RESPIRAHUB

Skrining TB Berbasis AI dari Suara Batuk

Laporan Trial Pertama

Model: Wav2Vec2-base Fine-tuned pada CODA TB Dataset
Dataset: 1.080 Pasien | 1.818 Segmen Audio | 7 Negara
Evaluasi: 10-Fold Stratified Cross-Validation

Februari 2026 | Dokumen Internal

Ringkasan Eksekutif

Kami baru saja menyelesaikan trial pertama model AI RespiraHub untuk mendeteksi tuberkulosis (TB) dari suara batuk. Hasilnya: **model berhasil membedakan batuk TB dari batuk non-TB dengan AUROC rata-rata 0,733.**

Angka ini memang masih di bawah target 0,85 yang dicapai studi Zambia (Lancet Digital Health, 2025). Tapi ini bukan kegagalan — ini adalah first run yang berhasil mengidentifikasi masalah teknis spesifik di pipeline data kami. Kami sudah tahu persis apa yang harus diperbaiki.

Dokumen ini menjelaskan apa yang kami lakukan, apa yang kami temukan, mengapa hasilnya seperti ini, dan langkah selanjutnya.

Apa yang Kami Uji?

Bayangkan seorang pasien datang ke puskesmas dengan keluhan batuk. Saat ini, tenaga kesehatan mengandalkan *anamnesis* (wawancara gejala) untuk memutuskan apakah pasien perlu tes TCM (Tes Cepat Molekuler). Masalahnya? Sensitivitas anamnesis hanya sekitar 42% menurut WHO — artinya lebih dari separuh kasus TB terlewat.

RespiraHub ingin menambahkan satu langkah sederhana: **rekam batuk pasien lewat smartphone, biarkan AI menganalisis, dan berikan rekomendasi apakah pasien perlu dirujuk ke TCM.**

Untuk menguji apakah ini mungkin, kami menggunakan dataset CODA TB — kumpulan rekaman batuk dari 1.080 pasien di 7 negara yang sudah dikonfirmasi status TB-nya melalui TCM. Kami melatih model AI (Wav2Vec2, model speech recognition dari Meta/Facebook) untuk belajar membedakan pola batuk TB+ versus TB-.

Hasil Trial Pertama

Kami mengevaluasi model menggunakan 10-fold cross-validation — artinya data dibagi 10, model dilatih pada 9 bagian dan diuji pada 1 bagian, berulang 10 kali. Setiap pasien hanya muncul di set uji tepat satu kali, sehingga tidak ada kebocoran data.

Performa per Fold

Fold	AUROC	Interpretasi	Train (seg)	Val (seg)
1	**0.7023**	Cukup	1.640	178
2	**0.7097**	Cukup	1.638	180
3	**0.8215**	Baik	1.635	183
4	**0.8590**	Sangat Baik	1.644	174
5	**0.6766**	Kurang	1.641	177
6	**0.7407**	Cukup	1.630	188

7	**0.7154**	Cukup	1.649	169
8	**0.6997**	Cukup	1.604	214
9	**0.6914**	Kurang	1.643	175
10	**0.7137**	Cukup	1.638	180
Rata-rata	**0.7330 ± 0.057**	**Di bawah target**	**~1.636**	**~182**

Perbandingan dengan Benchmark

Metode	AUROC	Catatan
Anamnesis (WHO)	~0.42 sensitivitas	Standar saat ini di puskesmas
Baseline MFCC + LR	~0.70	Sanity check (klasik)
RespiraHub Trial 1	**0.733**	First run, pipeline belum optimal
Zambia Study (target)	**0.852**	Lancet Digital Health, 2025

Mengapa Hasilnya Belum Optimal?

Setelah analisis mendalam, kami menemukan **satu akar masalah utama: jumlah data per pasien terlalu sedikit untuk proses agregasi yang efektif.**

Masalah: Terlalu Sedikit Segmen per Pasien

Setiap rekaman batuk dalam CODA TB berdurasi sekitar 0,5 detik. Model AI kami membutuhkan input audio sepanjang 3 detik. Untuk itu, kami menggabungkan beberapa rekaman batuk per pasien lalu memotongnya menjadi segmen 3 detik.

Masalahnya: banyak pasien yang hanya memiliki 2–3 rekaman batuk. Ketika digabungkan, totalnya hanya sekitar 1–1,5 detik — sisanya diisi dengan kesunyian (zero-padding). Artinya model belajar dari **kesunyian, bukan dari batuk.**

Metrik	Nilai
Rata-rata segmen per pasien	1,7
Median segmen per pasien	1,0
Pasien dengan hanya 1 segmen	**626 dari 1.080 (58%)**
Pasien dengan ≥3 segmen	182 dari 1.080 (17%)
Segmen minimum per pasien	1
Segmen maksimum per pasien	9

58% pasien hanya punya 1 segmen. Ini berarti model tidak bisa melakukan *soft voting* — teknik di mana beberapa prediksi dari pasien yang sama dirata-ratakan untuk menghasilkan keputusan yang lebih stabil. Dengan hanya 1 kesempatan prediksi per pasien, setiap kesalahan langsung berdampak pada AUROC keseluruhan.

Bukti: Fold dengan Lebih Banyak Segmen Lebih Baik

Fold 4 (AUROC tertinggi: 0,859) memiliki 174 segmen validasi untuk 108 pasien (rasio 1,6). Fold 3 (0,821) memiliki 183 segmen untuk 108 pasien (rasio 1,7). Meskipun perbedaannya kecil, fold-fold dengan performa terbaik konsisten memiliki komposisi pasien yang kebetulan memiliki lebih banyak rekaman batuk. Variance tinggi antar fold (0,057) juga menunjukkan bahwa hasil sangat bergantung pada komposisi data di setiap fold.

Rencana Perbaikan

Berdasarkan temuan di atas, kami akan menjalankan Trial Kedua dengan perbaikan berikut:

Perbaikan #1: Ubah Strategi Segmentasi

Sebelumnya: Gabungkan semua batuk per pasien → potong setiap 3 detik. Pasien dengan sedikit batuk menghasilkan segmen yang mayoritas berisi kesunyian.

Perbaikan: Setiap file batuk individual (0,5 detik) di-pad menjadi 3 detik. Satu file = satu segmen. Pasien dengan 8 file batuk sekarang memiliki 8 segmen (bukan 1–2 segmen seperti sebelumnya). Soft voting menjadi jauh lebih efektif.

Perbaikan #2: Turunkan Filter Hyfe

Sebelumnya: Hanya menggunakan file dengan Hyfe confidence score ≥0,8 (filter ketat). Banyak rekaman batuk yang terbuang.

Perbaikan: Turunkan threshold ke ≥0,5 untuk memasukkan lebih banyak data. Akan diuji apakah penurunan kualitas data terkompensasi oleh peningkatan kuantitas.

Perbaikan #3: Audio Augmentasi (jika diperlukan)

Jika setelah perbaikan #1 dan #2 AUROC masih di bawah 0,80, kami akan menambahkan augmentasi audio saat pelatihan: menambahkan noise latar, variasi kecepatan, dan reverb buatan. Ini melatih model untuk fokus pada pola batuk, bukan kondisi rekaman.

Kesimpulan

Trial pertama ini bukan kegagalan — ini adalah langkah diagnostik yang berhasil. Kami membuktikan bahwa model AI **bisa belajar membedakan batuk TB** (AUROC 0,733, di atas baseline 0,70 dan jauh di atas anamnesis 0,42). Yang perlu diperbaiki bukan modelnya, tapi cara kami menyiapkan data input.

Dengan perbaikan segmentasi yang sudah diidentifikasi, kami optimis Trial Kedua akan mendekati target AUROC 0,85 yang dicapai studi Zambia. Target ini realistis karena kami menggunakan dataset, model, dan metodologi yang sama — hanya pipeline preprocessing yang perlu dioptimalkan.

Konteks Penting untuk Tim

AUROC bukan sensitivitas. AUROC mengukur kemampuan model secara keseluruhan untuk membedakan dua kelas. Sensitivitas (kemampuan mendeteksi TB+) dan spesifisitas (kemampuan mengenali non-TB) adalah metrik terpisah yang ditentukan oleh pemilihan threshold. Target WHO: sensitivitas ≥90%, spesifisitas ≥70%.

Ini baru Phase 1A. Trial ini menggunakan data audio saja. Dalam roadmap lengkap, data anamnesis (gejala klinis) akan ditambahkan di fase berikutnya. Studi Zambia menunjukkan penambahan data klinis meningkatkan AUROC dari 0,852 menjadi 0,921.

Data dari luar Indonesia. CODA TB berisi rekaman dari 7 negara (mayoritas Afrika). Adaptasi ke konteks Indonesia akan dilakukan di Phase 2 melalui domain adaptation. Trial ini bertujuan memvalidasi bahwa teknologinya bekerja, bukan untuk langsung digunakan di lapangan.

Timeline Selanjutnya

Target Waktu	Aktivitas	Target Hasil
Minggu ini	Fix segmentasi + retrain (Trial 2)	AUROC ≥0,80
Minggu depan	Validasi pada data longitudinal	AUROC ≥0,80 (publishable)
Bulan depan	Mulai pengumpulan data batuk puskesmas	200+ rekaman unlabeled
2–3 bulan	Domain adaptation ke Indonesia	Model adapted

Dokumen ini bersifat internal. Disiapkan oleh Tim Engineering RespiraHub, Februari 2026.