DS&AI 2023
DATA SCIENCE AND ARTIFICIAL INTELLIGENCE CONFERENCE 2023
1ST - 3RD FEBRUARY 2023

Autonomous Surveillance of Infants' Needs Using CNN Deep Learning Model For Audio Cry Classification: Artificial Parenting

John Mitch Okwiri
Geofrey Owino

Sponsored by Artificial Intelligence JOURNAL
Google TensorFlow

Background

- Autonomous infant monitoring is crucial for ensuring infants well-being and development
- Traditional methods of monitoring can be time-consuming and not always accurate
- Research in using audio signals to detect infants needs
- Audio cries are a crucial means of communication and can indicate various needs
- CNNs have been widely used in image and audio classification tasks and have shown to be highly effective
- This project aims to develop an autonomous surveillance application using CNN deep learning models for audio cry classification to enhance infants care and well-being.

Problem

Did you know?

- 12% increase infant annual mortality rate
- 31% increase of baby misdiagnosis
- Increased cost on childcare services
- 2% increase of mother-baby hospital visitation
- Successive escalating health problems

Source: UNICEF & WHO

Study Objectives

General objective:

- To determine the infants' needs through audio cries using CNN classifier model

The specific objectives are:

- To transform infant cry to spectrum images
- To develop CNN model for spectrum image classification
- To train the CNN model and validate its classification performance.
- To deploy the model as an android App

Background Literature

Audio Analytics review

- An automatic classification of bird species using audio feature extraction *(Garima et al., 2016).*
- Binary classification of two infant cries classes using ANN model *(Singh et al., 2013).*

Model based review

- Music genre classification using Multilayer perceptron network *(Costa et al., 2015).*

Modelling of infant audio cries

- Binary audio pattern recognition of baby crying sound events *(Ntalampiras et al., 2015)*

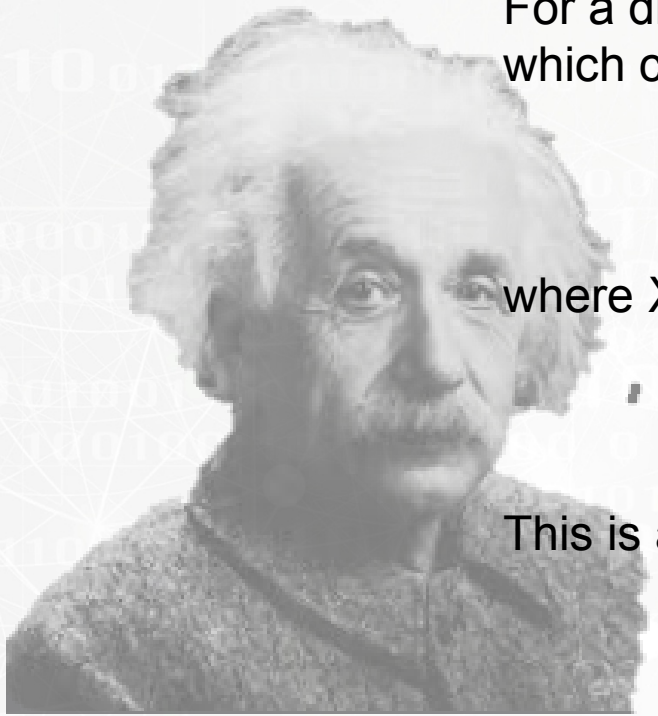
Methodology

❑ Data Preparation

- ✓ Trimming of the audio files to standard length.
- ✓ Transforming all audio files to .wave format.
- ✓ Setting up a common audio sample rate.
- ✓ Loading audio files and extracting features.
- ✓ Class Balance using GAN algorithm.
- ✓ Deliberate mixture of acoustics background with the
- ✓ audio classes.
- ✓ Storing the audio data in no-structured database server, postgres server.

Methodology

Fourier Transform



For a discrete sequence of audio data say; {xn} = x0, x1, ..., xn−1 which can be transformed in frequency domain signal as:

$$X_k = \sum_{n=0}^{N-1} x_n e^{\frac{-2\pi_i}{N} k_n}$$

where X_K is k^{th} transformed signal and k_n is the output index.

$$X_k = \sum_{n=0}^{N-1} x_n [\cos(\frac{-2\pi_i}{N} k_n) - i \sin(\frac{-2\pi_i}{N} K_n)$$

This is an assumption that the sound is linear scale.

Methodology

Mel Spectrum

A better transform is the Mel spectrum which converts frequency (f) in Hertz to Mels as:

$$m = 2595 log_{10}(1 + \frac{f}{700})$$

This is then fed as input image for computer vision as shown below

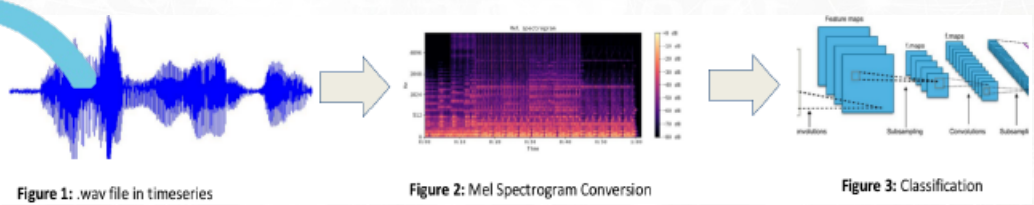


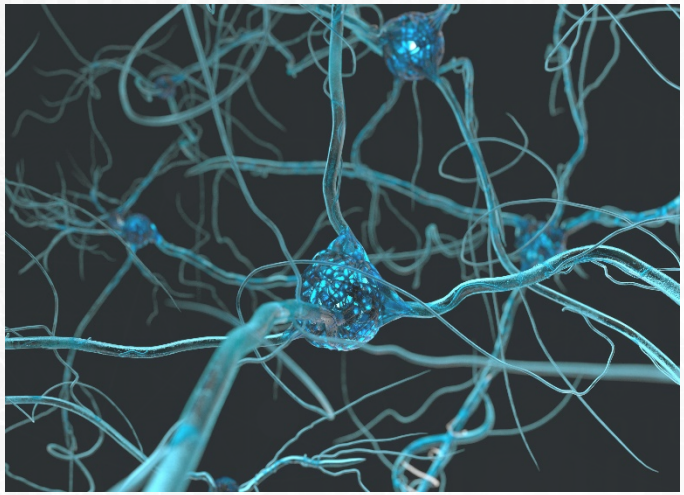
Figure 1: .wav file in timeseries

Figure 2: Mel Spectrogram Conversion

Figure 3: Classification

Methodology

Convolution Neural Network



CNN has been widely used in computer vision, audio
analytics and problems that perhaps involves very large
complex data sets.
CNN has the following layers;

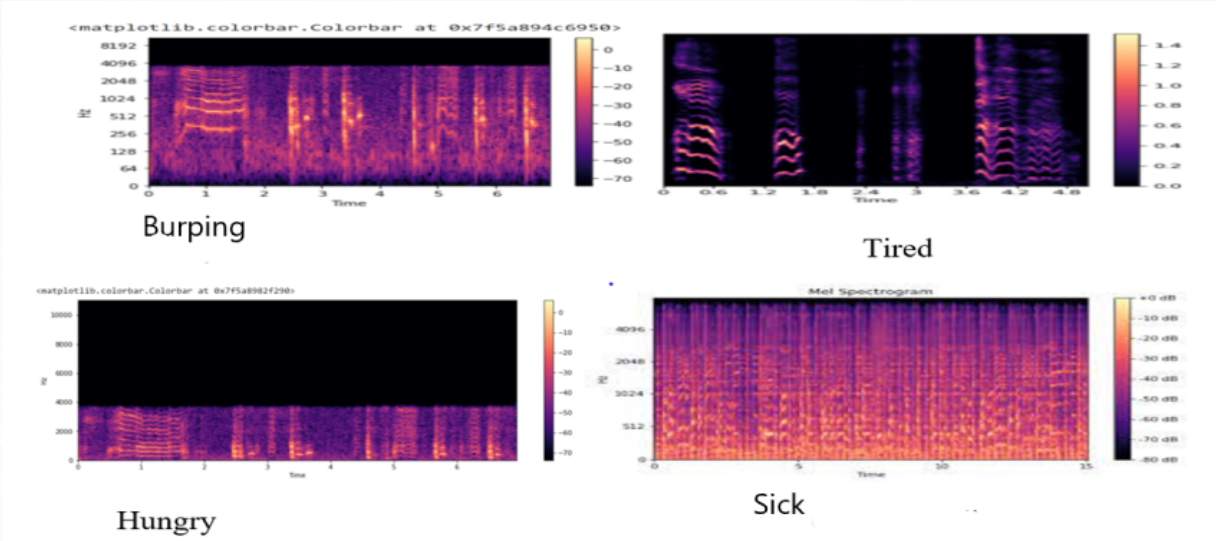
- Convolution layer
- Non-linearity layer
- Feature pooling or sub-sampling layer
- Fully connected layer
- Loss layer and softmax function

Results

Data Preparation

- Mel Spectrum Output

Every class was shown to have different Mel spectrum images.

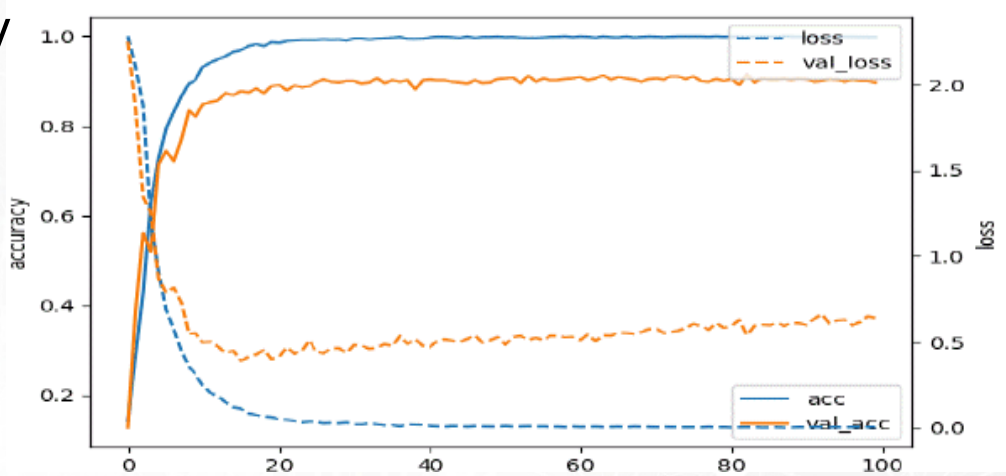


Results

Training and validation accuracy

The training and validation accuracy was increasing with epochs.

The training and validation loss was decreasing with epochs.

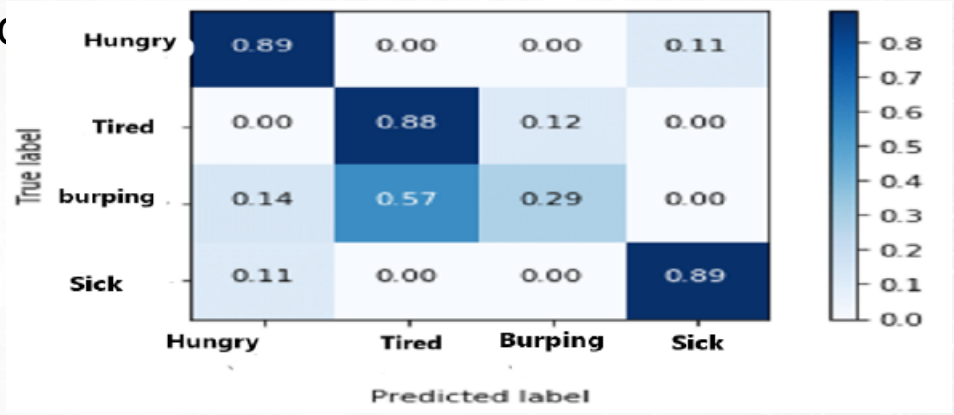


Results

Model Evaluation

- The model performed well as 0.89 of both hungry and
- discomfort class were classified to be true.

- The model classified few burping recor
- due to few data set.



Results

Model Comparison

Model	Training Accuracy	Testing Accuracy	Epochs	Computational time(minutes)
Decision Tree	0.70652	0.5671	100	73.02
Cat Boost Classifier	0.83696	0.6834	100	72.23
Cross Gradient Booster	0.83997	0.6439	100	74.31
Cross Gradient Booster	0.82609	0.6123	100	73.00
Support Vector Machine	0.80102	0.6906	100	72.21
Naive Bayes	0.73029	0.6308	100	76.04
Stochastic Gradient Descent	0.79348	0.7192	100	75.01
Mobile Resnet	0.9822	0.8340	100	66.87
CNN	0.94739	0.9175	100	32.49
RNN	0.82609	0.8717	100	79.36
Multi-layer perceptron	0.5783	0.6756	100	78.32

Discussion

- Higher accuracy – The convolutional neural networks did relatively well in generalizing the testing data thus having a higher accuracy. This promised the stability of the model thus affirming the operational stability of the model on the production environment.

- Reduced computational time – The convolution neural networks really possess the strength of down sampling with translational invariance. The dimension reduction has been scientifically embraced to reduce the computational time for the model. This brings a smooth significance in efficient operations of the model on the production environment.

- It is evident that ResNet transfer model performed pretty well but relatively failed on testing data as compared with the CNN model.

Conclusions

- This study confirms that the convolution neural networks indeed fitted the infant audio dataset well as compared with other models.

- The study also motivates the stage wise evaluation of the deep learning model during the training stage and also upon production. On the realization of the efficient model that was least affected by the covariance shift, light and stable, the study came to conclude that CNN model was very stable on deployment and faster on inferential stage.

- The inherent down sampling through convolving the mage pixels to arrive at receptor fields promoted the dimension reduction that facilitated the light weight of CNN model.

Future Work

5. **Monitor**
4. **Supply of the product to the bigger market (3 months)**
3. **Integration of the gadget with USSD message (6months)**
2. **Roll out of the product in the market (4months)**
1. **Piloting**

Future Work

- The study proposed use of ensemble deep learning models to improve on the stability of the models and classification power of the CNN model.
- The study recommends use of reinforcement learning models that could be customized for every infant so that the model could be evaluated based on the development of the baby.
- The study also proposes the deployment of the model into an IOT gadget that could accommodate some complex but more accurate like transformer models

THANK YOU!