.1 Potential of Analyzing the Infant Cry Audio Signal (CAS)

For newborns, crying is the most effective method of communicating with others because infants lack the linguistic ability of adults. In newborns, crying can be initiated for a variety of reasons, such as pain, hunger, anger, and frustration. These types of crying can often be differentiated by individuals who are familiar with an infant's cry audio signals (CASs) like mothers (Mukhopadhyay *et al.* (2013); Sagi (1981)). Some diseases also affect the acoustic features of crying. These differences can be observed in the CAS spectrogram of healthy infants in comparison to unhealthy infants for specific features, such as the duration of CAS, the frequency measures, and the melody contour (Michelsson & Michelsson (1999)). Investigating the characteristics of infant CAS yet has been addressed by two methods of spectral analysis by viewing the CAS spectrograms, and developing automatic system for CAS classification (Kheddache & Tadj (2019)). The present study adopted the latter method for further analyze of infant CASs.

2.2.2 Newborn Cry Diagnostic System (NCDS)

The NCDS is a CAS-based diagnostic system that was developed by several researchers (Alaie *et al.* (2016); Kheddache & Tadj (2012,1); Rosales-Pérez, Reyes-García, Gonzalez, Reyes-Galaviz, Escalante & Orlandi (2015)), and it functions in a similar way to an automatic speech recognition system. The system aims to elicit useful information from the CAS of infants to determine the relevant specific features in order to diagnose their diseases. Figure 2.1 shows a block diagram of the NCDS configuration. The NCDS consists of three stages: signal preprocessing, feature extraction, and classification. The details of these stages are discussed later in Section 2.3.2. This system is used to classify infant CASs by type, namely healthy or unhealthy, based on the pattern of the CAS that it receives. The NCDS installation cost is relatively low in comparison to other systems, and it could help address the lack of specialists in developing countries (Alaie *et al.* (2016)).

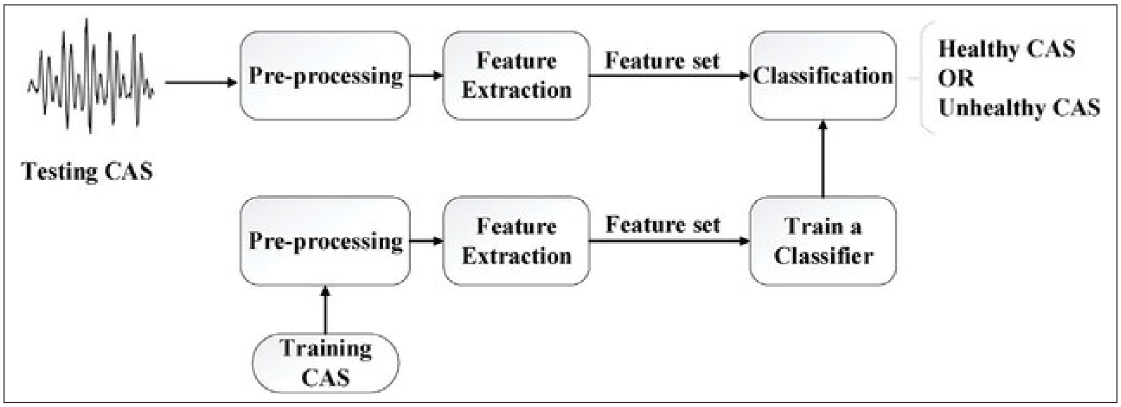


Figure 2.1 Block diagram of the NCDS

2.2.3 Respiratory distress syndrome (RDS) prevalence

RDS is a life-threatening pulmonary disorder caused by a deficiency of surfactants in the lungs of some newborns. The prevalence of this syndrome is 7% among newborns (Kumar & Bhat (1996)). The disorder is primarily related to premature birth; a premature infant's lungs are deficient in the amount and composition of surfactants. RDS is at the root of the mortality rate of early infants. Table 2.1 shows the rank of RDS among the causes of infant fatalities and the number of RDS-related deaths in Canada. No specific test exists for diagnosing RDS. Typically, the diagnosis is based on several physical examinations, such as a radiography of the chest, the measurement of oxygen levels via a blood test, and using echocardiography to eliminate other potential causes. However, it is important to note that other conditions may present symptoms that are similar to those of RDS.

Table 2.1 Rank and number of RDS among the fatal pathologies in Canada
Adapted from Canada (2021)

	2012	2013	2014	2015	2016
Rank of RDS among leading causes of infant death	12	12	11	10	11
Number of infant deaths due to RDS	26	25	29	29	21

2.2.4 Literature review of studies on infant CAS

Researchers suggest that the audio characteristics of infants CASs can be used to understand the infant's needs, or to diagnosis of a particular disease (Rosales-Pérez *et al.* (2015)). In general, the efforts made in the field of infant CAS processing can be divided into two groups. In the first group, the main purpose is to identify the condition of the baby, including pain, hunger (Rosales-Pérez *et al.* (2015)), discomfort (Rodriguez & Caluya (2017)), fear, sleepiness etc. In the second group, the purpose is to diagnose a particular disease. Regarding the purpose of this work, the pertinent researches to diagnosing pathology is focused in this literature review. Initial studies of the characterization of the CASs of unhealthy infants, as well as studies differentiating unhealthy infants from healthy ones, date back to the 1960s (Lederman, Cohen, Zmora, Wermke, Hauschildt & Stellzig-Eisenhauer (2002)). These studies include the observations of melody forms and the temporal features of diseases, such as Cri Du Chat Syndrome, Down syndrome, and brain damage (Wasz-Hockert (1968b)). Additionally, the pattern of occurrence of some characteristics, such as slide, glide, bi-phonation, and minimum and maximum pitches, have been anecdotally reported for diseases of the central nervous system, asphyxia, hydrocephalus, and malformation syndromes (Boukydis & Lester (2012)), for which accurate detection of fundamental frequency contour is a perquisite (Fort & Manfredi (1998)). These findings have resulted in research on the development of an automatic signal-based system for diagnosing hearing impairment (Hariharan *et al.* (2012a); Jam & Sadjedi (2009); Moller & Schonweiler (1999); Orozco-García & Reyes-García (2003); Rosales-Pérez *et al.* (2015); Wahid *et al.* (2016)), asphyxia (Hariharan *et al.* (2012a); Reyes-Galaviz *et al.* (2005);

Rosales-Pérez *et al.* (2015); Sahak *et al.* (2010b); Sahak, Mansor, Khuan, Zabidi & Yassin (2012); Santiago-Sánchez *et al.* (2009); Zabidi, Mansor & Lee (2017)), hyperbilirubinemia (Santiago-Sánchez *et al.* (2009)), hypothyroidism (Zabidi, Mansor, Khuan, Yassin & Sahak (2009b)), RDS (Lederman *et al.* (2002)), cleft palate (Lederman *et al.* (2002,0)), using the CASs of infants. More recently, the general case of differentiating healthy versus unhealthy infants for cardiac, neurological, respiratory, and blood diseases (Alaie *et al.* (2016)) and, preterm and full-term infants (Kheddache & Tadj (2019) has been successfully performed in our laboratory. The introduced methods for feature extraction for pathological purposes include Mel-cepstral coefficients (MFCCs) (Alaie *et al.* (2016); Reyes-Galaviz, Cano-Ortiz & Reyes-García (2008); Sahak *et al.* (2010b)), linear prediction cepstral coefficients (Orozco & García (2003)), short Fourier time transform (Hariharan, Sindhu & Yaacob (2012b)), and wavelet packet transform (Hariharan, Yaacob & Awang (2011)), in which MFCCs are the most common and efficient features, among others (Reyes-Galaviz & Reyes-Garcia (2004)).

2.2.5 Criteria for NCDS

The present study aimed to implement an NCDS using the short- term features of MFCCs and the long-term features of melody and rhythm based on the following criteria:

2.2.5.1 Generalizable

The system should be independent of individual characteristics and it should identify general discriminant patterns, thus requiring a reliable database containing the CAS of a sufficient number of infants. For this, the dataset is partitioned into ten-fold segments, in which the CAS samples of each of the infants are only used in one fold; hence, the samples in each fold are independent of the other folds.

2.2.5.2 Unbiased by region or language

Studies have shown that certain CAS characteristics for infants change, depending on the geographical region or the linguistic group of the parents (Mampe *et al.* (2009); Manfredi, Viellevoye, Orlandi, Torres-García, Pieraccini & Reyes-García (2019); Wermke, Ruan, Feng, Dobnig, Stephan, Wermke, Ma, Chang, Liu & Hesse (2017)). Hence, the NCDS should be free of regional or linguistic biases.

2.2.5.3 Robust to reason for crying

The NCDS should be able to determine the category of CAS without a priori knowledge of the reason for crying (hunger, pain, birth, etc.). The literature indicates that the prosodic and spectral characteristics of hunger, pain, and some other types of CASs are different (Chang & Li (2016); Michelsson *et al.* (1996); Rodriguez & Caluya (2017); Vempada *et al.* (2012)).

Melody refers to pitch variation as a function of time. Rhythm is defined by the presence of a pattern that repeats, periodically. Rhythm indicates durational characteristics. This research study aims to assess the classification performance of long-term features in the NCDS, and those features in combination with short-term features.

With the design aims mentioned above, we primarily explored the short-term and long-term features, and integrated these features at various levels of the NCDS (features that, in this case, have not been studied extensively in infant pathology), provided these configurations exhibit enhanced performance in speech recognition systems (Adami *et al.* (2003); Vicsi & Szaszák (2010)). For a quantitative representation, we used the MFCC features for the short-term scale, and melody (specifically tilt) and rhythm features for the long-term scale. The features of the MFCC, melody, and rhythm will be discussed in detail in Section 2.3.2.2.

2.3 Methodology

2.3.1 Dataset

The CASs from 117 at-term healthy and unhealthy infants were used in this experiment (78 healthy infants and 34 infants with RDS). Details about the number of datasets used in the experiment will be explained in Section 2.3.2.3, which addresses classification. In this study, a mixed population of both healthy infants and infants with RDS were used in expiration and inspiration episodes experiments, compromising 191 CASs (number of separated expiration episodes) and 185 CASs (number of separated inspiration episodes), respectively. The underlying reason for each cry (such as hunger, pain, or frustration) was masked from the system. In this study, the age of the infants was restricted to between 1 day and 53 days old, as infants are generally unable to voluntarily control their CASs during this period (Boukydis & Lester (2012)). The gestational age of all the subjects was at least 37.2 weeks.

The CASs of the newborns were recorded at two different hospitals, one in Lebanon (Al-Sahel and Al-Raee) and another in Montreal, Canada (Hôpital Sainte-Justine). The cause of crying, including hunger, pain, birth, wet diaper, etc., was annotated by medical experts at both hospitals. In the recording procedure, a two-channel sound recorder with a sampling frequency of 44.1 kHz and a resolution of 16 bits was placed at a distance of 10 cm-30 cm from the infant. The length of each record is in within 2-3 min.

2.3.2 NCDS design

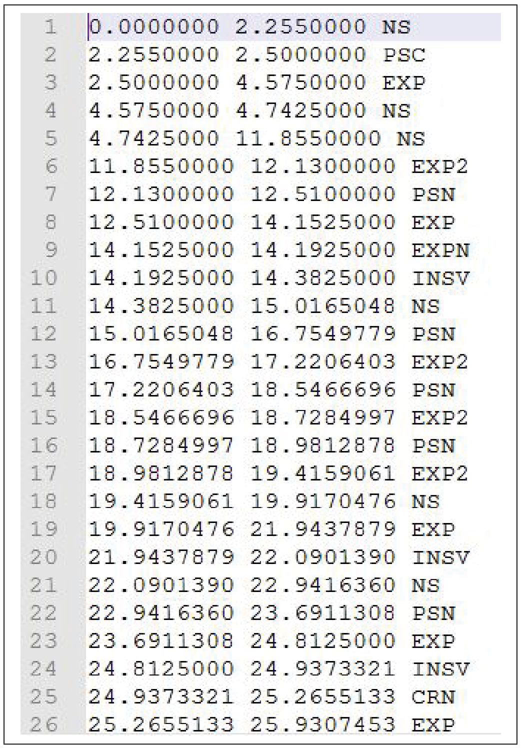
2.3.2.1 Preprocessing

Just as human language is composed of phonemes (sound units) placed in a specific order, a CAS follows a sequence of expirations and inspirations; behaviors, such as grunting and fussing, may coincide with these episodes. In our experiment, the infants' CASs were recorded in a hospital environment; therefore, the recordings include background noises (such as machine sounds or nurses' voices). Thus, the corpus of the recordings includes expirations and inspirations, in addition to some parts that are not useful for processing. During preprocessing, unwanted episodes were removed because only expiration and inspiration episodes are useful. This requires a segmented and labeled CAS corpus. The assigned labels and their descriptions are listed in Table 2.2.

Table 2.2 CAS labels in the database and their descriptions

Labels	**Description**
EXP	Voiced expiration segment during a period of cry
EXPN	Unvoiced expiration segment during a period of cry
INS	Unvoiced inspiration segment during a period of cry
INSV	Voiced inspiration segment during a period of cry
EXP2	Voiced expiration segment during a period of pseudo-cry
INS2	Voiced inspiration segment during a period of pseudo-cry
PSEUDOCRY	Any sound generated by the baby and it is not a cry
Speech	Sound of the nurse or parents talking around
Background	The kind of noise so low, it is characterized by a very low power-silence affected with little noise
BIP	The sound of the medical instruments next the baby
Noisy cry	Any sound heard with the cry: machine's beep sound, water, diaper, etc.
Noisy pseudo-cry	Any sound heard with the pseudo-cry
Noise	Like the sound caused by the mic moved by someone, the diaper, a door sound, speech + background, speech +beep.

Labels were assigned using WaveSurfer software. This program contains options for visualizing the waveform and spectrogram, as well as manual labeling. The associated manual annotation text file is also available for each recording. An example of this file for a portion of a CAS is shown in Figure 2.2. An example of the labels in a portion of a CAS, as well as the recorded CAS and the corresponding labels, are shown in Figure 2.3. We used these data, which were developed and segmented in our laboratory, for the preprocessing step found in (Abou-Abbas *et al.* (2016)) for automatic CAS episode detection. Previous studies investigating infant CAS have primarily focused on the expiration episode; however, it has also been proven that the inspiration episode is useful (Abou-Abbas *et al.* (2016)). The CAS segments are the input data for the proposed NCDS. In the labels for the CAS, two types of episodes exist, each for expirations and inspirations, namely, with vocalization (regular) and without vocalization (mute). Because this study focused on experimenting with prosodic features, only the vocalization (regular) segments were used, i.e., “EXP” and “INSV” from Table 2.2. This labeled corpus still requires further preprocessing in accordance with the type of feature extractor phase. These procedures, along with the feature extraction applications, will be discussed in later sections in this paper.



```
0.0000000 2.2550000 NS
2.2550000 2.5000000 PSC
2.5000000 4.5750000 EXP
4.5750000 4.7425000 NS
4.7425000 11.8550000 NS
11.8550000 12.1300000 EXP2
12.1300000 12.5100000 PSN
12.5100000 14.1525000 EXP
14.1525000 14.1925000 EXPN
14.1925000 14.3825000 INSV
14.3825000 15.0165048 NS
15.0165048 16.7549779 PSN
16.7549779 17.2206403 EXP2
17.2206403 18.5466696 PSN
18.5466696 18.7284997 EXP2
18.7284997 18.9812878 PSN
18.9812878 19.4159061 EXP2
19.4159061 19.9170476 NS
19.9170476 21.9437879 EXP
21.9437879 22.0901390 INSV
22.0901390 22.9416360 NS
22.9416360 23.6911308 PSN
23.6911308 24.8125000 EXP
24.8125000 24.9373321 INSV
24.9373321 25.2655133 CRN
25.2655133 25.9307453 EXP
```

Figure 2.2 Manual transcription of a CAS example in the database

2.3.2.2 Feature extraction

1. **Mel-frequency cepstral coefficients (MFCCs).** MFCCs are the most practical features for obtaining and separating patterns in CAS episodes (inspiration, expiration, etc.) (Abou-Abbas, Montazeri, Gargour & Tadj (2015b); Abou-Abbas *et al.* (2016)), as well as for determining the reasons for crying (e.g., pain or hunger) (Wahid *et al.* (2016)) and for

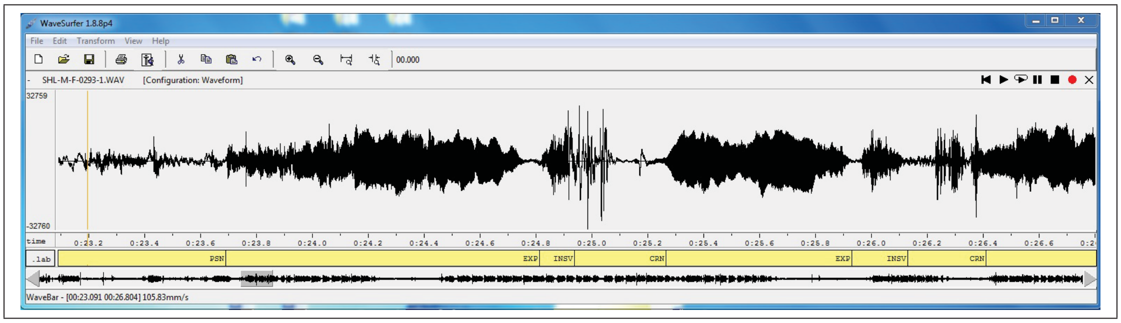


Figure 2.3 Example of a labeled CAS in the WaveSurfer software

diagnostic purposes (Alaie *et al.* (2016)). The present study followed the method developed by Jurafsky et al. (Jurafsky & Martin (2014)) for acquiring the MFCC features from CASs. The steps are discussed below:

A. Pre-emphasizing:

Initially, the CAS is filtered using a low-pass filter with a transfer function shown in equation 2.1:

$$H(z) = 1 - az^{-1} \tag{2.1}$$

This is performed to provide more weight to the higher frequencies of the CAS, which are generally smaller than those at the lower frequencies. The parameter, “a”, in the filter is set to 0.97 to correspond with the parameters used in our previous study (Alaie *et al.* (2016)).

B. Windowing:

Because the statistical properties of the signal are not constant over time, windowing results in a relatively stationary state for the CAS. The signal is separated into frames ranging from 10 ms to 50 ms. Several types of windowing functions are available, including rectangular, Hamming, and Hanning. However, rectangular windowing can generate skewed results during Fourier analyses. Hence, in this project, a Hamming window with a frame size of 10 ms was used, with a 30% overlap between consecutive frames.

C. Discrete Fourier Transform (DFT):

Next, the Fourier transform of each frame was calculated. The DFT conveys spectral information, i.e., the amount of energy possessed by the signal at different frequencies.

D. Mel Filter Bank and Log:

A mel is a unit of frequency that relates to the perceived pitch by the human ear (Jurafsky & Martin (2014)). It can be calculated directly from the frequency, using the following equation:

$$M(f) = 1125 \ln(1 + \frac{f}{700}) \tag{2.2}$$

The power spectrum is multiplied by overlapping triangular filters, called mel filters, to obtain the power spectrum of each of the mel bands. The cut-off frequencies of each filter correspond to the central frequencies of the neighboring filters. Below 1 kHz, the filters are separated linearly with a constant bandwidth. Above 1 kHz, the central frequency of each filter is 1.1-times greater than the central frequency of the previous filter, which produces logarithmic spacing. This configuration approximates the mel scale (the human perception of the pitch). In general, the number of filters used varies between 13 and 24. In the present study, the MFCC features were computed with 24 filter banks. Moreover, similar to previous work (Alaie *et al.* (2016)), we only considered information below the frequency of 4 k. As the final task in this block, the logarithm of the normalized energy of each band was calculated. First, the energy of each band was calculated by evaluating the sum of the coefficients of the power spectrum of the band. Subsequently, the energy was normalized according to the width of the band. The logarithm was applied to approximate the logarithmic response of the human ear to sound intensity.

E. The Cepstrum:

Inverse Discrete Fourier Transform (IDFT): The Mel cepstrum is obtained by calculating the IDFT of the logarithmic energies of each band. Most of the valuable information was found among the first factors produced by the transformation. Therefore, this step enables information compression.

In this study, we examined the performance of a linear support vector machine (SVM) classifier using 12–20 MFCCs in order to determine the most discriminative number of coefficients. Figure 2.4 shows the classifier accuracy for the inspiration episode dataset and the expiration episode dataset. Using the expiration episode dataset, 12 MFCC features would result in the highest accuracy. In the inspiration episode experiment, 13 MFCC features had the highest prediction rate. Thus, we extracted 12 MFCCs for expiration and 13 MFCCs for inspiration.

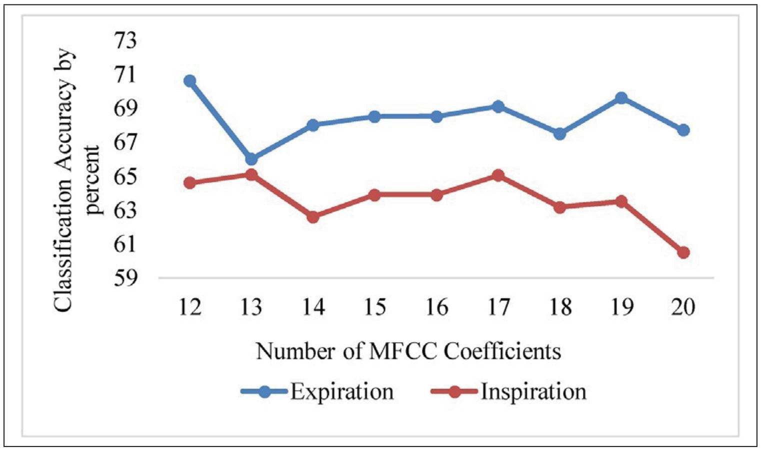


Figure 2.4 Classification accuracy using 12 to 20 MFCC coefficients

The preceding steps provide the static MFCCs, which we extracted 12 MFCCs for expiration and 13 MFCCs for inspiration respectively. Often energy feature also accompanies to the MFCCs; thereby we also added the energy of each frame to the static MFCC features. Eq. 2.3, below, is used to compute energy from time t_n to time t_{n+1}. In this equation, X is the main signal within the duration of t_n to t_{n+1}.

$$Energy = \sum_{t_n}^{t_{n+1}} X(t)^2 \tag{2.3}$$

Finally, we concatenated the dynamic MFCCs to the previously computed static MFCCs and energy feature. These dynamic MFCCs are called delta and delta–delta. Delta refers to the variation in the MFCC features over time. The temporal variation of the MFCC features was obtained and added to the primary feature vector, which has been shown to increase the accuracy of automatic speech recognition.

Similarly, the delta–delta features can be obtained from the delta features by computing the variation of the delta features over time. The delta features of the MFCC coefficients are calculated using Eq. 2.4 (Jurafsky & Martin (2014)):

$$D_n = \frac{\sum_{\Theta=1}^{\Phi} \Theta(f_{n+\Theta} - f_{n-\Theta})}{2\sum_{\Theta=1}^{\Phi} \Theta^2} \tag{2.4}$$

Eq. 2.4 uses the features from the previous frame and the next frame, in which "f" is the static feature or the static coefficient, and "Θ" is the number of frames. Since no value exists for the static coefficient before the first and after the last frames in the signal, the delta and delta–delta of the first and last frames in this signal are considered to be zero (Alaie *et al.* (2016)). Static and dynamic MFCCs, as well as energy feature are the features that were obtained. It is possible to feed this obtained matrix into the classifier; however, it is huge and contains redundancy (Zabidi *et al.* (2017)). If the size of the training data is huge (i.e., the matrix of the features), it may negatively affect the computing cost of the model being trained. Generally, the use of each of these methods is in accordance with the next stage in the designed system. Thus, the statistical measurements were applied to all the frames for the MFCC feature matrix to reduce unnecessary information and to represent the features that are most relevant to the classifier. Here, the range, mean, and standard deviation (Amaro-Camargo & Reyes-García (2007); Bhargava & Polzehl (2013)) were used, along with the median and interquartile range of the MFCC features.

2. **Tilt features.** Tilt features were used in this study to parameterize the melody features in the CAS. These features capture variations in the fundamental frequency (f_0) contour (Mary (2012)). They are defined by A_t and D_t in Eqs. 2.5 and 2.6 (Mary (2012)):

$$A_t = \left(\frac{|A_r| - |A_f|}{|A_r| + |A_f|}\right) \tag{2.5}$$

$$D_t = \left(\frac{|D_r| - |D_f|}{|D_r| + |D_r|} \right) \tag{2.6}$$

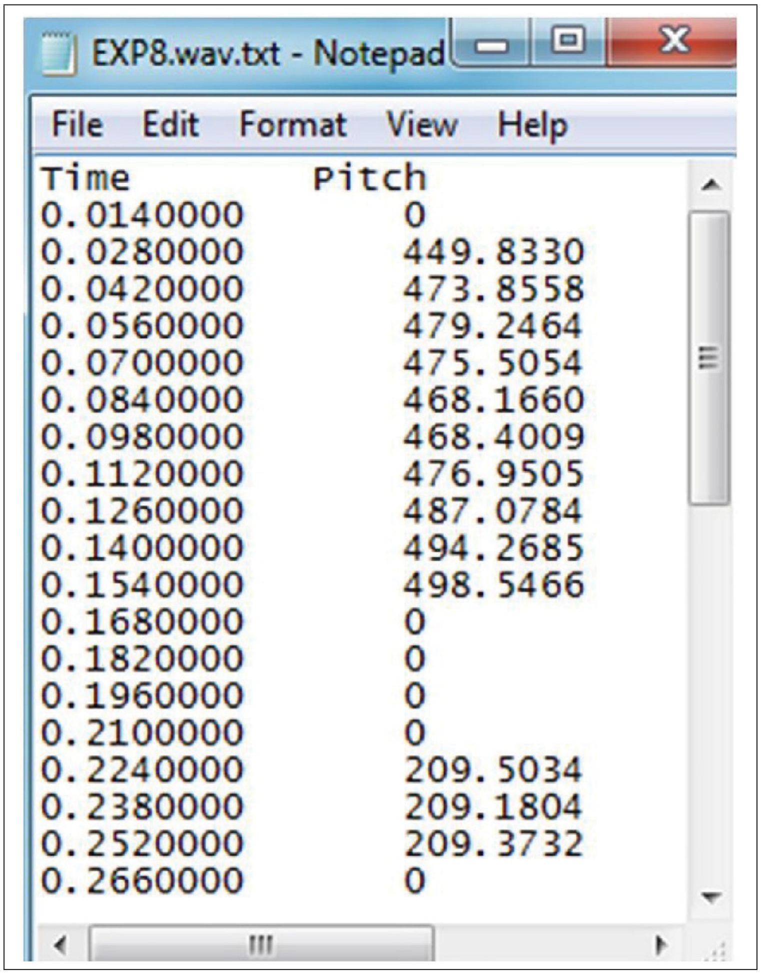
In a portion of a CAS, where the fundamental frequency is the highest, A_f and A_r capture the amplitudes of the f_0 contour when this contour is descending and ascending, respectively. Similarly, D_f and D_r capture the lengths of the descending and ascending portions of the contour, respectively.

To extract the melody feature of the tilt, a precise estimation of the fundamental frequency is required. However, the extraction and representation of f_0 is more difficult in newborn CASs than in adult voices, as the former are highly nonstationary and are in a higher octave range in comparison to adults; hence, a stationary model for the f_0 waveform does not suffice (Manfredi, Bandini, Melino, Viellevoye, Kalenga & Orlandi (2018); Mende *et al.* (1990); Moller & Schonweiler (1999)). In the present study, the f_0 contours were extracted using Praat acoustic analysis software as in practice it performs better that other available software tools (Orlandi, Bandini, Fiaschi & Manfredi (2017)). The specifications for the desired frequency ranges and the corresponding timescales were programmed in Praat scripting. An example of the obtained f_0 of an episode of a CAS is illustrated in Fig. 2.5. Subsequently, from the obtained f_0 contours, A_t and D_t were calculated for each segment from Eqs. 2.5 and 2.6. Next, the obtained values of A_t and D_t were concatenated for each CAS. A vector of tilt features was obtained for each CAS. Finally, the range, mean,standard deviation, median, and interquartile range of each of these feature vectors were calculated.

3. **Rhythm feature.** In speech recognition, vowel and consonant segments are the units for rhythm feature observations. We measured the temporal features of the two episodes of expiration and inspiration. The rhythm features we used are described below:

    - **A. Normalized Raw Pairwise Variability Index:** The raw Pairwise Variability Index (rPVI) characterizes the pattern of timing contrasts between successive extends for speech, which is applied to syllables or segments. The rPVI is described as follows (Fang, Li, Li, Shen & Shao (2012)):

$$rPVI = [\frac{\sum_{k=1}^{M-1} |d_k - d_{k+1}|}{m-1}] \tag{2.7}$$



```
Time          Pitch
0.0140000         0
0.0280000         449.8330
0.0420000         473.8558
0.0560000         479.2464
0.0700000         475.5054
0.0840000         468.1660
0.0980000         468.4009
0.1120000         476.9505
0.1260000         487.0784
0.1400000         494.2685
0.1540000         498.5466
0.1680000         0
0.1820000         0
0.1960000         0
0.2100000         0
0.2240000         209.5034
0.2380000         209.1804
0.2520000         209.3732
0.2660000         0
```

Figure 2.5 Example of f_0 contour extraction using Praat software

where, "d" is the duration of each episode and "m" is the number of episodes within a CAS record file. The normalized rPVI used for this experiment is as follows (Fang *et al.* (2012)):

$$nrPVI = 100 \times [\frac{\sum_{k=1}^{M-1} \left| \frac{d_k - d_{k+1}}{\frac{d_k + d_{k+1}}{2}} \right|}{m-1}] \tag{2.8}$$

- **B. Std:** The standard deviation of the episode durations contained in each CAS.
- **C. Varco:** The standard deviation of the episode durations divided by their mean duration in each CAS.

- **D. N events:** The number of episodes that occur in each CAS.
- **E. Total duration:** The total duration of each episode in each CAS.
- **F. Range:** The range (maximum less minimum) of the episode durations in each CAS.
- **G. Mean:** The average of all the episode durations in each CAS.

2.3.2.3 Classification and statistical evaluation

An SVM was the classification algorithm used in this experiment. In the multi-dimensional feature space, the goal of SVM is to obtain the border (hyperplane) with the furthest distance from the boundary feature points of each class. These boundary feature points are called support vectors, and they are used for training the classifier models. An SVM with a linear kernel was used in this experiment as a binary classifier. For training the SVM classifier, we used cross-validation to reduce the potential bias by dividing the dataset into testing and training parts. Ten folds were chosen to provide acceptable table trade-off between bias reduction and evaluation time. For each round of testing, the dataset was segmented into 10 parts: nine for training and one for testing. This procedure ended after 10 rounds, when each part had been used for testing once.

Table 2.3 shows the number of healthy and unhealthy CASs in each fold for the expiration and inspiration datasets. The CASs of each of the infants were sorted so they would be in one fold; hence, the CASs of an infant, with which the classifier was trained, were only used for the training data; in order to ensure that the folds were independent, they were not included in the testing fold. We also evaluated the statistical significance of the accuracy values achieved by each classifier using the MFCC feature set and the combined feature sets. The classifier results and p-values are reported in the following section.

2.4 Results

Several evaluations were performed to assess the effectiveness of each individual feature set and different combination of these features on the classification accuracy. For each feature set

Table 2.3 Number of used CASs for healthy and unhealthy infants in each of experiments

	fold	Healthy	RDS	Test data	Train data	Total
Expiration Dataset	1-5	10	10	20	171	191
	6-9	10	9	19	172	
	10	8	7	15	176	
Inspiration Dataset	1-4	10	10	20	165	185
	5-9	9	9	18	167	
	10	8	7	15	170	

combination, the array of observations associated with each predicator was normalized using Z-scores. In total, six groups of feature types were supplied to the SVM classifier for training:

Individual features:

- MFCCs only;
- Tilt features only; and
- Rhythm features only.

Combined features:

- Normalized MFCCs and tilt features;
- Normalized MFCCs and rhythm features; and
- All three feature types (normalized MFCCs, tilt, and rhythm).

The efficacy of each feature group was then evaluated using several measures. Although the accuracy measure is algorithm-performance explanatory, in order to more accurately present the efficiency of the algorithms used in this study, other measures, such as true positive rate (also known as recall), false positive rate,precision, and F-score, were evaluated. The aforementioned measurements are widely used to select the prime model in studies where the cost of the misclassification of true positive is crucial, such as when identifying sick infants from healthy ones or when stating the need of an infant based on his or her CAS (Osmani, Hamidi & Chibani

(2017); Rodriguez & Caluya (2017)). The accuracy of each group is defined by Eq. 2.9, where the Test Error Rate indicates the percentage of the test data misclassified by the trained model.

$$Accuracy = 1 - Test\ Error\ Rate \tag{2.9}$$

The true positive rate (recall), false positive rate, and precision measures are, respectively, represented in Equations 2.10 to 2.12:

$$TP\ Rate = \frac{\sum Labeled\ correctly\ as\ RDS}{\sum Labeled\ correctly\ as\ RDS + Labeled\ wrongly\ as\ healthy} \tag{2.10}$$

$$FP\ Rate = \frac{\sum Labeled\ wrongly\ as\ RDS}{\sum Labeled\ wrongly\ as\ RDS + Labeled\ correctly\ as\ healthy} \tag{2.11}$$

$$Precision = \frac{\sum Labeled\ correctly\ as\ RDS}{\sum Labeled\ correctly\ as\ RDS + Labeled\ wrongly\ as\ healthy} \tag{2.12}$$

The F-score is defined as a function of recall and precision, as shown in Eq. 2.13:

$$Fscore = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.13}$$

The overall classification results for individual features and the results of combining the short-term and long-term feature sets are shown in Tables 2.4 and 2.5. These results were obtained by taking the average of the classification results in 10 portions of the dataset. The used samples in each of the portions are independent of each other, and are from different infants. To measure the usefulness of adding more information to the system, and to assess whether the improvements in the classifier accuracy were statistically significant, the accuracy results of 10 iterations were compared using pairwise t-test statistical analysis with a significance threshold of p = 0.05. Table 2.6 shows the p-values obtained by comparing the MFCC-only classifier with the MFCC with rhythm, MFCC with tilt, and MFCC with tilt and rhythm feature sets. Moreover, the bar graphs shown in Fig. 2.6 present the statistical significance of adding each type of information from the MFCC feature set, and the error bars show the variability of the accuracy rate.

Fig. 2.6 shows that the differences between the groups of MFCCs and the combined feature sets are statistically significant, for both the expiration and inspiration experiments, except for the combination of the two feature sets of MFCC and rhythm, and the combination of the three feature sets of MFCC, rhythm, and tilt for the inspiration episode.

Table 2.4 Evaluation of SVM classifier performance using the proposed individual and combined feature sets for the expiration dataset

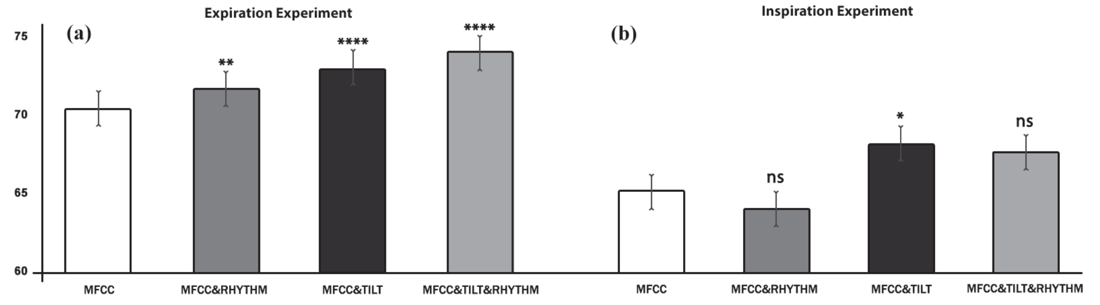
Expiration	Accuracy%	TP Rate%	FP Rate%	Precision%	F-score%
MFCC	70.60	43.00	3.10	93.00	58.80
Tilt	55.50	63.40	52.00	53.60	58.10
Rhythm	44.50	35.50	46.90	41.70	38.40
MFCC+Tilt	73.30	60.20	14.30	80.00	68.70
MFCC+Rhythm	72.20	46.20	3.10	93.40	61.80
MFCC+Tilt +Rhythm	73.80	60.20	13.20	81.10	69.10

Table 2.5 Evaluation of SVM classifier performance using the proposed individual and combined feature sets for the inspiration dataset

Inspiration	Accuracy%	TP Rate%	FP Rate%	Precision%	F-score%
MFCC	65.10	36.60	7.50	82.50	50.70
Tilt	60.70	71.10	49.20	58.20	64.00
Rhythm	50.90	32.20	31.00	50.00	39.20
MFCC+Tilt	68.40	43.00	6.90	85.50	57.00
MFCC+Rhythm	63.70	32.20	5.80	84.00	46.60
MFCC+Tilt +Rhythm	67.80	41.10	6.40	86.00	55.60

Table 2.6 P-values obtained by performing a t-test on the classifier accuracies

t test	Expiration Episode	Inspiration Episode
MFCC and (MFCC + rhythm)	0.001521	0.127377
MFCC and (MFCC + tilt)	1.98 * 10e-7	0.043246
MFCC and (MFCC+tilt+rhythm)	8.180*10e-6	0.061211



P value	Explanation	Summary
Less than 0.0001	Extremely Significant	****
Between 0.0001 and 0.001	Extremely Significant	***
Between 0.001 and 0.01	Very Significant	**
Between 0.01 to 0.05	Significant	*
More than 0.05	Not Significant	ns

Figure 2.6 (a), (b) Bar graphs showing the mean average of 10 runs for classifiers of MFCC, MFCC and rhythm, MFCC and tilt, and the three feature sets of MFCC, rhythm and tilt for the expiration and inspiration episodes. Moreover, the figure presents information about the error bars and the statistical difference of each category from the MFCC category only

2.5 Discussion

In this study, the NCDS was designed to assess three criteria: generalizable, unbiased by region, and robust in determining the reason for crying. We created a large dataset (78 healthy infants and 34 infants with RDS) collected from two different hospitals in Lebanon and Canada, with parents from different regions in those countries. This number of individuals in our dataset allowed us to use the CASs of a number of infants (see Table 2.3) in the training set, and, subsequently, to use the CAS of the remaining infants in the testing set in which the CASs of each infant is only used in one fold. The individual characteristics did not influence the classification, thus allowing the system to be generalizable. To determine if the NCDS met the criteria of being unbiased by region or language, the CAS collection from infants in different geographical regions in which the infants' parents are from different language communities would challenge the system's performance while using the melody features of tilt. The CASs were initiated for a variety of reasons, including birth, blood tests, completion of a shower, fear, hunger, weighing, etc. To use rhythm features to distinguish unhealthy (RDS) CASs from healthy CASs without regard to the reason for crying, we created the system to be blind to the reason for crying.

With regard to these challenges, the first three rows presented in Tables 2.4 and 2.5 show that the accuracy of the classifier was lower for the long-term features in comparison to the MFCC features. However, in terms of the F-score measure, in the expiration experiment, the classifier's performance using tilt features was almost as good as the performance when using MFCC features. Moreover, in the inspiration experiment, tilt features outperformed the MFCCs, while the tilt features are normally dependent on the linguistic group of the infants' parents (Mampe *et al.* (2009); Manfredi *et al.* (2019); Wermke *et al.* (2017)). In the present study, the good classifier performance results (in terms of the true positive rate and the F- score) using the tilt feature set is consistent with the previously reported changes in the minimum and maximum pitches as well as the occurrence of the melody types of flat, rising, and falling/rising of the CASs of infants with RDS in comparison to healthy infants (Boukydis & Lester (2012)). To compare the CASs of healthy infants with those with RDS, the rhythm features (Wasz-Hockert (1968b)), also included changes in duration; in our experiments, rhythm alone did not result in significant accuracy. This may be because the rhythm feature is not distinctive enough to describe the durational feature in the inspiration episodes of healthy and RDS infants. Further investigation and experiments are necessary to guarantee the usefulness of this feature.

Regarding the combined feature sets, as seen in the last three rows of Tables 2.4 and 2.5, in the expiration experiment, always adding the tilt and rhythm feature sets to the MFCC features increased the accuracy and the F-score measures. As seen in Table 2.4, the best classification result is for the combination of the tilt, rhythm, and MFCC feature sets. Thus, in the expiration experiments, the accuracy increased by more than 3% and the F-score measure increased by more than 10%. In the inspiration experiment, the best-obtained result among the combined features was for the MFCC and tilt feature set. The accuracy increased more than 3% and the F-score increased more than 6% for the classifier that only used the MFCC feature set in comparison to the classifier that used the MFCC and tilt feature sets. Thus, the long-term features proposed in this paper alone would not have an accuracy result and an F-score as significant as the short-term features of MFCCs, unless one considers the F-scores of the tilt features in the inspiration experiment. Moreover, we noticed that, in some cases, combining the long-term features with the MFCC features improved the classifier's performance in comparison to the classifier that only used the MFCC features (the combination of all three features in the expiration episodes, and the combination of the tilt feature set and the MFCCs in inspiration episodes). Fig. 2.7 shows the Receiver Operating Characteristic (ROC) space, which is the true positive rate versus the false positive rate obtained from classifier. The graph in Fig. 2.7 shows the changes in the classifier performance after combining the features. For the expiration episode dataset, the best result was obtained for the combination of all three feature sets; for the inspiration episode dataset, the best result was obtained for the combination of the MFCC and tilt features.

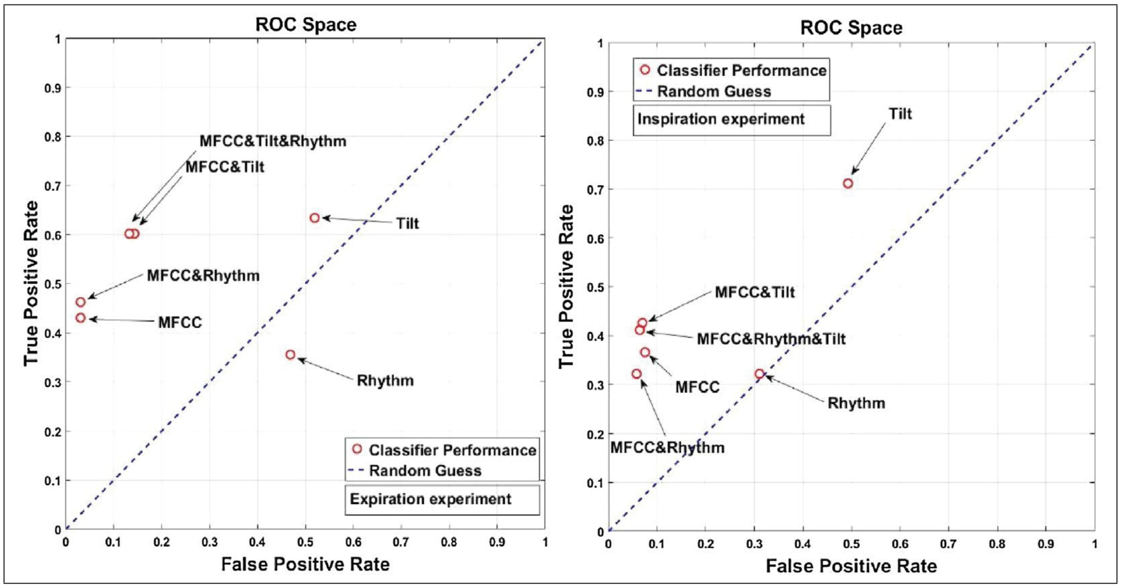


Figure 2.7 ROC spaces showing the performance of the classifiers using each of feature set for the inspiration and expiration episodes

In domains other than those used for infant CAS processing, long-term features have also been reported to contribute to the better performance of automatic speech recognition and automatic speaker recognition (Adami *et al.* (2003); Vicsi & Szaszák (2010)) when added to short-term features. In the present study, the t-test evaluation supported the usefulness of the combined features, in particular, for the expiration episode. In all the combination cases, the results were statistically significant except when the rhythm and MFCC features were combined, and when the MFCC, rhythm, and tilt features were combined for the inspiration episode. Consequently, in the expiration episode, adding long-term feature information consistently helped the classifier achieve a higher degree of accuracy; in contrast, in the inspiration episode, only the addition of the tilt was helpful.

Less operating time was required to extract the rhythm features than the tilt and MFCC features. The rhythm features were computed from the time domain directly without any frequency transformation; however, for the tilt and MFCC features, the CASs were transformed to the frequency domain and the information was extracted from that domain. Table 2.7 shows the feature extraction time for each set of features.

Table 2.7 Feature extraction estimated time for the MFCC, tilt, and rhythm features

Feature	**Ealapsed time(s)**
MFCC feature set	1469.47
Rhythm feature set	2.90
Tilt feature set	1163.89

In each evaluation stage, each group of features was normalized; they were then supplied directly to the classifier without any extra processing or manipulation. Hence, adding the tilt features to the MFCC features was more time-intensive than adding the rhythm features. As shown in Tables 2.4 and 2.5, adding the tilt features to the MFCC features resulted in a greater improvement in accuracy than adding the rhythm features. Additionally, Table 2.6 shows that the accuracy of the group with the combined MFCCs and tilt features was statistically more significant than the accuracy of the group with the combined MFCCs and rhythm features. Thus, the time-related cost was rewarded with higher accuracy. In terms of memory, the long-term features (tilt and rhythm) occupied less space than the short-term features. Concerning the MFCC features used in the present study, after extracting the features based on frames, the statistical measures of range, mean, standard deviation, median and interquartile range were computed. We found it a computationally efficient approach for feature extraction, however we acknowledge it leads to losing temporal granularity of MFCC feature. We hope to address this issue in future work.

Another potential future work is identifying other long-term patterns in CASs and determining a better way to combine the long-term and short-term features to improve the power of the NCDS. A method should also be proposed for managing the features that does not degrade the feature space or the system performance.

2.6 Conclusion

In this work, MFCC features (short-term features) and the long-term features of tilt and rhythm, as well as several combinations of these feature sets, were used to create an NCDS. The combination of long-term (melody and rhythm) and short-term (MFCCs) features was found to provide a better classification performance for differentiating the CAS of healthy infants from infants with RDS in comparison to using short-term features alone, particularly for the expiration episodes. The best improvements of the results (F-score) that we achieved were 10.3% and 6.3% in the expiration episode and inspiration episode experiments, respectively. Moreover, in the inspiration episode, the tilt feature alone resulted in the highest F-score in comparison to all the individual and combined feature sets. The results of this study demonstrate the importance of using long-term features as diagnostic markers for RDS. Furthermore, the expiration and inspiration episodes of infant CAS demonstrated distinctive prosodic patterns between groups of healthy infants and infants with RDS.

2.7 Acknowledgments

This research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) [RGPIN-2016-05067], and has been made possible through the funding provided by the Bill and Melinda Gates Foundation [OPP1025091]. We would also like to thank Habiba Dahmani, who developed the initial idea of investigating the usefulness of rhythm features in an NCDS, and Prasun Lala from the ETS library for his feedback on the manuscript.

2.8 Acknowledgments

The authors declare no conflict of interest.

CHAPTER 3

MACHINE LEARNING-BASED CRY DIAGNOSTIC SYSTEM FOR IDENTIFYING SEPTIC NEWBORNS

Fatemeh Salehianmatikolaie[1], Chakib Tadj[1]

[1] Department of Electrical Engineering, École de technologie supérieure,
1100 Notre-Dame West, Montreal, Quebec, Canada H3C1K3

Paper submitted in the Journal of Voice, 2021.

Résumé

Le traitement du signal audio des pleurs des nouveau-nés (Cry Audio Signal, CAS) fournit des informations utiles sur l’état de santé des nouveau-nés. Ces informations peuvent être utilisées pour diagnostiquer une maladie ou comprendre les besoins du nouveau-né.

Cet article analyse les CAS des nouveau-nés de moins de deux mois en utilisant des approches d’apprentissage automatique pour développer un système de diagnostic automatisé permettant d’identifier les nourrissons septiques des nourrissons en bonne santé. Les nourrissons septiques n’ont pas été étudiés dans ce contexte.

Les caractéristiques proposées comprennent les coefficients cepstraux de fréquence Mel et les caractéristiques prosodiques d’inclinaison, de rythme et d’intensité. Les performances de chaque ensemble de caractéristiques ont été évaluées à l’aide d’une série de classificateurs, notamment la machine à vecteurs de support (Support Vector Machines, SVM), l’arbre de décision et l’analyse discriminante. Nous avons également examiné la méthode du vote majoritaire pour améliorer les résultats de la classification ainsi que la manipulation des caractéristiques et le cadre des classificateurs multiples, ce qui n’a pas été rapporté auparavant dans la littérature relative au développement d’un système de diagnostic automatique basé sur le CAS du nourrisson.

Les meilleurs résultats F-score obtenus sont pour le cadre de la concaténation de tous les ensembles de caractéristiques en utilisant le SVM quadratique avec 86%, et le cadre de l'ensemble de caractéristiques d'inclinaison avec le discriminant quadratique avec 83,90% respectivement pour les deux ensembles de données des épisodes d'expiration et d'inspiration du CAS des nouveau-nés. Grâce à ces expériences, nous avons découvert que les nourrissons septiques pleurent différemment des nourrissons sains. La méthode que nous proposons peut donc être utilisée comme un outil non invasif pour identifier les nourrissons septiques des nourrissons sains uniquement sur la base de leur CAS.

Mots-clés: Sepsis, cri du nourrisson, coefficient cepstral de fréquence Mel, caractéristique prosodique, PCA, manipulation de caractéristiques, machine à vecteurs de support, arbre de décision, analyse discriminante, fusion de classificateurs.

3.1 Abstract

Processing the newborns' cry audio signal (CAS) provides useful information about the newborns' condition. This information can be used to diagnose the disease or, to understand the newborns' needs. This article analyzes the CASs of newborns under two-month-old using machine learning approaches for developing an automatic diagnostic system for identifying septic infants from healthy ones. Septic infants have not been studied in this context.

The proposed features include Mel frequency cepstral coefficients, and the prosodic features of tilt, rhythm, and intensity. The performance of each feature set was evaluated using a collection of classifiers including Support Vector Machine (SVM), decision tree, and discriminant analysis. We also examined the method of majority voting for improving the classification results as well as feature manipulation and multiple classifier framework, which has not previously been reported in the literature relating to developing an automatic diagnostic system based on the infant's CAS.

The best obtained F-score results are for the framework of the concatenation of all feature sets using quadratic SVM with 86%, and the framework of tilt feature set with quadratic discriminant with 83.90% respectively for the two datasets of expiration and inspiration episodes of newborns' CAS. Through these experiments, we found out that septic infants cry differently than healthy infants. Thus our proposed method can be used as a noninvasive tool for identifying septic infants from healthy ones only based on their CAS.

Keywords: Sepsis, infant cry, Mel-Frequency Cepstral Coefficient, prosodic feature, PCA, Feature Manipulation, Support Vector Machine, Decision tree, Discriminant analysis, Classifiers Fusion

3.2 Introduction

In recent years, the infant mortality rate in developed countries has decreased. However, this rate is still high in developing countries. Saving newborns' lives and promoting their health is of particular importance in the health of any nation and for further providing health services. In this paper, we set out a Newborn Cry Diagnostic System (NCDS) to see if we can apply audio signal processing techniques to investigate features of different domains and manipulate features to make decisions about categorizing newborns' cry audio signal (CAS) as septic or healthy. In this study by CAS, we refer to the sound waveform that the infant produces by pushing airflow from their lungs to the vocal track. In this section, we will discuss what CAS is, the types of NCDSs proposed by researchers, the problems they faced, and how we can apply them to sepsis pathology, which has not been studied before.

The act of crying for infants is their most prominent communication activity. Crying is their only weapon against the inconveniences like hunger, pain, discomfort, and infection that happens to them. Crying is a natural warning method to call on those around to help. Not responding properly to these warning signs can cause harm to the infant and his or her parents. A fair number of researchers indicated that infants' CAS holds a lot of information that, if properly analyzed, can be used to access messages sent from the newborn's brain (Boukydis & Lester (2012)). We also know that mothers and hospital staff who are constantly in contact with infants are able to distinguish several types of infant needs, only based on their CAS (Mukhopadhyay *et al.* (2013)). Further investigations on infants' CAS even revealed its reliability for diagnostic purposes (Michelsson & Michelsson (1999)). In (Boukydis & Lester (2012)), they anecdotally explained the characteristics of the CASs of infants affiliated with certain diseases such as asphyxia, deafness, etc. versus healthy ones. There are patterns in a CAS that warn about the pathology that is menacing for the health of the infant which may be clueless even in physical examinations by doctors (Abdulaziz & Ahmad (2010)). The infant CASs has been studied for decades (Manfredi *et al.* (2018)). Traditional popular approaches were based on visual inspections of the spectrogram of infant CASs (Boukydis & Lester (2012)). However, manually sorting the patterns in CAS and categorizing accordingly are not practical for human beings due to the huge amount of information for processing (Abou-Abbas *et al.* (2016)). Thus, this shortcoming has led to the development of various automatic classification systems. There have been works on developing an automatic system for recognizing the infant CASs from other surrounding sounds (Kim *et al.* (2013)), detecting different parts of CASs (such as episodes of expiration and inspiration) (Abou-Abbas *et al.* (2016); Aucouturier *et al.* (2011)), identifying the need of an infant (hunger, diaper, sleepy, etc.) (Abdulaziz & Ahmad (2010); Saha *et al.* (2013); Wahid *et al.* (2016)), and the very recent one is processing for diagnosing pathology task (Alaie *et al.* (2016); Orozco-García & Reyes-García (2003); Salehian Matikolaie & Tadj (2020)). In our work, we also concentrated on diagnostic pathology using an NCDS.

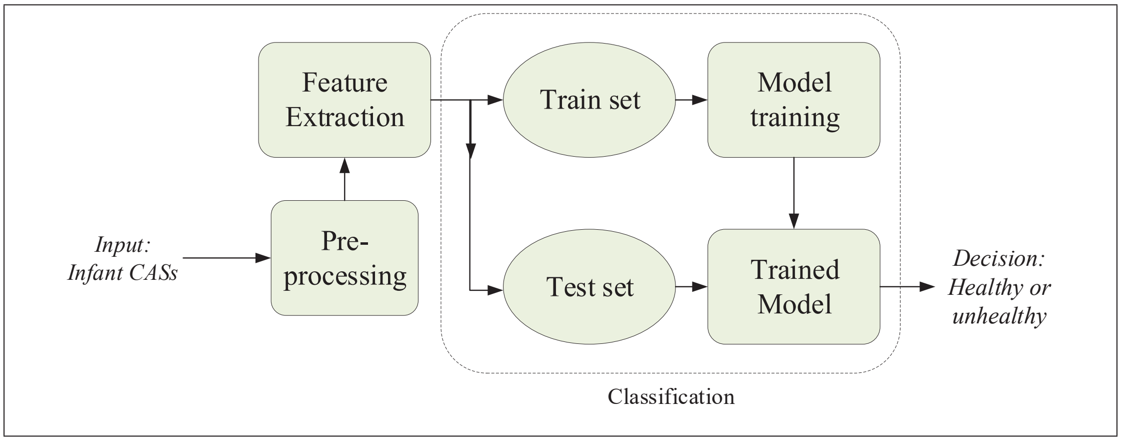


Figure 3.1 Block diagram of the Newborn Cry Diagnostic System (NCDS)

Figure 3.1 shows the block diagram of the NCDS. The NCDS framework like any identification system includes the phases of pre-processing, feature extraction, and a phase of training a model based on obtained features for classification. The aim of pre-processing step is to better help feature extraction. It includes applications such as pre-emphasizing, windowing, and finding the fundamental frequency. In phase of feature extraction, the methods such as Mel Frequency Cepstral Coefficient (MFCCs) (Alaie *et al.* (2016); Hariharan, Sindhu, Vijean, Yazid, Nadarajaw, Yaacob & Polat (2018); Kheddache & Tadj (2019); Martinez-Cañete, Cano-Ortiz, Lombardía-Legrá, Rodríguez-Fernández & Veranes-Vicet (2018); Rosales-Pérez *et al.* (2015); Salehian Matikolaie & Tadj (2020)), Linear Prediction Coding (LPC) (Hariharan *et al.* (2018); Martinez-Cañete *et al.* (2018); Rosales-Pérez *et al.* (2015)), patterns of fundamental frequency contour (Kheddache & Tadj (2019); Salehian Matikolaie & Tadj (2020)), resonance frequency (Kheddache & Tadj (2019)) are the most common ones. Furthermore in this phase extra analysis such as combining feature sets from different techniques for feature set representation such as integrating features of MFCC and LPC (Martinez-Cañete *et al.* (2018)), as well as merging MFCC, rhythm, and tilt features (Salehian Matikolaie & Tadj (2020)), or techniques for identifying the most relevant features such as F-ratio and Binary Particle Swarm Optimization (Sahak *et al.* (2012)), Orthogonal Least Square Algorithm (Sahak *et al.* (2012)) for improving the classification performance were suggested.

In the classification phase a variety of pattern recognition models have been studied including Support Vector Machine (SVM) (Alaie *et al.* (2016); Salehian Matikolaie & Tadj (2020)), Multilayer Perception Neural Network (Alaie *et al.* (2016); Sahak *et al.* (2012)), Probabilistic Neural Network (Alaie *et al.* (2016); Hariharan *et al.* (2012a); Kheddache & Tadj (2019)), decision tree (Rosales-Pérez *et al.* (2015)), Forest (Rosales-Pérez *et al.* (2015)) and k-nearest neighbor algorithm (Rosales-Pérez *et al.* (2015)).

The CASs of pathologies that were yet investigated by machine learning approaches to automatically identify sick infants from healthy ones includes cleft palate (Lederman *et al.* (2002,0)), hearing disorder (Hariharan *et al.* (2018,1); Orozco-García & Reyes-García (2003); Rosales-Pérez *et al.* (2015); Wahid *et al.* (2016)), hyperbilirubinemia (Santiago-Sánchez *et al.* (2009)), autism

(Orlandi, Manfredi, Bocchi & Scattoni (2012)), asphyxia (Hariharan *et al.* (2018); Reyes-Galaviz *et al.* (2005); Rosales-Pérez *et al.* (2015); Sahak *et al.* (2010b,1); Santiago-Sánchez *et al.* (2009); Wahid *et al.* (2016); Zabidi *et al.* (2017)), hypothyroidism (Zabidi *et al.* (2009b)) and respiratory distress (Lederman *et al.* (2002); Salehian Matikolaie & Tadj (2020)).

In this study, our contribution is twofold. One contribution is the way we evaluate and manipulate features and the way that we use these features to make a final decision. The second contribution is to look at the unstudied pathology of sepsis.

We performed four sets of experiments. In the first experiment, we considered each expiration episode and inspiration episode of infants' CAS as a sample. The expiration episode and inspiration episode are respectively any perceivable sound during exhalation and inhalation of infants during crying, and the silence episode is the soundless gap between inspiration and expiration episodes of CAS (Grau *et al.* (1995); Robb & Goberman (1997)). We refer to this experiment as the Single Episode (SE) experiment. We evaluated the performances of prosodic features of intensity, rhythm, tilt, and the commonly used feature of MFCCs using three sets of classifier families of SVM, discriminant analysis, and decision tree in each episode of expiration and inspiration.

In the second experiment, we used the predicted labels for episodes within each CAS from SE experiment to predict each CAS label using the majority voting technique. We call this experiment as All Episode (AE) experiment. We borrowed this idea from automatic environmental sound classification presented by (Abdoli, Cardinal & Lameiras Koerich (2019)).

In our previous study which was on analyzing the CASs of respiratory distress infants, we concatenated MFCC features with the two prosodic features of tilt and rhythm (Salehian Matikolaie & Tadj (2020)). In third experiment, we expanded on this idea, and concatenated the feature set of MFCC and the three prosodic feature sets of tilt, rhythm, and intensity, and then fed them to the classifiers. In the fourth experiment, we set up a framework to aggregate the prediction of the most competent classifiers for each individual set of features, and then to predict the CAS label using the majority voting technique. Our approach can handle classifying the CASs regardless of the duration of CASs, reasons of crying, and the ambient noise as will be discussed later in the paper.

Regarding our second contribution, according to our knowledge despite the frequent infant death due to sepsis, disappointingly, so far, there is no investigation on the connection between the CASs of infants with sepsis. In Canada alone in 2019, among the newborns' cause of death, sepsis is reported on rank 6 (Government of Canada (2020)). The rank of sepsis among leading to death has increased in recent years as shown in Table 3.1). Thus it would be very useful to have an automated infant cry system that can classify septic from healthy ones.

Sepsis is a serious disease that is usually caused by bacteria. Infants under two months are more prone to sepsis because their immune systems are not yet developed enough to fight off some sources of infection. In clinical findings, a set of certain symptoms are reported for sepsis. However newborns have few obvious symptoms; moreover, these symptoms may vary from child to child. Thus it would be very useful to have an automated infant cry system that can classify septic newborns from non-septic ones. Perhaps the reason that this important pathology of septic remained unstudied is that enough data did not exist. Thus having this dataset available in our lab lends support to delivering this work. An NCDS is a useful tool in saving the lives and promoting the health level of newborns specifically in developing countries where are suffering from the lack of pediatricians. The NCDS would address this issue as its installation cost is relatively low (Alaie *et al.* (2016); Manfredi *et al.* (2018)). Practical applications of the NCDS include its use for infant screening (Prathibha, Putta, Srinivas & Satish (2012)), infancy education (Ruvolo & Movellan (2008)), robot nursing (Yamamoto, Yoshitomi, Tabuse, Kushida & Asada (2013)), and as a medical assistant for pediatricians. Moreover, NCDS is a non-intrusive tool.

The paper is prepared as follows: Section 3.3 is assigned to describe the collection of data sets, information of the dataset, the participants, feature sets definitions, and explanations of the examined classifiers in this work; Section 3.4 is for reporting the results of the four implemented experiments including the SE experiment, the AE experiment, as well as the results of feature manipulation and use of multiple classifier framework. And Section 3.5 concentrates the discussion of the research developed including the usefulness of each feature set, the feature reduction schemes, the majority voting, and the computation cost of each framework.

Table 3.1 Leading cause of death related to sepsis among newborns in Canada
Adapted from Government of Canada (2020)

	2015	**2016**	**2017**	**2018**	**2019**
Sepsis rank	9	8	6	8	6
Number of infant death	31	32	43	47	38

3.3 Materials and Methods

3.3.1 Dataset Description

In this section, we described how the data was collected, the dataset details, the participants in our experiments, and the dataset pre-processing procedures.

3.3.1.1 Data Collection and Recording

The research group in our laboratory collected the CASs of groups of infants at Hôpital Sainte-Justine in Montréal, Canada, and hospitals of Al-Sahel Hospital and Al-Raee hospital in Lebanon. The hospitals' staff of mentioned hospitals recorded the CASs in the clinical medium. A 2-channel digital hand-held Olympus recorder system was posed at a distance of 10 to 30 centimeters from the infant. The sampling frequency of the recordings is 44.1 kHz, and the sample resolution is 16 bits. The duration of each sample is variable between two to three minutes.

Alongside the recording phase, they collected details of infants including the reason of crying, gestational age, birth weight, Apgar [1] result, gender, name of the hospital, type of disease, baby's age during the recording and prematurely state of the baby. The distinguishing feature of our

[1] The very first test taken from newborns for measuring the newborn's general health state.

database is that it contains the CASs of infants taken in a hospital medium with a variety of noises including the sound of the environment such as the sound caused by the microphone moved by someone, a door sound, speech, etc. Also, we have not limited our dataset to a certain reason of crying, it includes the CASs of infants initiated by a variety of reasons and are recorded at different times in a day (Abou-Abbas, Tadj & Fersaie (2017)). The reason of crying includes CAS initiated by hunger, discomfort, diaper, blood tests, shower, birth, collection of urine, etc.

Table 3.2 Example description of some CAS lables

Labels	**Description**
EXP	Voiced expiration segment during a period of cry
EXPN	Unvoiced expiration segment during a period of cry
INS	Unvoiced inspiration segment during a period of cry
INSV	Voiced inspiration segment during a period of cry
EXP2	Voiced expiration segment during a period of pseudo-cry
INS2	Voiced inspiration segment during a period of pseudo-cry
PSEUDOCRY	Any sound generated by the baby and it is not a cry
Speech	Sound of the nurse or parents talking around
Background	The kind of noise so low, it is characterized by a very low power-silence affected with little noise
BIP	The sound of the medical instruments next the baby
Noisy cry	Any sound heard with the cry: machine's beep sound, water, diaper, etc.
Noisy pseudo-cry	Any sound heard with the pseudo-cry
Noise	Like the sound caused by the mic moved by someone, the diaper, a door sound, speech + background, speech +beep.

3.3.1.2 Participants

The age range of infants in our dataset is from one day to 208 days. However, in the current experiment similar to our previous ones (Alaie *et al.* (2016); Kheddache & Tadj (2019); Salehian Matikolaie & Tadj (2020)), we excluded the CASs of infants whose age were more than 53 days. This is because infants above this age can control their voice (Boukydis & Lester (2012)). The groups of infants that were studied in this experiment are the infants affected by sepsis vs. healthy ones. In our dataset, there are the CASs of 17 infants with sepsis who were diagnosed by pediatricians through medical examinations. In general, there are 53 recordings from these infants. Each infant in our dataset has one, or more recordings. For another class that is healthy, there is numerous number of CAS, in which we only used an equal number of samples as sepsis class to observe the balanced dataset for precise diagnosis by our classification models.

Unlike other studies, we imposed no criteria to select data in our dataset. Our dataset was very variable in conditions. Firstly, we included all reasons of crying initiated for a variety of reasons, while the reason of crying affects the durational feature of CAS (Salehian Matikolaie & Tadj (2020)). Secondly, we considered a wide variety of newborns whose parents are from different linguistic groups. This is of importance as we have this knowledge that the unborn infants start learning the prosodic features such as rhythm, intensity, and melody from the last three months of pregnancy, and this affects the prosodic aspect of CAS production as it is discussed in (Manfredi *et al.* (2019)). Lastly, the CASs were recorded in hospitals which include ambient noises such as humans speech, the sound of the instrument, etc.

3.3.1.3 Dataset Pre-processing

In our experiment, the CASs underwent several pre-processing stages. Initially, the medical-related experts at hospitals annotated the different segments of CAS such as expiration episode, inspiration episode, etc. These assigned episodes (components of CAS) are explained in Table 3.2). The process of the CASs episodes' annotations was performed using the WaveSurfer software.

In categorization, the infants' CAS components are expiration and inspiration which are divided by silence (Grau *et al.* (1995); Robb & Goberman (1997)). We used the vocal segments of expiration and inspiration episodes of CAS which are explained in Table 3.2) as "EXP" and "INSV". This selection is based on the usefulness of these segments as explained by several researchers (A, E, Ca, J & J (2012); Abou-Abbas, Alaie & Tadj (2015a); Grau *et al.* (1995)).

3.3.2 Methodology

In this study, we used two datasets of expiration and inspiration episodes of infant CASs. In the SE experiment, we extracted features from several domains including tilt, rhythm, intensity, and MFCCs from each episode of expiration and inspiration datasets. Then we fed them to different models for classification. We examined each dataset separately. Figure 3.2 illustrates the scheme of SE experiment for a portion of CAS.

In the AE experiment, we used all predicted labels of each single episode in a CAS from SE experiment to predict the label of the single CAS using the majority voting method. Figure 3.3 shows the scheme of AE experiment. In another set of experiment similar to the method in our previous study (Salehian Matikolaie & Tadj (2020)), we concatenated all feature sets together, and in this paper, we also added the intensity features. In the final set of experiment, we used the best classifiers for each set of features and label the CAS based on the most predicted labels. The aim was to choose the framework which results in the most accurate recognition for identifying the CASs of septic infants.

3.3.2.1 CAS Feature Description

The extracted features are in the temporal, spectral, and in both domains. We adopted MFCC features, and the prosodic features of tilt, intensity, and rhythm. In the following paragraphs, we bring the description of each of these feature sets, and the details of the parameters we used.

3.3.2.2 Mel Frequency Cepstral Coefficients (MFCC)

Among several algorithms introduced in speech processing for characterization, MFCC is the most widely used method in both adult and infant voice processing (Salehian Matikolaie & Tadj (2020)).

A set of transformations are applied to the audio signal to acquire the MFCC coefficients such as filtering to reduce the impact of the vocal tract, applying the windowing technique to each frame

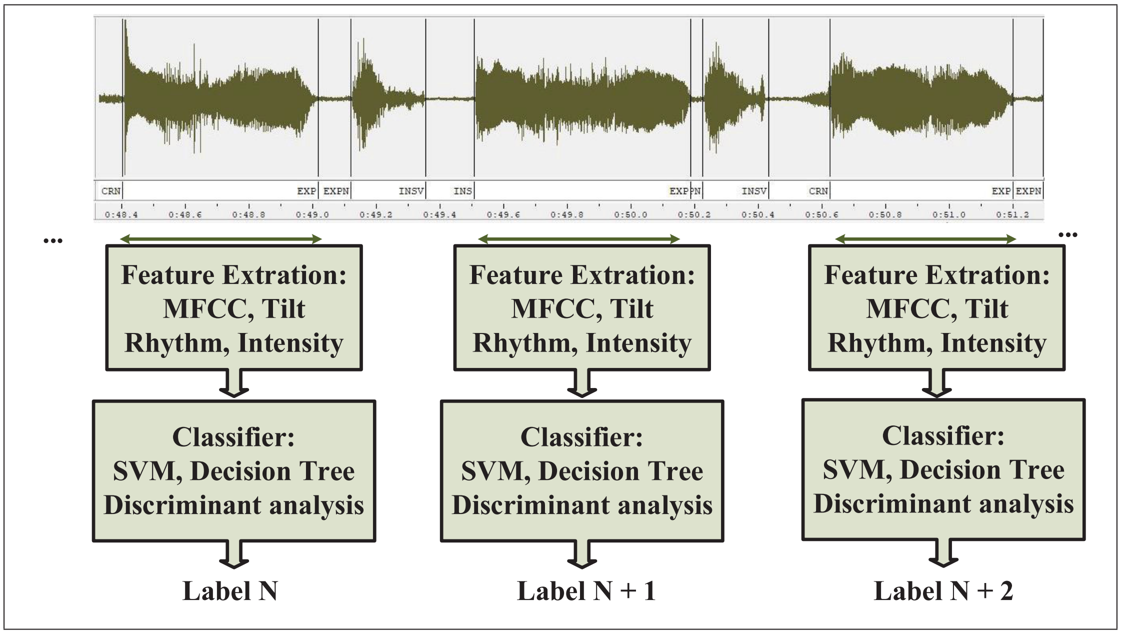

CRN
EXP
EXPN
INSV
INS
EXPN
INSV
CRN
EXP
EXPN
0:48.4
0:48.6
0:48.8
0:49.0
0:49.2
0:49.4
0:49.6
0:49.8
0:50.0
0:50.2
0:50.4
0:50.6
0:50.8
0:51.0
0:51.2
...
Feature Extration: MFCC, Tilt Rhythm, Intensity
Classifier: SVM, Decision Tree Discriminant analysis
Label N
Feature Extration: MFCC, Tilt Rhythm, Intensity
Classifier: SVM, Decision Tree Discriminant analysis
Label N + 1
Feature Extration: MFCC, Tilt Rhythm, Intensity
Classifier: SVM, Decision Tree Discriminant analysis
Label N + 2
...

Figure 3.2 Illustration of procedure of SE experiment in a portion of an infant CAS visualized using WaveSurfer software

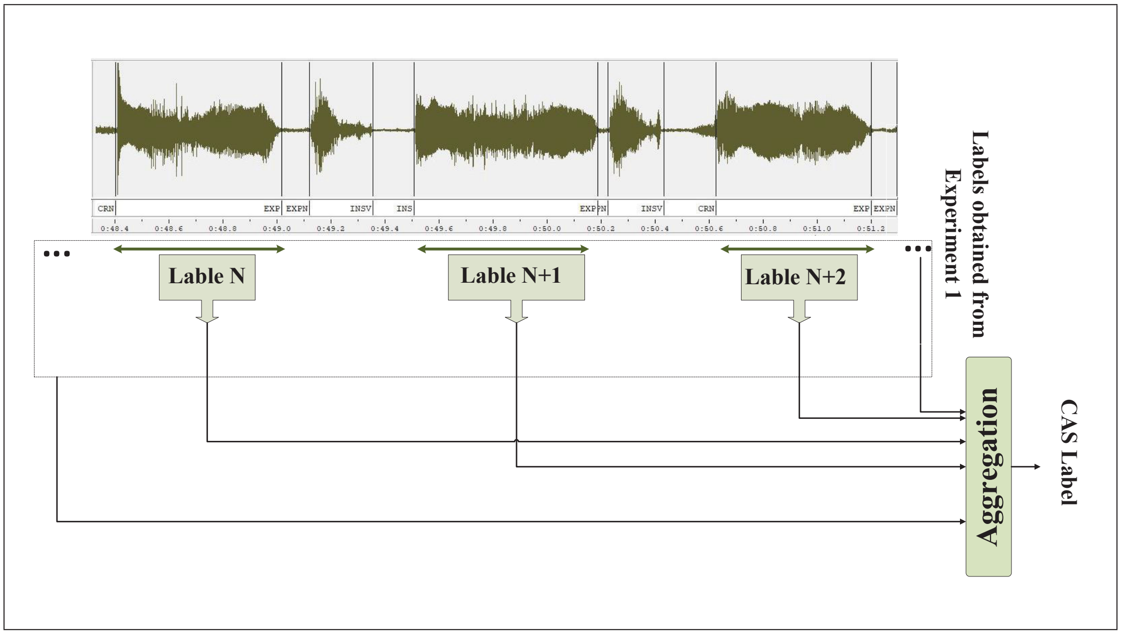


Figure 3.3 Illustration of procedure of AE experiment in a portion of an infant CAS visualized using the WaveSurfer software

to obtain the stationary audio signal, assessing the power spectrum sequence of the signal using FFT over the Mel scale and finally taking the log and the IDFT.

Mel frequency cepstrum shows the power spectrum of an audio signal using the linear cosine transform of the power spectrum logarithm at the Mel scale. The Mel scale is defined as equation 3.1.

$$M(f) = 1125\ln(1 + \frac{f}{700}) \tag{3.1}$$

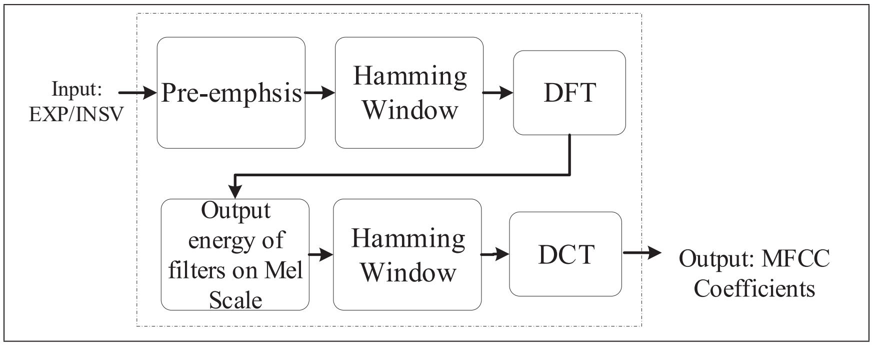

Input:
EXP/INSV
Pre-emphsis
Hamming Window
DFT
Output energy of filters on Mel Scale
Hamming Window
DCT
Output: MFCC Coefficients

Figure 3.4 The block diagram of MFCC features extraction

Where "f" is the frequency value and "$M(f)$" is the corresponding Mel value. The MFCC coefficients can be defined as the logarithmic cosine conversion of the energy obtained by applying the Mel Bank filter to the windowed signal spectrum. The steps for calculating the MFCC coefficients are shown in Figure 3.4.

The coefficients extracted from each frame contain only the static information of the frame, and this causes the effect of adjacent frames not to be considered, and due to the non-staticity of the newborns CASs, the feature vector of each frame should also reflect changes in spectral characteristics. Thus the feature vector of each frame also includes the time derivatives of the extraction coefficients. For further information on MFCC please read (Jurafsky & Martin (2014).

In our framework, we only analyzed the information less than the frequency of 4 kHz according to the result of our experiment in (Alaie *et al.* (2016)) for infant CASs. In the windowing stage, we used a hamming window with a frame size of 10 ms, with a 30% overlap between each consecutive frame. In our previous work (Alaie *et al.* (2016)), the results showed that the smaller frame length of 10 ms performs better than the frame size of 30 ms. Moreover, we set the number of filter bank channels to 24. These adjustments, that particularly suit infants CAS processing, are based on our previous experiments (Alaie *et al.* (2016); Salehian Matikolaie & Tadj (2020)).

3.3.2.3 Tilt Feature

Fundamental frequency (F_0) is defined as the harmony of the oscillation of the vocal folds (Boukydis & Lester (2012); Manfredi *et al.* (2018). The pattern of changes in F_0 repeatedly has been described to be relevant with some pathology (Boukydis & Lester (2012). The tilt feature represents changes in F_0 of the voice. The tilt features are based on the F_0 and was initially presented by (Mary (2012)) in an automatic speech recognition system, and also was successfully used in our previous study (Salehian Matikolaie & Tadj (2020)). Tilt parameters capture the changes of the F_0 using parameters called A_t and D_t. In the present study, we followed the method provided by (Mary (2012)). The parameters A_t and D_t are presented respectively by equations 3.2 and 3.3 :

$$A_t = \left(\frac{|A_r| - |A_f|}{|A_r| + |A_f|} \right) \tag{3.2}$$

$$D_t = \left(\frac{|D_r| - |D_f|}{|D_r| + |D_r|} \right) \tag{3.3}$$

Considering the contour of F_0 in a portion of CAS, A_r is the amplitude of the F_0 when the contour is rising to reach the peak of F_0, and A_f is alternatively the amplitude when the contour is declining. Correspondingly, D_f and D_r respectively are the measures of the distance of the rising and declining parts of the F_0 contour. This feature set is described in detail in (Mary (2012)).

For extracting the tilt features, the requirement is finding the accurate F_0 contour. Finding the F_0 in newborns' CAS is hindered by the high instability of the infants CAS (Manfredi *et al.* (2018)).

Among the popular software for extracting F_0, the most precise one is Praat software (Orlandi *et al.* (2017)). Thus we extracted the F_0 using Praat software. Table 3.3 shows an example of the result of F_0 extracting using Praat software. The values of A_t, D_t and the F_0 of each episode of the CAS were computed. Finally, the statistical measures of the range, mean, standard deviation, median and interquartile range of these values were put in the feature vector.

3.3.2.4 Rhythm Feature

In this study, we also investigated the usefulness of the duration feature which is a subset of rhythm feature. We calculated the duration of expiration and inspiration episodes within each CASs.

3.3.2.5 Intensity Feature

This feature was already used in the automatic identification of expiration and inspiration episodes of infant CASs (Abou-Abbas *et al.* (2017)). Intensity is the measure of the loudness of the signal. It measures the quantity of energy that the signal conveys per unit area. The intensity magnitude is measured by equation 3.4:

$$Intensity = 10log(\sum_{n=1}^{N} A^2(n)w(n)) \tag{3.4}$$

In this equation "w" and "A" respectively refer to the window function and the amplitude of the CASs. We used Praat software for precise estimation of the intensity of infant CAS. Table 3.3 shows the results of extracting this feature from a portion of CAS in our dataset. Like tilt feature extraction, the statistical measures of the range, mean, standard deviation, median and interquartile range of the values of intensity features were put in the feature vector.

Table 3.3 The evaluated values of F_0 and intensity by Praat software for a portion of CAS within the time period of 0.014 to 0.308

Time index	**F_0 value**	**Intensity value**
0.0140	0	–undefined–
0.0280	373.3105	–undefined–
0.0420	376.4588	–undefined–
0.0560	379.6858	77.751
0.0700	370.2263	77.362
0.0840	361.8400	76.333
0.0980	362.1973	75.891
0.1120	367.0559	75.978
0.1260	364.5674	76.924
0.1400	363.7566	78.619
0.1540	365.9141	79.855
0.1680	369.2621	80.186
0.1820	373.5069	79.616
0.1960	374.0275	78.628
0.2100	373.9442	78.098
0.2240	375.1161	78.809
0.2380	385.2669	80.151
0.2520	397.3219	80.028
0.2660	404.9940	78.564
0.2800	406.3105	77.089
0.2940	404.5215	75.652
0.3080	403.8883	74.393

3.3.3 Feature Reduction: Principle Component Analysis

Feature selection is used for reducing the dimensionality size of measuring space by eliminating the low effect or useless features. Principle Component Analysis (PCA) is one of the best methods for decreasing feature dimensionality linearly. It can identify key components and help the classifier to analyze a set of features that are more valuable in terms of conveying distinctive group information than just examining them all. This algorithm tries to represent the features in a way that highlights their similarities and differences. This technique defines new axes for the features and these new axes display the features. The first axis is supposed to be placed in a direction, which maximizes the data variance. In other words, in a direction in which the distribution of features is highest. Then the second axis is perpendicular to the aforementioned axis. For more information on PCA, the authors suggest reading (Ayesha, Hanif & Talib (2020)).

Besides the popular method of PCA, we experimented with the statistical metrics as a method of feature reduction. This includes the range, mean, standard deviation (Amaro-Camargo & Reyes-García (2007); Bhargava & Polzehl (2013)), median and interquartile range (Salehian Matikolaie & Tadj (2020)) for compressing the size of MFCCs vectors. In the evaluation section, we will compare the results and the cost of processing time of each method of PCA and statistical measures.

3.3.4 Classifiers

The classification approaches taken in this study are classification by a single episode called as SE experiment shown in Figure 3.2 and classification by the whole episodes in CAS called as AE experiment shown in Figure 3.3. In the SE experiment, each episode of CAS including expiration or inspiration (referred to "EXP" and "INSV" in Table 3.2 is considered a sample, and the model is trained to assign a label to it. While in the AE experiment, we used the majority voting technique to vote based on the number of the most predicted label in each CAS.

To develop a comparison we investigated the performance of 11 classifiers from three families to investigate the most credible functional one in identifying the CASs of unhealthy infants suffering from sepsis from healthy ones. In the following, we describe the three families of classifiers.

3.3.4.1 Support Vector Machine (SVM) Algorithm: Five Classifiers

Support vector machine, or SVM for short, is known as one of the best methods for classification and outlier detection. The basis of the linear SVM classifier is the linear classification of data. The SVM approach is to select the decision boundary in such a way that the minimum distance between each of the certain classes is maximized. This mechanism of selection makes classifiers' decisions in practice well tolerable to noise conditions. This method of selecting the border is based on points called support vectors (Wang (2005)). Here linear, cubic, quadratic, fine Gaussian and medium Gaussian SVM classifiers are included in this study.

3.3.4.2 Decision Trees Algorithm: Six Classifiers

This algorithm develops a set of conditions in form of tree construction to predict the class of a feature. The tree algorithm is based on minimizing the diversity at nodes. The lack of uniformity in the nodes is measurable using the criteria of impurity measure. The difference between types of the tree classifiers are due to the impurity measure, splitting method, and pruning tree nodes (Safavian & Landgrebe (1991)). In this paper, we evaluated the performance of six types of tree classifiers including simple, medium, complex, bagged, boosted, and reboosted trees.

3.3.4.3 Discriminant Analysis Algorithm: Three Classifiers

In this algorithm, the assumption is that different classes generate data based on different Gaussian distributions. In other words, every class is assumed to be a normally distributed cluster of data points. In this survey, we constructed the linear, quadratic, and subspace discriminant analysis algorithm.

After performing SE and AE experiments using the explained method, we put together the most competent classifiers for each feature set. The predicted labels by these classifiers then were fed to a majority voting block to predict the CAS class as healthy or septic. This idea is based on the assumption that the classifiers perform in a complementary way to enhance the predictive results.

3.4 Model Evaluation and Results

All the procedures of feature extraction and classification and evaluation stages were performed using Matlab. We utilized features from several domains and different classifiers with several kernels to capture the best result. For measuring each frameworks' ability to identify the CASs of infants with sepsis disease from healthy ones, we used the standard metrics in the pathology diagnostic field including specificity, recall, and F-measure (Wimalarathna, Ankmnal-Veeranna, Allan, Agrawal, Allen, Samarabandu & Ladak (2021)). The followings are the definitions of our evaluation measures:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{3.5}$$

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{3.6}$$

$$Fscore = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.7}$$

In our case, a "True Positive" would be correctly identifying the CAS of an infant with a septic pathology.

The performance of the classifiers was measured with 5-fold cross-validation. To ensure the validity of our model, we designed the distribution of CASs between the folds in a way to guarantee the independence of the folds. In other words, there are not samples of the same infants in more than one fold. Accordingly, in each iteration, the models learn on four folds (called trained folds) on the CASs of some infants and then chooses the one test fold which does not include any sample of the infants in the training folds. In each iteration, one fold becomes the test fold. We used two datasets in our research. The dataset includes the expiration and inspiration episodes of CAS. These episodes are called EXP and INSV in Table 3.2. Table 3.4 presents the number of samples in each fold for each dataset of expiration and inspiration.

Table 3.4 Number of samples in each class

	EXP Dataset		**INSV Dataset**	
	Class Healthy	Class Sepsis	Class Healthy	Class Sepsis
Fold One	507	507	140	140
Fold Two	517	517	141	141
Fold Three	524	524	139	139
Fold Four	523	523	132	132
Fold Five	453	453	109	109

3.4.1 Evaluation of MFCC Features

In this study, we decided to further investigate the MFCC features from the previous study that we presented in (Salehian Matikolaie & Tadj (2020)). We evaluated the results of the methods of dimension reduction techniques for MFCC features including statistical measures and PCA. These results are presented in Tables 3.5, 3.6 and 3.7 respectively showing the results of classification by families of discriminant analysis, decision tree and SVM models. We also compared the results of SE and AE experiments. As it was explained in the previous section, in AE experiment, a majority voting system is used to label the CAS based on the labels of its episodes resulted from SE experiment.

Regarding the Tables 3.5, 3.6 and 3.7 the AE experiment consistently outperformed the SE experiment in all evaluations, unless in sole case of using fine gaussian SVM classifier using PCA reduction method. We highlighted this result using ⊕ sign in Table 3.7. Meanwhile, the statistical measure resulted in better recognition power in all cases except in cases of using fine gaussian SVM for SE experiment, cubic SVM in SE experiment, and quadratic discriminant analysis for both AE and SE experiments. We highlighted these results using * sign in Tables 3.5 and 3.7. It is notable that these mentioned exceptional cases are related to the inspiration dataset.

In discriminant analysis family classifiers, as it shows in the Table 3.5, the best method for feature reduction for MFCC in the expiration dataset is the use of statistical measures which resulted in 81% F-score using subspace discriminant analysis classifier. However, in inspiration datasets, the best result is 83% F-score which belongs to using PCA techniques using the quadratic discriminant analysis classifier. Tables 3.6 and 3.7 illustrate the results obtained from decision tree and SVM families classifiers. In Table 3.6, for the expiration dataset and inspiration dataset the best F-score results are respectively 85.50% for the bagged tree, and 81.80% for both complex tree and medium tree classifiers. For SVM classifiers as shown in the Table 3.7, we see that the cubic SVM and medium gaussian SVM outperformed others respectively in the expiration dataset with 85.70% F-score and the inspiration dataset with 81.10% F-score.

Table 3.5 The classification results of discriminant analysis family classifiers using statistical measure and PCA method for MFCC features. Percentages refer to F-score. The results of the best frameworks are bolded. The * sign is used to indicate the results of the classification frameworks in which the PCA method resulted in a better recognition power than the statistical measure reduction method

Method		**EXP Dataset**		**INSV Dataset**	
		PCA	**Statistical Measurements**	**PCA**	**Statistical Measurements**
Linear Discriminant	SE	54.60%	65.50%	54.70%	65.20%
	AE	70.20%	77.30%	74.60%	80.85%
Quadratic Discriminant	SE	59.20%	65.10%	65.60% *	64.30%
	AE	72.40%	78.10%	**83.00**% *	79.20%
Subspace Discriminant	SE	54.50%	68.10%	54.20%	57.10%
	AE	70.40%	**81.00%**	72.80%	73.20%

Table 3.6 The classification results of decision tree family classifiers using statistical measure and PCA method for MFCC features. Percentages refer to F-score. The results of the best frameworks are bolded

Method		**EXP Dataset**		**INSV Dataset**	
		PCA	**Statistical Measurements**	**PCA**	**Statistical Measurements**
Simple Tree	SE	55.30%	65.60%	55.50%	55.60%
	AE	74.10%	85.00%	70.60%	74.60%
Medium Tree	SE	54.90%	62.80%	55.60%	60.60%
	AE	72.50%	82.30%	71.90%	**81.80**%
Complex Tree	SE	50.10%	60.40%	55.40%	60.40%
	AE	63.00%	81.80%	71.20%	**81.80**%
Bagged Tree	SE	56.60%	66.60%	62.60%	62.80%
	AE	80.10%	**85.50%**	77.30%	80.60%
Boosted Tree	SE	58.10%	67.10%	56.40%	59.30%
	AE	80.60%	81.30%	68.10%	77.20%
Ruboosted Tree	SE	55.20%	63.40%	56.90%	59.60%
	AE	74.60%	82.30%	76.20%	81.10%

3.4.2 Evaluation of Prosodic Features

Regarding the results obtained using tilt and intensity feature sets shown in Tables 3.8 and 3.9. In every case, the method of AE experiment resulted better than SE experiment. Among the

Table 3.7 The classification results of SVM family classifiers using statistical measure and PCA method for MFCC features. Percentages refer to F-score. The results of the best frameworks are bolded. The ⊕ sign is used to indicate the classification framework in which the SE experiment outperformed the AE experiment. The * sign is used to indicate the results of the classification frameworks in which the PCA method resulted in a better recognition power than the statistical measure reduction method

Method		**EXP Dataset**		**INSV Dataset**	
		PCA	**Statistical Measurements**	**PCA**	**Statistical Measurements**
Linear SVM	SE	54.20%	68.70%	56.50%	62.00%
	AE	73.60%	85.30%	74.20%	79.80%
Cubic SVM	SE	56.80%	66.30%	59.50% *	56.90%
	AE	77.30%	**85.70%**	77.10%	78.20%
Quadratic SVM	SE	55.60%	67.40%	57.00%	59.30%
	AE	77.60%	82.60%	75.70%	78.00%
Fine Gaussian SVM	SE	56.60%	64.00%	61.60% ⊕ *	56.60%
	AE	80.10%	81.50%	57.10%	63.60%
Medium Gaussian SVM	SE	53.90%	68.90%	59.60%	63.20%
	AE	76.40%	84.20%	71.30%	**81.10**%

classifiers for the tilt feature set, in the expiration dataset and inspiration dataset respectively boosted tree with 79% F-score and quadratic discriminant analysis with 83.9% F-score defeated other classifiers. In intensity feature set investigation, as shown in Table 3.9 we observed that cubic SVM is the best classifier for both expiration dataset and inspiration dataset with the F-score of 70.9% and 74.60% respectively. Table 3.10 shows the efficacy of the rhythm feature using different classifiers. This durational feature only was measured for the AE experiment as it requires longer length of CAS. In the expiration dataset and the inspiration dataset, the cubic SVM with 75.60% F-score and quadratic SVM with 77.70% F-score were the best classifiers respectively.

3.4.3 Evaluation of Feature Set Manipulation and Use of Multiple Classifiers

After acquiring the results of different classifiers using MFCC features and the prosodic features of intensity, tilt, and rhythm, we inspected the performance of two other frameworks. The first approach was to concatenate all features and see the best result obtained by which classifier. The second approach was the use majority voting technique which inputs were the results of the most capable classifiers for each feature set that outperformed in the AE experiment. Table 3.11 and 3.12, respectively show the results of mentioned frameworks for the expiration dataset and the inspiration dataset. In the majority voting framework, we only included the feature sets of MFCC, tilt, and rhythm as they consistently outperformed intensity features in both datasets.

In Table 3.11, it is shown that the feature set concatenation framework using quadratic SVM classifier outperformed the majority voting model which respectively resulted in the F-scores of 86% and 83.30%. However, in the inspiration dataset, there was not a major difference between these two methods as both methods resulted in about 82% F-score.

3.5 Discussion

In this research, we examined machine learning techniques to develop an NCDS for investigating the potential of newborns CASs for diagnosing septic infants from healthy ones. The sepsis pathology yet has not been studied while it is ranked as the 6th cause leading to the death among newborns in Canada on 2019 (Government of Canada (2020)). Several evaluations were carried out to develop a comparison between the performance of each framework. In total, four feature sets of MFCC, tilt, rhythm, and intensity were supplied to three families of classifiers including SVM, discriminant analysis, and decision tree. We also assessed the performance of the method of the concatenation of all feature sets together and the method of collecting the votes of the most accurate classifiers for each feature set, and then label the test sample using the majority voting method. Our system's input data were two datasets including episodes of expiration and inspiration of infant CASs that were separately examined.

As the results of the experiments show from Table 3.5 to 3.10, the technique of majority voting in the AE experiment enhanced the performance of the model in all cases by far, except in the only case of classification of inspiration episode dataset using the framework of MFCC and fine gaussian SVM classifier with the PCA feature reduction technique as shown in the Table 3.7

Table 3.8 The classification results of different classifiers using tilt features. Percentages refer to F-score. The results of the best frameworks are bolded

Classifier	**EXP Dataset**		**INSV Dataset**	
	SE	**AE**	**SE**	**AE**
Linear Discriminant	54.60%	67.00%	65.00%	76.00%
Quadratic Discriminant	47.00%	49.40%	66.90%	**83.90**%
Subspace Discriminant	54.90%	70.00%	58.70%	73.30%
Simple Tree	38.80%	45.80%	60.20%	69.20%
Medium Tree	53.70%	68.60%	52.40%	70.30%
Complex Tree	54.90%	74.70%	54.00%	68.70%
Bagged Tree	59.30%	78.70%	60.80%	76.50%
Boosted Tree	57.70%	**79.00**%	58.90%	74.30%
Ruboosted Tree	56.30%	74.80%	56.20%	70.20%
Linear SVM	55.50%	69.00%	61.30%	72.40%
Cubic SVM	55.90%	78.50%	60.80%	75.90%
Quadratic SVM	54.50%	74.20%	61.20%	74.10%
Fine Gaussian SVM	56.00%	75.60%	61.20%	71.60%
Medium Gaussian SVM	55.10%	70.10%	63.70%	71.70%

(highlighted by ⊕ sign). In Figure 3.5 we brought the minimum, maximum, and mean of the increase using the majority voting technique in the AE experiment in datasets of expiration and inspiration.

Table 3.9 The classification results of different classifiers using intensity features. Percentages refer to F-score. The results of the best frameworks are bolded

Classifier	**EXP Dataset**		**INSV Dataset**	
	SE	**AE**	**SE**	**AE**
Linear Discriminant	51.00%	62.90%	44.10%	59.00%
Quadratic Discriminant	45.80%	50.60%	49.30%	71.40%
Subspace Discriminant	52.40%	61.50%	48.90%	59.50%
Simple Tree	47.10%	57.10%	38.80%	60.90%
Medium Tree	50.60%	60.50%	47.60%	66.00%
Complex Tree	53.80%	68.00%	45.00%	58.30%
Bagged Tree	53.10%	65.70%	45.80%	62.20%
Boosted Tree	48.80%	60.10%	48.30%	65.20%
Ruboosted Tree	48.30%	58.20%	43.90%	61.70%
Linear SVM	50.30%	58.90%	52.30%	66.30%
Cubic SVM	58.70%	70.90%	53.20%	74.60%
Quadratic SVM	46.60%	56.70%	53.40%	69.50%
Fine Gaussian SVM	48.00%	58.70%	44.50%	58.00%
Medium Gaussian SVM	49.10%	60.20%	43.90%	52.60%

Thus the successive classification of episodes in CAS, and then use of majority voting to predict the CAS resulted quite assuring in enhancing the performance of NCDS. This idea was inspired by (Abdoli *et al.* (2019)) which was also successful in the domain of environmental sound classification.

Regarding MFCC features, we have analyzed the comparison of the use of two methods of feature reduction including PCA and statistical measures. These results are presented in Tables from 3.5 to 3.7. The results consistently show the superiority of the use of statistical measures over the PCA method in feature reduction in all families of classifiers in both datasets unless in some cases for classification of the inspiration dataset. These cases include quadratic discriminant for both experiments of SE and AE (Table 3.5, as well as cubic SVM and fine gaussian SVM in SE experiment (Table 3.7). These cases are marked using * in mentioned tables.

The importance of feature selection is based on the problem, dataset properties and their number, the interconnection condition among samples in the dataset, the desirable running time, and the

Table 3.10 The classification results of different classifiers using rhythm feature. Percentages refer to F-score. The results of the best frameworks are bolded

Classifier	EXP Dataset	INSV Dataset
	SE	SE
Linear Discriminant	41.80%	62.80%
Quadratic Discriminant	20.40%	77.00%
Subspace Discriminant	41.80%	62.80%
Simple Tree	64.40%	58.30%
Medium Tree	62.40%	64.30%
Complex Tree	70.70%	65.50%
Bagged Tree	65.50%	62.70%
Boosted Tree	62.40%	61.30%
Ruboosted Tree	61.70%	63.10%
Linear SVM	55.30%	44.40%
Cubic SVM	**75.60**%	15.50%
Quadratic SVM	55.20%	**77.70%**
Fine Gaussian SVM	37.20%	48.40%
Medium Gaussian SVM	17.30%	55.40%

Table 3.11 Best Classifiers for the expiration dataset. The results of the best frameworks are bolded

	EXP Dataset			
Feature Set	Classifier	Recall	Precision	F-score
MFCCs	Cubic SVM	**85%**	86.44%	85.70%
Tilt	Boosted Tree	78.30%	79.70%	79.00%
Intensity	Cubic SVM	71.50%	70.30%	70.90%
Rhythm	Cubic SVM	68.70%	83.90%	75.60%
All feature Concatenation	Quadratic SVM	83.90%	88.10%	**86.00**%
All feature Majority Voting	best classifiers in the AE experiment	71.80%	**99.10**%	83.30%

considered classifier scheme. Through these examinations, we found out that in all experiments for expiration dataset, and most cases for inspiration dataset the statistical measures are more powerful in terms of their discriminatory properties to represent the features that are most relevant to the classifiers experimented within this work including families of classifiers of

Table 3.12 Best Classifiers for inspiration dataset. The results of the best frameworks are bolded

	INSV Dataset			
Feature Set	**Classifier**	Recall	**Precision**	F-score
MFCCs	Quadratic Discriminant	**78.80**%	87.60%	83.00%
Tilt	Quadratic Discriminant	74.10%	**96.60**%	**83.90**%
Intensity	Cubic SVM	65.80%	82.00%	74.60%
Rhythm	Quadratic Discriminant	69.40%	86.50.00%	77.70%
All feature Concatenation	Quadratic Discriminant	76.20%	89.90%	82.80%
All feature Majority Voting	best classifiers in the AE experiment	71.70%	**96.60**%	82.30%

discriminant, decision tree, and SVM, compared to the use PCA algorithm. Moreover, as a feature reduction method, we noticed that statistical measures are a more low-cost approach in terms of computational resources compared to PCA. Table 3.13 shows the running time of feature extraction for each feature set. Thus the statistical measures not only saves the execution time, but actually, in the majority of cases, it elevates the predictive power of the model. The statistical method was applied successfully in (Amaro-Camargo & Reyes-García (2007); Bhargava & Polzehl (2013); Salehian Matikolaie & Tadj (2020)) in the domain of automatic emotion recognition in speech, and developing NCDSs for infants with deafness, asphyxia, and respiratory distress.

Regarding the prosodic features of tilt, intensity, and rhythm, the assessment of tilt and intensity features took nearly the same amount of time. However, tilt features showed better distinctive properties. The rhythm feature had the lowest computational cost. Rhythm is very simple and fast to extract, while it had better F-score results than intensity features. According to (Dietterich (2000)) an authoritative classifier has an error rate lower than the random guessing on an untrained dataset, thus the present study shows that septic infants of less than two months, cry differently than healthy ones in terms of spectral features, duration feature, the pattern of changes of the fundamental frequency and the energy of their CAS, which makes this method promising as a possible diagnostic tool.

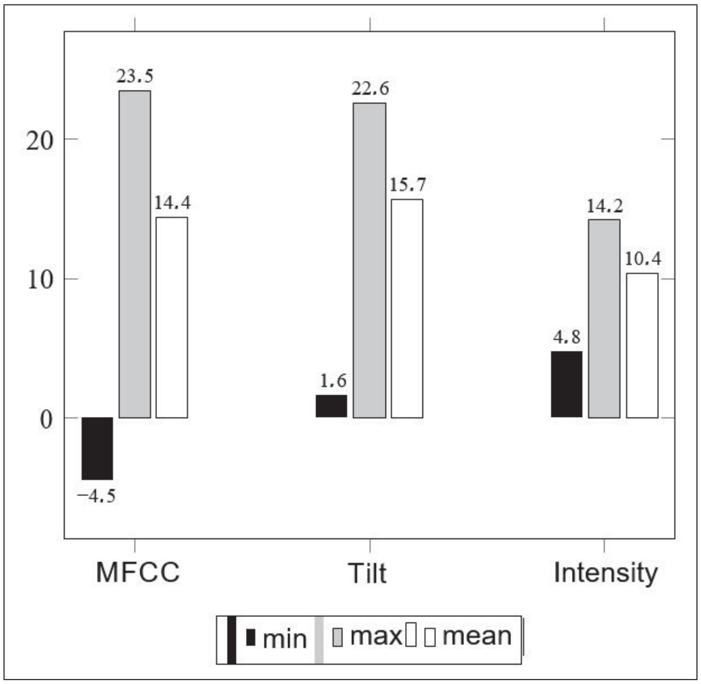


Figure 3.5 The minimum, maximum and mean of the improvement using the majority voting technique in the AE experiment in datasets of expiration and inspiration

For further analysis, we concatenated all feature sets together and fed them to each classifier. Unlike the promising results in our previous study in which we concatenated tilt, rhythm, and MFCC (Salehian Matikolaie & Tadj (2020)), the results of the concatenation of MFCC with tilt, rhythm, and intensity in both episodes were not improving in the present study. In a previous study, the control group was infants with respiratory distress. Thus the idea of feature manipulation for diagnosing septic infants from healthy infants did not reproduced the good results of training based on the individual feature set.

We also examined the idea of aggregating the results of the best classifiers for each feature set that were extracted from the same dataset and vote for the most recurred label. The intuition was to generate a framework in which the classifiers would complement their errors, thus would enhance the diagnostic power of the NCDS. Accordingly, the predicted labels achieved from the most competent classifiers for each feature set shown in Tables 3.11 and 3.12 were collected and

Table 3.13 Elapsed running time for extracting each feature set

	Feature set	**Elapsed time (min)**
1	MFCC + PCA	23.20
2	MFCC + stats	15.80
3	Tilt	10.30
4	Intensity	10.60
5	Rhythm	0.08

aggregated to predict the final result. In practice this framework could not enhance the NCDS performance and has a more computational cost, however, in the expiration dataset, it could improve the precision measure up to 99% (Table 3.11).

The unexpected performance of the multiple classifiers scheme might be explained by the fact that the integration of best classifiers was chosen globally. We generalized the model to predict for all test samples. However, in the case of noise existence around some test samples in the feature space, this scheme probably would not be able to guarantee the best prediction for those test samples. Thus we have to employ an approach that selects the outperforming classifiers locally. In every region of feature space, the competency of classifiers should be estimated based on local information. This approach is called the dynamic selection of classifier. We hope to address the shortcoming of our proposed multiple classifier scheme in the future work by experimenting the scheme of dynamic selection of classifiers, as well as the stacked classifier. The method should handle the feature sets that do not degrade the feature space or the system performance. In the future extraction phase, we also expect to examine the performance of other feature sets such as the auditory inspiration modulated feature set in the NCDS.

In our study, we made the generalization that the CASs are initiated by any reason, which in practice makes the task of diagnosing difficult as newborns cry rhythmically different for their different needs (Michelsson *et al.* (1996)). Moreover, the CASs in our dataset belong to infants from different geographical regions. Infant in one linguistic group was proven to have a similar pattern of (F_0) contour (Manfredi *et al.* (2017)). Thus the state of a more uniform database in terms of rhythmicity and melody by experience would probably help the overall performance of the NCDS. However, the motivation was to develop an NCDS to be able to make a precise decision under different situations, and be unbiased by reason of crying, the surrounding noise, and be flexible with the length of the sample.

As a final point, it is worthwhile to explain why we did not use the pervasive deep learning techniques in our study. While the use of deep learning techniques is becoming rapidly prevalent, there are yet classification problems that have the limitation of dataset shortage which massively hinders the use of such techniques (Wimalarathna *et al.* (2021)). Notably, there are fewer applications of deep learning in the infant diagnostic task based on CASs as well, due to the absence of enough CASs dataset. The number of infants and their CASs for each disease is often inadequate. Thus in case of enough number of the dataset, it is worth attempting deep learning techniques, however, there is no certainty that they work better than other classifiers for a given dataset (Fernandez-Delgado M., Cernadas E., Barro S. & Amorim D. (2014)), as the choice of a classifier is dataset-based.

3.6 Conclusion

The experiments presented here evaluates the functionality of our proposed NCDS for the unstudied disease of sepsis which is one of the most common leading to death factor in infant mortality. In our suggested NCDS, we used the well-known MFCC features and the prosodic features of tilt, rhythm, and intensity, in a configuration with different families of classifiers including SVM, decision tree, and discriminant analysis. These parameters were applied on CASs of groups of healthy and septic newborns. The obtained results show the strong contributions of the proposed features and classifiers to distinguish septic infants from healthy ones, only based on their CASs. The best accomplished F-score results are for the framework of the concatenation of all feature sets using quadratic SVM with 86%, and the framework of tilt feature set with quadratic discriminant analysis with 83.90% respectively for the two datasets of expiration and inspiration episodes of newborns' CAS. Thus we conclude that septic infants cry differently than healthy infants from the spectral, and temporal views. The scheme proposed in this study is promising to be used as a tool to assist pediatricians and address the lack of pediatricians in deprived areas.

3.7 Acknowledgments

This research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) [RGPIN-2016-05067] and has been made possible through the funding provided by the Bill and Melinda Gates Foundation [OPP1025091]. We would also like to thank Prasun Lala from the ETS library for his feedback on the manuscript.

3.8 Declaration of Competing Interest

The authors declare no conflict of interest.

CHAPTER 4

AUTOMATED NEWBORN CRY DIAGNOSTIC SYSTEM USING MACHINE LEARNING APPROACH

Fatemeh Salehianmatikolaie[1], Yasmina Kheddache, Chakib Tadj[1]

[1] Department of Electrical Engineering, École de technologie supérieure,
1100 Notre-Dame West, Montreal, Quebec, Canada H3C1K3

Paper submitted in Biomedical Signal Processing and Control, 2021.

Résumé

Les chercheurs ont constaté que les pleurs des nouveau-nés étaient un symptôme acoustique parmi les nouveau-nés en mauvaise santé. Cet article vise à développer un système non invasif de diagnostic des pleurs du nouveau-né (Newborn Cry Diagnostic System, NCDS) en utilisant les informations à différents niveaux du signal sonore des pleurs (Cry Audio Signal, CAS) des nourrissons. Le groupe de nouveau-nés en mauvaise santé dans notre expérience est composé de 34 cas cliniques.

Les techniques d'apprentissage automatique proposées comprennent l'extraction d'ensembles de caractéristiques tels que les coefficients cepstraux de fréquence de Mel (Mel-frequency cepstral coefficients, MFCC), les caractéristiques de modulation d'amplitude inspirées de l'audition (Auditory-inspired Amplitude Modulation, AAM) et les ensembles de caractéristiques de prosodie tels que l'inclinaison, l'intensité et le rythme. Les modèles d'apprentissage sont des réseaux neuronaux probabilistes et des algorithmes de machines à vecteurs de support.

Les ensembles de caractéristiques AAM et MFCC permettent d'extraire les modèles de bas niveau, tandis que l'ensemble des caractéristiques prosodiques (inclinaison, intensité et rythme) permet d'extraire les informations de haut niveau dans les CAS des nourrissons. L'ensemble de caractéristiques AAM n'a jamais été examiné dans le NCDS. L'innovation de cette étude est d'inclure l'ensemble des caractéristiques AAM dans le NCDS et de fusionner cet ensemble de caractéristiques avec les ensembles de caractéristiques MFCC et prosodie. Une autre innovation que nous reproduisons est une problématique du monde réel en incluant de nombreuses pathologies dans le groupe des personnes en mauvaise santé. Parmi les cadres proposés, la fusion de tous les ensembles de caractéristiques a amélioré les performances du système. Le meilleur résultat est celui de la fusion de l'AAM et du MFCC avec un F-measure de plus de 80%.

Les résultats de cette expérience ont révélé l'utilité des informations à différents niveaux du CAS des nouveau-nés, car elles varient entre les groupes sains et malsains. De plus, ces informations peuvent être capturées de manière non invasive par les méthodes d'apprentissage automatique dans le NCDS afin d'identifier les nouveau-nés en mauvaise santé de ceux en bonne santé.

Mots-clés: Pleurs du nourrisson, coefficient cepstral de fréquence de Mel, caractéristiques de modulation d'amplitude inspirées de l'auditoire, caractéristiques prosodiques, machine à vecteurs de support, réseaux neuronaux probabilistes, PCA, fusion de caractéristiques.

4.1 Abstract

Cry-researchers found newborns crying as an acoustic symptom among unhealthy newborns. This article aims to develop a non-invasive Newborn Cry Diagnostic System (NCDS) using the information at different levels of infants' Cry Audio Signal (CAS). The unhealthy newborns' group in our experiment consists of 34 clinical cases.

The proposed machine learning techniques include extracting feature sets of Mel Frequency Cepstral Coefficients (MFCC), Auditory-inspired Amplitude Modulation (AAM) features, and the prosody feature sets of tilt, intensity, and rhythm. The training models are probabilistic neural networks and support vector machines algorithms.

The feature sets of AAM and MFCC extract the low-level patterns, while the prosody feature set of tilt, intensity, and rhythm extracts the high-level information in infants' CAS. The AAM feature set has never been examined in NCDS. The innovation of this study is to include the AAM feature set in NCDS and fuse this feature set with the feature sets of MFCC and prosody.

Another innovation is that we reproduce the real-world problem by including many pathologies in the unhealthy group. Among proposed evaluated frameworks, the fusion of all feature sets improved the system performance. The best result relates to the fusion of AAM and MFCC with the F-measure of over 80%.

The results of this experiment revealed the usefulness of information at different levels within the newborns' CAS as they vary among healthy and unhealthy groups. Moreover, this information can be captured noninvasively by the machine learning methods in NCDS to identify unhealthy newborns frome healthy ones.

Keywords: Infant cry, Mel Frequency Cepstral Coefficient, Auditory-inspired Amplitude Modulation features, Prosodic feature, Support Vector Machine, Probabilistic Neural Networks, PCA, Feature Fusion.

4.2 Introduction

Until the first words begin, newborns use crying to communicate to attract the attention of the surrounding people. At first glance, all types of infants' Cry Audio Signals (CASs) seem the same; however, several investigations revealed the distinct cues in the infants' CASs at different states. In this reading by CAS, we refer to the sound waveform that the infant produces by pushing airflow from their lungs to the vocal track.

According to the subjective investigations, mothers and the hospital staff interacting with newborns can understand the needs of newborns only by listening to their CAS (Mukhopadhyay *et al.* (2013); Sagi (1981)). The time-domain investigation of newborns' CASs showed the different temporal morphology in different types of CASs (Wolff (1967)). The frequency-domain investigation also revealed the coarse information of the frequency spectrum properties of the newborns' CAS (Boukydis & Lester (2012)). Moreover, the cry-researchers found visual cues in the spectrographic investigation of newborns' CAS (Boukydis & Lester (2012)). Thus, these examinations provided shreds of evidence that contribute to the interpretation of the infant CAS.

Manually looking at the newborns' CASs in the domains mentioned above for exploring the cues is a tedious process. Hence automated computer-based analyses of newborns' CASs were developed. And this is where machine learning models were introduced to capture the statistics in the data.

Generally the works made in the domain of newborns' CAS analyzing includes several tasks such as automatic detection of newborns' CAS among other non-crying sounds in the environment (Kim *et al.* (2013)), automatic identification of segments in newborns' CAS such as the inhaling and exhaling segments (Abou-Abbas *et al.* (2016); Aucouturier *et al.* (2011)), identification of non-pathological reason of crying such as CAS initiated by hunger, pain, birth etc. (Abdulaziz & Ahmad (2010); Saha *et al.* (2013); Wahid *et al.* (2016)), and the identification of CASs of sick newborns form healthy ones (Alaie *et al.* (2016); Kheddache & Tadj (2013a); Lahmiri *et al.* (2021); Orozco-García & Reyes-García (2003); Salehian Matikolaie & Tadj (2020)). This research focused on the diagnostic computer-based model called Newborn Cry Diagnostic System (NCDS) in this write-up. The task of NCDS is to identity sick newborns from healthy ones based on their CAS.

To make a diagnosis classification decision, we designed an NCDS, a pipeline of three main stages of preprocessing, feature extraction, and classification. Figure 4.1 shows the diagram of the NCDS. After preparing the input CAS, the feature extraction block is to capture distinct statistics in the dataset. Then, in the classification stage, it tries to map the fed features to the specified class, delivering the predicted label for the given input sample.

The study on developing an NCDS is not as developed as other audio recognition systems due to the lack of newborns' CAS samples; however, several studies revealed the functionality of machine leanrning approaches in identifying sick newborns from healthy ones using their CASs. The studied pathologies include cleft palate (Lederman *et al.* (2002,0), hearing disorder (Hariharan *et al.* (2018,1); Orozco-García & Reyes-García (2003); Rosales-Pérez *et al.* (2015); Wahid *et al.* (2016), hyperbilirubinemia (Santiago-Sánchez *et al.* (2009)), autism (Orlandi *et al.* (2012)), asphyxia (Hariharan *et al.* (2018); Reyes-Galaviz *et al.* (2005); Rosales-Pérez *et al.*

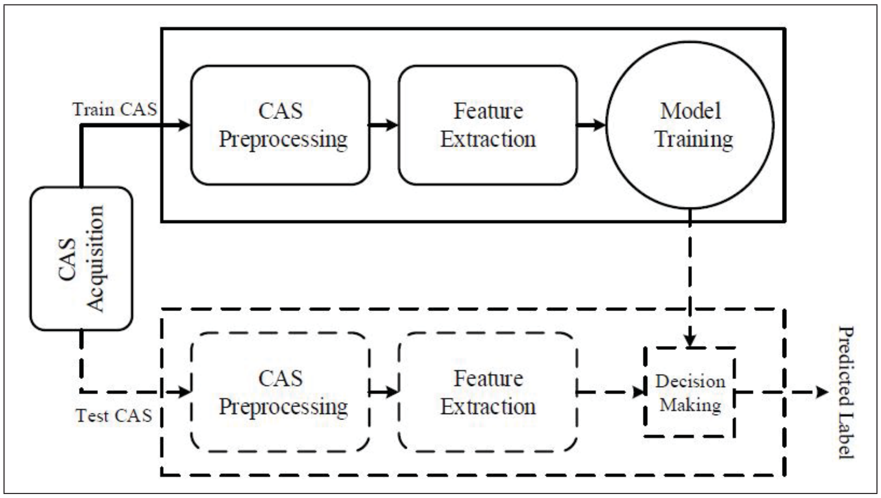


Figure 4.1 The scheme of implementing the Newborn Cry Diagnostic System (NCDS)

(2015); Sahak *et al.* (2010b,1); Santiago-Sánchez *et al.* (2009); Wahid *et al.* (2016); Zabidi *et al.* (2017)), hypothyroidism (Zabidi, Mansor, Khuan, Sahak & Rahman (2009a); Zabidi *et al.* (2009b)), respiratory distress (Lederman *et al.* (2002); Salehian Matikolaie & Tadj (2020)), and preterm newborns (Orlandi, Reyes Garcia, Bandini, Donzelli & Manfredi (2016)).

In the audio processing applications, the MFCC feature set has been the most popular and also the most practical feature set in the feature extraction phase (Salehian Matikolaie & Tadj (2020)). In the use of infant CAS for diagnostic purposes, the MFCC feature set has performed successfully in the configuration with learning algorithms such as feed-forward neural network model (García & García (2003)), Support Vector Machine (SVM) (Badreldine, Elbeheiry, Haroon, ElShehaby & Marzook (2018); Sahak *et al.* (2010a,1); Salehian Matikolaie & Tadj (2020)), multilayer perceptron (Wahid *et al.* (2016); Zabidi, Mansor, Khuan, Yassin & Sahak (2010a,1)), k-nearest neighbor (Wegener (2015)), Gaussian mixture model (Alaie *et al.* (2016)) etc.

The Linear Predictive Cepstral Coefficients (LPCC) are likewise one of the most robust and mainly used (Jurafsky & Martin (2014) tools in speech processing. The LPCC feature set in configuration with Probabilistic Neural Network (PNN) was proved to have a potent recognition accuracy (Hariharan *et al.* (2012a)). The comparison between MFCC and LPCC feature extraction techniques, however, showed better system accuracy using MFCC with the feed-forward neural network model (Orozco-García & Reyes-García (2003)), as well as hidden Markov models (Lederman *et al.* (2008)).

Another successful feature examined in NCDS is the energy entropy of wavelet packet transform. This feature set was supplied to the PNN favorably (Hariharan *et al.* (2011)). A set of the prosodic feature was as well studied in the analysis of infants' CASs. As melody concerns, results confirm the differences between the CASs of healthy infants versus sick ones. The density of melody types of plateau, rising, falling, symmetric and complex from CAS unit, as well as the features of the average of duration of CAS unit, average and standard deviation of the fundamental frequency, were determinative between full-term and preterm infant CASs (Manfredi *et al.* (2017)). The prosodic feature set including the statistical measures of fundamental frequency and the three formants of CAS was shown quite functional to detect the preterm newborns from full-term newborns in (Orlandi *et al.* (2016)).

The method of feature fusion of the prosodic feature set with the short-term feature set of MFCC was found considerably helpful to reduce the model's error rate (Ji *et al.* (2019); Salehian Matikolaie & Tadj (2020)). Our contribution to this research is twofold. First, it is of interest to study other feature sets as an addition or substitution of the MFCC feature set; thus, we examined the short-term feature set of Auditory-inspired Amplitude Modulation (AAM) for the first time in the NCDS. Our goal was to compare the functionality of the AAM feature set in NCDS compared to the most potent examined feature set of MFCC and explore the fusion potential of these feature sets. This idea was inspired by the improvement gained in the speech speaker verification system performance by fusion of AAM feature set with MFCC (Bouserhal *et al.* (2018); Sarria-Paja & Falk (2017)).

Besides the short-term feature sets, the prosodic feature sets of tilt, rhythm, and intensity are extracted in the feature extraction phase. And then, the performance of the prosodic feature set and its fusion with short-term feature sets were explored. Finally, the efficacy of the proposed feature sets was examined using the two learning algorithms of PNN and SVM as the classification phase of the NCDS.

In the present study, as mentioned earlier, we explored suggested feature sets among the healthy and unhealthy groups, including 34 pathologies. Hence our second contribution is that we investigated a large number of pathology in newborns. The majority of NCDSs was designed to identify the group of healthy infants from one group of pathology (Hariharan *et al.* (2012b); Lederman *et al.* (2008); Orlandi *et al.* (2012,1); Orozco-García & Reyes-García (2003); Sahak *et al.* (2010b,1); Salehian Matikolaie & Tadj (2020); Zabidi *et al.* (2017,0,0)), while in real-world problem the clinical state of the newborn is unspecified. Thus, we mainly do not know the potential disease that the newborn is suffering from before feeding the CAS to the NCDS.

The paper is prepared as follows: Section 4.3 is assigned to describe the collection of data sets, information of the dataset, the participants, the definition of the proposed feature sets, and explanations of the examined classifiers in this work; Section 4.4 is for reporting the results of running the SVM and PNN models using the three proposed feature sets, as well as the fusion of short-term feature sets and the fusion of all feature sets. And Section 4.5 concentrates on the discussion of the research developed, including the efficacy of each feature set, the use of joint feature sets, the classifier performance, and the computation cost of each framework.

4.3 Materials and Methods

4.3.1 Dataset Description

In this section, we described how the CASs of the newborns were collected, the dataset specifications, the dataset preprocessing procedures, and the participants in our experiments.

4.3.1.1 Dataset Acquisition

The first stage for developing an automatic recognition system is data acquisition. The medical staff of the hospitals of Al-Sahel, Al-Raee from Lebanon, and Ste-Justine from Canada collected the CASs of 769 newborns. In the recording procedure, a two-channel sound recorder with a sampling frequency of 44.1 kHz and a resolution of 16 bits was fixed at a distance between 10 and 30 cm of the newborn (Salehian Matikolaie & Tadj (2020)). The length of each record is in the range of two to three minutes. During recording, the noise of the medium, including the human talks and the medical machines' noise, was also captured. Thus our dataset resembles the real-world samples. The CASs in the database are either of healthy infants or ones affiliated with one of the diseases. There are 96 types of diseases in the database. For some diseases, the number of the infant is limited to one baby with several CASs.

Every CAS in the database has the following specifications: reason of crying, Apgar result [1], gestational age [2], birth weight, race variety, gender of the infant, and baby's age during recording.

4.3.1.2 Dataset prepration

The CASs in the database were labeled by the previous group in our lab (Abou-Abbas *et al.* (2016)). The designated labels and their descriptions are noted in Table 4.1. The labels were attached using the WaveSurfer software tool. Using the WaveSurfer software tool, it is possible to visualize the CASs' waveform and the spectrogram and give manual labeling access. The manual annotation file also is available for each recording. An example of this file for a portion of an audio CAS is shown in Figure 4.2 (Salehian Matikolaie & Tadj (2020)).

In our experiment, we used the segment of the newborns' CASs that are labeled with "EXP" as shown in Table 4.1. The significance of using "EXP" is due to the usefulness of the information in this segment as explained in our previous works (Salehian Matikolaie & Tadj (2020)).

4.3.1.3 Participated Dataset in our experiment

The development of NCDS is an age-dependent experiment (Salehian Matikolaie & Tadj (2020)). The age range of infants in our dataset is from one day to 208 days; however, in

[1] Apgar test is the very first test taken from newborns for measuring the newborn's general health state.

[2] Gestational age is in the range of 27 weeks and two days and 41 weeks and four days

Table 4.1 The descriptions of the CAS's labels in the database

Labels	Description
EXP	Voiced expiration segment during a period of cry
EXPN	Unvoiced expiration segment during a period of cry
INS	Unvoiced inspiration segment during a period of cry
INSV	Voiced inspiration segment during a period of cry
EXP2	Voiced expiration segment during a period of pseudo-cry
INS2	Voiced inspiration segment during a period of pseudo-cry
PSEUDOCRY	Any sound generated by the baby and it is not a cry
Speech	Sound of the nurse or parents talking around
Background	The kind of noise so low, it is characterized by a very low power-silence affected with little noise
BIP	The sound of the medical instruments next the baby
Noisy cry	Any sound heard with the cry: machine's beep sound, water, diaper, etc.
Noisy pseudo-cry	Any sound heard with the pseudo-cry
Noise	Like the sound caused by the mic moved by someone, the diaper, a door sound, speech + background, speech +beep.

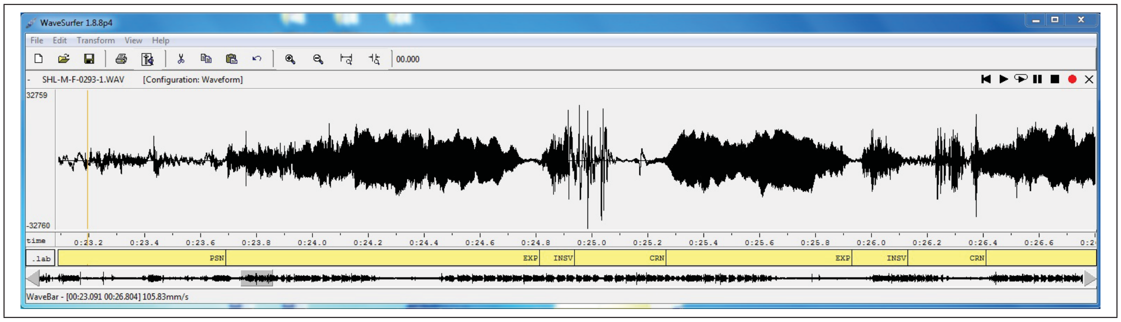


Figure 4.2 An illustration of a labeled CAS in our dataset in WaveSurfer software Medium

this experiment, similar to our previous ones (Alaie *et al.* (2016); Kheddache & Tadj (2019); Salehian Matikolaie & Tadj (2020)), , we used the samples of newborns younger than 53 days. This is because infants above this age can control their voices (Boukydis & Lester (2012)).

Table 4.2 represents the number of healthy and sick newborns in our dataset. Eighty-four newborns suffering from one of 34 pathologies are in the unhealthy group, and the healthy group contains 162 newborns. Each of these newborns in our dataset has a different number of samples.

In general, 632 CASs of full-term newborns were found eligible to cooperate in our experiment, among which 316 CASs are for healthy newborns, and 316 CASs belong to unhealthy newborns.

Table 4.2 This table shows the labels of pathology used in our experiment accompanied with the number of individual in that group. The label of healthy group is 17

Pathology Label	**Pathology Name**	**Num of Infants**	**Pathology Label**	**Pathology Name**	**Num of Infants**
17	Healthy	162	18	Hyperbilirubinemia	2
1	Ankyloglossia	3	19	Hypoglycemia	3
2	Apnea	3	20	Hypoglycemia	3
3	Asphyxia	3	21	Hypothermia	3
4	Aspiration	3	22	Intra Uterine Growth Retardation	3
5	Broncholities	3	23	Jaundice	1
6	Bronchopulmonary Dysplasia	2	24	Kidney Failure	3
7	Choanal Atresia	2	25	Meconium Aspiration Syndrome	3
8	Cleft lip and palate	1	26	Meningitis	3
9	Complex Cardio	3	27	Myelomeningocele	3
10	Cyanosis	2	28	RDS	3
11	Down Syndrome	3	29	Retraction	4
12	Duodenal Atresia	3	30	Seizure	3
13	Dyspnea	1	31	Sepsis	3
14	Fever	3	32	Tachypnea	3
15	Gastrochisis	1	33	Thrombose	3
16	Grunting	3	34	Vomit	4

4.3.2 Feature Sets Definition

In this paper, we study the suitability of various sets of features from different levels in infants' CASs, and they are also combined to arrive at a decision. These feature sets include MFCC, AAM, and prosody. MFCC and AAM are the short-term feature sets, while the prosody feature set is obtained by analyzing the more extended frame sizes of the CAS. The prosody feature set includes three subsets, including tilt features, intensity features, and rhythm features. In this section, we define these feature sets and the parameters that we computed.

4.3.2.1 MFCC Feature Set

The MFCC feature set is the most successful and well-known feature set, broadly used for speech and speaker recognition purposes. A Mel is a unit of measurement based on the sensed frequency of the human ear. The Mel scale has relatively linear frequency intervals below 1000 Hz and logarithmic intervals above 1000 Hz. An approximation of Mel for frequency can be represented as follows in the equation 4.1:

$$M(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{4.1}$$

"f" refers to the actual frequency in this equation, and "M(f)" is the perceived frequency. The main advantage of MFCC is its resistance to noise and spectrum estimation errors under different conditions.

For obtaining the MFCC coefficients, a set of applications are applied such as windowing, computing the DFT of the signal, applying the Mel filter banks and log, and finally taking the IDFT. In this work, we followed the same procedures as explained in our previous article (Salehian Matikolaie & Tadj (2020)). All the parameters were taken from our prior experience (Alaie *et al.* (2016); Salehian Matikolaie & Tadj (2020)) that were inquired to be beneficial to be used in NCDS.

4.3.2.2 AAM Feature Set

The AAM feature set has never been investigated in the infants' CAS analyzing system. The AAM feature set has been successfully tested in other acoustic recognition systems such as nonverbal human-produced audio events (Bouserhal *et al.* (2018)), speaker verification (Kinnunen *et al.* (2008)), and specifically was found to defeat the widely used feature set of MFCC (Sarria-Paja & Falk (2017)).

Figure 4.3 presents the stages for obtaining the AAM feature set. In our framework, we developed the method provided by (Sarria-Paja & Falk (2017)). Initially, the Short-Time Discrete Fourier

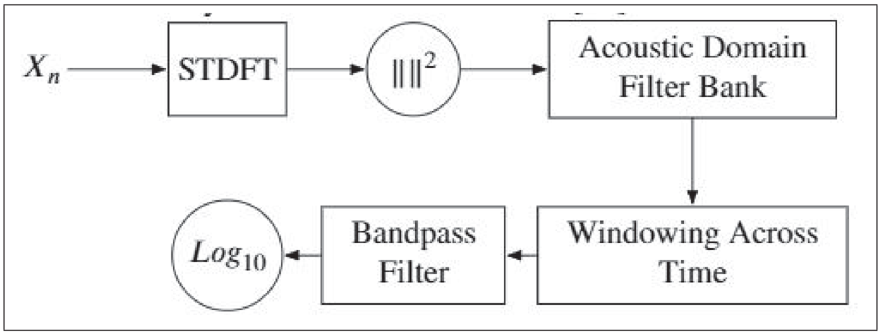


Figure 4.3 The block diagram of obtaining the AAM feature set

Transform (STDFT) is applied to the original CASs. Next, the square magnitude is calculated to reflect the human ear mechanism. These squared magnitudes of the captured acoustic frequency elements are then classified into 27 subbands based on the perceptual Mel scale.

A second transformation is then applied overtime to all subbands' magnitude CAS signals. Following that, a band-pass filter is applied. This is due to the physiological evidence of an auditory filterbank formation in the modulation domain. At the end of this process, the logarithm of the feature set is computed to reduce the massive volume of feature sets (Sarria-Paja & Falk (2017)). For more information about the details of each stage, the authors suggest reading (Sarria-Paja & Falk (2017)).

4.3.2.3 Prosody Feature Set

Humans naturally use various prosodic indications for identifying sounds (Mary (2019)). In the study of sound systems such as speaker recognition, language identification, emotion detection, and speech recognition, while the main focus was on short spectral information, several studies have shown the improvement of recognition power using prosody (Vicsi & Szaszák (2010)). Likewise, the fusion of prosody into NCDS may have the potential for a more robust system.

For the prosodical properties representation of infant CAS, we extracted the three subsets of tilt, rhythm, and intensity feature subsets. The definitions of these features are read in the following sections.

4.3.2.4 Tilt Feature Subset

Tilt features are used to explore how the fundamental frequency behaves in the CASs of healthy and unhealthy groups. Tilt features have been used favorably in tasks of automatic speaker, language, emotion and speech recognition (Mary (2019)) and newly in NCDS (Salehian Matikolaie & Tadj (2020)). In the NCDS, it was used for groups of RDS pathology versus healthy (Salehian Matikolaie & Tadj (2020)) . We explained the procedure of evaluating tilt features in our previous work (Salehian Matikolaie & Tadj (2020)). This feature set first was introduced by (Mary (2019)). The description of the tilt feature set is as follows.

The main parameters in tilt feature computations are A_t and D_t. Considering a portion of CAS, the oscillation of F_0 is captured by parameters A_t and D_t. In a which are calculated by equation 4.2 and 4.3.

$$A_t = \left(\frac{|A_r| - |A_f|}{|A_r| + |A_f|} \right) \tag{4.2}$$

$$D_t = \left(\frac{|D_r| - |D_f|}{|D_r| + |D_r|} \right) \tag{4.3}$$

The A_f and A_r capture the amplitude of the F_0 contour when F_0 contour is descending and ascending respectively. Similarly, D_f and D_r capture the length of F_0 contour descending and ascending respectively. For more details please read (Salehian Matikolaie & Tadj (2020)).

For evaluating the tilt feature subset tilt, an F_0 contour is required. However, obtaining an accurate F_0 in infant CASs is a major problem in infant CAS analysis. Among the well-known methods for F_0 contour extraction, Praat software was shown to be of the most accurate (Orlandi *et al.* (2017)).

After computing the precise F_0, the tilt feature parameters were extracted, and then the range, standard deviation, and mean of F_0 were computed. Finally, the range, mean, standard deviation, median and interquartile of each of these features' variation was measured.

4.3.2.5 Intensity Feature Subset

Intensity describes the height of the audio signal. Intensity measures the volume of energy that the audio waveform carries per unit area. The intensity measure is determined with the following equation 4.4:

$$Intensity = 10log(\sum_{n=1}^{N} A^2(n)w(n)) \tag{4.4}$$

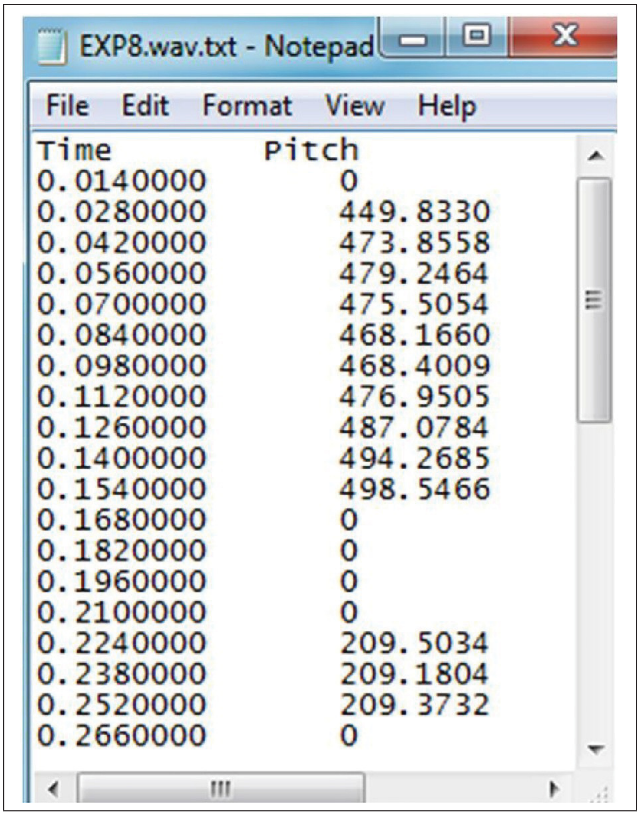
In the above equation, "w" is the window function, and "A" is the amplitude of the CASs.For extracting the intensity of the infant CASs, we used the Praat software. Then, the range, mean, standard deviation, median and interquartile of each feature's variation was measured.

Figure 4.4 shows the information of F_0 and the intensity with time index obtained for a portion of newborn CAS in our dataset.

4.3.2.6 Rhythm Feature Subset

Rhythm features capture the durational patterns of the audio. Rhythm features were quite successful in the language processing domain. In our previous work (Salehian Matikolaie & Tadj (2020)), we also found that newborns with RDS problems rhythmically cry differently from healthy ones. Accordingly, in this work, we also employed the rhythm feature subset to assess the distinctness of the behavior of the CASs of a multi-pathology group from the healthy group rhythmically. The rhythm feature subset includes the following parameters:

- **Normalized Raw Pairwise Variability Index**: The raw Pairwise Variability Index (rPVI) defines the behavior of timing contrasts between successive lengthens for speech, which is applied to syllables or segments. The rPVI's formula defines as (Salehian Matikolaie & Tadj



```
Time        Pitch
0.0140000    0
0.0280000      449.8330
0.0420000      473.8558
0.0560000      479.2464
0.0700000      475.5054
0.0840000      468.1660
0.0980000      468.4009
0.1120000      476.9505
0.1260000      487.0784
0.1400000      494.2685
0.1540000      498.5466
0.1680000    0
0.1820000    0
0.1960000    0
0.2100000    0
0.2240000      209.5034
0.2380000      209.1804
0.2520000      209.3732
0.2660000    0
```

Figure 4.4 An illustration of a labeled CAS in WaveSurfer software Medium

(2020)) in equation 4.5:

$$rPVI = [\frac{\sum_{k=1}^{M-1} |d_k - d_{k+1}|}{m-1}] \tag{4.5}$$

In which "d" is equal to the length of each "EXP", and "m" is the number of "EXP" within a CAS sample. The normalized PVI used in this work is defined as (Salehian Matikolaie & Tadj (2020))in equation 4.6:

$$nrPVI = 100 \times [\frac{\sum_{k=1}^{M-1} \left|\frac{d_k - d_{k+1}}{\frac{d_k + d_{k+1}}{2}}\right|}{m-1}] \tag{4.6}$$

- **Std**: It measures the standard deviation of the "EXP" length in each CAS (Salehian Matikolaie & Tadj (2020)).

- **Varco**: It measures the standard deviation of the "EXP" length divided by their mean length in each CAS (Salehian Matikolaie & Tadj (2020)).
- **N events**: It is the number of "EXP" that occur in each CAS (Salehian Matikolaie & Tadj (2020)).
- **Total duration**: It calculates the total length of each "EXP" in each CAS.
- **Range**: It equals the range of the "EXP" length in each CAS (Salehian Matikolaie & Tadj (2020)).
- **Mean**: It is the average of all the "EXP" length in each CAS (Salehian Matikolaie & Tadj (2020)).

4.3.3 Classification

In this study, we examined the functionality of the obtained feature sets using two learning algorithms of PNN and SVM as binary classifiers between the groups of healthy and multi-pathology.

4.3.4 Probabilistic neural network (PNN)

The efficient PNN classifier has been chosen to evaluate the proposed NCDS. It is widely used in classification problems in the medical field (Othman & Basri (2011); Sweeney, Musavi & Guidi (1994)). The PNN classifier that is ideal for real-time applications is computationally inexpensive. By means of the conjugate gradient method, it can learn new incoming training data without having to repeat the whole training process and without weight adaptation (Kheddache & Tadj (2012)).

4.3.5 Support Vector Machine (SVM)

The SVM is one of the supervised learning approaches widely used in audio classification problems. This method is effective and has proved high-grade performance compared to older machine learning methods in recent years. The principle of SVM is mainly to find the longest margin that yields the most distant between the feature points of each group. The boundary feature points are called support vectors and are then used for training (Salehian Matikolaie & Tadj (2020)). In this work, a linear kernel was employed to map the feature space.

4.4 Results

The inputs of the classifiers in our study are the vectors of the characteristics obtained at the feature extraction steps. To evaluate the efficiency of the studied system, five experiments were performed. The experiments consist of using the feature sets of:

1. AAM,
2. MFCC,
3. Prosody,
4. AAM + MFCC + Prosody, and,
5. AAM + MFCC.

These five vectors of feature characteristics were used for the training and test the two classes of infants' CASs (multi-pathology and healthy).

The test of the studied system was performed with five folds of cross-validation. The folds are independent of each other. Thus, there are no samples of the same infants in more than one fold. Fold (1) contains CAS samples from all studied pathologies, the fold (2) contains CAS samples from only a part of studied pathologies. Fold (3) contains CAS samples from 7 pathologies. Fold (4) contains CAS samples from 4 pathologies, and fold (5) contains CAS samples from only two pathologies. Table 4.3 shows the label of pathology's in each fold of our experiment.

Table 4.3 The distribution of the groups of pathology in each fold. The numbers relate to the label of pathology that was explained in Table 4.2

Fold	The pathology label
1	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,18,19, 20,21,22,23,24,25,26,27,28,29,30,31,32,33,34
2	5,10,13,15,16,18,19,22,24,25,27,30,31,34
3	5,13,18,22,27,30,34
5	18,22,27,30
5	27,30

To evaluate the performance of this system, the measures such as accuracy, specificity, sensitivity, F-score are calculated. The equation of the measures mentioned above are as following:

$$Accuray = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.7}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.8}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.9}$$

$$Fmeasure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4.10}$$

4.4.1 Evaluation of short term Feature sets

The short-term feature sets of AAM and MFCC resulted in the highest identification rates comparing to the prosodic feature set. As shown in Table 4.4, the AAM feature set consistently outperformed the MFCC feature set using the PNN model, however in converse, the MFFC feature set outperformed AAM using the SVM model.

Generally, the best results in terms of the criteria of accuracy, precision, specificity, sensitivity, and F-measure was obtained for the short-term feature set of MFCC using the SVM model.

Table 4.4 The results of feeding the feature sets of MFCC, AAM, and prosody individually to the PNN and SVM classifiers

Feature Set		**AAM**	**MFCC**	**Prosody**
Accuracy	*PNN*	**70.70%**	68.90%	52.10%
	SVM	*75.75%*	**76.50%**	61.50%
Precision	*PNN*	**67.60%**	66.60%	51.90%
	SVM	72.80%	**73.10%**	60.00%
Specificity	*PNN*	**61.70%**	61.30%	48.50%
	SVM	68.90%	**69.30%**	51.00%
Sensitivity	*PNN*	**81.60%**	76.60%	55.70%
	SVM	82.50%	**83.80%**	72.00%
Fmeasure	*PNN*	**73.70%**	71.10%	53.60%
	SVM	77.30%	**78.00%**	65.10%

Table 4.5 The results of feeding the joint feature sets of MFCC and AAM, and also the joint feature sets of all to the PNN and SVM classifiers

Feature Set		**AAM+MFCC +Prosody**	**AAM+MFCC**
Accuray	*PNN*	69.10%	**72.80%**
	SVM	*77.90%*	**78.70%**
Precision	*PNN*	65.60%	**69.60%**
	SVM	74.00%	**74.70%**
Specificity	*PNN*	59.30%	**62.50%**
	SVM	69.00%	**70.00%**
Sensitivity	*PNN*	81.50%	**83.10%**
	SVM	86.90%	**87.50%**
Fmeasure	*PNN*	72.60%	**75.60%**
	SVM	80.00%	**80.50%**

The best obtained F-measure for the proposed short-term feature sets of MFCC and AAM are 78% and 77.70% in the pipeline with the SVM training model.

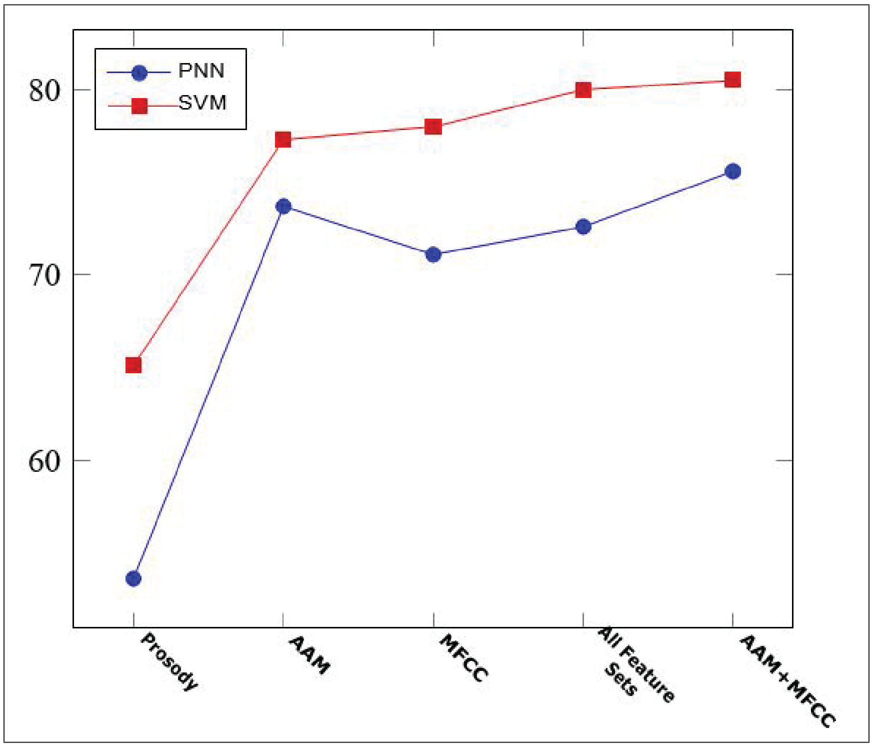


Figure 4.5 The changes in the F-measure values obtained by the SVM and PNN models for the five proposed feature sets

4.4.2 Evaluation of Prosodic Feature set

As noted in Table 4.4, prosodic feature sets were less influential than the short-term feature sets. Among the trained models, the prosodic feature set in configuration with SVM outperformed the PNN model. The F-measure using SVM and PNN were respectively 65.10% and 53.60% for the proposed prosodic feature set.

4.4.3 Evaluation of Feature sets' fusion

In this section, the results of using joint feature sets are explained. We fused the short-term feature sets and the fusion of all short-term and prosodic feature sets and supplied them to PNN and SVM.

According to Table 4.5, with the PNN classifier, the fusion of all feature sets always decreases the performance; however, with the SVM classifier, the performance consistency increased compared to only using one short term feature set. All measurements confirm the inefficacy of all feature set fusion with PNN, while the improved recognition power using the SVM model.

On the other hand, the use of joint short-term feature sets of AAM and MFCC consistently increased the classifiers' performance rates. The F-measure of using the fusion of AAM and MFCC with the SVM model is 85.50%, and with the PNN model is 75.60%.

4.5 Discussion

In this work, we intended to develop an NCDS that resembles a real-world problem. For this, we included a mixture of pathologies in the unhealthy group to be distinguished from the healthy group.

Table 4.6 The elapsed running time for extracting each feature set accompanying the number of features in each set

Feature Set Name	**Elapsed Time (sec)**	**Number of Features**
AAM	3954.50	200
MFCC	928.80	65
Prosody	572.80	38

Our other goal was to upgrade the NCDS identification power by introducing the AAM feature set for the first time in this system. Our proposed model outlined the feature extraction phase to capture MFCC, AAM, and prosodic feature sets. Our last goal was to investigate the potential of fusion of the mentioned feature sets to enhance the system performance. In the classification phase, we explored SVM and PNN that are learning algorithms from two different families. SVM and PNN were trained to solve the binary classification task of identifying unhealthy newborns from healthy ones. The models' performance is measured with 5-fold cross-validation. To do our due diligence on the reliability of our system for the real-world problem, in the evaluation part, we designed the folds to be independent of each other. Per each repetition of the 5-fold, the models are tested on the newborns' CASs that were not trained on before.

Concerning the short-term feature sets, the MFCC feature set, as expected, performed very well in NCDS. The MFCC feature set consistently was proven to be the most influential feature set in the audio processing applications by the multitude of studies (Ji *et al.* (2021)). The second examined short-term feature set is AAM, which we saw in the results section that provides essential statistical information to the models. According to several measures, as exposed in Table 4.4, the AAM feature set significantly differs among the group of healthy and unhealthy newborns and is comparatively equally powerful as the MFCC feature set. Moreover, the very assuring obtained results with the AAM feature set in NCDS is consistent with this feature set's high-grade performance in domains of acoustic recognition systems such as nonverbal human-produced audio events (Bouserhal *et al.* (2018), and speaker verification (Kinnunen *et al.* (2008)). Hence the AAM feature set can be used as a substitution of the MFCC feature set.

Concerning the prosodic feature set, the purpose was to examine whether the newborns' health condition affects the high-level information within their CAS. The results informed that the system could relatively identify unhealthy newborns; however, the obtained results are not as satisfying as the short-term feature sets of MFCC and AAM. This is why prosodic features are used as the supplemental features to aid the system performance (Salehian Matikolaie & Tadj (2020)). Furthermore, the research approved that the combination of these features would result in models' better performance based on experiences in similar systems (Adami *et al.* (2003); Vicsi & Szaszák (2010)).

Hence, the joint vector of all feature sets was fed to the classifiers. SVM using the baseline MFCC feature set resulted in the F-measure of 78%, and with all feature sets, it achieved a higher F-measure of 80%. So SVM could gain a better hyperplane that maximizes the margin between the two classes to distinguish more unhealthy infants. Conversely, PNN showed a different way of training all feature sets, as the F-measure decreased from 71.10% to 69.10% besides all the other evaluated measures as shown in Tables 4.4 and 4.5. Figure 4.5 also shows how PNN and SVM behavior changes as more feature sets are added.

We repeated our analysis for joint feature sets of AAM feature set and MFCC as they individually resulted in the most stable system decision. We found that the ideal system performance for both classifiers was obtained using joint short-term feature sets. These feature sets jointly increased the SVM and PNN performance in all examined criteria, with the F-measure of more than 80%. Therefore the optimal feature vector in our experiment is the combination of short-term feature sets of AAM and MFCC. Moreover, these evaluated measures answer our initially asked question that the AAM feature set can complement the popular MFCC feature set in the NCDS.

Moreover, in our study, we learned that the information in the short-term intervals of the newborns' CASs is more affected by the newborns' clinical state than in the longer intervals.

By experimenting with the results of two learning algorithms from different families, the goal was to pick the model which more reasonably achieves the best performance for our dataset in which SVM defeated PNN in all experiments, which are represented in Figure 4.5. To benefit from the excellent performance of the SVM classifier in future work, we hope to use it with different kernels and use multiple classifier schemes by experimenting with the scheme of dynamic selection of classifiers and the stacked classifier. Furthermore, the method should examine the feature sets that do not diminish the feature space or the system realization.

In terms of computational costs, Table 4.6 shows the elapsed time for extracting the sets of features used in our experiment. While AAM and MFCC feature sets help the models to achieve the best results, these techniques require far more execution time and mathematical resources than the prosodic feature set. Moreover, the AAM and MFCC feature sets contain more features than the prosodic feature set; thus, the system performance may improve when introducing more prosodic features. In our future work, we will address this by encompassing more prosodic feature sets such as intonation, pause patterns in the newborns CAS, etc.

4.6 Conclusion

In this research, we examined the importance of information at different levels of the newborn's CAS as cues to identify unhealthy newborns from healthy ones. For this, we extracted the standard short-term feature set of MFCC, and also, for the first time, we obtained the AAM feature set in NCDS and the prosodic feature set to capture the statistics in longer intervals of the newborns' CAS. The two classification models of SVM and PNN were trained using the feature sets mentioned above. Besides the three feature sets of MFCC, AAM, and prosody, we also explored the efficacy of feature sets fusion. Two feature vectors of the fusion of all feature sets and the fusion of the AAM and MFCC feature sets were also supplied to the classifiers. Our studied dataset includes newborns belonging to 34 groups of pathology versus healthy ones. Optimal system achievement relates to the fusion of AAM and MFCC feature sets with the F-measure of over 80% for both SVM and PNN.

This research informed us of the value of the information at a different level of newborns' CAS. The newborns affiliated with a pathology cry differently than healthy newborns, and these different patterns can statistically get captured using machine learning methods. This information at different levels is necessary to succeed in the upgrade of NCDS.

CONCLUSION AND RECOMMENDATIONS

Conclusion

In this thesis, an NCDS was introduced, which gives the possibility of examining the newborns' health condition noninvasively using the infants' CASs. The studies revealed some patterns in the newborns' CASs that warn about the menacing pathology for the infant's health, which may be clueless even in physical examinations by doctors. Hench, the NCDS is a valuable tool in saving lives and promoting the health level of newborns, specifically in developing countries where are suffering from the lack of pediatricians. The NCDS can address this issue as its installation cost is relatively low. Practical applications of the NCDS include its use for infant screening (Prathibha *et al.* (2012)), infancy education (Ruvolo & Movellan (2008)), robot nursing (Yamamoto *et al.* (2013)), and as a medical assistant for pediatricians. Our study initially focused on two critical pathologies of respiratory distresses and sepsis, ranking 11th and sixth leading causes of death in Canada. In the end, we came up with a comprehensive model encompassing 34 pathologies common among newborns.

The steps for developing an NCDS have been explained in detail in the previous chapters. In brief, the NCDS blocks are described as following:

1. The first step is preprocessing the infants' CASs proportionately with its following feature extraction technique—the preprocessing includes windowing, finding pitch contour, applying filters, etc.

2. The next step is to compact the preprocessed CASs of two healthy and sick newborns classes into the most discriminant descriptive way by feature extraction methods and then selecting the best representative groups of features.

3. The final stage is to evaluate the feature sets' efficacy by applying them to the learning models.

The literature review chapter revealed that the scientific community has vastly put effort into practicing short-term features in the NCDS. However, a few have worked on long-term features in NCDS; hence, this thesis was intended to upgrade the NCDS performance, with emphasis on using prosodic features. To the best of our knowledge, this work represents the first effort to use the prosodic features (long-term features) of tilt and rhythm and evaluate their fusion potential with the short-term features in NCDS. Moreover, this work compares the standard short-term feature set of MFCC with the AAM feature set in the NCDS and its fusion potential with the proposed features in this study. In addition, the AAM feature set is applied for the first time in the NCDS.

Various classifiers, including SVM, decision tree, discriminant analysis, and PNN, were employed in the evaluation stage. We also examined the majority voting method to enhance the classification results, which has not been reported in the literature on developing an NCDS.

Before concluding the performance of the proposed NCDS and the research accomplishments, we highlight our initial criteria for developing an NCDS. In general, concerning the tasks of the NCDS in this thesis, the ideal qualities of an NCDS is as following:

1. **Noise-independent** The NCDS should endure the environmental sounds that the CASs is recorded, such as speech and the sound of equipment. The intention was that the dataset resembles the real-world sample to increase the credibility of the performance of our proposed NCDS. In this research, we used the "EXP" and "INSV" segments of the infants' CAS in NCDS; however, these segments yet contain the surrounding noise.

2. **Generalizable**: The system should be independent of individual characteristics. It should distinguish group discriminant patterns (groups of healthy and pathology), thus requiring a reliable database containing the CASs of an adequate number of infants. As our dataset is small, it is essential to ensure the system does not learn the infants' personal vocal

characteristics. For adopting this criterion, it is necessary to perform cross-validation in the evaluation stage with this strict rule that the folds are independent. In each fold, the CAS samples of each of the infants are only used in one fold. Hence, the samples in each fold are independent of the other folds.

3. **Unbiased by region or language**: Unbiased by region or language: Investigations have found that some CAS characteristics for infants vary, depending on the geographical region or the linguistic group of the parents (Mampe *et al.* (2009); Manfredi *et al.* (2019); Wermke *et al.* (2017)). Therefore, the NCDS should be free of regional or linguistic prejudices. In order to support the principle that NCDS is unaffected by region or language, our dataset encompasses infants' CASs from parents of different linguistic groups. Our dataset is collected from hospitals in Lebanon and Canada.

4. **Robust to the reason for crying**: The NCDS should be able to identify the CAS category without a priori knowledge of the reason for crying (hunger, pain, birth, etc.). The literature indicates that the prosodic and spectral characteristics of hunger, pain, and some other types of CASs are different (Chang & Li (2016); Michelsson *et al.* (1996); Rodriguez & Caluya (2017); Vempada *et al.* (2012)). So we included all reasons for crying in our experiments.

Our research observed the abovementioned criteria in developing an NCDS with the ambitious goal of enhancing the system performance using prosodic features. Our proposed prosodic feature sets of rhythm, tilt and intensity could capture the statically distinct patterns in newborns' CASs among the group of healthy and unhealthy. Hence, the present research informed us of the value of the information at the high level of newborns' CAS. In other words, the newborns affiliated with a pathology cry differently than healthy newborns at the supra-segmental level.

However, in our study, we saw that prosodic features' distinctiveness is not as persuasive as the short-term features. The results indicated that the NCDS is much more successful with short-term features comparing to the prosodic features. This is why prosodic features in other audio processing domains are used as supplemental features to aid system performance (Adami *et al.* (2003); Vicsi & Szaszák (2010)). Our results regarding the pathology groups of infants with RDS and sepsis also approved that the more diverse the features are in the NCDS, the more powerful the NCDS is.

Another innovation of this research was to investigate the discriminating efficacy of the short-term feature set of AAM in the NCDS. We aimed to investigate the potent of the AAM feature set to be used instead of the baseline feature set of MFCC, its fusion potential, and to complement MFCC to provide more short-term information. The results indicated that the AAM feature set is relatively equivalently robust as the MFCC feature set. Regarding feature fusion, in our last experiment, the system performance is optimally enhanced with the fusion influence using the two short-term feature sets of AAM and MFCC.

Concerning our contribution in the classification stage, we compared classifying the whole CAS and classifying using majority voting for a final decision. In the majority voting experiment, the NCDS first classifies the "EXP" or "INSV" episodes individually within each CAS and then predicts the CAS label using the majority voting technique. The adapted majority voting technique significantly lowered the NCDS error rate. This result is consistent with the appliance of the majority voting technique in other audio processing domains such as automatic environmental sound classification presented by (Abdoli *et al.* (2019)).

Recommendations for future work

Some recommendations for potential future work associating improving the NCDS performance are suggested in the following.

In this thesis, techniques for extraction and representation of prosodic features of tilt and rhythm were presented. The main theoretical conclusion of this work was that the high-level information also called prosodic features or long-term features, conveys valuable information about newborns' clinical state. However, in the present work, we only introduced 38 prosodic features, which is far less than the number of short-term features. The results obtained with few prosodic features encourage utilizing more long-term features such as intonation and the pause patterns in the newborns' CAS within the NCDS.

Another realization of this work was that the results revealed that the fusion of feature sets increases the NCDS's recognition capability comparing to using the short-term feature set individually. In this work, we commonly concatenated the features from different information levels and supplied them to the model after normalization. However, another idea for upgrading the NCDS performance with various features might be possible by feeding these features in a manner more beneficial to the NCDS. This method can be using a multi-modal estimation procedure by using score fusion after obtaining each classifiers' probability score. The advantage of this method is that group features with more influence can impose more weight on the final decision. If $\{S_1, S_2, ..., S_n\}$ are the scores of each classifier and $\{W_1, W_2, ..., W_n\}$ are the allotted weights, the fusion score is obtained by the following equation:

$$S = \{W_1 S_1 + W_2 S_2 + ... + W_n S_N\} \tag{5.1}$$

In the third chapter, we discussed comparing the widely used feature selection method of PCA with the statistical measures. We also recommend the idea of using other powerful feature selection methods such as OneR, F-Ratio, BPSO, ReliefF, CNS, etc., to examine the best method for each short-term and long-term features.

A challenging issue in our work was the lack of a large dataset from different infants. Hence our methodology was to use traditional machine learning approaches to design the steps for the NCDS. Whereas recently, machine learning researchers overwhelmingly use the end-to-end deep learning methods due to their very assuring results. These new techniques can map the inputs to the related category with the most precise function without hand-designing features; however, this credence requires a large amount of dataset for the system to train itself and learn statistics within the data. Hence, to make our dataset eligible for using deep learning approaches, it is required to enlarge our dataset. Therefore, a much more extensive database is required to deliver a realistic evaluation of the classification performance.

STATEMENT OF ORIGINAL CONTRIBUTION

This research project's central initiation and motivation were to explore the hidden value of the information at the supera-segmental level of newborns' CASs for diagnostic purposes using machine learning methods. Supera-segmental information is also called prosodic features that are achieved by capturing the long-term patterns.

The original contributions to the research described in this thesis are summarized below:

1. To have an insight into the distinctness of patterns on the long-term intervals of newborns' CASs, we introduced the feature sets of tilt and rhythm for the first time in the NCDS.

2. In order to practice feature sets from different levels in a more helpful way in the NCDS, we manipulated feature sets by normalizing them with the z-score method, then concatenated them and represented them jointly to the learning algorithm.

3. In characterizing short-term information, we introduced the AAM feature set. The standard short-term feature set in NCDS has been the MFCCs. The use of the AAM feature set in the NCDS was not previously published in the literature reviews of newborns' CASs analysis.

4. In the decision-making stage, we proposed the use of multiple classifiers. We set up a multi-model framework to aggregate the prediction of the most competent classifiers for each set of features to predict the CAS group.

5. To upgrade the NCDS's performance, we introduced the majority voting techniques by predicting the episodes of "EXP" and "INSV" first and then decide the whole CAS class.

6. In terms of system credibility, we imposed rules to reduce biases in our data. Firstly, we practiced the independent folds in the cross-validation stage to make sure our NCDS could generalize across different individuals. Secondly, we masked the reason for crying to the system. We included any reason for crying (such as hunger, attention etc.) as the

long-term features can also represent the group differences in reasons for crying unrelated to the pathology. Thirdly, similar to the second practice, we masked the geographical group of newborns, as this also influences the long-term features' pattern. So we set up an NCDS solely sensitive to pathology patterns and free from the individual characteristics, nonpathology reason of crying, and regional or linguistic biases.

7. In terms of the importance of pathology that we addressed in the NCDS, we studied the CASs of septic newborns for the first time. We also restudied the CASs of newborns with RDS with our methodology. In the end, we came up with a comprehensive model encompassing 34 pathologies common among newborns.

LIST OF REFERENCES

A, V.-M., E, A.-T., Ca, R.-G., J, L.-I. & J, L.-B. (2012). Spectrographic cry analysis in newborns with profound hearing loss and perinatal high-risk newborns. *Cirugia y Cirujanos*, 80(1), 3–10. Consulted at https://europepmc.org/article/med/22472146.

Abdoli, S., Cardinal, P. & Lameiras Koerich, A. (2019). End-to-end environmental sound classification using a 1D convolutional neural network. *Expert Systems with Applications*, 136, 252–263. doi: 10.1016/j.eswa.2019.06.040.

Abdulaziz, Y. & Ahmad, S. M. S. (2010). Infant cry recognition system: A comparison of system performance based on mel frequency and linear prediction cepstral coefficients. pp. 260–263.

Abou-Abbas, L., Alaie, H. F. & Tadj, C. (2015a). Automatic detection of the expiratory and inspiratory phases in newborn cry signals. *Biomedical Signal Processing and Control*, 19, 35–43.

Abou-Abbas, L., Montazeri, L., Gargour, C. & Tadj, C. (2015b). On the use of EMD for automatic newborn cry segmentation. pp. 262–265.

Abou-Abbas, L., Tadj, C., Gargour, C. & Montazeri, L. (2016). Expiratory and inspiratory cries detection using different signals' decomposition techniques. *Journal of Voice*.

Abou-Abbas, L., Tadj, C. & Fersaie, H. A. (2017). A fully automated approach for baby cry signal segmentation and boundary detection of expiratory and inspiratory episodes. *The Journal of the Acoustical Society of America*, 142(3), 1318. doi: 10.1121/1.5001491.

Adami, A. G., Mihaescu, R., Reynolds, D. A. & Godfrey, J. J. (2003). Modeling prosodic dynamics for speaker recognition. 4, IV–788.

Alaie, H. F., Abou-Abbas, L. & Tadj, C. (2016). Cry-based infant pathology classification using GMMs. *Speech Communication*, 77, 28–52.

Amaro-Camargo, E. & Reyes-García, C. A. (2007). Applying statistical vectors of acoustic characteristics for the automatic classification of infant cry. pp. 1078–1085.

Aucouturier, J.-J., Nonaka, Y., Katahira, K. & Okanoya, K. (2011). Segmentation of expiratory and inspiratory sounds in baby cry audio recordings using hidden Markov models. *The Journal of the Acoustical Society of America*, 130(5), 2969–2977.

Aufa, B. Z., Suyanto, S. & Arifianto, A. (2020). Hyperparameter Setting of LSTM-based Language Model using Grey Wolf Optimizer. *2020 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 1–5. doi: 10.1109/ICoDSA50139.2020.9213031.

Ayesha, S., Hanif, M. K. & Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, 59, 44–58. doi: 10.1016/j.inffus.2020.01.005.

Badreldine, O. M., Elbeheiry, N. A., Haroon, A. N. M., ElShehaby, S. & Marzook, E. M. (2018). Automatic Diagnosis of Asphyxia Infant Cry Signals Using Wavelet Based Mel Frequency Cepstrum Features. *2018 14th International Computer Engineering Conference (ICENCO)*, pp. 96–100. doi: 10.1109/ICENCO.2018.8636151.

Bhargava, M. & Polzehl, T. (2013). Improving automatic emotion recognition from speech using rhythm and temporal feature. *Computer Vision and Pattern Recognition*.

Boukydis, C. Z. & Lester, B. M. (2012). *Infant crying: Theoretical and research perspectives*. Springer Science & Business Media.

Bouserhal, R. E., Chabot, P., Sarria-Paja, M., Cardinal, P. & Voix, J. (2018). Classification of nonverbal human produced audio events: A pilot study. *19th Annual Conference of the International Speech Communication, INTERSPEECH 2018, September 2, 2018 - September 6, 2018*, 2018-September(Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH), 1512–1516. doi: 10.21437/Interspeech.2018-2299.

Buder, E. H., Chorna, L. B., Oller, D. K. & Robinson, R. B. (2008). Vibratory regime classification of infant phonation. *Journal of Voice*, 22(5), 553–564.

Canada, S. (2021). Leading causes of death, infants.

Chang, C.-Y. & Li, J.-J. (2016). Application of deep learning for recognizing infant cries. pp. 1–2.

Chittora, A. & Patil, H. A. (2013). Data collection and corpus design for analysis of nonnal and pathological infant cry. pp. 1–6.

Chittora, A. & Patil, H. A. (2016). Spectral analysis of infant cries and adult speech. *International Journal of Speech Technology*, 19(4), 841–856.

CIA, C. Infant mortality rate - The World Factbook. Consulted at https://www.cia.gov/the-world-factbook/field/infant-mortality-rate/country-comparison.

Dahmani, H., Selouani, S.-A., Chetouani, M. & Doghmane, N. (2008). Prosody Modelling of Speech Aphasia: Case Study of Algerian Patients. pp. 1–6.

Dahmani, H., Selouani, S.-A., Doghmane, N., O'Shaughnessy, D. & Chetouani, M. (2014). On the relevance of using rhythmic metrics and SVM to assess dysarthric severity. *International Journal of Biometrics*, 6(3), 248–271.

DeCasper, A. J. & Spence, M. J. (1986). Prenatal maternal speech influences newborns' perception of speech sounds. *Infant behavior and Development*, 9(2), 133–150.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. *Multiple Classifier Systems*, (Lecture Notes in Computer Science), 1–15. doi: 10.1007/3-540-45014-9_1.

Fang, F.-q., Li, X.-g., Li, S.-m., Shen, D.-x. & Shao, S.-h. (2012). Based on improved PVI objective evaluation system of English sentences. pp. 1755–1759.

Fernandez-Delgado M., Cernadas E., Barro S. & Amorim D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15, 3133–3181.

Fort, A. & Manfredi, C. (1998). Acoustic analysis of newborn infant cry signals. *Medical engineering & physics*, 20(6), 432–442.

García, J. O. & García, C. A. R. (2003). Acoustic features analysis for recognition of normal and hypoacustic infant cry based on neural networks. pp. 615–622.

Golub, H. L. & Corwin, M. J. (1982). Infant cry: a clue to diagnosis. *Pediatrics*, 69(2), 197–201.

Golub, H. L. (1979). A physioacoustic model of the infant cry and its use for medical diagnosis and prognosis. *The Journal of the Acoustical Society of America*, 65(S1), S25–S26.

Government of Canada, S. C. (2020). Leading causes of death, infants. Last Modified: 2020-11-26, Consulted at https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310039501.

Grau, S. M., Robb, M. P. & Cacace, A. T. (1995). Acoustic correlates of inspiratory phonation during infant cry. *Journal of Speech, Language, and Hearing Research*, 38(2), 373–381.

Hariharan, M., Yaacob, S. & Awang, S. A. (2011). Pathological infant cry analysis using wavelet packet transform and probabilistic neural network. *Expert Systems with Applications*, 38(12), 15377–15382.

Hariharan, M., Chee, L. S. & Yaacob, S. (2012a). Analysis of infant cry through weighted linear prediction cepstral coefficients and probabilistic neural network. *Journal of medical systems*, 36(3), 1309–1315.

Hariharan, M., Sindhu, R., Vijean, V., Yazid, H., Nadarajaw, T., Yaacob, S. & Polat, K. (2018). Improved binary dragonfly optimization algorithm and wavelet packet based non-linear features for infant cry classification. *Computer Methods and Programs in Biomedicine*, 155, 39–51. doi: 10.1016/j.cmpb.2017.11.021. Publisher: Elsevier Ireland Ltd.

Hariharan, M., Sindhu, R. & Yaacob, S. (2012b). Normal and hypoacoustic infant cry signal classification using time–frequency analysis and general regression neural network. *Computer methods and programs in biomedicine*, 108(2), 559–569.

Jam, M. M. & Sadjedi, H. (2009). Identification of hearing disorder by multi-band entropy cepstrum extraction from infant's cry. pp. 1–5.

Ji, C., Xiao, X., Basodi, S. & Pan, Y. (2019). Deep Learning for Asphyxiated Infant Cry Classification Based on Acoustic Features and Weighted Prosodic Features. *2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pp. 1233–1240. doi: 10.1109/iThings/-GreenCom/CPSCom/SmartData.2019.00206.

Ji, C., Mudiyanselage, T. B., Gao, Y. & Pan, Y. (2021). A review of infant cry analysis and classification. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1), 8. doi: 10.1186/s13636-021-00197-5.

Jurafsky, D. & Martin, J. H. (2014). *Speech and language processing*. Pearson London.

Kheddache, Y. & Tadj, C. (2012). Newborn's pathological cry identification system. pp. 1024–1029.

Kheddache, Y. & Tadj, C. (2013a). Characterization of pathologic cries of newborns based on fundamental frequency estimation. *Engineering*, 5(10), 272.

Kheddache, Y. & Tadj, C. (2013b). Frequential characterization of healthy and pathologic newborns cries. *American Journal of Biomedical Engineering*, 3(6), 182–193.

Kheddache, Y. & Tadj, C. (2019). Identification of diseases in newborns using advanced acoustic features of cry signals. *Biomedical Signal Processing and Control*, 50, 35–44.

Kim, M. J., Kim, Y., Hong, S. & Kim, H. (2013). ROBUST detection of infant crying in adverse environments using weighted segmental two-dimensional linear frequency cepstral coefficients. pp. 1–4.

Kinnunen, T., Lee, K.-A. & Li, H. (2008). Dimension reduction of the modulation spectrogram for speaker verification. *Speaker and Language Recognition Workshop, Odyssey 2008, January 21, 2008 - January 24, 2008*, (Odyssey 2008: Speaker and Language Recognition Workshop).

Kumar, A. & Bhat, B. V. (1996). Epidemiology of respiratory distress of newborns. *The Indian Journal of Pediatrics*, 63(1), 93–98.

LaGasse, L. L., Neal, A. R. & Lester, B. M. (2005). Assessment of infant cry: acoustic cry analysis and parental perception. *Mental retardation and developmental disabilities research reviews*, 11(1), 83–93.

Lahmiri, S., Tadj, C., Gargour, C. & Bekiros, S. (2021). Characterization of infant healthy and pathological cry signals in cepstrum domain based on approximate entropy and correlation dimension. *Chaos, Solitons & Fractals*, 143, 110639. doi: 10.1016/j.chaos.2020.110639.

Lederman, D. & Lederman, D. (2002). *Automatic Classification of Infants ' Cry*.

Lederman, D., Cohen, A., Zmora, E., Wermke, K., Hauschildt, S. & Stellzig-Eisenhauer, A. (2002). On the use of hidden Markov models in infants' cry classification. pp. 350–352.

Lederman, D., Zmora, E., Hauschildt, S., Stellzig-Eisenhauer, A. & Wermke, K. (2008). Classification of cries of infants with cleft-palate using parallel hidden Markov models. *Medical & biological engineering & computing*, 46(10), 965–975.

Lester, B. M. & LaGasse, L. L. (2008). Crying A2 - Haith, Marshall M. In Benson, J. B. (Ed.), *Encyclopedia of Infant and Early Childhood Development* (pp. 332–343). San Diego: Academic Press.

Lester, B. M., Boukydis, C. Z., Garcia-Coll, C. T., Hole, W. & Peucker, M. (1992). Infantile colic: Acoustic cry characteristics, maternal perception of cry, and temperament. *Infant behavior and Development*, 15(1), 15–26.

Lieberman, P. (1985). The physiology of cry and speech in relation to linguistic behavior. In *Infant Crying* (pp. 29–57). Springer.

Lynip, A. W. (1951). The use of magnetic devices in the collection and analysis of the preverbal utterances of an infant. *Genetic psychology monographs*.

Mampe, B., Friederici, A. D., Christophe, A. & Wermke, K. (2009). Newborns' cry melody is shaped by their native language. *Current biology*, 19(23), 1994–1997.

Manfredi, C., Viellevoye, R., Orlandi, S., Torres-García, A., Pieraccini, G. & Reyes-García, C. A. (2019). Automated analysis of newborn cry: relationships between melodic shapes and native language. *Biomedical Signal Processing and Control*, 53, 101561.

Manfredi, C., Pieraccini, G., Viellevoye, R., Torres-Garcia, A. & Reyes-Garcia, C. (2017). Relationships between newborns-cry melody shapes and native language.

Manfredi, C., Bandini, A., Melino, D., Viellevoye, R., Kalenga, M. & Orlandi, S. (2018). Automated detection and classification of basic shapes of newborn cry melody. *Biomedical Signal Processing and Control*, 45, 174–181.

Martinez-Cañete, Y., Cano-Ortiz, S. D., Lombardía-Legrá, L., Rodríguez-Fernández, E. & Veranes-Vicet, L. (2018). Data Mining Techniques in Normal or Pathological Infant Cry. *Progress in Artificial Intelligence and Pattern Recognition*, (Lecture Notes in Computer Science), 141–148. doi: 10.1007/978-3-030-01132-1_16.

Mary, L. (2012). Automatic extraction of prosody for speaker, language and speech recognition. In *Extraction and Representation of Prosody for Speaker, Speech and Language Recognition* (pp. 19–33). Springer.

Mary, L. (2019). *Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition* (ed. 2). Springer International Publishing. doi: 10.1007/978-3-319-91171-7.

Mende, W., Wermke, K., Schindler, S., Wilzopolski, K. & Hock, S. (1990). Variability of the cry melody and the melody spectrum as indicators for certain CNS disorders. *Early Child Development and Care*, 65(1), 95–107.

Michelsson, K. & Wasz-Höckert, O. (1980). The value of cry analysis in neonatology and early infancy. *Infant communication: Cry and early speech*, 152–182.

Michelsson, K., Sirviö, P., Koivisto, M., Sovijärvi, A. & Wasz-Höckert, O. (1975). Spectrographic analysis of pain cry in neonates with cleft palate. *Neonatology*, 26(5-6), 353–358.

Michelsson, K., Christensson, K., Rothgänger, H. & Winberg, J. (1996). Crying in separated and non-separated newborns: sound spectrographic analysis. *Acta Paediatrica*, 85(4), 471–475.

Michelsson, K. (1971). Cry analyses of symptomless low birth weight neonates and of asphyxiated newborn infants. *Acta Pædiatrica*, 60(S216), 9–45.

Michelsson, K. & Michelsson, O. (1999). Phonation in the newborn, infant cry. *International journal of pediatric otorhinolaryngology*, 49, S297–S301.

Michelsson, K., Eklund, K., Leppänen, P. & Lyytinen, H. (2002). Cry characteristics of 172 healthy 1-to 7-day-old infants. *Folia Phoniatrica et Logopaedica*, 54(4), 190–200.

Moller, S. & Schonweiler, R. (1999). Analysis of infant cries for the early detection of hearing impairment. *Speech Communication*, 28(3), 175–193.

Mukhopadhyay, J., Saha, B., Majumdar, B., Majumdar, A. K., Gorain, S., Arya, B. K., Bhattacharya, S. D. & Singh, A. (2013). An evaluation of human perception for neonatal cry using a database of cry and underlying cause. pp. 64–67.

Nisar, S., Shahzad, I., Khan, M. A. & Tariq, M. (2017). Pashto spoken digits recognition using spectral and prosodic based feature extraction. pp. 74–78.

Orlandi, S., Manfredi, C., Bocchi, L. & Scattoni, M. L. (2012). Automatic newborn cry analysis: A Non-invasive tool to help autism early diagnosis. *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2953–2956. doi: 10.1109/EMBC.2012.6346583.

Orlandi, S., Reyes Garcia, C. A., Bandini, A., Donzelli, G. & Manfredi, C. (2016). Application of Pattern Recognition Techniques to the Classification of Full-Term and Preterm Infant Cry. *Journal of Voice*, 30(6), 656–663. doi: 10.1016/j.jvoice.2015.08.007.

Orlandi, S., Bandini, A., Fiaschi, F. F. & Manfredi, C. (2017). Testing software tools for newborn cry analysis using synthetic signals. *Biomedical Signal Processing and Control*, 37, 16–22.

Orozco, J. & García, C. A. R. (2003). Detecting pathologies from infant cry applying scaled conjugate gradient neural networks. pp. 349–354.

Orozco-García, J. & Reyes-García, C. A. (2003). A study on the recognition of patterns of infant cry for the identification of deafness in just born babies with neural networks. pp. 342–349.

Osmani, A., Hamidi, M. & Chibani, A. (2017). Machine Learning Approach for Infant Cry Interpretation. pp. 182–186.

Othman, M. F. & Basri, M. A. M. (2011). Probabilistic Neural Network for Brain Tumor Classification. *Modelling and Simulation 2011 Second International Conference on Intelligent Systems*, pp. 136–138. doi: 10.1109/ISMS.2011.32.

Pattnaik, P. & Dash, S. (2012). A Study on Prosody Analysis. *International Journal Of Computational Engineering Research*, 2(5). Consulted at /paper/A-Study-on-Prosody-Analysis-Pattnaik-Dash/543b5e0113e073953c2c1a65eb66e16cbae7e2a3.

Prathibha, A. M., Putta, R., Srinivas, H. & Satish, S. B. (2012). An eclectic approach for detection of infant cry and wireless monitoring of swinging device as an alternative warning system for physically impaired and better neonatal growth. *World Journal of Science and Technology*, 2(5), 62–65.

Ramus, F., Nespor, M. & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73(3), 265–292.

Rao, K. S., Reddy, V. R. & Maity, S. (2015). *Language Identification Using Spectral and Prosodic Features*. Springer.

Reyes-Galaviz, O. F. & Reyes-Garcia, C. A. (2004). A system for the processing of infant cry to recognize pathologies in recently born babies with neural networks.

Reyes-Galaviz, O. F., Verduzco, A., Arch-Tirado, E. & Reyes-García, C. A. (2005). Analysis of an infant cry recognizer for the early identification of pathologies. In *Nonlinear Speech Modeling and Applications* (pp. 404–409). Springer.

Reyes-Galaviz, O. F., Cano-Ortiz, S. D. & Reyes-García, C. A. (2008). Evolutionary-neural system to classify infant cry units for pathologies identification in recently born babies. pp. 330–335.

Robb, M. P. & Goberman, A. M. (1997). Application of an acoustic cry template to evaluate at-risk newborns: Preliminary findings. *Neonatology*, 71(2), 131–136.

Rodriguez, R. L. & Caluya, S. S. (2017). Waah: Infants Cry Classification of Physiological State Based on Audio Features. pp. 7–10.

Rosales-Pérez, A., Reyes-García, C. A., Gonzalez, J. A., Reyes-Galaviz, O. F., Escalante, H. J. & Orlandi, S. (2015). Classifying infant cry patterns by the Genetic Selection of a Fuzzy Model. *Biomedical Signal Processing and Control*, 17, 38–46. doi: 10.1016/j.bspc.2014.10.002.

Rutledge, J. C. (1995). Fundamentals of Speech Recognition, by Lawrence Rabiner and Bing-Hwang Juang. *ANNALS OF BIOMEDICAL ENGINEERING*, 23, 526–526.

Ruvolo, P. & Movellan, J. (2008). Automatic cry detection in early childhood education settings. pp. 204–208.

Ruíz, M. A., Altamirano, L. C., Reyes, C. A. & Herrera, O. (2010). Automatic identification of qualitatives characteristics in infant cry. pp. 442–447.

Safavian, S. R. & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3), 660–674. doi: 10.1109/21.97458. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics.

Sagi, A. (1981). Mothers' and non-mothers' identification of infant cries. *Infant behavior and Development*, 4, 37–40.

Saha, B., Purkait, P. K., Mukherjee, J., Majumdar, A. K., Majumdar, B. & Singh, A. K. (2013). An embedded system for automatic classification of neonatal cry. pp. 248–251.

Sahak, R., Mansor, W., Lee, Y. K., Yassin, A. I. M. & Zabidi, A. (2010a). Performance of combined support vector machine and principal component analysis in recognizing infant cry with asphyxia. pp. 6292–6295.

Sahak, R., Mansor, W., Lee, Y. K., Yassin, A. M. & Zabidi, A. (2010b). Orthogonal least square based support vector machine for the classification of infant cry with asphyxia. 3, 986–990.

Sahak, R., Mansor, W., Khuan, L. Y., Zabidi, A. & Yassin, A. I. M. (2012). Detection of asphyxia from infant cry using support vector machine and multilayer perceptron integrated with Orthogonal Least Square. pp. 906–909.

Salehian Matikolaie, F. & Tadj, C. (2020). On the use of long-term features in a newborn cry diagnostic system. *Biomedical Signal Processing and Control*, 59, 101889. doi: 10.1016/j.bspc.2020.101889.

Santiago-Sánchez, K., Reyes-García, C. A. & Gómez-Gil, P. (2009). Type-2 fuzzy sets applied to pattern matching for the classification of cries of infants under neurological risk. pp. 201–210.

Sarria-Paja, M. & Falk, T. H. (2017). Fusion of auditory inspired amplitude modulation spectrum and cepstral features for whispered and normal speech speaker verification. *Computer Speech & Language*, 45, 437–456. doi: 10.1016/j.csl.2017.04.004.

Selouani, S.-A., Dahmani, H., Amami, R. & Hamam, H. (2012). Using speech rhythm knowledge to improve dysarthric speech recognition. *International Journal of Speech Technology*, 15(1), 57–64.

Shrawankar, U. & Thakare, V. M. (2013). Techniques for feature extraction in speech recognition system: A comparative study. *arXiv preprint arXiv:1305.1145*.

Shriberg, E. & Stolcke, A. (2004). Prosody modeling for automatic speech recognition and understanding. In *Mathematical Foundations of Speech and Language Processing* (pp. 105–114). Springer.

Sirviö, P. & Michelsson, K. (1976). Sound-spectrographic cry analysis of normal and abnormal newborn infants. *Folia Phoniatrica et Logopaedica*, 28(3), 161–173.

Soltis, J. (2004). The signal functions of early infant crying. *Behav Brain Sci*, 27(4), 443–58; discussion 459–90.

Sweeney, W. P., Musavi, M. T. & Guidi, J. N. (1994). Classification of chromosomes using a probabilistic neural network. *Cytometry*, 16(1), 17–24. doi: 10.1002/cyto.990160104. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cyto.990160104.

Varallyay, G. J., Benyó, Z., Illényi, A., Farkas, Z. & Kovács, L. (2004). Acoustic analysis of the infant cry: classical and new methods. 1, 313–316.

Varallyay, G. & Benyó, Z. (2007). Melody shape-A suggested novel attribute for the biomedical analysis of the infant cry. pp. 4119–4122.

Varallyay Jr, G. (2006). Future prospects of the application of the infant cry in the medicine. *Periodica Polytechnica, Electrical Engineering*, 50(1), 47–62.

Vempada, R. R., Kumar, B. S. A. & Rao, K. S. (2012). Characterization of infant cries using spectral and prosodic features. pp. 1–5.

Verduzco-Mendoza, A., Arch-Tirado, E., García, C. A. R., Ibarra, J. L. & Bonilla, J. L. (2009). Qualitative and quantitative crying analysis of new born babies delivered under high risk gestation. In *Multimodal signals: cognitive and algorithmic issues* (pp. 320–327). Springer.

Vicsi, K. & Szaszák, G. (2010). Using prosody to improve automatic speech recognition. *Speech Communication*, 52(5), 413–426.

Wahid, N. S. A., Saad, P. & Hariharan, M. (2016). Automatic Infant Cry Classification Using Radial Basis Function Network. *Journal of Advanced Research in Applied Sciences and Engineering Technology*.

Wang, L. (Ed.). (2005). *Support Vector Machines: Theory and Applications*. Berlin Heidelberg: Springer-Verlag. doi: 10.1007/b95439.

Wasz-Hockert, O. (1968a). The infant cry: A spectrographic and auditory analysis. *Clinics in Development Medicine*, 29.

Wasz-Hockert, O. (1968b). A spectrographic and auditory analysis. *Clin. Develop. Med.*, 29.

Wasz-Höckert, O., Partanen, T. J., Vuorenkoski, V., Michelsson, K. & Valanne, E. (1964). The identification of some specific meanings in infant vocalization. *Experientia*, 20(3), 154–154. doi: 10.1007/BF02150709.

Wasz-Höckert, O., Michelsson, K. & Lind, J. (1985). Twenty-five years of Scandinavian cry research. In *Infant crying* (pp. 83–104). Springer.

Wegener. (2015). Comparison of Supervised-learning Models for Infant Cry Classification / Vergleich von Klassifikationsmodellen zur Säuglingsschreianalyse. *International Journal of Health Professions*, 2(1), 4–15. doi: 10.1515/ijhp-2015-0005.

Wermke, K., Mende, W., Manfredi, C. & Bruscaglioni, P. (2002). Developmental aspects of infant's cry melody and formants. *Medical engineering & physics*, 24(7), 501–514.

Wermke, K. & Mende, W. (2009). Musical elements in human infants' cries: in the beginning is the melody. *Musicae Scientiae*, 13(2_suppl), 151–175.

Wermke, K., Birr, M., Voelter, C., Shehata-Dieler, W., Jurkutat, A., Wermke, P. & Stellzig-Eisenhauer, A. (2011). Cry melody in 2-month-old infants with and without clefts. *The Cleft Palate-Craniofacial Journal*, 48(3), 321–330.

Wermke, K., Ruan, Y., Feng, Y., Dobnig, D., Stephan, S., Wermke, P., Ma, L., Chang, H., Liu, Y. & Hesse, V. (2017). Fundamental frequency variation in crying of Mandarin and German neonates. *Journal of Voice*, 31(2), 255. e25–255. e30.

Wimalarathna, H., Ankmnal-Veeranna, S., Allan, C., Agrawal, S. K., Allen, P., Samarabandu, J. & Ladak, H. M. (2021). Comparison of machine learning models to classify Auditory Brainstem Responses recorded from children with Auditory Processing Disorder. *Computer Methods and Programs in Biomedicine*, 200, 105942. doi: 10.1016/j.cmpb.2021.105942.

Wolff, P. H. (1967). The role of biological rhythms in early psychological development. *Bulletin of the Menninger Clinic*, 31(4), 197.

Wolff, P. H. (1969). The natural history of crying and other vocalizations in early infancy. *Determinants of infant behavior*, 81–111.

Yamamoto, S., Yoshitomi, Y., Tabuse, M., Kushida, K. & Asada, T. (2013). Recognition of a baby's emotional cry towards robotics baby caregiver. *International Journal of Advanced Robotic Systems*, 10(2), 86.

Zabidi, A., Mansor, W., Khuan, L. Y., Yassin, I. M. & Sahak, R. (2010a). The effect of f-ratio in the classification of asphyxiated infant cries using multilayer perceptron neural network. pp. 126–129.

Zabidi, A., Mansor, W., Khuan, L. Y., Yassin, I. M. & Sahak, R. (2011). Binary particle swarm optimization and F-ratio for selection of features in the recognition of asphyxiated infant cry. pp. 61–65.

Zabidi, A., Mansor, W. & Lee, K. Y. (2017). Optimal Feature Selection Technique for Mel Frequency Cepstral Coefficient Feature Extraction in Classifying Infant Cry with Asphyxia. *Indonesian Journal of Electrical Engineering and Computer Science*, 6(3), 646–655.

Zabidi, A., Mansor, W., Khuan, L. Y., Sahak, R. & Rahman, F. (2009a). Mel-frequency cepstrum coefficient analysis of infant cry with hypothyroidism. *2009 5th International Colloquium on Signal Processing Its Applications*, pp. 204–208. doi: 10.1109/CSPA.2009.5069217.

Zabidi, A., Mansor, W., Khuan, L. Y., Yassin, I. M. & Sahak, R. (2009b). Classification of infant cries with hypothyroidism using multilayer perceptron neural network. pp. 246–251.

Zabidi, A., Khuan, L. Y., Mansor, W., Yassin, I. M. & Sahak, R. (2010b). Detection of infant hypothyroidism with mel frequency cepstrum analysis and multi-layer perceptron classification. *2010 6th International Colloquium on Signal Processing its Applications*, pp. 1–5. doi: 10.1109/CSPA.2010.5545331.

Zeskind, P. S. & Lester, B. M. (1978). Acoustic features and auditory perceptions of the cries of newborns with prenatal and perinatal complications. *Child development*, 580–589.

Zeskind, P. S., Parker-Price, S. & Barr, R. G. (1993). Rhythmic organization of the sound of infant crying. *Developmental psychobiology*, 26(6), 321–333.