# A Comparison of Solicited and Longitudinal Cough Sounds for Tuberculosis Detection

Aprianto Dwi Prasetyo*, Bagus Tris Atmaja§, Dhany Arifianto* and Sakriani Sakti§
* Sepuluh Nopember Institute of Technology, Indonesia
E-mail: 6009231008@student.its.ac.id, arifianto@its.ac.id
§ Nara Institute of Science and Technology, Japan
E-mail: bagus.tris@naist.ac.jp, ssakti@is.naist.jp

***Abstract*—Tuberculosis (TB) caused an estimated 1.25 million deaths in 2023, exceeding the number affected by COVID-19. Existing diagnostic tools are costly and inaccessible in low-resource settings, leading to delays and ongoing transmission. A rapid and affordable screening method is urgently needed to identify and refer suspected TB cases for confirmation. One promising approach is deep learning using cough sounds, which offers a low-cost, scalable solution. However, building an effective deep learning model depends heavily on data, particularly the balance between data quality and quantity. Two types of datasets were evaluated: solicited (supervised recording) and longitudinal (unsupervised recording). Results show that supervised recordings achieved higher performance than unsupervised ones when using the same data size (79% vs. 71% of accuracy). However, increasing the amount of unsupervised data significantly improved performance, reaching 91% accuracy—highlighting the benefit of larger datasets. Interestingly, combining solicited and longitudinal data did not enhance performance, likely due to the small proportion of supervised data.**

## I. Introduction

Tuberculosis (TB) is a highly infectious disease and remains a major global health concern. In 2023, TB was responsible for an estimated 1.25 million deaths, making it likely the leading cause of death worldwide from a single infectious agent, surpassing COVID-19. In comparison, the number of deaths from COVID-19 officially reported to the World Health Organization (WHO) in 2023 was 320,000; it is currently unlikely that the total for 2024 will exceed this figure for TB [1]. Across the globe, the net reduction in the TB incidence rate between 2015 and 2023 was 8.3%, far from the WHO End TB Strategy milestone of a 50% reduction by 2025. The majority of TB cases occur in 30 high-burden countries, which collectively accounted for 87% of global TB cases in 2023. Some of these countries are classified as developing nations, where complex socio-economic challenges hinder access to public healthcare. These barriers make it difficult to detect, treat, and control TB effectively, often leading to inadequate care and continued transmission of the disease [2].

Existing TB diagnostic tools are often expensive and demand specialized expertise and infrastructure. In many developing countries, primary healthcare facilities lack the financial resources to implement these advanced methods for rapid screening. As a result, they must rely on slower, more basic, but affordable techniques such as sputum analysis [1]. This leads to prolonged diagnostic delays, increasing the risk of TB spreading quickly within communities. Therefore, there is a pressing need for an affordable and rapid screening method that can help identify individuals who are likely to have TB and refer them for more comprehensive confirmatory testing.

Using human voice as a biomarker for disease detection has been widely studied. Arifianto et al. [3] proposed a time-varying autoregressive method to analyze speech disorders. Sasou and Chen [4] combined glottal inverse filtering and self-supervised learning for detecting pathological voice. Atmaja and Sasou [5] improved the performance of F1-score from the previous research on the same test dataset using ensemble learning. For TB, various studies have shown the potential of using cough sounds to detect respiratory diseases (e.g., [6], [7]). These two studies used deep learning as the main model. In building a good deep learning model, several factors affect model quality, mainly four aspects of data quality: equilibrium of the data set, size of the data set, quality of the labels, and contamination of the data set [8].

The type of recording used for TB cough analysis may have an impact on the performance. As for the datasets to be used in this study, two types of recordings are included. The first data is solicited coughs, which were obtained in a controlled environment following a standardized procedure, although they account for only about 1.5% of the total dataset. The second type of data is longitudinal coughs, which were recorded in an unsupervised manner, where patients were asked to carry a study phone for two weeks to passively collect cough sounds in an outpatient setting [9]. Consequently, the quality of longitudinal coughs is generally lower compared to that of solicited coughs.

The effect of longitudinal data have ben studied previously in other domains, such as driving behavior and charisma prediction [10], [11]. However, the effect of longitudinal data on cough-based TB prediction is not available yet. This study proposes to deeply analyze longitudinal data and compare the results with solicited data, including the combination of both.

## II. Methods

### A. Datasets

The dataset used in this study was the CODA TB DREAM Challenge Dataset [12]. This open-access dataset was obtained via the Synapse platform, which permits access exclusively to authenticated users following a review by the data access

979-8-3315-7206-8/25/$31.00 

team. Only researchers with confirmed profiles—which include ORCID connection, identity verification, and Synapse Pledge acceptance—are granted access to regulated data by Synapse. This approach promotes both transparency and the protection of patient privacy and data security. To safeguard participant information, the dataset has been de-identified, with all demographic and clinical details anonymized. The dataset contain two data types based on how the data was obtained, the first one was solicited cough, which was obtained in a controlled environment and procedure. The other one was longitudinal coughs were recorded in an unsupervised approach, where the patient was asked to carry a study phone for two weeks and collect longitudinal cough sounds in an outpatient setting using the Hyfe algorithm [13] to detect and segment the cough. Using this algorithm, the probability of explosive sounds being human coughs is calculated and reported as a 0 to 1 probability score [9]. From this probability criteria, we select only cough data with probability score above 0.95 to ensure that no sound other was cough in the datasets. Selecting coughs based on their probability have been found beneficial in other cough-based disease detection [14]. In the end, for solicited cough samples, the number of cough samples was 9,232, and for longitudinal cough samples, the number of cough samples was 647,060. The cough sound class contained in this database was TB Positive and TB Negative based on the microbiological reference standard, as mentioned in the original datasets.

### *B. Acoustic Features and Preprocessing*

Before acoustic feature extraction, cough audio recordings undergo a series of preprocessing steps. Initially, the audio sampling is resampled to 16 kHz from 44.1 kHz. Then, we normalized the audio signal range to -1 and 1 with Min-Max Normalization, which is done by dividing each element by the maximum absolute value. After that, to ensure uniformity in input length across all audio samples, each clip was standardized to a fixed duration of 1 second. In cases where the original audio was shorter than the target duration, we applied repeat padding, wherein the audio signal is looped until the required length is achieved.

Mel spectrograms were used as the main acoustic feature in this research, for Mel spectrograms are computed by applying the short-time Fourier transform (STFT) with a frame size of 64 ms, a frame hop of 16 ms, and utilizing a Hann window function. For the mel scale, an 80-channel mel filterbank spanning 125 Hz to 7.6 kHz was used. Examples of the extracted Mel-spectrogram for Non-TB and TB are shown in Fig. 2.

### *C. Model and Training Setup*

For all experiments, the same model was used, a simple LSTM model [15] with some modification, the modified architecture illustrated in Fig. 1, The network consists of a two-layer long- and short-term memory (LSTM) network. The input is first normalized using batch normalization, then goes to the first LSTM Layer, and projects it to a hidden dimension of 1024, then batch normalized again and passed into a second

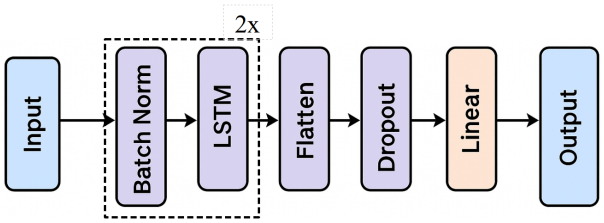



Fig. 1: Architecture of LSTM Network

LSTM layer, also with a hidden dimension of 1024. The output corresponding to the last LSTM layer is extracted and flattened. After that, the dropout layer with a rate of 0.1 was applied. Finally, the hidden states are passed through a fully connected layer that maps the 1024 hidden states to 2 outputs, representing Positive TB or Negative TB.

The training setup was the same for all experiments. The data for each evaluation (solicited, longitudinal, and mixed types) was divided into 90% data training and 10% data testing. The model was trained on a single GPU with 128 batch size, AdamW optimizer [16] was used with $\beta_1 = 0.8$, $\beta_2 = 0.999$, $\epsilon = 10^{-6}$ and learning rate $5e^{-4}$ exponentially decaying by $0.999875$ every epoch. For the loss function, the cross-entropy loss with calculated class weights is used, which is to address class imbalance by giving higher importance to minority classes. An early stopping mechanism with a patience value greater than 5 is applied to prevent overfitting. The model checkpoint with the lowest validation loss is selected as the final model and used for evaluation on the test dataset.

### *D. Evaluation Metric*

Multiple evaluation metrics were used to provide different insights into the classifier's performance, i.e., accuracy, ROC AUC, F1 score, sensitivity, and specificity.

**Accuracy** measures how often the classifier is correct overall. It is calculated by taking the total number of correct predictions (both true positives and true negatives) and dividing it by the total number of predictions. In other words, accuracy tells us how often the model gets things right. However, accuracy can be misleading, especially when the data is imbalanced (for example, when there are many more negative cases than positive ones). A model might get high accuracy just by always predicting the majority class. That's where other metrics like the F1-score come in.

**F1 Score** focuses on the balance between precision (how many of the predicted positives are actually correct) and recall (how many of the actual positives were correctly identified). It is especially useful when we care more about the *positive class* and want a balance between catching positives and avoiding false alarms.

**Sensitivity** is a performance measure used in binary classification, especially helpful when dealing with imbalanced data. It tells us how well the system correctly identifies positive cases (true positives). A higher sensitivity means fewer false negatives; in other words, the classifier is better at catching

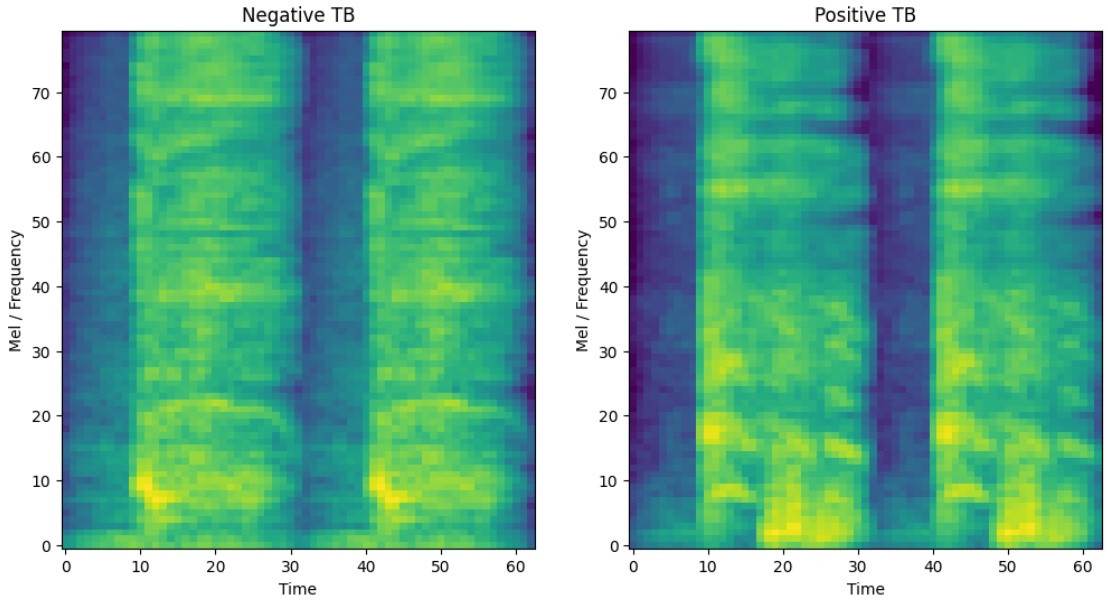



Fig. 2: Mel-Spectrogram Images of Negative TB and Positive TB Cough

the positive cases. This is very important in medical systems, where missing a sick patient (a false negative) could have serious consequences.

**Specificity**, on the other hand, measures how well the system correctly identifies negative cases (true negatives). It's the proportion of actual negative samples that are correctly predicted. A higher specificity means fewer false positives, in the context of a medical system, fewer people being incorrectly told they might be sick and needing further (often unnecessary) testing.

**The ROC curve** is a graph that shows how well a classifier works. It plots the actual positive rate (TPR, same as sensitivity) on the vertical axis, and the false positive rate (FPR, same as specificity) on the horizontal axis. The ROC curve helps us see the balance between detecting real positives (sensitivity) and avoiding false alarms (specificity). If the curve goes through the top-left corner, the classifier is perfect. If it lies along the $45\,\mathrm{deg}$ diagonal, it means the classifier is guessing randomly. Although ROC curves are helpful, they can be hard to use when comparing models that perform similarly. That's why we often use the area under the ROC curve (AUC) as a single number to compare models. AUC ranges from 0 to 1, higher values mean better performance, and a value close to 0.5 means the model is no better than random guessing.

## III. Results and Discussion

### *A. Solicited vs. Longitudinal Cough Sounds on the Same Number of Samples*

In this scenario, longitudinal cough sounds were reduced to achieve a sample size the same as solicited cough. We used random sampling to sample the longitudinal cough as many as the solicited cough. To report the average result, we ran the experiment 5 times for longitudinal training. Hence, the result for longitudinal data in Table I was an average of 5 times longitudinal training with a different sample set.

TABLE I: Result for Solicited vs. Longitudinal Cough Sounds on the Same Number of Samples

| Data | Accuracy | ROC AUC | F1 | Sensitivity | Specificity |
|---|---|---|---|---|---|
| Solicited | 0.79 | 0.80 | 0.70 | 0.82 | 0.78 |
| Longitudinal | 0.71 | 0.72 | 0.62 | 0.73 | 0.71 |

From Table I, when using the same data (9,232 samples), solicited cough recordings consistently outperform longitudinal data across all evaluated metrics. This performance difference is likely due to the controlled conditions under which solicited recordings were collected. These recordings were obtained with the supervision of study personnel, ensuring consistent and high-quality audio capture [9]. In contrast, longitudinal recordings are more susceptible to variability in recording quality. Factors such as inconsistent microphone placement, environmental background noise, and the potential presence of other individuals coughing near the participants can contribute to reduced audio quality [17] and, consequently, diminished model performance.

TABLE II: Result for Solicited vs. Longitudinal Cough Sounds on Different Number of Samples

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

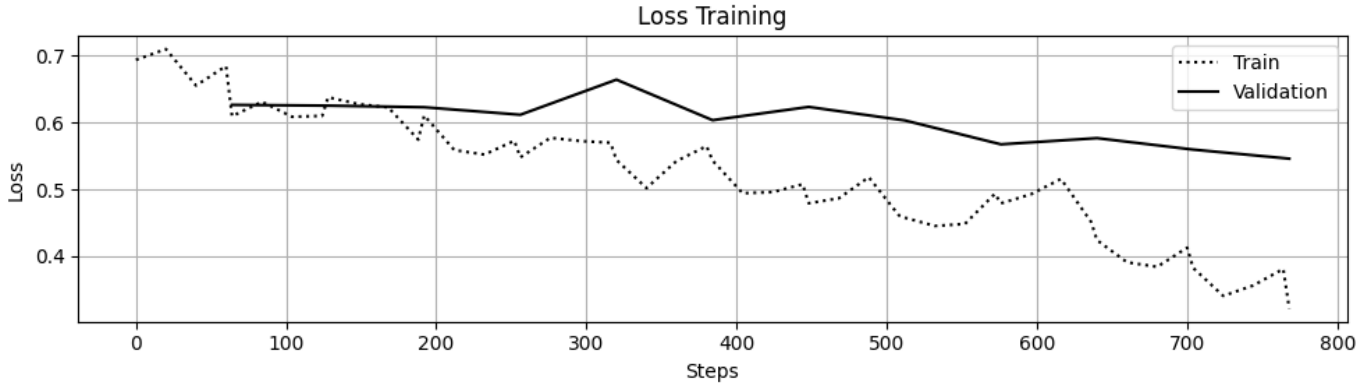



Fig. 3: Training and validation losses for longitudinal model using the same number of samples with solicited data

### *B. Solicited vs. Longitudinal Cough Sounds on Different Number of Samples*

In this scenario, longitudinal cough and solicited cough were trained on different numbers of samples, with each of them using full datasets for each partition. After training, the result was placed in Table II. Using a full longitudinal dataset (647,060 samples) leads to significantly improved performance, showing a gain of around 10% compared to solicited data alone. As previously mentioned, the longitudinal cough dataset exhibited significant variability, including differences in the positioning of the recording device relative to participants, background noise, and the presence of other individuals coughing nearby [9]. This variability contributes to the dataset's richness, which is beneficial for deep learning models, as they typically perform better with larger and more diverse data [18]. Furthermore, the variability in longitudinal data might enhance the model's ability to generalize and reduce overfitting. While solicited data offers higher quality due to controlled recording conditions, the scale and variability of longitudinal data become valuable assets when leveraging deep learning techniques, this is illustrated in Fig. 6 compared to Fig. 3 and Fig. 4. In Fig. 6, both the training and validation losses decrease together until the early stopping stops the training, indicating good generalization. In contrast, Fig. 3 and Fig. 4 show a different trend: while the training loss continues to decrease, the validation loss declines at a slower rate and eventually diverges, resulting in a larger gap between the two. This suggests that the model is overfitting to the training data in these cases. Furthermore, if longitudinal data from Table I was compared, the performance was increased up to 20%.

To better understand how well each model represents different classes, t-SNE was used to visualize the embeddings [19]. A random sample of 200 data points from each class was selected and plotted in a scatter plot. The embeddings generated by the solicited model (See Fig. 5a) did not show a clear separation between classes; points from different classes were scattered and overlapped, making it difficult to distinguish one class from another. In contrast, the embeddings produced by the longitudinal model (See Fig. 5b) showed much clearer separation, with data points forming distinct clusters based on class. This indicates that the longitudinal model captures more meaningful and discriminative features, allowing it to represent different classes more effectively than the solicited model.

TABLE III: Result for combining solicited and longitudinal cough sounds

| Data | Accuracy | ROC AUC | F1 | Sensitivity | Specificity |
|---|---|---|---|---|---|
| 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 |

The finding on using less qualified (unsupervised recording) but more quantified data that leads to higher performance than qualified data also supports the unreasonable effectiveness of data [20]: *"with very large sources, the data holds the details"*. As we can see in Tables II and III, the quantity of data is more important than the quality of data for cough-based TB detection.

### *C. Combining Solicited and Longitudinal Cough Sounds*

In this last scenario, we combine solicited and longitudinal data to train the model, which could obtain higher results than the previous longitudinal data training. The result was placed in Table III. From that Table, it is shown that combining solicited and longitudinal data does not result in significant performance improvements in most of the metrics. The only improvement is in the sensitivity score, in which the combined data attained 0.01% higher than the longitudinal data. In the specificity score, the combined data attained a lower score than longitudinal data, while in the other three metrics, the scores are the same. This is likely because solicited data represents only about 1.5% of the total dataset, making its influence minimal in the combined model training.

## IV. Conclusions

In this paper, we examined structured experiments on the detection of tuberculosis disease via cough sounds. Two dif-

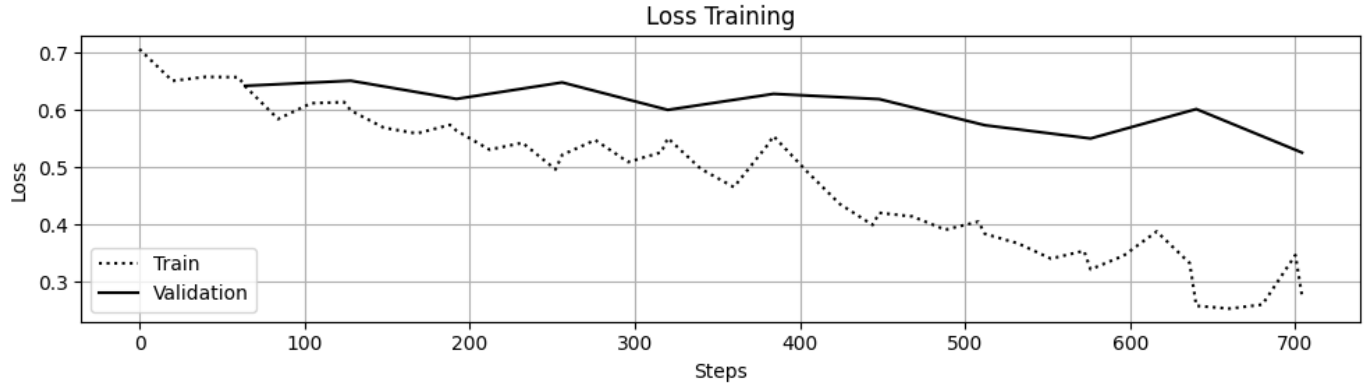
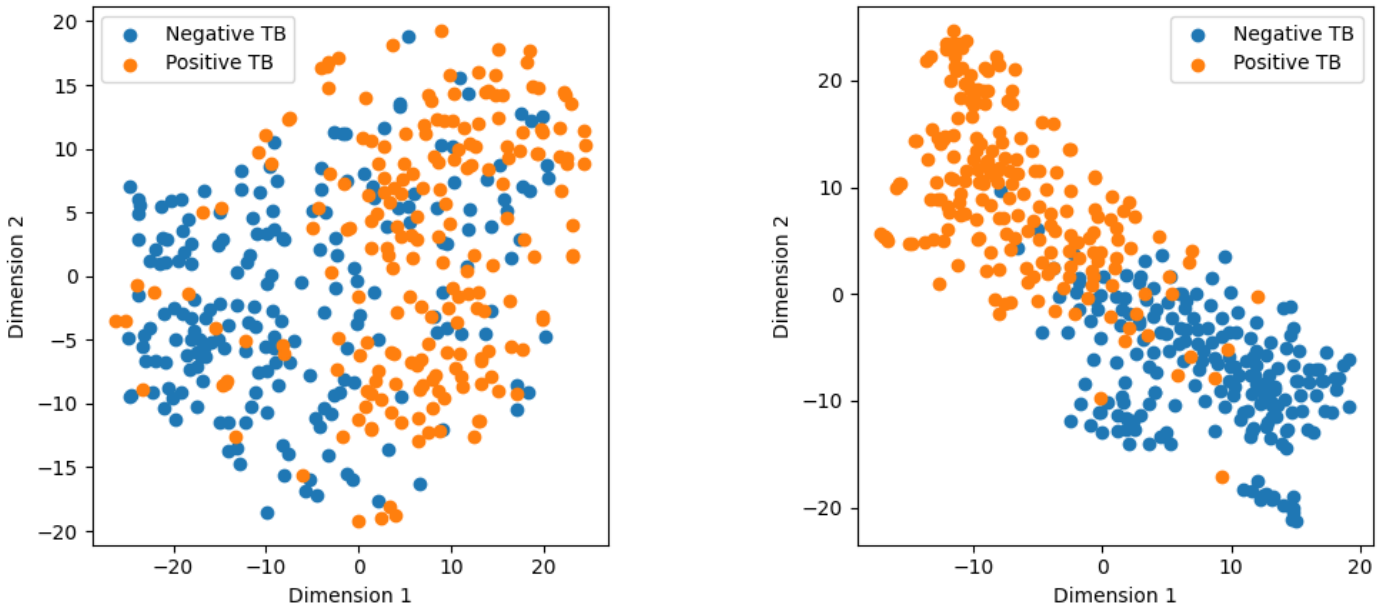



Fig. 4: Training and validation losses for solicited model




(a) Solicited Model

(b) Longitudinal Model

Fig. 5: TSNE visualization of embeddings for two different models

ferent types of data are evaluated, i.e., solicited (supervised recording) and longitudinal (unsupervised recording) data. Several insights could be gathered based on the findings. First, supervised recording obtained higher performance than unsupervised recording with the same amount of data. Second, when the number of unsupervised data is increased, the performance improves significantly (accuracy 91%), highlighting the contribution of bigger data. Third, the combination of solicited and longitudinal data did not improve the performance, probably due to the small number of solicited data. Future research could be directed toward standardizing supervised data collection, as well as evaluating other acoustic features and architectures.


## ACKNOWLEDGMENT

This paper is partly based on results obtained from projects JST Nexus 2025 and JSPS KAKENHI Grant Number 24K0296. The datasets used for the analyses described were contributed by Dr. Adithya Cattamanchi at UCSF and Dr. Simon Grandjean Lapierre at University of Montreal and were generated in collaboration with researchers at Stellenbosch University (PI Grant Theron), Walimu (PIs William Worodria and Alfred Andama); De La Salle Medical and Health Sciences Institute (PI Charles Yu), Vietnam National Tuberculosis Program (PI Nguyen Viet Nhung), Christian Medical College (PI DJ Christopher), Centre Infectiologie Charles Merieux Madagascar (PIs Mihaja Raberahona & Rivonirina Rakotoarivelo), and Ifakara Health Institute (PIs Issa Lyimo & Omar Lweno) with funding from the U.S. National Institutes of Health (U01 AI152087), The Patrick J. McGovern Foundation and Global Health Labs.

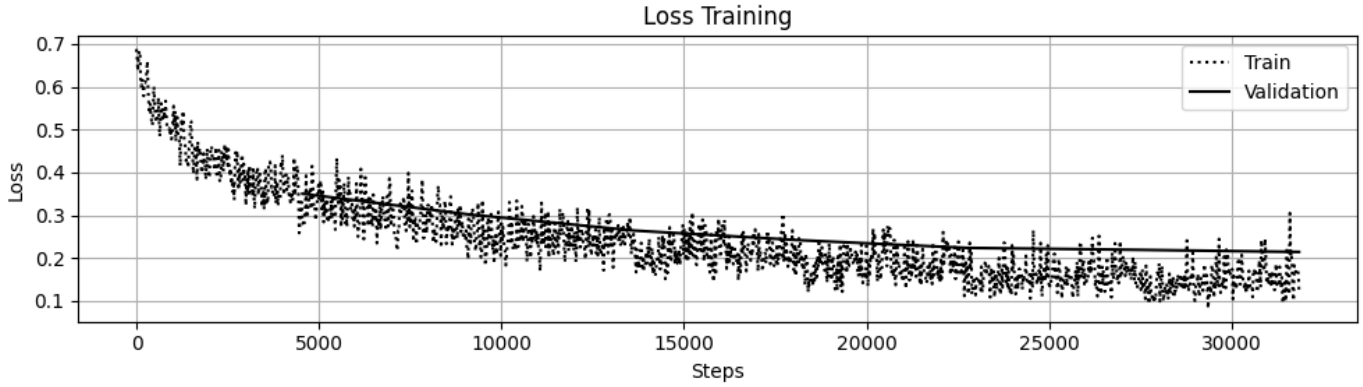



Fig. 6: Training and validation losses for longitudinal model using the full number of samples

## References


[1] World Health Organization, "Global tuberculosis report 2024," World Health Organization, Technical Report, 2024, Accessed: 2025-06-12. [Online]. Available: https://www.who.int/publications/i/item/9789240101531.

[2] N. Foster, A. Vassall, S. Cleary, L. Cunnama, G. Churchyard, and E. Sinanovic, "The economic burden of tb diagnosis and treatment in south africa," *Social Science & Medicine*, vol. 130, pp. 42–50, 2015.

[3] D. Arifianto, H. Setijono, and Sekartedjo, "Speech disorder analysis using time-varying autoregressive," *Midwest Symp. Circuits Syst.*, vol. 3, no. 2, pp. 191–194, 2004, ISSN: 15483746.

[4] A. Sasou and Y. Chen, "Comparison of GIF- and SSL-based Features in Pathological-voice Detection," in *INTERSPEECH 2023*, ISCA: ISCA, Aug. 2023, pp. 2893–2897.

[5] B. T. Atmaja and A. Sasou, "Pathological Voice Detection From Sustained Vowels : Handcrafted vs. Self-supervised Learning," *2025 IEEE Int. Conf. Acoust. Speech, Signal Process. Work.*, 2025.

[6] G. Frost, G. Theron, and T. Niesler, "TB or not TB? Acoustic cough analysis for tuberculosis classification," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2022-Septe, no. September, pp. 2448–2452, 2022, ISSN: 19909772. eprint: 2209.00934.

[7] W. Xu, H. Yuan, X. Lou, Y. Chen, and F. Liu, "DMR-Net Based Tuberculosis Screening With Cough Sound," *IEEE Access*, vol. 12, no. January, pp. 3960–3968, 2024, ISSN: 21693536.

[8] T. He, S. Yu, Z. Wang, J. Li, and Z. Chen, *From data quality to model quality: An exploratory study on deep learning*, 2019. arXiv: 1906.11882 [cs.LG].

[9] S. Huddart, V. Yadav, S. K. Sieberts, *et al.*, "A dataset of solicited cough sound for tuberculosis triage testing," *Scientific Data*, vol. 11, no. 1, p. 1149, 2024.

[10] W. Wang, C. Liu, and D. Zhao, "How much data are enough? A statistical approach with case study on longitudinal driving behavior," *IEEE Trans. Intell. Veh.*, vol. 2, no. 2, pp. 85–98, 2017, ISSN: 23798858.

[11] A. Hernández, C. Araya, J. García, and V. González, "Leader charisma and affective team climate: the moderating role of the leader's influence and interaction.," *Psicothema*, vol. 21, no. 4, pp. 515–20, 2009.

[12] S. Huddart, V. Yadav, S. K. Sieberts, *et al.*, "Solicited cough sound analysis for tuberculosis triage testing: The coda tb dream challenge dataset," *MedRxiv*, 2024.

[13] C. Chaccour, I. Sánchez-Olivieri, S. Siegel, *et al.*, "Validation of the Hyfe cough monitoring system: a multicenter clinical study," pp. 1–11, 2025.

[14] B. T. Atmaja, Zanjabila, Suyanto, W. A. Asmoro, and A. Sasou, "Cross-dataset COVID-19 transfer learning with data augmentation," *Int. J. Inf. Technol.*, p. d, Feb. 2025, ISSN: 2511-2104.

[15] A. Hassan, I. Shahin, and M. B. Alsabek, "Covid-19 detection system using recurrent neural networks," in *2020 International conference on communications, computing, cybersecurity, and informatics (CCCI)*, IEEE, 2020, pp. 1–5.

[16] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[17] P. Bottalico, J. Codino, L. C. Cantor-Cutiva, *et al.*, "Reproducibility of voice parameters: The effect of room acoustics and microphones," *Journal of Voice*, vol. 34, no. 3, pp. 320–334, 2020, ISSN: 0892-1997.

[18] Y. Ba, M. V. Mancenido, and R. Pan, *How does data diversity shape the weight landscape of neural networks?* 2024. arXiv: 2410.14602 [cs.LG].

[19] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *J. Mach. Learn. Res.*, vol. 164, pp. 2579–2605, 2008, ISSN: 02624079.

[20] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intell. Syst.*, vol. 24, no. 2, pp. 8–12, 2009, ISSN: 15411672.