Decoding Infant Cry Using Audio Data and Machine Learning Approaches

MSc Research Project
MSc in Data Analytics

Kalyani Phursule
Student ID: 23229268

School of Computing
National College of Ireland

Supervisor: Prof. Furqan Rustam

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Kalyani Phursule
Student ID:	23229268
Programme:	MSc in Data Analytics
Year:	2024-2025
Module:	MSc Research Project
Supervisor:	Prof. Furqan Rustam
Submission Due Date:	29/01/2025
Project Title:	Decoding Infant Cry Using Audio Data and Machine Learning Approaches
Word Count:	7851
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Kalyani Phursule
Date:	28th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Decoding Infant Cry Using Audio Data and Machine Learning Approaches

Kalyani Phursule
23229268

Abstract

This paper showcases a robust machine learning and deep learning based approach to classify the different infant cries. The significance of accurately determining why an infant is crying is from parents and caregivers perspective provides a motivation for this research. To accomplish this, several models such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, AdaBoost, and XGBoost, and deep learning models like, Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and ensemble models to improve the final model performance. The Synthetic Minority Over-sampling Technique (SMOTE) was implemented to handle imbalance in the dataset. Random Forest and XGBoost models returned highest accuracy, with Random Forest with SMOTE outperformed all the models by achieving 99.6% accuracy and 99% F1 score. Ensemble models also performed well but slightly less accurate than separate models trained with SMOTE. These competitive results highlight the importance of addressing data imbalance for better performance of a model, specifically for minority classes. Although the research is able to train machine learning models that could classify infant cries with significant accuracy, there still are challenges such as limitations of computational requirements of the models and real time audio processing. The future work involves transformer models, registration of pathological cry types, and optimization for low-resource environments targeting real-time applications. This research provides a fair basis for the further improvement of smart baby monitoring systems in adapting automated infant care.

Keywords: Baby cry classification, Machine learning, Ensemble learning, Deep learning, MFCC, Audio features extraction, SMOTE, Data Imbalance

1 Introduction

Languages provide an ease for humans to communicate their sentiments with each other. However, it is difficult for infants to express their feelings or discomfort due to lack of linguistic skill development. High pitched sound of screaming or crying is the primary means for kids to signal any unease or needs. Although parents remain most attentive, still sometimes it is difficult for them to know the exact reason for baby's cry. In order to improve infant well-being as well as parent's responsiveness to address infant needs promptly, it is significantly important to understand and interpret the cry of the baby. Parents and caregivers traditionally rely on individual experiences and instincts to guess or understand the reason of baby cry, this may be inaccurate and inconsistent. Employing advanced machine learning techniques can help to offer a reliable solution, which can help to accurately detect the cry reason (Pradhan et al. (2022)).

The automated cry classification system can help parents and caregivers to understand the exact reason for baby's unease, significantly reducing response time to handle baby's needs, improving infant care and detecting potential health risks at an early stage. These models can also be integrated with baby care instruments such as baby monitors which may lead to an intelligent tool in assisting parents. This study aims to explore the implementation of advanced machine learning techniques to differentiate the infant cries into classes : 'hungry', 'burping', 'discomfort', 'belly-pain' and 'tired'. The existing work in this field shows a variety of techniques to classify audio signals, including different feature extraction methods such as MFCC, Chrom feature extraction and Spectrograms, and machine learning models such as Support Vector Machine (SVM), Random Forest KNN, etc., (Aggarwal et al. (2023), Riadi et al. (2024)). Although deep learning methods specially those involving hybrid models and ensemble learning models are underexplored (G. A et al. (2024), Narayanan et al. (2024)). Moreover, the class imbalance issue in infant cry datasets remains inadequately addressed, leading to models that perform poorly when exposed to underrepresented classes. This gap motivates the need for a more comprehensive approach that not only leverages cutting-edge deep learning techniques but also addresses data imbalance issues.

Research Question and Objective:

The research question aiming in the study is: How can machine learning and deep learning models be effectively used to classify infant cries in different categories like hunger, discomfort and pain, especially when addressing class imbalance using SMOTE technique? The proposed solution explores different machine learning models from standard classifiers to deep learning models, also balancing the minority classes using Synthetic Minority Over-Sampling Technique (SMOTE) (Kumari and Mahto (2024), Zhang et al. (2023)). The study aims to fill that gap and allow to perform a deeper analysis of the mentioned models, their robustness against noise, and thus to develop better infant cry monitoring systems.

The paper is structured as follows: Section 1 gives an introduction of the topic, research problem, motivation and research question. A detailed overview of related previous work, including a literature review on audio classification, challenges, and techniques applied to the analysis of infant crying, is given in Section 2. Section 3: research methodology, as data collection, pre-processing, feature extraction, and model training. In Section 4 design specification details are provided, followed by the implementation of the machine learning models. Section 5, presents the results of the research with evaluation metrics to provide a comparative analysis of the various approaches. And in Section 6 the conclusion and future work are discussed, corresponding to the summary of the key findings along with the research outcome and outline for future work recommendations.

2 Related Work

In this section critical analysis of the previous works is presented, discussing their aims, contributions, limitations and their relationship with the current research on baby cry analysis. The literature review shows that a wide variety of methods in terms of machine learning, deep learning, hybrid and IoT-based approaches have been explored for baby cry analysis and classification aiming to enhance the identification and understanding of the infant cries.

2.1 Cry Identification Under Challenging Environment

Cry detection has been discussed in past studies even in harsh environments such as ambient noise and its applicable use in real life. According to Nimbarte et al. (2021), Zhang et al. (2023) and Mala and Darandale (2024) handling background noise while accurately identifying baby cries during real-world application is a critical issue. In such cases, it can be hard to generalize the models in terms of noise. Varying noise levels challenge models, affecting their accuracy and reliability. As pointed out by Narayanan et al. (2024), the frequent occurrence of false alarms further limits the practicality of these models. Jamal and Al-Azani (2023) proposed a hybrid feature approach, which combined prosodic and spectral acoustic features, attaining a significant improvement in the classification accuracy of pain-related cries but with poor scalability. Khandelwal et al. (2022) proposed a low-complexity Convolutional Recurrent Neural Network (CRNN) model with the aim of achieving computational efficiency in detecting baby cries in domestic environments, although at some cost to the accuracy of the results. Bella and Sanjaya (2023), showcased the use of augmented datasets can contribute to a model's robustness. This shows the need for developing models capable of differentiating cry types in challenging acoustic surroundings.

2.2 Pathological and Specialized Cry Detection

Kumari and Mahto (2024) have done a narrative review of techniques on pathological cry detection, emphasizing that one of the biggest challenges in this area is the unavailability of pathological datasets. You et al. (2023) implemented LSTM networks in classifying infant cries into different classes such as awake, diaper change, hunger, sleepy and discomfort. Inclusion of different types of cries into this approach makes the idea more practical for real-world applications. The pathological detection and multiple cry types being the focus provide background to the current research, which suggests that expanding the categories of cries beyond general needs is important to enhance practical utility in baby cry detection systems.

2.3 Traditional Machine Learning methods

In several studies classic machine learning methods were utilized to classify baby cries. Aggarwal et al. compared SVM, Random Forest, Decision Tree models based on spectral features in (2023) and SVM model achieved the best performance in the classification of cry. Their study was, however, limited by the use of a small dataset, which constrains the applicability of the results. Riadi et al. (2024) also conducted their research on different machine learning models, including SVM and Random forest. However, they found that the best accuracy is obtained using Random Forest. Similarly, Mahmoud et al. (2020) proposed an SSKNN approach for expanding the training set with unlabelled data to improve accuracy in hunger-related cries. While the feature extraction for these studies was quite detailed, all were bound by dataset limitations that extend to broader applicability. This points to the very critical challenge in the domain: the availability of enough diversity in datasets for model training and validation. The concentration on spectral features by these various studies underlines the need for research into richer datasets in an effort to improve robustness in classification.

2.4 Deep Learning Approaches

The deep machine learning methods have improved in modern times in dealing with some of the traditional machine learning methods' limitations regarding the analysis of baby cries. Narayanan et al. (2024) leveraging spectrograms and building upon a more advanced modelling of the Long Short-Term Memory (LSTM) network onward and achieving higher accuracy as a result of improved features processing, high values of sensitivity and specificity. Those LSTM models robustly performed adequately in a noisy environment compared with Aggarwal et al.'s submission in 2023. Similarly, Gülmez et al. (2024) applied CNN for the automatic classification of cry sounds, augmenting the data to improve the accuracy of the performance. Özseven (2022) reviewed deep learning methods and discussed that these techniques are increasingly being used as compared to traditional approaches. In spite of the promising results of the deep learning studies, they often had to struggle with computational demands, which thus reflects the trade-off between model complexity and real-world feasibility. The effectiveness of LSTM and CNN in these studies shows that deep learning can greatly raise the bar in terms of performance for a classification problem, especially when combined with techniques like data augmentation. However, this computational burden calls for lighter architectures that retain high accuracy, a focus area for the current research.

2.5 Hybrid and Transfer Learning Methods

Deep learning combined with other methods is another aspect that has also found its way into baby cry detection. G. A et al. (2024) CNN combined with LSTM hybrid models was applied by Reddy et al. (2023)Reddy et al., in 2023 to detect infant cries from noise with an accuracy of 99%. Sharma and Malhotra (2020) proposed the intelligent Infant Cry Classifier IICC using CNN and Decision Trees. While hybrid models often outperform single architectures by complementing each other's strengths, the general weakness of such studies is usually computation-intensive. Another way of trying to mitigate dataset limitations has involved the use of transfer learning. Anjali et al. (2022), in their work, performed transfer learning using VGG16 and attained the highest accuracy compared to the rest of the models they have compared. This would ensure that Zhang et al. (2023) presented a BCRNet model with built-in transfer learning and feature fusion strategies to avoid overfitting. The power in this approach is leveraging potentiality from pre-trained networks. There does stand the problem of their sensitivity with regard to real-world conditions and different noisy conditions. These few works on hybrid and transfer learning techniques have informed the present research in a manner that could hopefully indicate how it can improve model accuracy with much less training data, even while attention is given to real-world performance.

2.6 The Internet of Things - IoT and its application in real life

Although many research works are directed to practical applications of baby cry detection systems, the scope of their applicability is further extended by integrating them with the IoT. Younis et al. (2024) addressed this using Vision Transformers and CNN with IoT sensors for real-time capture of the cry signals. Although this has increased the responsiveness of the system because of IoT integration, a major bottleneck is the requirements of high computational resources limit its extension to resource-constrained settings. Similarly, Kolandaisamy et al. proposed a Raspberry Pi and wireless sensor networks-based baby cry detection system that could be implemented practically with regard to various real-world monitoring scenarios in 2022. In 2021, Visvesvaran et al. (2021), came up with some add-ons that may develop a more feasible IoT-based intelligent baby monitoring system whereby features of automatic cradle swinging with alerting through email to the parents may be carried out. Therefore, the feasibility of an integration approach is also very limited, as identified in two discussed studies about cry analysis integration with IoT technologies into practical complexities such as hardware integration challenges and scalability issues. These findings from various implementations point to the need for lightweight and scalable solutions that can be used effectively within residential environments—a focus of the present research project on practical cry detection.

2.7 Gaps and Limitation

The literature reviewed indicates the extensive application of machine learning and deep learning models, from classic methods of SVM and Random Forest to complex methods including LSTM and Vision Transformers, in infant cry classification. These proposed machine learning models are simpler and have lesser computational requirements. However, in the presence of complex cry signals and background noise, most of the machine learning-based methods handle them poorly. While on the other hand, deep learning-based methods provide superior accuracy, consuming lots of data and greater computational resources. Hybrid approaches with the incorporation of traditional machine learning combined with deep learning techniques present encouraging results, leveraging strengths from both methodologies. In turn, data augmentation, transfer learning, and hybrid models achieve better results but, unfortunately, often at the expense of higher computational requirements besides the processing issues introduced by real-world noise. The noticeable gaps are on one hand robust models that keep accuracy high without excessive demands for more resources and, on the other hand, good generalization in diversified and noisy conditions. This research will contribute to bridging these gaps by seeking more efficient hybrid models and techniques that are robust against noise, hence improving practical applicability in realistic conditions for infant cry classification systems.

Table 1 provides summary of the recent literature.

Authors	**Year**	**Method**	**Findings**
B. N. Pradhan et al.	2022	Pre-processing, MFCC, Spectrogram, Machine Learning Classifiers	EfficientNet and ResNet were used for the classification of baby cries to provide a reliable system for hearing-impaired caregivers. Performance considerably deteriorated in noisy environments.

Gaurav Aggarwal et al.	2023	Random Forest, SVM, Decision Tree, Spectral Features	SVM outperformed all the other classifiers. As the cry signals are analyzed in detail, the generalization of the results is not possible with a little dataset.
Nita Nimbarte et al.	2023	MFCC, KNN	The proposed cry detection system worked well for a residential environment with high accuracy and poorly robust when handling diversified background noise.
S. Narayanan et al.	2024	LSTM, Spectrogram	The LSTM-based approach showed very high accuracy and specificity but low practical applicability due to frequent false alarms.
S. A. Younis et al.	2024	Vision Transformers, CNN, IoT-Enabled Sensors	High accuracy is achieved by the Vision Transformers: Resource requirements were a bottleneck for low-resource settings.
Putri Agustina Riadi et al.	2024	SVM, Random Forest, Naïve Bayes, MFCC	Random Forest performed the best. The study provided very good feature extraction but did not explore deep learning methods satisfactorily enough.
I. Gülmez et al.	2024	ANN, CNN, Data Augmentation	Data augmentation improved accuracy but the small dataset hindered generalization to diverse populations.
P. Kumari and K. Mahto	2024	Pathological Cry Detection, ML, DL	Focused on the detection of pathological cries. The limited dataset and availability of pathological samples restricted the applicability of findings.
B. M. Mala et al.	2024	Leaky Bi-LSTM, Improved Artificial Rabbit Optimization (IARO)	Highly accurate, computationally intensive, and where the model performance suffered with changing environmental noise conditions.
G. Anjali et al.	2024	Transfer Learning (VGG16)	Achieved high accuracy using transfer learning. Performance degraded seriously under a noisy and uncontrolled environment.
T. Özseven	2024	Computer-Aided Diagnoses Review	Detailed review upheld the transition from traditional methods to deep learning. Few experimental validations made it less practically relevant.
G. A. Senthil et al.	2024	CNN, LSTM, Hybrid Approach	Hybrid techniques provide high accuracy but are computationally expensive. Ethical and privacy considerations were adequately addressed.
Bella, V. et al.	2024	Data Augmentation, MFCC, LSTM	Time-stretching and pitch-shifting helped the model significantly. The small dataset limited generalization.
Kolandaisamy et al.	2024	Raspberry Pi, Wireless Sensor Networks	Practical implementation for the purpose of baby monitoring. Its complexity and dependence on various hardware components limited scalability.
Khandelwal et al.	2024	Convolutional Recurrent Neural Network (CRNN)	Developed a low-complexity CRNN, focused on practical application at the cost of accuracy.
Zhang, Ke et al.	2024	BCRNet, Transfer Learning, Feature Fusion	Transfer learning and feature fusion techniques have avoided overfitting, thereby increasing the accuracy. It had some drawbacks in the form of computational demands regarding real-time applications.
Reddy, Keshavagari et al.	2024	Hybrid CNN and LSTM	High accuracy in distinguishing the different types of baby cries but lacking analysis under different noise conditions.
Jindal, S. et al.	2024	Acoustic Features, Spectrogram, Conventional ML and DL	The acoustic features along with spectrograms showed satisfying performances. Limitations in the used dataset have influenced the robustness of the models.
Visvesvaran et al.	2024	IoT-Based Smart Baby Monitoring	IoT-based enhanced baby monitoring system is presented here. Its affordability and maintenance are at a challenge.
Ashwini et al.	2020	CNN, SVM, KNN, Bayesian Hyperparameter Tuning	CNN for feature extraction followed by SVM classification. The best performing was SVM with Bayesian tuning. Real-time applications are very limited since this study had a limited dataset and utilized few spectrograms.
Mahmoud et al.	2020	Semi-Supervised K-Nearest Neighbor (SSKNN), Google AudioSet	Semi-supervised approach: Increasing the size of the training set led to improved classification performance, though at some computational cost.
Sharma et al.	2020	Intelligent Infant Cry Classifier (IICC), CNN, Decision Trees	IICC was targeted to identify the reason for an infant crying. Although it uses robust pre-processing in order to achieve high-quality data, generalizability remained low because of the dataset size.
Jamal et al.	2023	Prosodic and Spectral Features, KNN, ANN, CNN	While the hybrid feature set improved accuracy, the computational intensity of the methods limited scalability.

You, Weijie et al.	2023	LSTM, MFCC, Spectral Features	This study can classify multiple cry types with a good accuracy, but it is restricted to dataset size and computational demand.

Table 1: Summary of literature review

3 Methodology

This study uses a quantitative research model aimed to classify infant cries into hunger, discomfort and pain categories. The research process consists of various stages, data collection, data pre-processing, feature extraction, model selection, model training, and evaluation. Based on section 2, each phase is corresponded to research gaps through continuously improving noise robustness and classification accuracy. This research mostly relied on experimental simulations for the training and testing of models with a combination of classical machine learning and modern deep learning models similar to the other related studies (Aggarwal et al. (2023), G. A et al. (2024), Narayanan et al. (2024)). Figure 1 outlines each phase of the research including data pre-processing, MFCC feature extraction, data splitting, data imbalance handling with SMOTE, model training, evaluation and results interpretation.

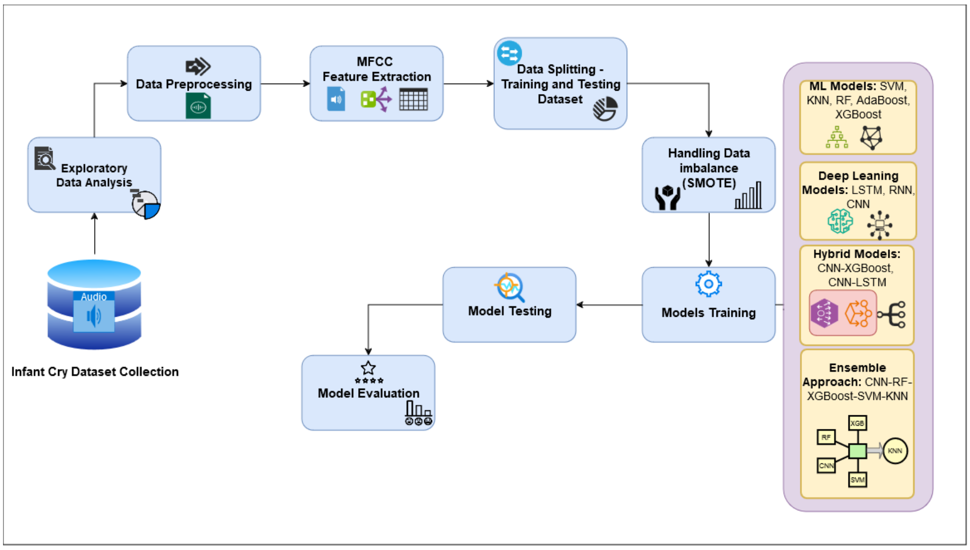

Figure 1: Research Methodology

3.1 Data Collection and Overview

This study used a Kaggle data set available at [1]. It included audio samples of baby cries, each labeled according to the type of cry , labels included - *hungry*, *discomfort*, *tired*, *belly pain*, and *burping*. Exploratory Data Analysis (EDA) was performed to understand the dataset and find potential issues with the data, which involved analysing class distribution, MFCC feature distribution and audio lengths of different classes.

[1] Kaggle url: https://www.kaggle.com/datasets/warcoder/infant-cry-audio-corpus

3.1.1 Class Distribution

The original dataset had an imbalance of the classes, with the highest count labeled category *hungry* from Figure 2. This was balanced out using SMOTE in the later stages to create synthetic samples of the minority classes.

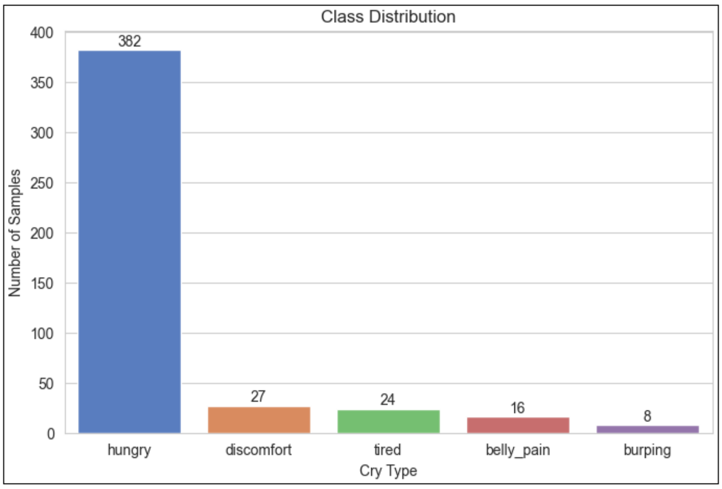

Figure 2: Class Distribution

3.1.2 Distribution of First 5 MFCC Features

By analyzing the MFCC feature distribution from Figure 3, the sounds variability was studied in this scenario. The first five MFCC coefficients were plotted, revealing the underlying patterns and correlations in the audio data.

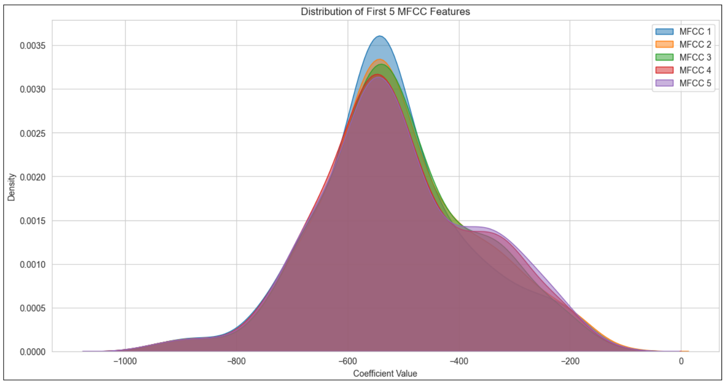

Figure 3: Distribution of first five MFCC

3.1.3 Distribution of duration of audio by class

Also, the proportion of audio duration was analyzed for different cry types. It differed slightly from class to class, and some outliers were indicative of small time length differences of audio samples [Figure 4].

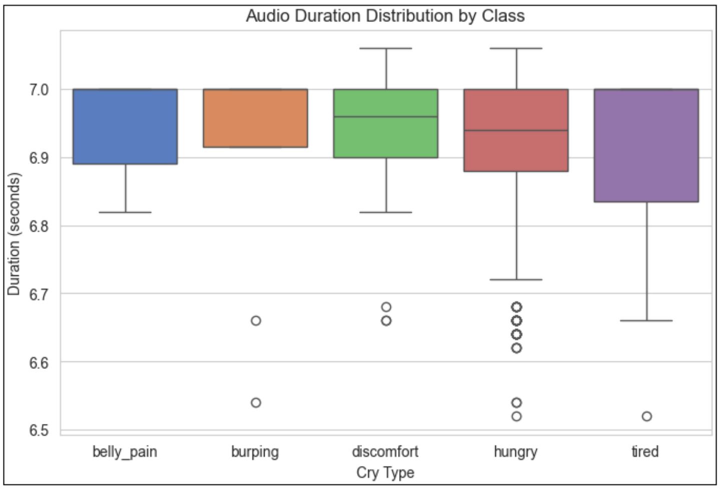

Figure 4: Distribution of duration of audios

3.2 Data Preprocessing

Data is loaded using Python libraries to clean and manipulate further. In preprocessing steps, data is cleaned by removing noise from audio signals. This data is then standardized using StandardScaler.

3.3 MFCC Feature Extraction

The time based audio feature Mel Frequency Cepstral Coefficients (MFCCs) was extracted from the audio signals. This resulted in numerical data to represent input of the sound signals, which are effective for analyzing and classifying various baby cries Riadi et al. (2024).

3.4 Data Splitting

The dataset was split into an 80-20 split for training and test sets, respectively. In this respect, the former served for model development, whereas the latter was held out for evaluation purposes. Besides, 5-fold cross-validation was applied to ensure the robustness and avoid overfitting for better generalization of performance of the model.

3.5 Handling Data Imbalance

As it can be seen from the dataset, the class distribution is extremely unbalanced (Hungry:382, Discomfort:27, Tired: 24, Belly Pain: 16, Burping: 8) where hungry is the majority. This, however, created an imbalance, and consequently Synthetic Minority Over-sampling Technique (SMOTE) was applied. As per (Kumari and Mahto (2024), Zhang et al. (2023)) the SMOTE technique creates samples for minority classes to balance the data, this balanced data further used in the model training results in improved performance.

3.6 Reshaping Data for Deep Learning Models

This high dimensional MFCC data was required to be reshaped while training deep learning models such as LSTM and CNN. 3D input (for 2D input images, each of the MFCC features were reshaped to have 3D array so as to meet the input requirements of the deep learning architectures). This restructuring enabled LSTM and CNN models to effectively learn temporal dependencies present in the audio signals.

3.7 Model Training

Based on the literature review of previous work (G. A et al. (2024),Zhang et al. (2023),Bella and Sanjaya (2023)), various classic and advanced machine learning models have been chosen. The following models are trained as part of this research:

3.7.1 Traditional Machine Learning Models

Several classical machine learning models were compared in this work using 5-fold cross-validation to ensure reliable performance estimation. The SVM algorithm was selected based on its strong performance in scenarios where data is high dimensional and was previously applied to different but related audio classification tasks successfully, such as in Aggarwal et al. (2023). K-Nearest Neighbors: KNN is a non-parametric model, and it was used due to its simplicity. KNN works well when the number of instances or observations is small, in which one can see how close neighboring labeled instances are affecting the classification results Nimbarte et al. (2021). Followed by Random Forest model, which is capable of avoiding overfitting using multiple decision trees and has already shown great results in the field of cry classification Riadi et al. (2024). Similarly, boosting was utilized through AdaBoost to achieve progressive improvements in the classification rate by focusing iteratively on misclassified samples Younis et al. (2024). Finally, XGBoost is a computationally efficient gradient boosting scheme, included here for demonstrated scalability and great performance found in prior contexts of audio classification by Narayanan et al. (2024).

3.7.2 Deep Learning Models

As discussed by Anjali et al. (2022), Long Short Term Memory (LSTM) model is efficient in time dependent features in audio signals, so it is chosen to handle the baby cry classification. Mala and Darandale (2024) showcased the strong capability of Recurrent Neural Network (RNN) in analyzing sequential sound signals, which are explored in the study. Additionally, Convolutional Neural Network (CNN) was employed as it can extract spatial features from audio data, which helps to understand the acoustic structure of the infant cries Gülmez et al. (2024).

3.7.3 Hybrid Models

Hybrid models CNN-LSTM and CNN-XGBoost are employed for infant cry classification. The proposed hybrid models use both CNNs for feature extraction on spectrograms and LSTMs for temporal sequence learning (Zhang et al. (2023), Bella and Sanjaya (2023)). CNN detects spatial patterns in the audio frequency distribution, whereas LSTM or XGBoost incorporates time-dependent information or combines multiple classifiers to improve generalization performance.

3.7.4 Ensemble Approach

Ensemble Model combines the strengths of different algorithms to bring up better and generalized classifier (G. A et al. (2024),Reddy et al. (2023)). In the proposed model Convolutional Neural Network (CNN), Random Forest, Support Vector Machine (SVM) and K-nearest Neighbors are employed together to improve overall accuracy and stability by reducing individual weaknesses of each model.

3.8 Evaluation Methodology

The main evaluation strategy was cross-validation for unbiased performance of the models. All the models were assessed with some evaluation metrics, such as accuracy, precision, recall, and F1 score combined with 5-fold cross-validation. These have been used to assure comparability and consistency in various machine learning algorithms in a coherent way (Bella and Sanjaya (2023),Younis et al. (2024)). Misclassification patterns are analyzed using confusion matrices G. A et al. (2024). All deep learning models in this study were trained using the categorical cross-entropy loss function combined with early stopping. The model would stop training if it could not improve the results of validation loss.

4 Design Specification

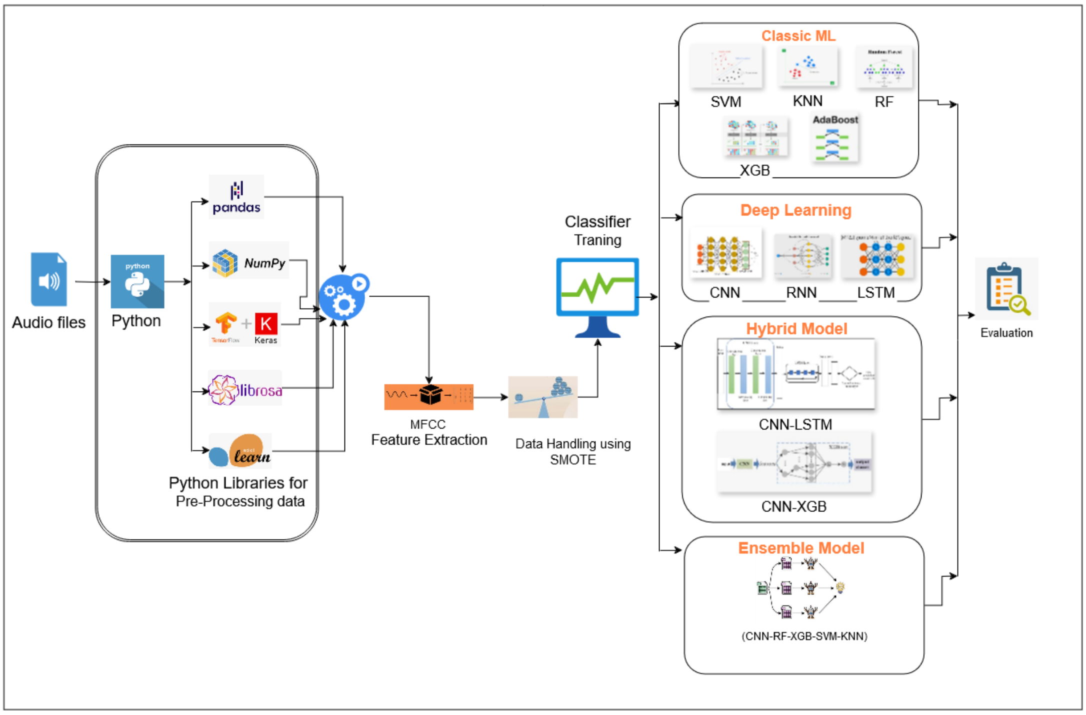

Figure 5: Design Specification Components

The overall design of the Baby Cry Classification System integrated all components discussed in Section 3 such as pre-processing, feature extraction with machine learning and deep learning models, and augmentation of the data based on previous discussions from related work in the field. It basically employed the classical machine learning approach along with deep learning models in its framework for a strong classification system that can be deployed on real-time applications with efficiency. This work relies fundamentally on the key technique of Mel Frequency Cepstral Coefficients for feature extraction. MFCC is usually employed in audio signal processing due to its spectral properties being close to the perceptions of human hearing. Also, SMOTE has been utilized in this regard for overcoming data imbalance by artificially creating new synthetic samples of the minority classes, so as to improve classifier performance. Component-wise details are shown in Figure 5:

4.1 Algorithm and Model Description

4.1.1 MFCC Feature Extraction

MFCC extracts coefficients that are representative of the spectral properties of the audio, reducing the raw audio data's complexity while retaining most of the features relevant to human perception. According to Aggarwal et al., 2023, and Riadi et al., 2024, MFCC feature extraction is done by first breaking down the audio into overlapping frames and applying a window function and then computing the Fourier transform and Mel scale filterbanks. Figure 6 shows the steps involved in MFCC feature extraction process.

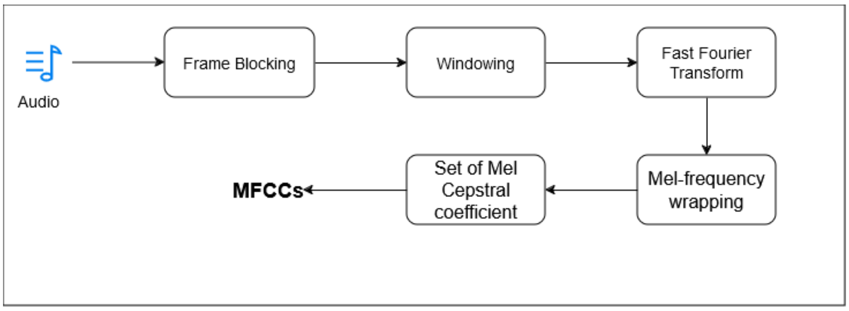

Figure 6: MFCC Feature Extraction Process

4.1.2 Handling Imbalanced Data using SMOTE

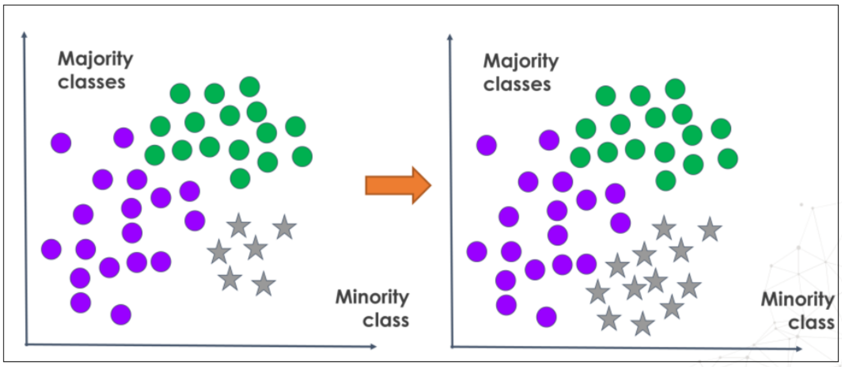

Figure 7: Random Over Sampling using SMOTE

Synthetic Minority Over-Sampling Technique (SMOTE) created synthetic samples for the minority classes to handle imbalanced datasets. Therefore, it avoids any model biases toward the majority classes in improving the overall accuracy of classification. It works by generating synthetic samples between existing instances of the minority class based on their nearest neighbors, this functionality can be seen in below Figure 7 [2].

4.1.3 Hybrid Models

The hybrid architecture is a combination of traditional and deep learning models. Traditional models such as SVM and Random Forest were considered because of their simplicity and effectiveness in low-dimensional feature space. The deep learning models, such as CNN, LSTM, and RNN were chosen for capturing higher temporal and spatial relationships (Zhang et al. (2023), Bella and Sanjaya (2023)). Aiming to propose a robust classifier, CNN-LSTM and CNN-XGBoost models were trained to accommodate the strengths of these individual models, such as capturing temporal patterns and convolutional feature extraction.

CNN + LSTM: Figure 8 shows design of this hybrid model, which combined CNN for feature extraction with LSTM for capturing temporal relationships in the data.

CNN + XGBoost: This model combined the feature extraction capability of CNN with the classification strength of XGBoost. The design flow is as presented in Figure 9 .

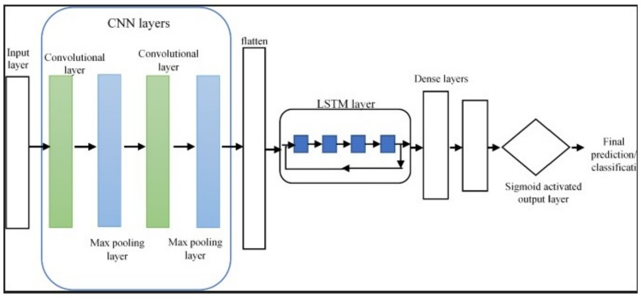

Figure 8: CNN + LSTM Hybrid Model

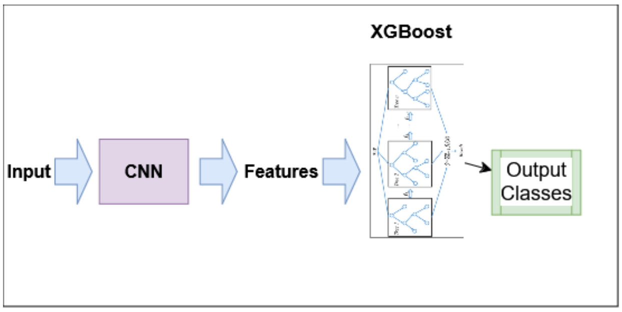

Figure 9: CNN + XGBoost Hybrid Model

4.1.4 Ensemble Approach

The ensemble model combines multiple classifiers (CNN, Random Forest, XGBoost, SVM, KNN in Figure 10) to enhance performance by reducing variance and bias. The ensemble approach aggregates the predictions from multiple models, thereby increasing reliability(G. A et al. (2024), Reddy et al. (2023)). The majority vote or weighted averaging method was used to determine the final prediction.

[2] Image Source: `https://www.blog.trainindata.com/oversampling-techniques-for-imbalanced-data/`

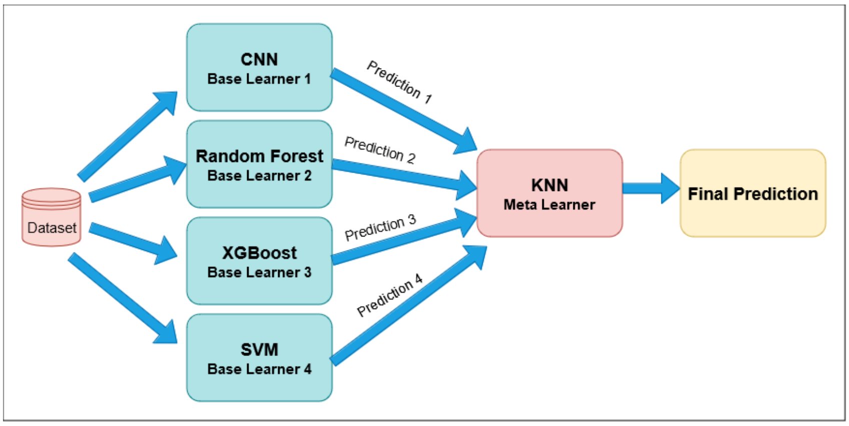

Figure 10: Ensemble Model (CNN, Random Forest, XGBoost, SVM, KNN)

5 Implementation

The implementation step for the proposed system of baby cry classification required incorporating data pre-processing, feature extraction, training, and testing into a workable pipeline. It describes models developed and tools utilized during implementation.

5.1 Model Training

The models used for infant cry classification are implemented as follows:

5.1.1 Traditional Models

Model Name	Parameters
SVM	Kernel: Linear, Random State: 42, Cross-Validation: 5-fold
KNN	Neighbors (n_neighbors): 5, Random State: 42, Cross-Validation: 5-fold
Random Forest	Number of Estimators: 100, Random State: 42, Cross-Validation: 5-fold
AdaBoost	Number of Estimators: 100, Random State: 42, Cross-Validation: 5-fold
XGBoost	Number of Estimators: 100, Evaluation Metric: mlogloss, Label Encoder: False, Random State: 42, Cross-Validation: 5-fold

Table 2: Parameters for Classic ML Models

It uses the classical machine learning approach with five different classifiers - SVM, KNN, Random Forest, AdaBoost, and XGBoost, along with SMOTE to deal with class imbalance, and rigorous 5-Fold Cross-Validation. The SVM, as applied here with a linear kernel and a fixed random state, constitutes an easy and effective solution to class separation within a high-dimensional feature space. KNN was adopted for its memory-based learning principles that classify points by finding those closest to the instance and thereby achieve a favorable tradeoff between simplicity and ease of use, given an equal seed and n_neighbors set to 5. Below Table 2 shows parameters used to implement these model.

5.1.2 Deep Learning Models

Deep learning models for CNN, LSTM, and RNN were applied for baby cry classification using parameters presented in Table 3. CNN makes use of convolutional and pooling layers to capture temporal features of the input and to reduce overfitting. The LSTM was designed to efficiently model temporal dependencies in data by making use of stacked layers together with dropout, balancing classes in the dataset. A simple RNN using stacked SimpleRNN layers was implemented to capture sequential patterns. All the models will make use of Adam optimizer, softmax outputs, and use categorical cross-entropy as the loss, with the implication of all models assuring robust performance in multi-class problems.

Model Name	**Parameters**
CNN	Timesteps:6, Conv1D Filters:64, Kernel Size:3, MaxPooling Pool Size:2, Dropout Rate:0.5, Dense Output Layer:Softmax activation, Optimizer:Adam, Loss Function:Categorical Cross-Entropy, Epochs:20, Batch Size:32
LSTM	LSTM-Units (Layer 1):128, LSTM Units (Layer 2):64, Dropout Rate:0.3, Dense Layer:32 units, ReLU activation, Output Layer:Softmax activation, Optimizer:Adam, Loss Function:Categorical Cross-Entropy, Epochs:50, Batch Size:32
RNN	SimpleRNN Units:128 (Layer 1), 64 (Layer 2), Dropout Rate:0.3, Dense Layer:32 units, ReLU activation, Output Layer:Softmax activation, Optimizer:Adam, Loss Function:Categorical Cross-Entropy, Epochs:50, Batch Size:32

Table 3: Parameters for Deep Learning Models

5.1.3 Hybrid Models

This hybrid approach proposes CNN-XGBoost and CNN-LSTM with respective parameters Table 4. These models consolidate many features, which contribute to refining the classification of baby's crying. The CNN-XGBoost model integrates convolution features together with an extreme boosting technique. Further, CNN-LSTM performs spatial feature learning-based convolutional neural networks coupled with time based feature identification using long short-term memory. LSTM incorporated dropout to avoid overfitting and get back multi-class classification with adaptive learning.

Model Name	**Parameters**
CNN-XGBoost	Conv1D Filters:64, Kernel Size:3, MaxPooling Pool Size:2, XGBoost Estimators:100, Cross-Validation:5-fold
CNN-LSTM	Conv1D Filters:64,Kernel Size:3,MaxPooling Pool Size:2,LSTM Units:64,Dropout Rate:0.5,Dense Output Layer:Softmax activation,Optimizer:Adam,Loss Function:Categorical Cross-Entropy,Epochs:20,Batch Size:32

Table 4: Parameters for Hybrid Learning Models

5.1.4 Ensemble Approach

Ensemble model implemented using a combination of CNN, Random Forest, XGBoost, SVM, and KNN, parameters are as shown in Table 5. It provides the improvement of predictive performance due to CNN feature extraction combined with ensemble and stacking in a classification model. Its base classifiers are as follows: The Random Forest uses 100 estimators in order to make robust predictions based on this ensemble of decision trees, SVM -linear kernel- is applied since it works efficiently in high-dimensional feature spaces, while the XGBoost used will utilize 100 estimators along with a maximum depth of 5 to allow fine-grain boosts. Therefore, diverse perspectives from the base classifier models will be achieved. For these purposes, a KNN meta-classifier uses k = 5 neighbors; output from base learners balances between bias and variance within StackingClassifier. This setup, which is evaluated with 5-fold cross-validation, uses the complementary strengths in learning optimally to ensure robust and interpretable classification.

Model Name	Parameters
Ensemble Approach Model	Feature Extraction:Trained CNN,Base Learner 1:Random Forest (100 estimators),Base Learner 2:SVM (linear kernel),Base Learner 3:XGBoost (100 estimators, max depth = 5),Meta-Learner:KNN (5 neighbors),Stacking Framework:StackingClassifier with 5-fold CV

Table 5: Parameters for Ensemble Approach Model

5.2 Tools and Languages Used

The programming language Python was used for the model building due to its flexibility in handling data processing, model training, and evaluation. Development environment used was Jupyter Notebook as it has an user friendly interface which helps in coding, visualizations, and debugging. The model development was done with the help of multiple Python libraries. To address the problem of converting audio data into mathematical format Librosa was used to facilitates noise reduction and MFCC feature extraction. For machine learning models, cross-validation as well as SMOTE for data imbalance Scikit-learn was used. Deep learning models were implemented using TensorFlow and Keras. NumPy and Pandas were instrumental for data manipulation. Visualizations were created using Matplotlib and Seaborn for EDA and performance metrics such as accuracy and confusion matrix.

6 Evaluation

The evaluation of the baby cry classification system was done by comparing the model performances based on their accuracy, F1-score, precision and recall. The following table presents the comparison results:

Trained Models	Without SMOTE				With SMOTE			
	Accuracy	F1 Score	Precision	Recall	Accuracy	F1 Score	Precision	Recall
SVM (5-fold CV)	0.767	0.749	0.732	0.767	0.977	0.976	0.978	0.977
KNN (5-fold CV)	0.838	0.775	0.72	0.838	0.844	0.803	0.87	0.844
Random Forest (5-fold CV)	0.849	0.78	0.721	0.849	0.996	0.996	0.996	0.996
AdaBoost (5-fold CV)	0.759	0.736	0.739	0.759	0.294	0.208	0.259	0.294
XGBoost (5-fold CV)	0.844	0.777	0.721	0.844	0.989	0.989	0.989	0.989
LSTM	0.79	0.741	0.699	0.79	0.754	0.731	0.71	0.754
RNN	0.826	0.757	0.699	0.826	0.717	0.702	0.688	0.717
CNN	0.793	0.741	0.718	0.793	0.685	0.658	0.64	0.685
CNN-XGBoost	0.793	0.741	0.718	0.793	0.983	0.983	0.984	0.983
CNN-LSTM	0.783	0.687	0.612	0.783	0.37	0.433	0.571	0.37
Ensemble Model (CNN, RF, XGB, SVM, KNN)	0.728	0.67	0.62	0.728	0.707	0.684	0.679	0.707

Table 6: Comparison of Models with and without SMOTE

6.1 Experiment 1 : Traditional Machine Learning Models

Traditional models of machine learning compared in this paper include the support vector machine, K-Nearest Neighbor, Random Forest, AdaBoost, and XGBoost. All these models have been trained with 5-fold cross-validation in order to further robust the results in performance evaluations. Of these compared models, Random Forest stands at the highest among them, preceded by XGBoost. The SVM also performed very well, while AdaBoost lagged, as it is sensitive to unbalanced data. The RF model with SMOTE outperforms the XGBoost model with SMOTE on each of the metrics: accuracy, F1 score, precision, and recall show near-perfect scores in 0.996, from Table 6. In contrast, the XGBoost has performed a little inferior at 0.989 for all the metrics of this study; hence, their classification performance can also be reckoned as very good.

The results emphasize that Random Forest is robust and works effectively on the resampled dataset. Table 7 presents the classification report for the best performing models (Random Forest and XGBoost) . Figures 11 and 12 show the confusion matrix.

Classes	**Random Forest**				**XGBoost**			
	Precision	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**
belly_pain	0	0	0	4	0	0	0	4
burping	0	0	0	2	0	0	0	2
discomfort	0	0	0	7	0	0	0	7
hungry	0.79	0.99	0.88	72	0.78	0.97	0.86	72
tired	0	0	0	7	0	0	0	7
Accuracy	0.77				0.76			
Macro Avg	0.16	0.20	0.18	92	0.16	0.19	0.17	92
Weighted Avg	0.62	0.77	0.69	92	0.61	0.76	0.68	92

Table 7: Classification Report for Random Forest and XGBoost Models

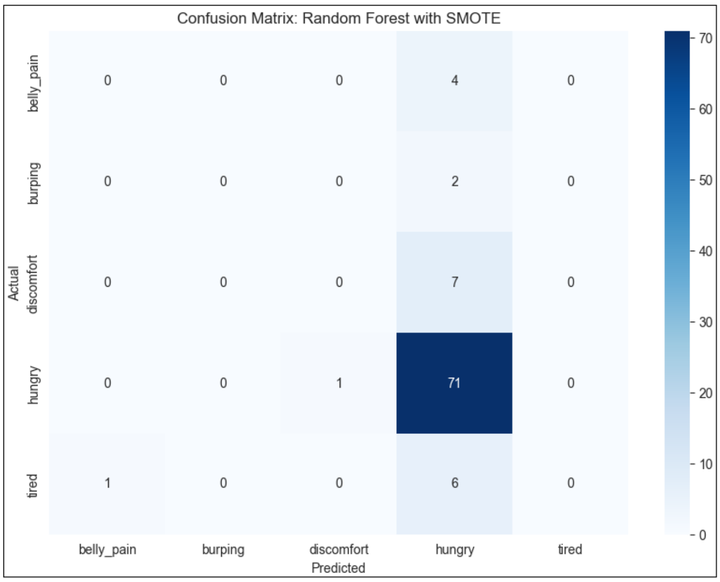

Figure 11: Confusion Matrix - Random Forest

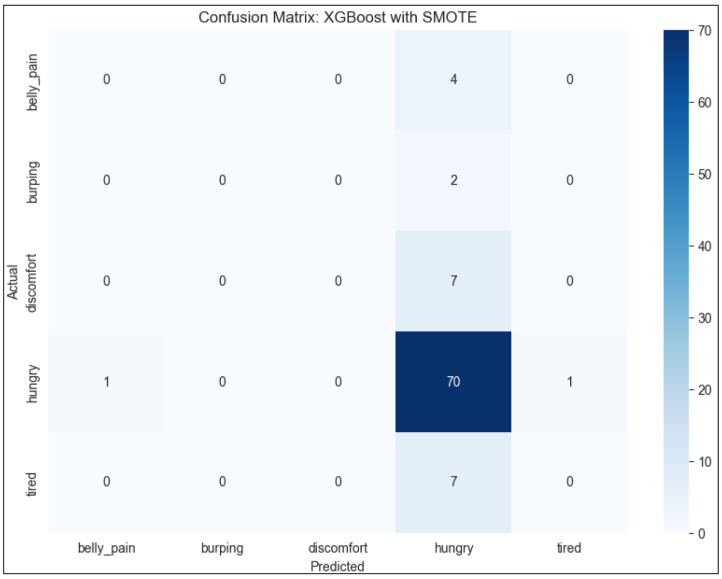

Figure 12: Confusion Matrix - XGBoost

6.2 Experiment 2: Deep Learning Models

In this paper, deep learning models such as LSTM, RNN, and CNN are trained in order to capture the temporal and spatial features of the baby cries. Deep learning models take raw MFCC features as input. Among deep learning models, LSTM achieved the highest accuracy(0.754), F1-score (0.731) Precision (0.71) and Recall(0.754), followed by RNN and CNN. Deep learning models showed very poor performance compared to traditional models with the SMOTE technique, thus, it holds very strong potential when hybrid approaches are considered. The classification report and confusion matrix for LSTM models are shown in Table 8 and Figure 13.

Classes	**LSTM**				**CNN-XGBoost**			
	Precision	**Recall**	**F1-Score**	**Support**	**Precision**	**Recall**	**F1-Score**	**Support**
belly_pain	0.25	0.20	0.22	5	0	0	0	4
burping	0	0	0	3	0	0	0	2
discomfort	0.09	0.12	0.11	8	0.25	0.14	0.18	7
hungry	0.84	0.87	0.85	115	0.79	0.94	0.86	72
tired	0	0	0	7	0	0	0	7
Accuracy	0.74				0.75			
Macro Avg	0.24	0.24	0.24	138	0.21	0.22	0.21	92
Weighted Avg	0.71	0.74	0.73	138	0.64	0.75	0.69	92

Table 8: Classification Report for LSTM and CNN-XGBoost Models

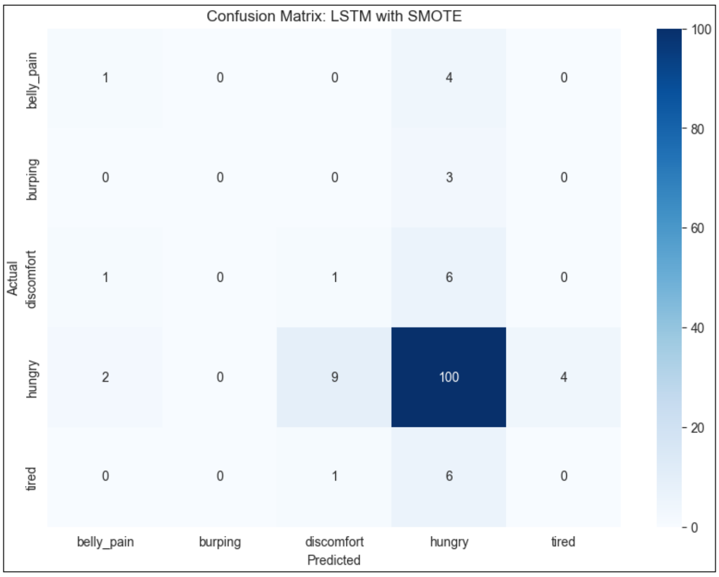

Figure 13: Confusion Matrix - LSTM

6.3 Experiment 3: Hybrid Models

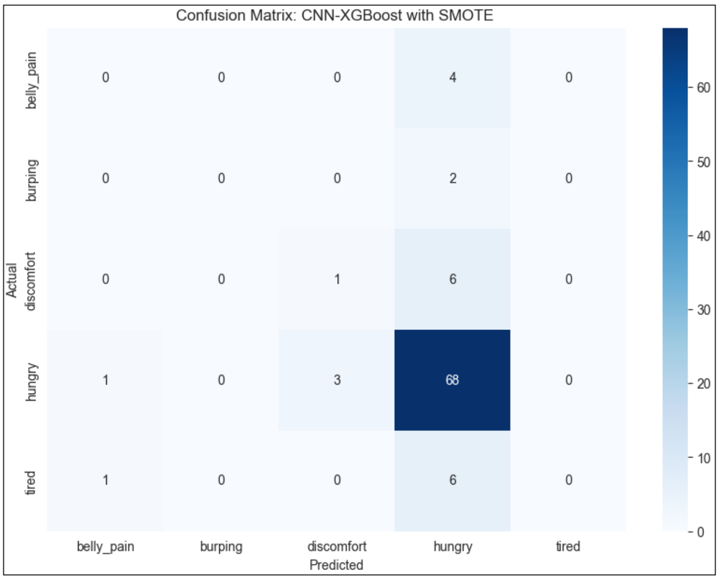

Figure 14: Confusion Matrix - CNN-XGB

Hybrid models combined the feature extraction capabilities of CNN with ensemble methods, such as XGBoost or LSTM for temporal dependencies. CNN-XGBoost outperformed CNN-LSTM, showing the combination of CNN's spatial feature extraction with XGBoost's ensemble learning capabilities. The CNN-XGBoost with SMOTE model has performed very well, with an accuracy of 0.983, F1 Score-0.983, precision-0.984, and recall of 0.983 [Table 6], which is a bit lower than the individual performances of Random Forest and XGBoost. The results of CNN-XGBoost still show its robust multi-metric performance due to the combination of CNN's feature extraction and XGBoost classification strengths. Table 8 and Figure 14 presents the classification report and confusion matrix for the CNN-XGBoost model.

6.4 Discussion

The outcomes of this experiment show classic models Random Forest and XGBoost models trained on the balanced data obtained from SMOTE outperforming the models based on the original dataset, achieving accuracies and F1-scores of 99% and 98%, respectively. This highlights that imbalanced data handling is crucial for the dataset used. The generated synthetic data helped in better detection of minority cry types like *'burping'* and *'belly-pain'*, which leads to the conclusion, that better balancing gives better infrequent class detection. The deep learning models (LSTM, RNN and CNN) performed moderately with 75-80% accuracy, when hybrid models like CNN-XGBoost gave even better results, as it combine the strengths of both CNN and XGBoost, feature extraction and ensemble decision making. The proposed ensemble (CNN, Random Forest, XGBoost, SVM, and KNN) model did not perform well compared to the chosen individual models trained with SMOTE. These findings highlight the importance of data augmentation with hybrid and ensemble approaches for improving the robustness of the system for baby cry classification. This work can be utilized in the future innovations of infant monitoring systems and integration for real-world application.

7 Conclusion and Future Work

This study is dedicated to building a reliable baby cry classifier with the purpose of determining different classes of baby cries using a variety of machine learning and deep learning models, after which the major objectives focused on exploring different models regarding overcoming class imbalance challenges and enhancing the accuracy of classification by leveraging appropriate feature extraction techniques besides oversampling methods. This is evidenced by the experimental work, which has been able to successfully prove that models trained on SMOTE-augmented datasets perform quite well, with Random Forest and XGBoost outperforming other models in terms of accuracy, precision, recall, and F1 score.

Some key inferences from this work include that SMOTE to address the imbalance significantly improves the performance of models, and the CNN ensembled with XGBoost yields high performance compared to using each technique in a separate manner. These were some of the disadvantages too, in terms of high computational costs and problems while dealing with real-time cry signals, as those are problematic to most of the deep learning techniques such as LSTMs.

Future Work: Based on the literature reviewed, a few suggestions can be given about the future research to enlarge the domain of baby cry classification: Investigating transformer-based models would be a good suggestion for future work based on recent successes in sequence modeling. Besides, more extensive datasets, including even pathological types of cries, may serve to enhance the robustness and applicability of the models in real-world applications. The development of an edge-compatible version of the classification system that can perform real-time processing on low-power devices will help bridge the gap between research and practical commercial applications.

References

Aggarwal, G., Jhajharia, K., Izhar, J., Kumar, M. and Abualigah, L. (2023). A machine learning approach to classify biomedical acoustic features for baby cries, *Journal of Voice* **ISSN 0892-1997**.

Anjali, G., Sanjeev, S., Mounika, A., Suhas, G., Reddy, G. P. and Kshiraja, Y. (2022). Infant cry classification using transfer learning, *TENCON 2022 - 2022 IEEE Region 10 Conference (TENCON)*, IEEE, pp. 1–7.

Bella, V. and Sanjaya, S. (2023). Refining baby cry classification using data augmentation (time-stretching and pitch-shifting), mfcc feature extraction, and lstm modeling, *2023 7th International Conference on New Media Studies (CONMEDIA)*, IEEE, pp. 250–256.

G. A, S., S, G., Tharagarani, G., P, S. and B, S. (2024). An automated mood analysis of crying infants through sound recognition using hybrid deep learning, *2024 Third International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*, IEEE, pp. 1–6.

Gülmez, I., Kayan, M. Y. and Demirci, M. F. (2024). Automatic recognition of baby crying sounds, *2024 32nd Signal Processing and Communications Applications Conference (SIU)*, Mersin, Turkiye, pp. 1–4.

Jamal, A. and Al-Azani, S. (2023). A machine-learning approach for children's pain assessments using prosodic and spectral acoustic features, *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pp. 1–6.

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

Kolandaisamy, R., Al-Mashhadani, A. F. S., Nandy, T. and Keat, T. F. (2022). Speech signal processing based on machine learning and complex processors for baby cry detection system, *Journal of Positive School Psychology* **6**(2): 2193–2207.

Kumari, P. and Mahto, K. (2024). A narrative review on different novel machine learning techniques for detecting pathologies in infants from born baby cries, *Journal of Voice: Official Journal of the Voice Foundation* . Preprint.

Mahmoud, A. M., Swilem, S. M., Alqarni, A. S. and Haron, F. (2020). Infant cry classification using semi-supervised k-nearest neighbor approach, *2020 13th International Conference on Developments in eSystems Engineering (DeSE)*, pp. 305–310.

Mala, B. M. and Darandale, S. S. (2024). Effective infant cry signal analysis and reasoning using iaro based leaky bi-lstm model, *Computer Speech Language* **86**.

Narayanan, S., Sabarimalai, M., Linga, M. and Cenkeramaddi, R. (2024). Spectrogram and lstm based infant cry detection method for infant wellness monitoring systems, *Proceedings of the 2024 IEEE International Conference on Human-Systems Integration (HSI)* pp. 1–4.

Nimbarte, N., Ramteke, K., Khan, H. and Wairagade, S. (2021). New born baby cry analysis and classification, *Proceedings of the 2021 International Conference on Signal Processing and Communication (ICSPC)* pp. 1–6.

Pradhan, B. N., Shah, G. R., Ankalaki, S., Shorya, K., Venkatesh, K. and Kushwaha, A. (2022). A machine learning approach to classify biomedical acoustic signals, *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, Bengaluru, India, pp. 1–6.

Reddy, K. S., Rithani, M., Rangarajan, P. K. and Mohan, G. B. (2023). A comparative analysis: Enhancing baby cry detection with hybrid deep learning techniques, *2023 International Conference on Next Generation Electronics (NEleX)*, pp. 1–6.

Riadi, P. A., Faisal, M. R., Kartini, D., Nugroho, R. A., Nugrahadi, D. T. and Magfira, D. B. (2024). A comparative study of machine learning methods for baby cry detection using mfcc features, *Journal of Electronics, Electromedical Engineering, and Medical Informatics* **6**(1): 73–83.

Sharma, A. and Malhotra, D. (2020). Speech recognition based iicc - intelligent infant cry classifier, *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 992–998.

Visvesvaran, C., Nishanth, S., Sudha, R., Kingson Kumar, M., Karthikeyan, J. and Mohammed Rinish, A. (2021). Iot based smart baby monitoring, *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 1–6.

You, W., Xiao, X., Tu, D., Dai, Z. and Huo, Z. (2023). Analysis of multiple types of baby cries based on lstm, *2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pp. 1141–1146.

Younis, S., Sobhy, D. and Tawfik, N. (2024). Evaluating convolutional neural networks and vision transformers for baby cry sound analysis, *Future Internet* **16**(7): 242.

Zhang, K., Ting, H.-N. and Choo, Y.-M. (2023). Baby cry recognition by bcrnet using transfer learning and deep feature fusion, *IEEE Access* **11**: 126251–126262.

Özseven, T. (2022). A review of infant cry recognition and classification based on computer-aided diagnoses, *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, IEEE, pp. 1–11.