Biomedical Signal Processing and Control 73 (2022) 103434

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Automated newborn cry diagnostic system using machine learning approach

Fatemeh Salehian Matikolaie *, Yasmina Kheddache, Chakib Tadj

Department of Electrical Engineering, École de Technologie Supérieure, Université du Québec, Montréal, QC H3C 1K3, Canada

ARTICLE INFO

Keywords:
Infant cry
Mel frequency cepstral coefficient
Auditory-inspired amplitude modulation features
Prosodic feature
Support vector machine
Probabilistic neural networks
PCA
Feature fusion

ABSTRACT

Researchers have found that crying is an acoustic symptom among unhealthy newborns. This study aims to develop a non-invasive newborn cry diagnostic system (NCDS) using information at different levels of the cry audio signal (CAS) of infants. The unhealthy newborn group in our experiment consists of 34 clinical cases. The proposed machine learning (ML) techniques include the extraction of feature sets of Mel frequency cepstral coefficients (MFCC), auditory-inspired amplitude modulation (AAM) features, and a prosody feature set of tilt, intensity, and rhythm features. The training models are probabilistic neural networks and support vector machine algorithms. The feature sets of AAM and MFCC extract low-level patterns, whereas the prosody feature set of tilt, intensity, and rhythm extracts high-level information in an infant CAS. The AAM feature set in the NCDS has never yet been examined. As an innovative aspect of this study, the AAM feature set is included in the NCDS, and this feature set is fused with the feature sets of MFCC and prosody. As another innovation, we reproduce real-world problems by including many pathologies in the unhealthy group. Among the evaluated frameworks proposed, the fusion of all feature sets improves the system performance. The best result is obtained with the fusion of AAM and MFCC with an F-measure of over 80%. The results of this experiment reveal the usefulness of information at different levels within the CASs of newborns, which vary among healthy and unhealthy groups. Moreover, to identify unhealthy newborns, this information can be captured noninvasively by applying ML methods to the NCDS.

1. Introduction

Until they speak their first words, newborns use crying to attract the attention of people around them. At first glance, all types of cry audio signals (CASs) of infants seem to be the same; however, several investigations have revealed distinct cues in their CASs under different states. By CAS, we refer to the sound waveform produced by the infant when pushing airflow from the lungs to the vocal track.

According to subjective investigations, mothers and hospital staff interacting with newborns can understand the needs of newborns by simply listening to their CAS [1,2]. The time-domain investigation of CASs in newborns has shown different temporal morphologies in different types of CASs [3]. The frequency-domain investigation also revealed the coarse information of the frequency spectrum properties of the CAS in newborns [4]. Moreover, researchers found visual cues in the spectrographic investigation of such CASs [4]. Thus, these examinations provided evidence contributing to the interpretation of the CAS in infants.

Manually looking at the CASs of infants in the domains mentioned above for exploring potential cues is a tedious process. Hence, automated computer-based analyses of such CASs have been developed. Machine learning (ML) models were introduced to capture the statistics within the data.

In general, the studies conducted in the domain of newborn CAS analysis include several tasks such as the automatic detection of infant CASs among other non-CAS sounds within the environment [5,6], the automatic identification of segments in a newborn CAS such as the inhaling and exhaling segments [7,8], and the identification of a non-pathological reason for crying such as a CAS initiated by hunger, pain, or birth, among other factors [9–11]. Other tasks include the identification of CASs of sick newborns from healthy newborns [12–17]. This research focuses on a diagnostic computer-based model called a newborn cry diagnostic system (NCDS). The task of NCDS is to identify sick newborns from healthy infants based on their CAS.

To make a diagnosis classification, we designed an NCDS, which consists of three main stages: preprocessing, feature extraction, and

* Corresponding author.
E-mail address: fatemeh.salehianmatikolaie.1@ens.etsmtl.ca (F. Salehian Matikolaie).

https://doi.org/10.1016/j.bspc.2021.103434
Received 19 September 2021; Received in revised form 9 November 2021; Accepted 28 November 2021
Available online 18 December 2021

classification. Fig. 1 shows a diagram of the NCDS. After preparing the input CASs, the feature extraction block captures distinct statistics that define the group characteristics of the CASs of healthy and unhealthy infants in the dataset. The classification stage then maps the features to the specified class and delivers the predicted label for the given input CAS.

The NCDS is not as developed as other audio recognition systems owing to a lack of CAS samples of newborns; however, several studies have revealed the functionality of ML approaches in identifying sick newborns from healthy infants based on their CASs. The pathologies studied include cleft palates [18,19], hearing disorders [20,21,12,9,22,23], hyperbilirubinemia [24], autism [25], asphyxia [21,9,24,26–29,22,23], hypothyroidism [30,31], respiratory distress syndrome (RDS) [18,14], and preterm newborns [32].

In audio processing applications, the feature set of Mel frequency cepstral coefficients (MFCCs) is the most popular and practical feature set in the feature extraction phase [14]. In the use of infant CASs for diagnostic purposes, the MFCC feature set has been successfully applied in a configuration with learning algorithms such as a feedforward neural network (FFNN) model [33], support vector machine (SVM) [34,27,35,14], multilayer perceptron [36,37,9], k-nearest neighbor [38], and Gaussian mixture model [13].

Linear predictive cepstral coefficients (LPCCs) are among the most robust and commonly used tools in speech processing [39]. The LPCC feature set in configuration with a probabilistic neural network (PNN) was proven to have a potent recognition accuracy [40]. The comparison between MFCC and LPCC feature extraction techniques, however, showed a better system efficiency than when using MFCC with the FFNN model [12], or with hidden Markov models [19]. Another successful feature examined in the NCDS is the energy entropy of the wavelet packet transform. This feature set is supplied to the PNN [41].

A set of prosodic features was also studied in the analysis of infant CASs. In terms of melody, the results confirm the differences between the CASs of healthy and sick infants. The density of the melody types of plateau, rising, falling, and complexity form the CAS unit, as well as the features of the average duration of the CAS unit, i.e., the average and standard deviation of the fundamental frequency (F_0), were determinative between full-term and pre-term infant CASs [42]. The prosodic feature set including the statistical measures of F_0 and the three formants of CAS was shown in [32] to be quite functional for detecting preterm newborns from full-term newborns.

The method of feature fusion of the prosodic feature set with the short-term feature set of MFCC was found to be considerably helpful in reducing the error rate of the model [14,43].

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 the 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 speech speaker verification system performance through the fusion of the AAM and MFCC feature sets [44,45].

In addition to the short-term feature sets, the prosodic feature sets of tilt, rhythm, and intensity were extracted during the feature extraction phase. 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 PNN and SVM learning algorithms during the classification phase of the NCDS.

In the present study, we explored the suggested feature sets among the healthy and unhealthy groups, including 34 pathologies. Hence, our second contribution is that we investigated a large number of pathologies in the unhealthy group. The majority of NCDSs were designed to identify the group of healthy infants from one group of pathology [12,19,20,25,27–32,14], whereas in real-world problems, the clinical state of the newborn is unspecified. Thus, we mainly do not know the potential disease that a newborn is suffering from before feeding the CAS to the NCDS.

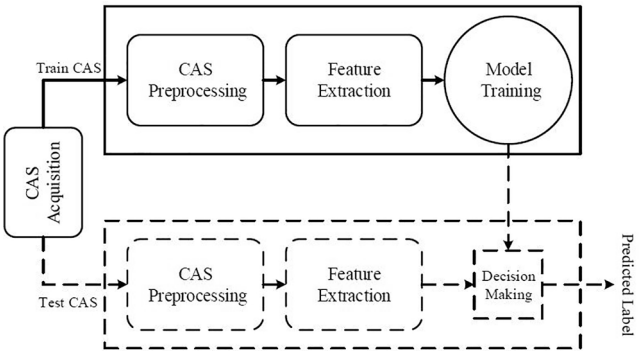

Fig. 1. Scheme of implementing the NCDS.

The paper is organized as follows: Section 2 describes the collection of datasets, information of the datasets, the participants, definition of the proposed feature sets, and descriptions of the classifiers examined in this study; Section 3 reports 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. Section 4 focuses 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.

2. Materials and methods

2.1. Dataset description

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

2.1.1. Dataset acquisition

The first stage for developing an automatic recognition system is data acquisition. The medical staff of the hospitals of Al-Sahel and Al-Raee in Lebanon and Ste-Justine in Canada collected the CASs of 769 newborns. During 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 of between 10 and 30 cm from the newborn [14]. The length of each record was within the range of 2–3 min. During the recording, environmental noises, including human speech and noises from the medical machinery, were also captured. Thus, our dataset resembles that of real-world samples. The CASs in the database are either healthy infants or those affiliated with one of the diseases. There were 96 types

Table 1
Descriptions of the CAS labels of infants in the database.

Labels	Description
EXP	Voiced expiration segment during a period of the CAS
EXPN	Unvoiced expiration segment during a period of the CAS
INS	Unvoiced inspiration segment during a period of the CAS
INSV	Voiced inspiration segment during a period of the 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 that is not a CAS
Speech	Sound of the nurse or parents talking nearby
BIP	The sound of the medical instruments next to the baby
Noisy CAS	Any sound heard along witha CAS: beep sounds of the machine, water, diapers, etc.
Noisy pseudo-CAS	Any sound heard along with a pseudo-CAS
Noise	Similar to sounds caused by a mic being moved, diapers, door sounds, speech + background, speech + beeps.

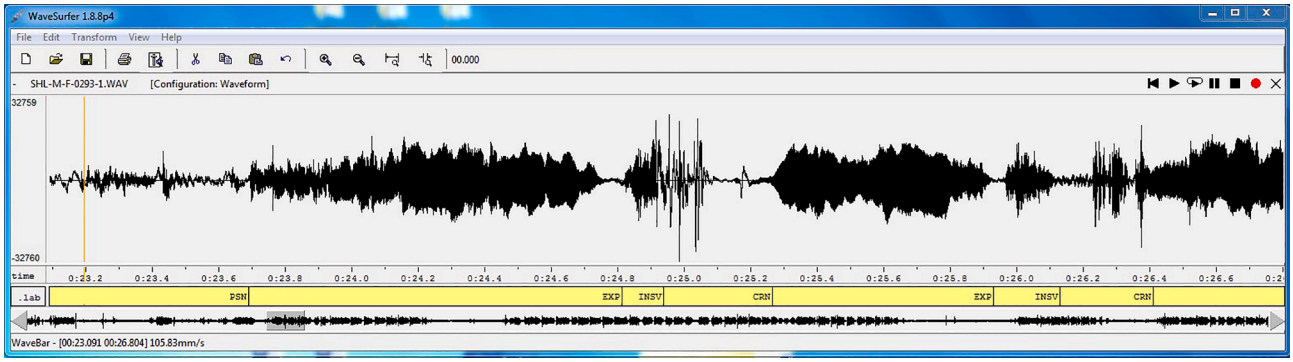

Fig. 2. Illustration of a labeled CAS in our dataset in WaveSurfer software.

of disease in the database. For certain pathologies, the number of infants is limited to one infant with several CASs.

Every CAS in the database has the following attributes: the reason for crying, the Apgar score [1], gestational age [2] birth weight, race, gender, and age during recording.

2.1.2. Dataset preparation

The CASs of infants in the database were labeled by the previous group in our lab [7]. The designated labels and their descriptions are presented in Table 1. The labels were attached using WaveSurfer software. Using the WaveSurfer software tool, it is possible to visualize the waveform of the CASs and the spectrogram and provide manual labeling access. A manual annotation file is also available for each recording. An example of this file for a portion of the CAS is shown in Fig. 2 [14].

In our experiment, we used the segments of the CASs of newborns labeled as 'EXP' in Table 1. The significance of using 'EXP' is the usefulness of the information in this segment, as described in our previous study [14].

2.1.3. Dataset used in our experiment

The development of the NCDS is age-dependent [14]. The age range of the infants used in our dataset was from 1 to 208 days; however, during this experiment, similar to our previous studies [46,14,13], we used samples from newborns younger than 53 days. This is because infants above this age can control their own voices [4].

Table 2 shows the numbers of healthy and sick newborns in our dataset. Eighty-four newborns with one of 34 pathologies are in the unhealthy group, whereas the healthy group contains 162 newborns. Each of these newborns in our dataset have a different number of samples. In general, 632 CASs of full-term newborns were found to be eligible to be used in our experiment, among which 316 were healthy newborns, and 316 were unhealthy.

2.2. Definitions of the feature sets

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

[1] An Apgar test is the very first test taken from newborns for measuring their general state of health.

[2] Gestational age is within the range of 27 weeks and 2 days and 41 weeks and 4 days.

2.2.1. 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 of below 1000 Hz and logarithmic intervals of above 1000 Hz. An approximation of Mel for the frequency can be represented as follows:

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

In Eq. (1), f refers to the actual frequency in this equation, and $M(f)$ is the perceived frequency. The main advantage of the 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 discrete fourier transform of the signal, applying the Mel filter banks and log, and finally taking the

Table 2
Labels of the pathology used in our experiment accompanied with the number of individuals in that group. For example, the label of the healthy group is 17.

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

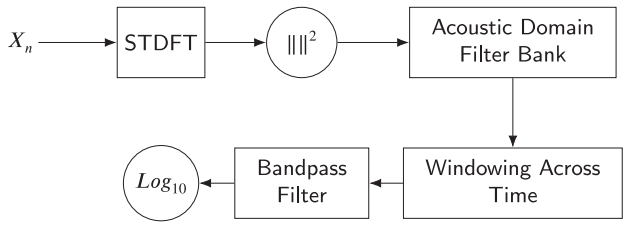

Fig. 3. Block diagram of obtaining the AAM feature set.

inverse discrete fourier transform. In this study, we followed the procedures described in our previous paper [14]. All parameters were taken from our prior experience [14,13] which were found to be beneficial for use in the NCDS.

2.2.2. AAM feature set

The AAM feature set has yet to be investigated in infant CAS analyzing system. The AAM feature set has been successfully tested in other acoustic recognition systems such as nonverbal human-produced audio events [45], and speaker verification [47], and was specifically found to outperform the widely used feature set of MFCCs [44].

Fig. 3 shows the stages for obtaining the AAM feature set. Using our framework, we applied the method described in [44]. Initially, the short-time discrete Fourier transform (STDFT) is applied to the original CASs. Next, the square magnitude was calculated to reflect the mechanism of the human ear. These squared magnitudes of the captured acoustic frequency elements were then classified into 27 subbands based on the perceptual Mel scale.

A second transformation was then applied over time to the magnitude of the CAS signals of all subbands. Subsequently, a bandpass filter was applied. This is due to the physiological evidence of the auditory filter bank formation within the modulation domain. At the end of this process, the logarithm of the feature set is computed to reduce its massive volume [44].

For more information regarding the details of each stage, the authors suggest referring to [44].

2.2.3. Prosody feature set

Humans naturally use various prosodic indications to identify sounds [48]. In the study of sound systems such as speaker recognition, language identification, emotion detection, and speech recognition, although the main focus has been on short spectral information, several studies have shown improvements in the recognition power using prosody [49]. Likewise, the fusion of prosody into the NCDS may have the potential to achieve a more robust system.

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

2.2.3.1. Tilt feature subset. We used tilt features to explore how distinctively F_0 behaves in the CASs of healthy and unhealthy infants' groups. Tilt features have been used favorably in tasks of automatic speaker, language, emotion, and speech recognition [48] and newly applied in the NCDS [14]. In the NCDS [14] such features are used for groups of infants with RDS pathology versus healthy infants. We described the procedure of evaluating the tilt features in our previous study [14]. This feature set was first introduced in [48]. The description of the tilt feature set is as follows.

The main parameters in the tilt feature computations are A_t and D_t. Considering a portion of the CAS, the oscillation of F_0 is captured by the parameters A_t and D_t calculated using Eqs. (2) and (3).

$$A_t = \left(\frac{|A_r| - |A_f|}{|A_r| + |A_f|}\right) \tag{2}$$

Time	Pitch	Intensity
0.0140000	0	--undefined--
0.0280000	394.0919	--undefined--
0.0420000	395.5019	--undefined--
0.0560000	403.5974	73.986
0.0700000	390.3895	72.125
0.0840000	379.7228	70.390
0.0980000	371.5754	69.579
0.1120000	353.4577	68.236
0.1260000	0	67.435
0.1400000	0	67.563
0.1540000	0	67.622
0.1680000	682.7644	67.561
0.1820000	687.9321	67.472
0.1960000	700.2967	67.258
0.2100000	693.6510	66.802
0.2240000	696.6400	66.613
0.2380000	701.7672	--undefined--
0.2520000	700.7368	--undefined--
0.2660000	0	--undefined--

Fig. 4. Illustration of a labeled CAS in WaveSurfer software tool.

$$D_t = \left(\frac{|D_r| - |D_f|}{|D_r| + |D_f|}\right) \tag{3}$$

In above equations, A_f and A_r are the measures of the amplitude of the F_0 contours when they are descending and ascending, respectively. Similarly, D_f and D_r capture the length of the F_0 contour when they are descending and ascending, respectively. More details on the tilt features can be found in [14].

To assess the tilt feature subset, an F_0 contour is required. However, obtaining an accurate F_0 in the infant CASs is a significant problem [50]. Among the well-known methods for F_0 contour extraction, Praat software was proved to be among the most accurate [51]. After computing the precise F_0, the tilt feature parameters were extracted, and the range, standard deviation, and mean of F_0 were then computed.

2.2.3.2. Intensity feature subset. The intensity represents the height of the audio signal. Intensity measures the volume of energy that the audio waveform carries per unit area. The intensity measure was determined using Eq. (4):

$$Intensity = 10log(\sum_{n=1}^{N} A^2(n)w(n)) \tag{4}$$

In the above equation, w is the window function, and A is the amplitude of the CASs.

To extract the intensity of the infant CASs, we used Praat software. The range, mean, standard deviation, median, and interquartile of each feature variation were then measured.

Fig. 4 shows the information of the F_0 and intensity with the time index obtained for a portion of the CASs of newborns within our dataset.

2.2.3.3. Rhythm feature subset. Rhythm features capture the duration patterns of an audio clip. Rhythm features have been quite successful in the language-processing domain. In our previous study [14], we found that the CAS of newborns with RDS differs rhythmically from that of healthy infants. Accordingly, in this research, 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 the timing contrasts

between successive lengths of speech, which is applied to syllables or segments. The formula of the rPVI is defined in [14] as:

$$rPVI = [\frac{\sum_{k=1}^{M-1} |d_k - d_{k+1}|}{m-1}] \quad (5)$$

in which d is equal to the length of each 'EXP' and m is the number of 'EXP' length within a CAS sample. The normalized rPVI used in this study is defined as [14]:

$$nrPVI = 100 \times [\frac{\sum_{k=1}^{M-1} \left| \frac{d_k - d_{k+1}}{\frac{d_k + d_{k+1}}{2}} \right|}{m-1}] \quad (6)$$

- **Std**: The standard deviation of the 'EXP' length in each CAS [14].
- **Varco**: This measures the standard deviation of the 'EXP' length divided by their mean length in each CAS [14].
- **N events**: This is the number of 'EXP' length that occur in each CAS [14].
- **Total duration**: This is the total length of each 'EXP' in every CAS [14].
- **Range**: This is the range (the maximum minus the minimum) of the 'EXP' length in each CAS [14].
- **Mean**: This is the average of all 'EXP' length in each CAS [14].

2.3. Classification

In this study, we examined the functionality of the obtained feature sets using two learning algorithms, PNN and SVM, as binary classifiers between healthy and multi-pathological groups. The PNN and SVM models were trained to classify the CAS feature sets identified by the corresponding states. These feature sets were obtained during the feature-extraction stage.

2.3.1. Probabilistic neural network (PNN)

The efficient PNN classifier was chosen to evaluate the proposed NCDS. This classifier is widely used in classification problems within this field [52,53]. The PNN classifier that is ideal for real-time applications, however, is computationally inexpensive. Using the conjugate gradient method, new incoming training data can be learned without having to repeat the entire training process and without a weight adaptation [54,55].

2.3.2. Support vector machine (SVM)

An SVM is a supervised learning approach widely used in audio classification problems. This method is effective and has been proven to be superior to older ML methods in recent years. As the principle of an SVM, the aim is to find the longest margin that yields the greatest distance between the feature points of each group. The boundary feature points are called support vectors and are then used for training [14,56–58]. In this study, a linear kernel is employed to map the feature space.

3. Model evaluation and results

The inputs of the classifiers in our study are the vectors of the characteristics obtained through the feature extraction steps. Five experiments were conducted to evaluate the efficiency of the system. The experiments used the following feature sets:

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

These five vectors of feature characteristics were used for training and testing the two classes of infant CASs (multiple pathologies and healthy).

The test of the studied system was conducted using a fivefold cross-validation. The folds are independent of each other. Thus, there were no samples of the same infants in more than one fold. Fold (1) contains CASs from all studied pathologies, and fold (2) contains CASs from only a portion of the studied pathologies. Fold (3) contains CASs from seven pathologies. Fold (4) contains CASs from four pathologies, and fold (5) contains CASs from only two pathologies. Table 3 shows the pathology labels in each fold of our experiment. The labels and their corresponding pathologies are shown in Table 2.

To evaluate the performance of the NCDS, measures such as the accuracy, specificity, sensitivity, and F-score were calculated. The equations for the measures mentioned above are as follows:

$$Accuray = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$Fmeasure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (10)$$

Table 3
Distribution of the groups of pathologies in each fold. The numbers relate to the label of the pathology described in Table 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
4	18,22,27,30
5	27,30

3.1. Evaluation of short-term feature sets

The short-term feature sets of the AAM and MFCC resulted in the highest identification rates compared to the prosodic feature set. As shown in Table 4, the AAM feature set consistently outperformed the MFCC feature set when using the PNN model; by contrast, the MFFC feature set outperformed the AAM when using the SVM model.

In general, the best results in terms of accuracy, precision, specificity, sensitivity, and F-measure were obtained for the short-term feature set of the MFCC when using the SVM model.

Table 4
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 5
Results of feeding the joint feature sets of MFCC and AAM, as well as their joint feature sets, 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 F-measure obtained for the proposed short-term feature sets of the MFCC and AAM are 78% and 77.70% when applying the SVM learning model.

3.2. Evaluation of prosodic feature set

As noted in Table 4, the prosodic feature set was less influential than the short-term feature sets. Among the trained models, the prosodic feature set in the configuration with the SVM outperformed the PNN model. The F-measure using SVM and PNN were 65.10% and 53.60%, respectively, for the proposed prosodic feature set.

3.3. Evaluation of feature set fusion

In this section, the results of using the 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 the PNN and SVM.

According to Table 5 with the PNN classifier, the fusion of all feature sets always hinders the performance; however, with the SVM classifier, the performance consistency increases compared to the use of only one short-term feature set. All measurements confirm the inefficacy of all feature sets when fused with the PNN, whereas the recognition power is improved using the SVM model.

For the use of joint short-term feature sets of AAM and MFCC, this fusion consistently increases the performance rates of the classifiers. The F-measure when using the fusion of the AAM and MFCC feature sets with the SVM model was 85.50%, whereas that of the PNN model was 75.60%.

4. Discussion

In this study, we intended to develop an NCDS that resembles the real-world problem. For this, we included a mixture of pathologies in the unhealthy group for distinguishing those of the healthy group. Our other goal was to upgrade the NCDS identification ability by introducing the AAM feature set for the first time into this system. Our proposed model outlines the feature extraction phase to capture the 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. During the classification phase, we explored the SVM and PNN, which are learning algorithms from two different families. The SVM and PNN were trained to solve the binary classification task of identifying unhealthy newborns from healthy newborns. The performance of the models was measured through a 5-fold cross-validation. To ensure the reliability of our system for the real-world problem, during the evaluation part, we designed the folds to be independent of each other. For each 5-fold repetitions, the models were tested on newborn CASs that were not previously trained.

Concerning the short-term feature sets, the MFCC feature set, as expected, performed extremely well in the NCDS. The MFCC feature set was consistently proven to be the most influential feature set in audio processing applications based on a multitude of studies [59]. The second examined short-term feature set is the AAM, which as we saw in the results section provides essential statistical information to the models. According to several measures, as listed in Table 4, the AAM feature set significantly differs among the groups of healthy and unhealthy newborns and is comparatively as powerful as the MFCC feature set. Moreover, the highly reliable results obtained with the AAM feature set used in the NCDS are consistent with high-grade performance of this feature set in the domain of acoustic recognition systems such as nonverbal human-produced audio events [45], and speaker verification [47]. Hence, the AAM feature set can be used as a substitute for the MFCC feature set.

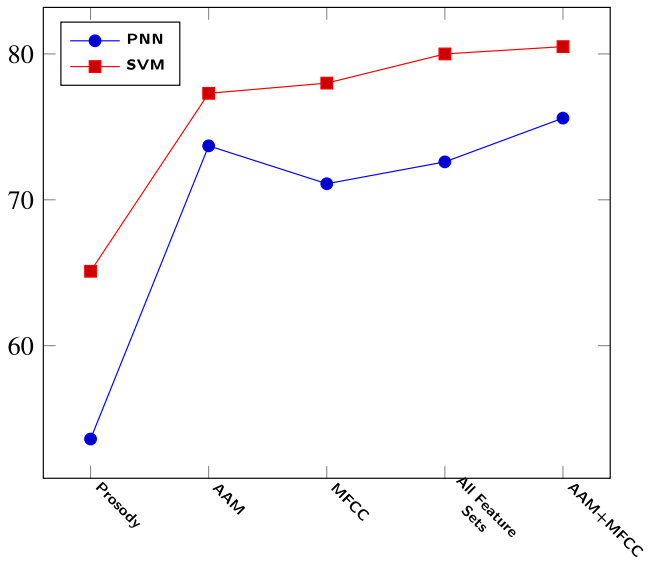

Fig. 5. Changes in the F-measure values obtained by the SVM and PNN models for the five proposed experiments.

Concerning the prosodic feature set, the purpose was to examine whether the health condition of newborns affected the high-level information within their CAS. The results indicate that the system can identify unhealthy newborns with relative accuracy; however, the results obtained are not as satisfying as the short-term feature sets of the MFCC and AAM. For this reason, prosodic features are used as supplemental features to improve the system performance [14]. Furthermore, the research proved that the combination of these features will result in a better performance of the models based on experience with similar systems [60,49]. Hence, the joint vectors of all feature sets were fed into the classifiers. The SVM using the baseline MFCC feature set resulted in an F-measure of 78%, and with all feature sets achieved a higher F-measure of 80%. Thus, the SVM could gain a better hyperplane maximizing the margin between the two classes to distinguish more unhealthy infants.

Conversely, the PNN showed a different way of training all feature sets. The F-measure decreased from 71.10% to 69.10%, as shown in Tables 4 and 5. Fig. 5 also shows how the behaviors of the PNN and SVM change as more feature sets are added.

We repeated our analysis for joint feature sets of the AAM feature set and the MFCC because 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 performances of the SVM and PNN for all criteria examined, with an F-measure of more than 80%. Therefore, the optimal feature vector in our experiment is a combination of the short-term feature sets of the AAM and MFCC. Moreover, these evaluated measures answer our initially asked question regarding whether the

Table 6
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

AAM feature set can complement the popular MFCC feature set when applied to the NCDS.

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

By experimenting with the results of two learning algorithms from different families, the goal was to select the model that achieves the best performance for our dataset in which the SVM defeated the PNN in all experiments, as shown in Fig. 5. To benefit from the excellent performance of the SVM classifier, in a future study, we hope to use it with different kernels, and apply multiple classifier schemes by experimenting with the dynamic selection scheme of the classifiers as well as a stacked classifier.

In terms of computational costs, Table 6 shows the elapsed time required to extract the sets of features used in our experiment. Whereas the AAM and MFCC feature sets help the models achieve the best results, these techniques require a far longer execution time and more 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 and pause patterns in the CASs of newborns.

Another future investigation will be to propose a learning algorithm that addresses the deficiency of unbalanced data between groups of diseases and healthy newborns in our dataset. The Adaboost classifier successfully solves this issue [61], and thus we plan to examine this technique in our forthcoming studies.

5. Conclusion

In this study, we examined the importance of using information at different levels of CAS in newborns as cues to identify unhealthy newborns from healthy ones. For this, we extracted the standard short-term feature set of the MFCC, and for the first time obtained the AAM feature set in the NCDS and the prosodic feature set to capture the statistics at longer intervals of the CASs of newborns. The SVM and PNN classification models were trained using the feature sets mentioned above. In addition to the three feature sets of the MFCC, AAM, and prosody, we also explored the efficacy of the feature set fusion. Two feature vectors of the fusion of all feature sets and the fusion of the AAM and MFCC feature sets were supplied to the classifiers. Our study dataset included newborns belonging to 34 groups of pathologies versus a healthy group. The optimal achievement of the system is related to the fusion of the AAM and MFCC feature sets with an F-measure of over 80% for both the SVM and PNN models.

This research helped establish the importance of the information at different levels of CASs in newborns. Newborns affiliated with a pathology cry differently than healthy newborns, and these different patterns can be statistically captured using ML methods. Such information at different levels is necessary to achieve an upgrade of the NCDS.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was 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]. Declaration of Competing Interest The authors declare no conflict of interest.

References

[1] J. Mukhopadhyay, B. Saha, B. Majumdar, A.K. Majumdar, S. Gorain, B.K. Arya, S. D. Bhattacharya, A. Singh, An evaluation of human perception for neonatal cry using a database of cry and underlying cause, IEEE (2013) 64–67.
[2] A. Sagi, Mothers' and non-mothers' identification of infant cries, Infant Behavior and Development 4 (1981) 37–40.
[3] P.H. Wolff, The role of biological rhythms in early psychological development, Bull. Menninger Clin. 31 (4) (1967) 197.
[4] C.Z. Boukydis, B.M. Lester, Infant crying: Theoretical and research perspectives, Springer Science & Business Media, 2012.
[5] M.J. Kim, Y. Kim, S. Hong, H. Kim, ROBUST detection of infant crying in adverse environments using weighted segmental two-dimensional linear frequency cepstral coefficients, IEEE (2013) 1–4.
[6] R. Cohen, D. Ruinskiy, J. Zickfeld, H. IJzerman, Y. Lavner, Baby Cry Detection: Deep Learning and Classical Approaches, in: W. Pedrycz, S.-M. Chen (Eds.), Development and Analysis of Deep Learning Architectures, Studies in Computational Intelligence, Springer International Publishing, Cham, 2020, pp. 171–196. doi:10.1007/978-3-030-31764-5_7.
[7] L. Abou-Abbas, C. Tadj, C. Gargour, L. Montazeri, Expiratory and inspiratory cries detection using different signals' decomposition techniques, J. Voice.
[8] J.-J. Aucouturier, Y. Nonaka, K. Katahira, K. Okanoya, Segmentation of expiratory and inspiratory sounds in baby cry audio recordings using hidden Markov models, J. Acoust. Soc. Am. 130 (5) (2011) 2969–2977.
[9] N.S.A. Wahid, P. Saad, M. Hariharan, Automatic infant cry classification using radial basis function network, J. Adv. Res. Appl. Sci. Eng. Technol.
[10] Y. Abdulaziz, S.M.S. Ahmad, Infant cry recognition system: a comparison of system performance based on mel frequency and linear prediction cepstral coefficients, IEEE (2010) 260–263.
[11] B. Saha, P.K. Purkait, J. Mukherjee, A.K. Majumdar, B. Majumdar, A.K. Singh, An embedded system for automatic classification of neonatal cry, IEEE (2013) 248–251.
[12] J. Orozco-García, C.A. Reyes-García, A study on the recognition of patterns of infant cry for the identification of deafness in just born babies with neural networks, Springer (2003) 342–349.
[13] H.F. Alaie, L. Abou-Abbas, C. Tadj, Cry-based infant pathology classification using GMMs, Speech Commun. 77 (2016) 28–52.
[14] F. Salehian Matikolaie, C. Tadj, On the use of long-term features in a newborn cry diagnostic system, Biomed. Signal Process. Control 59 (2020), 101889, https://doi.org/10.1016/j.bspc.2020.101889. URL: http://www.sciencedirect.com/science/article/pii/S1746809420300458.
[15] Y. Kheddache, C. Tadj, Characterization of pathologic cries of newborns based on fundamental frequency estimation, Engineering 5 (10) (2013) 272.
[16] S. Lahmiri, C. Tadj, C. Gargour, S. Bekiros, Characterization of infant healthy and pathological cry signals in cepstrum domain based on approximate entropy and correlation dimension, Chaos Solitons Fractals 143 (2021), 110639, https://doi.org/10.1016/j.chaos.2020.110639 https://www.sciencedirect.com/science/article/pii/S0960077920310304.
[17] S. Lahmiri, C. Tadj, C. Gargour, Biomedical diagnosis of infant cry signal based on analysis of cepstrum by deep feedforward artificial neural networks, IEEE Instrum. Meas. Mag. 24 (2) (2021) 24–29, https://doi.org/10.1109/MIM.2021.9400952.
[18] D. Lederman, A. Cohen, E. Zmora, K. Wermke, S. Hauschildt, A. Stellzig-Eisenhauer, On the use of hidden Markov models in infants' cry classification, IEEE (2002) 350–352.
[19] D. Lederman, E. Zmora, S. Hauschildt, A. Stellzig-Eisenhauer, K. Wermke, Classification of cries of infants with cleft-palate using parallel hidden Markov models, Med. Biol. Eng. Comput. 46 (10) (2008) 965–975.
[20] M. Hariharan, R. Sindhu, S. Yaacob, Normal and hypoacoustic infant cry signal classification using time-frequency analysis and general regression neural network, Comput. Methods Programs Biomed. 108 (2) (2012) 559–569.
[21] A. Rosales-Pérez, C.A. Reyes-García, J.A. Gonzalez, O.F. Reyes-Galaviz, H. J. Escalante, S. Orlandi, Classifying infant cry patterns by the Genetic Selection of a

Fuzzy Model, Biomed. Signal Process. Control 17 (2015) 38–46, https://doi.org/10.1016/j.bspc.2014.10.002. http://www.sciencedirect.com/science/article/pii/S1746809414001517.
[22] M. Hariharan, R. Sindhu, V. Vijean, H. Yazid, T. Nadarajaw, S. Yaacob, K. Polat, Improved binary dragonfly optimization algorithm and wavelet packet based non-linear features for infant cry classification, in: Computer Methods and Programs in Biomedicine 155, Elsevier Ireland Ltd., 2018, pp. 39–51, https://doi.org/10.1016/j.cmpb.2017.11.021.
[23] A. Díaz-Pacheco, C. Reyes-García, V. Chicatto-Gasperín, Granule-based fuzzy rules to assist in the infant-crying pattern recognition problem, Sadhana – Acad. Proc. Eng. Sci. 46 (4). doi:10.1007/s12046-021-01736-8.
[24] K. Santiago-Sánchez, C.A. Reyes-García, P. Gómez-Gil, Type-2 fuzzy sets applied to pattern matching for the classification of cries of infants under neurological risk, Springer (2009) 201–210.
[25] S. Orlandi, C. Manfredi, L. Bocchi, M.L. Scattoni, Automatic newborn cry analysis: a Non-invasive tool to help autism early diagnosis, in: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2012, pp. 2953–2956, https://doi.org/10.1109/EMBC.2012.6346583, iSSN: 1558–4615.
[26] O.F. Reyes-Galaviz, A. Verduzco, E. Arch-Tirado, C.A. Reyes-García, Analysis of an infant cry recognizer for the early identification of pathologies, in: Nonlinear Speech Modeling and Applications, Springer, 2005, pp. 404–409.
[27] R. Sahak, W. Mansor, Y.K. Lee, A.M. Yassin, A. Zabidi, Orthogonal least square based support vector machine for the classification of infant cry with asphyxia, vol. 3, IEEE, 2010, pp. 986–990.
[28] R. Sahak, W. Mansor, L.Y. Khuan, A. Zabidi, A.I.M. Yassin, Detection of asphyxia from infant cry using support vector machine and multilayer perceptron integrated with Orthogonal Least Square, IEEE (2012) 906–909.
[29] A. Zabidi, W. Mansor, K.Y. Lee, Optimal feature selection technique for mel frequency cepstral coefficient feature extraction in classifying infant cry with asphyxia, Indon. J. Electr. Eng. Comput. Sci. 6 (3) (2017) 646–655.
[30] A. Zabidi, W. Mansor, L.Y. Khuan, I.M. Yassin, R. Sahak, Classification of infant cries with hypothyroidism using multilayer perceptron neural network, IEEE (2009) 246–251.
[31] A. Zabidi, W. Mansor, L.Y. Khuan, R. Sahak, F. Rahman, Mel-frequency cepstrum coefficient analysis of infant cry with hypothyroidism, in: 2009 5th International Colloquium on Signal Processing Its Applications, 2009, pp. 204–208. doi:10.1109/CSPA.2009.5069217.
[32] S. Orlandi, C.A. Reyes Garcia, A. Bandini, G. Donzelli, C. Manfredi, Application of pattern recognition techniques to the classification of full-term and preterm infant cry, J. Voice 30 (6) (2016) 656–663. doi:10.1016/j.jvoice.2015.08.007. https://www.sciencedirect.com/science/article/pii/S0892199715001733.
[33] J.O. García, C.A.R. García, Acoustic features analysis for recognition of normal and hypoacustic infant cry based on neural networks, Springer (2003) 615–622.
[34] O.M. Badreldine, N.A. Elbeheiry, A.N.M. Haroon, S. ElShehaby, E.M. Marzook, Automatic Diagnosis of Asphyxia Infant Cry Signals Using Wavelet Based Mel Frequency Cepstrum Features, in: 2018 14th International Computer Engineering Conference (ICENCO), 2018, pp. 96–100, iSSN: 2475-2320. doi:10.1109/ICENCO.2018.8636151.
[35] R. Sahak, W. Mansor, Y.K. Lee, A.I.M. Yassin, A. Zabidi, Performance of combined support vector machine and principal component analysis in recognizing infant cry with asphyxia, IEEE (2010) 6292–6295.
[36] A. Zabidi, W. Mansor, L.Y. Khuan, I.M. Yassin, R. Sahak, The effect of f-ratio in the classification of asphyxiated infant cries using multilayer perceptron neural network, IEEE (2010) 126–129.
[37] A. Zabidi, W. Mansor, L.Y. Khuan, I.M. Yassin, R. Sahak, Binary particle swarm optimization and F-ratio for selection of features in the recognition of asphyxiated infant cry, Springer (2011) 61–65.
[38] Wegener, Comparison of Supervised-learning Models for Infant Cry Classification/Vergleich von Klassifikationsmodellen zur Säuglingsschreianalyse, Int. J. Health Profess. 2(1) (2015) 4–15. doi:10.1515/ijhp-2015-0005. https://www.sciendo.com/article/10.1515/ijhp-2015-0005.
[39] D. Jurafsky, J.H. Martin, Speech and language processing, vol. 3, Pearson London, 2014.
[40] M. Hariharan, L.S. Chee, S. Yaacob, Analysis of infant cry through weighted linear prediction cepstral coefficients and probabilistic neural network, J. Med. Syst. 36 (3) (2012) 1309–1315.
[41] M. Hariharan, S. Yaacob, S.A. Awang, Pathological infant cry analysis using wavelet packet transform and probabilistic neural network, Expert Syst. Appl. 38 (12) (2011) 15377–15382.
[42] C. Manfredi, G. Pieraccini, R. Viellevoye, A. Torres-Garcia, C. Reyes-Garcia, Relationships between newborns-cry melody shapes and native language, Firenze University Press, 2017.
[43] C. Ji, X. Xiao, S. Basodi, Y. Pan, Deep Learning for Asphyxiated Infant Cry Classification Based on Acoustic Features and Weighted Prosodic Features, in: 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), 2019, pp. 1233–1240, 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00206.
[44] M. Sarria-Paja, T.H. Falk, Fusion of auditory inspired amplitude modulation spectrum and cepstral features for whispered and normal speech speaker verification, Comput. Speech Language 45 (2017) 437–456, https://doi.org/10.1016/j.csl.2017.04.004. URL: http://www.sciencedirect.com/science/article/pii/S0885230816303382.
[45] R.E. Bouserhal, P. Chabot, M. Sarria-Paja, P. Cardinal, J. Voix, Classification of nonverbal human produced audio events: A pilot study, in: 19th Annual Conference of the International Speech Communication of Proceedings of the Annual Conference of the International Speech Communication Association, International Speech Communication Association, Hyderabad, India, 2018, pp. 1512–1516, https://doi.org/10.21437/Interspeech.2018-2299.
[46] Y. Kheddache, C. Tadj, Identification of diseases in newborns using advanced acoustic features of cry signals, Biomed. Signal Process. Control 50 (2019) 35–44.
[47] T. Kinnunen, K.-A. Lee, H. Li, Dimension reduction of the modulation spectrogram for speaker verification, in: Speaker and Language Recognition Workshop, Odyssey 2008, January 21, 2008 - January 24, 2008, Odyssey 2008: Speaker and Language Recognition Workshop, International Speech Communication Association, Stellenbosch, South africa, 2008.
[48] L. Mary, Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition, 2nd Edition, SpringerBriefs in Speech Technology, Springer International Publishing, 2019. doi:10.1007/978-3-319-91171-7. https://www.springer.com/gp/book/9783319911700.
[49] K. Vicsi, G. Szaszák, Using prosody to improve automatic speech recognition, Speech Commun. 52 (5) (2010) 413–426.
[50] C. Manfredi, A. Bandini, D. Melino, R. Viellevoye, M. Kalenga, S. Orlandi, Automated detection and classification of basic shapes of newborn cry melody, Biomed. Signal Process. Control 45 (2018) 174–181.
[51] S. Orlandi, A. Bandini, F.F. Fiaschi, C. Manfredi, Testing software tools for newborn cry analysis using synthetic signals, Biomed. Signal Process. Control 37 (2017) 16–22.
[52] W.P. Sweeney, M.T. Musavi, J.N. Guidi, Classification of chromosomes using a probabilistic neural network, Cytometry 16(1) (1994) 17–24, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cyto.990160104. doi:10.1002/cyto.990160104. https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.990160104.
[53] M.F. Othman, M.A.M. Basri, Probabilistic Neural Network for Brain Tumor Classification, in: Modelling and Simulation 2011 Second International Conference on Intelligent Systems, 2011, pp. 136–138, iSSN: 2166-0670. doi:10.1109/ISMS.2011.32.
[54] Y. Kheddache, C. Tadj, Newborn's pathological cry identification system, IEEE (2012) 1024–1029.
[55] M. Kusy, R. Zajdel, Probabilistic neural network training procedure based on Q(0)-learning algorithm in medical data classification—SpringerLink.https://link.springer.com/article/10.1007/s10489-014-0562-9.
[56] S. Aydin, S. Demirtaş, S. Yetkin, Cortical correlations in wavelet domain for estimation of emotional dysfunctions, Neural Comput. Appl. 30 (4) (2018) 1085–1094, https://doi.org/10.1007/s00521-016-2731-8. https://doi.org/10.1007/s00521-016-2731-8.
[57] S. Aydin, N. Arica, E. Ergul, O. Tan, Classification of obsessive compulsive disorder by EEG complexity and hemispheric dependency measurements, Int. J. Neural Syst. 25 (3) (2015) 1550010, https://doi.org/10.1142/S0129065715500100.
[58] S. Aydin, N. Güdücü, F. Kutluk, A. Öniz, M. Özgören, The impact of musical experience on neural sound encoding performance, Neurosci. Lett. 694 (2019) 124–128, https://doi.org/10.1016/j.neulet.2018.11.034. URL:https://www.sciencedirect.com/science/article/pii/S0304394018308036.
[59] C. Ji, T.B. Mudiyanselage, Y. Gao, Y. Pan, A review of infant cry analysis and classification, EURASIP J. Audio Speech Music Process. 2021 (1) (2021) 8, https://doi.org/10.1186/s13636-021-00197-5.
[60] A.G. Adami, R. Mihaescu, D.A. Reynolds, J.J. Godfrey, Modeling prosodic dynamics for speaker recognition, vol. 4, IEEE, 2003, pp. IV–788.
[61] S. Aydin, Cross-validated Adaboost Classification of Emotion Regulation Strategies Identified by Spectral Coherence in Resting-State, Neuroinformatics doi:10.1007/s12021-021-09542-7. doi: 10.1007/s12021-021-09542-7.