Freeze and Learn using KAN for Infant Cry Classification

Arth J. Shah, Vishnu Vardhan G V S and Hemant A. Patil
Speech Research Lab @ DA-IICT, Gandhinagar, Gujarat, India
{202101154, 202411071, hemant_patil}@daiict.ac.in

***Abstract*—Advancements in technology have significantly benefited healthcare services, particularly using artificial intelligence (AI)-based methods. Understanding the reasons behind infant crying can improve early diagnosis and treatment. This study aims to develop a system that provides information about infants health condition by performing two types of classification: binary classification (Healthy *vs.* Pathological) and multi-class classification (four pathological and four healthy classes). Proposed approach uses acoustic cepstral features using Freeze and Learn approach via continual learning by employing LCNN and KAN-Linear network. The optimal proposed methodology achieved a highest accuracy of 92.76 % for binary classification and 86.91 % for eight-class classification outperforming the base model of LCNN. These results demonstrate the effectiveness of employing KAN in distinguishing between healthy and pathological infant cries, contributing to improved early diagnosis.**

***Index Terms*—KANLinear, continual learning, pathology, infant cry classification.**

I. Introduction

Infancy is the stage between birth and the development of language, during which infants rely on crying, facial expressions, and babbling to communicate [1]. Understanding and analyzing their cries is crucial, as studies indicate that newborns can differentiate between two languages within just four days of birth [2]. While biometric systems using infant fingerprints have been introduced to enhance security, they face significant challenges. Infants often keep their fists closed or suck on their fingers, making fingerprint acquisition difficult due to moisture and movement [3]. An important concern during early infancy is the high risk of undiagnosed medical conditions, which can lead to severe health complications or even fatalities. Many newborns succumb to illnesses within the first few months of life due to a lack of early detection and treatment. One of the leading causes of infant mortality is Sudden Infant Death Syndrome (SIDS) [4], along with birth asphyxia and congenital disorders. Addressing these challenges through improved diagnostic techniques, early intervention, and better monitoring can significantly enhance infant care and survival rates. Gathering infant cry recordings is a challenging task, as many parents or guardians are reluctant to grant permission due to privacy concerns and apprehension about data collection [5]. Additionally, large-scale datasets for infant cry classification remain scarce due to issues, such as data imbalance, ethical considerations, and confidentiality concerns. Despite these limitations, there is a growing demand for improved methods to analyze infant cries, given the significance of this field [6].

Some of the existing works in infant cry analysis currently rely heavily on the availability, quality, and size of the dataset [7], [8]. In such cases, due to the challenges involved in collecting large amounts of clinically validated infant cry data, researchers resort to augmentation of their datasets using synthetic data which are then injected into the original samples to increase the size of dataset and balance class distributions leading to poor results and poor clinical insights.

A. Related Works

Acoustic analysis plays a crucial role in speech processing, enabling the identification of speech disorders, emotions, and pathologies. Infant cry classification leverages acoustic cues to detect abnormalities linked to neurological or physiological conditions, making it an essential tool for early diagnosis. Previous studies employing cepstral features like Mel Frequency Cepstral Coefficients (MFCC) [9], Linear Frequency Cepstral Coefficients (LFCC) [10], demonstrate the importance of temporal and frequency-based features in identifying cry patterns [11], [12]. Models such as Convolutional Neural Networks (CNN) excels in handling subtle patterns due its learnable activation functions instead of integer weights, making it highly effective for infant cry classification [5]. However, existing methods limit their ability to analyze spectral variations and also often overlook or miss crucial variations in spectral energy distribution [13]. Infant cry analysis has long been studied as a non-invasive biomarker for detecting neurological and physiological conditions in newborns. Traditional methods have focused on fundamental frequency, pitch, formant analysis, and cepstral coefficients like MFCC and LFCC [11]. In this paper, we employ the continual learning approach in the context of infant cry classification and investigate the effective approach of freezing one and enabling the other and vice versa when training the model. The proposed method introduces continual learning method called Freeze and Learn using Light CNN (LCNN) [14] and Kalmogorov Arnold Networks (KAN) [15] where the cepstral features are given as input to two functional blocks namely Encoder and Classifier.

II. Proposed Methodology

Continual Learning techniques allow models to learn from unseen data without losing what they have already learned.
In this paper, we investigate and findout when it is more

979-8-3315-7206-8/25/$31.00

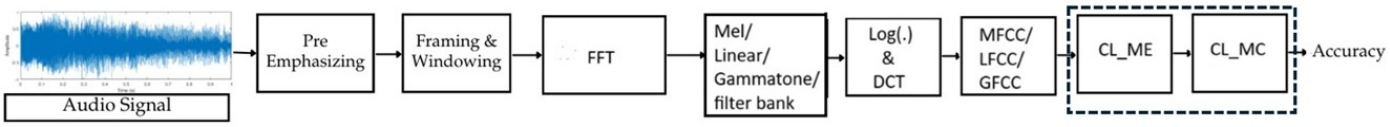

Fig. 1. Functional Block Diagram

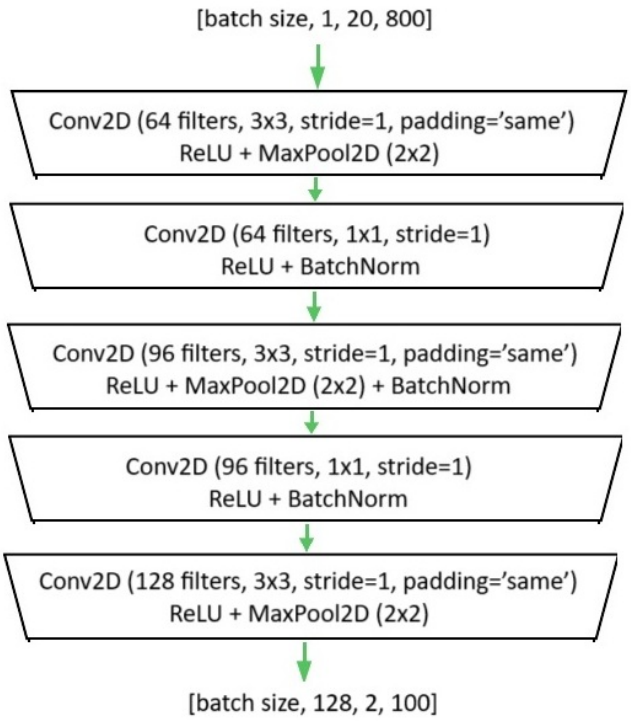

Fig. 2. Encoder Architecture (LCNN)

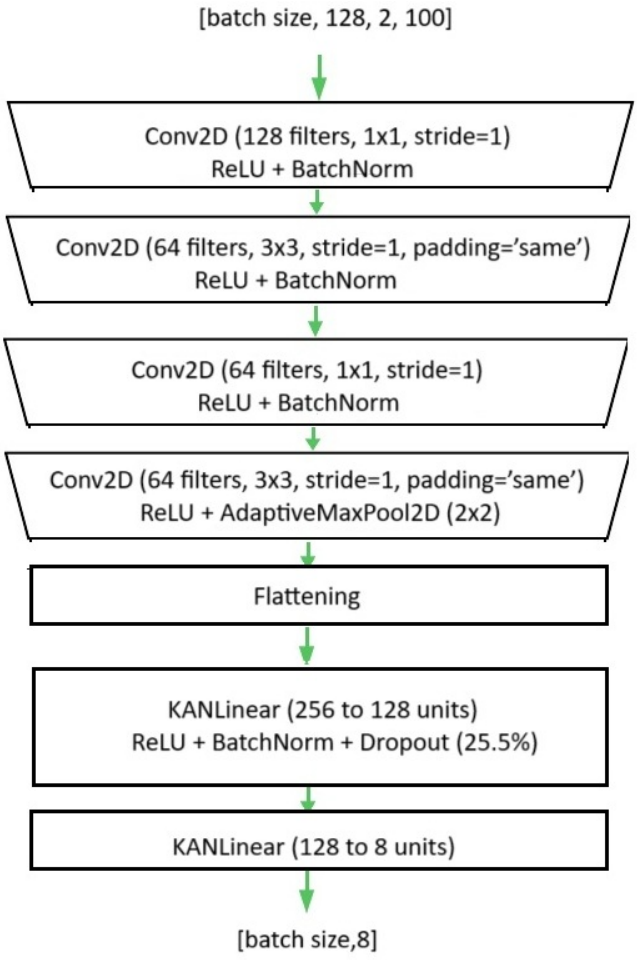

Fig. 3. Classifier Architecture (LCNN + KANLinear)

effective to apply continual learning across the entire model or to update only some layers while freezing others. Traditional methods on infant cry classification relied heavily on computational requirements and as an alternative, pre-trained models are taken and fine-tuned on novel datasets incorporating transfer learning. Nevertheless, these methods lead to the concept of catastrophic forgetting where models unlearn previously learned data. Therefore, we in this study aim to explore the aspect of whether it is beneficial to use a continual learning strategy which updates all layers or selectively retrain only a few of them aiming to understand and achieve optimality by balancing model's performance when exposed to new data.

We do this by taking two detectors for detecting the type of cry pattern (in LCNN classifier) and dividing into 2 logical sections: Module Encoder (ME) and Module Classifier (MC). MC consists of KANLinear layer instead of normal Linear layer. Using three approaches: Updating all model weights, updating only 1 section while freezing other and vice versa. Finally, we compare the performances of these configuration.

A. Intuition of the approach

Freeze and Learn approach is a kind of continual learning approach where we selectively freeze one section of model and estimate the performance of another and vice versa. This method is employed to address the issue of catastrophic forgetting in Multi-Layer Perceptron (MLP) and neural networks in which the model learns new current data and forgets what it had learned in the past, thereby unable to preserve long-term dependency and long-term learned data. We believe that this method is beneficial not only for continual learning tasks but also in transfer learning and domain adaptation

scenarios where maintaining balance between past and present knowledge is needed.

B. Workflow Setup

Consider a discrete time cry signal sampled at 16kHz frequency and belonging to class y = 0,1,2,3,4,5,6,7 where 0 means Asphyxia, 1 means Deaf, 2 means Hyperbilirubinemia, 3 means Hypothyroidism, 4 means Hypothyroidism, 5 means Hunger, 6 means Pain, 7 means shower. The detector is divided into two modules:
Encoder (CL_ME): The detector is trained on dataset which updates the weights of ME while keeping those of MC frozen.
Classifier (CL_MC): The detector is trained on the dataset which updates the weights of MC and freezing those of ME.
The detector is chosen to be **LCNN** (Light CNN) model which operates on cepstral features (MFCC, LFCC, GFCC) extracted from input audio and integrates several convolution layers with varying kernel sizes and strides to capture different levels of abstraction and enhance feature extraction. ME consists of 5 convolutional layers and MC consists of 4 convolutional layers along with subsequent KANLinear layer.

C. KANLinear Layer

In this paper, we have replaced normal nn.Linear layer in LCNN classifier achitecture with KANLinear and it works based on the priniciple of KART.
Kolmogorov-Arnold Representation Theorem (KART) [16], [17]: It states that if f is a multivariate continuous function on a bounded domain, then f can be written as a finite composition of continuous functions of a single variable and the binary operation of addition[16]. More specifically, for a smooth $f : [0, 1] \to R$,

$$f(x) = \sum_{q=1}^{2k+1} \Phi_q \left(\sum_{p=1}^{n} \Phi_{q,p}(x_p) \right), \tag{1}$$

where $\Phi_{q,p} : [0, 1] \to R, and \quad \Phi_q : R \to R$. Here, the multivariate function is addition and every other function is written using univariate functions and summed up. Replacing dense layers in LCNN with KANLinear offers merits in terms of interpretability, efficiency, and robustness. KANLinear provides an interpretable mapping, enabling researchers to analyze feature transformations explicitly. Additionally, Linear layer may introduce potential overfitting. KANLinear mitigates these issues by leveraging a structured, low rank functional decomposition, reducing the number of learnable parameters while maintaining expressivity.

KANLinear Layer is designed to learn nonlinear relationships in features via PReLU and enhances classification performance by making weight transformations adaptive and learnable and to implement it we considered by taking the base activation function to be PReLU, grid size of 10 and spline order as 4 for optimal performance. The number of input dimensions is 128 which come from the previous convolutional layers. For initializing the base weights, we used Xavier uniform initialization to balance activation variances to ensure stable training.
The KANLinear output is computed as:

$$\text{Output} = (\text{base weight} \times \text{activation}) + (\text{spline weight} \times \text{B-spline basis})$$

,where
Base Weight: Standard linear transformation.
Spline Weight: A learnable function over grid points which approximates smooth nonlinear mappings.
The number of B-spline basis functions equals to grid size + spline order.
In this case the number of B-Spline functions will be **14** (Grid Size = **10** and Spline Order = **4**).
Resulting output shape: (batch,out_features).

III. Experimental Setup

A. Datasets used

For this study, we use Baby Chillanto 2.0 [18], which is a property of the Instituto Nacional de Astrofisica Optica y Electronica – CONACYT, Mexico, the latest infant cry classification dataset, chosen for its extensive collection of labeled infant cry samples covering multiple health conditions. Unlike older datasets, Baby Chillanto 2 provides a diverse range of cry types, making it ideal for training robust classification models. This dataset is particularly significant due to its up-to-date cry patterns, diverse categories—including both healthy and pathological cries—and real-world applicability in early diagnosis of infant health conditions. It consists of eight classes: normal, deaf, hunger, shower, asphyxia, hyperbilirubinemia, hypothyroidism, and pain, each representing a distinct infant cry condition. The dataset is organized into subfolders, each containing '.wav' files corresponding to a specific class. These raw audio files are sampled at 16 kHz, varying in length to reflect real-world variability in cry duration and intensity [19]. Additionally, the dataset supports entropy-based feature extraction across different frequency bands, enabling more precise analysis of infant cries. Since the size of the dataset is relatively small, so we have segmented every .wav file for 2 second duration to increase the size of dataset as shown in Table 3.

TABLE I
NUMBER OF FILES IN DATASET (BEFORE - AFTER) SEGMENTATION

	Train	**Test**	**Val**
0_Asphyxia	4 - 114	1 - 27	1 - 33
1_Deaf	36 - 355	9 - 82	7 - 63
2_Hyperbilirubinemia	6 - 167	2 - 41	1 - 8
3_Hypothyroidism	32 - 271	8 - 109	7 - 99
4_Hunger	23 - 132	6 - 31	5 - 36
5_Normal	14 - 184	4 - 43	3 - 46
6_Pain	17 - 87	5 - 21	3 - 9
7_Shower	2 - 15	1 - 5	1 - 5

B. Training Setup

We trained all models for 20 epochs by monitoring validation loss computed by calculating cross entropy function with batch size of 64 and Adam optimizer with learning rate of 0.003. The data is being tested to do binary classification and 8-class classification where in binary classification all the abnormal classes like Asphyxia, Deaf, Hyperbilirubinemia, Hypothyroidism are clubbed into one class named **Pathological** and normal classes like Normal, Hunger, Shower, Pain are clubbed into another class named **Healthy**.

C. Comparative Features Used

In this study, we extracted 20 cepstral coefficients using a 25ms frame size with a 10ms overlap. Hamming window is applied for windowing to minimize spectral leakage.

1) Mel Frequency Cepstral Coefficients (MFCC): MFCC is used to replicate human auditory perception by applying the Mel scale, which is non-linearly spaced—denser at lower frequencies and more spread out at higher frequencies. It extracts vocal tract characteristics by applying a logarithmic filterbank, emphasizing speech phonetics over pitch information. This makes MFCC particularly useful for analyzing infant cry patterns.

2) Linear Frequency Cepstral Coefficients (LFCC): LFCC utilizes linearly spaced filters instead of the Mel scale, ensuring equal emphasis on both low- and high-frequency components. It captures detailed spectral information based on raw signal characteristics rather than human auditory perception, making it especially beneficial for pathology detection in infant cry abnormalities. Additionally, LFCC is less affected by nonlinear distortions in speech, allowing it to effectively analyze subtle spectral variations, which is crucial for medical and diagnostic applications.

3) Gammatone Frequency Cepstral Coefficients (GFCC): GFCC employs gammatone filterbanks that mimic the human cochlear response, allowing it to capture both formant and pitch information, unlike MFCC. It is particularly robust to noise, making it highly effective for speech enhancement, speaker recognition, and infant cry classification. GFCC excels at identifying fine-grained spectral details, especially in noisy environments, making it a valuable feature extraction method for challenging acoustic conditions.

IV. Experimental Results

From the experimental results, we can see that adding KANLinear layer clearly provides better results when compared to normal LCNN classifier having normal Linear layer. It is so because KANs avoid catastrophic forgetting by using spline functions' locality and thus, only a sample of a few nearby spline coefficients will get affected, leaving far-away coefficients intact [16]. In contrast, for LCNN, which employs activations, such as ReLU/SiLU, any local change can disrupt the information stored in distant regions.

MFCC consistently outperforms both LFCC and GFCC across most configurations, achieving the highest binary test accuracy of 0.9276 under the MC condition and the highest 8-class test accuracy of 0.8468 under the ME condition. Comparing conditions, the MC setup yields the best binary test accuracy overall (0.9276 with MFCC), while the ME setup leads in 8-class accuracy (0.8468 with MFCC), implying that different conditions slightly affect performance, with ME being marginally better for multi-class tasks. MFCC emerges as the most reliable feature set for both binary and multi-class classification using LCNN on 2-second audio segments, while LFCC remains a solid alternative with better EER performance.

The classifier module (MC) where only LCNN is used consists of Linear layer. In both cases, encoder module's (ME) is not varied. Results reveal that the proposed LCNN+KANLinear model outperformed a conventional LCNN, achieving higher classification accuracy. This performance gain is attributed to the model's ability to learn data-specific activation behaviors through KANLinear, capturing intricate and variable patterns in cry signals more effectively than static activation functions.

A. Delta and Double - Delta Features

These features symbolize and capture the dynamic characteristic properties which is the rate of change of signal over time and are derived from the static cepstral features which include MFCC, LFCC, GFCC. Delta features are typically calculated as the difference between consecutive static features as mentioned and are approximated to the first and second derivatives of those static features. These are calculated to provide temporal context and improve model performance.
The results for delta and double delta features, as shown for LCNN with 2-second audio segments, show higher testing accuracies, particularly with MFCC, achieving up to 0.9081 in the MC condition for binary classification. This indicates that the temporal dynamics provided by delta features are highly informative, enabling the model to better distinguish patterns in the audio data. However, the accuracy drops notably in 8-class classification, suggesting that the increased complexity of multiple classes challenges the model's ability to leverage these dynamics effectively.
In contrast, double delta features, which represent the second-order acceleration of feature changes, exhibit a slight trade-off in accuracy compared to delta features but offer improved equal error rates (EER), with values as low as 0.4402 for MFCC in ME. This improvement in EER suggests that double delta features enhance the separation between genuine and impostor scores, making them valuable for tasks prioritizing reliability over raw accuracy, such as speaker verification. Across feature sets, MFCC continues to outperform LFCC and GFCC, while conditions like ME show better results than MC or ALL, likely due to reduced variability. Overall, these

results underscore the complementary roles of delta and double delta features, with delta excelling in accuracy and double delta contributing to robustness, particularly when paired with MFCC.

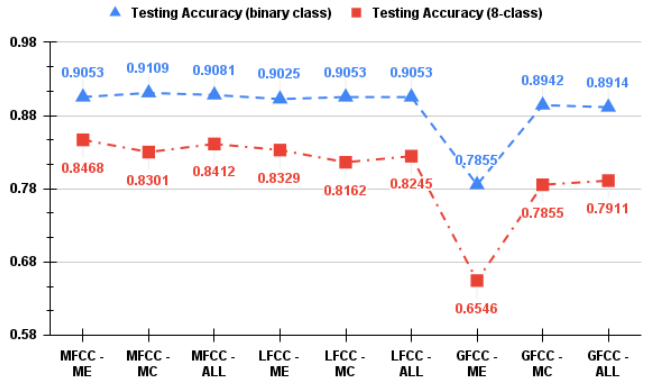

Fig. 4. Obtained results using LCNN

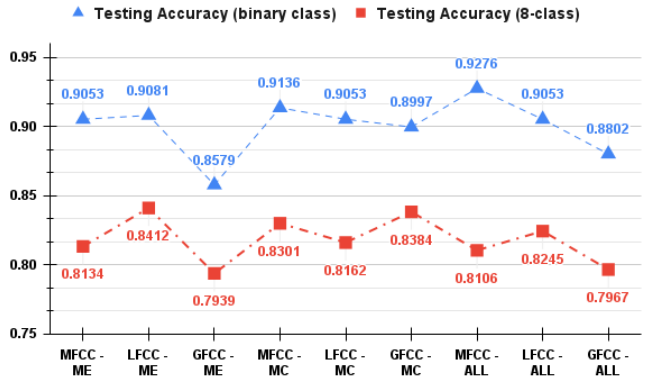

Fig. 5. Obtained results using LCNN KANLinear combination

V. Summary and Conclusions

The proposed methodology introduces a hybrid model combining a Lightweight Convolutional Neural Network (LCNN) with Kolmogorov-Arnold Network Linear (KANLinear) layers for infant cry classification, employing a Freeze and Learn strategy. In this setup, the early convolutional layers of LCNN, responsible for capturing fundamental acoustic features, are frozen, while the higher-level KANLinear layers are trained to adaptively learn complex, task-specific representations. KANLinear replaces traditional fixed activation functions with learnable, spline-based functions at each neuron, allowing the model to dynamically shape its non-linear mappings based on the data. This approach addresses the highly variable, non-stationary nature of infant cries, which often challenge conventional deep learning models in small dataset scenarios. Experimental analysis demonstrated that the proposed LCNN

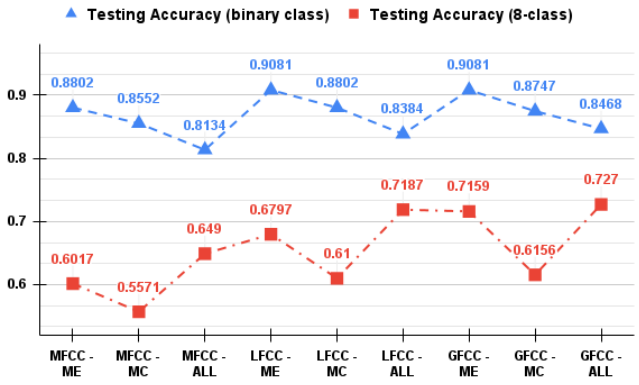

Fig. 6. Obtained delta results for LCNN

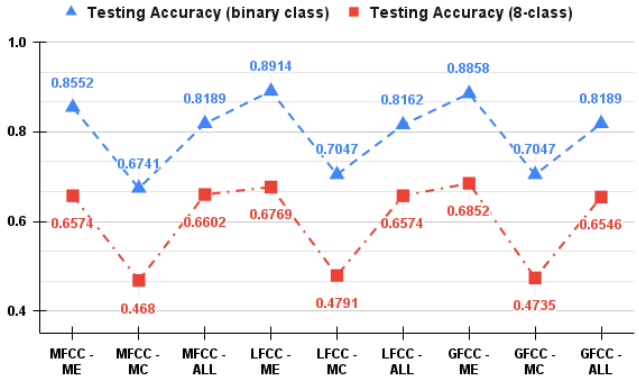

Fig. 7. Obtained double-delta results for LCNN

combined with KANLinear network achieved superior accuracy compared to a standard LCNN architecture. The improvement is attributed to KAN's ability to learn flexible, data-specific activation functions that capture intricate patterns in cry signals more effectively than static functions like ReLU. Future work involves extending this adaptability to the convolutional layers themselves by integrating learnable activation functions at each convolution block, potentially enhancing the model's feature extraction capacity. Additionally, understanding the theoretical properties of KAN in tasks like infant cry classification and broader machine learning contexts remains an open research question. Investigating aspects such as KAN's generalization behavior, optimization dynamics, and interpretability in time-varying audio signals could significantly contribute to both the application domain and the theoretical foundations of adaptive neural networks.

VI. Future Work

This approach can be extended by integrating different variants of KAN which include BSRBF-KAN, Chebyshev-KAN, Wav-KAN etc. Also, employing entropy-based approach to capture randomness and complexity of crying patterns such that more informative and discriminative features can

be extracted. Additionally, employing different variants of the Kolmogorov-Arnold Network (KAN) by tweaking the hyperparameters which could enhance the model's flexibility and learning capacity like varying adaptive grid sizes, spline orders, or layer-wise activation learning within KAN variants may lead to improved performance and a deeper understanding of the underlying data dynamics.

VII. Acknowledgment

We like to thank Dr. Emilio Arch-Tirado and his INR-Mexico group and Dr. Carlos A. Reyes-Garcia, for their dedication to the collection and processing of the Infant Cry data base.

References

[1] P. D. Eimas, E. R. Siqueland, P. Jusczyk, and J. Vigorito, "Speech perception in infants," *Science*, vol. 171, no. 3968, pp. 303–306, 1971.

[2] J. Mehler, P. Jusczyk, G. Lambertz, N. Halsted, J. Bertoncini, and C. Amiel-Tison, "A precursor of language acquisition in young infants," *Cognition*, vol. 29, no. 2, pp. 143–178, 1988.

[3] J. J. Engelsma, D. Deb, K. Cao, A. Bhatnagar, P. S. Sudhish, and A. K. Jain, "Infant-id: Fingerprints for global good," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3543–3559, 2021.

[4] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[5] A. J. Shah, H. Chaudhari, and H. A. Patil, "Infant cry classification using modified group delay cepstral coefficients," in *International Conference on Pattern Recognition (ICPR)*, Springer, 2024, Kolkata, India, pp. 275–289.

[6] Reyes-Galaviz, Orion Fausto and Cano-Ortiz, Sergio Daniel and Reyes-García, Carlos Alberto, "Validation of the cry unit as primary element for cry analysis using an evolutionary-neural approach," in *2008 Mexican International Conference on Computer Science, Baja California, Mexico,*, 2008, pp. 261–267.

[7] L. Le, A. N. M. Kabir, C. Ji, S. Basodi, and Y. Pan, "Using transfer learning, svm, and ensemble classification to classify baby cries based on their spectrogram images," in *2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW)*, 2019, Monterey, CA, USA, pp. 106–110.

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

[9] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[10] X. Zhao and D. Wang, "Analyzing noise robustness of mfcc and gfcc features in speaker identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, Vancouver Canada, pp. 7204–7208.

[11] S. P. Dewi, A. L. Prasasti, and B. Irawan, "The study of baby crying analysis using MFCC and LFCC in different classification methods," in *2019 IEEE International Conference on Signals and Systems (ICSigSys)*, 2019, Bandung, Indonesia, pp. 18–23.

[12] A. Abbaskhah, H. Sedighi, and H. Marvi, "Infant cry classification by MFCC feature extraction with MLP and CNN structures," *Biomedical Signal Processing and Control*, vol. 86, pp. 105–261, 2023.

[13] A. M. Toh, R. Togneri, and S. Nordholm, "Spectral entropy as speech features for speech recognition," *Proceedings of PEECS*, vol. 1, p. 92, 2005.

[14] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE transactions on information forensics and security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[15] Z. Liu, Y. Wang, S. Vaidya, *et al.*, "KAN: kolmogorov-arnold networks," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, OpenReview.net, 2025. [Online]. Available: https://openreview.net/forum?id=Ozo7qJ5vZi.

[16] Z. Liu, Y. Wang, S. Vaidya, *et al.*, "Kan: Kolmogorov-arnold networks," 2024.

[17] V. I. Arnold, "On the representation of functions of several variables as a superposition of functions of a smaller number of variables," *Collected works: Representations of functions, celestial mechanics and KAM theory, 1957–1965*, pp. 25–46, 2009.

[18] O. F. Reyes-Galaviz, S. D. Cano-Ortiz, and C. A. Reyes-García, "Evolutionary-neural system to classify infant cry units for pathologies identification in recently born babies," in *2008 7^{th} Mexican International Conference on Artificial Intelligence*, 2008, Cambridge, England, pp. 330–335.

[19] H. Chaudhari, A. J. Shah, and H. A. Patil, "Cross lingual speech representation for infant cry classification," in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2024, Macau, China, pp. 1–5.