Acoustic Intelligence Framework for Decoding Infant Cry Patterns in Early Health Monitoring

Ayesha Nikitha[1], Sanda Supriya[2], Sathanuri Nandini[2], Gandra Ruthvik[2], Peddaveni Abhiram [2]

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

[1,2]Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India.

ABSTRACT

Understanding baby cries is essential, as infants cannot communicate their needs verbally. Traditionally, caregivers relied on experience, observation, and intuition to interpret cries. However, this approach is subjective, inconsistent, and may result in incorrect decisions regarding needs such as hunger, pain, or discomfort. These limitations highlight the necessity for an accurate and automated solution. The proposed system introduces an intelligent approach for baby cry classification using audio signal processing techniques. The system analyzes cry audio signals and extracts meaningful features using Mel-Frequency Cepstral Coefficients (MFCC), which effectively represent the frequency characteristics of sound. These features are then used to train multiple machine learning algorithms, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree Classifier (DTC), AdaBoost (ADB), and Linear Discriminant Analysis (LDA), enabling comparative performance evaluation. To further enhance classification accuracy, a Convolutional Neural Network (CNN), a deep learning model, is implemented as the primary approach. The CNN model is capable of learning complex patterns and relationships within the extracted features, leading to improved prediction performance over traditional machine learning techniques. The system's performance is assessed using evaluation metrics such as accuracy, precision, recall, and F1-score. Finally, the research is completed by integrating feature extraction, model training, evaluation, and prediction into a user-friendly interface, allowing efficient and reliable identification of baby cry types to support caregivers in making timely and informed decisions.

Key words: Neonatal Diagnosis, Spectrogram Analysis, Feature Extraction, Healthcare Monitoring, Infant Cry Classification.

1. INTRODUCTION

Babies cry as an innate reflex to express their needs, such as hunger, discomfort, and sleeplessness. Understanding these cries accurately and quickly is critical to babies' healthy development and happiness [1]. However, inexperienced parents, new caregivers, and some healthcare professionals may have difficulty understanding why babies cry. This can lead to prolonged crying, increased stress levels, and negatively impacted cognitive development [2]. Studies on the automatic classification of baby crying sounds have significantly contributed to this field. Analysis and classification of baby crying sounds using sound processing methods and machine learning algorithms constitute the basis of research in this field. This literature review examines important studies on the classification and detection of baby crying sounds as shown in Fig. 1. It discusses, in detail, the methodologies, datasets, feature extraction methods, and classification algorithms used in these studies. Various studies on the classification of baby cries have developed different methods using sound processing methods and machine learning algorithms.

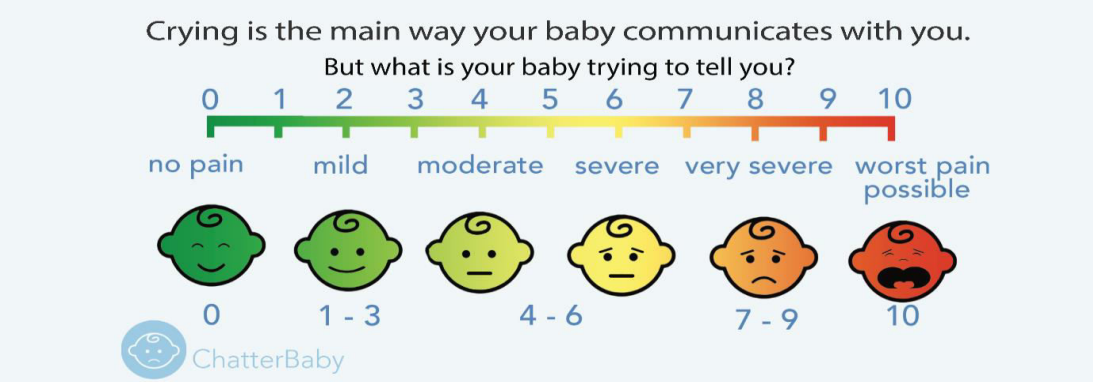

Fig. 1. Baby cry symptoms analysis.

Infant communication primarily relies on crying, a natural means to signal their needs or discomfort. However, the inability of babies to verbally express their needs presents a challenge for parents, who often must guess the reasons behind their cries [3]. This situation can lead to delays in responding to the infant's needs, causing stressful moments for parents and prolonged discomfort for the child. In response to this issue, technology offers solutions to help identify the reasons for crying more accurately [4]. Consequently, distinguishing between cries with distinct meanings using associated auditory features is imperative. This can enable the interpretation of a baby's needs and provide parents with appropriate ways to soothe their child [5]. It can also reduce stress for parents and caregivers by helping them avoid misinterpreting their baby's cries.

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

2.1 Traditional Machine Learning with Acoustic Feature Engineering

Early research in infant cry classification primarily relied on handcrafted acoustic features combined with classical machine learning models. Mekhfioui et al. [6] utilized prosodic and cepstral features such as MFCCs to distinguish between different crying causes including hunger, pain, and discomfort. Their system incorporated an embedded microphone for real-time acquisition and a display module for output visualization, along with cloud integration for remote monitoring and analysis. Similarly, Khalilzad et al. [11] explored cry signals as biomarkers for detecting neonatal pathologies such as sepsis and respiratory distress syndrome (RDS). They combined harmonic ratio (HR) and Gammatone frequency cepstral coefficients (GFCCs) and employed MLP and SVM classifiers, achieving up to 95.3% accuracy after feature fusion and hyperparameter tuning. These studies highlight the effectiveness of feature engineering in extracting discriminative information from cry signals.

2.2 Deep Learning Models for Infant Cry Classification

With the advancement of deep learning, researchers have increasingly adopted CNN-based architectures for improved performance. Li et al. [7] proposed a hybrid model integrating ResNet with

transformer mechanisms and SE attention modules to enhance feature representation from MFCC inputs, achieving an accuracy of 93% while reducing training time. Likewise, Liang et al. [9] combined CNN and LSTM architectures to classify infant needs such as hunger, emotional requirements, and pain using MFCC features extracted from 1607 audio samples. Their study demonstrated that hybrid deep architectures outperform traditional ANN models. Furthermore, Herlea et al. [8] conducted a comparative analysis of deep learning architectures such as ResNet and EfficientNet using spectrogram and MFCC representations, emphasizing their role in enhancing baby monitoring systems.

2.3 Multimodal and Hybrid Feature Fusion Approaches

Recent works have explored combining multiple feature domains to improve classification robustness. Zayed et al. [10] developed a diagnostic system that integrates prosodic features (HR), cepstral features (GFCC), and image-based spectrogram features extracted using pretrained CNN models. This multimodal fusion significantly improved classification accuracy for detecting neonatal diseases such as sepsis and RDS. Similarly, Khalilzad et al. [11] demonstrated that combining spectral and short-term features enhances discrimination capability across pathology classes. These hybrid approaches indicate that leveraging complementary feature representations leads to better performance than single-domain methods.

2.4 Transfer Learning and Pretrained Model Adaptation

Transfer learning has emerged as an effective strategy for improving classification accuracy with limited datasets. Tsalera et al. [12] investigated retraining pretrained CNN models such as GoogLeNet, SqueezeNet, ShuffleNet, VGGish, and YAMNet for sound classification tasks. Their work analyzed the impact of hyperparameters like learning rate, batch size, and optimizer on performance and computational efficiency. Additionally, Singh et al. [14] proposed a system combining Inception-v3 CNN with LSTM layers for neonatal activity classification using video data, demonstrating the versatility of transfer learning in healthcare monitoring applications.

2.5 Self-Supervised Learning and Advanced Representation Learning

To overcome limitations of manual feature extraction, recent studies have adopted self-supervised learning (SSL) techniques. Shayegh et al. [15] utilized SSL models such as wav2vec 2.0, WavLM, and HuBERT to extract deep representations directly from raw cry signals. These models eliminate the dependency on handcrafted features while capturing complex temporal patterns, achieving effective classification of healthy and pathological conditions such as sepsis and RDS. This approach represents a significant advancement toward fully automated and scalable cry analysis systems.

2.6 General Audio Event Detection and Evaluation Frameworks

Beyond infant cry analysis, general audio event detection frameworks provide valuable insights into classification methodologies. Villegas et al. [13] proposed a deep learning-based system for detecting environmental sound events using acoustic feature extraction and comprehensive evaluation metrics such as precision, recall, F1-score, and ROC-AUC. Their work emphasizes the importance of detailed performance evaluation, including confusion matrix analysis and false prediction assessment, which is also critical in infant cry classification systems.

3. PROPOSED METHODOLOGY

3.1 Overview

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.

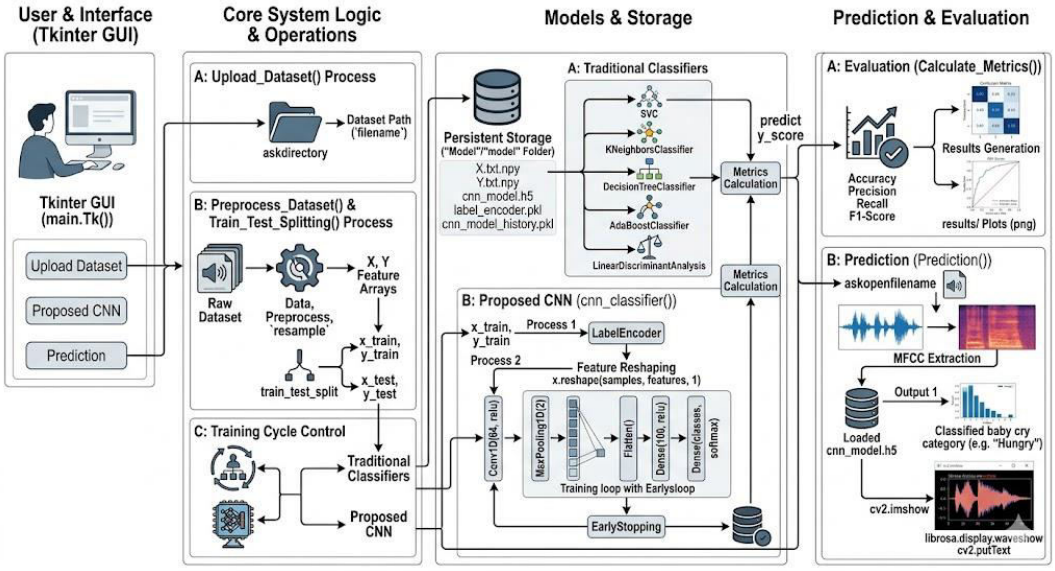

Fig. 2. Proposed system architecture of baby cry classification

The system begins with dataset collection by organizing baby cry audio recordings into a structured directory where each subfolder represents a class label such as Hunger, Pain, or Sleepy, containing .wav files recorded under realistic conditions with environmental noise and variability, while also storing metadata like age, device, and sampling rate; preprocessing then involves loading each audio file, removing corrupted or empty samples, extracting MFCC features using Librosa, and converting variable-length signals into fixed-length feature vectors (e.g., by averaging), followed by saving the processed feature matrix (X) and encoded labels (Y) as .npy files and applying LabelEncoder for numerical conversion, with optional handling of class imbalance through resampling or augmentation techniques like time-stretching or pitch shifting; finally, baseline model building is performed by training multiple classical classifiers such as SVM, KNN, DTC, ADB, and LDA using a consistent 80–20 train-test split to ensure fair comparison, generating predictions and probability scores for evaluation (including ROC analysis), and saving trained models using joblib or pickle while optionally tuning hyperparameters through grid search or cross-validation to establish strong and reproducible benchmarks.

The proposed system extends the pipeline by introducing a Conv1D CNN model that operates on MFCC features reshaped into (samples, features, 1), using a lightweight architecture consisting of Conv1D, MaxPooling1D, Flatten, and Dense layers with softmax activation to balance performance and overfitting, trained using categorical cross-entropy and Adam optimizer while monitoring validation performance with techniques like validation split and early stopping; the trained model and its history are saved for reuse, and for enhanced performance, a hybrid approach is implemented by extracting deep embeddings from the penultimate CNN layer and feeding them into an ensemble classifier such

as ExtraTrees or AdaBoost, with final predictions obtained through stacking or weighted averaging to combine deep learning and ensemble strengths; all models including baselines, CNN, and hybrid are evaluated consistently using accuracy, precision, recall, F1-score (macro), confusion matrices, and ROC-AUC curves, with results visualized and stored for analysis, along with statistical significance testing to validate improvements and detailed per-class error analysis; finally, the system supports real-time prediction on unseen audio through a GUI-based pipeline where a selected .wav file undergoes MFCC extraction, classification using the trained model and label encoder, and result display along with waveform visualization, while deployment readiness is ensured by packaging required model files and optionally extending the system into a edge-based solution for continuous monitoring.

3.2 Convolutional Neural Network

CNN are 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.

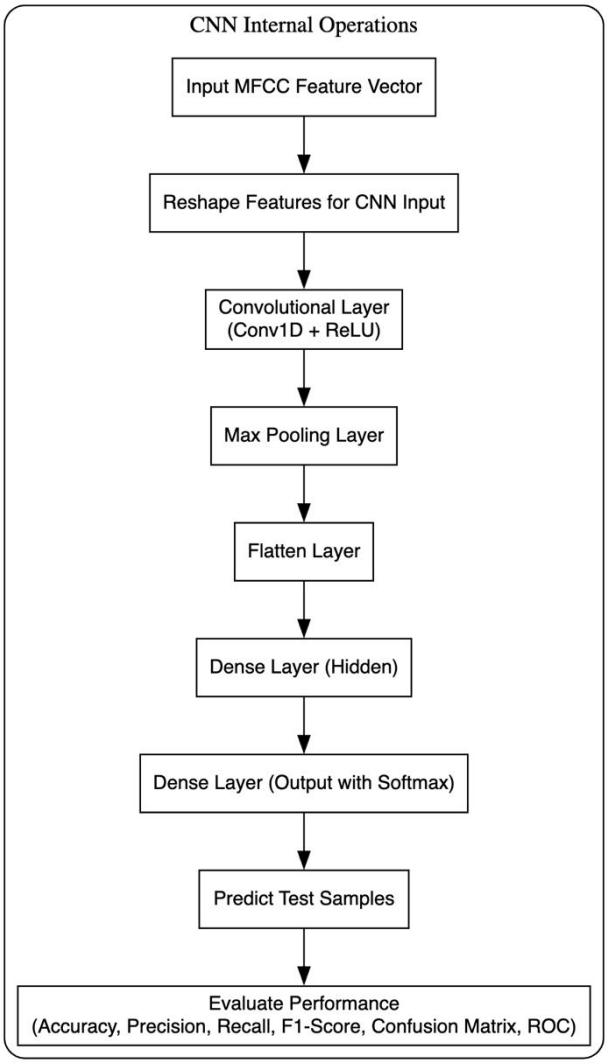

Fig. 3. Internal working flow of CNN.

MFCC features are extracted from raw audio signals. These features encode frequency and temporal information of baby cries and are reshaped to match the expected input dimensions of the CNN, typically as a one-dimensional array with a channel dimension. The reshaped MFCC vectors pass through convolutional layers. Each convolutional filter scans the input and detects local patterns, such as characteristic frequency modulations or amplitude changes that correspond to specific cry types. ReLU activation introduces non-linearity to capture complex relationships. Pooling layers reduce the dimensionality of feature maps by selecting maximum values within small regions. This step reduces computational complexity, provides robustness to minor variations, and retains the most salient features relevant for distinguishing cry types.

The output feature maps from convolutional and pooling layers are flattened into a one-dimensional vector, preparing the data for fully connected layers. Flattening allows integration of local patterns into a global feature representation suitable for classification. Dense layers process the flattened feature vector and learn higher-level representations. The hidden dense layer applies a non-linear transformation, enabling the network to model complex interactions among features. The final output layer uses a softmax activation to produce class probabilities for each cry type.

The CNN is trained using backpropagation with a categorical cross-entropy loss function. During training, the network adjusts filter weights iteratively to minimize the difference between predicted probabilities and actual cry labels. Validation data helps monitor overfitting and model generalization. For unseen test samples, MFCC features are extracted and reshaped. The trained CNN outputs probability distributions across all cry categories. The predicted label is the category with the highest probability, effectively classifying the baby cry. Predicted labels are compared with actual labels to compute accuracy, precision, recall, and F1-score. Confusion matrices visualize misclassifications, while ROC curves can depict class separability. CNNs often outperform traditional models in identifying subtle spectral-temporal differences due to their ability to learn hierarchical feature representations automatically.

4. RESULTS ANALYSIS

The 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.

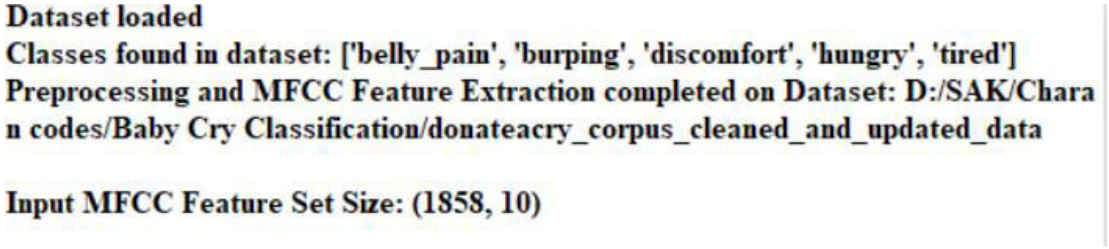

Fig. 4. MFCC feature extraction completed

The 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.

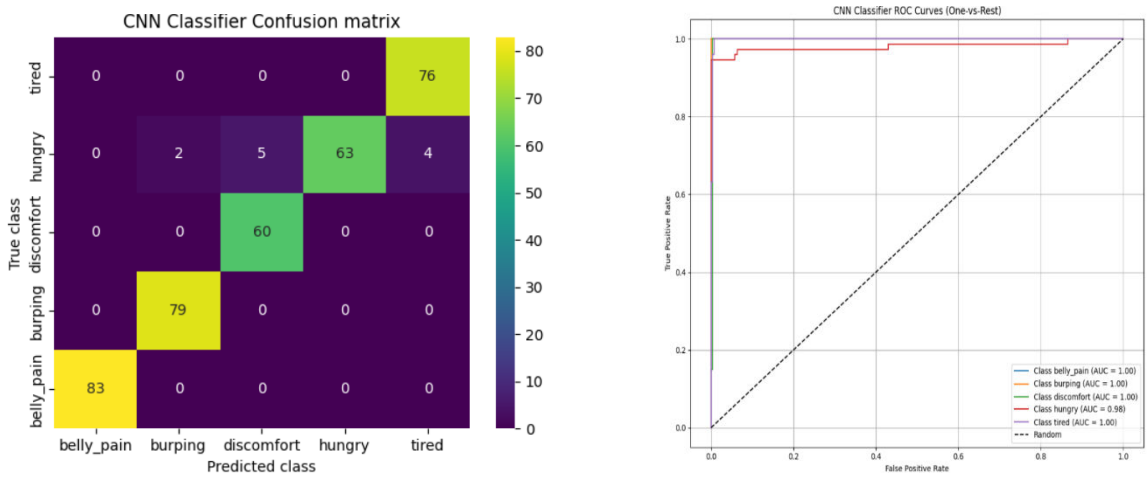

Fig. 5. Illustration of Confusion matrix and ROC obtained using Proposed CNN

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.

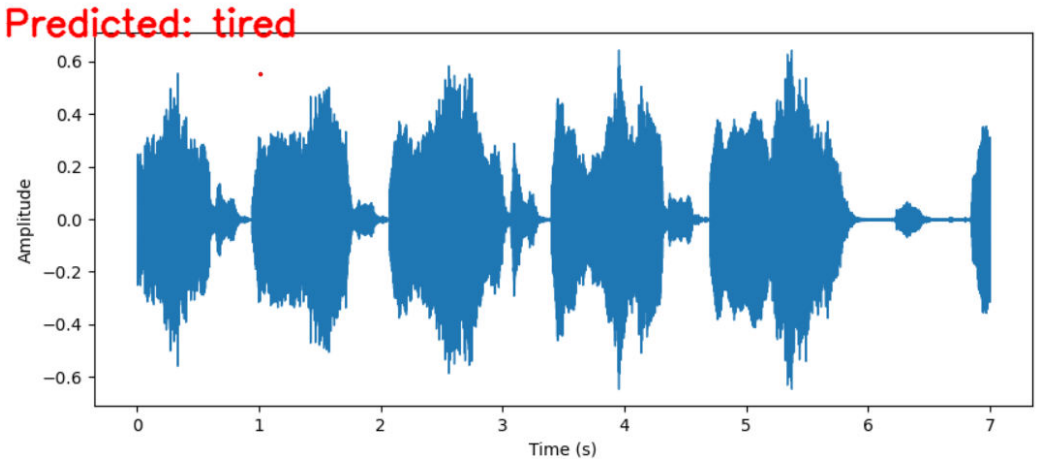

Fig. 6. Prediction on test sample obtained using Proposed CNN

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.

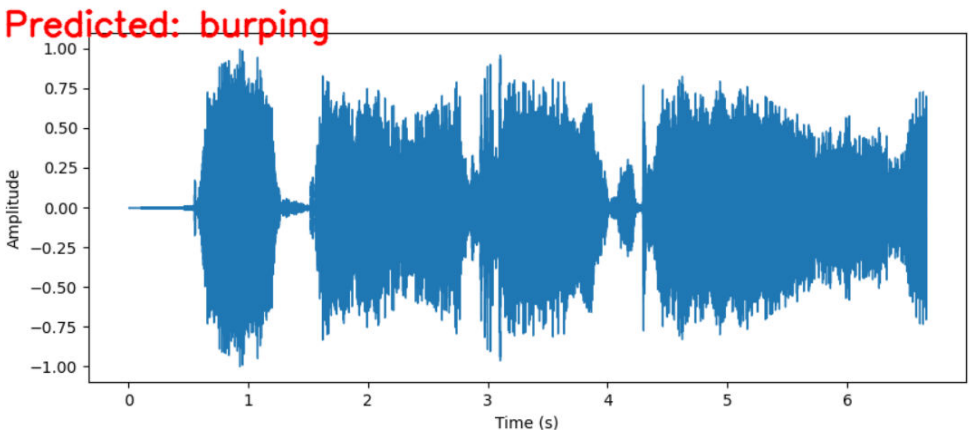

Fig. 7. Prediction on test sample obtained using Proposed CNN

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 successfully demonstates a CNN-based baby cry classification system for early infant condition analysis, demonstrating the effectiveness of DL techniques in handling complex audio signals. The system integrates a complete workflow including dataset collection, MFCC-based feature extraction, preprocessing, ML baseline models, and a proposed CNN model, all supported through a GUI for user interaction. The implementation ensures smooth processing from raw audio input to final prediction. Experimental results indicate that traditional ML models such as SVM, KNN, DT, ADB, and LDA show comparatively lower performance due to overlapping and nonlinear characteristics of baby cry signals. In contrast, the proposed CNN model achieves higher accuracy, precision, recall, and F-score, proving its ability to learn deep temporal and spectral patterns from audio features. The generated CM, ROC curves, and waveform-based prediction outputs further confirm the consistency and reliability of the system. Moreover, the system effectively classifies unseen audio samples into categories such as hungry, discomfort, burping, tired, and belly pain, demonstrating its applicability in real-time scenarios. The integration of ML and DL approaches enhances the overall robustness of the system. Thus, the proposed model serves as an intelligent support tool for caregivers and healthcare monitoring, contributing towards improved infant care and forming a strong base for future smart monitoring 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]. Mekhfioui M, Fadel W, Hammouch FE, Laayati O, Bouchouirbat M, El Bazi N, Satif A, Boujiha T, Chebak A. Development of a Baby Cry Identification System Using a Raspberry Pi-Based Embedded System and Machine Learning. *Technologies*. 2025; 13(4):130. https://doi.org/10.3390/technologies13040130

[7]. Li F, Cui C, Hu Y. Classification of Infant Crying Sounds Using SE-ResNet-Transformer. *Sensors*. 2024; 24(20):6575. https://doi.org/10.3390/s24206575

[8]. Herlea DM, Iancu B, Ardelean E-R. A Study of Deep Learning Models for Audio Classification of Infant Crying in a Baby Monitoring System. *Informatics*. 2025; 12(2):50. https://doi.org/10.3390/informatics12020050

[9]. Liang Y-C, Wijaya I, Yang M-T, Cuevas Juarez JR, Chang H-T. Deep Learning for Infant Cry Recognition. *International Journal of Environmental Research and Public Health*. 2022; 19(10):6311. https://doi.org/10.3390/ijerph19106311

[10]. Zayed Y, Hasasneh A, Tadj C. Infant Cry Signal Diagnostic System Using Deep Learning and Fused Features. *Diagnostics*. 2023; 13(12):2107. https://doi.org/10.3390/diagnostics13122107

[11]. Khalilzad Z, Hasasneh A, Tadj C. Newborn Cry-Based Diagnostic System to Distinguish between Sepsis and Respiratory Distress Syndrome Using Combined Acoustic Features. *Diagnostics*. 2022; 12(11):2802. https://doi.org/10.3390/diagnostics12112802

[12]. Tsalera E, Papadakis A, Samarakou M. Comparison of Pre-Trained CNNs for Audio Classification Using Transfer Learning. *Journal of Sensor and Actuator Networks*. 2021; 10(4):72. https://doi.org/10.3390/jsan10040072

[13]. Villegas-Ch W, Govea J. Application of Deep Learning in the Early Detection of Emergency Situations and Security Monitoring in Public Spaces. *Applied System Innovation*. 2023; 6(5):90. https://doi.org/10.3390/asi6050090

[14]. Singh H, Kusuda S, McAdams RM, Gupta S, Kalra J, Kaur R, Das R, Anand S, Pandey AK, Cho SJ, et al. Machine Learning-Based Automatic Classification of Video Recorded Neonatal Manipulations and Associated Physiological Parameters: A Feasibility Study. *Children*. 2021; 8(1):1. https://doi.org/10.3390/children8010001

[15]. Shayegh SV, Tadj C. Deep Audio Features and Self-Supervised Learning for Early Diagnosis of Neonatal Diseases: Sepsis and Respiratory Distress Syndrome Classification from Infant Cry Signals. *Electronics*. 2025; 14(2):248. https://doi.org/10.3390/electronics14020248