PREPRINT · MACHINE LEARNING FOR HEALTH

Site-invariant representations for cough-audio screening: removing the confound does not recover disease signal

A leave-one-country-out evaluation of adversarial domain-invariance on 2,739 crowdsourced cough recordings from nine countries.

Abdoulie Balisa BSc Statistics, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
abalisa@st.knust.edu.gh 5 September 2026 github.com/Balisa50/cough-tb-invariance

ABSTRACT

Cough-audio classifiers for respiratory disease report high in-domain accuracy and degrade sharply at new recording sites. Recent work attributes this failure to learned representations that organise by recording device and source dataset rather than by disease status, with predicted probability tracking country-level prevalence. That work diagnosed the problem and observed that device-diverse training improves transfer, without isolating the mechanism or verifying that the site information had in fact been removed. We implement a correction combining randomised device simulation with an adversarial site head trained through a gradient reversal layer, and we audit the outcome using a probe trained on frozen features to recover the recording site, so that invariance is measured directly rather than inferred from classification accuracy.

We evaluate on 2,739 COUGHVID recordings from nine countries under leave-one-country-out cross-validation, reporting specificity at 90 percent sensitivity as the operating point. The intervention removes site information: probe leakage falls from 0.018 to zero in eight of nine folds (Wilcoxon signed-rank, $p = 0.012$). Disease performance does not change (mean AUC 0.464 to 0.489, $p = 0.359$). A null model that discards the audio entirely and predicts each country's base rate achieves an AUC of 0.741, exceeding every fold measured in this study. We conclude that the invariance mechanism functions as intended, and that this corpus contains no cross-country disease signal for it to preserve.

1 Introduction

Automated cough analysis has been proposed as a triage tool for tuberculosis and other respiratory conditions, motivated by the low cost of audio capture relative to microbiological testing. Reported accuracies are frequently high. Performance at sites not represented in training is considerably lower, and the discrepancy has generally been attributed to population differences rather than to acquisition conditions.

Zhang et al. [1] evaluated classical and deep-learning cough classifiers for tuberculosis across three independent datasets and reached a different conclusion. Audio representations organised by recording device and source dataset rather than by TB status, and predicted probability tracked country-level prevalence in CODA. Within-dataset performance reached an AUC of 0.755, while external performance fell frequently below 0.6. A baseline using clinical variables alone generalised more consistently, at 0.655 to 0.711, which is what led them to conclude that acquisition-specific variability is a stronger driver of poor generalisation than population shift.

They also report that device-diverse training improves transfer, which points towards the intervention evaluated here. What that observation does not establish is whether the device dependence can be removed rather than diluted, and whether performance survives its removal. If site information can be taken out of the representation, reported performance should survive the operation. If it cannot be removed, or if performance collapses once it is, then the reported performance was the confound. We are not aware of a public implementation that performs the removal, verifies it, and re-measures.

This paper reports such an attempt. The contribution is threefold. First, we implement domain-invariant training for cough audio and release the code. Second, we introduce an explicit audit of the invariance claim in the form of a site probe, so that the mechanism is verified rather than assumed. Third, we report a negative result on COUGHVID under a strict evaluation protocol, and we quantify the confound using a prevalence-only null model that has not previously been reported for this task.

2 Related work

Domain-adversarial training was introduced by Ganin and Lempitsky [2], who placed a gradient reversal layer between a shared encoder and a domain classifier, so that the encoder is optimised to make the domain unpredictable while the task head is optimised normally. The approach has since been applied widely in vision and speech, though rarely with independent verification that the domain information was in fact removed.

In respiratory audio, most published work reports within-corpus performance under random splits. Where cross-corpus evaluation is performed, it is typically presented as a robustness check rather than as the primary result. Zhang et al. [1] is the exception, and provides the motivation for the present work.

Crowdsourced cough corpora, of which COUGHVID [3] is the largest public example, were assembled during the COVID-19 pandemic and carry self-reported labels. CODA TB [4] provides microbiologically confirmed tuberculosis labels across seven countries, but requires institutional credentialling for access.

3 Methods

3.1 Device simulation

Each clip is re-coloured on every training epoch as though captured by a different handset. The simulator applies a random band-pass filter within the range occupied by mobile microphones, a spectral tilt, two resonant peaks at randomised centre frequencies, gain variation, an additive noise floor, and occasional soft clipping. Because the simulated device changes at every epoch, no consistent device signature remains available for identifying a site.

An initial implementation randomised only the spectral tilt and held resonances fixed. Devices distinguishable by their resonant peak remained fully distinguishable under that version. The augmentation family must therefore span the family of transformations that the hardware applies; any component left uncovered survives as an intact shortcut. Time-stretching and pitch-shifting were excluded deliberately, since those operations alter the cough itself rather than the recording channel.

3.2 Adversarial site head

A second classification head predicts the recording site from the shared embedding. Its gradient is negated during backpropagation by a gradient reversal layer [2], so that the encoder is optimised to make site membership unpredictable from its own features. Reversal strength follows the schedule of Ganin and Lempitsky, rising from zero over the course of training, so that the encoder acquires useful features before the adversary begins to remove them.

3.3 The site probe

Classification accuracy cannot establish whether site information was removed. After training, we therefore freeze the encoder and fit a fresh classifier on its features with the single objective of recovering the recording site. If that probe performs at chance, the information is absent. If it succeeds, the encoder has concealed site membership from its own adversary without discarding it, and the invariance claim fails regardless of the reported disease accuracy.

Every fold reports *site leak*, defined as probe accuracy above the majority-class baseline, which is 0.332 on this manifest. An improvement in classification accuracy unaccompanied by a reduction in site leak does not constitute invariance.

3.4 Architecture and training

The encoder comprises four convolutional blocks over 64-band log-mel spectrograms, computed from half-second windows selected by signal energy, and produces a 128-dimensional embedding. Training runs for twelve epochs at batch size 32 with a learning rate of 0.001, on CPU. The probe is trained for 40 epochs on frozen features. Folds are deterministic under a fixed seed, which we verified by reproducing all nine baseline folds to three decimal places across independent runs.

4 Evaluation protocol

Two constraints are enforced in code rather than by convention, since both are failure modes that yield optimistic and publishable results.

- **No participant spans a split.** A single participant contributes multiple clips. Splitting clips at random places the same larynx on both sides of the partition, and measures memorisation rather than generalisation.
- **The test site is unseen.** Leave-one-country-out is the only partition under which identifying the site confers no advantage.

The primary metric is specificity at 90 percent sensitivity, following the World Health Organization target product profile for tuberculosis triage tests [5], which specifies 70 percent specificity at that sensitivity. AUC is reported alongside, but is not the headline figure, since the two measures can diverge, and in these results they do.

Each fold additionally reports a prevalence-only null: the AUC obtained by assigning every participant the base rate of their own country while ignoring the audio entirely. This quantifies the confound identified by Zhang et al., and provides the reference against which any reported performance must be judged.

The protocol was validated before any real corpus was obtained, using a synthetic cohort constructed with no disease signal, in which the label is reachable only through site prevalence. A same-site random split scored 0.634 and appeared successful; the prevalence-only null scored 0.637; leave-one-site-out scored 0.499. The protocol therefore identified apparent skill as an artefact of geography in a setting where ground truth was known by construction.

5 Data

COUGHVID [3] is a corpus of cough recordings collected during 2020 on participants' own devices. The release used here, Zenodo record 4048312, contains 20,072 submissions; later releases of the corpus are larger, and the counts below refer to this one. Relative to the seven controlled sites of CODA, device variation in COUGHVID is the phenomenon of interest rather than a proxy for it, which makes it a more demanding test of invariance rather than a less demanding one.

The corpus provides neither a site field nor participant identifiers. We derive country by reverse-geocoding the latitude and longitude present on 58 percent of rows, and we set `participant_id` to the submission identifier, since each uuid corresponds to one anonymous recording. Requiring a status label, a detected cough (the corpus `cough_detected` score above 0.8, as recommended in the data descriptor), and a location reduces the corpus to 4,409 clips. Restricting to COVID-19 against healthy, and excluding countries with fewer than 60 clips or 5 positive cases, yields the analysis set shown in Table 1.

COUNTRY	CLIPS	POSITIVES	PREVALENCE
Turkey	830	34	0.041
Spain	605	46	0.076
Uzbekistan	488	141	**0.289**
United States	231	6	**0.026**
Switzerland	201	19	0.095
France	115	11	0.096
Brazil	110	5	0.045
Argentina	96	18	0.188
Iran	63	16	0.254
Total	**2,739**	**296**	**0.108**

Table 1. The analysis set. Prevalence ranges from 0.026 in the United States to 0.289 in Uzbekistan, an elevenfold spread. This range determines the magnitude of the shortcut available to any model capable of identifying the recording country.

6 Results

Nine folds were run under two arms over identical partitions. The baseline arm disables both the adversary and the augmentation; the intervention arm enables both.

Figure 1. All eighteen fold measurements on a single AUC axis. The shaded region lies below chance. No measurement reaches the prevalence-only null at 0.741. France and Uzbekistan show the largest movement under the intervention; both carry fewer than 20 positive cases, so these differences fall within sampling noise.

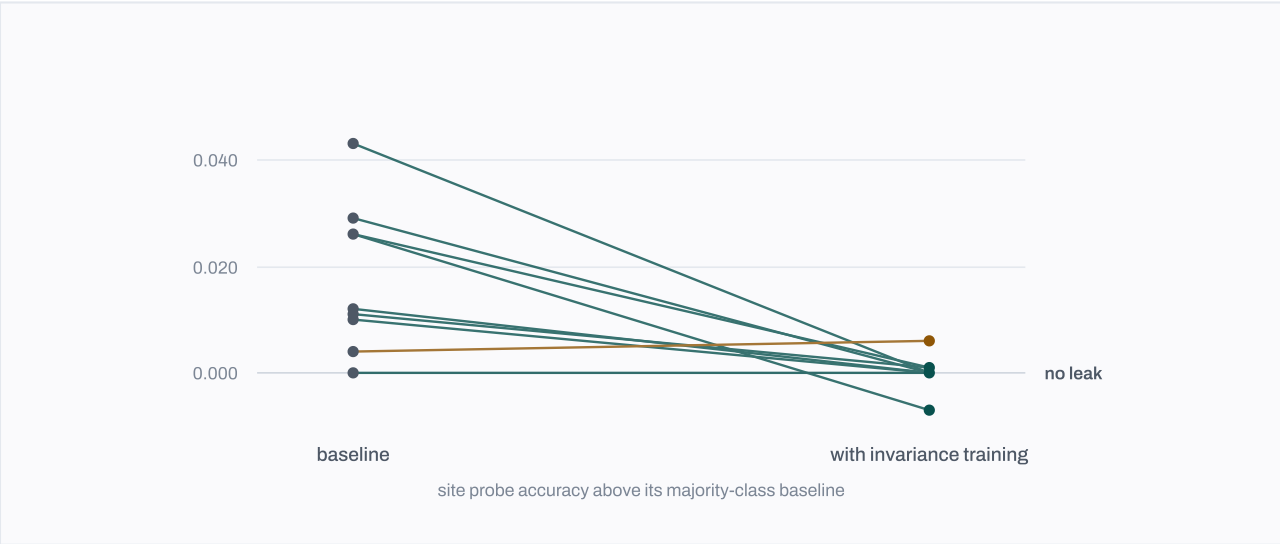

Figure 2. Site leakage by fold, before and after the intervention. Eight of nine folds fall to zero or below. The United States, shown in ochre, is the single exception, rising from 0.004 to 0.006; both values are small enough to represent noise about zero. Turkey, the leakiest fold at 0.043, is also the fold whose AUC declines most once the leak is removed.

	Baseline			With invariance training		
HELD OUT	AUC	SPEC@90	LEAK	AUC	SPEC@90	LEAK
Argentina	0.368	0.141	0.029	0.431	0.051	0.000
Brazil	0.392	0.124	0.011	0.396	0.210	0.001
Switzerland	0.428	0.033	0.026	0.368	0.093	0.001
Spain	0.519	0.193	0.012	0.469	0.116	0.000
France	0.523	0.077	0.026	0.637	0.510	-0.007
Iran	0.523	0.043	0.010	0.592	0.021	0.000
Turkey	0.549	0.121	**0.043**	0.509	0.113	0.000
United States	0.449	0.120	0.004	0.427	0.231	0.006
Uzbekistan	0.420	0.017	0.000	0.571	0.207	0.000
Mean	**0.464**	**0.096**	**0.018**	**0.489**	**0.173**	**0.000**

Table 2. Complete fold-level results. Paired Wilcoxon signed-rank tests across the nine countries give $p = 0.359$ for AUC, $p = 0.301$ for specificity at 90 percent sensitivity, and $p = 0.012$ for site leak. Only the reduction in leak reaches significance. AUC improved in five of nine folds; leak fell in eight of nine.

6.1 Divergence between AUC and the operating point

Iran gains AUC under the intervention, from 0.523 to 0.592, while its specificity at 90 percent sensitivity halves, from 0.043 to 0.021. Spain loses on both measures. Brazil is nearly unchanged on AUC and nearly doubles specificity. Because these measures move independently, a study reporting AUC alone could support several different conclusions from the same runs, depending on which folds it emphasised. Fixing the operating point in advance removes that latitude.

The strongest single fold in the study, France at 0.510 specificity, remains below the 0.70 required by the WHO target product profile, and rests on eleven positive cases.

7 Discussion

HYPOTHESIS 1: SUPPORTED

Site information can be removed from the learned representation.

`probe leak 0.018 to 0.000`
`8 of 9 folds, p = 0.012`

HYPOTHESIS 2: NOT SUPPORTED

Cough audio carries disease signal that generalises across countries.

`AUC 0.464 to 0.489, p = 0.359`
`country base rate alone, 0.741`

The invariance mechanism operates as designed on real audio. A probe trained specifically to defeat the claim cannot recover the recording country above its majority-class baseline. This constitutes stronger evidence than an improvement in classification accuracy, because it tests the mechanism directly rather than its intended consequence.

The corpus contains no cross-country disease signal. A mean AUC of 0.489 is indistinguishable from chance, and the prevalence-only null of 0.741 exceeds every individual fold measured. On these data, knowing a participant's country predicts their COVID-19 status substantially better than a convolutional network applied to their cough.

The magnitude of the null model merits attention, since it bears on the interpretation of prior results. Zhang et al. characterised cross-dataset AUC below 0.6 as degradation. Measured against a null that discards the audio, a stronger characterisation is warranted: on crowdsourced recordings the audio contributes nothing that transfers, and reported performance is largely a measurement of epidemiology. We suggest that a prevalence-only null be reported as standard in this literature, since without it a model that has learned only geography cannot be distinguished from one that has learned pathology.

These results also demonstrate that the protocol discriminates on real data, having previously been demonstrated only on a synthetic cohort. A within-site random split on this corpus would have produced apparently acceptable performance. Held out to an unseen country, performance falls below the null.

8 Limitations

- **This is not a test of tuberculosis screening.** COUGHVID labels record self-reported COVID-19 status, and only 4 percent of the corpus carries an expert diagnosis. A negative result here constrains claims about crowdsourced COVID-19 cough audio; it does not constrain claims about tuberculosis.
- **The positive class is small.** With 296 positives distributed across nine countries, several folds contain fewer than 20 positive cases, giving AUC confidence intervals wide enough to include chance. Fold-level differences should not be interpreted individually; the paired tests across folds carry the inference.
- **No participant identifiers exist.** Each uuid corresponds to one anonymous submission, so grouping operates at the level of the recording. A repeat submitter cannot be detected, and the guarantee weakens from "no participant spans a split" to "no recording spans a split". This is a property of the corpus that no protocol can repair.
- **Country is a proxy for device rather than a measurement of it.** Two countries may share a handset market, which would make the held-out unit weaker than intended. The site probe reports the separability actually achieved, which carries that uncertainty explicitly rather than assuming it away.
- **Location is available on 58 percent of rows**, reducing the usable set to 2,739 of 20,072. Whatever determines a participant's willingness to grant location permission may not be independent of other characteristics.
- **A single architecture and hyperparameter configuration was evaluated.** A larger encoder or a longer schedule might extract signal that this configuration cannot, although the margin separating measured performance from the prevalence null is wide enough that closing it would require more than parameter tuning.

9 Conclusion

Adversarial domain-invariance removes recording-site information from cough-audio representations, verified by a probe rather than inferred from accuracy. On COUGHVID this removal produces no change in disease classification, because the corpus contains no cross-country signal to preserve. A null model using

country base rates alone outperforms every configuration tested.

Establishing whether the method benefits a corpus that does contain disease signal requires microbiologically confirmed labels. CODA TB [4] provides 2,143 participants across seven countries, four of them in Africa, with culture-confirmed outcomes. The method, the evaluation protocol and the manifest abstraction are dataset-agnostic by construction: incorporating a corpus requires implementing a single function that emits four columns, with no other modification. The apparatus is complete, and data access is the remaining constraint.

10 Code and data availability

Source code, the evaluation protocol, complete fold-level results in JSON, and a checksum-verified downloader for COUGHVID are available at github.com/Balisa50/cough-tb-invariance. All experiments run on CPU. A synthetic pathway reproduces the protocol validation without requiring any download. COUGHVID is distributed under CC BY 4.0 from Zenodo record 4048312.

References

1. Zhang, W., Teijeiro, T., Thevenot, J. and Atienza, D. Why ML-based cough models do not generalize: a systematic cross-dataset evaluation for tuberculosis screening. *arXiv:2608.25846*, 26 August 2026.
2. Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 1180–1189.
3. Orlandic, L., Teijeiro, T. and Atienza, D. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8:156, 2021. Release used here: Zenodo record 4048312, CC BY 4.0.
4. CODA TB DREAM Challenge dataset. Synapse `syn31472953`.
5. World Health Organization. *High-priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting*. Geneva, 2014.

All figures and tables are generated programmatically from `results/coughvid_folds.json` in the accompanying repository rather than transcribed, so that published values cannot diverge from those the code produced. The commitment to report a null result as clearly as a positive one was recorded in the repository README before any real corpus was obtained.