Infant Vocal Signal Manifold Structuring through Temporal Convolutional Acoustic Encoding

M. Ramana Kumar[1*], K. Sunil Kumar[2*], Sriramula Geetha[3], Jangiti Srivani[3], Patha Srikar[3]

[1]Associate Professor, [2]Assistant Professor, [3]UG Student, [1,2,3]Department of Computer Science and Engineering

[1,2,3]Kommuri Pratap Reddy Institute of Technology, Ghanpur, Ghatkesar, 501301, Telangana, India.

*Correspondence: M. Ramana Kumar (ramana.reah@gmail.com), K. Sunil Kumar (suneelkumaa20@gmail.com)

ABSTRACT

Interpreting infant cries is a critical yet challenging task, as babies lack the ability to communicate their needs verbally. Traditionally, caregivers depend on personal experience, observation, and intuition to understand these cries; however, such methods are inherently subjective, often inconsistent, and can lead to misinterpretation of essential needs like hunger, pain, or discomfort. To overcome these limitations, the proposed system presents an intelligent and automated approach for baby cry classification using advanced audio signal processing techniques. The system processes recorded cry signals and extracts discriminative acoustic features through Mel-Frequency Cepstral Coefficients (MFCC), which effectively capture the underlying frequency patterns of audio. These extracted features are then utilized to train and evaluate multiple machine learning models, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree Classifier (DTC), AdaBoost (ADB), and Linear Discriminant Analysis (LDA), enabling a comprehensive comparative analysis of their performance. To further improve classification accuracy and robustness, a Convolutional Neural Network (CNN) is employed as the primary model, leveraging its capability to automatically learn complex feature representations and temporal patterns within audio data. The system's effectiveness is measured using key performance metrics such as accuracy, precision, recall, and F1-score, ensuring reliable evaluation. Additionally, the entire framework is integrated into a user-friendly interface that seamlessly combines feature extraction, model training, evaluation, and real-time prediction. This end-to-end solution provides a consistent, efficient, and accurate method for identifying infant cry types, thereby assisting caregivers in making timely and informed decisions for better infant care.

Keywords: Neonatal Signal Analysis, Pediatric Healthcare Systems, Acoustic Signal Processing, Feature Learning, Clinical Decision Support

1. INTRODUCTION

Babies cry as a natural and instinctive means of expressing their needs, including hunger, discomfort, and lack of sleep. Accurately and promptly interpreting these cries is essential for ensuring an infant's healthy development and overall well-being [1]. However, inexperienced parents, new caregivers, and even some healthcare professionals often find it challenging to correctly identify the underlying reasons for crying. This difficulty can result in prolonged distress, elevated stress levels for both the infant and caregiver, and potential negative effects on cognitive and emotional development [2]. In recent years, research on the automatic classification of infant crying sounds has gained significant attention, offering promising solutions to this problem. These studies primarily focus on analyzing and classifying cry signals using advanced audio signal processing techniques combined with machine learning algorithms.

This literature review provides a comprehensive overview of key contributions in this domain, highlighting various approaches for the detection and classification of infant cries, as illustrated in Fig 1. It examines the methodologies adopted in previous studies, including the types of datasets used, feature extraction techniques such as MFCC, and the application of diverse classification models. The existing body of research demonstrates the effectiveness of integrating sound processing methods with machine learning to develop reliable and automated systems for infant cry analysis.

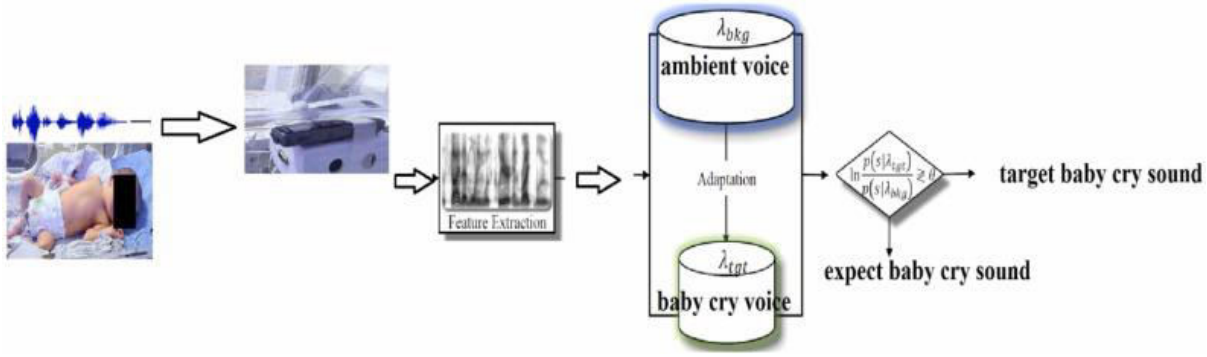

Fig. 1. Automated detection of infant cry.

Infants primarily communicate through crying, which serves as a natural and essential mechanism for expressing their needs and discomfort. Since babies are unable to convey their feelings verbally, parents and caregivers are often required to interpret these cries through observation and intuition, which may not always be accurate [3]. This uncertainty can result in delayed responses to the infant's needs, leading to increased stress for caregivers and prolonged discomfort for the child. To address these challenges, technological advancements have introduced intelligent solutions aimed at improving the accuracy of cry interpretation [4]. In the context, it becomes crucial to differentiate between various types of cries based on their distinct acoustic characteristics. By analyzing these auditory features, it is possible to better understand the underlying needs of the infant and provide appropriate and timely care [5]. Such systems not only enhance the effectiveness of infant care but also help reduce stress and anxiety among parents and caregivers by minimizing the chances of misinterpretation.

The improper segregation and disposal of waste have become a critical environmental issue in both developing and developed nations, especially in India, where urban areas generate more than 62 million tons of solid waste annually, and only a fraction is efficiently recycled. Manual waste sorting methods are inefficient, labor-intensive, and prone to human error, leading to contamination and low recycling rates. The challenge lies in developing an automated, intelligent system capable of identifying and categorizing waste materials accurately in real time. Thus, the problem is to design and implement a deep learning–based classification system that can automatically recognize and classify waste into appropriate categories using image data.

2. LITERATURE SURVEY

Gul et al. [6] constructed a large dataset of patient videos by labeling each frame with a set of patient actions and the patient's positions. They retrained the back-bone CNN model with 23,040 labeled images of patient's actions for 32 epochs. Across each frame, the proposed model allocated a unique confidence score and action label for video sequences by finding the recurrent action label. The present study shows that the accuracy of abnormal action recognition is 96.8%. They proposed approach differentiated abnormal actions with improved F1-Score of 89.2% which is higher than state-of-the-art techniques. Mohammed et al. [7] compiled convolutional neural network (CNN) methods which have the potential to automate the manual, costly and error-prone processing of medical images. They

attempted to provide a thorough survey of improved architectures, popular frameworks, activation functions, ensemble techniques, hyperparameter optimizations, performance metrics, relevant datasets and data preprocessing strategies that can be used to design robust CNN models. We also used machine learning algorithms for the statistical modeling of the current literature to uncover latent topics, method gaps, prevalent themes and potential future advancements. The statistical modeling results indicate a temporal shift in favor of improved CNN designs, such as a shift from the use of a CNN architecture to a CNN-transformer hybrid. Arooj et al. [8] evaluated on the public UCI heart-disease dataset comprising 1050 patients and 14 attributes. By gathering a set of directly obtainable features from the heart-disease dataset, They considered this feature vector to be input for a DCNN to discriminate whether an instance belongs to a healthy or cardiac disease class. To assess the performance of the proposed method, different performance metrics, namely, accuracy, precision, recall, and the F1 measure, were employed, and our model achieved validation accuracy of 91.7%. The experimental results indicate the effectiveness of the proposed approach in a real-world environment.

Ozcan et al. [9] proposed method includes data augmentation, feature extraction, hyperparameter tuning, and model training steps. In the first step, various data augmentation techniques were applied to increase the training data's diversity and strengthen the model's generalization capacity. The MFCC method was used in the second step to extract meaningful and distinctive features from the sound data. MFCC represents sound signals based on the frequencies the human ear perceives and provides a strong basis for classification. The obtained features were classified with an artificial neural network (ANN) model with optimized hyperparameters. The hyperparameter optimization of the model was performed using the grid search algorithm, and the most appropriate parameters were determined. Khalilzad et al. [10] focused on the detection of infants suffering from sepsis by developing a simplified design using acoustic features and conventional classifiers. The features for the proposed framework were MFCC, Spectral Entropy Cepstral Coefficients (SENCC) and Spectral Centroid Cepstral Coefficients (SCCC), which were classified through K-nearest Neighborhood (KNN) and SVM(SVM) classification methods. The performance of the different combinations of the feature sets was also evaluated based on several measures such as accuracy, F1-score and Matthews Correlation Coefficient (MCC). Bayesian Hyperparameter Optimization (BHPO) was employed to tailor the classifiers uniquely to fit each experiment. The proposed methodology was tested on two datasets of expiratory cries (EXP) and voiced inspiratory cries (INSV). Carollo et al. [11] focused Infant cry is an adaptive signal of distress that elicits timely and mostly appropriate caring behaviors. Caregivers are typically able to decode the meaning of the cry and respond appropriately, but maladaptive caregiver responses are common and, in the worst cases, can lead to harmful events. To tackle the importance of studying cry patterns and caregivers' responses, this review aims to identify key documents and thematic trends in the literature as well as existing research gaps. To do so, we conducted a scientometric review of 723 documents downloaded from Scopus and performed a document co-citation analysis.

Mishra et al. [12] evaluated primary causes of fatal head trauma in infants and young children, occurring in about 33 per 100,000 infants annually in the U.S., with mortality rates being between 15% and 38%. Survivors frequently endure long-term disabilities, such as cognitive deficits, visual impairments, and motor dysfunction. Diagnosing SBS remains difficult due to the lack of visible injuries and delayed symptom onset. Existing detection methods—such as neuroimaging, biomechanical modeling, and infant monitoring systems—cannot perform real-time detection and face ethical, technical, and accuracy limitations. This study proposes an inertial measurement unit (IMU)-based detection system enhanced with machine learning to identify aggressive shaking patterns. Martínez et al. [13] evaluated of pain in patients depends mainly on the continuous monitoring of the medical staff when the patient is unable to express verbally his/her experience of pain, as is the case of patients under sedation or babies. Therefore, it is necessary to provide alternative methods for its evaluation and detection. Facial

expressions can be considered as a valid indicator of a person's degree of pain. Consequently, this paper presents a monitoring system for babies that uses an automatic pain detection system by means of image analysis. Gulzar et al. [14] exploited to gauge health parameters, and machine learning techniques are investigated to predict the health conditions of patients to assist medical practitioners. Since these healthcare systems deal with large amounts of data, significant development is also noted in the computing platforms. The relevant literature reports the potential impact of ICT-enabled systems for improving maternal and infant health. This article reviews wearable sensors and AI algorithms based on existing systems designed to predict the risk factors during and after pregnancy for both mothers and infants.

3. PROPOSED METHODOLOGY

The research pipeline begins with a labelled dataset of baby-cry .wav files organized by class folders; this dataset is uploaded and inspected. Next, audio preprocessing and feature extraction convert raw waveforms into compact numeric descriptors in this research MFCC vectors are computed and saved. Using these features, traditional baseline models (SVM, KNN, DTC, ADB, LDA) are trained and evaluated to set performance baselines. The proposed model is a Conv1D based CNN that ingests MFCC feature sequences and learns hierarchical temporal patterns, trained with one-hot labels and validated during training. Models are compared using a suite of quantitative metrics (accuracy, precision, recall, F1), confusion matrices and ROC curves, and visualized/saved for analysis as demonstrated in Fig. 2. Finally, the saved CNN and label encoder are used to predict new unseen audio files predictions are shown in the GUI alongside a waveplot annotated with the predicted class.

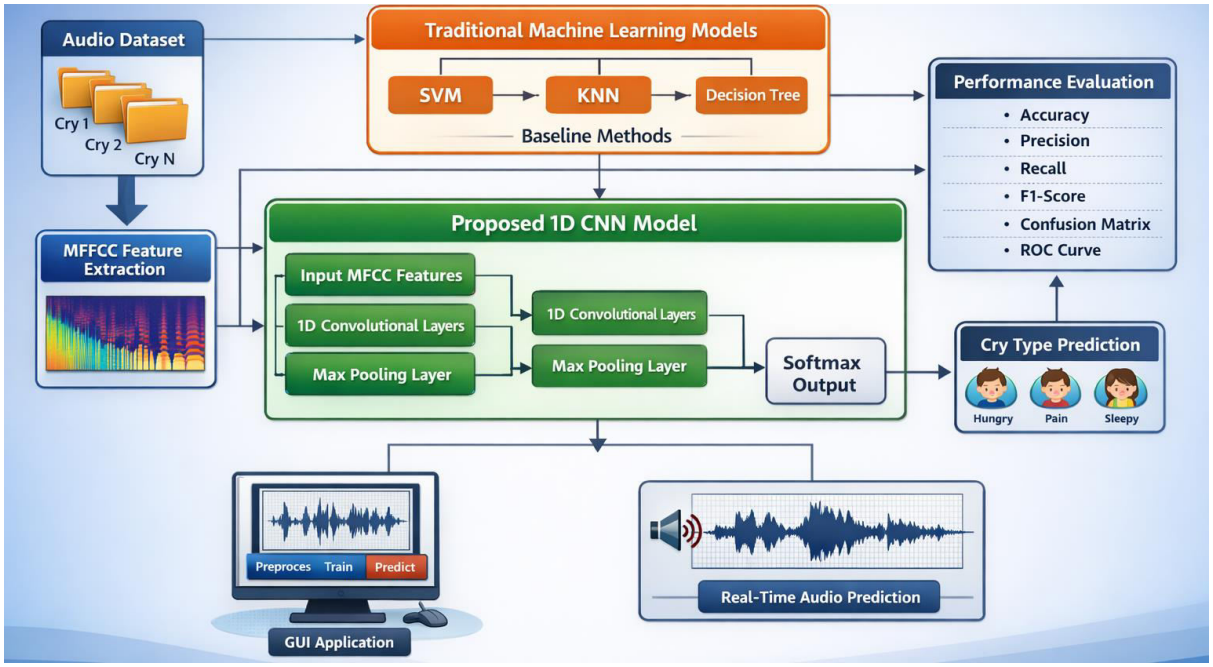

Fig. 2. Proposed system architecture of baby cry classification

The workflow begins with dataset collection, where baby cry audio recordings are organized into a structured directory with class-specific subfolders such as Hunger, Pain, and Sleepy, each containing .wav files recorded under realistic conditions to capture environmental variability. Metadata such as infant age, recording device, and sampling rate is also maintained to support analysis consistency. During preprocessing, each audio file is validated to remove corrupt or empty signals, followed by feature extraction using Librosa to compute MFCCs. These variable-length features are converted into fixed-length vectors through aggregation (e.g., mean across frames), forming a uniform feature matrix. The processed features (X) and corresponding labels (Y) are stored as .npy files for efficient reuse, while textual labels are encoded using LabelEncoder for model compatibility. To address class imbalance, techniques such as resampling or data augmentation (e.g., time-stretching and pitch shifting) may be applied. Baseline models including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree Classifier (DTC), AdaBoost (ADB), and Linear Discriminant Analysis (LDA) are trained using a consistent train-test split to ensure fair comparison, with predictions, probability scores, and trained models saved for evaluation and reproducibility.

The proposed system employs a Conv1D-based Convolutional Neural Network (CNN) designed to process MFCC feature sequences reshaped appropriately for deep learning input. The architecture consists of convolutional and pooling layers followed by dense layers with softmax activation, trained using categorical cross-entropy loss and optimized with Adam. To enhance performance, early stopping and validation monitoring are incorporated, and both the trained model and its history are stored. A hybrid approach is further introduced by extracting deep embeddings from the CNN and feeding them into an ensemble classifier such as ExtraTrees or AdaBoost, with final predictions obtained through stacking or weighted averaging. Model performance is evaluated using accuracy, precision, recall, F1-score, confusion matrices, and ROC-AUC curves, ensuring comprehensive analysis across all classes. For deployment, the trained model is integrated into a GUI-based prediction system, where unseen audio inputs are processed, classified, and visualized through waveform plots annotated with predicted labels, enabling practical and user-friendly real-time inference

CNN model

CNN is a class of deep learning models designed to automatically extract hierarchical features from structured data, such as images, signals, or sequential data. CNNs are particularly effective at capturing local patterns, spatial hierarchies, and correlations in input features. Instead of relying on handcrafted feature extraction, CNNs learn filters during training that highlight essential characteristics of the data. In the context of baby cry classification, CNNs process MFCC feature vectors derived from audio signals to identify patterns in the spectral-temporal domain that differentiate cry types such as Hunger, Pain, Sleepy, and Discomfort.

CNNs combine convolutional layers, which detect local feature patterns, with pooling layers that reduce dimensionality and introduce translation invariance as demonstrated in Fig. 3. Fully connected layers at the end of the network integrate the extracted features to perform classification. The model is trained using backpropagation with a categorical cross-entropy loss, allowing it to minimize prediction errors iteratively. This deep learning approach can capture subtle differences between cry categories that traditional machine learning models might miss, improving classification accuracy.

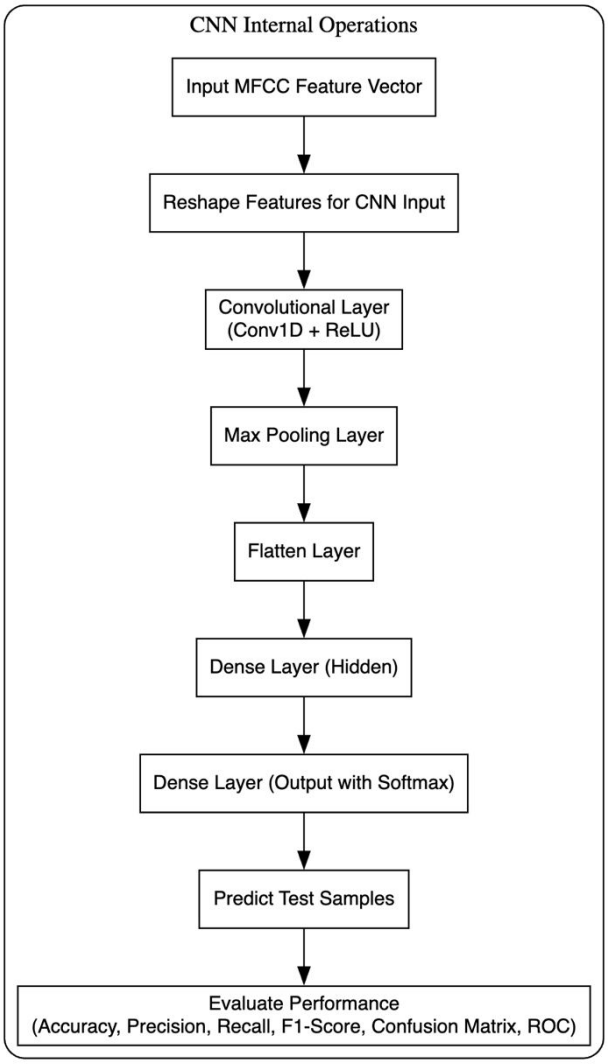

Fig. 3. Internal working flow of CNN.

The proposed Convolutional Neural Network (CNN) model processes Mel-Frequency Cepstral Coefficient (MFCC) features extracted from raw baby cry audio signals, which are reshaped into a one-dimensional format with an additional channel dimension to suit the network input. These features pass through convolutional layers where multiple filters identify local spectral and temporal patterns such as frequency shifts and amplitude variations associated with different cry types. The application of ReLU activation introduces non-linearity, enabling the model to learn complex relationships within the data. Max pooling layers are then used to reduce the dimensionality of feature maps, retain the most significant information, and enhance robustness to minor variations in the input signal. The resulting feature maps are flattened into a one-dimensional vector, preparing the data for classification through fully connected dense layers that learn higher-level feature representations.

During training, the CNN utilizes backpropagation with a categorical cross-entropy loss function to iteratively optimize its parameters and minimize prediction errors. Validation data is used to monitor model performance and prevent overfitting. For unseen test samples, MFCC features are extracted and passed through the trained network to generate probability distributions across all cry categories, with the highest probability indicating the predicted class. The model's performance is evaluated using metrics such as accuracy, precision, recall, and F1-score, while confusion matrices and ROC curves provide deeper insights into classification performance and class separability. Overall, CNN-based

approaches demonstrate superior capability in capturing intricate spectral-temporal characteristics compared to traditional machine learning models.

4. Results Description

Fig. 4. shows MFCC feature extraction interface confirms the successful preprocessing of the baby cry audio dataset, where the system automatically extracts MFCC from all input audio samples. After loading the dataset, the application identifies five cry categories such as belly pain, burping, discomfort, hungry, and tired and completes MFCC extraction, resulting in a structured feature matrix of size (1858, 10). This indicates that each audio sample is transformed into a compact 10-dimensional MFCC feature vector, making the data suitable for efficient training and evaluation of both traditional machine learning classifiers and the proposed CNN model for early baby health monitoring.

```
Dataset loaded
Classes found in dataset: ['belly_pain', 'burping', 'discomfort', 'hungry', 'tired']
Preprocessing and MFCC Feature Extraction completed on Dataset: D:/SAK/Chara
n codes/Baby Cry Classification/donateacry_corpus_cleaned_and_updated_data

Input MFCC Feature Set Size: (1858, 10)
```

Fig. 4. MFCC feature extraction completed

Fig. 5. shows proposed CNN classifier results that clearly demonstrate a substantial improvement in baby cry classification performance compared to all traditional machine learning models. The confusion matrix shows an almost perfectly diagonal structure, indicating that the CNN accurately identifies each cry category—belly pain, burping, discomfort, hungry, and tired with minimal misclassification, as most samples are correctly mapped to their respective classes. This highlights the CNN's ability to learn discriminative temporal and spectral patterns directly from MFCC feature sequences rather than relying on handcrafted decision boundaries. The ROC curves further validate this superiority, with AUC values reaching 1.00 for four classes and 0.98 for the hungry class, and all curves positioned far above the random baseline. This reflects excellent class separability, strong probability estimation, and high robustness in multiclass prediction. Overall, the results confirm that the CNN model effectively captures the complex and nonlinear acoustic characteristics of baby cries, making it highly suitable for accurate early health monitoring and real-world deployment.

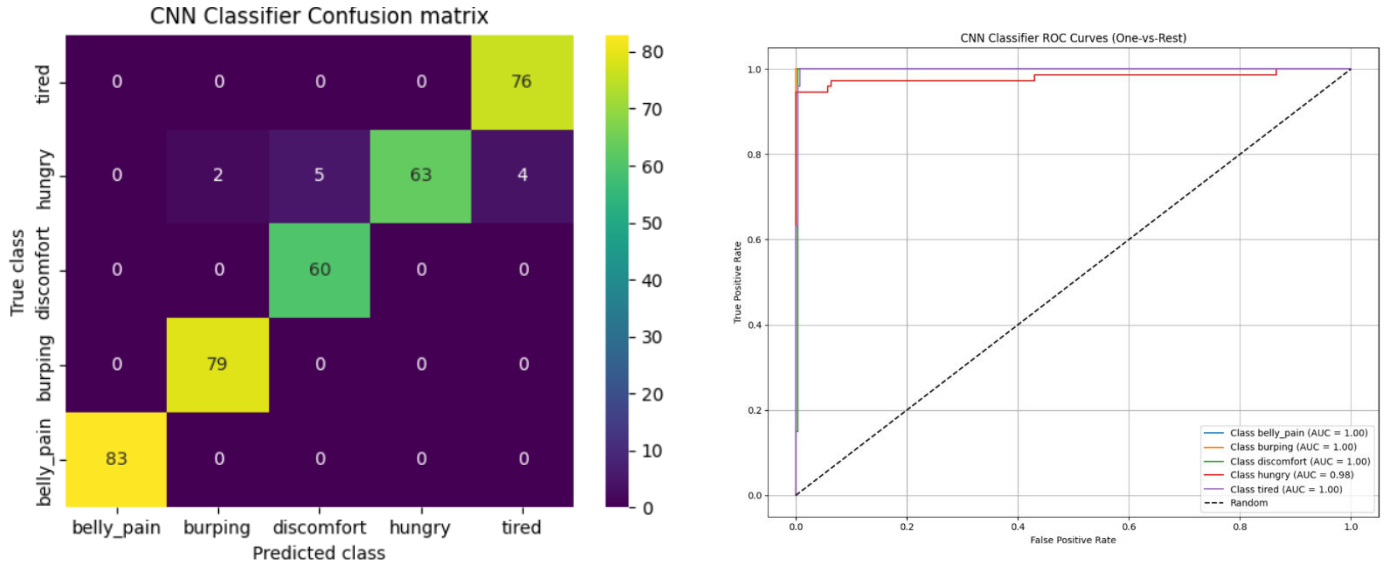

Fig. 5. Confusion matrix and ROC obtained using Proposed CNN

The Fig. 6. shows test audio prediction output using the proposed CNN model demonstrates the system's ability to accurately classify an unseen baby cry signal. After extracting MFCC features from the input audio, the trained CNN model analyzes the temporal and spectral patterns and predicts the cry category as "tired," which is clearly displayed on the waveform visualization. The amplitude–time plot provides an intuitive representation of the cry signal while simultaneously confirming the model's decision, thereby validating the effectiveness of the proposed CNN for real-time baby cry classification and early health monitoring.

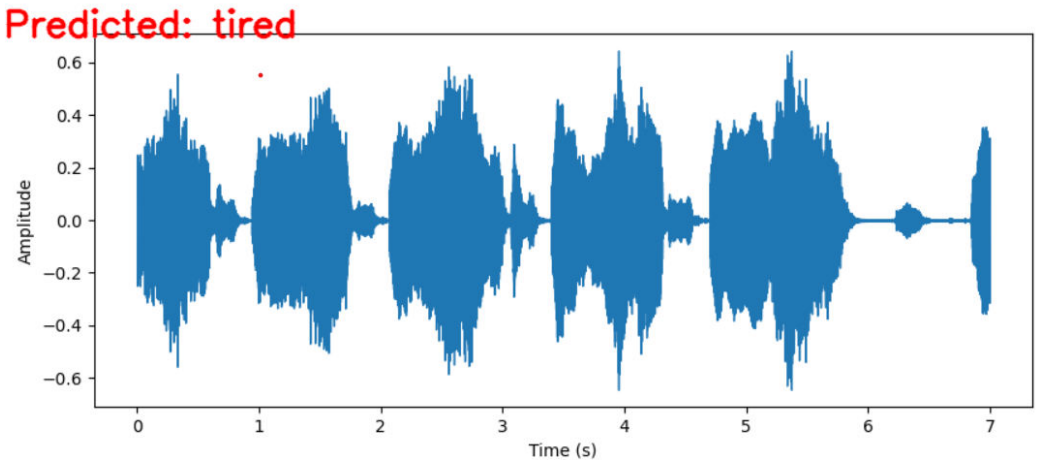

Fig. 6. Prediction on test sample obtained using Proposed CNN

The Fig. 7. depicts the prediction result on the test audio sample using the proposed CNN model shows that the system has correctly identified the baby cry category as "burping." After extracting MFCC features from the input audio signal, the CNN effectively captures distinctive temporal and spectral patterns associated with burping cries. The displayed waveform, along with the predicted label, confirms the model's strong classification capability and demonstrates its effectiveness for accurate and real-time baby cry analysis in early health monitoring applications.

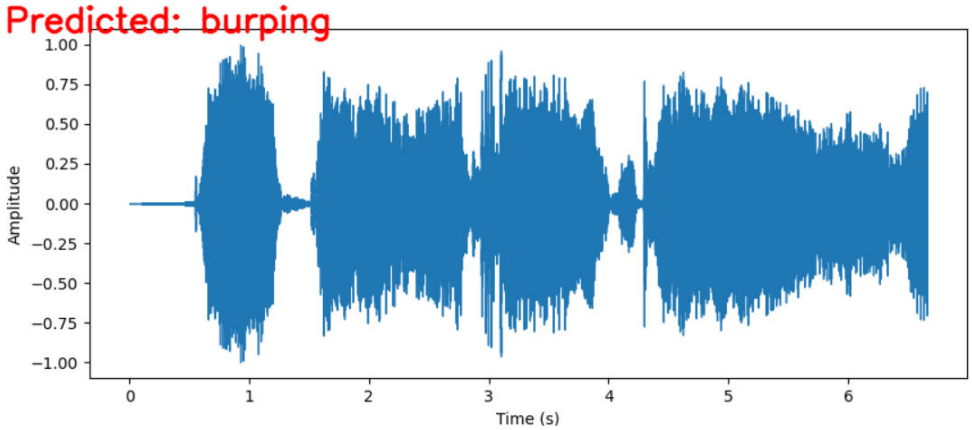

Fig. 7. Prediction on test sample obtained using Proposed CNN

Table 1: Performance comparison for the SVM, DTC, KNN, ADB, LDA and Proposed CNN Model

Algorithms Name	Accuracy	Precision	Recall	F-score
SVM	46.50%	51.65%	46.10%	47.03%
DTC	61/55%	60.74%	60.25%	57.60%
KNN	28.49%	20.50%	26.25%	20.08%
ADB	52.68%	54.94%	52.43%	53.23%

LDA	54.30%	53.11%	53.92%	53.16%
CNN	97.04%	96.96%	97.02%	96.83%

Table 1 presents a detailed performance comparison of the traditional machine learning classifiers such as SVM, DTC, KNN, ADB, and LDA against the proposed CNN model for baby cry classification. Among the existing methods, KNN exhibits the weakest performance with an accuracy of 28.49%, indicating poor discrimination of overlapping MFCC features, while SVM also shows limited effectiveness with an accuracy of 46.50% and relatively low recall. DTC, ADB, and LDA demonstrate moderate performance, achieving accuracies in the range of 52–61%, suggesting that tree-based, boosting, and linear methods can partially capture cry-specific patterns but struggle with complex acoustic variations. In contrast, the proposed CNN model significantly outperforms all baseline algorithms, achieving an accuracy of 97.04%, precision of 96.96%, recall of 97.02%, and F-score of 96.83%. This substantial improvement highlights the CNN's superior ability to learn discriminative temporal and spectral representations from MFCC features, thereby providing highly reliable and robust baby cry classification suitable for early health monitoring applications.

5. CONCLUSION

The research effectively presents a CNN-based infant cry classification system designed for early detection and analysis of infant conditions, highlighting the capability of deep learning techniques in processing complex audio signals. The proposed framework incorporates a comprehensive pipeline that includes dataset acquisition, MFCC-based feature extraction, data preprocessing, implementation of baseline machine learning models, and a CNN-based deep learning model, all integrated within a user-friendly graphical interface for seamless interaction. The system ensures efficient processing from raw audio input to final prediction output. Experimental findings reveal that conventional machine learning algorithms such as SVM, KNN, DTC AdaBoost, and LDA exhibit comparatively lower performance due to the nonlinear and overlapping nature of infant cry signals. In contrast, the proposed CNN model demonstrates superior performance in terms of accuracy, precision, recall, and F1-score, owing to its ability to capture intricate temporal and spectral patterns within the audio data. Furthermore, evaluation tools such as confusion matrices, ROC curves, and waveform-based prediction visualizations validate the consistency and reliability of the system. The model successfully classifies unseen audio samples into categories including hunger, discomfort, burping, tiredness, and abdominal pain, indicating its effectiveness in real-time applications. By combining machine learning and deep learning approaches, the system achieves enhanced robustness and reliability. The proposed solution serves as an intelligent decision-support tool for caregivers and healthcare monitoring, contributing to improved infant care and establishing a strong foundation for future smart healthcare systems.

REFERENCES

[1]. Lahti, K.; Vänskä, M.; Qouta, S.R.; Diab, M.Y.; Perko, K.; Punamäki, R.L. Maternal experience of their infants' crying in the context of war trauma: Determinants and consequences. Infant Ment. Health J. 2019, 40, 717–734.

[2]. Halpern, R.; Coelho, R. Excessive crying in infants. J. Pediatr. 2016, 92, S40–S45.

[3]. Sharma, K.; Gupta, C.; Gupta, S. Infant weeping calls decoder using statistical feature extraction and Gaussian mixture models. In Proceedings of the 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 6–8 July 2019; pp. 1–6.

[4]. Maghfira, T.N.; Basaruddin, T.; Krisnadhi, A. Infant cry classification using CNN-RNN. J. Phys. Conf. Ser. 2020, 1528, 012019.

[5]. Franti, E.; Ispas, I.; Dascalu, M. Testing the Universal Baby Language hypothesis - automatic infant speech recognition with CNNs. In Proceedings of the 2018 41st International Conference on Telecommunications and Signal Processing (TSP), Athens, Greece, 4–6 July 2018; pp. 1–4.

[6]. Gul MA, Yousaf MH, Nawaz S, Ur Rehman Z, Kim H. Patient Monitoring by Abnormal Human Activity Recognition Based on CNN Architecture. *Electronics*. 2020; 9(12):1993. https://doi.org/10.3390/electronics9121993

[7]. Mohammed FA, Tune KK, Assefa BG, Jett M, Muhie S. Medical Image Classifications Using Convolutional Neural Networks: A Survey of Current Methods and Statistical Modeling of the Literature. *Machine Learning and Knowledge Extraction*. 2024; 6(1):699-735. https://doi.org/10.3390/make6010033

[8]. Arooj S, Rehman Su, Imran A, Almuhaimeed A, Alzahrani AK, Alzahrani A. A Deep Convolutional Neural Network for the Early Detection of Heart Disease. *Biomedicines*. 2022; 10(11):2796. https://doi.org/10.3390/biomedicines10112796

[9]. Ozcan T, Gungor H. Baby Cry Classification Using Structure-Tuned Artificial Neural Networks with Data Augmentation and MFCC Features. *Applied Sciences*. 2025; 15(5):2648. https://doi.org/10.3390/app15052648

[10]. Khalilzad Z, Kheddache Y, Tadj C. An Entropy-Based Architecture for Detection of Sepsis in Newborn Cry Diagnostic Systems. *Entropy*. 2022; 24(9):1194. https://doi.org/10.3390/e24091194

[11]. Carollo A, Montefalcone P, Bornstein MH, Esposito G. A Scientometric Review of Infant Cry and Caregiver Responsiveness: Literature Trends and Research Gaps over 60 Years of Developmental Study. *Children*. 2023; 10(6):1042. https://doi.org/10.3390/children10061042

[12]. Mishra RK, AlAnsari K, Cole R, Nazarian A, Potter IY, Vaziri A. The Development of a Wearable-Based System for Detecting Shaken Baby Syndrome Using Machine Learning Models. *Sensors*. 2025; 25(15):4767. https://doi.org/10.3390/s25154767

[13]. Martínez A, Pujol FA, Mora H. Application of Texture Descriptors to Facial Emotion Recognition in Infants. *Applied Sciences*. 2020; 10(3):1115. https://doi.org/10.3390/app10031115

[14]. Gulzar Ahmad S, Iqbal T, Javaid A, Ullah Munir E, Kirn N, Ullah Jan S, Ramzan N. Sensing and Artificial Intelligent Maternal-Infant Health Care Systems: A Review. *Sensors*. 2022; 22(12):4362. https://doi.org/10.3390/s22124362