# SELF-SUPERVISED LEARNING FOR INFANT CRY ANALYSIS

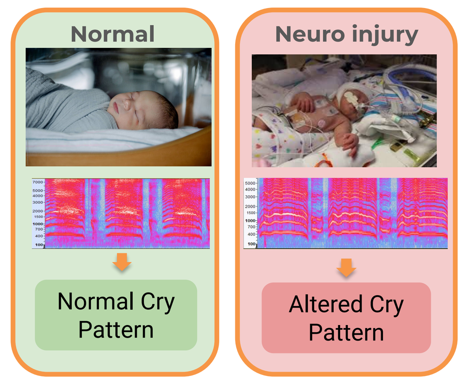
Arsenii Gorin[1], Cem Subakan[2,3,4], Sajjad Abdoli[1], Junhao Wang[1], Samantha Latremouille[1], Charles C Onu[1,2]
[1]Ubenwa Health, Montréal, CA [2]Mila - Quebec AI Institute, Montréal, CA [3]Université Laval, Québec City, CA [4]Concordia University, Montréal, CA

## Summary


- **SSL for cry-based detection of neurological injury and triggers** (pain, hunger, and discomfort)
- **Large database of cry recordings** with clinical indications of more than a thousand newborns
- **SSL pre-training (SimCLR)** of CNN on large audio (VGGSound) outperforms supervised pre-training and training from scratch
- **SSL-based domain adaptation (DA)** using unlabeled infant cries further improves results (especially with limited annotated data)
- **Replay of the original data** is important for efficient DA


## Context

- **Birth asphyxia (respiratory distress)** is a common cause of severe health problems, including neurological injury and death
- Cry characteristics extensively studied for its detection (clinical and ML)




- Annotating large clinical data is costly and time-consuming
- Unlabeled audio and SSL reduce the cost of clinical solutions

## Dataset used in this study

**Curated subset of larger database collected by Ubenwa (2020-22)**

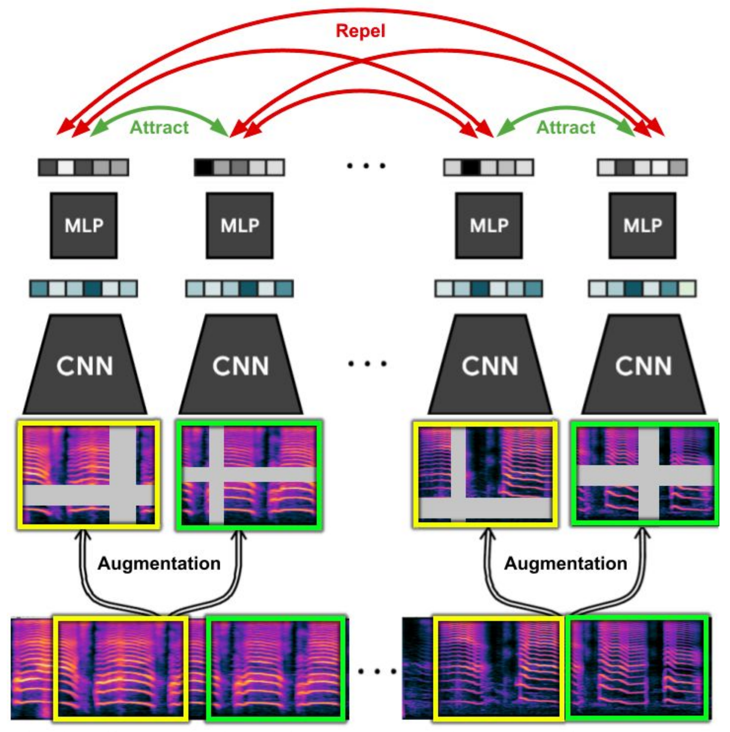
- **3 countries, 4 hospitals, 23.6 hours of cry**
- 2,022 recordings with clinical annotations
- 1,149 recordings with cry trigger labels (pain, hunger, discomfort)

*Neuro injury data split*

| | Healthy | | | Neuro injury | | |
|---|---|---|---|---|---|---|
| | Train | Val | Test | Train | Val | Test |
| Records | 1360 | 247 | 238 | 92 | 40 | 45 |
| Patients | 885 | 165 | 163 | 75 | 33 | 38 |
| Hours | 10.3 | 1.9 | 2.0 | 0.8 | 0.3 | 0.3 |

## SimCLR pre-training

- Pre-train CNN14 on large generic audio - VGGSound (550 hours)
- SimCLR maximizes similarity between distorted copies of audio spectrograms created by
  - Random chunk
  - SpecAugment masking
- Good results in image, music and audio classification




Adapted from Chen, T, et al. "Big self-supervised models are strong semi-supervised learners.", 2020

## Evaluation protocol

- Add linear classifier for target task (neuro injury or triggers)
- Train using labeled data
  - ***Linear probing*** - keep pre-trained model frozen
  - ***Linear+BN*** - also update batch-norm layers
  - ***End-to-end*** - update all parameters with smaller LR for encoder

## Supervised vs SimCLR pre-training

***Is it better to pre-train with supervised or self-supervised objective?***

***Do we need to update the encoder?***

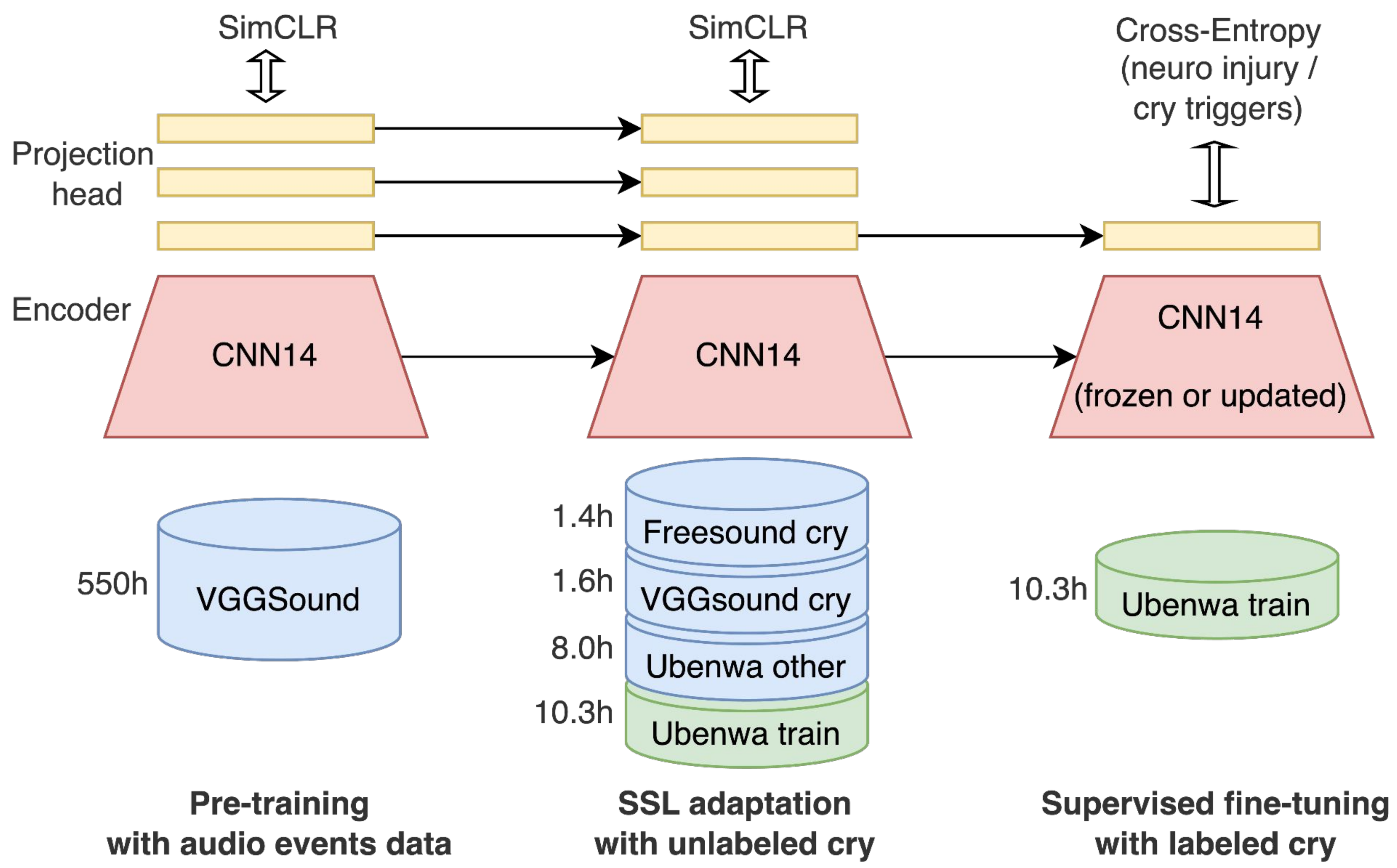
*AUC on neuro injury (mean and standard error for 10 randomized runs)*

| Pre-training | Linear | Linear+BN | End-to-end |
|---|---|---|---|
| Supervised | **75.5 ± 0.6** | 75.9 ± 0.7 | 80.0 ± 0.7 |
| SimCLR | 71.3 ± 0.9 | **78.5 ± 1.2** | **83.9 ± 0.6** |
| Trainable parameters | 4,096 | 24,482 | 76,116,450 |

- SSL performs significantly better than supervised pre-training
- Need to at least fine-tune BN or better the whole network
- Any pre-training is better than training from scratch (*74.6 ± 1.7*)

## SSL cry domain adaptation (DA)

***Can we further adapt generic CNN using unlabeled cry?***




- Instead of directly fine-tuning pre-trained model on labeled data **adapt the pre-trained model with SimCLR using cry data**
- We can use data without clinical annotations (FreeSound, etc)

## SSL domain adaptation (DA) results

- **SimCLR pre-training** with VGGSound
- **SimCLR cry adaptation** with train (10h) and larger cry data (21h)

*Neuro injury performance with domain adaptation (AUC)*

| Domain adaptation | Linear | Linear+BN | End-to-end |
|---|---|---|---|
| None - VGGSound only | 71.3 ± 0.9 | 78.5 ± 1.2 | **83.9 ± 0.6** |
| 10h cry (train) | 78.8 ± 0.5 | 78.0 ± 0.7 | 80.8 ± 0.8 |
| 21h cry (train + extra) | 79.8 ± 0.4 | 81.3 ± 0.5 | 81.3 ± 0.7 |
| 21h cry + VGGSound replay | 80.8 ± 0.5 | **83.3 ± 0.6** | **85.0 ± 0.9** |

DA significantly improves linear probing;
Using only train data, DA performs similar to updating BN of non-adapted model

DA with replay results in superior performance for all evaluations

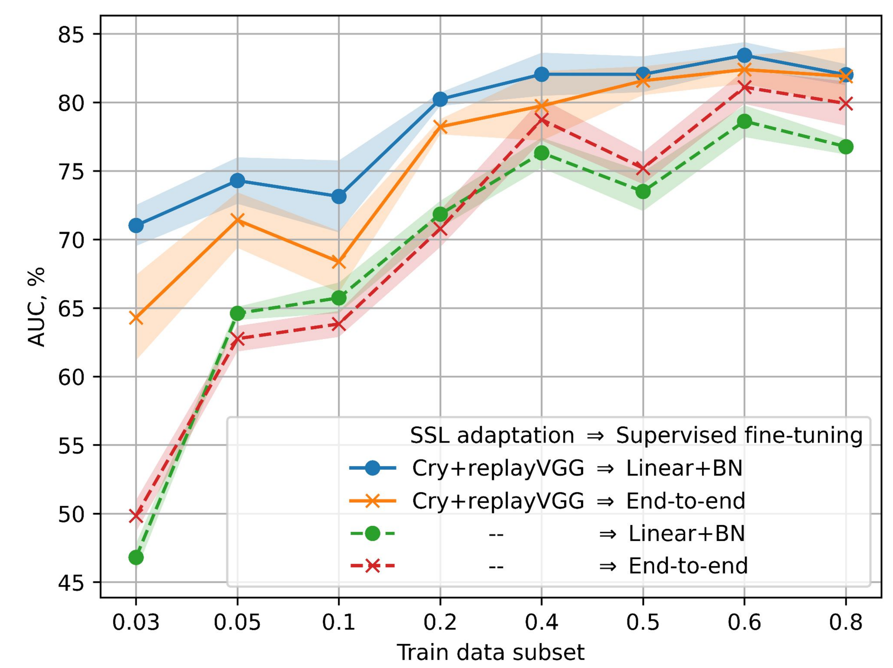
DA without replay degrades end-to-end fine-tuning

*Performance on cry trigger task (~2x less labeled data)*

| Domain adaptation | Linear | Linear+BN | End-to-end |
|---|---|---|---|
| None - VGGSound only | 65.9 ± 0.8 | 69.5 ± 0.7 | 69.0 ± 0.9 |
| 10h cry | 71.7 ± 0.5 | **75.4 ± 0.8** | 72.4 ± 1.4 |
| 21h cry | 74.5 ± 0.4 | 74.7 ± 0.4 | 72.0 ± 1.8 |
| 21h cry + VGGSound replay | 74.2 ± 0.4 | **75.6 ± 0.6** | 74.4 ± 0.7 |

## Fine-tuning using subsets of labeled data

***What happens with pre-trained and cry adapted model if we only have a small portion of labeled data for fine-tuning?***




- Adapted model yields >70% AUC with only 3% of data
- With 20% of labeled data, adapted model outperforms supervised baseline
- Linear+BN performs better than end-to-end fine-tuning with limited labeled data

## Conclusion

- SSL pre-training on generic audio significantly enhances performance of cry based neuro injury detection
- SimCLR pre-training on VGGSound outperforms supervised pre-training when further fine-tuned end-to-end on neuro injury
- Straightforward SSL domain adaptation improves linear evaluation, but replay of VGGSound is important for best transferability