Weakly Supervised Noisy Label Cleaning for Cough-Based TB Screening

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

***Abstract*—Noisy labels in cough-based TB screening datasets arise from the mismatch between microbiological TB diagnosis and cough sound characteristics. TB-positive labels are assigned by sputum-based tests confirming *Mycobacterium tuberculosis*, but a substantial fraction of these patients, particularly subclinical cases and those under treatment, produce coughs that are difficult to distinguish from non-TB coughs. As a result, some TB-labeled samples can confuse the classifier during training, reducing model performance. We propose a weakly supervised, training-free label cleaning pipeline that separates TB-labeled samples into *Acoustic-TB* and *Label-only-TB* subsets without requiring prior knowledge of noise rates or reliable ground truth. Evaluated on three cough datasets (CODA, CIDRZ, and TBScreen) under stratified group cross-validation, label cleaning consistently improves AUROC across datasets, with gains of ∼20–22% (CODA: 64% to 86%; CIDRZ: 66% to 88%; TBScreen: 69% to 89%), outperforming random-removal and Confident Learning baselines, and generalizing across two frozen audio encoders (Qwen3-Omni and HeAR). These results suggest that reducing label noise can substantially improve model performance.**

***Index Terms*—tuberculosis screening, cough sound analysis, noisy label detection, weakly supervised learning, audio representation learning, label cleaning**

I. Introduction

Tuberculosis (TB) remains one of the leading causes of death from infectious disease globally, with 10.6 million new cases and 1.3 million deaths reported in 2022 [1], with persistent high-burden clusters in sub-Saharan Africa and Southeast Asia, including Indonesia [2]. Even confirmatory diagnostics such as Xpert MTB/RIF remain costly to scale at the primary care level in high-burden settings (e.g., a per-test unit cost of approximately USD 27 in West Java, Indonesia) [3], leaving a critical need for affordable, scalable screening tools. Cough-based AI screening on smartphones has emerged as a promising low-cost alternative [4], with systems such as HeAR [5], TBscreen [6], and the CODA TB DREAM Challenge [7]. However, reported AUC values range around 60–80% [7], [8], [9], failing to meet the WHO screening threshold of $\geq$90% sensitivity at $\geq$70% specificity.

We identify a key contributor: *label ambiguity* arising from the mismatch between microbiological TB diagnosis and cough sound characteristics. TB labels are assigned by sputum culture or Xpert MTB/RIF, confirming *Mycobacterium tuberculosis* in sputum rather than whether the cough contains distinctive TB-related sound patterns. Two specific sources compound this ambiguity: (1) prevalence-survey analyses indicate that approximately 50% of microbiologically confirmed TB cases are subclinical [10], producing no clearly distinguishable cough characteristics; and (2) patients undergoing treatment exhibit dramatically reduced cough frequency (median cough episodes per hour fell from 2.3 pretreatment to 0.18 by end of treatment) [11], yet retain TB labels. As a consequence, some TB-positive patients produce coughs that sound similar to non-TB coughs despite carrying confirmed positive labels.

These acoustically ambiguous TB samples can act as noisy labels during classifier training. Label noise consistently degrades classifier performance, with reported accuracy drops of 5–13% at 20–50% noise levels in medical imaging, dermatological diagnosis, and remote sensing [12], [13], [14], and causes deep networks to memorize incorrect label associations rather than genuine discriminative patterns [15]. We therefore hypothesize that label noise is a major contributor to the low AUC reported for existing TB cough classifiers.

To address this, we propose a weakly supervised, training-free label cleaning pipeline that separates TB-labeled samples into acoustically consistent and acoustically ambiguous subsets from the geometry of frozen audio embeddings alone, requiring no knowledge of which samples carry reliable ground-truth and no estimate of noise proportion.

II. Noisy Labels and Weakly Supervised Label Cleaning

A. Noisy Labels on TB Cough Datasets

Throughout this paper, "noisy label" is used in the machine-learning sense of supervision inconsistent with the observable feature distribution of its class, not in the clinical sense of an incorrect diagnosis. We do not dispute the microbiological confirmation: labels assigned by sputum smear microscopy, GeneXpert MTB/RIF, or mycobacterial culture establish the presence of *Mycobacterium tuberculosis* in sputum, but not whether the cough carries acoustically distinguishable TB-specific characteristics. A confirmed TB patient whose cough carries no such signature therefore supplies supervision that contradicts the acoustic evidence and behaves as label noise during training even though the diagnosis is correct: a feature–label mismatch rather than mislabelling. We accordingly name the two TB subgroups by the acoustic evidence they carry rather than by label validity: *Acoustically Consistent TB* for

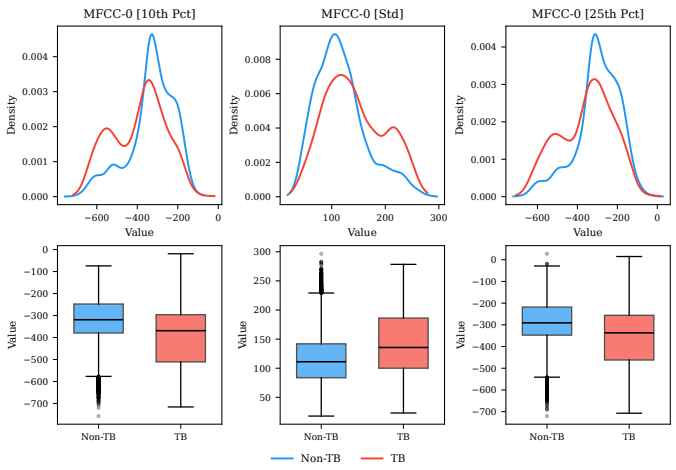


Fig. 1. Example acoustic characteristics of TB and non-TB cough samples.

confirmed cases whose cough matches the TB class, and *Acoustically Ambiguous TB* for confirmed cases whose cough does not.

Fig. 1 illustrates the resulting overlap using representative statistics of the zeroth MFCC coefficient (MFCC-0): the 10th percentile, 25th percentile, and standard deviation. The probability density function plots show substantial overlap between TB and non-TB distributions, while the boxplots indicate only modest differences in median values with heavily overlapping interquartile ranges. These observations motivate separating TB samples into the two subsets prior to classifier training.

B. Proposed Method

Algorithm 1 provides the full pseudocode. Given raw cough recordings with TB and non-TB labels from multiple datasets, the pipeline proceeds through four stages to produce a cleaned label set. All stages operate without any gradient-based training.

1) Audio Feature Extraction: The audio tower of Qwen3-Omni-30B-A3B-Thinking [16] is used as a frozen feature extractor. The Qwen3-Omni audio encoder is built on the AuT backbone, containing approximately 0.6B parameters. Raw waveforms are processed through a Mel frontend, producing frame-level hidden states $h_t \in \mathbb{R}^{2048}$. Temporal mean pooling across all T frames is used to aggregate frame-level representations into a fixed-length utterance embedding $z \in \mathbb{R}^{2048}$. No fine-tuning is applied; encoder weights are fixed throughout the entire pipeline.

In parallel, we use Google HeAR (Health Acoustic Representations) [5] as a second frozen feature extractor, also without fine-tuning. The two encoders are chosen for complementary reasons: HeAR is pretrained on a large corpus of non-speech health-related sounds, including cough, making it well matched to respiratory and TB cough analysis, whereas Qwen3-Omni is pretrained on a very large general audio–language corpus, providing broad acoustic coverage. The identical cleaning pipeline is applied to each encoder's embeddings.

Algorithm 1 Weakly Supervised Label Cleaning

Require: Embeddings $Z \in \mathbb{R}^{N\times D}$, labels $y \in \{0,1\}^N$
Ensure: Acoustically Consistent TB mask $\mathcal{M}_{\text{ac}}$, Acoustically Ambiguous TB mask $\mathcal{M}_{\text{aa}}$

1:
2: **Stage 1: Preprocessing**
3: $X \leftarrow \text{PCA}(Z,\ 256)$
4:
5: **Stage 2: Fisher Projection**
6: Fit LDA on $(X,\ y)$ {Eqs. 1–2}
7: $p_i \leftarrow w^\top x_i$ for all i {1-D projection}
8: **if** mean$(p[y{=}1]) <$ mean$(p[y{=}0])$ **then**
9: $\quad p \leftarrow -p$ {orient: higher = more TB-like}
10: **end if**
11:
12: **Stage 3: KDE Posterior and TB Likelihood**
13: Fit class-conditional KDE on $p[y{=}1]$, $p[y{=}0]$ {Eq. 3}
14: $s_i \leftarrow 1 - P(\text{NTB} \mid p_i)$ {TB-likelihood, Eqs. 4–5}
15:
16: **Stage 4: GMM Split (TB only)**
17: Fit GMM($k \in \{1,2,3\}$) on $\{s_i : y_i{=}1\}$, select k by BIC
18: **if** $\Delta_{\text{BIC}} > 0$ **and** effect size $r > 0.5$ **and** valid separation **then**
19: $\quad \mathcal{M}_{\text{ac}} \leftarrow$ TB in higher-mean component
20: $\quad \mathcal{M}_{\text{aa}} \leftarrow$ TB in lower-mean component
21: **else**
22: $\quad$ **return** all TB as Acoustically Consistent
23: **end if**
24:
25: **return** $\mathcal{M}_{\text{ac}}, \mathcal{M}_{\text{aa}}$

2) PCA Reduction: Each embedding z is projected via PCA to 256 dimensions, retaining 98.3% of total variance [17] (Algorithm 1, line 3). This step removes low-variance noise components while preserving the dominant acoustic structure.

3) Fisher Linear Discriminant Projection: After PCA reduction, embeddings are projected onto a single discriminative axis using Fisher Linear Discriminant Analysis (LDA) [18] (Algorithm 1, lines 6–9). The within-class scatter matrix is computed as:

$$S_W = \sum_{c\in\{TB,NTB\}} \sum_{x_i\in c} (x_i - \mu_c)(x_i - \mu_c)^\top, \qquad (1)$$

where c indexes the two classes, x_i is the i-th PCA-reduced embedding, and μ_c is the mean embedding of class c. The Fisher discriminant direction is then:

$$w \propto S_W^{-1}(\mu_{TB} - \mu_{NTB}), \qquad (2)$$

where $w \in \mathbb{R}^{256}$ is the projection vector, and μ_{TB}, μ_{NTB} are the class centroids. Each sample is projected as $p_i = w^\top x_i$, with sign adjusted so that larger values correspond to greater TB-likeness. This one-dimensional axis becomes the basis for likelihood estimation.

4) KDE-Based TB Likelihood Estimation: After Fisher projection, class-conditional densities are estimated on the one-dimensional projected scores using Gaussian kernel density estimation (KDE) with Silverman bandwidth selection (Algorithm 1, lines 12–13):

$$p(p_i \mid TB) = \hat{f}_{TB}(p_i), \qquad p(p_i \mid NTB) = \hat{f}_{NTB}(p_i), \quad (3)$$

where p_i denotes the Fisher-projected score of sample i, and $\hat{f}_{TB}$, $\hat{f}_{NTB}$ are KDE models fitted separately on TB and non-TB samples. Using class priors π_{TB} and π_{NTB} estimated from label frequencies, the posterior probability that a sample belongs to the non-TB class is computed using Bayes' rule:

$$P(\text{NTB} \mid p_i) = \frac{\hat{f}_{NTB}(p_i)\pi_{NTB}}{\hat{f}_{TB}(p_i)\pi_{TB} + \hat{f}_{NTB}(p_i)\pi_{NTB}}. \quad (4)$$

Samples with high $P(\text{NTB} \mid p_i)$ despite carrying TB labels are considered acoustically inconsistent with the TB class. The acoustic TB-likelihood score is defined as:

$$s_{TB}(p_i) = 1 - P(\text{NTB} \mid p_i), \quad (5)$$

where $P(\text{NTB} \mid p_i)$ is the non-TB posterior. A high s_{TB} indicates strong acoustic similarity to TB cough; a low s_{TB} indicates acoustic similarity to non-TB cough.

5) GMM Label Split: Only TB samples are subject to splitting (Algorithm 1, lines 16–21). A GMM is fit to the acoustic TB-likelihood score distribution $\{s_{TB}(p_i) : y_i = 1\}$ with $k \in \{1, 2, 3\}$ components; the optimal number is selected by BIC. A valid split requires three conditions: (i) BIC favors k=2 over k=1 ($\Delta_{\text{BIC}} > 0$), (ii) sufficient separation between component means (effect size $r = |\mu_1 - \mu_0|/\bar{\sigma} > 0.5$), and (iii) the components straddle the decision boundary ($\mu_0 < 0.5 < \mu_1$) or exhibit strong separation ($r > 1.5$). The higher-mean component is designated *Acoustically Consistent TB*; the lower-mean component is *Acoustically Ambiguous TB*.

III. Experimental Setup

A. Datasets

Three cough datasets are used, hereafter referred to as DB1–DB3:

- **DB1 (CODA):** From the CODA TB DREAM Challenge [7], containing cough recordings from TB and non-TB subjects collected across multiple clinical sites. Although the dataset provides both solicited and automatic cough recordings, this study only uses solicited coughs.
- **DB2 (CIDRZ):** A real-world clinical cough dataset [5] collected in a high-burden TB setting. Since CIDRZ recordings are unsegmented (full-length audio without cough boundaries), we built a cough detection module by modifying the approach of Jansen van Vüren et al. [19]: we retain the LSTM backbone but adapt it to predict a binary cough indicator, extracting individual cough events and yielding approximately 5–8 cough segments per recording. On held-out validation audio the detector attains a frame-level AUC of 0.995 and average precision of 0.992; at the event level it recovers 97.0% of annotated cough events (coverage) with 94.2% of detected frames falling inside annotated cough regions (purity).
- **DB3 (TBScreen):** The dataset from Sharma et al. [6], with microbiologically verified TB and non-TB subjects. Unlike CODA, this work uses the *automatic* (passively recorded) coughs rather than solicited coughs. Because the number of automatic cough files varies widely across patients and devices, we apply a balancing rule: whenever a patient contributes more than 35 cough files for a given recording device, 35 files are randomly selected, preventing high-volume patients and devices from dominating the dataset.

B. Classifier Evaluation

Three classifiers are evaluated on the same embedding space: Logistic Regression (LR), Multi-Layer Perceptron (MLP), and Histogram-based Gradient Boosting (HistGradientBoostingClassifier, HGB). Each classifier uses the original Qwen3-Omni (2048-dim) or HeAR (512-dim) embeddings, followed by StandardScaler normalization and PCA reduction to 256 dimensions. No external held-out test set is used; instead, stratified group 5-fold cross-validation is adopted for all three datasets, since label noise may concentrate in an arbitrary subset of the data and evaluating on every fold avoids conclusions drawn from one possibly-contaminated split. Participant ID is the grouping variable, so no participant appears in both training and validation folds; with respect to participant leakage this is stricter than the single train/test split used by most prior cough TB studies.

Metrics are computed independently per fold and averaged, then averaged again across the three classifiers; this two-stage averaging reduces sensitivity to any particular classifier choice or fold partition.

C. Evaluation Metrics, Baselines, and Subsets

We report four metrics: AUROC as the primary discrimination metric given class imbalance [20], UAR (unweighted average recall, i.e., balanced accuracy), sensitivity (proportion of TB cases correctly identified), and specificity (proportion of non-TB cases correctly identified). All configurations use identical raw audio embeddings as classifier input; the only difference is the training data, ensuring that any performance difference is attributable solely to label quality. We compare four configurations: (i) *Original*, using all samples with their original labels; (ii) *Random Cleaning*, which removes exactly as many TB samples as the number flagged as Acoustically Ambiguous TB (noisy) by the proposed pipeline, but selects them uniformly at random, isolating the effect of merely shrinking the TB set; to reduce variance from the random draw, this baseline is repeated 15 times and the metrics are averaged; (iii) *Confident Learning* [21], an established noisy-label baseline that prunes likely-mislabeled samples from the same embedding space, evaluated with both of its pruning strategies, prune-by-noise-rate (PBNR) and the default confident-learning (CL) ranking, using the official `cleanlab` implementation;

and (iv) *Proposed Cleaning*, which excludes the Acoustically Ambiguous TB samples identified by the GMM split.

IV. Results and Discussion

A. Distribution Analysis and Label Splitting

Fig. 2 presents the Fisher LDA projection per dataset. Class separation is modest across all three datasets, reflecting differences in collection protocols and label quality and consistent with the low Original-subset AUROC reported in Table I. On the raw axis, TB and non-TB distributions overlap substantially in every dataset, indicating that a portion of TB-labelled samples sits near the non-TB cluster prior to splitting.

Fig. 3 shows the acoustic TB-likelihood score distributions after GMM splitting. Acoustically Consistent TB samples (red) are separated from Acoustically Ambiguous TB samples (orange) within the TB population across all three datasets. Non-TB samples (blue) cluster at low likelihood values across all datasets. The per-dataset likelihood scores reflect the underlying label quality: in every dataset a sizeable subset of TB-labelled samples falls at low likelihood, overlapping the non-TB distribution, while the remaining Acoustically Consistent TB samples concentrate at higher likelihood. Across all three datasets, GMM splitting on Qwen3-Omni embeddings identified 5,284 samples as Acoustically Consistent TB (57.8%) and 3,855 as Acoustically Ambiguous TB (42.2%) within the TB population; on HeAR embeddings it identified 5,488 Acoustically Consistent TB (60.0%) and 3,651 Acoustically Ambiguous TB (40.0%). Per-dataset breakdown (Qwen / HeAR): DB1 (CODA), 1,454 / 1,453 Acoustically Consistent TB of 2,773 (52.4% / 52.4%); DB2 (CIDRZ), 562 / 578 of 1,279 (43.9% / 45.2%); DB3 (TBScreen), 3,268 / 3,457 of 5,087 (64.2% / 68.0%). The proposed filtering strategy remains conservative, preserving many TB-labelled samples instead of aggressively removing acoustically ambiguous coughs. However, some retained TB coughs still overlap acoustically with non-TB coughs, suggesting that better cough-specific acoustic features are needed in future work.

B. Classification Performance

Table I presents classification performance for both encoders across all datasets and configurations. On the *Original* configuration, classifiers trained on all samples with original labels stay well below the WHO threshold under both encoders: with Qwen3-Omni they reach only 64% AUROC on CODA, 66% on CIDRZ, and 69% on TBScreen, and HeAR shows the same low range (67%, 64%, 60%).

The *Random Cleaning* baseline, which removes the same number of TB samples uniformly at random, leaves AUROC essentially unchanged for both encoders (within $\pm 2\%$ of Original), confirming that the gains are not an artifact of simply shrinking the TB set. Both Confident Learning variants, CL (PBNR) and CL (CL), improve over Original (to roughly 73–82% AUROC) but consistently trail the proposed method, with PBNR generally the stronger of the two; this indicates that generic noisy-label pruning captures only part of the acoustic mismatch.

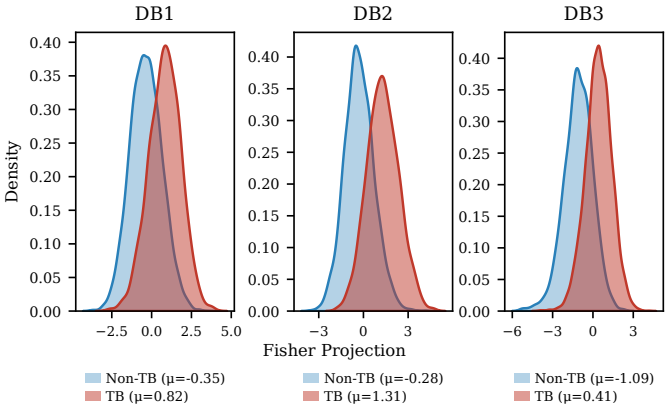


Fig. 2. Fisher LDA projection after PCA-256, shown per dataset.

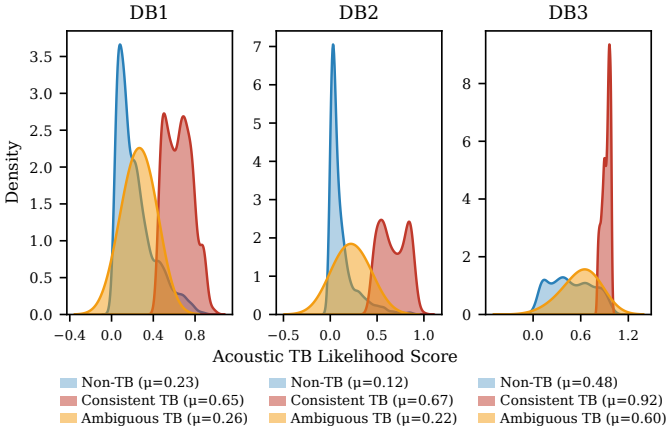


Fig. 3. KDE TB-likelihood score distributions after GMM split, shown per dataset. *Consistent TB* and *Ambiguous TB* abbreviate Acoustically Consistent TB and Acoustically Ambiguous TB.

Proposed Cleaning delivers the largest gains on every dataset and encoder. With Qwen3-Omni, CODA improves by ∼22% (64% to 86%), CIDRZ by ∼22% (66% to 88%), and TBScreen by ∼20% (69% to 89%); HeAR shows the same pattern (CODA 67% to 89%, CIDRZ 64% to 89%, TBScreen 60% to 83%). This pattern confirms the central hypothesis: classifiers trained on original labels were confused by acoustically inconsistent TB samples that provided contradictory supervision. Once these Acoustically Ambiguous TB samples are removed, the remaining Acoustically Consistent TB and non-TB samples form a cleaner training set where the decision boundary aligns with genuine acoustic differences. The consistency of the improvement across two independent frozen encoders indicates that it stems from label cleaning rather than from any particular embedding. Notably, TBScreen, which here uses passively recorded *automatic* coughs rather than solicited coughs, starts at a much lower Original AUROC than reported for solicited recordings, and benefits from cleaning to the same extent as the real-world datasets, indicating that acoustic label ambiguity is pronounced once naturalistic

TABLE I
CLASSIFICATION PERFORMANCE FOR BOTH AUDIO ENCODERS (QWEN3-OMNI AND HEAR), MEAN ± STD (%) ACROSS FOLDS

Dataset	Method	AUROC		UAR		Sens		Spec	
		Qwen	HeAR	Qwen	HeAR	Qwen	HeAR	Qwen	HeAR
CODA	Original	64 ± 1	67 ± 1	60 ± 1	63 ± 1	54 ± 3	59 ± 2	66 ± 1	66 ± 2
	Random Cleaning	63 ± 1	66 ± 1	59 ± 1	62 ± 1	54 ± 2	58 ± 1	65 ± 1	66 ± 1
	CL (PBNR)	79 ± 2	82 ± 2	70 ± 2	73 ± 2	64 ± 2	69 ± 2	76 ± 1	77 ± 2
	CL (CL)	75 ± 2	79 ± 2	67 ± 1	71 ± 2	61 ± 2	67 ± 2	73 ± 1	75 ± 1
	Proposed (Ours)	$\mathbf{86 \pm 2}$	$\mathbf{89 \pm 3}$	$\mathbf{76 \pm 3}$	$\mathbf{80 \pm 3}$	$\mathbf{71 \pm 3}$	$\mathbf{78 \pm 3}$	$\mathbf{82 \pm 3}$	$\mathbf{83 \pm 3}$
CIDRZ	Original	66 ± 0	64 ± 1	59 ± 1	59 ± 0	45 ± 3	42 ± 2	74 ± 1	75 ± 3
	Random Cleaning	66 ± 1	65 ± 1	60 ± 1	60 ± 1	48 ± 2	48 ± 2	72 ± 1	72 ± 1
	CL (PBNR)	75 ± 2	75 ± 1	66 ± 2	66 ± 1	53 ± 4	56 ± 3	79 ± 1	77 ± 2
	CL (CL)	73 ± 1	73 ± 1	65 ± 2	65 ± 1	52 ± 5	51 ± 3	77 ± 1	78 ± 2
	Proposed (Ours)	$\mathbf{88 \pm 3}$	$\mathbf{89 \pm 2}$	$\mathbf{77 \pm 3}$	$\mathbf{80 \pm 2}$	$\mathbf{68 \pm 5}$	$\mathbf{73 \pm 2}$	$\mathbf{86 \pm 2}$	$\mathbf{87 \pm 2}$
TBScreen	Original	69 ± 1	60 ± 2	64 ± 1	56 ± 1	72 ± 2	63 ± 2	56 ± 2	50 ± 3
	Random Cleaning	68 ± 3	60 ± 2	62 ± 2	56 ± 1	76 ± 2	65 ± 1	49 ± 5	47 ± 2
	CL (PBNR)	75 ± 5	79 ± 5	68 ± 3	72 ± 4	77 ± 3	75 ± 3	58 ± 4	68 ± 5
	CL (CL)	76 ± 4	72 ± 4	68 ± 3	66 ± 3	79 ± 3	72 ± 3	56 ± 5	61 ± 3
	Proposed (Ours)	$\mathbf{89 \pm 5}$	$\mathbf{83 \pm 6}$	$\mathbf{80 \pm 6}$	$\mathbf{75 \pm 5}$	$\mathbf{88 \pm 4}$	$\mathbf{77 \pm 4}$	$\mathbf{73 \pm 8}$	$\mathbf{74 \pm 6}$

All methods use identical embeddings within each encoder; only the training labels differ. CL (PBNR) and CL (CL) are the prune-by-noise-rate and confident-learning variants of Confident Learning [21]. **Bold** = best value in each column, per dataset and encoder.

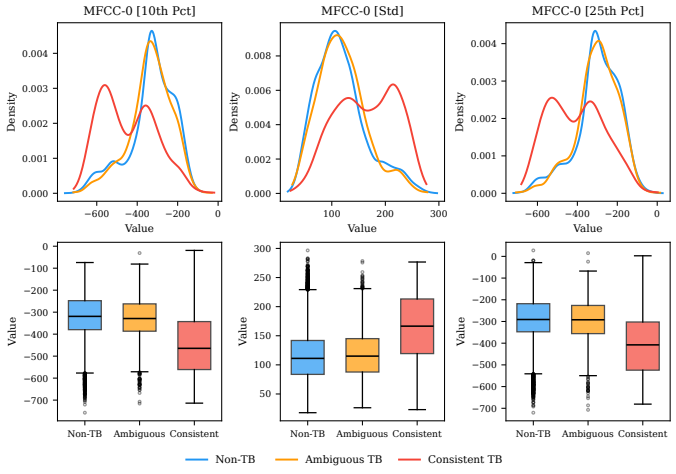


Fig. 4. Example acoustic characteristics after label cleaning. *Consistent* and *Ambiguous* abbreviate Acoustically Consistent TB and Acoustically Ambiguous TB.

coughs are used.

Despite these gains, no configuration reaches the WHO screening benchmark of ≥90% sensitivity at ≥70% specificity stated in Sec. I. The highest sensitivity observed is 88% (TBScreen, Qwen3-Omni, at 73% specificity), while CODA and CIDRZ reach 71% and 68% sensitivity respectively. We further note that a sensitivity computed after a portion of the positive set has been removed is not directly comparable to the WHO criterion, which is defined over all microbiologically confirmed cases; the values above should therefore be read as an upper bound rather than as screening sensitivity.

C. Acoustic Characteristics After Splitting

Fig. 4 describes the three resulting groups from an acoustic perspective. After GMM splitting, the three groups exhibit different MFCC-0 characteristics. Acoustically Consistent TB samples show lower MFCC-0 energy and higher standard deviation than non-TB samples, whereas Acoustically Ambiguous TB samples overlap the non-TB distributions across all three statistics: despite carrying TB labels, they are acoustically close to non-TB coughs. The separation between Acoustically Consistent TB and non-TB is also wider than the separation between all TB and non-TB before cleaning (Fig. 1), in line with the AUROC gains in Table I.

V. CONCLUSION

We hypothesized that label noise arising from the mismatch between microbiological TB diagnosis and cough acoustics is a substantial contributor to the low AUC reported for existing TB cough classifiers. To test this, we proposed a weakly supervised, training-free label cleaning pipeline that separates Acoustically Consistent TB samples from Acoustically Ambiguous TB samples, requiring no knowledge of noise rates or reliable ground-truth identification. Evaluated on three cough datasets (CODA, CIDRZ, and TBScreen), the approach improves AUROC by ∼20–22% after cleaning (CODA: 64% to 86%; CIDRZ: 66% to 88%; TBScreen: 69% to 89%), outperforming both a random-removal baseline and Confident Learning, and generalizing across two frozen audio encoders (Qwen3-Omni and HeAR). These results indicate a substantial impact of label ambiguity on downstream performance. Evaluation is nonetheless restricted to the three cohorts used here, as no further TB cough corpus with microbiological confirmation is currently publicly available. Future work will characterise the Acoustically Ambiguous subgroup against the clinical metadata accompanying these cohorts, in particular the predicted enrichment for subclinical presentation [10] and for patients under treatment [11]; quantify the pipeline on data with label noise injected at a known rate; and explore integrating acoustic TB-likelihood scores as soft labels in

downstream model training and extending the framework to multi-dataset co-training for more robust TB screening. Code is available at https://github.com/arkiven4/pub-wsnlc2026.

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.

REFERENCES

[1] Y. Yang, Y. Ruan, X. Li, J. Xiong, and Y. Zheng, "Global, regional, and national burden of tuberculosis and attributable risk factors for 204 countries and territories, 1990–2021: A systematic analysis for the Global Burden of Diseases 2021 study," *BMC Public Health*, vol. 24, no. 1, p. 3120, 2024.

[2] S. Xie, H. Xiao, L. Xu, F. Zhang, and M. Luo, "Decadal trends and regional disparities in tuberculosis burden: A comprehensive analysis of global, african, and southeast asian data from the GBD 1990–2021," *Frontiers in Public Health*, vol. 13, p. 1 467 509, 2025.

[3] M. Nadjib, S. Dewi, E. Setiawan, T. Y. Miko, and S. Putri, "Cost and affordability of scaling up tuberculosis diagnosis using Xpert MTB/RIF testing in west java, indonesia," *PLoS ONE*, vol. 17, no. 3, e0264912, 2022.

[4] M. Pahar, M. Klopper, B. Reeve, R. Warren, G. Theron, and T. Niesler, "Automatic cough classification for tuberculosis screening in a real-world environment," *Physiological Measurement*, vol. 42, no. 10, p. 105 014, 2021.

[5] S. Baur et al., "HeAR – health acoustic representations," *arXiv preprint*, 2024.

[6] M. Sharma et al., "TBscreen: A passive cough classifier for tuberculosis screening with a controlled dataset," *Science Advances*, vol. 10, no. 1, eadi0282, 2024.

[7] D. Jaganath, S. K. Sieberts, M. Raberahona, S. Huddart, L. Omberg, et al., "Accelerating cough-based algorithms for pulmonary tuberculosis screening: Results from the CODA TB DREAM challenge," *Open Forum Infectious Diseases*, vol. 12, no. 10, ofaf572, 2025.

[8] G. P. Kafentzis and E. Selisios, "Tuberculosis screening from cough audio: Baseline models, clinical variables, and uncertainty quantification," *arXiv preprint*, 2026.

[9] A. D. Prasetyo, B. T. Atmaja, D. Arifianto, and S. Sakti, "A comparison of solicited and longitudinal cough sounds for tuberculosis detection," in *2025 APSIPA ASC*, 2025, pp. 1657–1662.

[10] B. Frascella et al., "Subclinical tuberculosis disease—a review and analysis of prevalence surveys to inform definitions, burden, associations, and screening methodology," *Clinical Infectious Diseases*, vol. 73, no. 3, e830–e841, 2021.

[11] A. Proaño, M. A. Bravard, J. W. López, G. O. Lee, et al., "Dynamics of cough frequency in adults undergoing treatment for pulmonary tuberculosis," *Clinical Infectious Diseases*, vol. 64, no. 9, pp. 1174–1181, 2017.

[12] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical Image Analysis*, vol. 65, p. 101 759, 2020.

[13] A. Hekler, J. S. Utikal, A. H. Enk, et al., "Effects of label noise on deep learning-based skin cancer classification," *Frontiers in Medicine*, vol. 7, p. 177, 2020.

[14] C. Lin et al., "Deep learning network intensification for preventing noisy-labeled samples for remote sensing classification," *Remote Sensing*, vol. 13, no. 9, p. 1689, 2021.

[15] H. Song, M. Kim, D. Park, Y. Shin, and J. G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8135–8153, 2023.

[16] Qwen Team, "Qwen3-Omni technical report," *arXiv preprint*, 2025.

[17] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2065, p. 20 150 202, 2016.

[18] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[19] J. Jansen van Vüren et al., *Cough activity detection for automatic tuberculosis screening*, arXiv:2603.11241, 2026.

[20] E. Richardson, R. Trevizani, J. A. Greenbaum, H. Carter, M. Nielsen, and B. Peters, "The receiver operating characteristic curve accurately assesses imbalanced datasets," *Patterns*, vol. 5, no. 6, p. 100 994, 2024.

[21] C. G. Northcutt, L. Jiang, and I. L. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *Journal of Artificial Intelligence Research*, vol. 70, pp. 1373–1411, 2021.