Potential use of AI for Healthcare and Well-being: Study Case of Voice Technology

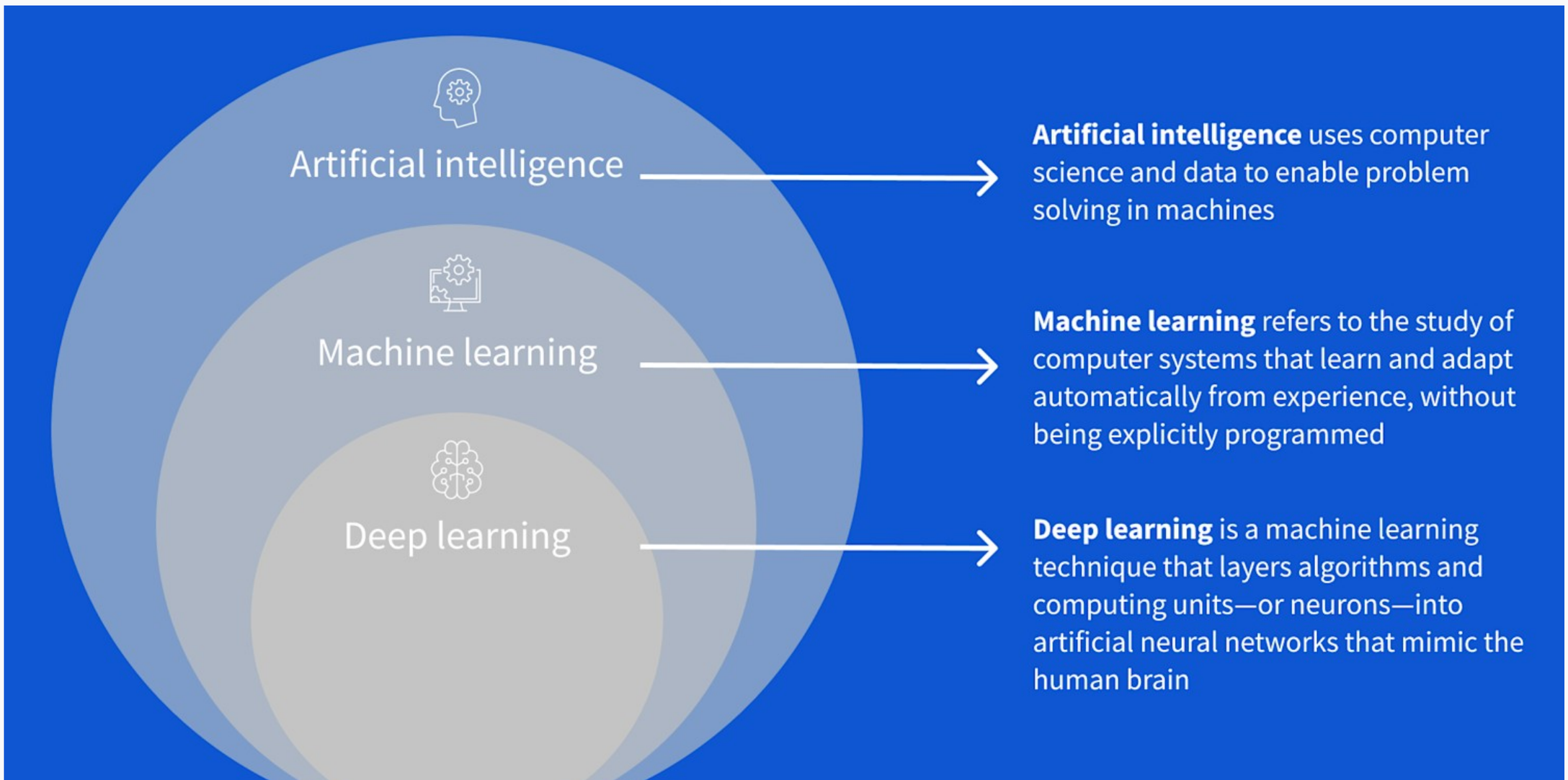
Bagus Tris Atmaja
Email: bagus.tris@naist.ac.jp

version 2025/2/17

Overview

- AI, Deep Learning, Machine Learning
- Healthcare and Well-being
- My research on voice technology for healthcare and well-being:
  - Pathological Detection
  - Cough segmentation
  - TB classification
  - Dementia prediction
- Demo (nkululeko toolkit, if time permits)

AI, ML, DL



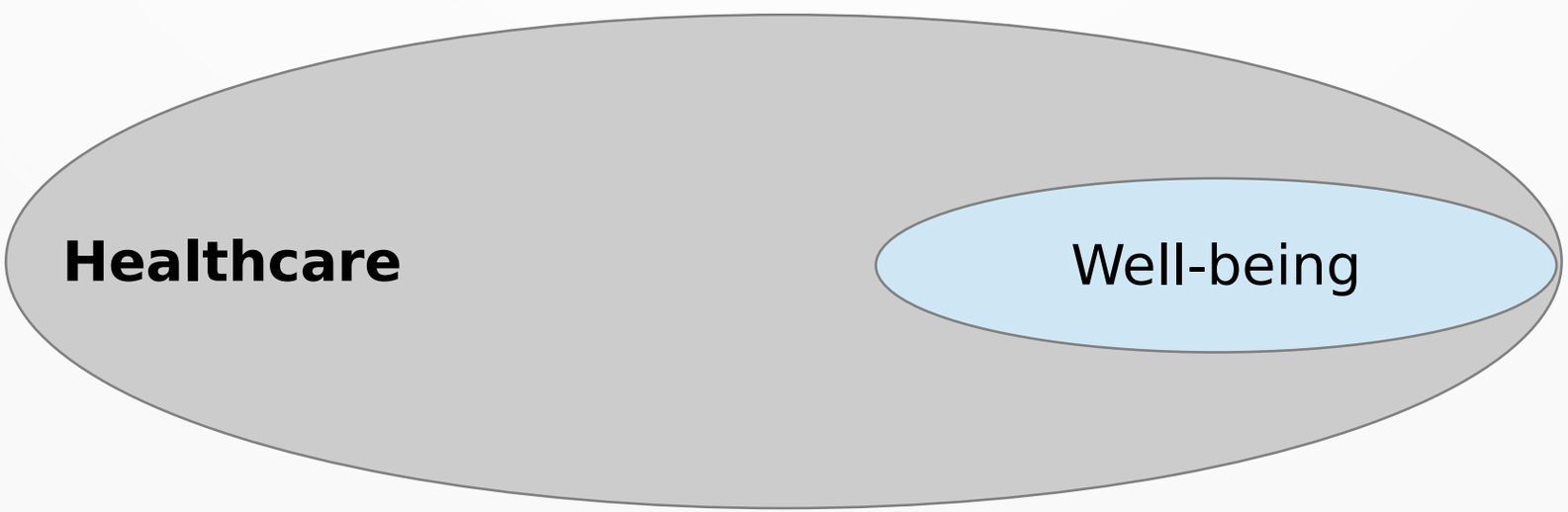
https://www.coursera.org/articles/ai-vs-deep-learning-vs-machine-learning-beginners-guide

Healthcare and well-being

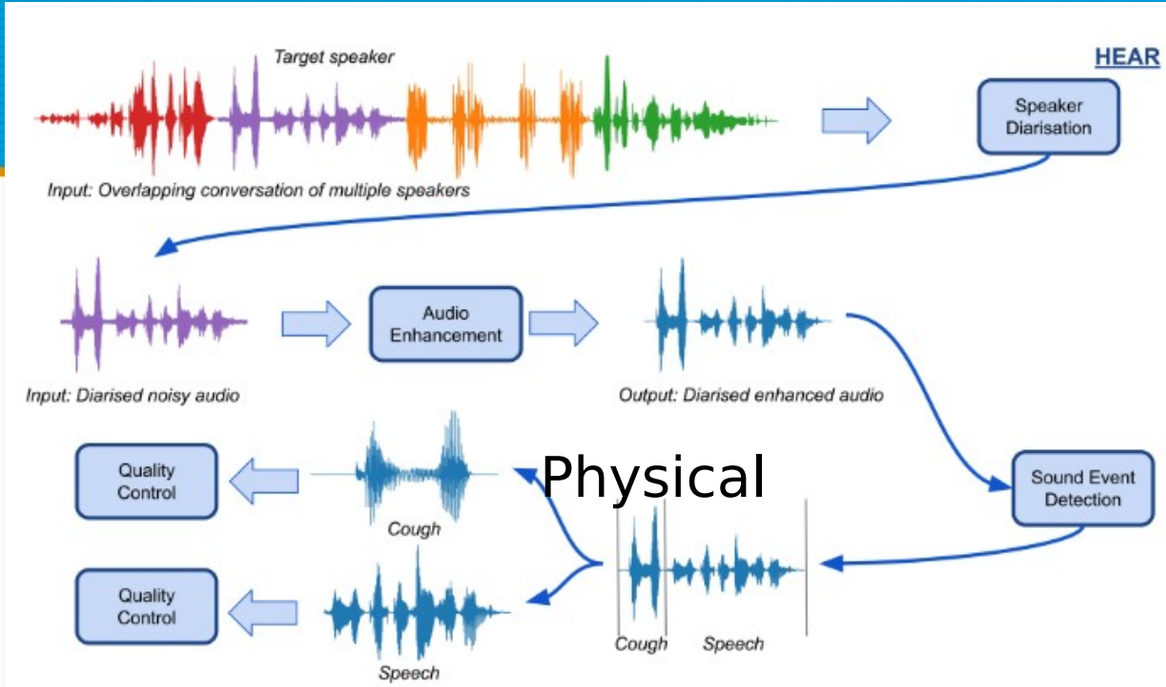
Healthcare → improvement or maintenance of health via the prevention, diagnosis, treatment, rehabilitation, and palliative care (easing suffering)

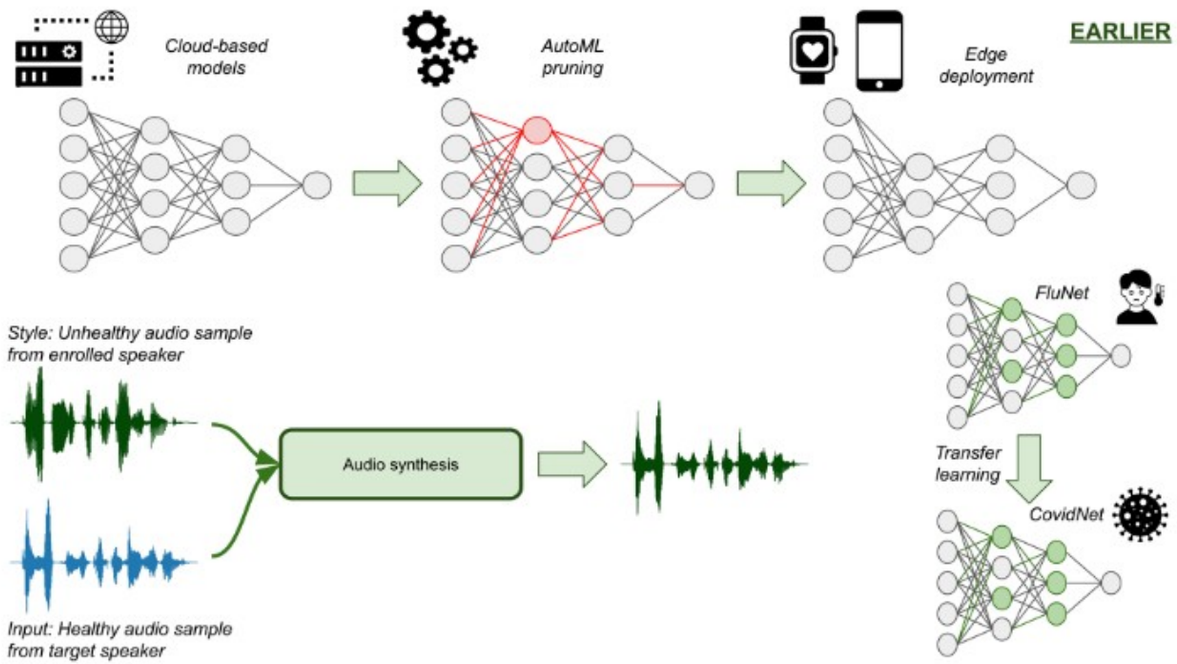
Healthcare
- Physical
- Mental → Well-being

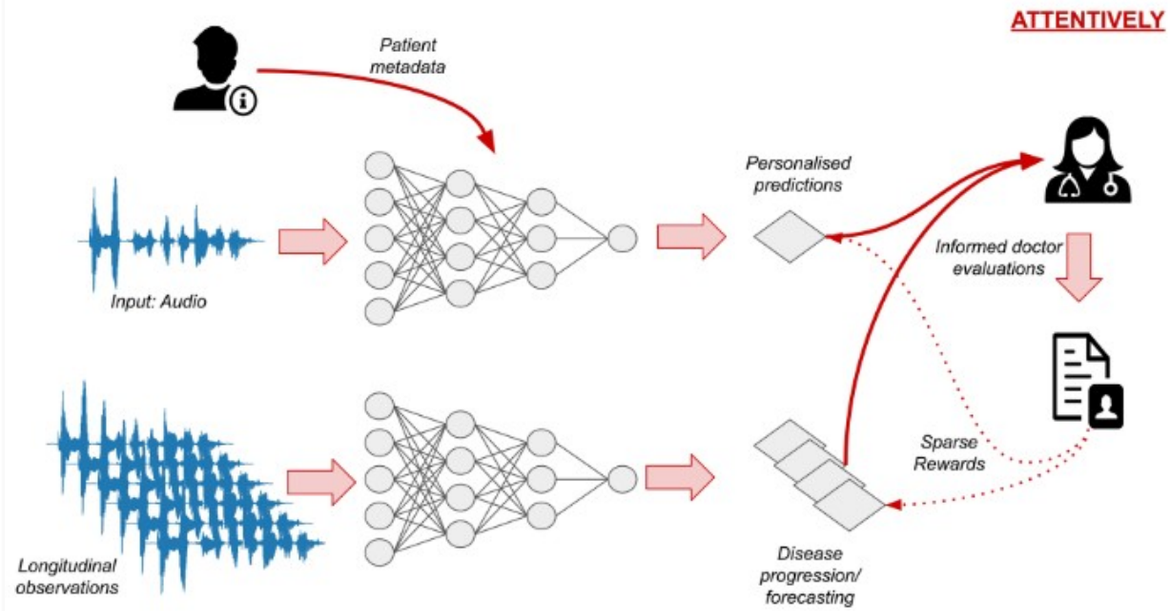
Well-being: a complex state of feeling good about life, encompassing physical, mental, emotional, and social health factors

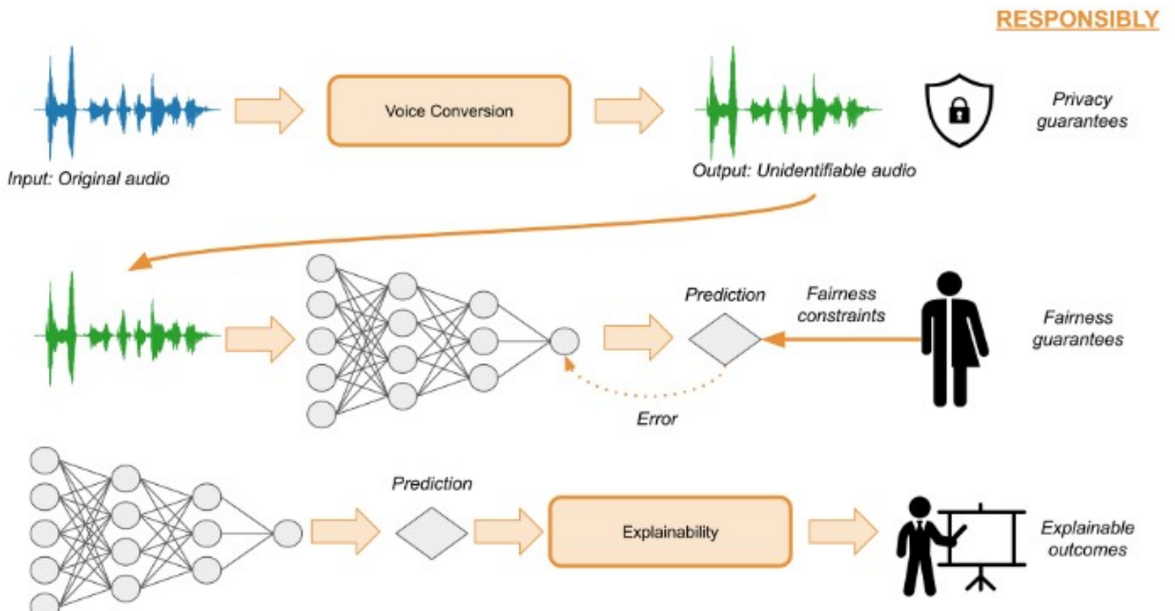


Four Pillars of Healthcare Audio AI (Andreas et al., 2023)





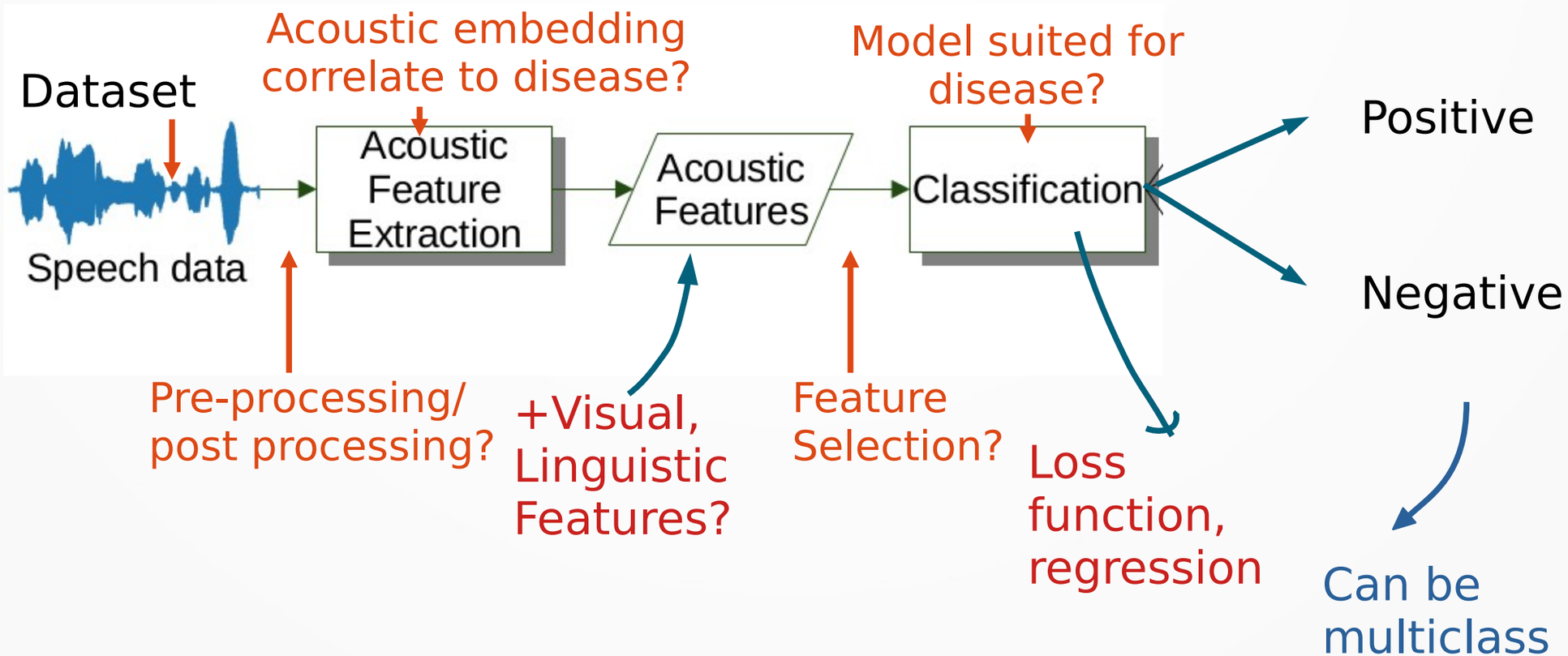




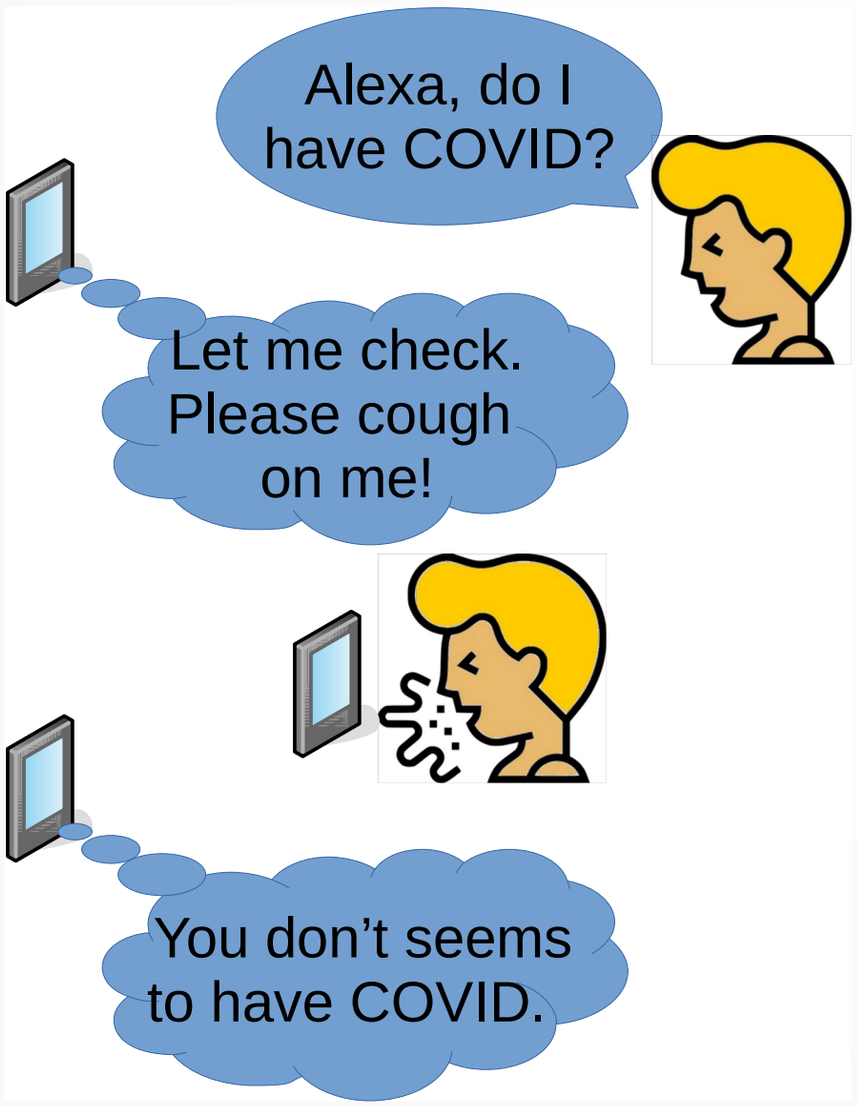
Research on audio-based healthcare and well-being

- Speech Emotion Recognition
- COVID-19 Detection
- Cough Segmentation
- Depression Detection
- Laughter classification (happy vs. evil laughter)
- Pathological speech detection
- Dementia prediction
- TB detection

Typical Workflow



COVID-19 Detection



Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-025-02433-z

ORIGINAL RESEARCH

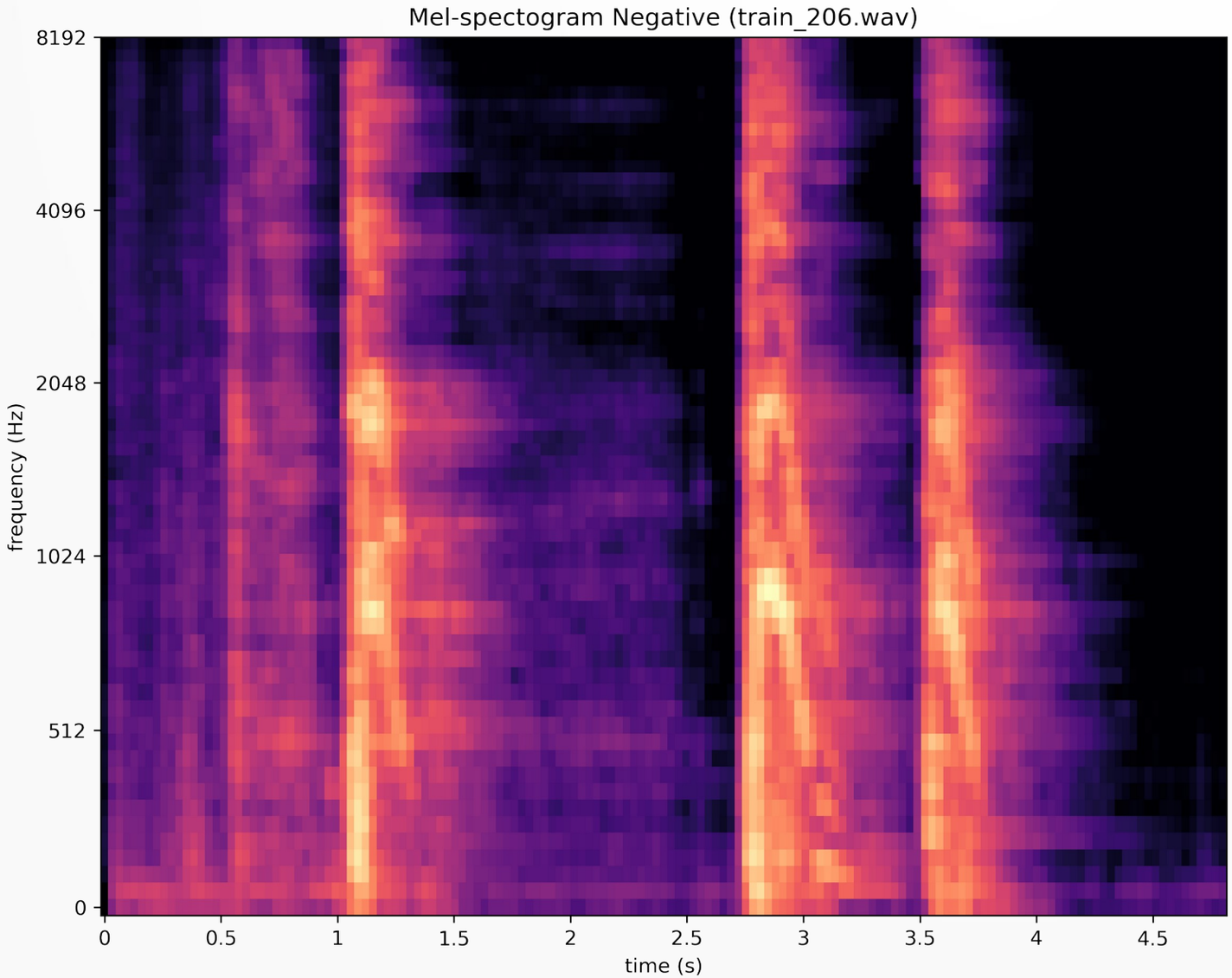
Cross-dataset COVID-19 transfer learning with data augmentation

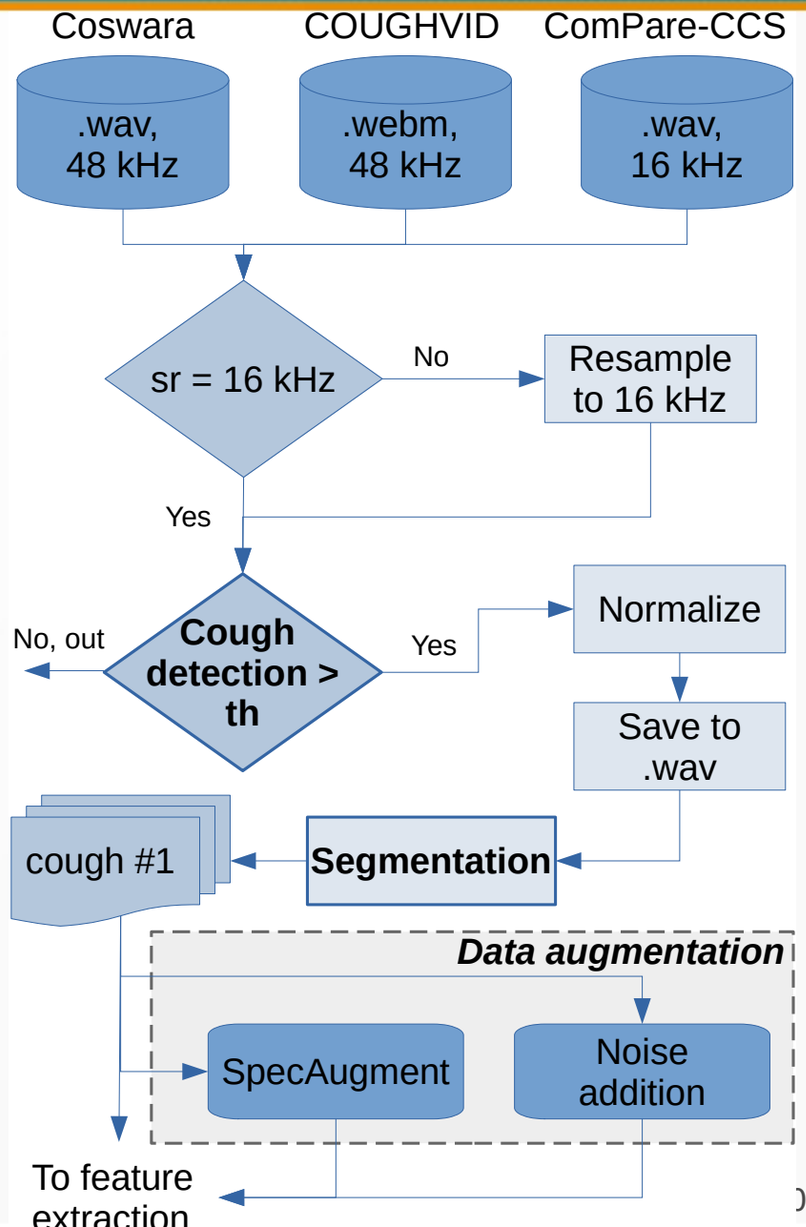
Bagus Tris Atmaja[2] · Zanjabila[1] · Suyanto[1] · Wiratno Argo Asmoro[1] · Akira Sasou[2]

Received: 29 August 2024 / Accepted: 20 January 2025

Atmaja, B. T., Zanjabila, Suyanto, Asmoro, W. A., & Sasou, A. (2025). Cross-dataset COVID-19 transfer learning with data augmentation. International Journal of Information Technology. https://doi.org/10.1007/s41870-025-02433-z

Acoustic Feature Extraction





COVID-19 detection result

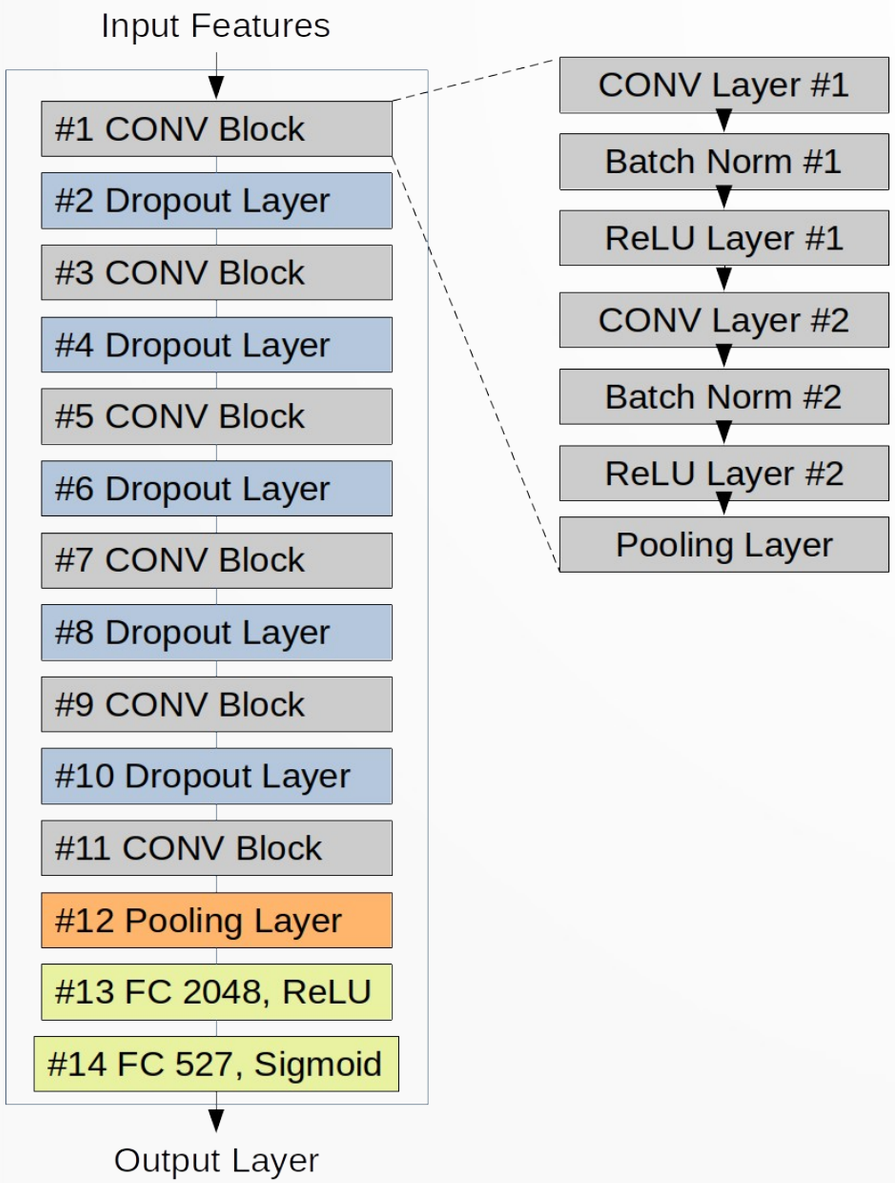


Table 4 UA results on different data augmentation methods

Segmentation method	Augmentation method	UA (%)
Hysteresis comparator	Without augmentation	81.68
	SpecAugment	**86.36**
	Noise addition	81.40
	SpecAugment + Noise addition	83.19

Reference	Data Aug	Feature	Classifier	Ensemble	UA
Baseline [31]	×	openSMILE, openXBOW, DeepSpctrum, AuDeep	SVM, End2You	√	73.90
Casanova et al. [23]	√	log mel spectrogram	CNN14	√	75.90
Illium et al. [49]	√	log mel spectrogram	Vision Transformer	×	76.90
Solera-Urena et al. [50]	×	TDNN-F, VGGish, PASE+	SVM	√	69.30
Suyanto et al. [26]	×	log mel spectrogram	CNN14	×	83.19
Ours	√	log mel spectrogram	CNN14+HPO	×	**88.19**

Cough Segmentation

Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-023-01626-8

Check for updates

ORIGINAL RESEARCH

Comparing hysteresis comparator and RMS threshold methods for automatic single cough segmentations

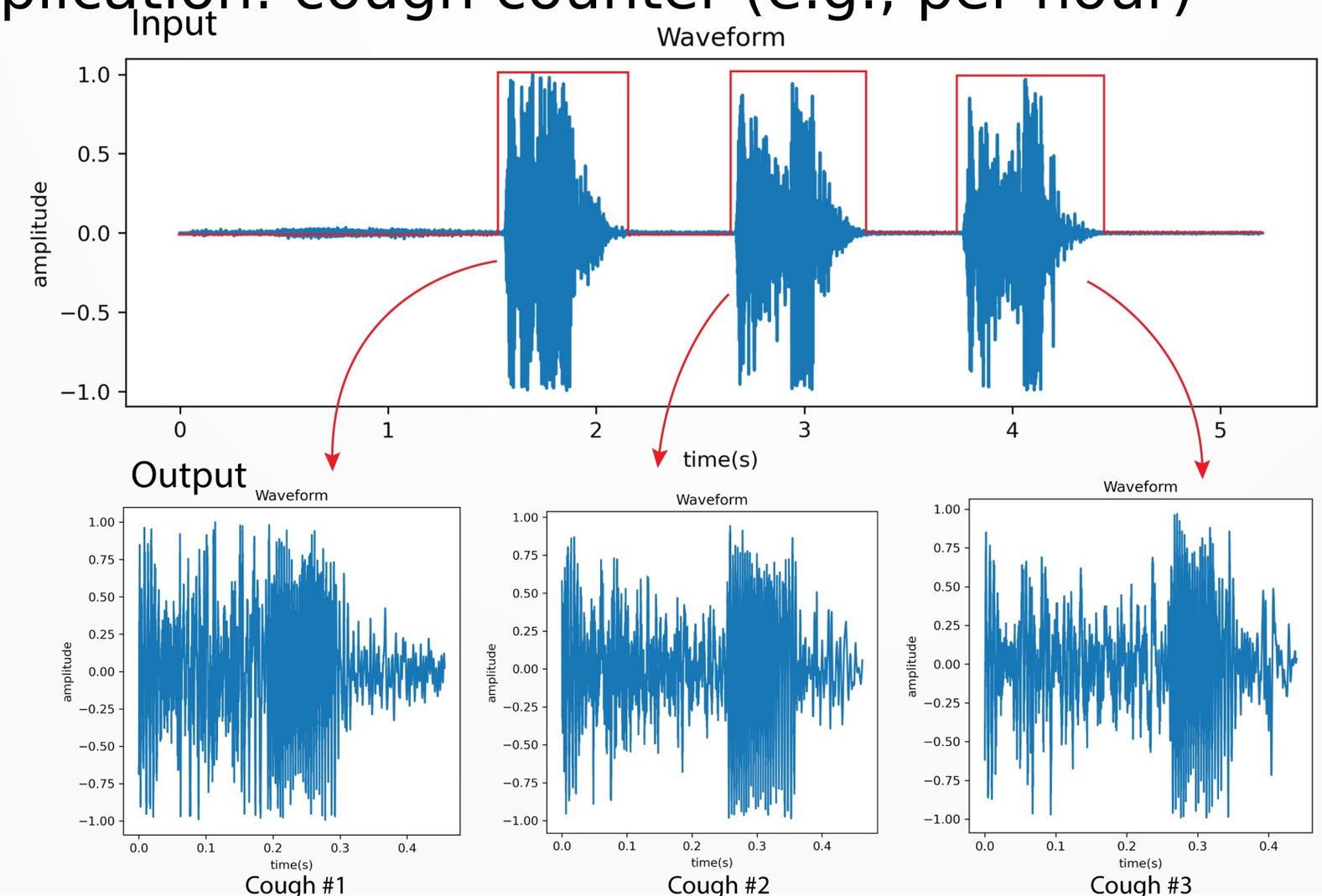
Bagus Tris Atmaja[1,2] · Zanjabila[2] · Suyanto[2] · Akira Sasou[1]

Received: 30 July 2023 / Accepted: 6 November 2023

Atmaja, B. T., Zanjabila, Suyanto, & Sasou, A. (2023). Comparing hysteresis comparator and RMS threshold methods for automatic single cough segmentations. International Journal of Information Technology, 0123456789.

Cough Segmentation

- Segmenting multiple cough into individual cough
- Application: cough counter (e.g., per hour)



Cough Segmentation result

Segmentation method	Fleiss' Kappa	Interpretation
Manual segmentation	0.345	fair
Hysteresis comparator	0.246	fair
RMS threshold	0.486	moderate

Segmentation method	N	Single-cough	Precision
Manual segmentation	121	60	49.59%
Hysteresis comparator	120	88	73.33%
RMS threshold	150	105	70.00%

Depression Detection

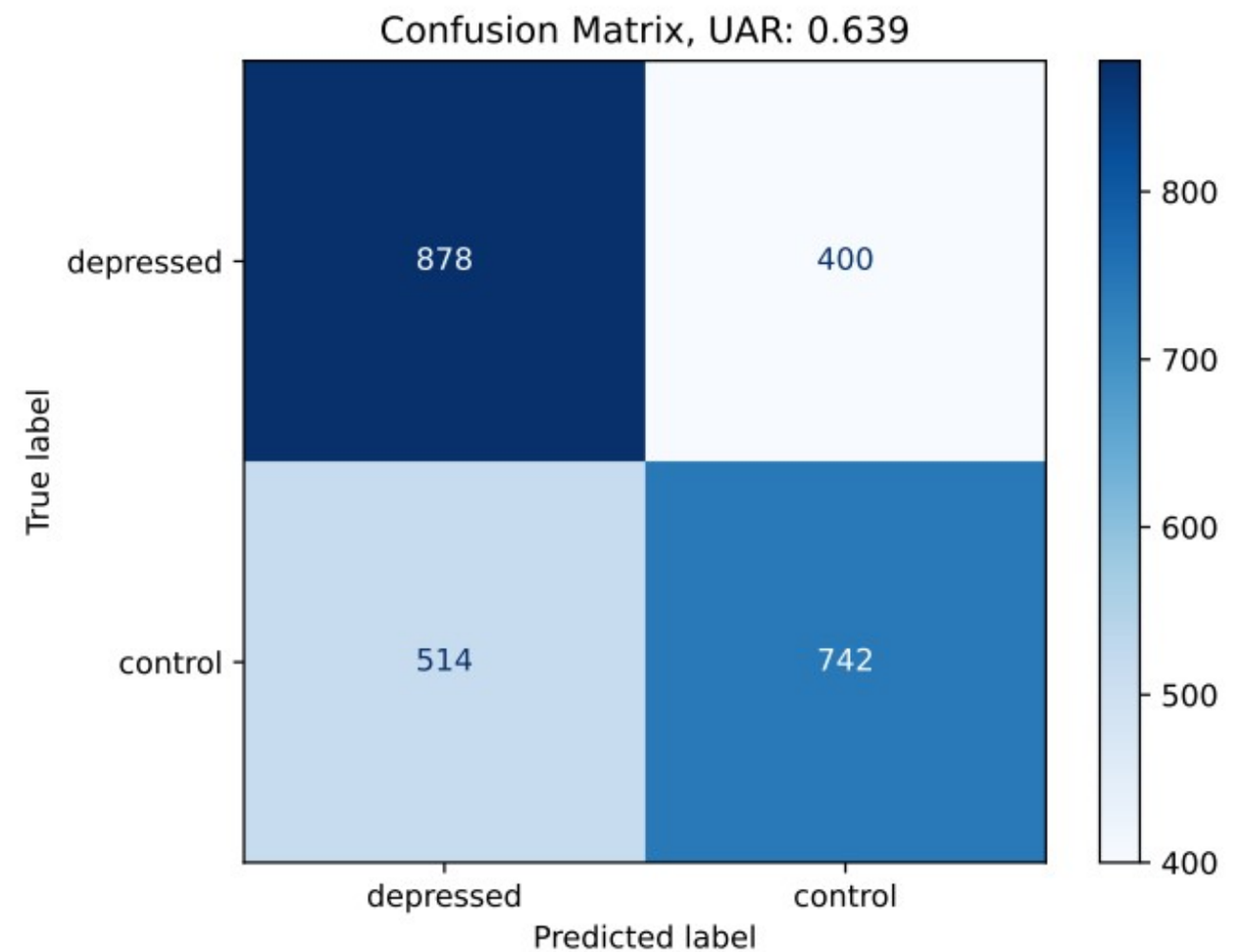
CHECK YOUR AUDIO DATA: NKULULEKO FOR BIAS DETECTION

Felix Burkhardt[1,2], *Bagus Tris Atmaja*[3], *Anna Derington*[1], *Florian Eyben*[1], *Björn Schuller*[1,4,5]

[1]audEERING GmbH, Germany
[2]Technical University of Berlin, Germany
[3]National Institute of Advanced Industrial Science and Technology, Japan
[4]Chair EIHW, Universität Augsburg, Germany
[5]GLAM, Imperial College London, UK

Burkhardt, F., Atmaja, B. T., Derington, A., & Eyben, F. (2024). Check Your Audio Data : Nkululeko for Bias Detection. Oriental COCOSDA, 1–6. https://doi.org/10.1109/O-COCOSDA64382.2024.10800580

Depression Detection



Feature	Androids	UASpeech
Duration	.067	Unclear-Mixed: **1.912**
Age	.128	Spastic-Athetoid: .169
Gender: χ^2 p	$\sim$ **0**	$\sim$ **0**
PESQ	.407	Unclear-Mixed: **2.014**
MOS	.459	Unclear-Athetoid: .385
SDR	.019	Unclear-Mixed: **1.169**
Arousal	**.666**	Spastic-Athetoid: .256
Valence: C.d	.415	Athetoid-Mixed: .289

Table 1. Results of the influence of various predicted influencing factors with respect to the target labels.

Laughter Classification

Performance-weighted Ensemble Learning for Speech Classification

Bagus Tris Atmaja
AIST
Tsukuba, Japan
b-atmaja@aist.go.jp

Akira Sasou
AIST
Tsukuba, Japan
a-sasou@aist.go.jp

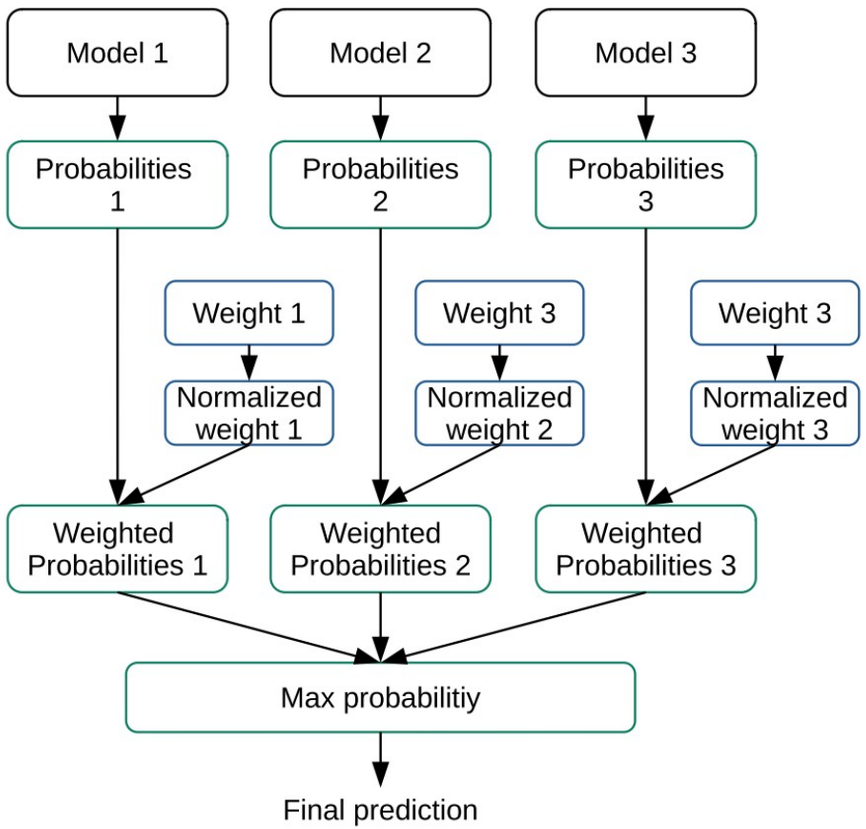
Felix Burkhardt
audEERING GmbH
Gilching, Germany
fburkhardt@audeering.com

***Abstract*—Ensemble learning is a useful technique to combine several models to improve classification performance. Previous**

the model usually takes the raw audio input to predict the class label.

Atmaja, B. T., Sasou, A., & Burkhardt, F. (2025). Performance-Weighted Ensemble Learning for Speech Classification. 2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 0044–0048. https://doi.org/10.1109/ICAIIC64266.2025.10920862

Laughter classification results



Task-dataset, features	Mean		Uncertainty		UA-weighted		WA-weighted	
	UA	WA	UA	WA	UA	WA	UA	WA
LC-Laughter os+praat	69.4	65.2	69.4	65.2	69.4	65.2	73.0	69.6

Pathological Voice Detection

Pathological Voice Detection From Sustained Vowels: Handcrafted vs. Self-supervised Learning

Bagus Tris Atmaja
AIST
Tsukuba, Japan
b-atmaja@aist.go.jp

Akira Sasou
AIST
Tsukuba, Japan
a-sasou@aist.go.jp

***Abstract*—Pathological voice detection aims to detect voice disorders from speech samples. With the recent development of**

of its datasets. Saarbrüecken Voice Database (SVD) [6], [7] perhaps is the first publicly available voice disorder dataset.

Atmaja, B. T., & Sasou, A. (2025). Pathological Voice Detection From Sustained Vowels : Handcrafted vs. Self-supervised Learning. 2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), 1–5. https://doi.org/10.1109/ICASSPW65056.2025.11011272

Pathological Speech Detection

- Voice disorder detection determines whether a voice sample contains pathology or not.
- The task is binary classification with F1-score and area under curve (AUC) were the most cited metrics
- Identification further classifies the type of voice disorder if detected as pathological, e.g., if a voice falls into one type of structural, neurogenic, functional, and psychogenic.

Performance of different vowel

TABLE I
PERFORMANCE OF DIFFERENT VOWELS (CLASSIFIER: XGB)

Vowels	UA	WA	F1	F1-ma	F1-we	AUC
			os			
a	**0.771**	0.773	0.817	0.760	0.777	**0.854**
i	0.609	0.700	0.799	0.606	0.662	0.609
u	0.754	**0.779**	**0.831**	**0.756**	**0.778**	0.836
aiu	0.706	0.742	0.806	0.710	0.739	0.819
			hubert			
a	0.717	0.753	0.814	0.722	0.749	0.717
i	**0.729**	**0.763**	**0.822**	**0.734**	**0.760**	**0.729**
u	0.655	0.695	0.770	0.658	0.691	0.656
aiu	0.575	0.647	0.751	0.574	0.626	0.575

Performance of different features

Features	Vowel	UA	WA	F1	F1-ma	F1-we
os	/a/	0.771	0.773	0.817	0.760	0.777
praat	/a/	**0.806**	**0.784**	0.815	**0.778**	**0.789**
w2v	/i/	0.732	0.758	0.815	0.733	0.757
hubert	/i/	0.729	0.763	**0.822**	0.734	0.760
wavlm	/u/	0.714	0.753	0.816	0.720	0.748
ft-w2v	/i/	0.721	0.758	0.819	0.727	0.754
ft-hubert	/i/	0.682	0.663	0.704	0.657	0.671
ft-wavlm	/i/	0.680	0.726	0.798	0.686	0.719

Performance of ensemble learning

Method	UA	WA	F1	F1-ma	F1-we	AUC
early fusion						
os+praat	0.780	0.789	0.833	0.774	0.791	0.872
w2v+h+w	0.731	0.753	0.808	0.730	0.753	0.809
late fusion						
os+praat	0.797	0.795	0.833	0.784	0.798	**0.876**
w2v /a/+/u/	0.752	0.789	0.844	0.761	0.785	0.831
os+praat+w2v	0.814	0.816	0.852	0.804	0.818	0.869
os+praat+h	**0.841**	**0.842**	**0.874**	**0.831**	**0.842**	0.867

Benchmark

F1-score

Method	F1 score
CAR-HMM [16]	0.7527
modidifed CPC [16]	0.7421
os /u/ (ours)	0.8306
hubert /i/ (ours)	0.8221
os+praat+hubert (ours)	**0.8739**

AUC

Method	Vowel/Phrases	Augmentation	AUC
Perturbation [1]	/a/	No	0.78
MFCC raw speech [1]	phrases	No	0.86
MFCC extracted vowel [1]	phrases	No	0.84
ComParE [3]	phrases	No	0.72
wav2vec2 + CNN [14]	phrases	No	0.87
Transformers [30]	phrases + /aiu/	Yes	0.91
os+praat + XGB (ours)	/a/	No	0.88

Dementia Prediction From Speech Signal Using Optimized Prosodic Features

Bagus Tris Atmaja, Sakriani Sakti

Presented at APSIPA-ASC 2025, Singapore

Why we need to detect dementia as early as possible:

- Early detection of dementia, cognitive decline, is vital for enhancing patient care and management
- Conventional clinical assessments and cognitive tests require expert clinicians and are time-consuming
- Developing rapid and reliable in-house detection methods is highly desirable
- We proposed optimized prosodic features (15 features), with a focus on pause-related features, for dementia classification

Proposed Acoustic Features

- pause_lognorm_mu: Location parameter (μ) of the log-normal distribution fitted to pause duration[1]
- pause_lognorm_sigma: Shape parameter (σ) of the log-normal distribution fitted to pause duration[1]
- pause_lognorm_ks_pvalue: P-value from Kolmogorov-Smirnov test for goodness of fit of the log-normal distribution[1]
- pause_mean_duration: Mean duration of pauses between speech segments [2]
- pause_std_duration: Standard deviation of pause duration[2]
- pause_cv: Coefficient of variation (CV) of pause duration (std/mean)
- Proportion of pause duration: Proportion of total pause duration relative to speaking time[2]

[1] P. Pastoriza-Domınguez, I. G. Torre, F. Di´eguez-Vide, et al., Speech pause distribution as an early marker for Alzheimer's disease, Jan. 2021

[2] R. Haulcy and J. Glass, "CLAC: A speech corpus of healthy english speakers," in Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 1, 2021, pp. 201–205,

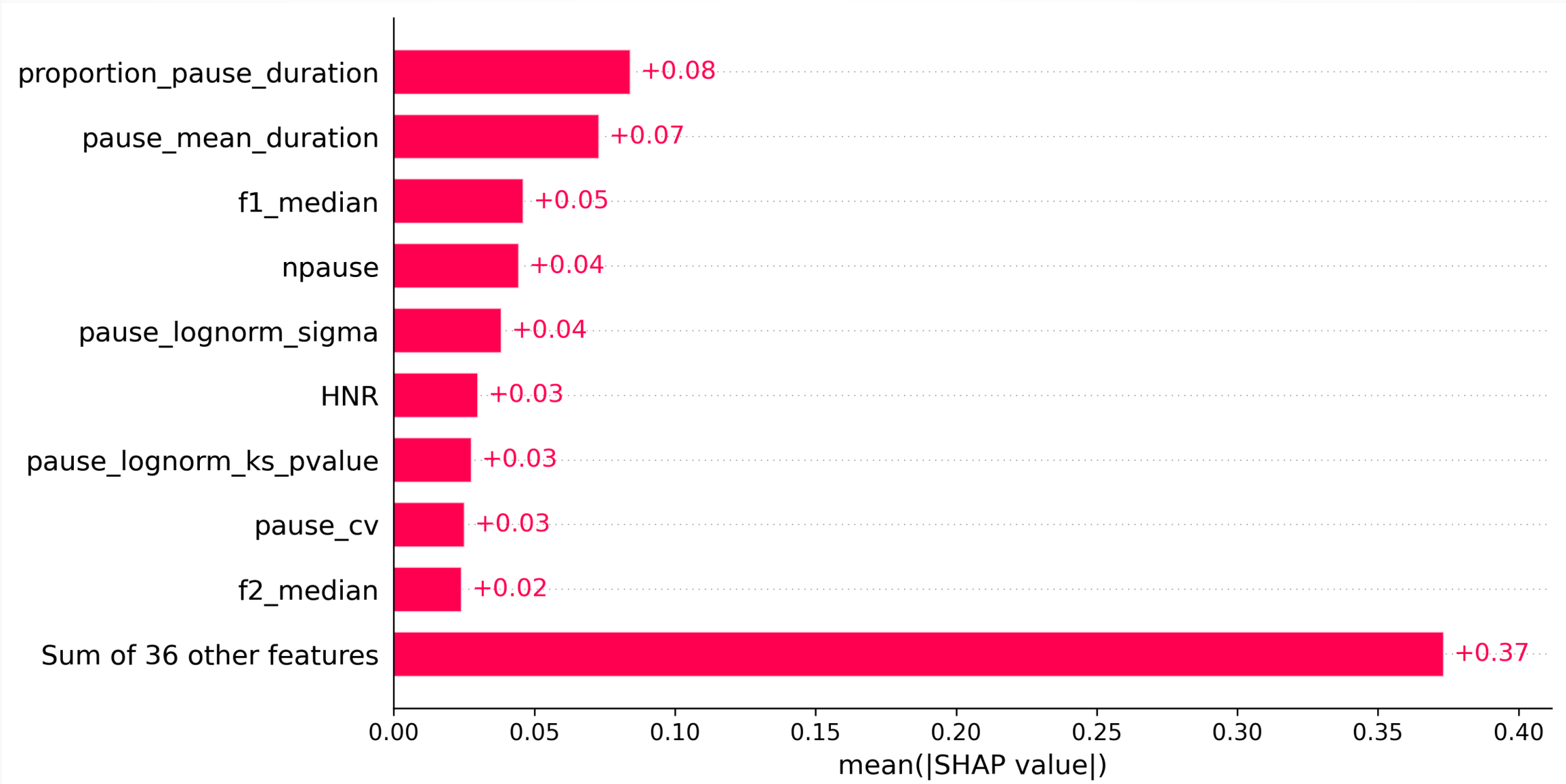
Datasets

- **DementiaNet**[1] contains public figures samples (WAV) with a confirmed dementia diagnosis and other public figures with no cognitive decline (control) over the age of eighty.
- **DementiaBank**[2] is a shared database of multi-media interactions for the study of communication in dementia. We analyzed subset of DementiaBank namely **Pitt Corpus** in MP3 format.
- **Mixed** of both datasets above

Model → XGB Feature importance → SHAP, XGB

[1] https://github.com/shreyasgite/dementianet
[2] https://talkbank.org/dementia/
[2] https://www.tensorflow.org/datasets/catalog/dementiabank

Results: DementiaNet



Results: DementiaBank & Mixed

DementiaBank

Feature	XGB Importance
f3_median	0.07530
pause_lognorm_ks_pvalue	0.00803
pause_lognorm_mu	0.00550
npause	0.00537
duration	0.00417
nsyll	0.00290
f1_mean	0.00278
localabsoluteJitter	0.00248
proportion_pause_duration	0.00236
ppq5Jitter	0.00115
meanF0Hz	0.00066
ASD_speakingtime_nsyll	0.00042
pF	0.00042
apq3Shimmer	0.00030
pause_mean_duration	0.00012

Mixed Dataset

Feature	XGB Importance
f1_median	0.02894
proportion_pause_duration	0.01306
f3_median	0.00915
pause_lognorm_ks_pvalue	0.00867
meanF0Hz	0.00697
ASD_speakingtime_nsyll	0.00605
f1_mean	0.00540
stdevF0Hz	0.00516
npause	0.00387
pF	0.00359
localdbShimmer	0.00343
f3_mean	0.00294
pause_lognorm_mu	0.00242
f4_mean	0.00198
HNR	0.00161

Results: 45 vs. 15 features

Dataset	UA	WA	F1-score	n_feat-model
			Baseline	
DementiaNet	0.746	0.771	0.820	45
DementiaBank	0.669	0.666	0.707	45
Mixed dataset	0.611	0.601	0.619	45
			Optimized	
DementiaNet	0.796	0.812	0.847	15-shap
DementiaBank	0.748	0.745	0.776	15-xgb
Mixed dataset	0.721	0.720	0.712	15-xgb

Benchmark:

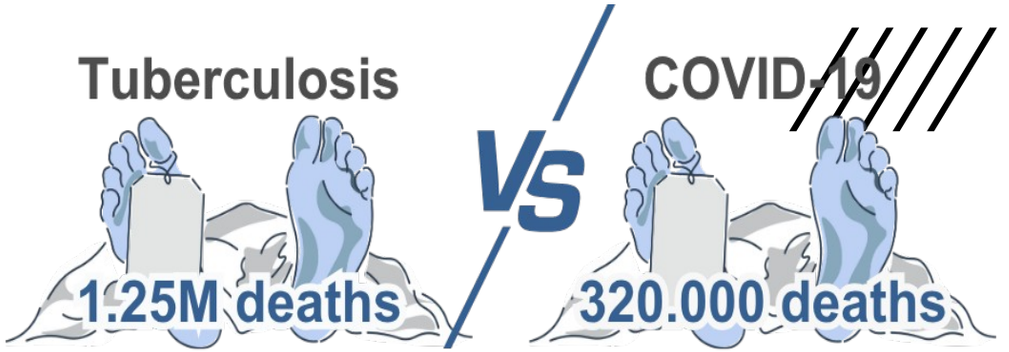
Method	n_feat	Dataset	UA	WA	F1-score
PRAAT [18]	7	Elderly	-	0.582	0.621
IS10 [19]	75	Pitt Corpus	-	0.731	-
MFCC++ [20]	44	Pitt Corpus	-	0.876	0.875
eGeMAPS [21]	88	ADReSSo	0.730	0.746	0.750
This study	15	Pitt Corpus	0.748	0.745	0.776
This study	15	DementiaNet	0.796	0.812	0.847

A COMPARISON OF SOLICITED AND LONGITUDINAL COUGH SOUNDS FOR TUBERCULOSIS DETECTION

APRIANTO DWI PRASETYO*, BAGUS TRIS ATMAJA§, DHANY ARIFIANTO* AND SAKRIANI SAKTI§

Presented at APSIPA-ASC 2025, Singapore

Introduction



- Tuberculosis (TB) remains one of the world's deadliest infectious diseases, causing ~1.25M deaths in 2023, surpassing COVID-19.
- Early and objective detection is essential to reduce global TB mortality

Research Gap

- Previous studies emphasize model performance but overlook how data quality impacts outcomes.
- Factors such as dataset balance, size, label reliability, and recording contamination remain underexplored.

Objective

- Investigate how data characteristics and recording conditions affect deep learning performance in TB detection.

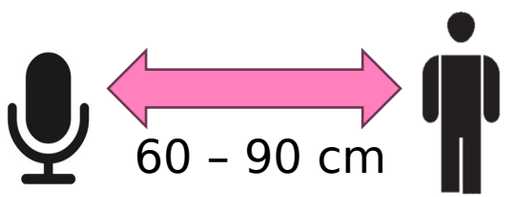
Dataset

- CODA TB DREAM Challenge dataset: 700K+ cough samples, 2,143 participants.
- Two distinct data types:
  - Solicited coughs: Collected under controlled, standardized conditions.
  - Longitudinal coughs: Collected passively in real-world settings..

This research systematically analyzes and compares solicited and longitudinal cough data, examining how each and their combination influence deep learning model robustness for real-world TB screening.

Solicited vs Longitudinal Coughs

Solicited Cough

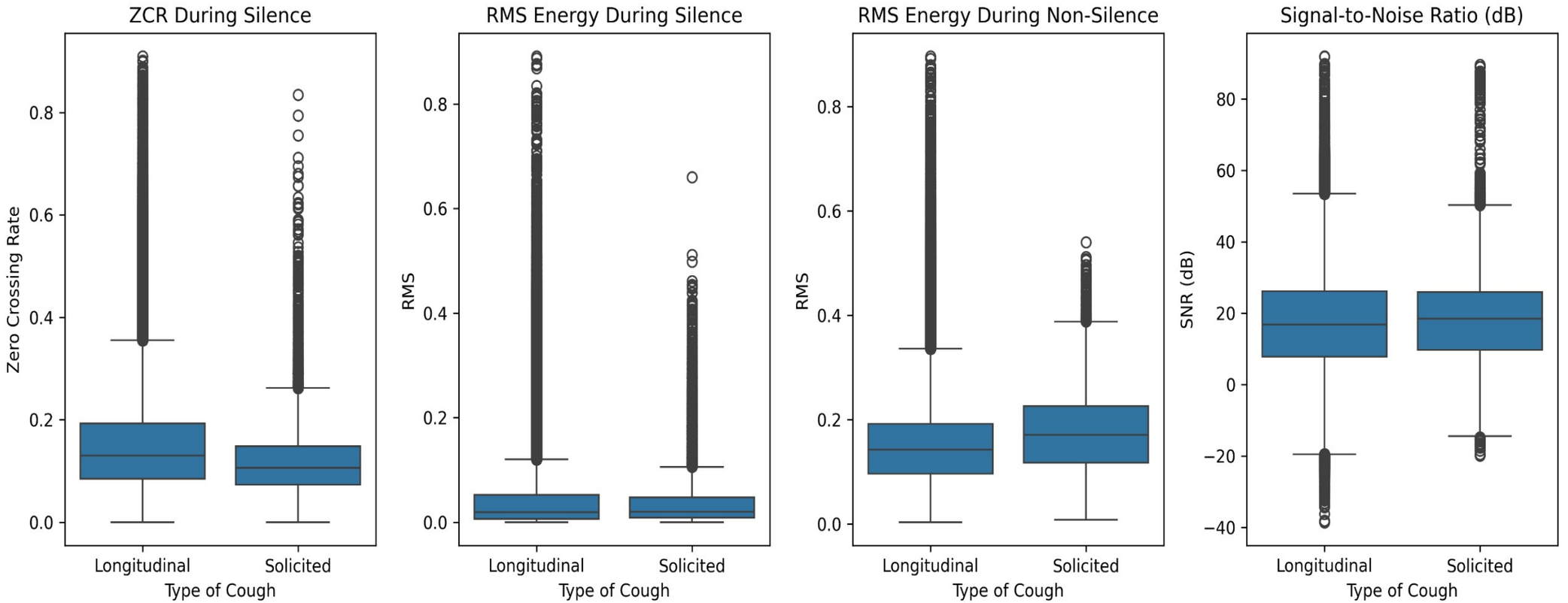


- Smartphones were positioned on tripods in rooms within the clinic in a dedicated room **without background noise**.
- Patient **Asked** to Cough
- Total 9,232 Coughs Data

Longitudinal Cough

- A subset of participants (n = 565) were also asked to carry a study phone for two weeks and collect longitudinal coughs sounds in an outpatient setting
- Contrarily to the solicited cough, longitudinal cough recording was **unsupervised** by study personnel and **was performed in community setting**
- Coughs are automatically detected and recorded by the system.
- Total 647,060 Coughs Data

Solicited vs Longitudinal Coughs



- ZCR & RMS: Longitudinal coughs show higher and more variable values, indicating persistent background noise and unstable environments. Solicited data remains lower and consistent, reflecting cleaner recordings.
- SNR: Slightly lower and more variable in longitudinal data; higher and stable in solicited.
- Insight: Longitudinal data is noisier but richer in diversity, while solicited data offers cleaner, more uniform signals.

Preprocessing and Feature Extraction

- Audio resampled from 44.1 kHz to 16 kHz.
- Normalization: Signal normalized to –1 to 1 using Min–Max normalization (dividing by max absolute value).
- Length Standardization: Each clip fixed to 1 second duration. Shorter clips extended using repeat padding.
- Feature Type: Mel spectrogram. Frame size: 64 ms, Hop length: 16 ms, Window: Hann function, 80 mel channels

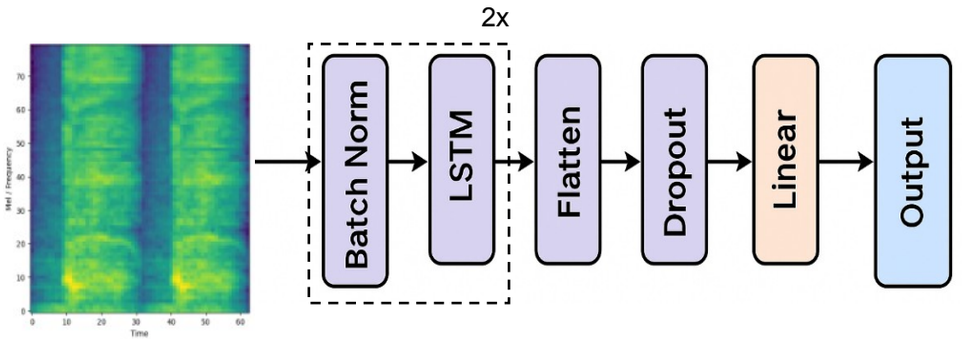


Figure 3: Architecture of LSTM Network used.

Training Configuration

- Dataset split: 90 % : 10 % (for solicited, longitudinal, and mixed data)
- Optimizer: AdamW
- Loss Function: Weighted Cross-Entropy
- Evaluation Metrics
  - Accuracy: Overall correct predictions (TP + TN / Total).
  - F1 Score: Harmonic mean of precision and recall
  - Sensitivity: Measures correct identification of positive TB cases. High sensitivity → fewer false negatives.
  - Specificity: Measures correct identification of non-TB cases. High specificity → fewer false positives.
  - ROC-AUC: Summarizes trade-off between sensitivity and specificity. AUC ≈ 1 = strong classifier; AUC ≈ 0.5 = random guessing.

Solicited vs. Longitudinal Cough on the Same Number of Samples

Data	Accuracy	F1	Sensitivity	Specificity	ROC-AUC
Solicited	**0.79**	**0.7**	**0.82**	**0.78**	**0.80**
Longitudinal	0.71	0.62	0.73	0.71	0.72

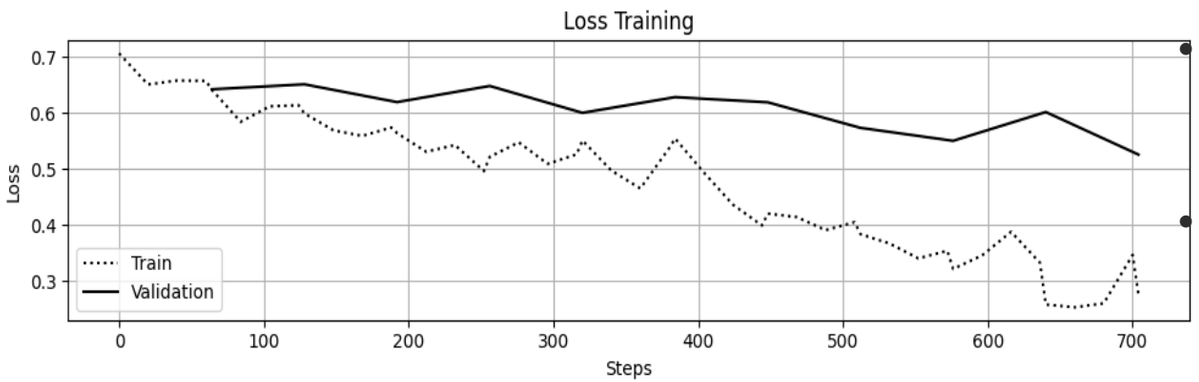


Figure 4: Training and validation losses for longitudinal model

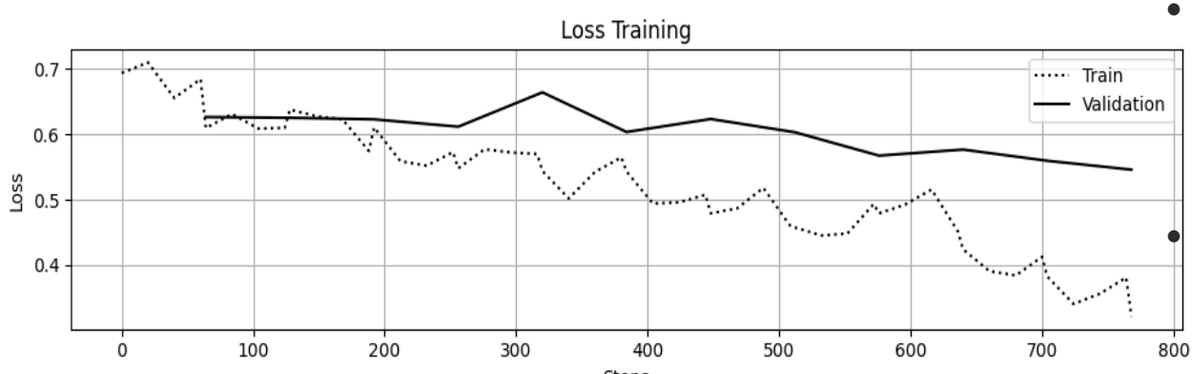


Figure 5: Training and validation losses for solicited model

Each longitudinal model was trained 5× with different random subsets; results are averaged for fairness.

Solicited data performs better due to controlled clinical recordings with consistent equipment, positioning, and low noise → cleaner, high-fidelity signals.

Longitudinal data because higher acoustic variability from noise, device differences, and inconsistent recording conditions → lower model performance.

Fig. 4 and Fig. 5 show that while training loss keeps decreasing, validation loss declines more slowly and eventually diverges, indicating overfitting of the model to the training data in these cases.

Solicited vs. Longitudinal Cough on Different Number of Samples

Data	Accuracy	F1	Sensitivity	Specificity	ROC-AUC
Solicited	0.79	0.7	0.82	0.78	0.80
Longitudinal	**0.91**	**0.93**	**0.91**	**0.92**	**0.91**

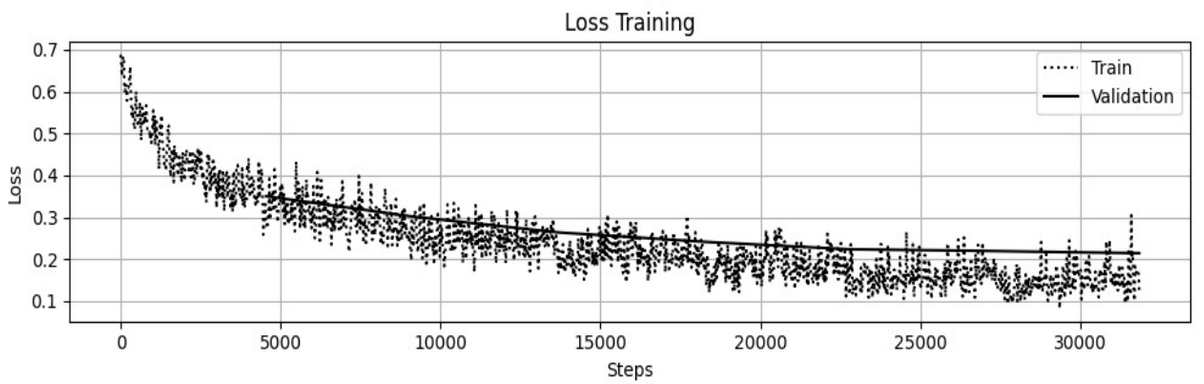


Figure 6: Training and validation losses for longitudinal model

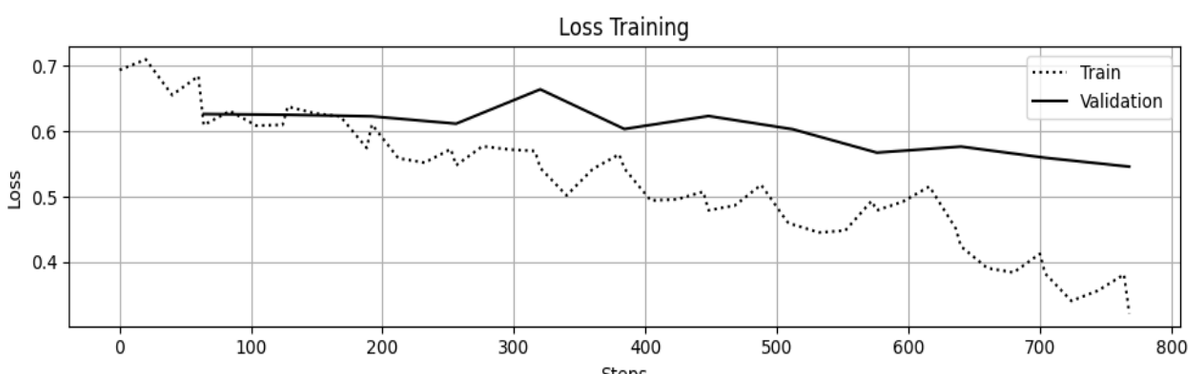


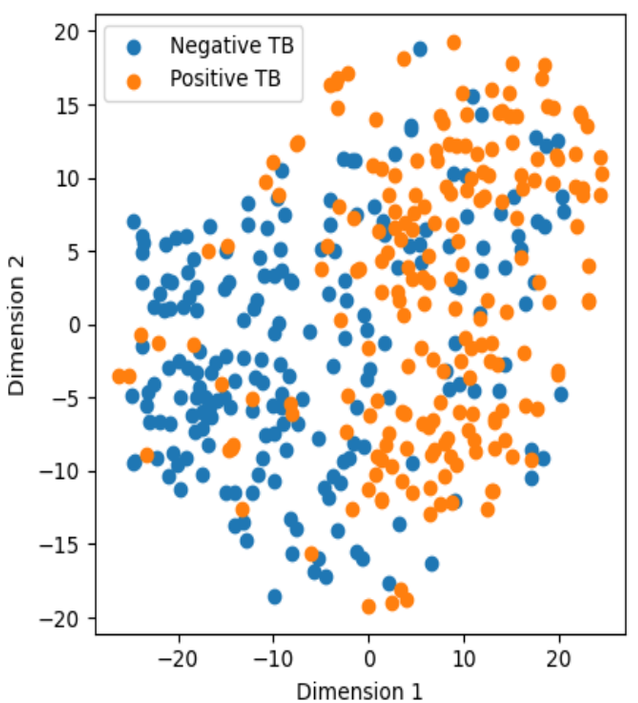
Figure 5: Training and validation losses for solicited model

- Training on the full longitudinal dataset (647,060 samples) improved performance by ~10% over solicited data → scale advantage.
- Despite higher noise and recording variability, this diversity enhances generalization and reduces overfitting.
- Solicited data is cleaner but lacks real-world variation.
- Figure 6: Training and validation losses decrease together until early stopping → strong generalization.

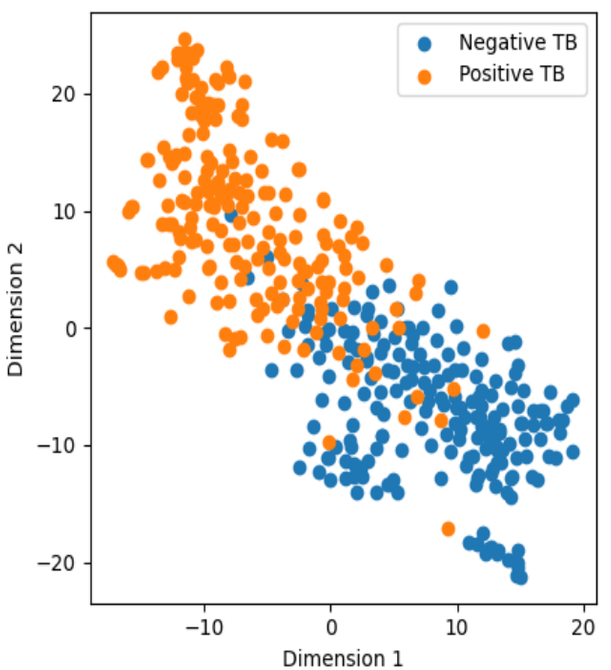
Combining Solicited and Longitudinal Cough

Data	Accuracy	F1	Sensitivity	Specificity	ROC-AUC
Solicited	0.79	0.80	0.70	0.82	0.78
Longitudinal	0.91	0.91	0.93	0.91	0.92
Combined	**0.91**	**0.91**	**0.93**	**0.92**	**0.91**

- Combining solicited and longitudinal data gave only marginal gains over longitudinal-only training.
- Minimal impact likely due to solicited data forming only ~1.5% of the total dataset.



(a) Solicited Model



(b) Longitudinal Model

- Solicited model: High class overlap; scattered points → weak separability and noisy feature learning.
- Longitudinal model: Clear, compact clusters → stronger class-specific embeddings and better discrimination.
- Improved clustering in the longitudinal model shows enhanced feature learning and better class representation overall.

Figure 7: TSNE visualization of embeddings for two different models.

TB Conclusion

- Supervised (solicited) data achieved higher accuracy with equal sample size, but large-scale longitudinal data (~647k samples) improved performance to **91%**, demonstrating the value of dataset scale.
- Merging both types offered no meaningful gain since solicited data accounted for only ~1.5% of total samples.
- Representation analysis (t-SNE)
  - Solicited model = High overlap among classes, weak separability, and noisier learned features.
  - Longitudinal model = Clear, compact clusters showing strong class-specific embeddings and superior discriminative learning.
- Large, diverse unsupervised data enhances both model generalization and feature representation.
- For Future work, Focus on standardized supervised data collection and exploration of advanced acoustic features and model architectures.

Take home messages

- Applying AI to healthcare and well-being is hard, but there is potential use of it
- Some models work, others didn’t
- Only certain cases (of diseases and well-being) could be predicted from voice, NOT all cases
- Robustness and generalization is main goal instead of performances only
- Large data matters (Effective!!)
- Next: Demo

Demo

- Speech Emotion Recognition (Polish Dataset)
- Laughter Classification
- Dementia Prediction