ICLAS

Infant Cry Longitudinal Analysis System

Lakshya Gupta, Aditya Tomar, Apurv Bhushan

https://github.com/la-dev05/ICLAS-Infant-Cry-Longitudinal-Analysis-System

Context

- Till date, detection of infant health issues remains mostly reactive.
- Parents and caregivers have little ability to identify any potential health issues before seeking medical attention.
- Research suggests subtle physiological and neurological changes often manifest as small but persistent shifts in an infant cry's acoustic characteristics.
- Usually, changes in cry pitch, intensity, prosody, and spectral characteristics can take place hours or even days before any visible symptoms appear.

Resources:
https://arxiv.org/pdf/2102.02909
https://link.springer.com/article/10.1186/s13636-021-00197-5

Problem Statement

Existing infant cry monitoring systems primarily **treat each crying event as an isolated instance** and **aim to predict a single immediate cause of distress**.

However, infant crying behavior evolves over time and varies significantly across individuals. Population-level cry classification models **fail to capture these longitudinal and personalized behavioral patterns**.

As a result, current systems are unable to model developmental trends, detect individual-level changes, or provide continuous longitudinal monitoring of infant well-being.

To the best of our knowledge, no existing academic or commercial system currently performs personalized longitudinal infant cry monitoring.

Literature Review

Ji et al. (2021) – Infant Cry Analysis & Classification

- **Core Contribution:**
  - A comprehensive review of infant cry research.
  - Surveys five pipeline stages: data acquisition, pre-processing, feature extraction, feature selection, and classification.
  - Reviews both traditional ML approaches **(SVM, KNN, GMM)** and deep learning methods **(CNN, RNN, CNN-RNN, Capsule Network)**.
- **Dataset used:**
  - Made use of Baby Chillanto (2287 samples)- insufficient for deep learning models highlighting a critical data scarcity problem across the field.
  - Dunstan Baby Language (DBL) Dataset.
  - Donate A Cry Corpus

Resources:
Paper Link: https://doi.org/10.1186/s13636-021-00197-5>
Dataset Links:
https://www.dunstanbaby.com/
https://github.com/gveres/donateacry-corpus

•••

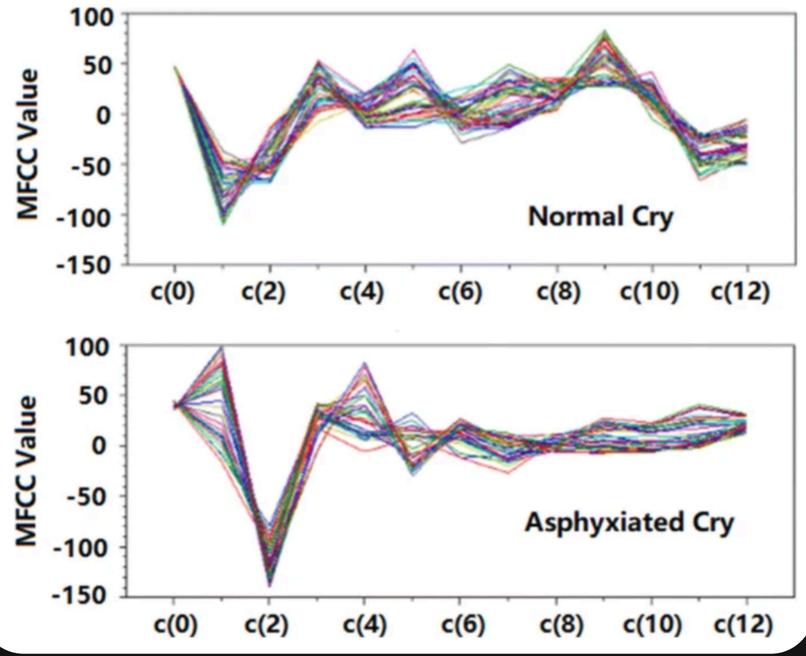
Literature Review

Ji et al. (2021) – Infant Cry Analysis & Classification

- **Key classification results**
  - Asphyxia detection: (Accuracy)
    - PNN/GRNN- **99%**
    - SVM- **97.7%**
    - deep FFNN- **96.74%**
    - CNN-RNN- **94.97%** (on 5-class Dunstan Baby Language database)
    - Autism screening via SVM + MFCC- **96%**

- **Relevance to our work:**
  - Confirms MFCC as the dominant feature across tasks, directly validating our extraction strategy.
  - Highlights dataset size as the main bottleneck; the field's largest private dataset has fewer than 20,000 samples.

Fig. 4



Literature Review

Budaghyan et al. (2023)- CryCeleb2023

- **Core Contribution:**
  - First labeled dataset of infant cries organized by **infant identity:** each cry is tagged to a specific baby, not just a cry type.
  - It covers 786 infants, 26,000 audio files, and 6.5 hours of manually segmented cry expirations.
  - Alongside hosted a public ML competition: where contestants were asked to develop a system capable of determining whether two distinct cry recordings originated from the same infant.
  - Established a baseline using ECAPA-TDNN: a state-of-the-art adult speaker verification model

Resources:
Paper Link: https://arxiv.org/abs/2305.00969

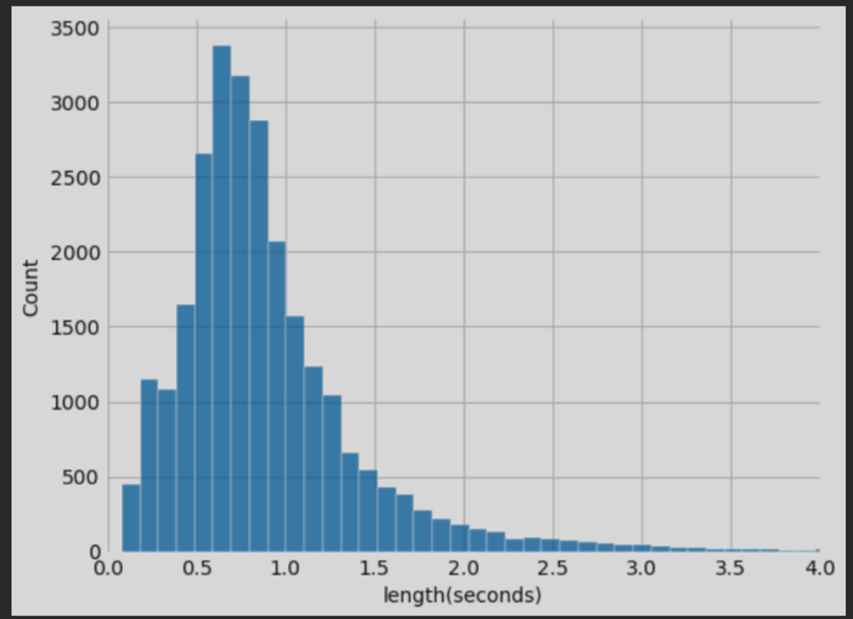


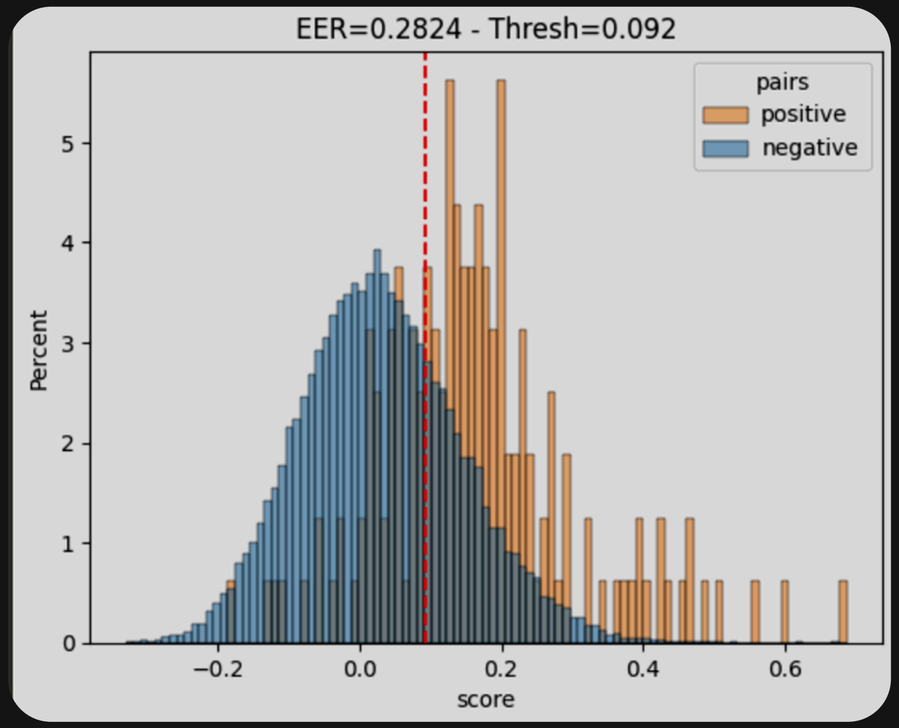
Fig. 1: Histogram of cry sound durations.

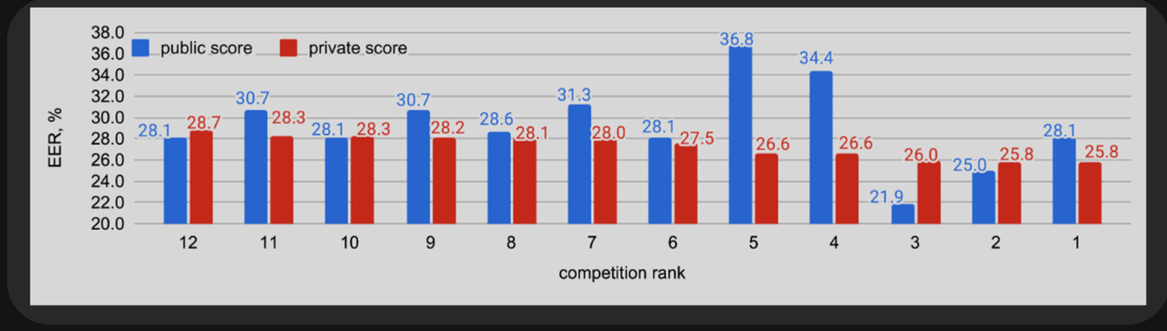
Fig. 3: CryCeleb challenge verification task. Given two recordings, predict if they belong to the same infant

Literature Review

Budaghyan et al. (2023)- CryCeleb2023

- **Key Findings:**
  - Numerical Results:
    - Baseline ECAPA-TDNN (no fine-tuning): **37.92% EER**
    - Fine-tuned on CryCeleb: **28.2% EER**
    - Best competition result: **25.8% EER** (59 participants, 11 teams beat the baseline)
  - This showed that adult speaker verification architectures can transfer to infant cry, but a significant gap remains.
- Relevance to our work:
  - CryCeleb2023 is the largest available labeled infant cry dataset





Literature Review

Liu et al. (2024)- Infant Cry and Snoring Detection

- **Core Contribution:**
  - Introduced a publicly available benchmark for infant cry and snoring detection.
  - Three-subset structure for diverse training conditions:
    - A real strongly labeled subset with event-based labels annotated manually
    - A weakly labeled subset with only clip-level event annotations
    - A synthetic subset generated and labeled with strong annotations.
  - The dataset combines multiple source recordings across diverse environments, consolidated into 10-second clips.

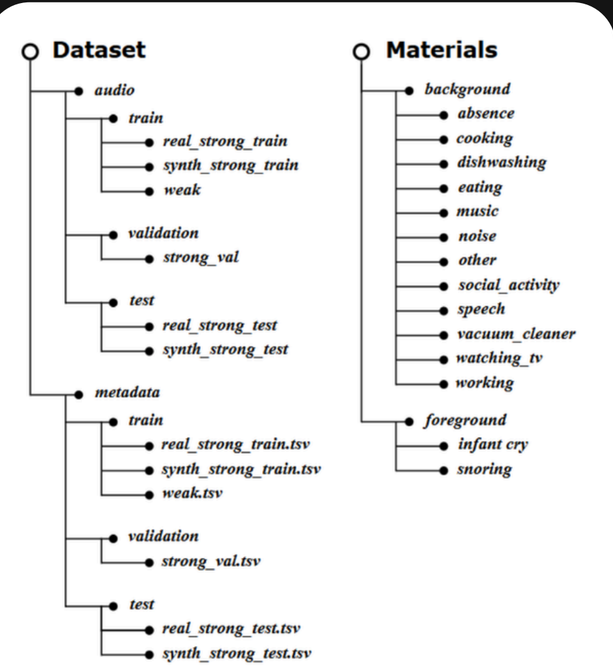


Table 2: Overall Statistics of ICSD dataset (#clips, 10s/clip).

Set	Train		Validation		Test	
	InfantCry	Snoring	InfantCry	Snoring	InfantCry	Snoring
Weakly labeled	1699	1577	189	176	None	None
Real strongly labeled	338	305	43	39	43	39
Synthetic strongly labeled	4000	4000	500	500	500	500

Resources:
Paper Link: https://arxiv.org/pdf/2408.10561

Dataset

PRIMARY DATASET

CryCeleb 2023
26093 samples, 6.5 hours of audio

ICSD Dataset
14000 samples

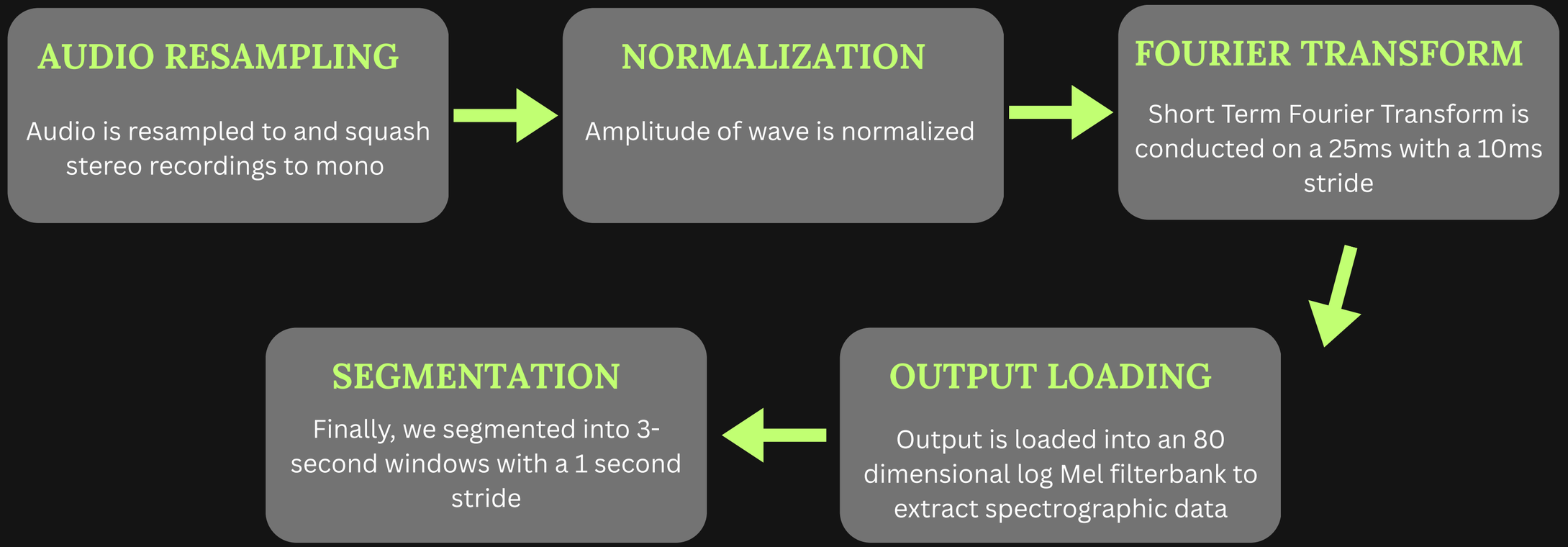
Baby Chillanto Database

Kaggle

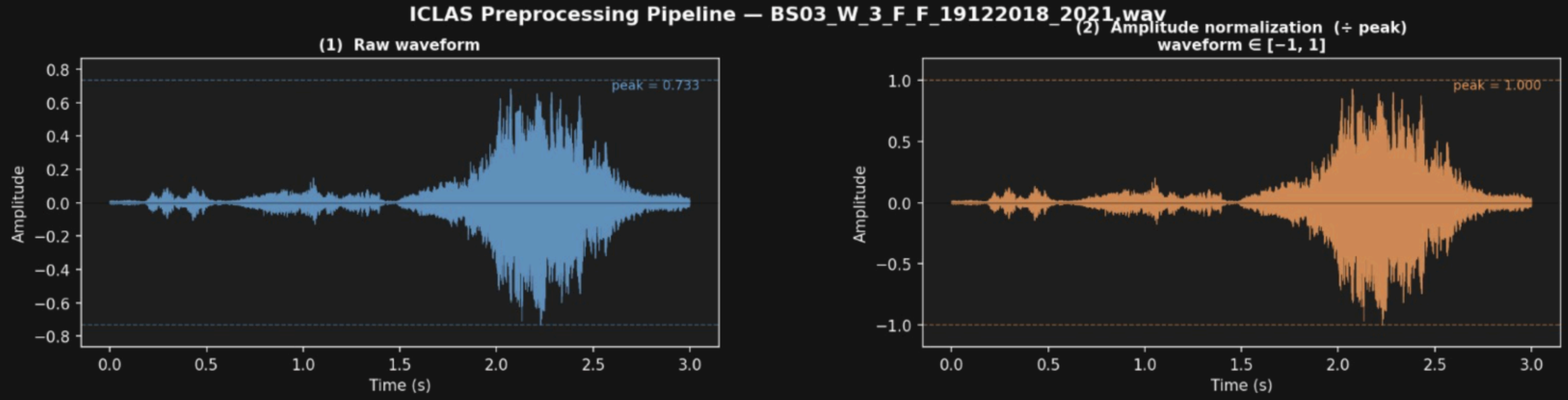
Donate-a-Cry Corpus

OSF

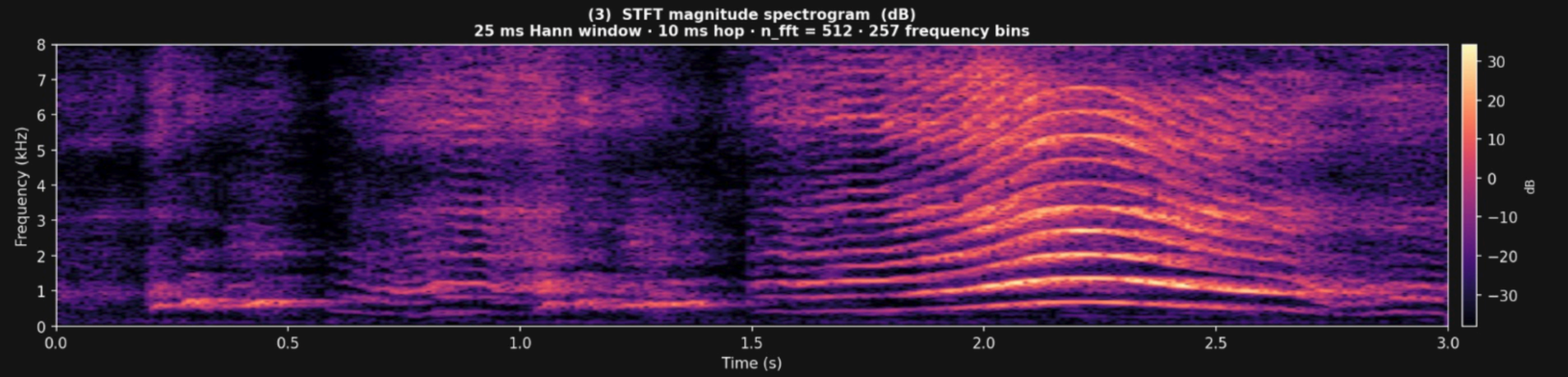
Features Preprocessing

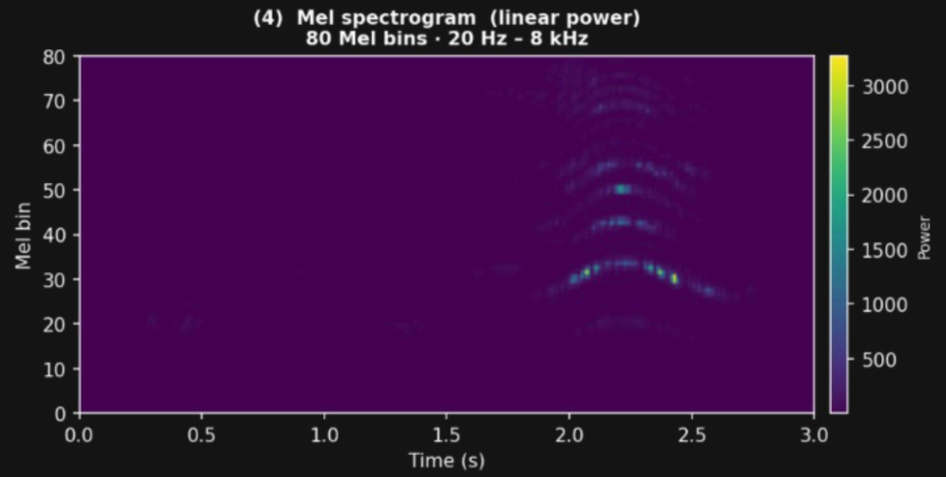


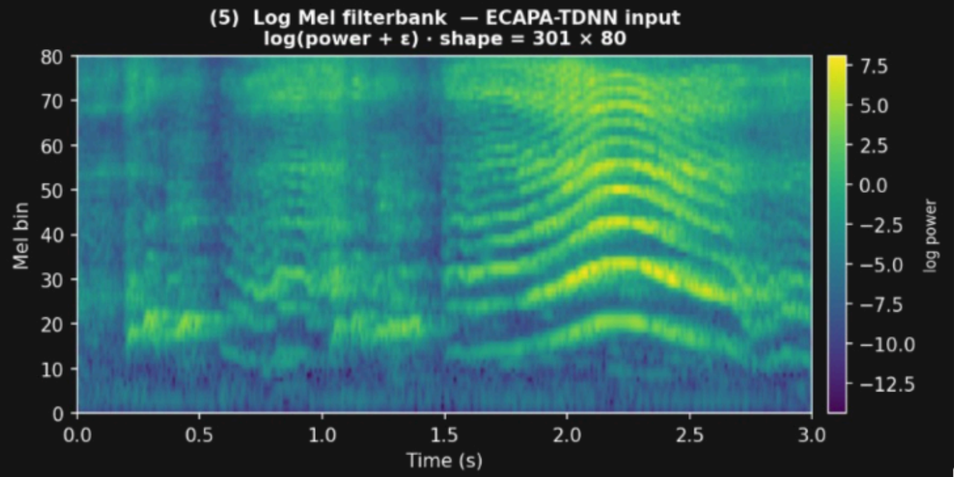
Features Preprocessing



Features Preprocessing







ML Methodology

OUR GOAL

To monitor changes in an individual infant's cry over time and flag cries that deviate significantly from that specific infant's own personalized baseline acoustic pattern.

5 STAGES

- Stage 0: Input and Preprocessing
- Stage 1: Baseline Feature extractor
- Stage 2: Personalized Anomaly Detection
- Stage 3 - Gated Continual Updating
- Stage 4 - Longitudinal Deviation Tracking

ML Methodology

Stage 1: Baseline Feature Extractor

ECAPA-TDNN BACKBONE

Initialized from the pretrained *VoxCeleb* Checkpoint

Trained on VoxCeleb1 and VoxCeleb2 datasets

Encompassing 7000 adult speakers.

FINE-TUNING

- CryCeleb 2023
- Baby Chillanto
- Donate-a-cry Corpus

ML Methodology

Stage 1: Baseline Feature Extractor

2 Training Losses were Considered

- Contrastive NT-Xent loss following the SimCLR framework
- Generalised End-to-End loss.

Partial fine-tuning strategy

- Lower SE-Res2Block layers were kept frozen
- The upper blocks, the pooling layer, and the fully connected head were updated during training.

OPTIMIZER

We made use of AdamW

Learning Rate: 1e-5

Weight Decay: 1e-4

FINAL OUTPUT

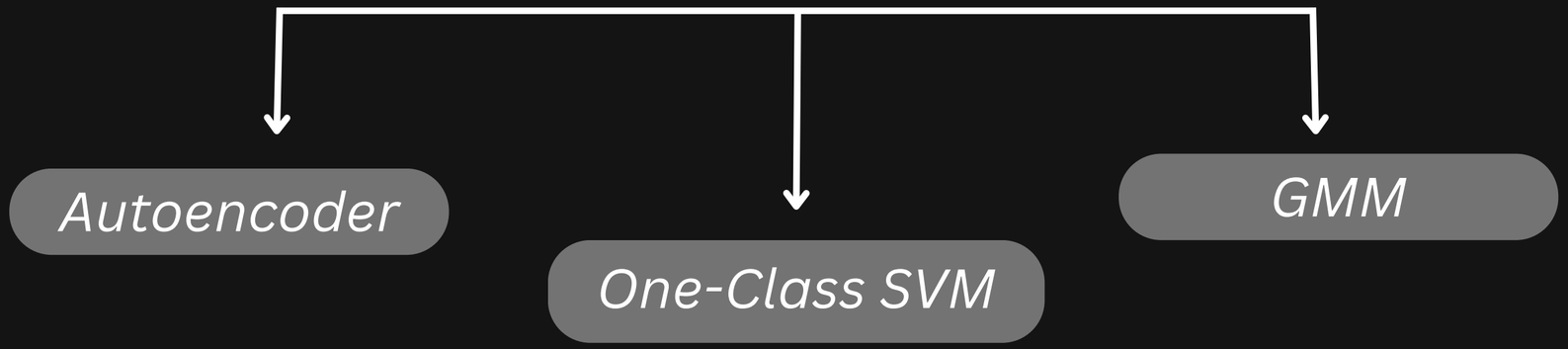
192-dimensional L2-normalised embedding vector representing the acoustic identity of the input cry segment.

...

ML Methodology

Stage 2: Personalized Anomaly Detection

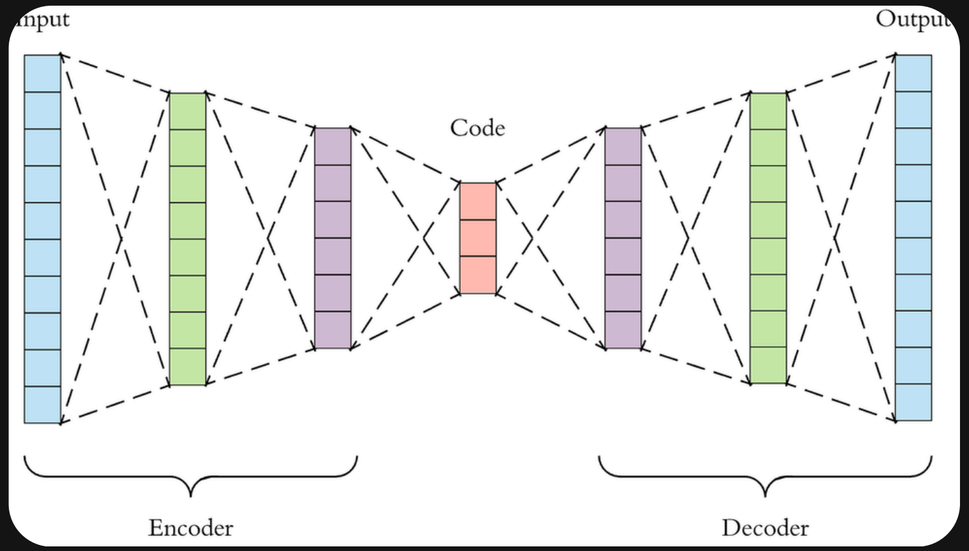
We evaluated three architecturally distinct options and compared them



ML Methodology

Stage 2: Personalized Anomaly Detection: Autoencoder

- The architecture follows a symmetric encoder-decoder structure.
- The encoder compresses the input embedding through three fully connected layers, reducing dimensionality from 192 to 128 to 64, and finally to a 32-dimensional bottleneck latent representation, with ReLU activations applied at each stage.
- The decoder then mirrors this process in reverse, expanding from 32 back through 64 and 128 before reconstructing a 192-dimensional output embedding.



ML Methodology

Stage 2: Personalized Anomaly Detection: Autoencoder

ANOMALY SCORE

Defined as the Mean Squared Error between the original input embedding and its reconstruction.

INTUITION

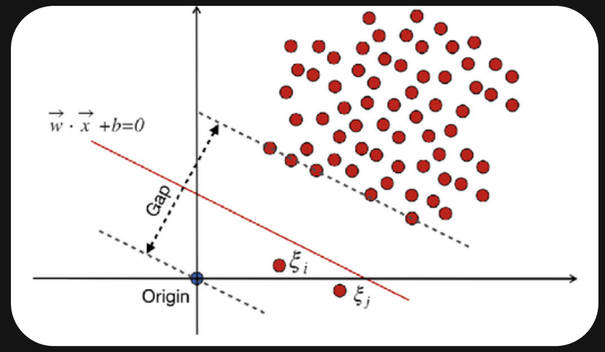
A model trained on a specific infant's baseline cry embeddings will learn to reconstruct those embeddings accurately

Producing a high reconstruction error when presented with an embedding that deviates meaningfully from that learned distribution.

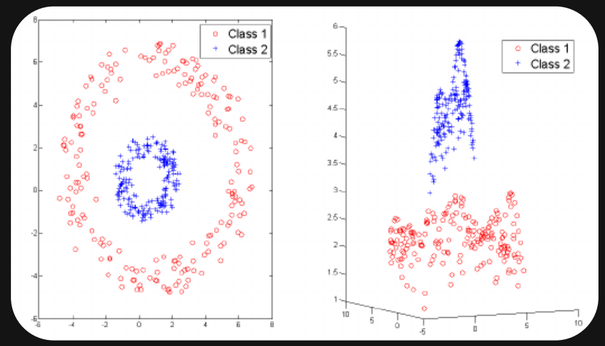
ML Methodology

Stage 2: Personalized Anomaly Detection: One-Class SVM

- The SVM learns a decision boundary, specifically a hypersphere in the embedding space, that encloses the normal distribution of a given infant's baseline cry embeddings.
- Any new embedding that falls outside this boundary is classified as anomalous.
- The model uses a Radial Basis Function kernel, which is well suited to handling the non-linear boundaries that arise in a 192-dimensional space.



Support Vector Machine



RBF Kernel

ML Methodology

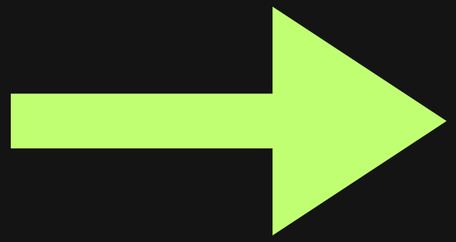
Stage 2: Personalized Anomaly Detection: One-Class SVM

PARAMETERS

ν = [0.05, 0.1]
Controls the fraction of training points treated as outliers during fitting

The gamma parameter is set to scale, which automatically adjusts to:

$\gamma = 1/(N \times V)$
Where, N = number of features
V = Variance of Training Data

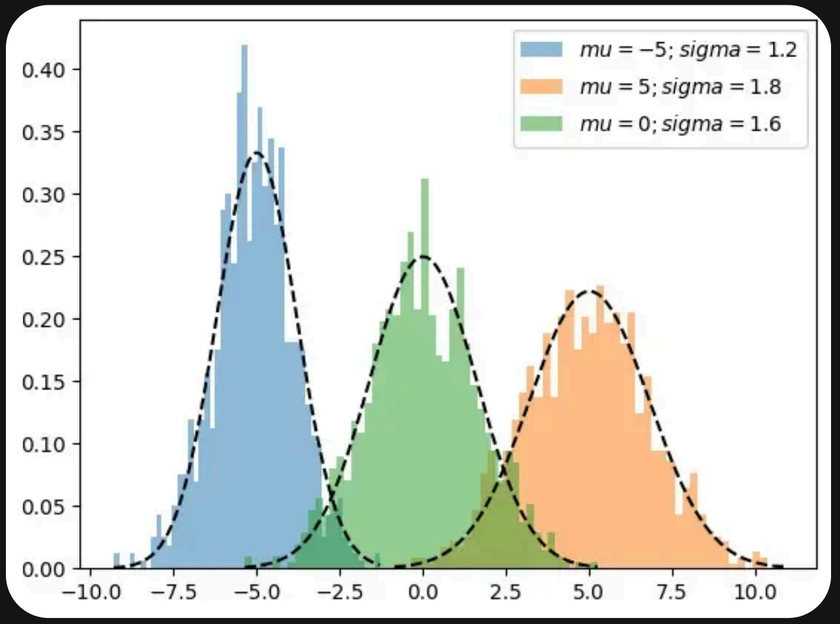


This ensures the kernel is calibrated to the actual spread of the embedding distribution.

ML Methodology

Stage 2: Personalized Anomaly Detection: Gaussian Mixture Model

- A unimodal model such as the OC-SVM, which fits a single decision boundary, may flag the intense cry as anomalous simply because it falls far from the cluster centre of the calmer cries.
- The GMM addresses this directly by explicitly modelling a multimodal normal distribution.
- It fits between two and five mixture components per infant, with the number of components K selected by Bayesian Information Criterion on held-out baseline clips rather than fixed in advance.



ML Methodology

Stage 2: Personalized Anomaly Detection: Gaussian Mixture Model

The fitting algorithm is Expectation-Maximisation, the standard approach for GMM fitting, and the model operates on the full 192-dimensional embedding space.

ANOMALY SCORE

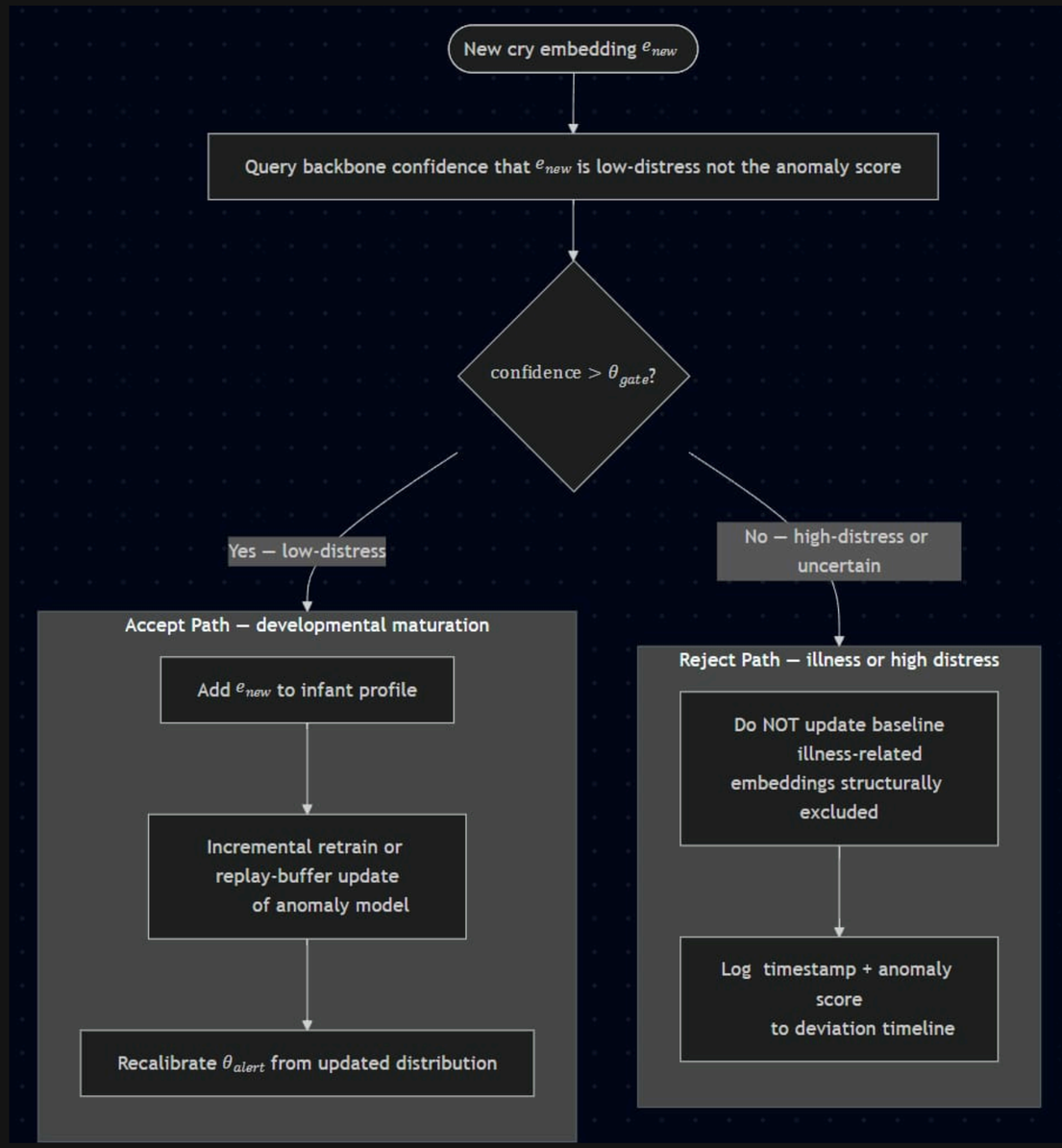
$$S(e) = -\log p(e \mid GMM)$$

A high negative log-likelihood indicates that the embedding falls outside all learned normal cry modes and is therefore anomalous

ML Methodology

Stage 3: Gated Continual Updating

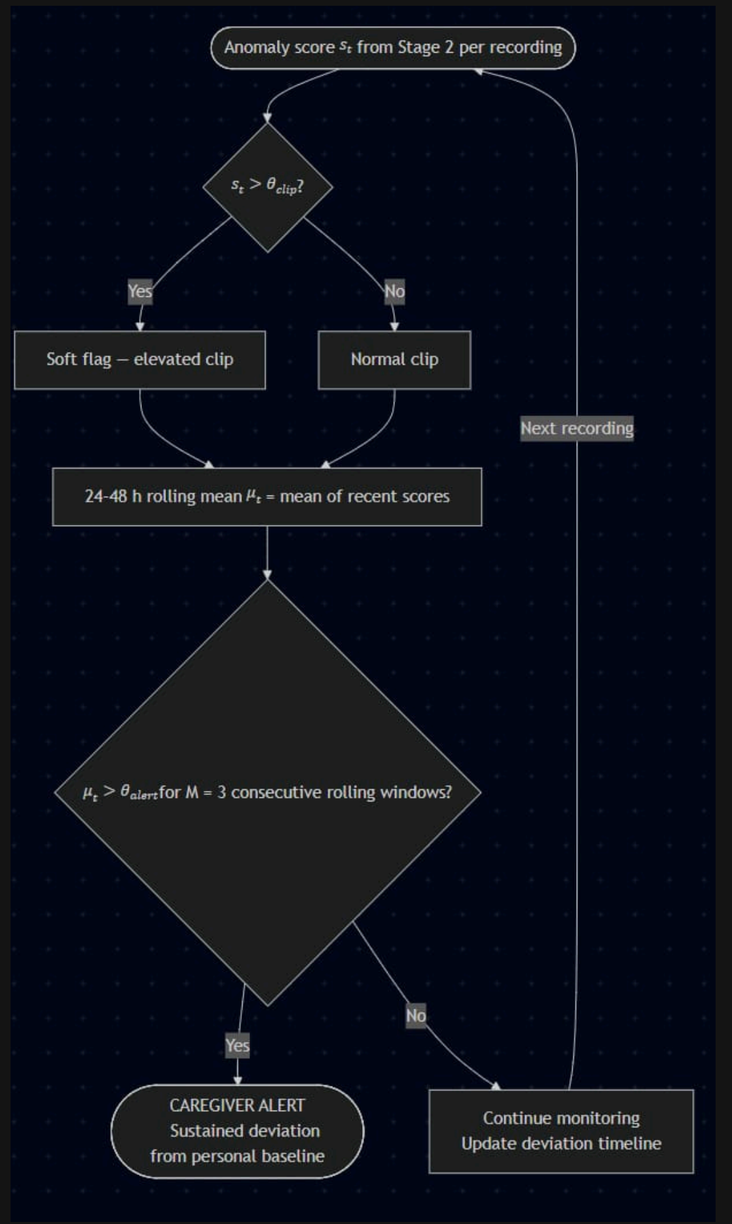
- The gate works by querying the backbone's confidence that a new incoming embedding is low-distress and developmentally normal before deciding whether to accept it into the baseline.
- If the confidence score exceeds the gate threshold, the anomaly model is updated through incremental retraining or a replay buffer
- If the confidence score falls below the gate threshold, the embedding is rejected from the baseline update entirely.
- The anomaly model remains unchanged, and instead the event is logged as a timestamped anomaly score entry on the infant's deviation timeline.



ML Methodology

Stage 4: Longitudinal Deviation Tracking

- The output is not a single classification decision but a per-infant time-series of anomaly scores aggregated across recording sessions.
- Each new cry clip is passed through the ECAPA backbone to produce an embedding.
- It is then scored by the infant's personal anomaly model to produce a scalar anomaly score.
- These scores accumulate over time, building a longitudinal deviation timeline.



Performance Metrics

ECAPPA-TDNN BACKBONE

Our backbone outperformed the CryCeleb model by 4% on the infant EER speaker identification metric (21.38% vs 25.04%)

GATE + ANOMALY TIMELINE

For Sanity Check

GATE TRAINING

Based on AUC for identifying whether its a cry or not

TRAIN PER-INFANT AUTOENCODER

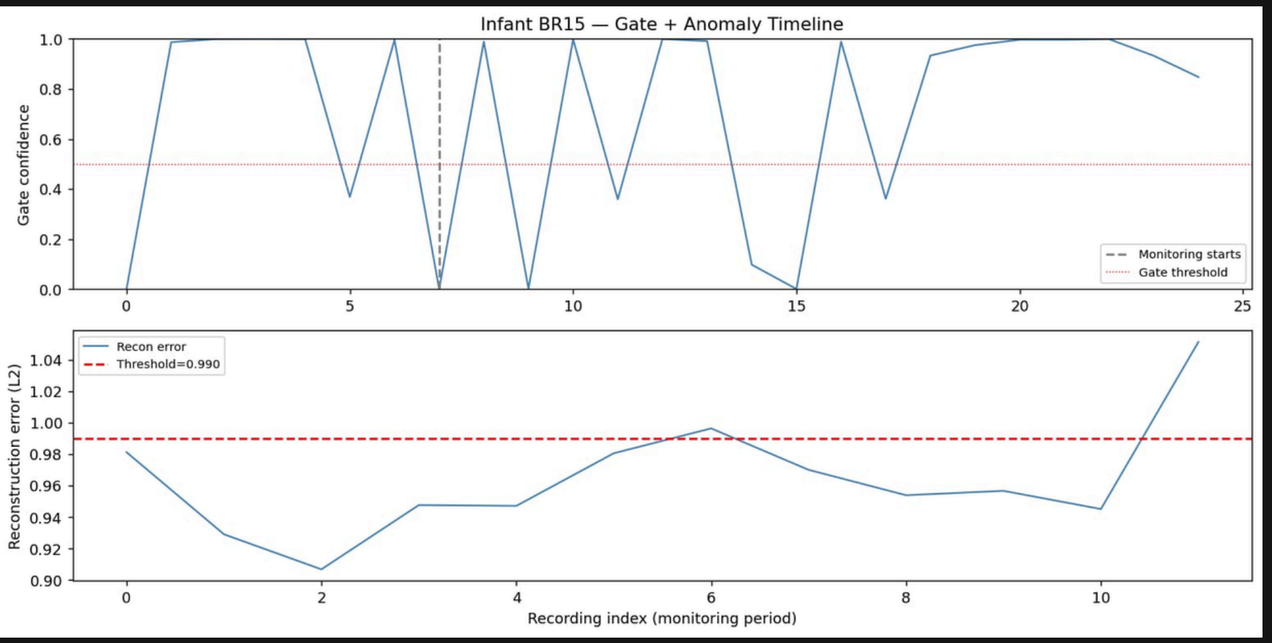
- AUC per infant and mean = 0.9055 +- 0.0495
- Average Precision (averaged over recording sessions) per infant and mean = 0.9550 +- 0.0277
- F1 score per infant and mean = 0.8719 +- 0.0887
- Recall per infant and mean = 0.903

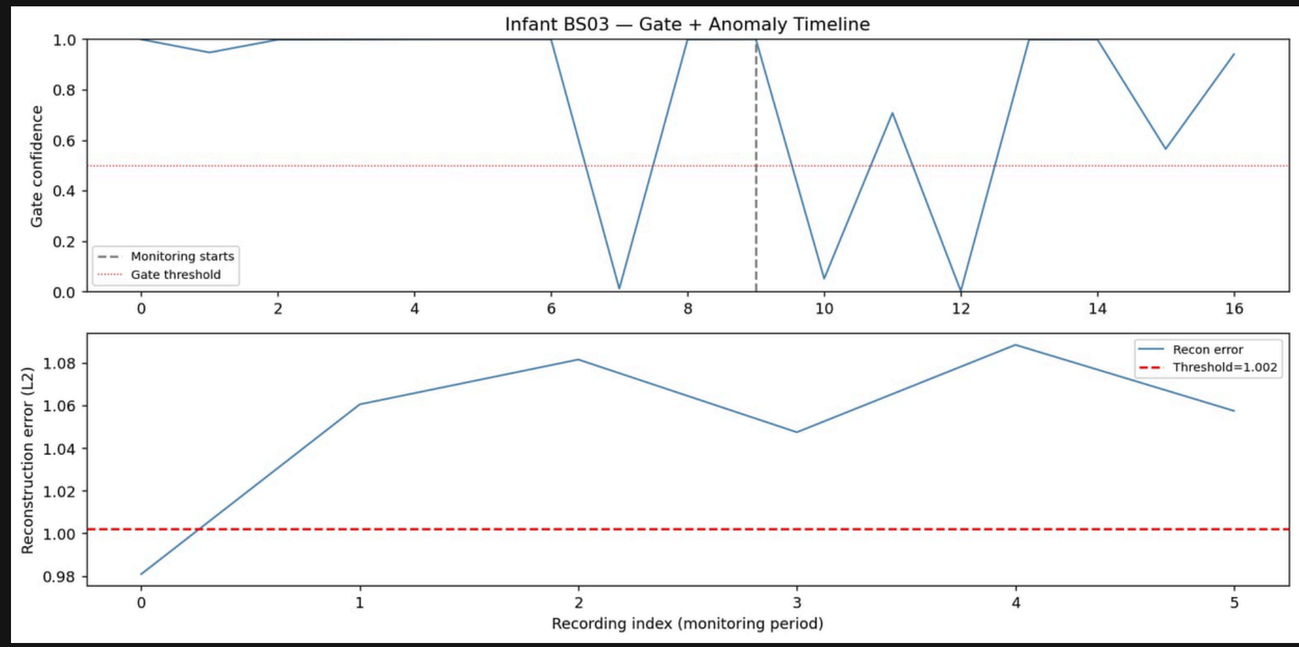
Performance Metrics

Infant	N_recs	AUC	AP	F1	Recall	FPR
BR07	15	0.923	0.964	0.937	0.9	0.067
BR15	18	0.908	0.96	0.818	0.72	0.111
BS03	8	0.885	0.976	0.903	0.84	0.125
CA12	35	0.902	0.927	0.797	0.98	0.686
GL13	22	0.894	0.942	0.847	1	0.818
KA02	15	0.925	0.978	0.935	1	0.467
LC10	16	0.919	0.944	0.952	1	0.312
LC26	29	0.956	0.971	0.916	0.98	0.276
MR30	36	0.966	0.976	0.949	0.94	0.056
PA08	11	0.771	0.926	0.88	0.88	0.545
PA27	9	0.969	0.995	0.943	1	0.667
PJ14	11	0.891	0.973	0.571	0.4	0
RB01	38	0.899	0.925	0.842	0.96	0.421
SA20	28	0.94	0.963	0.884	0.84	0.107
SB23	9	0.931	0.987	0.931	0.94	0.444
TA05	23	0.909	0.948	0.84	1	0.826
TM04	27	0.807	0.881	0.875	0.98	0.481

Performance Metrics

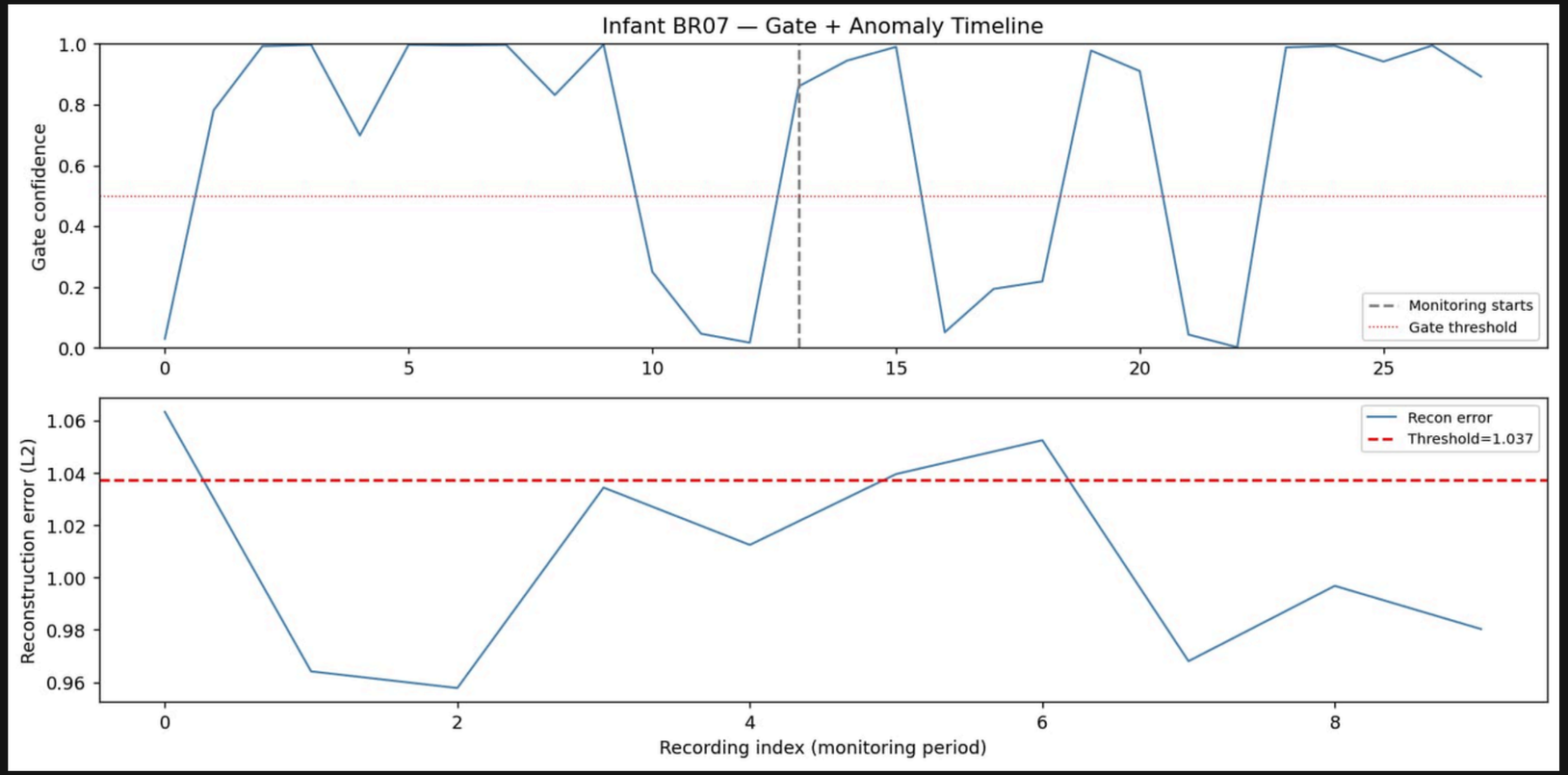
Gate Training Results





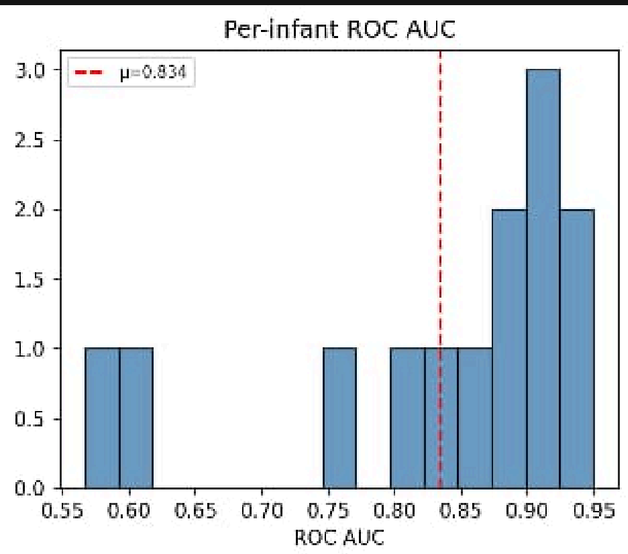
Performance Metrics

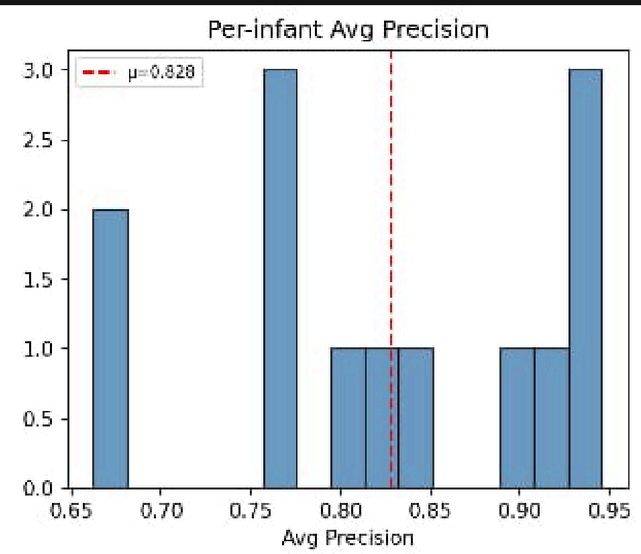
Gate Training Results

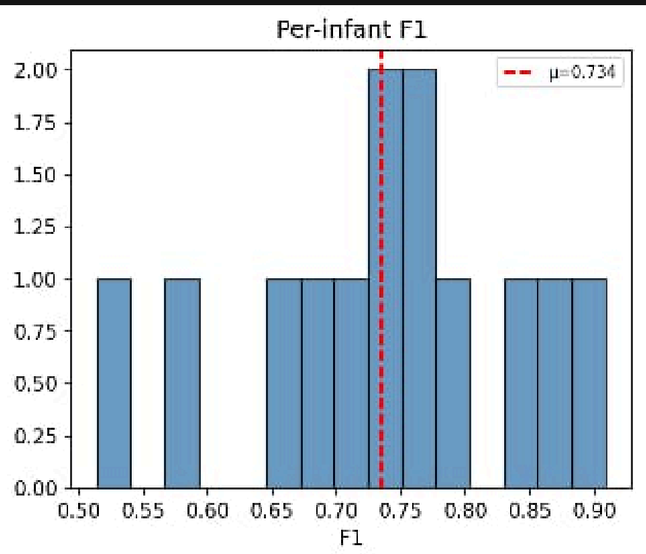


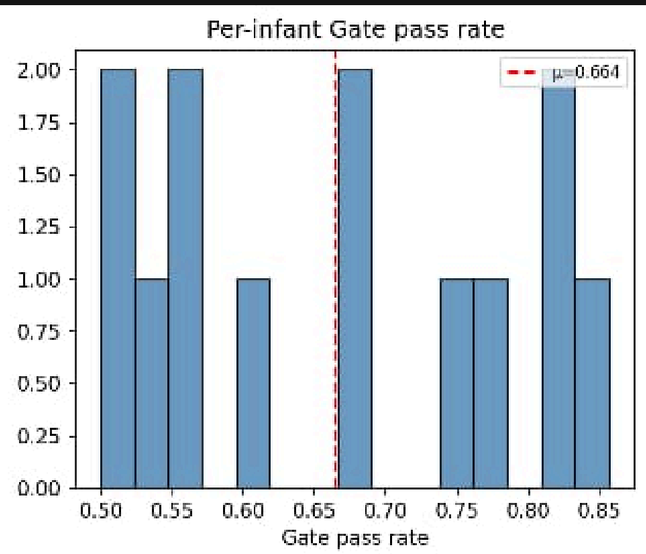
Performance Metrics

Train per-infant Autoencoder Results





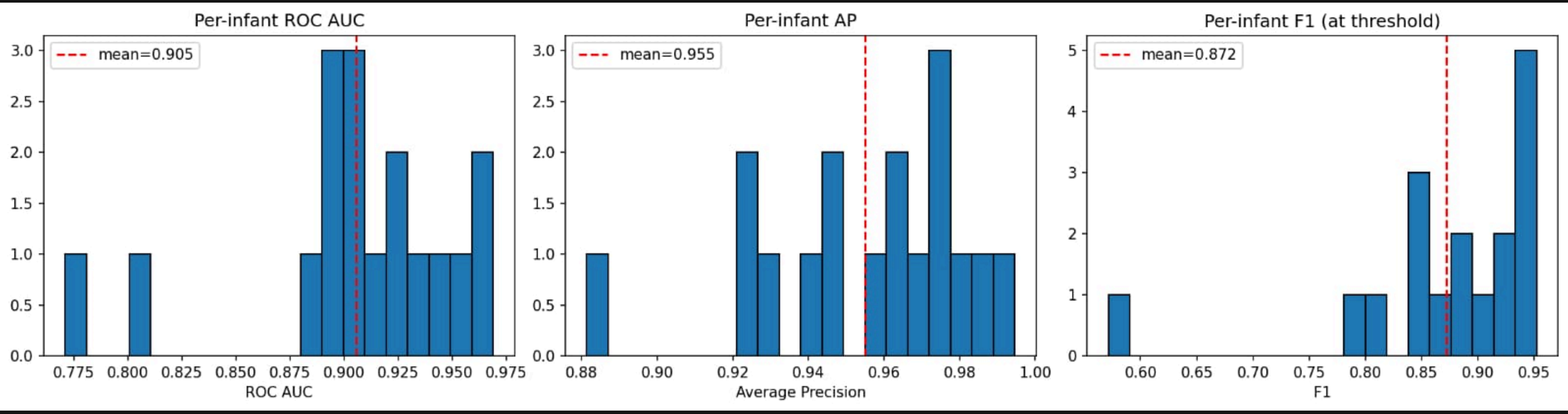




Continuous Time

Performance Metrics

Train per-infant Autoencoder Results



Discrete Time

Challenges and Future Work

- Due to the difficulty in accessing infant cry recordings, we had to limit the time the model has to learn an infant’s baseline.
- Our model currently requires complex computational hardware to run. A workaround to this must be found, or a leaner model should be developed.
- Variance in recording hardware in the real world may disrupt our models ability to make accurate predictions.
- Conduct a more extensive validation study of our model on data from real-life baby monitors in homes, hospitals.
- The false positive rate shows a high variance and is unusually high for some infants (TA05, CA12). While not an immediate concern, lowering this rate is essential for practical deployment of the model.

Thank You!