Fast Autocorrelation Feature-Based Infant Cry Detector for Resource-Efficient Affordable Edge Cry Sound Analysis Systems

Sivaranjini Perikamana Narayanan
Dept of ICT, University of Agder, Norway
Dept of EE, IIT Palakkad, India
E-mail: sivaranjin@student.uia.no
122213002@smail.iitpkd.ac.in

M. Sabarimalai Manikandan
Department of Electrical Engineering
Indian Institute of Technology Palakkad
Palakkad, Kerala-678623, India
E-mail: msm@iitpkd.ac.in

Linga Reddy Cenkeramaddi
Department of ICT
The University of Agder
Grimstad, 4879 Norway
E-mail: linga.cenkeramaddi@uia.no

2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA) | 979-8-3503-6086-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIEA61579.2024.10664677

***Abstract*—The effectiveness and efficiency of automated cry sound analysis systems highly rely on the accurate and reliable determination of cry sound portions in audio in the presence of different kinds of background sounds with lower computational loads. In this paper, we present a fast and resource-efficient infant cry detection (Infant-Cry-Detect) method using five features extracted from the autocorrelation function (ACF) of the sound and machine learning (ML) classifiers. On a large amount of wide variety of cry sound patterns and six background sounds, the effectiveness of ACF features is studied with six ML classifiers such as decision tree (DT), random forest (RF), Naive Bayes (NB), multi-layer perceptron (MLP), light gradient boosting machine (LGBM) and k-nearest neighbors (KNN) in terms of sensitivity (SE), specificity (SP), model size, and processing time. Evaluation results on untrained sounds show that the NB-based Infant-Cry-Detect method achieves a higher SE of 99.29% in accurately detecting a 1-second cry sound with lower processing time (1.11 ms) and model size (912 kB) as compared to the other five ML classifiers. Further, it is noted that the RF-based Infant-Cry-Detect method outperforms in terms of overall accuracy of 93.77% with higher computational cost (detection time of 8.064 ms and model size of 349.7 MB) as compared to the other five ML classifiers. Results demonstrate that the proposed ACF feature extraction not only achieves higher sensitivity in the presence of various background sounds but it can also reduce the overall computational load as compared to computationally expensive deep learning networks.**

***Index Terms*—Infant Health and Wellness Monitoring, Infant Cry Detection, Infant Cry Analysis and Classification, Edge-AI Cry Sound Analytics, New Born Baby Signal Diagnostic System**

I. Introduction

Accurate and reliable infant cry sound detection plays a major role in automatic cry analysis applications including cry-induced vital sign variation prediction, infant pain analysis, infant hunger prediction, infant discomfort sleepiness or infant sleep disorders, and stranger anxiety notification [1]–[4]. Therefore, automated detection of infant cry sounds or newborn baby cry sounds is highly demanded not only for the above-mentioned cry sound analysis applications but also for timely notifying parents or caregivers who are away from infants whenever the infant cries due to health problems, stranger anxieties, hunger, insect bites and stings, skin irritations, or diaper rash [3], [4].

Under background sound-free, silent, or controlled home environments, accurate detection of various kinds of infant cry sound patterns can be achieved by using simple short-term energy (STE) or short-term magnitude (STM) features. However, the presence of different kinds of background sounds is unavoidable in home or indoor environments and thus accurate and reliable infant cry detection is still a challenging task due to the overlapping nature of spectral and temporal characteristics of the cry sounds and background sounds, including the vocal sounds and non-verbal vocal sounds (produced by human livings in-home) and non-vocal background sounds such as music, songs, TV shows and varieties of machine-induced or home-appliance induced sounds [5], [6]. In home environments, background sounds such as speech, music, songs, and fan sounds are the prominent sounds when the home is located in transportation vehicle-free environments, whereas vehicle and train sounds can be prominent when the home is located in a zone of vehicle movement or railway stations and tracks.

In addition, exploring lightweight cry sound detection methods is essential for reducing the overall computational costs and improving the battery life of the battery-operated cry sound analysis systems. Therefore, in this paper, we present a simple and accurate infant cry detection (Infant-Cry-Detect) method based on autocorrelation function (ACF) features and machine learning classifiers. To the best of our knowledge, the ACF features have not been explored for cry sound detection except for determining the period of a signal buried in noise [7]. The key contributions of this paper are summarized below.

- A lightweight feature extraction approach is presented using ACF to discriminate cry sounds from background sounds such as speech, music, songs, vehicle, train, and fan sounds using simple ACF features.
- The performance of six machine learning classifiers is studied using the ACF features.
- The training and testing databases are created with a wide variety of cry sound patterns and background sounds

TABLE I
EXISTING CRY SOUND DETECTION METHODS

Classifier	Feature	Database	Performance

GMM, HMM [2]	FFT-MFCC, WE-DCT, EMD-MFCC	NCDS	BER: 8.98% (FFT-MFCC-GMM);
			11.03% (EMD-MFCC-GMM);

1D CNN [3]	MFCC	Own Database	PR:98.78; RR:98.78; F1:98.77 (500 ms)
FFNN [3]	MFCC	Own Database	PR:98.39; RR:98.39; F1:98.38 (500 ms)
MC-SVM [3]	MFCC	Own Database	PR:97.86; RR:97.87; F1:97.86 (500 ms)
KNN [8]	MFCC	Own Dataset	PR: 92.9%, RR: 90.5%, ACC: 94.1%
MLP [8]	MFCC	Own Dataset	PR: 92.7%, RR: 90.48%, ACC: 94.5%
GCN [9]	Spectrogram Features	Baby Chillanto, Baby2020	ACC: 94.39%
DWS [10]	Mel Spectrogram	AudioSet	F1: 73.8%
SVM [4]	DSF + AF	Own Dataset	PR: 67.2%; RR: 55.2%; F1: 0.613
RNN [11]	LPC + MFCC	Baby Cry Reason Classification, Baby Cry Detection Dataset	ACC: 94%
ANN [11]	LPC + MFCC	Baby Cry Reason Classification, Baby Cry Detection Dataset	ACC: 72%
CNN [11]	LPC + MFCC	Baby Cry Reason Classification, Baby Cry Detection Dataset	ACC: 98.4%
LSTM [11]	LPC + MFCC	Baby Cry Reason Classification, Baby Cry Detection Dataset	ACC: 80%
SVM [12]	MFCC + Log-Mel Spectrogram + ZCR	Baby Chillanto	PR: 96.12%; RR: 93.05%; F1: 94.56%; ACC: 96.96%
Signal Processing Techniques [13]	Short-Time Energy, Short-Time Zero-Crossing	Own Dataset	Classification ACC: 70%
HMM [14]	MFCC, Deltas and Delta-deltas, Fundamental Frequency, Aperiodicity Features	Own Dataset	ACC: 89.2%
GMM [15]	Statistical Features	Donate a Cry	ACC: 81.27%

Note: GMM: Gaussian Mixture Model; HMM: Hidden Markov Model; EMD: Empirical Mode Decomposition; WE-DCT: Wavelet Energy-based DCT; NCDS: Newborn Cry-based Diagnostic System Database; CNN: convolutional Neural Networks; FFNN: Feed Forward Neural Network; SVM: Support Vector Machine; MC-SVM: Multi-Class Support Vector Machine; DNN: Deep Neural Network; GCN: Graph Convolutional Neural Network; KNN: K-Nearest Neighbors; MLP: Multi-Layer Perceptron; DWS: Depth-wise Seperable; RNN: Recurrent Neural Network; ANN: Artificial Neural Network; LSTM: Long Short-Term Memory Networks; MFCC: Mel-Frequency Cepstral Coefficients; DSF: Deep Spectrum features; AF: Acoustic Features; LPC: Linear-Predictive Coding; ZCR: Zero-Crossing Rate; PR: Precision Rate; RR: Recall Rate; F1: F1-Score; PR-AUC: Precision-Recall Area Under the Curve; ACC: Accuracy

to demonstrate the accuracy and robustness of the ACF features-based Infant-Cry-Detect method.

The rest of the paper is organized as follows. Section II presents the existing sound detection methods and their limitations. Section III presents ACF-based feature extraction with the description of ACF features and machine learning classifiers used in this study. Section IV presents evaluation results using untrained sounds with the performance of six Infant-Cry-Detect methods in terms of standard benchmark metrics. Finally, the conclusions are drawn in Section IV.

II. EXISTING METHODS

In this section, we present a summary of some of the most popular cry sound detection and audio classification methods, that used the Mel frequency cepstrum coefficient (MFCC) and spectrogram features. We further highlight the limitations of the existing traditional machine learning (ML) classifier-based methods and computationally expensive deep learning network (DNN) based methods by considering the resource constraints of affordable infant cry sound analysis systems. Existing cry detection methods are summarized in Table I. In past studies, most cry sound detection methods explored traditional ML classifiers or deep learning architectures with spectral image inputs including the short-time Fourier transform (STFT), spectrogram, Mel spectrogram, and time-frequency representation of sounds. Some of the key limitations of existing methods are summarized below.

- The cry sound datasets included only limited baby cry sound patterns or cry sounds recorded using one type of audio acquisition with dedicated microphone(s).
- The background datasets included only short-duration background sounds (like door knocking and opening, laughing) and only a few common background sounds such as fans, wind, speech, or music. However, other common sounds encountered in most living environments should also be considered to ensure the robustness of the algorithm in practical application scenarios such as locations of homes in roadway zones, near railway stations or track zones, or in the presence of TV sounds.
- Accuracy and robustness of existing methods were not tested using untrained cry sounds and background sounds. Further, failure cases of cry sound detection were not investigated for the specific background sounds.

In addition, resource constraints of affordable cry sound analysis systems need to be addressed because most DNN-based cry sound analysis systems demand high-speed processors due to the large size of the DNN model. Audio signals are processed with a high sampling rate of 16 kHz or 44.1 kHz, further increasing the processor requirements. In addition to the demands of high computational resources (high-speed processor and memory), continuous processing of audio signals can incur more energy consumption, demanding frequent device charging. Therefore, there is a need to explore lightweight cry sound detectors to maximize the longevity of battery-operated devices.

III. MATERIALS AND METHODS

The main objective of this paper is to explore lightweight cry sound detection by considering the limited processor speed, memory space, and battery capacity of affordable infant cry sound analysis systems. By analyzing the characteristics

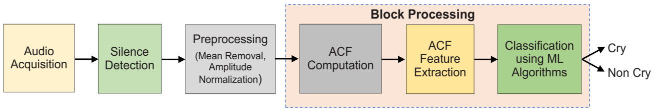

Fig. 1. Block diagram of the infant cry detection method using ACF features.

of sounds using ACF, we present a set of ACF features that can discriminate cry sounds from background sounds. With a focus on resource-efficiency improvement, we investigate the performance of the Infant-Cry-Detect method using six ML classifiers. A block diagram of the Infant-Cry-Detect method is shown in Fig. 1. The major steps of the method are silence detection, preprocessing, ACF feature extraction, and classification. The algorithm of the ACF feature-based infant cry detection is presented in Algorithm 1 with mathematical expressions, ACF features, and decision rules for silence detection before feature extraction. We briefly describe each of the steps and their significance in the subsequent subsections.

A. Preprocessing- Mean Removal & Amplitude Normalization

Mean removal and amplitude normalization are essential for improving the robustness of systems under different kinds of mean-shifted audio and varying amplitude ranges [16], [17]. For a discrete-time signal $x[n]$ with a length of N samples, the zero-mean signal $x_1[n]$ is computed as

$$x_1[n] = x[n] - \frac{1}{N}\sum_{i=0}^{N-1} x[i]. \tag{1}$$

The amplitude normalization is performed as

$$x_2[n] = \frac{x_1[n]}{\max_{i=0}^{N-1}|x_1[i]|}, \tag{2}$$

where $x_2[n]$ denotes the mean-removed, amplitude-normalized audio signal, which will be divided into overlapping blocks for detecting cry sounds in a 1-second audio signal.

In this study, each 1-second processed audio segment is divided into blocks with a block length of 100 ms and an overlap of 20 ms for capturing time-varying characteristics of sounds. Before performing feature extraction, random noise with an amplitude level of 0.2 is added to the audio segment to achieve better decorrelation under the presence of slowly varying components in the audio signal, which is one of the key contributions of this paper.

B. Autocorrelation Function (ACF) Based Features

In signal analysis applications, the ACF is widely used for finding the fundamental periods of signals such as speech, photoplethysmogram (PPG), music, etc. [7], [17]. ACF reflects the self-similarity of the signal [17]. For a discrete-time signal with length of N samples, the ACF r_{xx} is computed as [17]

$$r_{xx}[k] = \sum_{n=0}^{N-k-1} \frac{x[n]x[n+k]}{N} \quad k = 0, 1, ..., N-1, \tag{3}$$

Algorithm 1 ACF-Based Infant Cry Detection

```
Input: Audio signal x of 1 second duration
Output: Infant Cry, Non-infant Cry or Silence
Step-00: Read audio signal x and resample it to a uniform sampling rate of $F_s = 16000$ Hz.
Step-01: Silence detection
if $\max(x) < 0.05$ then
    Detect: "Silence"
end if
Step-02: Perform mean subtraction and amplitude normalization
Step-03: Add random noise $w[n]$ with amplitude of 0.2 to the normalized audio $x_2[n]$ to obtain noisy audio, $x_3[n] = x_2[n] + 0.2w[n]$.
Step-04: Process the 1 second audio signal $x_3[n]$. Divide the signal into blocks with a block duration of 100 ms and an overlap of 20 ms. Then, check the silence block using the maximum amplitude threshold criterion as defined in step-04(a).
Step-04(a):
if $\max(x_b[m]) < 0.1$ then
    Flag = 0 and NB=NB-1, where NB=12 (number of non-zero blocks)
end if
Step-04(b): Find the number of zero-crossings in the signal $x_b[m]$, $m = 0, 1, 2, ....M - 1$.
Step-04(c): Compute the partial ACF sequence $r_{xx}[k]$ for the lag number from 0 to $N/4$, using equation (3).
Step-04(d): Normalize the ACF sequence $r_{xx}[k]$ from -1 to 1.
Step-04(e): Find the first zero-crossing point (FZCP), the maximum amplitude ($r_{\max}$) and maximum amplitude location ($k_{\max}$).
Step-04(f): Determine the distance between the first and second positive zero-crossings of the ACF function ($\mathrm{ZCP}_{12}$).
Step-04(g): Compute the decay rate using equation (4).
Step-04(h): Using the different trained ML models generated with ACF features, predict whether the input sound block is a crying sound or not.
if $x \in$ "InfantCry" then
    Flag = 1
else
    Flag = 0
end if
Step-05: Detect as Silence Segment
if NB $< 4$ then
    Detect: "Silence"
end if
Step-06: Detect as Cry Sound or Non-Cry Sound
if $\mathrm{sum(flag)} < \mathrm{NB}/2$ then
    Detect: "Non-Cry Sound"
else if $\mathrm{sum(flag)} >= \mathrm{NB}/2$ then
    Detect: "Cry Sound"
end if
```

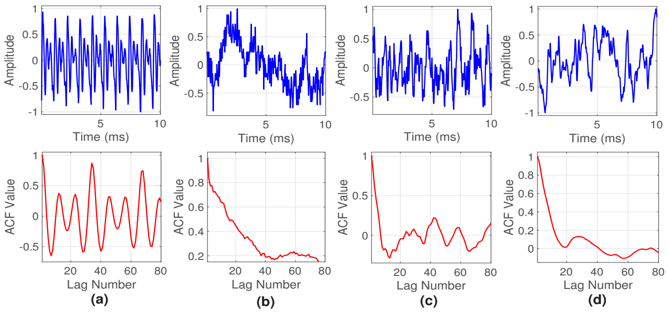

Fig. 2. Illustrates different audio signals and their ACF for (a) infant cry, (b) speech, (c) vehicle sound, and (d) fan sound.

where $r_{xx}[k]$ denotes the autocorrelation sequence and k denotes autocorrelation lag number [17]. The ACF function shows clear differences between periodic nature sounds and non-periodic sounds. The maximum energy of the ACF function is concentrated near the zero-lag point [17]. The ACF function for periodic signals is a slow, monotonically decreasing function whereas it rapidly decreases to zero for

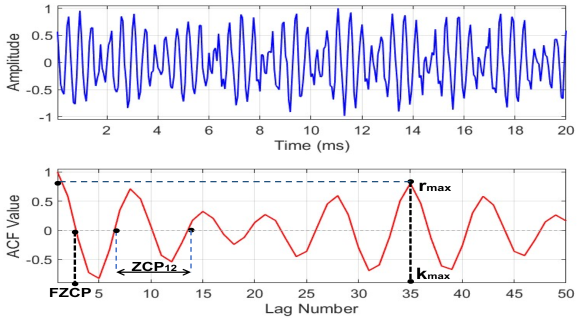

Fig. 3. Infant cry sound and corresponding ACF with essential features such as first zero-crossing point (FZCP), maximum peak amplitude ($r_{\max}$), lag of maximum peak (k_{max}), and distance between first and second positive zero-crossing points (ZCP_{12}).

high-noise signals [17]. Fig. 2 shows the ACF for infant cry sound, speech, vehicle sound, and fan sound. From visual inspection, it can be observed that ACF characteristics are different for different kinds of sounds. Hence, different features can be extracted from the ACF to discriminate different kinds of sounds. The various features extracted from the normalized ACF sequence are marked in Fig. 3. The extracted ACF features are the following.

- **FZCP:** Lag number corresponding to the first zero-crossing point (FZCP) of ACF sequence $r_{xx}[k]$. FZCP reflects the randomness of sounds, i.e., close to zero indicates a highly random signal.
- **AMP_MAX** ($r_{\max}$): Local maximum amplitude of the ACF sequence $r_{xx}[k]$, where k is defined from the first zero-crossing point to the end of the ACF sequence. $r_{\max}$ reflects the redundancy of successive samples, i.e., low amplitude indicates a periodic signal with noises.
- **MAXAMP_LOC** ($k_{\max}$): Lag number corresponding to the maximum amplitude of the ACF sequence $r_{xx}[k]$. $k_{\max}$ reflects the periodicity of the signal.
- **ZCP**$_{12}$: The distance between the first and second positive zero-crossings of the ACF sequence $r_{xx}[k]$.
- **DR**: The decay rate (DR) computed as [17]

$$\mathrm{DR} = \frac{\sum_{i=1}^{k_{\max}} r_{xx}[i]^2}{\sum_{i=1}^{L/4} r_{xx}[i]^2}, \tag{4}$$

  where L is the length of the ACF.
- **NZC:** The number of zero-crossings of the audio signal that helps to discriminate high-frequency sounds from cry-sounds [7].

We investigate the performance of the above-mentioned ACF features for discriminating cry sounds from non-cry sounds.

C. Silence Detection for Improving Energy Efficiency

Infant cry detection is performed for every 1-second segment. In practice, the segment may contain silence in sound-free environments. Furthermore, some of the 1-second background sound may contain more silence portions (low sound activity) with a few high sound activity portions. In such cases, silence detection can not only improve detection accuracy but also reduce overall energy consumption in continuous cry sound monitoring applications. Therefore, we present two stages of silence detection with a global amplitude threshold of 0.05 for the entire 1-second signal and a local amplitude threshold of 0.1 for the 100 ms block signal. For the second stage of silence detection, the decision rule is made based on the maximum of blocks of the segment which satisfies the local threshold-based decision rule.

D. Machine Learning Classifiers

We investigate the performance of ACF features trained with six ML classifiers, namely, decision tree (DT), random forest (RF), Naive Bayes (NB), multi-layer perceptron (MLP), light gradient boosting machine (LGBM), and k-nearest neighbors (KNN) [18]. We obtained six trained models using the same cry and non-cry sound datasets. Infant cry detection is formulated as a binary classification problem: cry sounds and non-cry sounds.

IV. Results and Discussions

In this section, we present the performance of the autocorrelation feature-based infant cry detection method using six classifiers (DT, RF, NB, MLP, LGBM, and KNN) in terms of accuracy (ACC), sensitivity (SE), specificity (SP), processing time and model size on the wide variety of cry sound and non-cry sound patterns.

A. Database Description and Performance Metrics

In this study, seven sound classes such as infant cry, speech, music, songs, fan, train, and vehicle sounds are considered which are encountered in indoor and outdoor environments [5], [16]. For creating the training and testing datasets, the infant cry sounds were collected from standard infant cry sound databases such as the ESC dataset [19] and Donate-a-cry corpus [20], and public multimedia websites such as YouTube. The clean speech signals were collected from publicly available databases such as the LJ Speech database [21], Indic TTS Malayalam Speech Corpus [22] and Microsoft Scalable Noisy Speech Dataset (MS-SNSD) [23]. The train and vehicle noises were partially collected from the iNoise Indian Noise database [24] which contained different outdoor and indoor sounds. The music, song, fan sounds, speech, vehicle, and train sounds were collected from YouTube sources. The training and validation databases contain different varieties and patterns of cry sounds and non-cry sounds that can be found in practical scenarios. The audio signal is resampled with a rate of 16000 Hz. The ML models are created using a training database with 10000 cry sounds and 624681 non-cry sounds. The test database consists of untrained 121208 1-second cry sounds and 150000 1-second non-cry sounds with 25000 sounds from speech, fan, music, song, vehicle, and train classes each.

The performance of the infant cry detection method is evaluated by using the following benchmark metrics.
The accuracy (ACC) is computed as [3]

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}}. \tag{5}$$

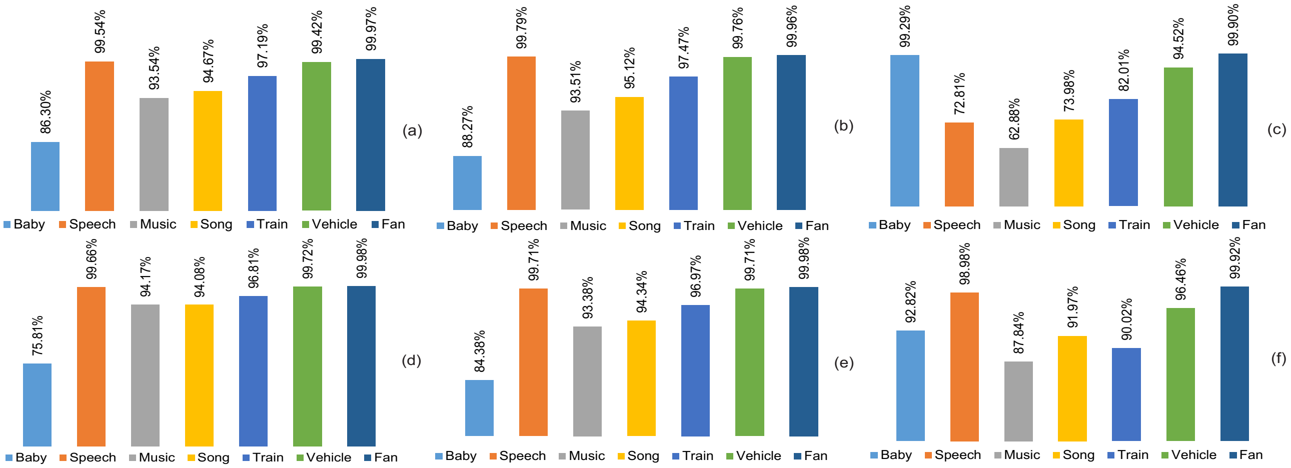

Fig. 4. Detection accuracy of different cry sounds and non-cry sounds for ML classifiers (a) DT, (b) RF, (c) NB, (d) MLP, (e) LGBM, and (f) KNN.

TABLE II
PERFORMANCE COMPARISON OF THE ML CLASSIFIERS ON THE UNSEEN DATASET WITH AND WITHOUT THE HANNING WINDOW

Classifier	**W/O Hanning Window**			**With Hanning Window**		
	SE (%)	SP (%)	ACC (%)	SE (%)	SP (%)	ACC (%)
DT	86.30	97.34	92.83	82.96	97.57	91.62
RF	88.27	97.56	93.77	85.18	97.75	92.63
NB	99.29	80.49	88.16	99.28	79.93	87.81
MLP	75.81	97.35	88.56	79.48	97.30	90.04
LGBM	84.38	97.30	92.03	81.83	97.63	91.19
KNN	92.82	94.08	93.56	91.84	94.26	93.27

The specificity (SP) is computed as

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}. \tag{6}$$

The recall rate or sensitivity (SE) is computed as [3]

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{7}$$

where the TP denotes the true positive (instances of correctly detecting cry sound), FN denotes the false negative (instances of wrongly detecting cry sound as non-cry), FP denotes the false positive (instances of wrongly detecting non-cry as cry sound) and TN denotes the true negative (instances of correctly detecting non-cry sounds).

B. Sound-Wise Detection Performance

For untrained sounds, the performances of the six ML models are shown in Fig. 4 in terms of detection accuracy of cry sounds and non-cry sounds. Among the six non-cry sounds considered in this study, the fan and vehicle sounds are detected with a detection accuracy above 99.9% and 94%, respectively, by all the six ML classifiers. The NB classifier had an exceptional performance in cry sound detection as compared to other classifiers with low detection accuracy for non-cry sounds. Meanwhile, the detection accuracy is low (62.88%) for music sounds as compared to the other classifiers.

C. Performance With and Without Hanning Window Function

In general, the Hanning function is used in most spectral features-based sound analyses to reduce the effect of boundary artifacts in spectral feature estimation [3]. Table II presents the performance of the six ML-based infant cry detection models with and without the Hanning window. Although the specificity of the models is the same before and after the application of the Hanning window, the sensitivity of the models is reduced except for the MLP classifier. Our study demonstrates that there is no improvement in detection performance with the use of the windowing function.

D. Performance Comparison - Six ML Models

Table II summarizes the performance of the six Infant-Cry-Detect methods. The testing was done on 96990 infant cry and 140800 non-cry sound segments. The NB-based method correctly classified 99.29% infant cry segments. However, the specificity of the NB-based method is 80.49%, which is comparatively less than other classifiers. Similarly, the RF-based method had an SP of 97.56%, which is the highest among the 6 classifiers but has an SE of 88.27% which is lesser than NB and KNN-based methods. Among the six classifiers, the KNN-based method had the best overall performance in terms of SE of 92.82% and SP of 94.08%. However, we have to evaluate the performance of the classifiers in terms of processing time and model size in addition to sensitivity and specificity.

E. Overall Performance of Six Classifiers

Table III summarizes the overall performance of the six trained models using the ACF features on the same test datasets. Results on untrained sounds show that the NB-based Infant-Cry-Detect method achieves a higher SE of 99.29% in accurately detecting a 1-second cry sound with lower processing time (1.11 ms) and model size (912 kB) as compared to the other five ML classifiers. Further, it is observed that the RF-based method outperforms in terms of the overall accuracy of 93.77% (SE=88.27% and SP=97.56%)

TABLE III
PERFORMANCE COMPARISON OF THE INFANT CRY DETECTION METHODS IN TERMS OF COMPUTATIONAL COMPLEXITY

Classifier	ACC (%)	Processing Time (ms)	Model Size (MB)
DT	92.83	1.071	4.00
RF	93.77	8.064	349.7
NB	88.16	1.111	0.000912
MLP	88.56	1.248	0.027
LGBM	92.03	1.770	0.343
KNN	93.56	2.831	36.5

with a higher computational cost (detection time of 8.064 ms and model size of 349.7 MB) as compared to the performance of the other five ML classifiers based methods. Among the six ML classifiers, the KNN-based cry sound detector provides promising overall performance in terms of SE of 92.82% and SP of 94.08% with comparable computational loads and processing time. Results demonstrate that the proposed ACF feature extraction not only achieves higher sensitivity in the presence of various background sounds but can also reduce overall computational loads as compared to computationally expensive deep learning networks.

V. CONCLUSION

In this paper, we presented six infant cry detection methods using ACF features. The performance of six ML classifiers was compared in terms of sensitivity, specificity, accuracy, processing time, and model size. Evaluation results showed that the RF-based cry detector outperforms the other five ML-based methods in terms of overall accuracy of 93.77%. Results further demonstrate that the DT-based method with an accuracy of 92.83% had a lower processing time of 1.07 ms as compared to the other methods. The NB-based method with an accuracy of 88.16% required the least memory space of 912 Bytes. It is also noted that the LGBM-based cry detection method had the best overall performance in terms of accuracy (92.03%), processing time (1.770 ms), and memory space (343 kB). Evaluation results showed that the ACF features can be capable of discriminating cry sounds from other sounds such as speech, music, song, vehicle, train, and fan sounds. The ACF feature-based infant cry detection methods are simple and fast compared to the computationally expensive DNN-based cry detection methods. In the future, we further study the performance of the ACF-based cry sound detection method with various other short burst sounds.

REFERENCES

[1] B. Lv, Y. Liu, S. Xu, and X. Shen, "Emotion recognition of infant cries using multi-scale CNN-BLSTM," in *8th Int. Conf. on Intelligent Computing and Signal Processing (ICSP)*, 2023, pp. 1659–1663.

[2] L. Abou-Abbas, C. Tadj, C. Gargour, and L. Montazeri, "Expiratory and inspiratory cries detection using different signals' decomposition techniques," *Journal of voice*, vol. 31, no. 2, pp. 259–e13, 2017.

[3] K. Manikanta, K. Soman, and M. S. Manikandan, "Deep learning based effective baby crying recognition method under indoor background sound environments," in *4th Int. Conf. on Computational Syst. & Information Tech. for Sustainable Solution (CSITSS)*, 2019, pp. 1–6.

[4] X. Yao, M. Micheletti, M. Johnson, E. Thomaz, and K. de Barbaro, "Infant crying detection in real-world environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 131–135.

[5] U. Ganapathi and M. S. Manikandan, "Convolutional neural network based sound recognition methods for detecting presence of amateur drones in unauthorized zones," in *Machine Learning, Image Processing, Network Security & Data Sciences*. Springer, 2020, pp. 229–244.

[6] A. Dayal *et al.*, "Lightweight deep convolutional neural network for background sound classification in speech signals," *The Journal of the Acoustical Society of America*, vol. 151, no. 4, pp. 2773–2786, 2022.

[7] P. Kathirvel, M. S. Manikandan, S. Senthilkumar, and K. P. Soman, "Noise robust zerocrossing rate computation for audio signal classification," in *3rd Int. Conf. on Trendz in Inf. Sciences & Computing (TISC2011)*, 2011, pp. 65–69.

[8] S. Cabon, B. Met-Montot, F. Porée, O. Rosec, A. Simon, and G. Carrault, "Automatic extraction of spontaneous cries of preterm newborns in neonatal intensive care units," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1200–1204.

[9] J. Chunyan, M. Chen, L. Bin, and Y. Pan, "Infant cry classification with graph convolutional networks," in *IEEE 6th Int. Conf. on Computer and Communication Systems (ICCCS)*, 2021, pp. 322–327.

[10] T. Khandelwal, R. K. Das, and E. S. Chng, "Is your baby fine at home? Baby cry sound detection in domestic environments," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 275–280.

[11] Z. Firas, A. A. Nashaat, and G. Ahmad, "Optimizing infant cry recognition: A fusion of LPC and MFCC features in deep learning models," in *7th Int. Conf. on Advances in Biomedical Engineering (ICABME)*, 2023, pp. 01–06.

[12] K. Zhang, H.-N. Ting, and Y.-M. Choo, "Baby cry recognition by BCRNet using transfer learning and deep feature fusion," *IEEE Access*, vol. 11, pp. 126 251–126 262, 2023.

[13] L. Liu, Y. Li, and K. Kuo, "Infant cry signal detection, pattern extraction and recognition," in *2018 International Conference on Information and Computer Technologies (ICICT)*, 2018, pp. 159–163.

[14] G. Naithani *et al.*, "Automatic segmentation of infant cry signals using hidden markov models," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 1, pp. 1–14, 2018.

[15] K. Sharma, C. Gupta, and S. Gupta, "Infant weeping calls decoder using statistical feature extraction and gaussian mixture models," in *10th Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2019, pp. 1–6.

[16] S. Soni, S. Dey, and M. S. Manikandan, "Automatic audio event recognition schemes for context-aware audio computing devices," in *7th Int. Conf. on Digital Inf. Processing & Communications (ICDIPC)*, 2019, pp. 23–28.

[17] C. Manoj, S. Magesh, A. S. Sankaran, and M. S. Manikandan, "Novel approach for detecting applause in continuous meeting speech," in *2011 3rd International Conference on Electronics Computer Technology*, vol. 3, 2011, pp. 182–186.

[18] S. Śmigiel, "ECG classification using orthogonal matching pursuit and machine learning," *Sensors*, vol. 22, no. 13, p. 4960, 2022.

[19] K. J. Piczak, "ESC: dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1015–1018.

[20] [Online]. Available: https://github.com/gveres/donateacry-corpus

[21] K. Ito and L. Johnson, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[22] N. Srivastava, R. Mukhopadhyay, P. K R, and C. V. Jawahar, "IndicSpeech: Text-to-speech corpus for Indian languages," in *Proc. 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020.

[23] C. K. A. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A scalable noisy speech dataset and online subjective test framework," 2019.

[24] S. K. Kopparapu and Others, "iNoise Indian noise database," 2020. [Online]. Available: https://dx.doi.org/10.21227/w3xm-jn45