JASA ARTICLE

Use of psychoacoustic spectrum warping, decision template fusion, and neighborhood component analysis in newborn cry diagnostic systems

Zahra Khalilzad[a] and Chakib Tadj
Department of Electrical Engineering, École de Technologie Supérieur, Université du Québec, Montréal, Québec H3C 1K3, Canada

ABSTRACT:
Dealing with newborns' health is a delicate matter since they cannot express needs, and crying does not reflect their condition. Although newborn cries have been studied for various purposes, there is no prior research on distinguishing a certain pathology from other pathologies so far. Here, an unsophisticated framework is proposed for the study of septic newborns amid a collective of other pathologies. The cry was analyzed with music inspired and speech processing inspired features. Furthermore, neighborhood component analysis (NCA) feature selection was employed with two goals: (i) Exploring how the elements of each feature set contributed to classification outcome; (ii) investigating to what extent the feature space could be compacted. The attained results showed success of both experiments introduced in this study, with 88.66% for the decision template fusion (DTF) technique and a consistent enhancement in comparison to all feature sets in terms of accuracy and 86.22% for the NCA feature selection method by drastically downsizing the feature space from 86 elements to only 6 elements. The achieved results showed great potential for identifying a certain pathology from other pathologies that may have similar effects on the cry patterns as well as proving the success of the proposed framework. © *2024 Acoustical Society of America.*
https://doi.org/10.1121/10.0024618
(Received 16 June 2023; revised 9 January 2024; accepted 10 January 2024; published online 2 February 2024)
[Editor: Yannis Stylianou] Pages: 901–914

I. INTRODUCTION

Thanks to advances in both engineering and medical research, it is now known that pathologic newborns cry diversely compared with healthy newborns and the cry characteristics could differ across different pathologies.[1] These observations sparked the idea for the design of various newborn cry diagnostic systems (NCDSs). So far, NCDS architectures have served the purpose of diagnosing newborns with a certain pathology from the healthy,[2–4] detected healthy newborns from a collective of pathologies,[1,5–7] and very recently differentiated between two pathology groups.[8] In this study, the NCDS was taken one step further to detect a certain group of pathologies among an ensemble of other pathologies. For simplicity, we refer to this assemblage of other pathologies as non-septic.

NCDSs benefit from a vast range of tools that help enhance their final diagnostic performance by improving different stages of the NCDS design. A NCDS design entails three main components, which are namely preprocessing, feature extraction and manipulation, and finally classification, which is vital to all of the audio classification applications.

Crying is a manifestation of the newborn's health since it is the product of an extensive number of organs working together in harmony, and malfunction in any of these organs would be reflected in the generated cry signal.[9] Early studies showed that spectrograms of the healthy newborns followed consistent patterns, whereas the spectrograms of newborns diagnosed with pathologies would have certain acoustic attributes that makes them distinguishable.[10] In this regard, many NCDS designs focused on the extraction and selection of the features that would effectively capture and represent these attributes. In the feature extraction step of the NCDS, a wide range of features from time, frequency, and time-frequency domains were employed, which include but are not limited to Mel frequency cepstral coefficients (MFCCs), gammatone-frequency cepstral coefficients (GFCCs), linear predictive coding (LPC), F0 contour, auditory amplitude modulation, resonance frequency, prosodic features such as rhythm and tilt, entropy-based features (e.g., spectral and approximate entropy), and features inspired by analysis of music, such as harmonic ratio, spectral centroid (SC), and spectral flux.

Among these features, Mel-frequency-based features, more specifically, MFCCs, are the most prevalent and often used as a baseline in many designs to ensure comparability with the works of other researchers. The reason behind the success and prevalence of MFCCs is because of their good discriminative performance;[11] however, the role of cepstral analysis is often not emphasized enough. Cepstral analysis facilitates the discrimination between the source and the filtering in audio analysis tasks since it is a homomorphic transformation. In speech analysis, the basic study of the speech components includes the segregation of how each component affects the final outcome, which translates to

[a]Email: zahra.khalilzad.1@ens.etsmtl.ca

declaring the functions of vocal tract impulse response, glottal pulse, and vocal cord timing. In the study of cry signal analysis for different applications, cepstral analysis was proven highly successful for the mentioned reasons. MFCCs were used to detect asphyxia,[12,13] hearing impairment,[14] sepsis[4,15] cleft palate,[2] respiratory distress syndrome,[3] and hypothyroidism.[16] There are also a number of studies that focus on separating the healthy infants from a collective of pathologies, where MFCCs have served successfully as well.[1] Another cepstral feature set that has recently gained attention are the GFCCs, owing to their better noise robustness, cost-efficiency, and better discriminative performance.[11,17] GFCCs were utilized for speaker identification,[18] emotion recognition based on both newborn cry signals and adult speech,[19,20] and finally detection/discrimination of pathologies based on newborn cry signals.[11] Inspired by this pattern for combining the psychoacoustic frequency warping with cepstral analysis, the idea of using another scale, named bark, along with the gammatone (GT) was worth exploring. Bark-frequency cepstral coefficients (BFCCs) were employed to identify the reason for crying in newborns,[21–23] detection of high-risk prematurity in newborns,[24] enhanced emotion detection from speech signal,[25] and automatic speech recognition[26] and other audio analysis applications. The SC is derived from the study of timbre in musical application and tone measurement in audio signals and utilized for the detection of Alzheimer's from an electroencephalogram (EEG),[27] and in NCDS applications it is used to detect pathologies[15] and developmental disorders[28] and for understanding the reason for crying.[29] Spectral crest is often utilized in feature sets along other spectral features in several studies, with the purpose of visualizing music emotion,[30] detection of hunger from stomach sounds,[31] epileptic seizure detection,[32] and audio fingerprinting.[33]

The next step of the NCDS design is classification, which has been developed with many different methods and classifiers. The support vector machine (SVM),[34] multilayer perceptron (MLP),[8] *K*-nearest neighborhood (KNN),[15] random forest (RF),[35] decision trees,[4] probabilistic neural network (PNN),[6] deep feedforward neural network (DFFNN), convolutional neural networks (CNNs),[5] long short-term memory (LSTM) networks,[11] and many other classification approaches are among the means employed in NCDS design. Although some of these studies compared the results of the mentioned classifiers, very few focused on combining the outcomes of the classifiers to form a final decision, and to the best of the authors' knowledge, there is no prior study in the field of NCDS designs that fulfills this purpose. Decision fusion (DF) has a wide range of applications in healthcare,[36,37] signal processing,[38] image analysis,[39] biological system activities,[40] disease monitoring,[41] drug-target interactions,[42] and many more. DF is lucrative for its role in enhancing combination of different data sources and non-uniform data,[43] enhanced decision making,[44] better performance,[45] and finally, diminution of noise, cost, information drop, and ambiguity.[46–48] There are multiple approaches for combining the outcomes of different classifiers and feature sets: among them, decision template fusion (DTF) was selected for this study since it proved to have better performance in experimental studies, especially on smaller sample sizes. It was shown that DF by the employment of DTs is independent of uncertain surmise and more immune to overtraining.[49]

The contribution of this study can be seen from three main aspects. (i) Distinguishing a certain group of pathology from a conglomeration of other pathologies that are closely related is unprecedented in the study of NCDS. Here, we distinguished sepsis from 31 other pathology groups such as respiratory distress syndrome (RDS), meningitis, etc. (ii) Extracting the crest and SC from bark and equivalent rectangular bandwidth (ERB) spectrum and combining them with cepstral analysis is novel in NCDS designs. (iii) Employment of DTF in NCDS to combine the result of a neural network (NN) classifier and SVM and KNN, which are all trained on different features, is novel in the decision-making stage of NCDS designs.

The importance of newborn sepsis is several-fold; it was among the top 10 mortality causes of newborns worldwide, accounting for around 3×10^6 deaths in children under 5 yr old.[50] The diagnosis of sepsis is complex and based on studying different medical cues: feeding difficulty, fever, convulsion, hemodynamic aberrations, apnea lasting longer than 20 s, and lethargy.[51,52] The study and monitoring of these cues require time and medical equipment; however, time is the most crucial element in treating sepsis, to the extent that once a newborn is suspected of sepsis, antibiotic treatment could be started even without validatory results.[53] Furthermore, availability of medical monitoring and test equipment is not evenly distributed throughout the world, and sadly, the areas that suffer from higher newborn mortality rates are struggling with a lack of sufficient professionals and equipment.[50] Therefore, the design of a diagnostic system that is non-intrusive and non-complex while being time-efficient and not requiring high computational power or state-of-the-art hardware is of high importance. This study presents a NCDS that while delivering acceptable performance, maintains simplicity and non-invasiveness.

This study is composed of four main sections. An introduction (Sec. I) is presented where the problem is highlighted and a short review of literature presents the novelty of the proposed study. Section II expounds the dataset and the proposed methodology, including description of features, classifiers, fusion technique, feature selection model, and evaluation measures. After that, Sec. III provides the results of the experiments and compares and discusses these results. Finally, Sec. IV presents the conclusion of the study.

II. DATA AND METHODOLOGY

A. Dataset description

Before presenting the details of the dataset, it should be highlighted that the most challenging factor in developing the NCDS is data collection and then taking the measures to gain the ethical approvals regarding those data. Even after meeting all the requirements, the occurrence of any

https://doi.org/10.1121/10.0024618

pathology could not be anticipated in any particular time span, which means, for example, over a 2-yr data collection phase, one might or might not encounter a newborn diagnosed with meningitis. Therefore, any acquired data are priceless and should be considered for in-depth analysis.

The dataset in this study was collected from newborns with various origins, races, gestational ages (less than 3 months old), cry stimuli, weights, genders, and pathologies. This was made possible with the collaboration of Saint Justine Hospital of Montreal, Canada and Al-Raee and Al-Sahel Hospitals located in Lebanon. The cry signals were recorded in the presence of noise in both public and private maternity rooms and neonatal intensive care units (NICUs) in the hospital environment, with no predefined conditions. These signals have different lengths from 1 to 4 min, with an average of 90 s, including both useful and unwanted information like staff chatter, equipment beeps, and crying from other newborns. The equipment used for data collection was a digital 2-channel handheld recorder with a 44.1 kHz sampling frequency and 16-bit resolution. The recorder was positioned 10 to 30 cm away from the newborn's mouth for the recording process. Up to 5 recordings were collected from each participant. A detailed overview of the database is represented in Table I.

The reason behind putting a limit on the age of newborns is due to the fact that a cry utterance below 53 days of age is only effectuated due to biological rhythms and the newborn has no control over it.[54] This may be related to the development of the vocal tract, which takes place after 3 months of age, when the supralaryngeal reconfiguration occurs, and therefore, no specific incrementing or decrementing pattern was observed in the average fundamental frequency of the cry signals.[55,56]

The most difficult part in any biomedical problem is data collection and curation. Obtaining the consent of newborns' guardians to record the cry signals and then achieving their consent to include that cry signal in the database are highly challenging. Afterwards, obtaining the ethical and technical approvals from the relevant regulatory bodies (e.g., ethical committees) to add samples to a database is an arduous and toilsome process that might even lead to losing some of the acquired data. More significantly, even after obtaining all the requirements, the occurrence of a certain pathology during the defined course of data collection is unpredictable. This means that one cannot guarantee or predict that there will absolutely be one or more newborns suffering from, for example, meningitis, admitted to the same hospital designated for data collection in a 24-month time span. Therefore, any acquired data are invaluable, and all efforts should be made to enable their investigation. For this study, we decided to include all of the available pathologic recordings as a large general group of "pathologic" newborns for the following reasons. (i) It was observed that each pathology has a unique pattern that distinguishes it from the rest, and hence, being able to sketch a methodology that can identify and candidate the septic newborns beyond these differences, seemed very interesting. (ii) Although there seems to be a limited number of participants from each pathology group (only 1 baby in some cases), there are other factors in play that make each recording from the same newborn different from the others to some extent. It should be noted that we believe that some leakage might be present between the training and test sets due to the similarities in the acoustic structures of the test and training sets, but we believe that there are also sufficient meaningful differences between the recordings of the same newborn to enable us to further analyze them. The data collection took place during different emotional and physical statuses of the newborn, and at the same time, it should be noted that in addition to the

TABLE I. An overview of the database participants.

Gender	Female and male	
Babies' ages	Less than 53 days old	
Weight	0.98 to 5.2 kg	
Origin	Canada, Haiti, Portugal, Syria, Lebanon, Algeria, Palestine, Bangladesh, Turkey	
Race	Caucasian, Arabic, Asian, Latino, African, Native Hawaiian, Quebec	
Cry stimulus	Discomfort, sleepiness, wet diaper, pain, fear, colic, reflux, birth cry, hunger	
Pathology	No. of babies	No. of files
Ankyloglossia	1	3
Apnea	1	3
Asphyxia	1	3
Aspiration	1	3
Bronchiolitis	4	12
Bronchopulmonary dysplasia	1	2
Choanal atresia	1	3
Cleft lip and palate	1	3
Complex cardio	1	3
Cyanosis	2	6
Down syndrome	1	3
Duodenal atresia	1	3
Dyspnea	4	10
Fever	1	3
Gastroschisis	2	4
Grunting	2	6
Hyperbilirubinemia	15	43
Hypoglycemia	2	7
Hypothermia	1	3
Intrauterine growth retardation	1	3
Kidney failure	1	3
Meconium aspiration syndrome	2	6
Meningitis	2	6
Myelomeningocele	1	3
Respiratory distress	33	102
Retraction	1	4
Seizure	1	3
Sepsis	17	53
Tachypnea	2	6
Tetralogy of Fallot	1	2
Thrombose	1	4
Vomit	4	12

diagnostic purposes, newborn cry signals have also been translated in terms of emotions.[21,57] There are numerous studies and even recent desktop and mobile applications[58] that can take different crying samples of the same newborn and categorize them based on the emotional needs of the infant: for example, it was shown that a pain cry is substantially different from the hunger or discomfort cry, and they can be efficiently distinguished by using the correct methodology. It should be also noted that we have at least two recordings from each participant that were used separately for training and testing purposes. As a final point, it should be pointed out that, as shown in Sec. III, the attained results in this study fall within the range of other existing literature and do not show a surprising increase to suggest any concerns.

B. Pre-processing

In Sec. II A, it was mentioned that there is no guarantee that a newborn diagnosed with a certain pathology group would be observed over a prespecified time span. Therefore, upon acquiring data from a certain pathology group, it is desirable to make deeper use of the data by any possible means. The cry signals in our dataset were segmented based on the physiological differences of the acoustic activities during a cry utterance, and different labels were assigned to each segment. The two main acoustic activities, for example, are EXP, which refers to an expiratory cry, and INSV, which refers to a voiced inspiratory cry unit. These labels for the bounded segments were appointed by the means of WAVESURFER (version 1.8.8) (WaveSurfer.js, Stockholm, Sweden) software by the group of researchers in our lab. Furthermore, the outlier samples and those with a length of less than 17 ms (equal to the length of two overlapping windows of 10 ms with a 30% overlap) were omitted. This condition was applied to ensure having a reliable analysis of the dataset.

The number of samples in each group of our study is presented in Table II; it should be noted that the numbers in Table II denote the values before selecting an equal number of samples in order to have an approximately balanced dataset. In total, 2264 samples (1132 from each class) were selected, which formed a training dataset with 1585 samples (789 septic and 796 non-septic) and a test dataset with 679 samples (343 septic and 336 non-septic).

C. Feature extraction

Feature extraction has the highest significance in the design of a NCDS framework as it can change the course of the following steps and affect the final decision. Moreover, the nature of a cry signal is dynamic, non-stationary, and disparate from both speech and music to some extent, while including noise. Therefore, the extraction of features that can represent the cry signal both from the spectral and short-term perspective and originate from the domains of speech processing and music analysis would be of the essence. Moreover, as was previously mentioned, since the cry signal is emanated in the nature of speech generation, employing the cepstral analysis would be inevitable. Consequently, this study combines psychoacoustic-based warping of the spectrum with cepstral analysis for the short-term analysis of the signal and studies the dynamic nature of the signal through delta and delta-delta coefficients of the bark and GT scales. Additionally, in order to capture the spectral properties of the cry signal and explore it from the musical perspective, SC and crest features were extracted.

The bark and ERB or GT scales were developed as psychoacoustic-based spectral measures. The bark scale for frequency f is given by Eq. (1) and the ERB scale by Eq. (2):

$$\text{Bark}\,(f) = 6\,\ln\left[\frac{f}{600} + \left[(f/600)^2 + 1\right]^{0.5}\right], \tag{1}$$

$$\text{ERB}\,(f) = 21.4\,\log\,(\,0.00437f + 1). \tag{2}$$

The ERB scale was chosen since it assists the study of lower frequencies with higher resolution. In addition, it was shown in several studies that ERB scaling resulted in better performance of non-speech classification problems, which is accompanied by more robustness and lower computational costs when compared to the triangular bands that are conventionally employed in MFCC feature extraction.[59,60] In order to attain GFCCs, the cry signal is first windowed into overlapping Hamming filters of 10 ms with 3 ms overlap length; since windowing enhances the performance of the feature extraction step and the non-stationarity of the signal could be neglected in such short frames.

The physiological and psychophysical study of the peripheral auditory system inspired the design of gammatone filters (GFs), which represent the modelling of the cochlea.

A bank of 64 filters was utilized for the extraction of the GFCCs. The magnitudes of the decimated outputs are then loudness-compressed by a cubic root operation, Eq. (3):

$$G_m[i] = \left|\,|g|_{\text{decimate}}[i, m]\right|^{1/3}, \quad i = 0, 1, N-1, m = 0, 1, M-1. \tag{3}$$

N is the total number of GFs, M denotes the number of frames, and $G_m[i]$ represent the time-frequency representation of the input signal. The index m is then removed for simplicity. The GFCCs are then obtained through the application of

TABLE II. Specifications of the dataset.

Sepsis status	No. of participants	No. of segmented training files	No. of segmented test files	Available time (s)	Average duration of samples (s)	No. of samples selected
Septic	17	789	343	1773.66	0.71	1132
Non-septic	110	796	336	10712.28	0.52	1132

https://doi.org/10.1121/10.0024618

a discrete cosine transform (DCT) to the GFs, yielding Eq. (4),

$$\mathrm{GFCC_j} = \sqrt{\frac{2}{N}} \sum_{i=1}^{N-1} G[i] \cos\left(\frac{j\pi}{2N}(2i+1)\right), \quad j = 0, 1, \ldots, N-1, \tag{4}$$

where GF[k] denotes the loudness-compressed response of the GFs, and the number of filters is given by N.[17]

The width of the critical bands of the human auditory system equals 1 bark, and hence, a more direct correlation with the spectral information processing of the human auditory system is achieved when the spectral energy is warped over the bark scale.[59] The process of extracting BFCCs is identical to GFCC feature set, with only the bark scale being the difference. Similar to GFCCs, the BFCC feature set is constituted of 39 elements.

SC is an indicator of how the signal's spectrum looks and where the majority of its mass lies. The average of SC is shown to be a powerful discriminator in audio signals, especially in the field of musical applications.[61] In order to calculate the SC of a given window i, we should take the weighted average of the frequency bins, as shown in Eq. (5),

$$\mathrm{SC(i)} = \frac{\sum_{k=1}^{H/2} f(k)\,|s_k(i)|}{\sum_{k=0}^{H/2} |s_k(i)|}, \tag{5}$$

where $|s_k(i)|$ is the amplitude at the corresponding bin k, H is the number of points in the Fourier transform, and $f(k)$ is the frequency at the kth bin.[62] Note that the frequencies have been mapped to the bark scale prior to the computation of the SC; therefore, we name this feature set the bark spectral centroid (BSC).

Finally, we extracted the equivalent rectangular bandwidth-based spectral crest (ERBS crest) which points out the level of peakiness in the spectrum of the signal.

The crest feature set is a highly informative audio descriptor in musical applications that represents the harmonicity of a spectrum. The crest value is associated with discerning how peaky the spectrum is, where the higher values correspond to the presence of a loud peak compared to the overall curvature of the spectrum. It was shown that crest is rather independent from other pitch-based features and dynamics. In order to obtain the crest feature set, for the ith frame of the signal, short-time Fourier transform (STFT) is applied to yield the power spectrum. The STFT results in k frequency bins across the signal's spectrum with s_k amplitudes. The crest is calculated by the ratio of the maximum value (loudest magnitude) to the arithmetic average of the window's power spectrum, as given in Eq. (6),

$$\mathrm{Crest(i)} = \frac{\max\{s_k(i)\}}{\frac{1}{K}\sum_{k=1}^{K} s_k(i)}, \tag{6}$$

where K denotes the number of STFT bins. In order to form a compact feature set and ensure the comparability along samples of different sizes, the average, standard deviation, H-spread, and median of the crests were calculated to form a feature set with four elements. The same statistical measures were applied for construction of the BSC feature set, which also has four elements.

For higher clarity of the feature space formation, Table III is presented, which summarizes the information given in this section.

D. Classification

In the classification step, all of the feature sets were fed to the three selected classifiers so that their performances would be tested, and also the best classifier + feature sets would be selected for the fusion step. All the classifiers benefitted from validation. SVM and KNN were validated using a stratified fivefold cross-validation. The validation secures the classifiers against overfitting and increases their reliability. Finally, all of the classifiers were optimized using random search. As for the MLP classifier, the validation process is different from those of the SVM and KNN classifiers. The validation process for the MLP consists of assigning a portion of the data for validation and then employing that data to validate the neural network during the training process. For this study, the validation was performed at every 50 iterations and the validation data were shuffled at each epoch. Shuffling means that we input the data to the system in a random order. The shuffling of the data is a vital step in training the classifier and is done with the purpose of variance reduction, enhanced ability of generalization, and preventing the overfitting. We shuffled the data succeeding each epoch to lower the possibility of creating batches that do not correctly represent our dataset.

1. MLP

A MLP classifier has four main components. First, the extracted features are fed to the input of the network, and then they are conveyed forward across the layers. A backpropagation method is employed in order to update the weights of the network, and an optimization function assists the tuning of the weights' update.[63] The decision in a MLP network is made based on having the minimum distance from the decision boundary hyperplane.[64] In this study, the root mean square propagation (RMSprop) optimization function updates the backpropagation weights by the means of minimizing the distance to the decision boundary

TABLE III. Feature space specifications.

Feature set	Components	Vector size
GFCC	13 coefficients, 13 deltas, 13 delta-deltas	39
BFCC	13 coefficients, 13 deltas, 13 delta-deltas	39
BSC	Mean, standard deviation, H-spread, median	4
ERBS crest	Mean, standard deviation, H-spread, median	4

hyperplane.[65] In order to further improve this classifier, random search hyperparameter optimization was employed. The number of input layer neurons was set according to the feature vectors' sizes. The hidden layer consisted of 128 fully connected neurons, which is accompanied by a normalization layer. In order to specify whether the neurons would fire during the process of learning, a hyperbolic tangent activation function was added to the layers. The output layer was made of a fully connected layer with two nodes that represent the two classes of septic versus non-septic and a sigmoid function that is in charge of translating the raw outputs of all the layers into class probabilities. Finally, the classification layer generates the final decision of class labels based on the class probabilities. The learning rate was set equal to 0.001, and the number of epochs was a total of 120; validation data included a 15% random share of all data. Thirty percent of the data were randomly selected for testing and separated from the dataset, and finally, 55% of the data were randomly chosen for training.

2. SVM

As mentioned before, SVMs are one of the most well-known classifiers and have a wide range of applications, especially in the analysis of the audio signals. SVMs are precise, versatile, and capable of dealing with linear and non-linear data. In order to classify data points, the SVM attempts to build a hyperplane that is able to separate the data points of the two classes as far as possible, and if the data are not divisible linearly, the radial basis function (RBF) kernel, which computes the Euclidean distance, is chosen.[66]

3. KNN

KNNs are known for their simplicity and effectiveness. As the name suggests, the basis of classifying the data points is measuring the distance from the neighbors, where each point would be placed in the same class as its neighbors with the lowest distance. There are three elements in a KNN: the distance measure (which can be Minkowski, standard Euclidean, Euclidean, Jaccard, Hamming, cosine, Chebyshev, and Manhattan), the number of neighboring data points K, and sets of labeled data for training and testing.[67]

E. Fusion using decision templates

Suppose that in a classification problem with l classifiers $\{C_1, C_2, \ldots, C_l\}$ and $X = [x_1, x_2, \ldots, x_n]^T$ denotes the n-dimensional input feature vector, which corresponds to the m class labels $W = \{w_1, w_2, \ldots, w_m\}$. Each ith classifier will produce an output where $C_i(X) = [c_{i,1}(X), c_{i,2}(X), \ldots, c_{i,m}(X)]^T$. Here, $c_{i,j}(X)$ represents the posterior probability that the ith classifier suggests that X belongs to the class ω_j.

In order to fuse the outputs of the classifiers, an $l \times m$ decision profile (DP) is constructed, as shown in Eq. (7):

$$DP(X) = \begin{bmatrix} c_{1,1}(X) & \cdots & c_{1,m}(X) \\ \vdots & \ddots & \vdots \\ c_{l,1}(X) & \cdots & c_{l,m}(X) \end{bmatrix}. \tag{7}$$

Each column j shows the possibility that a collective of l classifiers declare that X corresponds to the class label ω_j. Finally, the result of fusion would be in the form of a vector of length m, as shown in Eq. (8):

$$C(X) = [d_1(X), d_2(X), \ldots, d_m(X)]^T. \tag{8}$$

For $d_i(X)$ denotes the possibility that the result of fusion declares the input X to belong to class ω_i. The final decision is made based on a certain rule of fusion, such as minimum (min), maximum (max), median, product, and sum operating each corresponding column of the DP matrix to yield the decision templates (DTs). Here, the minimum rule was chosen, which is given in Eq. (9):

$$d_j(X) = \min_{i=1:l} c_{i,j}(X), \quad j = 1, 2, \ldots, m. \tag{9}$$

The reason behind the selection of the min rule was that it was shown that min/max and product outperformed other fusion rules and the min/max rule showed the best performance for uniformly distributed data. It was also proved that in case of a binary classification, the performances of the min and max fusion rules are equal.[68] This led us to the selection of the minimum as the fusion rule.

Thereafter, the DTs are calculated, as shown in Eq. (10), where z_j denotes the samples that are from class ω_i in the training set Z and the number of z_j is given by N_z:

$$DT_i = \frac{1}{N_z} \sum DP(z_j), \quad z_j \in Z,\ z_j \in \omega_i. \tag{10}$$

The input's labels are decided based on a similarity measure between the DP and different DTs. In this study, the squared Euclidean distance was selected as the similarity measure. Equation (11) shows the calculation for determining the output labels based for a given sample, P,

$$\mathrm{d_E} = \sum_{\mathrm{j=1}}^{\mathrm{m}} \sum_{\mathrm{k=1}}^{\mathrm{l}} \left(\mathrm{c_{k,j}(P)} - \mathrm{dt_i(k,j)}\right)^2, \tag{11}$$

where $\mathrm{d_E}$ represents the Euclidean distance measure between the DP and each $\mathrm{DT_i}$, and $dt_i(k, j)$ stands for the element marking the intersection of column j and row k.[49,69] Figure 1 shows the design of our NCDS employing the DTF technique.

F. NCA feature selection

As a final experiment, neighborhood component analysis (NCA) was implemented to determine which elements of the feature sets contributed the most to the final classification results. NCA is non-parametric and aims to enhance the

https://doi.org/10.1121/10.0024618

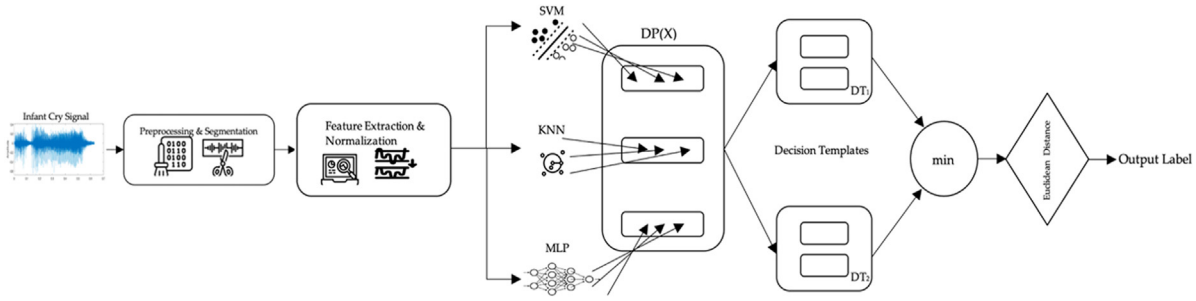

FIG. 1. (Color online) Design of the NCDS employing DTF.

accuracy of the classification to its peak performance. The general performance of the NCA can be explained as a KNN classifier, where K = 1 and neighbors are chosen randomly so that there is a probability for each point in the feature space to be chosen as the reference point. The goal is to learn a classifier that predicts the true label y of x based on the features fed to the input by selecting a random point, Ref(x), from the training set as the reference point and deciding the label of the point x based on this reference point, Ref(x).

The chance of any given point x_j to be picked as the reference point is evaluated based on a weighted distance function, d_w, which is given by Eq. (12),

$$d_w(x_i, x_j) = \sum_{r=1}^{p} w_r^2 |x_{ir} - x_{jr}|, \tag{12}$$

where w_r denotes the weight for the r th feature and p denotes the feature dimension of x_i. In order for the nearest neighbor classifier to perform desirably, one suitable way is to maximize its leave-one-out accuracy. This would not be practical since the selection of the nearest neighbor as the reference point in the leave-one-out accuracy would result in a non-differentiable function. Therefore, the approximation where the reference point is in the form of a probability distribution would be effective. Equation (13) presents the probability p_{ij} that a given point x_i gets x_j as the reference point,

$$p_{ij} = \begin{cases} \dfrac{\kappa\left(d_w(x_i, x_j)\right)}{\sum_{k \neq i} \kappa\left(d_w(x_i, x_k)\right)} & \text{if } i \neq j, \\ o & \text{otherwise}, \end{cases} \tag{13}$$

where κ is a kernel function $\kappa(z) = \exp(-z/\sigma)$ that affects the probability of any given point being selected as the reference point through the kernel width σ, which is determined by an input. In other words, if $\sigma \to \infty$, all of the points in the training set have equal probability to be chosen as the reference point, and if $\sigma \to 0$, only the nearest neighbor has the chance of being the reference point. Hence, the probability of the correct classification of the query point x_i is given by Eqs. (14a) and (14b),

$$p_i = \sum_j p_{ij} y_{ij}, \tag{14a}$$

where

$$y_{ij} = \begin{cases} 1, & y_i = y_j \\ 0 & otherwise. \end{cases} \tag{14b}$$

Now the leave-one-out classifier's accuracy can be approximated by Eq. (15):

$$\mathrm{F(w)} = \frac{1}{N} \sum_i p_i = \frac{1}{N} \sum_i \sum_j p_{ij} y_{ij}. \tag{15}$$

In order to prevent the classifier from overfitting, a positive-valued regularization term, λ, is added to the object function that can affect the influence of the weights and will be tuned via cross-validation. The objective function can now be written as Eq. (16):

$$F(w) = \sum_i \sum_j p_{ij} y_{ij} - \lambda \sum_{l=1}^{p} w_r^2. \tag{16}$$

The $1/N$ term is ignored since it does not affect the solution vector. Finally, in order to maximize the objective function, its derivative with respect to the feature weights is taken, as shown in Eq. (17):

$$\begin{aligned} \frac{\partial F(w)}{\partial w_r} &= \sum_i \sum_j y_{ij} \left[\frac{2}{\sigma} p_{ij} \left(\sum_{k \neq i} p_{ik} |x_{ir} - x_{kr}| - |x_{ir} - x_{jr}| \right) w_r \right] - 2\lambda w_r \\ &= 2 \left(\frac{1}{\sigma} \sum_i \left(p_i \sum_{k \neq i} p_{ik} |x_{ir} - x_{kr}| - \sum_j p_{ij} y_{ij} |x_{ir} - x_{jr}| \right) - \lambda \right) w_r \\ &= 2 \left(\frac{1}{\sigma} \sum_i \left(p_i \sum_{k \neq j} p_{ij} |x_{ir} - x_{jr}| - \sum_j p_{ij} y_{ij} |x_{ir} - x_{jr}| \right) - \lambda \right) w_r. \end{aligned} \tag{17}$$

The above equation is the basis of the NCA feature selection.[70] In this study, in each of the feature sets, the features

that accounted for more than 80% of the final classification results were extracted from the set. Then, these features were concatenated in a single vector and fed to the classifier to determine the role of NCA.

G. Evaluation measures

The framework of this study was designed and developed with the goal of identifying septic infants from a collective of several other pathologies for the first time. The features and classifiers are used from diverse natures and origins, and in order to compare their performance several evaluation measures are introduced in this study. The first measure that is used in any binary classification problem is accuracy due to its simplicity of calculation and being straightforward. Accuracy is computed from the ratio of correct predictions over all the samples. However, this measure is not illuminating enough to cover all aspects of system performance, and more measures are needed to study other aspects of the problem.[71] Therefore, two other measures, namely precision and specificity were studied alongside accuracy. Specificity shows the rate of true negative (TN) cases, which translates to the number of the cases that were correctly marked as non-septic, and precision, or positive predictive value (PPV), denotes how well the NCDS predicts an actual presence of septic cases.[72] The F-score measure is highly instructive as it summarizes these measures into a single value from calculating the harmonic mean of the PPV and the true positive rate (TPR).[73]

These measures evaluate the performance of the system from the problem-solving perspective; however, the system can also be assessed with regard to its classification performance. Therefore, one final evaluation measure is added to our evaluation criteria, which is Matthews' correlation coefficient (MCC). MCC helps elucidate all the information from a contingency matrix (TN, true positive [TP], false negative [FN], and false positive [FP]) as they are all taken into account for the calculation of MCC, as shown in Eq. (18). The value of MCC can be anything in the range of [−1, +1], where the negative value represents a misclassification, zero signifies random classification, and the higher positive values translate into better classification performance:[74,75]

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}. \quad (18)$$

The receiver operating characteristic (ROC) curve assesses the performance of a binary classification problem. In a diagnostic test like the presented study, the results should be classified into a differently diverse category such as the presence of sepsis or its absence. However, since these results are rather ordinal or continuous, we should set a reference (or threshold) to decide the presence of sepsis. An ROC curve is designed for this purpose. It functions via connecting the coordinates where the horizontal axis represents the FP rate (given by 1 − specificity) and the vertical axis represents the sensitivity for all threshold values. The ROC has the benefit of not being altered by pervasiveness, as opposed to the single evaluation measures such as specificity or sensitivity. Moreover, several experiments can be observed and compared at the same time.[76] Therefore, we have included the ROC curves for the NCA and DTF experiments in this study.

TABLE IV. Results for the classification of the BFCC feature set with KNN, SVM, and MLP classifiers.

Classifier	Accuracy	Recall	Specificity	Precision	F1-score	MCC
KNN	67.01	61.90	72.01	68.42	65.00	0.34
SVM	70.99	71.73	70.26	70.26	70.99	0.42
MLP	83.51	79.46	87.46	86.13	82.66	0.67

III. RESULTS AND DISCUSSION

The NCDS in this study was designed and developed with the purpose of identifying the septic newborns among an ensemble of other pathologies for the first time in NCDS designs. The features employed for the NCDS were the ERBS crest, BSC, GFCC, and finally, BFCC. These features were fed to three classifiers, namely SVM, KNN, and MLP. As it was discussed before, in this section, the result of classifying each feature set with different classifiers will be presented first in Tables IV to Table VII. In order to fuse the results of different features fed to various classifiers, the set of feature + classifier that resulted in the highest accuracy measure was selected to form the DPs and DTs. These sets are highlighted in each of the tables below. As for the feature sets of BSC and ERBS crest, the low dimensionality of the feature vectors prevented the MLP classifier from converging, which was expected, and MLP was more suitable for BFCC and GFCC features sets that had a larger size.[8]

The results for the classification of data with BFCC feature set are given in Table IV. The BFCC feature set showed great potential, with the highest accuracy of 83.51% and an MCC of 0.67 with the MLP classifier. By taking a look at Table V, it can be seen that the combination of GFCC + MLP was outperformed by the BFCC feature set with the same classifier. Moreover, Table IV shows that all of the classifiers had positive values for MCC, and hence, the classification was performed successfully. The SVM and KNN classifiers were less efficient in terms of all evaluation criteria; therefore, the set of BFCC + MLP was chosen for the DT and DP calculations of the next experiment.

TABLE V. Results for the classification of the GFCC feature set with KNN, SVM, and MLP classifiers.

Classifier	Accuracy	Recall	Specificity	Precision	F1-score	MCC
KNN	84.09	79.46	88.63	87.25	83.18	0.68
SVM	76.14	75.30	76.97	76.20	75.75	0.52
MLP	82.92	81.25	84.55	83.74	82.48	0.66

https://doi.org/10.1121/10.0024618

TABLE VI. Results for the classification of the BSC feature set with KNN and SVM classifiers.

Classifier	Accuracy	Recall	Specificity	Precision	F1-score	MCC
KNN	63.18	40.48	85.42	73.12	52.11	0.29
SVM	53.61	63.69	63.69	43.73	52.58	0.08

Table V presents the results for the GFCC feature set. The best evaluation measures were achieved through the combination of the GFCC and KNN classifier, with 84.09% for the accuracy and 0.68 for the MCC measure. It is also worth mentioning that this set remarkably achieved the highest values across all of the experiments from the first part. Another point worth highlighting is that not only does GFCC have the overall best performance among feature + classifier combinations, but also it has shown interestingly higher performance with simpler classifiers (SVM and KNN) compared to all of the other combinations in this study. This could signify higher efficiency of the GFCC compared to other feature sets.

The results of evaluating the NCDS with the BSC and ERBS crest feature sets are given in Table VI and Table VII, respectively. It should be highlighted that even though both feature sets had lower performances compared to the GFCC and BFCC feature sets, they only have low dimensions of four elements, which proves their favourable outcomes. The ERBS crest feature set had better performance than the BSC feature set overall; however, each of these feature sets responded better to different classifiers. The highest results achieved for the BSC feature set was via the KNN classifier, with 63.18% and 0.29 for the MCC, which was the combination selected for the next step. The SVM + ERBS crest combination had the highest values across evaluation measures of accuracy and MCC, with 77.91% and 0.62, respectively.

In order to fuse the outputs of the highlighted feature + classifier sets, the corresponding posterior probabilities of training and test datasets, as well as the training labels, were recorded to form the DPs and DTs. The result of the DTF technique for fusion is presented in Table VIII along with the best feature + classifier sets for a clearer interpretation.

Figures 2 and 3 illustrate a comparison of how each feature set and its corresponding evaluation measures were impacted by the fusion. As it can be interpreted from Table VII, Fig. 2, and Fig. 3, the result of the fusion framework enhanced the results in all cases, with an average of 11.49% for accuracy and 13.67% for the F-score. There is only one exception to this conclusion, where the specificity measure for GFCC + KNN set was higher by 0.29%, which is negligible. Even the best results of the first step of these

TABLE VII. Results for the classification of the ERBS crest feature set with KNN and SVM classifiers.

Classifier	Accuracy	Recall	Specificity	Precision	F1-score	MCC
KNN	61.56	36.31	86.30	72.19	48.32	0.26
SVM	77.91	100.00	56.27	69.14	81.75	0.62

TABLE VIII. Results for the DTF technique showing the best feature + classifier sets selected.

Classifier	Feature set	Accuracy	Recall	Specificity	Precision	F1-score	MCC
SVM	ERBS Crest	77.91	100.00	56.27	69.14	81.75	0.62
MLP	BFCC	83.51	71.73	87.46	86.13	82.66	0.67
KNN	BSC	63.18	40.48	85.42	73.12	52.11	0.29
KNN	GFCC	84.09	79.46	88.63	87.25	83.18	0.68
	Fusion	88.66	88.99	88.34	88.20	88.59	0.77

experiments had a 4.57% and 0.09 enhancement in the accuracy and MCC, respectively, with the DF method. The DF method imposes negligible computational cost on the system and is very fast since its calculations only take less than a second. The above results prove the high potential of this method for the design of multimodal NCDS, as presented in this study, where both spectral and short-term features were extracted and employed from musical and speech processing origins. Moreover, as was mentioned before, the enhancement is consistent across different evaluation measures.

Another experiment was carried out to study the role of feature selection and to evaluate to what extent the feature space could be compacted. Each feature set was analyzed with the NCA method, and the features that had the highest contribution to the final classification results were selected.

The NCA revealed some details worth explaining here. The study of both BFCC and GFCC showed that the most significant information in these feature sets belongs to the first 13 coefficients and not their deltas; this was shown in another study on a similar subject,[8] where only 13 GFCC coefficients resulted in around 93% accuracy of classification. Furthermore, the BSC and ERBS crest feature sets both had their 3rd elements selected. The 3rd element in both feature sets belonged to the H-spread, or interquartile range. It can be deduced that this statistical measure has high potential in representing and summarizing spectral data for the infant cries. As for the BSC feature set, the 4th element was also selected, which denotes the median statistical measure.

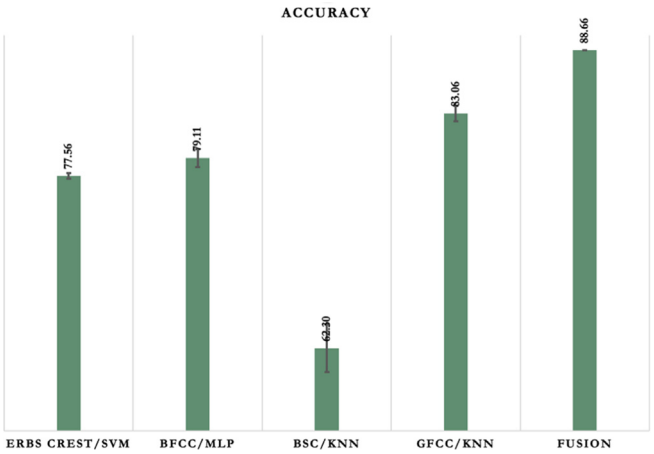

FIG. 2. (Color online) Average of the accuracy measures across different experiments selected for the DTF with 95% confidence intervals.

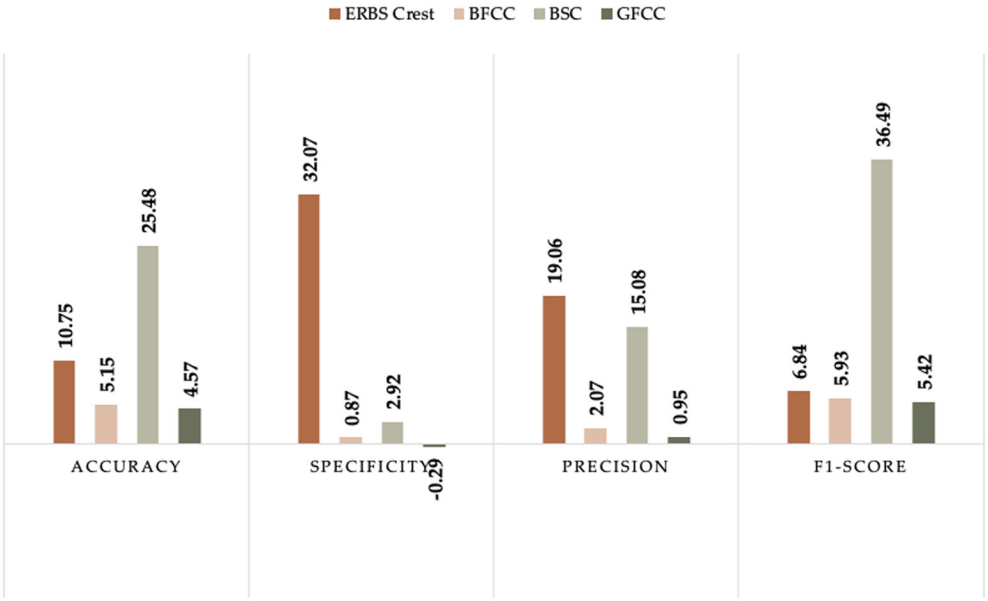

FIG. 3. (Color online) Comparison of the evaluation measures of the selected feature + classifier sets to the DTF technique.

In order to evaluate the system performance with these feature sets, the selected elements from each vector were all concatenated in a single vector and fed to the classifiers of this study. Once more, due to the low dimensionality of the feature vector, the MLP classifier did not converge. The result of the classification of the NCA-selected feature vector with SVM and KNN classifiers is presented in Table IX.

Although the results are lower than DF method, they still represent a high potential and the success of the NCA. In comparison to the GFCC feature set, the accuracy was enhanced by 2.13% and the MCC by 0.05. The enhancement suggests several points. First, the use of NCA will not have a detrimental effect on the final performance of the NCA. Moreover, it has shown that combination of the features from the domains of speech and music would improve the NCDS. Finally, the NCA feature set has only six elements and could obtain an accuracy of more than 86%, which is exceptional. Figure 4 shows a comparison between the results of the two frameworks represented in this study, DF and NCA feature selection. As can be seen from the graphs, not only are the results of the two frameworks compatible, but also the results for specificity and precision measures show better performance with the NCA feature selection method. It can be discussed that the feature selection leads to a more uniform structure of the feature space: extracting the essence of what each feature set represented and combining these elements formed a more powerful indicator in terms of these two measures.

TABLE IX. Results for the NCA feature selection method with KNN and SVM classifiers.

Classifier	Accuracy	Recall	Specificity	Precision	F1-score	MCC
KNN	86.22	80.12	92.19	90.94	85.18	0.73
SVM	78.20	99.23	62.10	70.98	81.12	0.60

Figure 5 and Fig. 6 represent the ROC curves for the assessment of the two methods' performance. As can be deduced, the DTF method outperformed the NCA selection in terms of the area under the curve (AUC). However, it can be seen that DTF did not treat both classes in the same manner: the AUCs for the septic and non-septic classes are different. The behavior of the DTF method was expected since it originated from the fact that we calculated the Euclidean distances with respect to the DPs of each class, which is not the case in a KNN classifier.

Through these graphs, we can see that the single evaluation measures did not highlight all the aspects of the results obtained through our experiments, and thus, employing the ROC curves is indispensable. In addition, it was shown that the DTF method does not have a symmetric performance for both classes as it is based on the DP of each class through various classifiers to form a final decision. With that in mind, we can see that the performance of the DTF for both classes was superior to the NCA method, since the AUC for the NCA with KNN classifier is 0.9358, whereas the AUCs for the DTF method are 0.9483 and 0.9377 for the presence and absence of sepsis, respectively.

This study served three purposes: (i) Assessing the role of decision-level fusion in NCDS designs for the first time; (ii) assessing the role of NCA feature selection in forming a highly compacted feature set while keeping acceptable performance, which is novel in NCDS designs; and (iii) distinguishing a certain pathology (sepsis) amid a collective of other pathologies, which is unprecedented in cry analysis studies.

In addition to the fact that many pathologies remain unexplored or not well-studied in the field of cry diagnostic applications, the NCDS itself has a great potential for further development compared to other audio recognition applications. One of the main areas that could play a part in this development is investigating whether a fusion of different

https://doi.org/10.1121/10.0024618

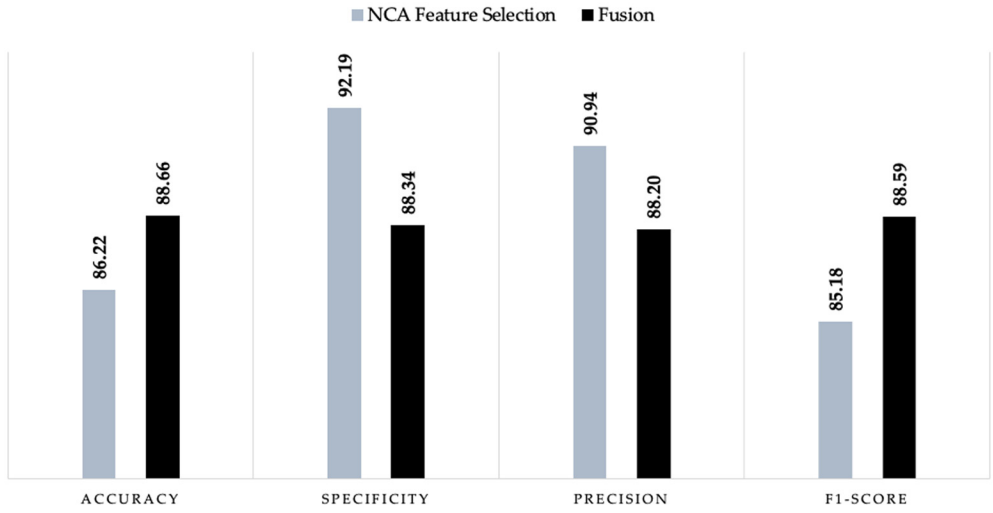

FIG. 4. (Color online) Comparison of the evaluation measures for the fusion framework and the NCA feature selection method.

modalities would contribute to the enhancement of the final decision made by the NCDS, which was the purpose of this study. This framework opens the door for employing features and classifiers from various modalities without the need for complicated designs and advanced technology. The importance of keeping the design simple arises from the fact that, unfortunately, the regions that are reported as having higher newborn mortality rates suffer from lack of adequate medical equipment and professionals and are listed among low-income and middle-income areas. Therefore, if the NCDS can classify as candidates the newborns with higher risk of suffering from certain pathologies, especially sepsis, and rule out the others, the existing equipment and experts can tend to the newborns marked with higher risk.

Although sepsis is closely entangled with newborn mortality rates,[50] the number of newborn cry studies targeting sepsis is scant. In order to address this research gap, the researchers in our lab made efforts to study sepsis from different perspectives. Matikolaie *et al.*[4] utilized prosodic features to distinguish between healthy and septic infants and attained 86% for the best F-score. Khalilzad *et al.*[15] introduced an entropy-based framework by extracting the spectral entropy cepstral coefficients and then having a fuzzy entropy as the feature selection means for the identification of septic infants from the healthy group, obtaining 88.51% for the accuracy regarding the expiratory cries. In another study, Khalilzad *et al.*[8] differentiated between RDS and septic cries through the combination of music-derived features

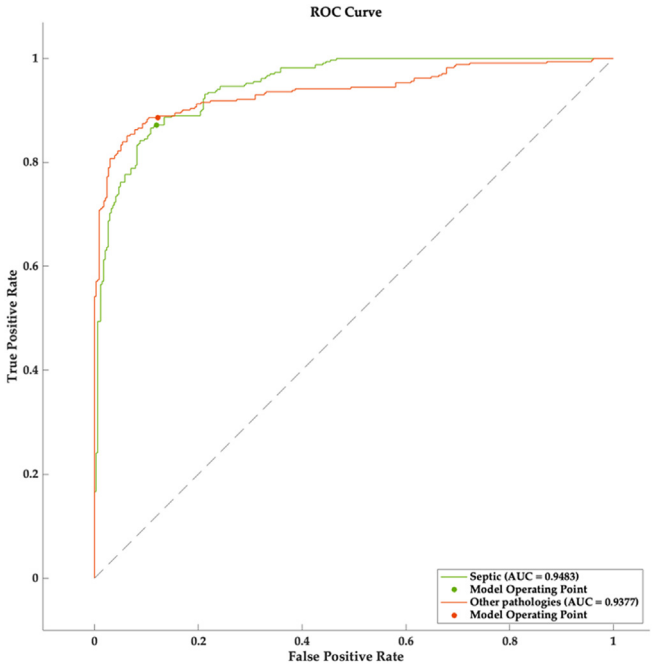

FIG. 5. (Color online) The receiver operating characteristic (ROC) curve for the DTF experiments.

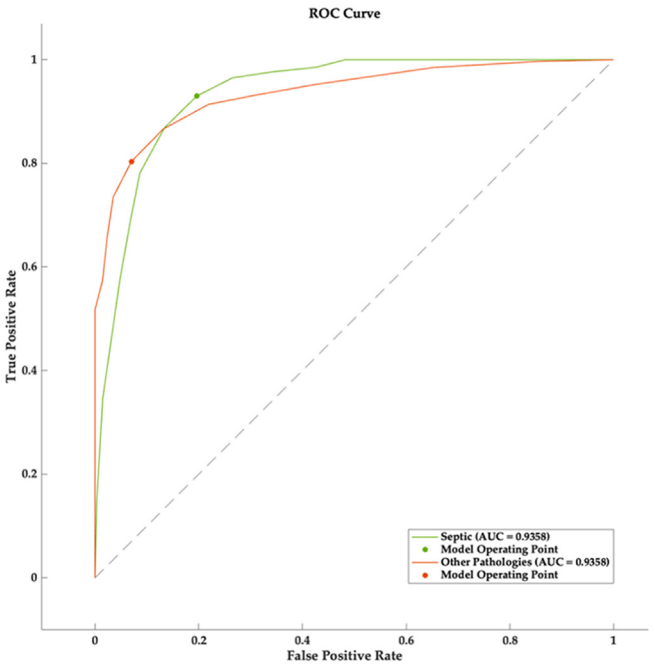

FIG. 6. (Color online) The receiver operating characteristic (ROC) curve for the NCA selection experiments.

TABLE X. Comparison of different works employing fusion and feature selection techniques.[a]

Study	Goal	Features	Fusion/feature selection	Machine/deep learning methods	Best outcome
Dar *et al*. (Ref. 77)	Detecting pulmonary abnormalities from the respiration sound	BFCC, SC, and spectral flux	Simple concatenation of features	Hierarchical attention network, CNN, RF	92.4% accuracy by HAN
Ebrahimpour *et al*. (Ref. 78)	Recognition of hand-written digits in Persian and English	Characteristic Loci.	DTF and PCA	MLP, decision tree, RBF	Accuracy: 97.99%
Fernandes *et al*. (Ref. 79)	Identifying underwater targets based on acoustical recordings from a hydrophone	GTCC, LPC, and MFCC	NCA	KNN	Accuracy: 83.3%
Khalilzad *et al*. (Ref. 15)	Detecting septic newborns from healthy newborns via their cry signals	MFCC, spectral entropy cepstral coefficients, SC cepstral coefficients	Fuzzy entropy feature selection	SVM, KNN	Accuracy: 91.81%
Khalilzad *et al*. (Ref. 11)	Detecting pathologic newborns based on their cry signal	MFCC, GFCC	Canonical correlation analysis-based feature fusion	LSTM, SVM	Accuracy: 99.86%
Li *et al*. (Ref. 38)	Detecting breast cancer via microwave breast screening	PCA scores	DTF, concatenation, PCA	SVM	Average error: 0.01
This study	Detecting septic newborns among other pathologic newborns	ERBS crest, BSC, GFCC, BFCC	NCA feature selection, DTF, concatenation	SVM, KNN, MLP	Accuracy: 88.66%

[a]HAN, Hierarchical attention network; PCA, principal component analysis.

of harmonic ratio (HR) and GFCC features that yielded 95.29% for accuracy.

Up to this point, the NCDS performance was compared regarding its performance with different features and classifiers, before and after applying the DF method and NCA feature selection. Also, the studies that scrutinized sepsis via cry signals were compared in terms of the methods, their purposes, and their performance with the accuracy or F-score measures. This framework could also be compared to the existing literature in terms of the methods employed here and in other NCDS designs. Table X presents a short comparison of the proposed framework with other similar works in the literature.

As can be interpreted through all the aforementioned studies, each of the tools that was implemented in the proposed framework has shown great performance with different applications. The results of our framework also suggest the promising potential of studying DTF and NCA feature selection methods for further studies in NCDS development. In the future, it would be great to explore the role of features from different modalities and more classifiers with the proposed framework here. Moreover, it would be fruitful to investigate how fusion at each level would affect the outcome of the system. Finally, it would be an interesting subject to compare different fusion rules such as maximum, product, etc., for the DTF technique.

IV. CONCLUSION

The cry signal is a powerful biomarker for studying the physical health and needs of a newborn. This study aimed to introduce a simple yet effective framework that was capable of capturing different aspects of the septic cry in comparison to a variety of other pathologies. The study of cry signals was performed independent of newborns' race, gender, weight, and the reason for their crying. The cry signal is different from both speech and music, yet shares so many common attributes with both. Via implementing features from different modalities and properties, both aspects of the cry were studied, and each of the four introduced feature sets of BSC, ERBS crest, GFCC, and BFCC showed desirable performance individually. Then, through the DTF technique these feature sets were fused, and the outcome surpassed the results of all the individual feature sets by an average of 11.49% for the accuracy measure, reaching up to 88.66%, which marks a notable increment and potential.

In order to achieve a more simplistic design and take a deeper look at each of the introduced feature sets, the NCA feature selection method was employed, where each of the feature sets was analyzed, and the indices that contributed the most to the final result were chosen. Next, all of the selected indices were concatenated to form a single feature vector that achieved 86.22% for the accuracy measure.

This study aimed to design an unsophisticated NCDS that served as an alert system to the medical experts for prioritizing the newborns with higher risk of being diagnosed with the fatal pathology of sepsis. The proposed framework showed that septic newborns could be effectively distinguished among a collective of other pathologies only based on their cries. Therefore, this framework could be employed as a non-invasive tool for diagnosis of septic from non-septic.

ACKNOWLEDGMENTS

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) (RGPIN-2016-05067) and was made possible through the funding provided by the Bill and Melinda Gates Foundation (OPP1025091).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Ethics Approval

The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee, École de Technologie Supérieure #H20100401 (approval date: 4 March 2022).

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to the restrictions of the ethical committee of École de Technologie Supérieure, which makes the data subject to private use only.

[1]H. F. Alaie, L. Abou-Abbas, and C. Tadj, "Cry-based infant pathology classification using GMMs," Speech Commun. **77**, 28–52 (2016).
[2]R. M. Massengill, Jr, "Cry characteristics in cleft-palate neonates," J. Acoust. Soc. Am. **45**(3), 782–784 (1969).
[3]F. S. Matikolaie and C. Tadj, "On the use of long-term features in a newborn cry diagnostic system," Biomed. Signal Process. Control **59**, 101889 (2020).
[4]F. S. Matikolaie and C. Tadj, "Machine learning-based cry diagnostic system for identifying septic newborns," J. Voice (published online 2022).
[5]S. Lahmiri, C. Tadj, C. Gargour, and S. Bekiros, "Deep learning systems for automatic diagnosis of infant cry signals," Chaos, Solitons Fractals **154**, 111700 (2022).
[6]F. S. Matikolaie, Y. Kheddache, and C. Tadj, "Automated newborn cry diagnostic system using machine learning approach," Biomed. Signal Process. Control **73**, 103434 (2022).
[7]S. Lahmiri, C. Tadj, and 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), 24–29 (2021).
[8]Z. Khalilzad, A. Hasasneh, and C. Tadj, "Newborn cry-based diagnostic system to distinguish between sepsis and respiratory distress syndrome using combined acoustic features," Diagnostics **12**(11), 2802 (2022).
[9]A. Fort and C. Manfredi, "Acoustic analysis of newborn infant cry signals," Med. Eng. Phys. **20**(6), 432–442 (1998).
[10]K. Michelsson, P. Sirviö, and O. Wasz-Höckert, "Pain cry in full-term asphyxiated newborn infants correlated with late findings," Acta Paediatr. **66**(5), 611–616 (1977).
[11]Z. Khalilzad and C. Tadj, "Using CCA-fused cepstral features in a deep learning-based cry diagnostic system for detecting an ensemble of pathologies in newborns," Diagnostics **13**(5), 879 (2023).
[12]A. Zabidi, W. Mansor, and K. Y. Lee, "Optimal feature selection technique for Mel frequency cepstral coefficient feature extraction in classifying infant cry with asphyxia," Indones. J. Electr. Eng. Comput. Sci. **6**, 646–655 (2017).
[13]N. S. A. Wahid, P. Saad, and M. Hariharan, "Automatic infant cry classification using radial basis function network," J. Adv. Res. Appl. Sci. Eng. Technol. **4**(1), 12–28 (2020).
[14]M. M. Jam and H. Sadjedi, "Identification of hearing disorder by multiband entropy cepstrum extraction from infant's cry," in *2009 International Conference on Biomedical and Pharmaceutical Engineering*, Singapore (IEEE, New York, 2009), pp. 1–5.
[15]Z. Khalilzad, Y. Kheddache, and C. Tadj, "An entropy-based architecture for detection of sepsis in newborn cry diagnostic systems," Entropy **24**(9), 1194 (2022).
[16]A. Zabidi, W. Mansor, L. Y. Khuan, I. M. Yassin, and R. Sahak, "Classification of infant cries with hypothyroidism using multilayer perceptron neural network," in *2009 IEEE International Conference on Signal and Image Processing Applications*, Kuala Lumpur, Malaysia (IEEE, New York, 2009), pp. 246–251.
[17]X. Valero and F. Alias, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," IEEE Trans. Multimedia **14**(6), 1684–1689 (2012).
[18]S. Admuthe and P. H. Patil, "Feature extraction method-MFCC and GFCC used for speaker identification," Int. J. Sci. Res. Dev. **3**(04), 1261–1264 (2015).
[19]E. Garg and M. Bahl, "Emotion recognition in speech using gammatone cepstral coefficients," Int. J. Appl. Innov. Eng. Manag. **3**(10), 285–291 (2014).
[20]U. Kumaran, S. Radha Rammohan, S. M. Nagarajan, and A. Prathik, "Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN," Int. J. Speech Technol. **24**, 303–314 (2021).
[21]L. Liu, W. Li, X. Wu, and B. X. Zhou, "Infant cry language analysis and recognition: An experimental approach," IEEE/CAA J. Automatica Sin. **6**(3), 778–788 (2019).
[22]T. S. N. Sriraam and G. Pradeep, "Pre-term neonates cry pattern recognition using bark frequency cepstral coefficients," in *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE 2019)*, Bangalore, India (IEEE, New York, 2019), pp. 335–338.
[23]T. N. Maghfira, T. Basaruddin, and A. Krisnadhi, "Infant cry classification using CNN–RNN," J. Phys Conf. Ser. **1528**, 012019 (2020).
[24]S. Tejaswini, N. Sriraam, and G. Pradeep, "Identification of high risk and low risk preterm neonates in NICU: Pattern recognition approach," in *Biomedical and Clinical Engineering for Healthcare Advancement*, edited by N. Sriraam (IGI Global, Hershey, PA, 2020), pp. 119–140.
[25]S. Lalitha, S. Tripathi, and D. Gupta, "Enhanced speech emotion detection using deep neural networks," Int. J. Speech Technol. **22**, 497–510 (2019).
[26]D. Kamińska, T. Sapiński, and G. Anbarjafari, "Efficiency of chosen speech descriptors in relation to emotion recognition," EURASIP J. Audio Speech Music Process. **2017**(1), 3.
[27]N. Kulkarni and V. Bairagi, "Extracting salient features for EEG-based diagnosis of Alzheimer's disease using support vector machine classifier," IETE J. Res. **63**(1), 11–22 (2017).
[28]A. Oren, A. Matzliach, R. Cohen, and H. Friedman, "Cry-based detection of developmental disorders in infants," in *2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, Eilat, Israel (IEEE, New York, 2016), pp. 1–5.
[29]A. Osmani, M. Hamidi, and A. Chibani, "Machine learning approach for infant cry interpretation," in *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, Boston, MA (IEEE, New York, 2017), pp. 182–186.
[30]D. Kim, P. V. Ho, and Y. Lim, "A new recognition method for visualizing music emotion," Int. J. Electr. Comput. Eng. **7**(3), 1246 (2017).
[31]A. Maria and A. S. Jeyaseelan, "Development of optimal feature selection and deep learning toward hungry stomach detection using audio signals," J. Control Autom. Electr. Syst. **32**(4), 853–874 (2021).
[32]D. P. Dash and M. H. Kolekar, "EEG-based epileptic seizure detection using least square SVM with spectral and multiscale key point energy features," in *Soft Computing for Problem Solving 2019: Proceedings of SocProS 2019*, edited by A. K. Nagar, K. Deep, J. C. Bansal, and K. N. Das (Springer, New York, 2020), Vol. 1, pp. 323–335.
[33]A. Ramalingam and S. Krishnan, "Gaussian mixture modeling using short time Fourier transform features for audio fingerprinting," in *2005 IEEE International Conference on Multimedia and Expo*, Baltimore, MD (IEEE, New York, 2005), pp. 1146–1149.
[34]P. D. R. Vincent, K. Srinivasan, and C.-Y. Chang, "Deep learning assisted neonatal cry classification via support vector machine models," Front. Public Health **9**, 670352 (2021).
[35]C.-Y. Chang, S. Bhattacharya, P. Raj Vincent, K. Lakshmanna, and K. Srinivasan, "An efficient classification of neonates cry using extreme gradient boosting-assisted grouped-support-vector network," J. Healthc. Eng. **2021**, 7517313.
[36]L. Terveen and W. Hill, "Beyond recommender systems: Helping people help each other," HCI New Millennium **1**(2001), 487–509 (2001).
[37]M. Niemeijer, M. D. Abramoff, and B. Van Ginneken, "Information fusion for diabetic retinopathy CAD in digital color fundus photographs," IEEE Trans. Med. Imaging **28**(5), 775–785 (2009).
[38]Y. Li, E. Porter, A. Santorelli, M. Popović, and M. Coates, "Microwave breast cancer detection via cost-sensitive ensemble classifiers: Phantom

and patient investigation," Biomed. Signal Process. Control **31**, 366–376 (2017).
[39]M. Velikova, P. J. Lucas, M. Samulski, and N. Karssemeijer, "A probabilistic framework for image information fusion with an application to mammographic analysis," Med. Image Anal. **16**(4), 865–875 (2012).
[40]J. Synnergren, B. Olsson, and J. Gamalielsson, "Classification of information fusion methods in systems biology," In Silico Biol. **9**(3), 65–76 (2009).
[41]M. L. Fung, M. Z. Chen, and Y. H. Chen, "Sensor fusion: A review of methods and applications," in *2017 29th Chinese Control and Decision Conference (CCDC)*, Chongqing, China (IEEE, New York, 2017), pp. 3853–3860.
[42]L. Peng, B. Liao, W. Zhu, Z. Li, and K. Li, "Predicting drug-target interactions with multi-information fusion," IEEE J. Biomed. Health Inform. **21**(2), 561–572 (2017).
[43]J. M. Fontana, M. Farooq, and E. Sazonov, "Automatic ingestion monitor: A novel wearable device for monitoring of ingestive behavior," IEEE Trans. Biomed. Eng. **61**(6), 1772–1779 (2014).
[44]S. O'Regan and W. Marnane, "Multimodal detection of head-movement artefacts in EEG," J. Neurosci. Methods **218**(1), 110–120 (2013).
[45]S. Acharya, A. Rajasekar, B. S. Shender, L. Hrebien, and M. Kam, "Real-time hypoxia prediction using decision fusion," IEEE J. Biomed. Health Inform. **21**(3), 696–707 (2017).
[46]B. Lelandais, S. Ruan, T. Denoeux, P. Vera, and I. Gardin, "Fusion of multi-tracer PET images for dose painting," Med. Image Anal. **18**(7), 1247–1259 (2014).
[47]H. Zhong and J. Xiao, "Enhancing health risk prediction with deep learning on big data and revised fusion node paradigm," Sci. Program. **2017**, 1901876.
[48]É. Bossé and B. Solaiman, *Information Fusion and Analytics for Big Data and IoT* (Artech House, Norwood, MA, 2016).
[49]L. I. Kuncheva, J. C. Bezdek, and R. P. Duin, "Decision templates for multiple classifier fusion: An experimental comparison," Pattern Recognit. **34**(2), 299–314 (2001).
[50]World Health Organization, "Newborn mortality (2021)," available at https://www.who.int/news-room/fact-sheets/detail/levels-and-trends-in-child-mortality-report-2021 (Last viewed June 10, 2023).
[51]J. L. Wynn and H. R. Wong, "Pathophysiology and treatment of septic shock in neonates," Clin. Perinatol. **37**(2), 439–479 (2010).
[52]Mayo Clinic, ARDS, https://www.mayoclinic.org/diseases-conditions/ards/symptoms-causes/syc-20355576 (Last viewed August 16, 2022).
[53]J. Ruiz-Contreras, L. Urquía, and R. Bastero, "Persistent crying as predominant manifestation of sepsis in infants and newborns," Pediatr. Emerg. Care **15**(2), 113–115 (1999).
[54]B. M. Lester and C. Z. Boukydis, *Infant Crying: Theoretical and Research Perspectives* (Springer, New York, 1985).
[55]K. Lind and K. Wermke, "Development of the vocal fundamental frequency of spontaneous cries during the first 3 months," Int. J. Pediatr. Otorhinolaryngol. **64**(2), 97–104 (2002).
[56]V. Fisichelli, S. Karelitz, R. Fisichelli, and J. Cooper, "The course of induced crying activity in the first year of life," Pediatr. Res. **8**(12), 921–928 (1974).
[57]B. Saha, P. K. Purkait, J. Mukherjee, A. K. Majumdar, B. Majumdar, and A. K. Singh, "An embedded system for automatic classification of neonatal cry," in *2013 IEEE Point-of-Care Healthcare Technologies (PHT)*, Bangalore, India (IEEE, New York, 2013), pp. 248–251.
[58]Ubenwa, "Your baby's cry is a window to their health," https://www.ubenwa.ai/ubenwa-app.html (Last viewed July 20, 2023).
[59]J. O. Smith and J. S. Abel, "Bark and ERB bilinear transforms," IEEE Trans. Speech Audio Process. **7**(6), 697–708 (1999).
[60]T. Gulzar, A. Singh, and S. Sharma, "Comparative analysis of LPCC, MFCC and BFCC for the recognition of Hindi words using artificial neural networks," Int. J. Comput. Appl. **101**(12), 22–27 (2014).
[61]N. Kulkarni, "Use of complexity based features in diagnosis of mild Alzheimer disease using EEG signals," Int. J. Inf. Technol. **10**(1), 59–64 (2018).
[62]W. Brent, "Physical and perceptual aspects of percussive timbre," Ph.D. dissertation, University of California, San Diego, 2010.
[63]F. Murtagh, "Multilayer perceptrons for classification and regression," Neurocomputing **2**(5-6), 183–197 (1991).
[64]G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning* (Springer, New York, 2013).
[65]G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning, Lecture 6a, Overview of mini-batch gradient descent," **14**(8), 2 (2012), https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.
[66]S. Winters-Hilt, A. Yelundur, C. McChesney, and M. Landry, "Support vector machine implementations for classification & clustering," BMC Bioinformatics **7**(2), S4 (2006).
[67]X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," Knowl. Inf. Syst. **14**, 1–37 (2008).
[68]L. I. Kuncheva, "A theoretical study on six classifier fusion strategies," IEEE Trans. Pattern Anal. Mach. Intell. **24**(2), 281–286 (2002).
[69]A. Mi, L. Wang, and J. Qi, "A multiple classifier fusion algorithm using weighted decision templates," Sci. Program. **2016**, 3943859 (2016).
[70]W. Yang, K. Wang, and W. Zuo, "Neighborhood component feature selection for high-dimensional data," J. Comput. **7**(1), 161–168 (2012).
[71]M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," Int. J. Data Mining Knowl. Manag. Process. **5**(2), 01 (2015).
[72]W. Zhu, N. Zeng, and N. Wang, "Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations," in *NESUG 2010 Proceedings: Health Care and Life Sciences*, Baltimore, MD (SAS, Cary, NC, 2010), p. 67.
[73]P. Flach and M. Kull, "Precision-recall-gain curves: PR analysis done right," Adv. Neural Inf. Process. Syst. **28**, 838–846 (2015).
[74]D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics **21**(1-13), 6 (2020).
[75]M. Vihinen, "How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis," BMC Genomics **13**(4), S2–S10 (2012).
[76]T. Fawcett, "An introduction to ROC analysis," Pattern Recognit. Lett. **27**(8), 861–874 (2006).
[77]J. A. Dar, K. K. Srivastava, and S. A. Lone, "Spectral features and optimal hierarchical attention networks for pulmonary abnormality detection from the respiratory sound signals," Biomed. Signal Process. Control **78**, 103905 (2022).
[78]R. Ebrahimpour and S. Hamedi, "Hand written digit recognition by multiple classifier fusion based on decision templates approach," World Acad. Sci. Eng. Technol. **57**, 560–565 (2009).
[79]R. P. Fernandes and J. A. Apolinário, Jr, "Underwater target classification with optimized feature selection based on genetic algorithms," in *Proc. Simpósio Brasileiro de Telecomunicações e Processamento De Sinais*, Brazil (2020).