Automated detection of neonatal asphyxia

Gladys INABEZA
Carnegie Mellon University
ginabeza@andrew.cmu.edu

Jacob RAST
Carnegie Mellon University
jrast@andrew.cmu.edu

Peter SAUER
Carnegie Mellon University
psauer@andrew.cmu.edu

Pamely ZANTOU
Carnegie Mellon University
pzantou@andrew.cmu.edu

Abstract

Birth asphyxia is a neonatal condition that is not only hard to identify, but is also the leading cause of neonatal deaths. Identifying neonates, or babies, with birth asphyxia would thus save numerous lives. Current models able to be put into practice use convolutional neural networks to identify neonates with asphyxia using Mel Frequency Cep Coefficients (MFCCs). In this project, we outline several methods that achieved competent results that performed the same as our implementation of a baseline. In particular, we show that the performances of Linear Networks, Bi-directional LSTMs, and ResNet18 have similar Unweighted Average Recall (UAR) scores. However, upon bootstrapping and applying Tanh() distortion and Room Reverberation, UAR increases from the baseline UAR of 0.6 to 0.7. We thus propose that researchers in the field of Pathological Cry Analysis should focus less on model sophistication and more on bootstrapping techniques to increase data samples or generative models and tasks that can rely on fewer data samples to make high-accuracy results.

1 Introduction

Birth asphyxia is one of the most prevalent neonatal diseases in the world. It is among the most common mechanisms of death in neonates [17]— causing 23% of all neonatal deaths [15] — and a leading cause of death in many areas of the world. While birth asphyxia is preventable with proper identification, diagnosis requires expensive technology or highly trained medical professionals [7] both of which are lacking in remote or rural locations. As a result, low-income areas and developing countries have a disproportionate amount of infants suffering from neonatal asphyxia-induced disability and death [citation]. In this work, we propose methods to extend the state-of-the-art in automated cry-based diagnosis of neonatal asphyxia.

Previous studies have shown that neonatal asphyxia can be assessed using crying sound waves coming from infants, as speech and breathing are controlled both by the lungs [21]. In machine learning research, many cutting edge techniques have been developed to perform infant cry classification. These methods include using various forms of Deep Neural Networks (DNNs)[13], the use of SVM or transfer learning from the African startup Ubenwa. A baseline sensitivity of 85% and specificity of 89% was reported in 2018 [21]. In this work, we propose the use of transformer architectures to advance the state of the art in infant cry detection. Cry detection performance will be compared against the transfer learning model proposed by Ubenwa.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

2 Literature Review

2.1 Classification of Neonatal Asphyxia Using Various Machine Learning Models

There have been several machine learning classifiers and features used as a way to analyze biomakers. For infant cry classification most successful techniques include SVMs, DNNs with transfer learning and DNNs without transfer learning. Recently DNNs approaches, especially CNNs based ones have achieved better performances ranging from 85% to 98% accuracy. For example, [21] used a neural transfer learning method based on Residual Neural Network (ResNet) to achieve a sensitivity of 85%. [12] combined a ResNet 50 with a semi-supervised Graph Convolution Network (GCN) and achieved 94.39% accuracy on the Baby Chillanto dataset. [12] argues that GCNs architecture solved the challenge of data limitation and label uncertainty. CNN-RNN and Probabilistic Neural Networks [9] achieved respectively 94.97% and 98% accuracy for infant cry classification.

Before performing any classification tasks, it is essential to process cry signals and extract and select useful features. Automated infant cry research involves signal processing, feature extraction and feature selection. The method used for any of these operations can lead to better performance or not.

Cry processing consists in removing noise and audio segmentation. This stage is critical to accurately classify infant cry signals. High-pass FIR filter, filter-and-sum beamformer, and OMLSA post-filter are used to reduce noise from audio data. On the other hand, Voice Activity Detection technique is used for audio data segmentation [13]. After cleaning the audio data the next step is feature extraction.

Cry is a mixture of acoustic and prosodic information at different levels [13]. Feature extraction consists in retrieving discriminative features from audio signals which will be later fed to a machine learning model. [2] show a feature representation of the cry signal influence more the performance of the model than the choice of the classifier. Infant cry feature extraction is performed both in time domain and frequency domain. However, frequency domain features has been proven to be more robust in the litterature. MFCCs, LPCCs, and LFCCs are the most common features used for infant cry analysis. Other Cepstral coefficient features have been used (Weighted Linear Prediction Cepstral Coefficients, Linear Predictive Coding and have proven to be better than time-domain features. On other occasions, Wavelet packet transformations, log-Mels and Short-time Fourier Transforms (STFTs) are used instead of MFCCs. [24] states that MFCCs and log-Mels might lead to a poor representation of infant cry signals due to its frequency range, 1500 Hz to 3000 Hz. In their work they suggested a differential time-frequency log-Mel spectrogram features representations for infant cry signals which outperform simple MFCCs and log-Mels. Other feature augmentation methods incorporate the use of prosodic features, which provides metadata for the acoustic features extracted from MFCCs and image features which provide acoustic and prosodic information and a strong representation of the signal in time domain.

2.2 Transformers based infant cry classification

There are a few research exposing the use of transformers to classify infant cry. [24] used Vision Transformer (ViT) transformer fed with differential time-frequency log-Mel spectrogram features extracted from the CRIED database. They proved that combining time and frequency domain feature representation of infant cry signals leads to better classification. In fact, they combined log-Mel features, the deltas of the log-Mel spectrogram over time and a 3D feature representation of the infant cry signals. They the Unweighted Average Recall (UAR) as metric to measure the performance of their model. The model outperforms most of the existing models using the CRIED dataset with an UAR of 81.2%. Another notable effort was the use of CNN coupled with a self-attention (CNSA) with multiple feature groups by [16] to classify infant and mother vocalizations.

2.3 Classification of Other Diseases Involving Analysis of Speech or Cries

In addition to identifying asphyxia, speech recognition tasks have been used for the diagnosis of a large number of diseases. Some diseases associated with pathological cries include: asphyxia, hypo-acoustic (hearing disorder), hypothyroidism, hyperbilirubinemia, cleft palate, respiratory distress syndrome, ankyloglossia with a deviation of the epiglottis and larynx [10]. Vocal biomarkers have also been associated with: Parkinson's Disease, Alzheimer's Disease, Multiple Sclerosis, and Rheumatoid

Arthritis, Cardiometabolic and Cardiovascular Diseases, COVID-19, and Mental Health Conditions [6].

3 Methods

This section describes our approach to designing and evaluating our selected models for detecting perinatal asphyxia in infant cry. We present the source tasks selected along with representative datasets. We further describe preprocessing steps and the choice of models.

3.1 Dataset

For this project, we intend to use the Baby Chillanto Infant Cry database, which contains 1,049 recordings of healthy neonates, 870 deaf-baby cries, and 340 cry recordings of infants with perinatal asphyxia. Note that, because of the imbalanced data points, different evaluation metrics–particularly a form of recall–will be used to evaluate this dataset. The Baby Chillanto Infant Cry database contains five type of cry signals including deaf, asphyxia, normal and pain. Audio recordings were one second long and sampled at frequencies between 8kHz to 16kHz with 16-bit PCM encoding [13]. This dataset is chosen because it is considered to be the "standard" for cry analysis and because several studies discussed in the literature review use this dataset. Other alternative datasets used in cry analysis include the Dunstan Baby Database, "Donate-a-Cry," "CRIED," "iCOPE," "ChatterBaby," "SPLANN," and "DA-IICT Cry." Of these databases, only the "ChatterBaby" database has more samples (13373 in comparison to 2268 recordings), and the only other database with recordings of babies with asphyxia is the "DA-IICT Cry" database [13]

3.2 Pre-Processing

To reproduce previous research studies[13], the raw audio files from the Baby Chillanto Infant Cry database were converted to Mel-frequency cepstral coefficients (MFCCs). The training, validation, and testing dataset were split by 60 / 20 / 20, respectively.

Moreover, to increase the robustness of our models, the following data augmentations were performed on the training dataset:

Time and Frequency Masking: Time and Frequency Masking were applied to the training set on each epoch. For each mask, 0-10% of the respective mask was applied; for example, if an MFCC contained 1000ms of a recording, 0-10% of the 1000ms of the recording was time-masked (0 to 100 ms of the recording).

In addition, Cepstral Mean Normalization was applied to all datasets to normalize the data.

3.3 Evaluation Metrics

We will use unweighted average recall (UAR) to determine which model is best. Here, UAR is the average recall between both class, where recall is defined as such:

$$Recall = \frac{TP}{TP + FN}$$

where TP is the true positive rate of a class, and FN is the false negative rate of a class. This metric is preferred over typical accuracy measurements because the dataset used is imbalanced in number of classes, and because it is used in previous research over traditional accuracy measurements [19].

3.4 Model Selection

3.4.1 SVMs

A support vector machine model was used as the baseline model in the Ubenwa publication. Our project did not implement an SVM but as it is a baseline the theory is reviewed here.

Briefly, an SVM finds a hyperplane in a hypothesis space different from the input data $\mathbf{x}$. Given a dataset $D = \{(x_1, y_1), (x_2, y_2), ...(x_n, y_n)\}$, the affine hyperplane an SVM produces will maximize the margin between different classifications with the following function:

$$f(\mathbf{w}, \mathbf{x}) = K(\mathbf{w}, \mathbf{x}) + c$$

Where $\mathbf{w}$ is a set of weights, $\mathbf{x}$ is the input, K is the kernel function that defines a dot product in this space, and c is a constant value to allow the SVM to. From the previous studies by Ubenwa, the optimal kernel function found for the Chillanto Infant Cry dataset was the polynomial kernel:

$$K(\mathbf{w}, \mathbf{x}) = (\mathbf{w} \cdot \mathbf{x})^d$$

Where d is a hyper-parameterized integer to denote the polynomial's degree of freedom. The optimized degree is not given but is shown to be tested from one to eight[18].

From Ubenwa, the UAR is determined to be 84.4% [19]. In this paper, an SVM is used as a baseline because an SVM's feasibility in training a small number of samples and its relative simplicity to other Machine Learning models make it a prime candidate for being implemented onto a smartphone. In other words, while there are other models that exist and have better performance in terms of UAR and accuracy, these models are significantly more complex and likely require high computation costs to train. In context with this report, since we are interested in Deep Learning models, this reference provides a baseline to consider but is not worth implementing.

3.4.2 ResNet18

As another baseline, we adopt a residual network (Resnet) architecture, specifically, the *resnet18* architecture [11] due to the model's robustness and historical success in other machine learning tasks while not being as computationally expensive as other Resnet architectures such as *resnet34* or *resnet50*. This model was also implemented by Ubenwa and was shown to achieve 80.0 to 86.5%, depending on the transferred learning task [19]. We also initialize the weights by randomly sampling from a Glorot distribution according to $U(-k, k)$, where $k = \frac{\sqrt{6}}{n_i + n_o}$ (n_i and n_o represent the number of input and output neurons for the respective layer).

Consider the convolution layer that learns the mapping function $F(\mathbf{x})$. A residual block adds a skip connection so that this output layer becomes $F(\mathbf{x}) + x$. Adding this skip connection addresses the degradation problem of gradients when creating deep learning models [11].

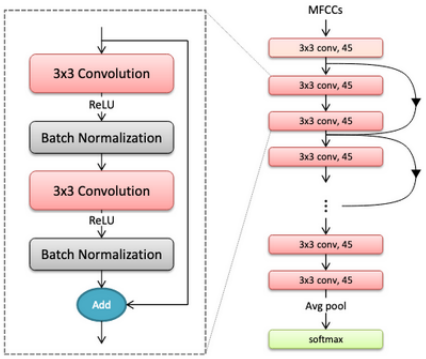
This model serves as another baseline because it not only serves similar functions to the SVM model (i.e., *resnet18* is a relatively simple model in comparison to other machine learning models and may possibly be able to be run on a smartphone) but also because it is more similar to other models compared to in this report. In contrast to an SVM, *resnet18* is a neural network that can be optimized via an optimizer, loss function, etc. This property will allow for better comparisons to other proposed models, as most of our proposed improvements in pathological cry classification.

3.4.3 Transformers

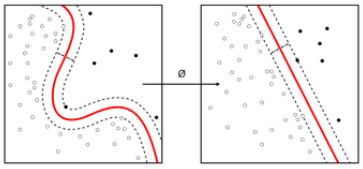
Transformers are deep learning models based on a self-attention mechanism that learns relevant information from a sequence of elements. The self-attention block in transformers models captures long-term dependency between one element in a sequence to other elements.[23, 14].

Although self-attention is the key characteristic defining transformers, their incredible accuracy also comes from their (self) supervised pre-training. In general, transformers — specifically the network weights — are first trained on a large-scale dataset. The weights trained on a large-scale dataset are then fine-tuned to perform tasks on small datasets. [26]

One can think that transformers' performance comes from large datasets. While this might be true, Xu et al showed that transformers can be optimized on small datasets with proper weight initialization and DT-Fixup, an optimization strategy they proposed. [25]



(a) ResNet architecture for infant cry classification



(b) Support Vector Machine

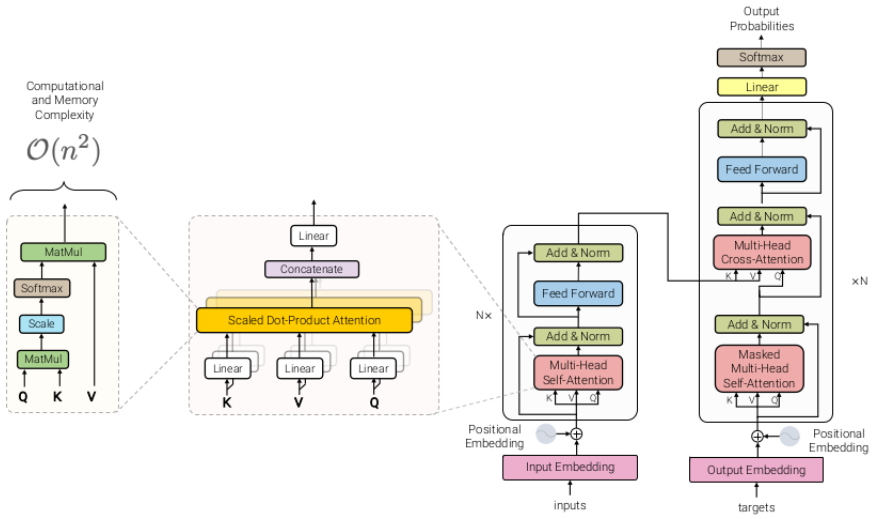


Figure 2: Architecture of the standard Transformer (Vaswani et al., 2017)

4 Results

4.1 Resnet18 Baseline Reimplementation

Details of the reimplementation can be found in tables 1, 2, and 3.

A record of model training of the baseline can be found in Figure 3.

4.2 LAS Model

While we were interested in applying an Attention-based Bi-directional LSTM based of the Listen, Attend, and Spell neural architecture [5] and the transfer learning process using the Common Voice Dataset [3], this model was ultimately unable to be implemented due to issues in the attention module. Nevertheless, we have obtained sufficient results from other models tested to make a sound conclusion.

4.3 Results for the Pre-trained Transformer Model

The pre-trained Transformer model used was a word2vec model which is trained from 960 hours of LibreSpeech audio [4]. However, with this model, we could not obtain sufficient results, retrieving a UAR of 0.5 at every epoch. This is shown in table 4.

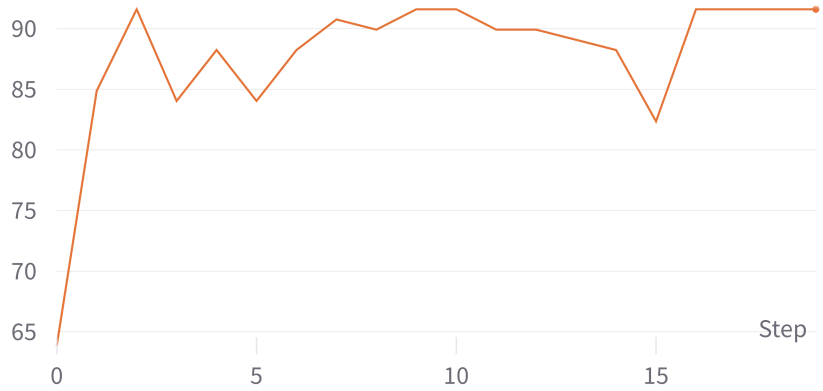


Figure 3: Validation accuracy vs. epoch for replication model

4.4 Bi-Directional LSTM Model without Attention

We used a bidirectional LSTM to perform the classification. Specifically, we adapt the sequence to sequence model built in HW3P2 to our binary classification task. We have extracted features from the data with a CNN and added a final linear layer for classification. After training for 30 epochs, we got an UAR of 0.61 before augmenting our initial dataset and 0.67 after. 7 shows the results for of training the Bi-Directional LSTM after augmenting the data with TanH() distortion.

4.5 Linear Neural Networks

For our linear neural networks, we performed a grid-search to determine what set of neurons increased or decreased UAR. The number of hidden layers searched for were: 64, 128, 256, 512, 1024, and 2048. One to three neuron layers were assessed. A constant dropout of 0.25 was applied, with Xavier Uniform Weight Initialization, Batch Normalization, and with a GELU activation.

Results showed that, regardless of linear layer combination, UAR ranged from 0.595 to 0.605.

4.6 Bootstrapping

The Baby Chillanto dataset is small. Thus, augmenting the data is an essential to achieve good performance. To augment we have opted for bootstrapping the dataset with two techniques: tanh() distortion and room reverberation.

4.6.1 Tanh() Distortion

It adds a rounded soft clipping (kind of distortion) to the audio signal. The distortion amount is proportional to the loudness of the input and its gain. The Tanh function is symmetric. After adding a distortion, the positive and negative parts of the signal are squashed. Below are respectively the and with a distortion G added:

Tanh function without distortion

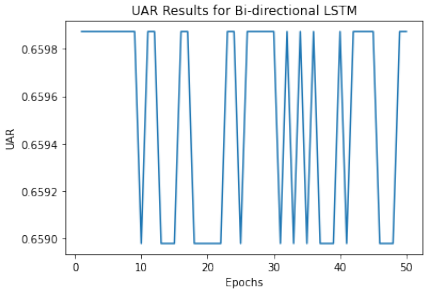
$$\frac{sinh(x)}{cosh(x)} = \frac{e^x - e^- x}{e^x + e^- x}$$

Tanh function with distortion G added:

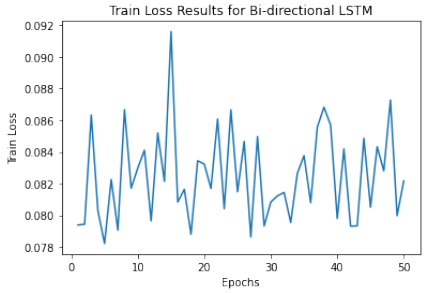
$$\frac{e^{x*(a+G)} - e^{x*(b-G)}}{e^{x*G} - e^{x*-G}}$$

4.6.2 Room Reverberation

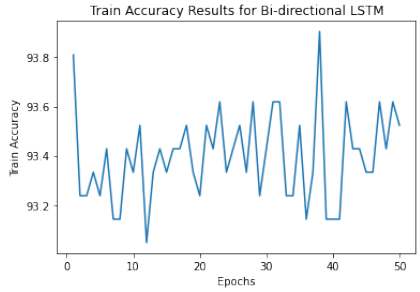
I consists in creating audio signals by simulating the reverberation of sound inside rooms. For example using Room Impulse Response (RIR) an audio signal can be generated as if has been uttered in conference room. To perform room reverberation, we use the RIR data. The RIR is cleaned up and



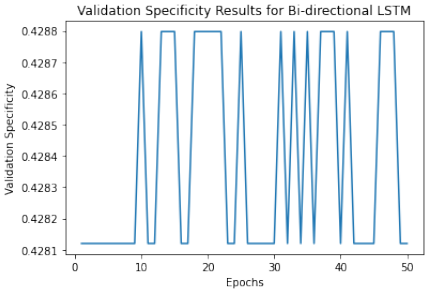
(a) UAR results for Bi-directional LSTM



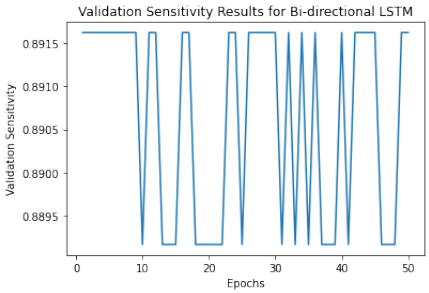
(b) Train loss results for Bi-Directional LSTM



(c) Train accuracy results for Bi-Directional LSTM



(d) Validation specitivity results for Bi-Directional LSTM



(e) Validation sensitivity results for Bi-Directional LSTM

Figure 4: Results of training Bi-Directional LSTM on augmented data with TanH() distortion

used to extract the main impulse, normalize the signal power, then flip along the time axis. Finally, the signal is convolved with a RIR filter.

Many parameters are used to reverberate an audio signals. Among them we have the which tells us the way the sound will be generated within the room.

An impulse response measurement can tell us a great deal about a room and the way sound will be reproduced within it. It can show us what kinds of treatment will be helpful and whether treatments have been correctly applied to achieve the best results.

The results of bootstrapping data augmentations are shown in table 5. Notably, this larger dataset's performance increased the UAR from 0.6 to 0.7.

4.7 Summary

Below are tables describing the results as mentioned from sections 4.1 - 4.6. While these results are not mimicking the results from the Ubenwa paper [21], our results are still conclusive since we can still determine what aspects of Pathological Cry Data Analysis improved model performance.

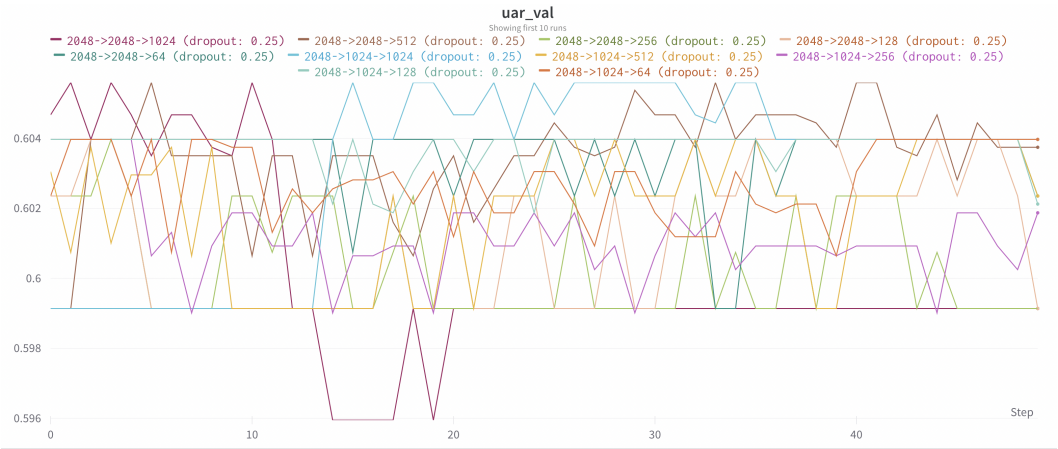


Figure 5: 10 examples of ablation results using the linear model. This figure illustrates how UAR did not effectively change regardless of layer configuration.

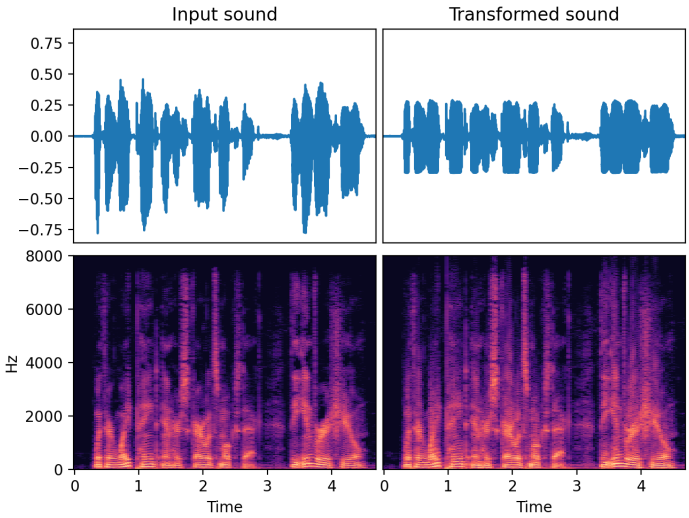
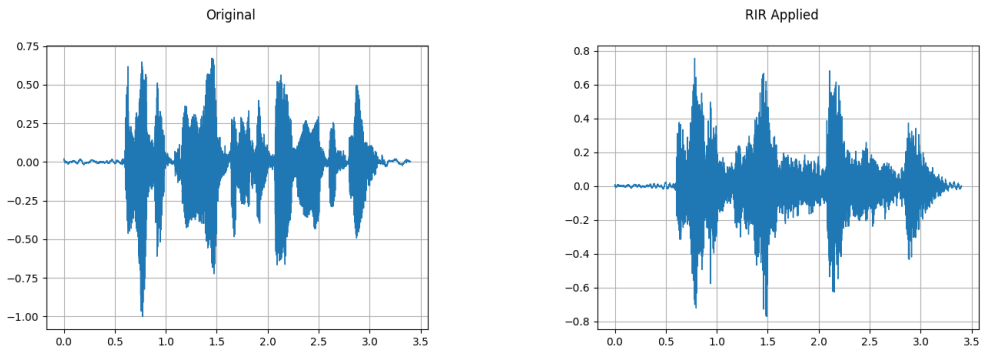


Figure 6: Input-output from a Tanh() distortion [1]



(a) Audio signal without RIR

(b) Audio signal with RIR

Figure 7: Room reverberation data augmentation

Table 1: Pre-processing

Implementation	Downsampling	Spectral conversion
Onu et al [20]	8 kHz	MFCC, 40 frames, 30ms window , 10ms shift
Re-implementation	8 kHz	MFCC, 40 frames, 50ms window , 25ms shift

Table 2: Performance

Implementation	Accuracy	Sensitivity
Specificity	UAR	
Onu et al [20]	0.86	0.84
0.889	0.865	
Re-implementation	0.91	1
0.376	0.688	

Table 3: Architecture

Implementation	Layers
Onu et al [20]	Conv, Residual (conv, relu, batchnorm) x6, Conv, linear
Re-implementation	Conv, ReLU, MaxPool, Conv, batchnorm, maxpool, conv, batchnorm, maxpool, linear

Table 4: Performance

Model	UAR
Attention-based Bi-Directional LSTM	NA
Bi-Directional LSTM	0.61
Pre-trained Transformer	0.5
Linear	0.605
ResNet18	0.60

Table 5: Performance

Dataset	UAR
Ubenwa unmodified	0.605
Tanh	0.70
Reverb	0.69
Tanh and Reverb	0.696

5 Conclusion and Future Work

Voice-based diagnosis of disease is an emerging area of research with potential for huge impact. The ability to deploy these models in virtually any environment (include outside of the clinic) and by untrained individuals offers strong research direction.

In this report, we demonstrate a strong need for larger and more diverse biomedical audio datasets. New model development has outstripped the pace of data collect to the extent that state-of-the-art models require massive data corups. This exciting development indicates that simply through more rigorous data collection techniques many lives could be saved.

There are several avenues for future work in the area of transformers and their performance on small datasets. Of these possibilities, we identify three that are particularly interesting and promising. The first avenue is to investigate and apply optimization techniques that can help increase the performance of transformers on small datasets, such as DT-Fixup and T-fixup, which effectively remove warm-up and layer normalization in exchange for attention blocks and, for DT-Fixup, MLPs. This method of weight initialization has shown to accomodate transformers to smaller datasets. [27] The second

avenue is to redefine the task itself to better accommodate for the lack of data. For example, one could look to apply few-shot learning approaches, where the model is trained on a few examples with supervised information to then classify new data points [22]. The third avenue is to explore and develop models that do not necessarily require large amounts of data to achieve high-accuracy results. For example, generative adversarial networks (GANs) have been shown to be able to learn complex distributions from a limited number of samples [8]. These potential avenues provide exciting possibilities for future work we believe that could lead to significant advances in pathological cry analysis research.

6 Acknowledgements

The Baby Chillanto Data Base is a property of the Instituto Nacional de Astrofisica Optica y Electronica – CONACYT, Mexico.

We like to thank Dr. Carlos A. Reyes-Garcia, Dr. Emilio Arch-Tirado and his INR-Mexico group, and Dr. Edgar M. Garcia-Tamayo for their dedication of the collection of the Infant Cry database. c. And a citation to the article “Orion Fausto Reyes-Galaviz, Sergio Daniel Cano-Ortiz, Carlos Alberto Reyes-García. "Evolutionary-Neural System to Classify Infant Cry Units for Pathologies Identification in Recently Born Babies". Proceedings of the Special Session MICAI 2008, Pg. 330- 335. Eds. Alexander Gelbukh & Eduardo Morales. IEEE Computer Society. ISBN: 978-0-7695-3441-1.

We like to thank Professor David Vernon for paving the way for this work through his excellent case study on the Ubenwa solution, his useful advice and insights during this project. Kindly find more details in his Artificial Intelligence and Machine Learning in Africa course.

7 GitHub Link

Here is the GitHub Repository that contains our code: `https://github.com/pamely/idl-project/tree/master`

References

[1] Tanhdistortion. https://iver56.github.io/audiomentations/waveform_transforms/tanh_distortion/. Accessed: 2022-12-15.

[2] Franz Anders, Mario Hlawitschka, and Mirco Fuchs. Comparison of artificial neural network types for infant vocalization classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:54–67, 2020.

[3] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *CoRR*, abs/1912.06670, 2019.

[4] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477, 2020.

[5] William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. Listen, attend and spell. *CoRR*, abs/1508.01211, 2015.

[6] Guy Fagherazzi, Aurélie Fischer, Muhannad Ismael, and Vladimir Despotovic. Voice for health: the use of vocal biomarkers from research to clinical practice. *Digital biomarkers*, 5(1):78–88, 2021.

[7] National Center for Biotechnology Information. Birth asphyxia.

[8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[9] M Hariharan, Lim Sin Chee, and Sazali Yaacob. Analysis of infant cry through weighted linear prediction cepstral coefficients and probabilistic neural network. *Journal of medical systems*, 36(3):1309–1315, 2012.

[10] M. Hariharan, J. Saraswathy, R. Sindhu, Wan Khairunizam, and Sazali Yaacob. Infant cry classification to identify asphyxia using time-frequency analysis and radial basis neural networks. *Expert Systems with Applications*, 39(10):9515–9523, 2012.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[12] Chunyan Ji, Ming Chen, Bin Li, and Yi Pan. Infant cry classification with graph convolutional networks. *arXiv preprint arXiv:2102.02909*, 2021.

[13] Chunyan Ji, Thosini Bamunu Mudiyanselage, Yutong Gao, and Yi Pan. A review of infant cry analysis and classification. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1):1–17, 2021.

[14] Salman H. Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *CoRR*, abs/2101.01169, 2021.

[15] Joy E Lawn, Hannah Blencowe, Shefali Oza, Danzhen You, Anne CC Lee, Peter Waiswa, Marek Lalli, Zulfiqar Bhutta, Aluisio JD Barros, Parul Christian, et al. Every newborn: progress, priorities, and potential beyond survival. *The lancet*, 384(9938):189–205, 2014.

[16] Jialu Li, Mark Hasegawa-Johnson, and Nancy L McElwain. Analysis of acoustic and voice quality features for the classification of infant and mother vocalizations. *Speech Communication*, 133:41–61, 2021.

[17] Newborn Mortality. World health organization.

[18] Charles C. Onu. Harnessing infant cry for swift, cost-effective diagnosis of perinatal asphyxia in low-resource settings. In *2014 IEEE Canada International Humanitarian Technology Conference - (IHTC)*, pages 1–4, 2014.

[19] Charles C. Onu, Jonathan Lebensold, William L. Hamilton, and Doina Precup. Neural transfer learning for cry-based diagnosis of perinatal asphyxia. *CoRR*, abs/1906.10199, 2019.

[20] Charles C. Onu, Jonathan Lebensold, William L. Hamilton, and Doina Precup. Neural transfer learning for cry-based diagnosis of perinatal asphyxia. September 2019.

[21] Charles C Onu, Innocent Udeogu, Eyenimi Ndiomu, Urbain Kengni, Doina Precup, Guilherme M Sant'Anna, Edward Alikor, and Peace Opara. Ubenwa: Cry-based diagnosis of birth asphyxia. *arXiv preprint arXiv:1711.06405*, 2017.

[22] Archit Parnami and Minwoo Lee. Learning from few examples: A summary of approaches to few-shot learning, 2022.

[23] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *CoRR*, abs/2009.06732, 2020.

[24] Hai-tao Xu, Jie Zhang, and Li-rong Dai. Differential time-frequency log-mel spectrogram features for vision transformer based infant cry recognition. *Proc. Interspeech 2022*, pages 1963–1967, 2022.

[25] Peng Xu, Dhruv Kumar, Wei Yang, Wenjie Zi, Keyi Tang, Chenyang Huang, Jackie Chi Kit Cheung, Simon J. D. Prince, and Yanshuai Cao.

[26] Peng Xu, Dhruv Kumar, Wei Yang, Wenjie Zi, Keyi Tang, Chenyang Huang, Jackie Chi Kit Cheung, Simon J.D. Prince, and Yanshuai Cao. Optimizing deeper transformers on small datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2089–2102, Online, August 2021. Association for Computational Linguistics.

[27] Peng Xu, Wei Yang, Wenjie Zi, Keyi Tang, Chengyang Huang, Jackie Chi Kit Cheung, and Yanshuai Cao. Optimizing deeper transformers on small datasets: An application on text-to-sql semantic parsing. *CoRR*, abs/2012.15355, 2020.